NONPARAMETRIC LEAST SQUARES ESTIMATION IN INTEGER-VALUED GARCH MODELS

Dissertation
zur Erlangung des akademischen Grades
doctor rerum naturalium (Dr. rer. nat.)

vorgelegt dem Rat der Fakultät für Mathematik und Informatik der
Friedrich-Schiller-Universität Jena

von Maximilian Wechsung, M. Sc.,
geboren am 10. April 1990 in Berlin.

Jena, Oktober 2019

Gutachter:

1. Prof. Dr. Michael Neumann, Friedrich-Schiller-Universität Jena
2. Prof. Dr. Alexander Meister, Universität Rostock
3. Prof. Dr. Konstantinos Fokianos, Lancaster University

Tag der öffentlichen Verteidigung: 29. Oktober 2019

Contents

"I pass with relief from the tossing sea of Cause and Theory to the firm ground of Result and Fact."

Winston Churchill, The Story of the Malakand Field Force

Acknowledgments

Writing this thesis was a challenging endeavor which I could only accomplish with the help of many supporters. I want to express my sincere appreciation for Michael Neumann. With his enduring, patient support as my supervisor, and his illuminating advice as a teacher, he helped me navigate through Cause and Theory. That the want of firm ground let me sometimes tumble yet never fall is to a great extent owed to the support I received from my friends and family. I am especially grateful for the relationship with my girlfriend Selina Schmid, which is a constant source of love, joy, and inspiration. For her invaluable emotional support, a significant share of the credit for this work belongs to her.

List of symbols

$\mathbb{N}$	natural numbers, $\{0,1,2,\ldots\}$;
$\mathbb{N}_+$	positive integers, $\{1,2,3,\ldots\}$;
$\mathbb{R}^\infty$	$\mathbb{R}\times\mathbb{R}\times\mathbb{R}\times\ldots$;
$\mathbb{R}^n$	n-dimensional euclidean space;
$\mathscr{B}^n$	borelian σ-field on $\mathbb{R}^n$;
$\mathscr{B}(X)$	borelian σ-field on a metric space (X,d);
$\mathscr{B}^\infty$	cylindrical σ-field on $\mathbb{R}^\infty$, cf. Shiryayev (1984, page 144);
$x\wedge y$	minimum of two real numbers x and y;
$x\vee y$	maximum of two real numbers x and y;
$a_n \lesssim b_n$	$a_n \le C\,b_n$, for some $C>0$ independent of n;
$a_n = O(b_n)$	$\limsup_{n\to\infty}\left\lvert\frac{a_n}{b_n}\right\rvert<\infty$;
$a_n = o(b_n)$	$\limsup_{n\to\infty}\left\lvert\frac{a_n}{b_n}\right\rvert=0$;
$a_n \asymp b_n$	$0<\liminf_{n\to\infty}\left\lvert\frac{a_n}{b_n}\right\rvert\le\limsup_{n\to\infty}\left\lvert\frac{a_n}{b_n}\right\rvert<\infty$;
$\mathbb{E}_{\lvert\mathscr{F}}(X)$	$\mathbb{E}(X\,\lvert\,\mathscr{F})$, conditional expectation of the random variable X with respect to the sub σ-field $\mathscr{F}$, cf. Shiryayev (1984, page 211);
$\mathbb{E}_{\lvert Y=y}(X)$	$\mathbb{E}(X\,\lvert\,Y=y)$, conditional expectation of the random variable X with respect to the random variable Y, cf. Shiryayev (1984, page 218);
a.s. (P)	almost surely with respect to the probability measure P, in case of unambiguity, we just write “a.s.”.

Zusammenfassung

In dieser Arbeit betrachten wir ein Poisson-Regressionsmodell für Zähldaten. Angenommen, wir beobachten eine Reihe von Zähldaten, die jeweils bedingt auf die Information über ihre Vergangenheit einer Poissonverteilung folgen. Die Intensitäten dieser Verteilungen werden als nicht beobachtbar angenommen, jedoch unterstellen wir einen funktionalen Zusammenhang zwischen der Intensität zu einem Zeitpunkt und dem vorangegangenen Paar von Intensität und Zählbeobachtung. In der Fachliteratur wurden bisher parametrische Modelle dieser Art behandelt, beispielsweise das lineare INGARCH(1,1)-Modell oder das etwas komplizeritere log-lineare Modell. In diesen Fällen wurde für den Partiellen Maximum Likelihood Schätzer (partial maximum likelihood estimator) die Konvergenzrate $n^{-1/2}$ bewiesen.

Unser Ziel ist es, ein Modell für den bivariaten Prozess aus Zähldaten und Intensitäten zu betrachten, in dem die Regressionsfunktion nicht durch einen endlichdimensionalen Parameter identifiziert werden kann. Um in diesem nichtparametrischen Modell einen Ansatz zur Schätzung der Regressionsfunktion zu finden, müssen wir an diese eine Kontraktionsbedingung stellen. Davon ausgehend analysieren wir einen Kleinste-Quadrate-Ansatz, der in ähnlicher Form schon von Meister und Kreiß (2016) in einem verwandten Modell untersucht wurde. In unserer Analyse werden wir beweisen, dass der univariate Zählprozess gleichmäßig mischend ist. Diese Eigenschaft nutzen wir, um anschließend bekannte Resultate aus der klassischen Theorie der empirischen Prozesse anzuwenden. Typischerweise ist in dieser Art von Schätzproblemen die Größe der Klasse aller möglichen Regressionsfunktionen entscheidend für die Konvergenzrate eines Schätzers. Dieser Effekt ist auch im vorliegenden Modell zu beobachten.

Da der zunächst untersuchte Schätzer numerisch nicht besonders gut zugänglich ist, werden wir im zweiten Teil der Arbeit einen modifizierten sogenannten Näherungsweise Kleinste-Quadrate-Schätzer betrachten. Dieser wird auf seine asymptotische Güte hin theoretisch untersucht, wobei wir im Gegensatz zum ersten Teil Techniken aus der Theorie der Martingale verwenden werden. Der modifizierte Schätzer ist in der Tat numerisch gut zu approximieren. Zum abschluss der Arbeit illustrieren wir dieses Verfahren in einer Simulationsstudie.

Abstract

In this thesis we consider Poisson regression models for count data. Suppose we observe a time series of count variables. Given the information about the past, each count variable has a Poisson distribution with a random intensity. The time series of intensities is unobservable, but we impose a functional relationship between the current intensity and the preceding pair of intensity and count observation. In the literature some consideration has been given to parametric models of the linear INGARCH(1,1) type or more involved ones like the log linear model. In these cases $n^{-1/2}$-consistency of the partial maximum likelihood estimator has been proven.

Suppose that the relationship between a count variable and the respectively preceding pair of count and intensity variables is given by a link function that cannot be characterized by a finite-dimensional parameter. We call this model a nonparametric integer valued GARCH model. In order to obtain a suitable estimation equation in this nonparametric model, a contractive condition has to be imposed on the true link function. We analyze the rate of convergence of a least squares estimator that is inspired by the work of Meister and Kreiß (2016). We prove uniform mixing of the univariate count process and use the derived properties to apply some classical tools from empirical process theory. The size of the class of admissible functions determines the rate of convergence, which is a common property of nonparametric models.

Since this estimator is computationally rather impractical, we also analyze the behavior of an approximate least squares estimator. In contrast to the analysis of the first estimator, the examination of the approximate least squares estimator's asymptotic quality is based on the exploitation of martingale properties instead of mixing. The approximate least squares estimator is indeed computable, and we take the opportunity to conduct experiments to illustrate the proposed statistical procedure. An exposition of the experimental results will conclude this thesis.

Chapter 1

Introduction

The general theme of this thesis is the statistical analysis of time dependent count data. This kind of time series arises in many branches of empirical research. Typically invoked examples are epidemiological data, e.g. when the number of reported cases of a certain disease in a series of successive time intervals is counted. Further instances include data from meteorology (e.g. counting weather events) or social science (e.g. number of applicants for social security), to name but a few. We depict two examples of epidemiological time series of counts in Figure 1.0.1 at the and of this chapter.

The aim of the statistical analysis of such a data set is to find a description that captures its phenomenological main features. This includes observable effects of some explanatory variables as well as the inherent dynamics of the observed process of responses. We adopt the established perspective of seeking the best prediction for the next response variable given the complete information about the past (Kedem and Fokianos, 2002, page 5). This information includes past realizations of all explanatory variables, all response variables, and possible unobservable innovation processes. Let the symbol $\mathscr{F}_t$ denote this information at a given time t. The response variable at time t is called Y_t. The best prediction of Y_t given the information $\mathscr{F}_{t-1}$ is written $\mathbb{E}[Y_t|\mathscr{F}_{t-1}]$. The aim is to express $\mathbb{E}[Y_t|\mathscr{F}_{t-1}]$ as a function of all involved variables (explanatory, response, innovation). This can be achieved using a generalized linear regression model. In this context the usual approach is to assume that the marginal conditional distributions of the responses given the past information, $Y_t|\mathscr{F}_{t-1}$, are some exponential family distributions with natural parameters θ_t (Kedem and Fokianos, 2002, page 6). The natural parameter is a function of $\mathbb{E}[Y_t|\mathscr{F}_{t-1}]$ (ibid., page 8).

Having made the assumption regarding the marginal distribution, we consider two possibilities to introduce an evolutionary dynamic. Cox (1981) distinguished

two different approaches. In a *parameter driven* model, the parameter evolves independently of the past responses, $\theta_t = \tilde{m}(\theta_{t-1}, \tilde{\varepsilon}_{t-1})$. Here, $\{\tilde{\varepsilon}_t\}$ is an unobservable stationary innovation process that is assumed to be independent of past explanatory and response variables. The time dependence is entirely carried by the innovation process. In contrast, *observation driven* models impose a relation of the form $\theta_t = m(\boldsymbol{Z}_{t-1}, \varepsilon_{t-1})$, where $\boldsymbol{Z}_t$ is the sequence of past explanatory and response variables up to time t, and ε_t is some random innovation variable. The case $\varepsilon_t = \boldsymbol{\theta}_t := (\theta_t, \theta_{t-1}, \ldots)$, which leads to $\theta_t = m(\boldsymbol{Z}_{t-1}, \boldsymbol{\theta}_{t-1})$, resembles the definition of observation driven models given by Sim et al. (2016) and captures many contributions in the literature on time series of counts, e.g. Rydberg and Shephard (2000); Davis et al. (2003); Heinen (2003); Ferland et al. (2006); Jung and Tremayne (2011). We adopt this notion of observation driven models.

It has been acknowledged that parameter driven models have some disadvantages. Among others, Shephard (1996) and Davis et al. (2003) pointed out that estimation and forecasting in parameter driven models require substantially more effort than in observation driven models. This problem has received some attention, for example by Durbin and Koopman (1997, 2000). In contrast, observation driven models offer elegant solutions to capture the series' inherent dynamics. They are in line with established principles in the statistical modelling and analysis of time series (Shephard, 1996), and they are often expected to be more parsimonous (Neumann, 2011; Fokianos et al., 2009). For example, the classes of ARCH and GARCH models introduced by Engle (1982) and Bollerslev (1986), respectively, have become a popular tool in modeling econometric time series with changing volatility (Enders, 1995). Moreover, count data are often overdispersed, which is illustrated by our examples in Figure 1.0.1. Observation driven models can be used to properly account for this phenomenon while upholding the popular assumption of a Poisson distribution (Heinen, 2003). In this thesis we turn our focus to observation driven models for time series of counts.

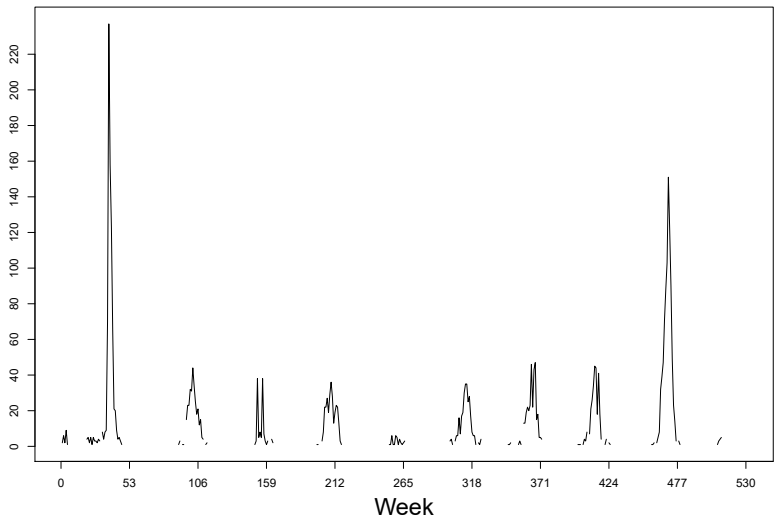
A common model for the marginal conditional distributions of a time series of counts is the Poisson distribution (Kedem and Fokianos, 2002, page 140). In this case we suppose that $Y_t|\mathscr{F}_{t-1}$ follows a Poisson distribution with intensity λ_t. This distribution belongs to the exponential family, and the natural parameter is $\theta_t = \log \lambda_t$ (ibid., page 142). As we seek an observation driven model, we set $\lambda_t = m(\boldsymbol{Z}_{t-1}, \boldsymbol{\lambda}_{t-1})$, where $\boldsymbol{\lambda}_{t-1}$ denotes the series $(\lambda_{t-1}, \lambda_{t-2}, \ldots)$ and m is some measurable function which we call *link function*. In order to focus on the inherent dynamics of the process, we disregard any potential explanatory variables and suppose that $m(\boldsymbol{Z}_{t-1}, \boldsymbol{\lambda}_{t-1}) = m(Y_{t-1}, \ldots, Y_{t-p}; \lambda_{t-1}, \ldots, \lambda_{t-q})$ for some $p, q \in \mathbb{N}_+$. As we assume a Poisson distribution for $Y_t|\mathscr{F}_{t-1}$, the marginal conditional variance at time t equals the conditional mean. Hence, the last model equation bears a striking similarity to the GARCH(p,q) model which is constituted by a pro-

cess $\{X_t\}$ with $X_t|\mathscr{F}_{t-1} \sim N(0,\sigma_t^2)$ and $\sigma_t^2 = \alpha_0 + \alpha_1 X_{t-1}^2 + \ldots \alpha_p X_{t-p}^2 + \beta_1 \sigma_{t-1}^2 + \ldots + \beta_q \sigma_{t-q}^2$. Therefore, our Poisson model could be called integer-valued GARCH(p,q) (INGARCH(p,q)) model. This term has been introduced by Ferland et al. (2006). They showed existence and stationarity of an INGARCH(p,q) process with linear link function. Without proof they postulated asymptotic normality of the conditional maximum likelihood estimator for the parameters. They illustrated their model with a data set of reported campylobacteriosis cases in Quebec. Rydberg and Shephard (2000) used the linear INGARCH(1,1) model in a financial market context to model the number of trades in a short time interval. They referred to it as BIN(1,1) model. A log-linear specification for m in the same model was analyzed by Davis et al. (2003). For the case $p = q = 1$ and a linear link function $m(y,\lambda) = \alpha_0 + \alpha_1 y + \beta_1 \lambda$, Fokianos et al. (2009) proved consistency and asymptotic normality of the conditional maximum likelihood estimator for $(\alpha_0, \alpha_1, \beta_1)'$. Later, Fokianos and Tjøstheim (2012) extended this result to functions of the form $m(y,\lambda) = f(\lambda) + g(y)$ for rather general functions f and g that are supposed to belong to some finite-dimensional class.

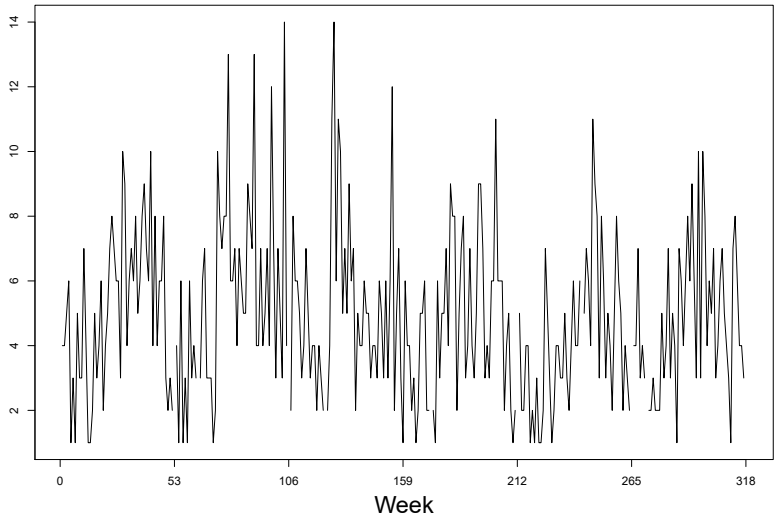
All of the mentioned contributions have in common that they assume a function m that can be identified by a finite-dimensional parameter. To our knowledge, the first one who considered a purely nonparametric INGARCH(1,1) model was Neumann (2011). He proved absolute regularity of the count process $\{Y_t\}$ in this model with the assumption that m satisfies a contraction property. This result was generalized by Doukhan and Neumann (2018). However, neither of the two contributions considered estimation of m. The aim of this thesis is to propose and analyze an estimation procedure for INGARCH(1,1) models with link functions m that cannot be described by a finite dimensional parameter. Meister and Kreiß (2016) already accomplished this goal for a nonparametric GARCH(1,1) model. The similarity between GARCH(1,1) and INGARCH(1,1) models allow us to adapt their underlying idea for an estimator. However, we pursue a slightly different strategy in the asymptotic analysis on the basis of the results of Neumann (2011). His contractive assumption will be essential to our approach.

In Chapter 1 we will formally introduce the stochastic process that constitutes our statistical model. We will state and partially prove the processes' main features that are essential for our statistical analysis. Chapter 2 is concerned with a formal definition of a nonparametric least squares estimator of m and its asymptotic analysis. In Chapter 3 we propose a modified estimation approach that requires less severe model assumptions and is more suitable for computation. This will be demonstrated in a simulation study concluding the thesis.

Figure 1.0.1: Two examples of epidemiological time series of count data. Both data sets show indications of overdispersion. Figure (a) shows the number of reported influenca infections in Berlin-Mitte on a weekly basis in the period ranging from the first week in 2009 to week 53 in 2018. The sample mean value is $\hat{\mu} = 15.79$ and the sample variance is $\hat{\sigma}^2 = 809.06$. Figure (b) shows the number of reported cases of campylobacteriosis in Berlin-Mitte on a weekly basis. In this case the measurements range from 2001 (week one) to 2018 (week 53). Sample mean: $\hat{\mu} = 4.91$, sample variance: $\hat{\sigma}^2 = 6.56$. Source: Robert Koch-Institut (2019)



(a) Reported influenza infections 2009–2018



(b) Reported cases of campylobacteriosis 2013–2018

Chapter 2

The model: definition and fundamental properties

2.1 Nonparametric INGARCH(1,1) processes: existence and first properties.

We start the formal examination of our statistical problem by defining the actual object of interest. As has become clear from the introduction, we consider a time series of counts and corresponding intensity parameters. Such an object can be formally defined as a bivariate stochastic process. The first component contains the parameter value, and the second one contains the counts.

DEFINITION 2.1.1. For $\mathbb{T} \in \{\mathbb{N}, \mathbb{Z}\}$, let $\{(\lambda_t, Y_t)\}_{t\in\mathbb{T}}$ be a stochastic process that is defined on a probability space $(\Omega, \mathscr{F}, \mathbb{P})$ and takes values in $\mathbb{R}\times\mathbb{N} = \mathbb{R}\times\{0,1,2,\ldots\}$. We define $\mathscr{F}_t := \sigma\{\lambda_s, Y_s : s \le t\}$ to be the σ-field generated by the process up to time t. The bivariate process $\{(\lambda_t, Y_t)\}_{t\in\mathbb{T}}$ is called a *nonparametric INGARCH(1,1)* process if there exists a $(\mathscr{B}\otimes 2^{\mathbb{N}} - \mathscr{B})$-measurable function $m : [0,\infty)\times\mathbb{N} \to [0,\infty)$ such that

$$\begin{aligned} \mathbb{P}^{Y_t|\mathscr{F}_{t-1}} &= Poiss(\lambda_t) \\ \lambda_t &= m(\lambda_{t-1}, Y_{t-1}). \end{aligned} \tag{IG}$$

The processes $\{Y_t\}$ and $\{\lambda_t\}$ are called *count process* and *intensity process* respectively. The function m is called *link function*.

REMARK 2.1.2. In the case of $\mathbb{T} = \mathbb{N}$, the non-parametric INGARCH(1,1) process is well defined. It can be constructed in the following way. Let π be a probability distribution on $\mathscr{B}\otimes 2^{\mathbb{N}}$. Suppose that for $B_k \in \mathscr{B}$, $k \in \mathbb{N}$, and $(\lambda, y) \in [0,\infty)\times\mathbb{N}$ the

function $P_0\big((\lambda,y);B\times\{k\}\big)$ is given by

$$P_0\big((\lambda,y);B\times\{k\}\big) := \mathbb{1}_{\{m(\lambda,y)\in B\}}\frac{\big(m(\lambda,y)\big)^k}{k!}e^{-m(\lambda,y)},$$

where m is some measurable function. This function is measurable for fixed $B\times\{k\}$. For every pair (y,λ), the function $P_0\big((\lambda,y);\cdot\big)$ can be extended to a measure $P\big((\lambda,y);\cdot\big)$ on the field of finite unions of disjoint rectangles,

$$\mathscr{C} := \Big\{\sum_{k=0}^{n}(B_k\times A_k)\colon n\in\mathbb{N},\, B_k\in\mathscr{B},\, A_k\in 2^{\mathbb{N}}\Big\},$$

such that for fixed A_k and B_k the function $P\big((\lambda,y);\sum_{k=0}^{n}(B_k\times A_k)\big)$ is measurable. For any (λ,y), the measure can be further extended uniquely to a measure $P\big((\lambda,y);\cdot\big)$ on the σ-field $\sigma(\mathscr{C})=\mathscr{B}\otimes 2^{\mathbb{N}}$ (Shorack, 2017, page 89). By Halmos' approximation theorem (Shorack, 2017, page 15), the function $P\big((\lambda,y);A\big)$ is measurable for fixed $A\in\mathscr{B}\otimes 2^{\mathbb{N}}$. It follows from general Markov chain theory that there exists a Markov chain with initial distribution π and transition kernel $P=\big\{P\big((\lambda,y);A\big)\colon(\lambda,y)\in[0,\infty)\times\mathbb{N},\, A\in\mathscr{B}\otimes 2^{\mathbb{N}}\big\}$ (Meyn and Tweedie, 2009, Theorem 3.4.1). This Markov chain is an example of a nonparametric INGARCH(1,1) process with index set $\mathbb{T}=\mathbb{N}$.

DEFINITION 2.1.3. For a subset $D\subset\mathbb{R}$, a function $m\colon D\times\mathbb{N}\to D$ is called *semi-contractive* if there exists a number $0\le\ell<1$ such that for all $\lambda_1,\lambda_2\in D$

$$\sup_{y\in\mathbb{N}}|m(\lambda_1,y)-m(\lambda_2,y)|\le\ell\,|\lambda_1-\lambda_2|. \tag{C_*}$$

The function m is called *contractive* if there exist numbers $L_1,L_2\ge 0$ such that $L_1+L_2<1$ and

$$|m(\lambda_1,y_1)-m(\lambda_2,y_2)|\le L_1|\lambda_1-\lambda_2|+L_2|y_1-y_2|, \tag{C^*}$$

for all $\lambda_1,\lambda_2\in D$ and all $y_1,y_2\in\mathbb{N}$.

Let the function $m\colon[0,M]\times\mathbb{N}\mapsto[0,M]$ be contractice or semi-contractive. Either property implies continuity of $\lambda\mapsto m(\lambda,y)$ for every $y\in\mathbb{N}$. Since the second variable is discrete, we conclude that the bivariate function $(\lambda,y)\mapsto m(\lambda,y)$ is $(\mathscr{B}\otimes 2^{\mathbb{N}}-\mathscr{B})$-measurable and therefore eligible to serve as a link function of a nonparametric INGARCH(1,1)-process.

Throughout the thesis, we assume to observe realizations $Y_0,\dots,Y_n$ of the count process of a nonparametric INGARCH(1,1) process with link function m. The link function is assumed to satisfy at least the semi-contractive condition (C_*) and to have bounded range. It is our goal to find an estimator $\hat{m}_n$ for m on the basis

of $Y_0, \dots, Y_n$. This means that we treat the intensities as hidden variables. The first question we have to answer before we proceed is whether there is a chance of success in this endeavor: is there a substantial loss of information arising from the fact that the intensities are hidden variables? From a probabilistic point of view, this comes down to examining the relation between the σ-fields $\sigma\{Y_s : s \leq t\}$ and $\mathscr{F}_t$. A first step towards the answer of that fundamental question is the observation that the information delivered by λ_t is determined entirely by the previous values of the bivariate process via the equation

$$\lambda_{t+1} = m(\lambda_t, Y_t).$$

This argument can be applied repeatedly, yielding in the second step

$$\lambda_{t+1} = m\big(m(\lambda_{t-1}, Y_{t-1}), Y_t\big).$$

In the current notation, displaying a k-fold application of this substitution argument would result in a rather clumsy expression. To circumvent these difficulties, we adopt the notation of Meister and Kreiß (2016). For a function $g : \mathbb{R} \times \mathbb{N} \to \mathbb{R}$ and an arbitrary natural number k, we define the function $g^{[k]} : \mathbb{R} \times \mathbb{N}^{k+1} \to \mathbb{R}$ by setting

$$\begin{aligned}
g^{[0]}(x, y_0) &:= g(x, y_0) \\
g^{[1]}(x, y_0, y_1) &:= g(g^{[0]}(x, y_0), y_1) = g(g(x, y_0), y_1) \\
g^{[2]}(x, y_0, y_1, y_2) &:= g(g^{[1]}(x, y_0, y_1), y_2) = g(g(g(x, y_0), y_1), y_2) \\
&\vdots \\
g^{[k]}(x, y_0, \dots, y_k) &:= g(g^{[k-1]}(x, y_0, \dots, y_{k-1}), y_k).
\end{aligned}$$

Using this notation to display a k fold repetition of the above stated substitution argument, we see that

$$\lambda_{t+1} = m^{[k]}(\lambda_{t-k}, Y_{t-k}, \dots, Y_t).$$

Hence, the complete information that is delivered by the current intensity λ_t could be recovered by an observation of an intensity variable λ_{t-k} at an arbitrary time in the past, together with all count variables $Y_{t-k}, \dots, Y_t$ observed between that very time point and the present. Moreover, the more steps k we go back into the past, the less information is contributed by the intensity variable λ_{t-k} compared to the information of $Y_{t-k}, \dots, Y_t$. This fact is owed to the semi-contractive property of the link function m. It can be technically demonstrated by a comparison between the value of λ_{t+1} and $m^{[k]}(0, Y_{t-k}, \dots, Y_t)$. The absolute difference of these two

values indicates how far off the true value of the current intensity we are if we entirely disregard the intensity variable λ_{t-k}. By a repeated application of the semi-contractive property (C_*) of m, we obtain

$$\begin{aligned}
&\left|\lambda_{t+1} - m^{[k]}(0, Y_{t-k}, \ldots, Y_t)\right| \\
&= \left|m^{[k]}(\lambda_{t-k}, Y_{t-k}, \ldots, Y_t) - m^{[k]}(0, Y_{t-k}, \ldots, Y_t)\right| \\
&= m\big(m^{[k-1]}(\lambda_{t-k}, Y_{t-k}, \ldots, Y_{t-1}), Y_t\big) - m\big(m^{[k-1]}(0, Y_{t-k}, \ldots, Y_{t-1}), Y_t\big) \\
&\leq \ell \left|m^{[k-1]}(\lambda_{t-k}, Y_{t-k}, \ldots, Y_{t-1}) - m^{[k-1]}(0, Y_{t-k}, \ldots, Y_{t-1})\right| \\
&\quad \vdots \\
&\leq \ell^k \left|m^{[0]}(\lambda_{t-k}, Y_{t-k}) - m^{[0]}(0, Y_{t-k})\right| \\
&\leq \ell^{k+1} |\lambda_{t-k} - 0|.
\end{aligned}$$

Looking $k+1$ steps back into the past, the share of information about λ_{t+1} that is exclusively carried by λ_{t-k}, and hence cannot be recovered by observing the counts from $t-k$ up to t, decreases geometrically in $k+1$. Heuristically, taking into account the entire past of the count process should allow us to ignore the intensity variables completely. In order to formulate this idea in a rigorous fashion, we need to operate with a process $\{(\lambda_t, Y_t)\}_{t\in\mathbb{Z}}$ that allows us to consider limits $t \to -\infty$. We have not yet addressed the question of existence of such a two-sided INGARCH(1,1) process. The first part of the next lemma will resolve this matter by establishing the existence of a stationary distribution. This is a result by Doukhan and Neumann (2018, Corollary 2.1). The second part is adopted from Theorem 3.1 in Neumann (2011). It states formally that the entire past of the count process carries the same amount of statistical information as the bivariate process. The proof of this claim relies on the just introduced successive substitution argument.

LEMMA 2.1.4. *(i) Suppose that the one-sided nonparametric INGARCH(1,1) process $\{(\lambda_t, Y_t)\}_{t\in\mathbb{N}}$ has a semi-contractive link function with range $[0, M]$ and domain $[0, M] \times \mathbb{N}$. Then the bivariate process is a time homogeneous Markov chain with a unique stationary distribution π. Moreover, let $x \in [0, M]$ be arbitrary and let the process start with $\lambda_0 = x$. Then the marginal distributions of the bivariate process converge weakly to the stationary distribution, $\mathbb{P}^{(\lambda_t, Y_t) \mid \lambda_0 = x} \rightsquigarrow \pi$, as $t \to \infty$.*

(ii) Let $m : [0, M] \times \mathbb{N} \to [0, M]$ be a semi-contractive function. There exists a two-sided nonparametric INGARCH(1,1) process $\{(\lambda_t, Y_t)\}_{t\in\mathbb{Z}}$ with link function m. As in Definition 2.1.1, let $\mathscr{F}_t$ be the σ-field generated by the bivariate process up to time t. Then, $\mathscr{F}_t = \sigma\{Y_s : s \leq t\}$.

Proof. (i) The time homogeneous Markov property follows directly from the model assumption $\lambda_t = m(\lambda_{t-1}, Y_{t-1})$. The transition functions have been established in Remark 2.1.2. They do not depend on the time index. The existence of a unique stationary distribution π is stated as Corollary 2.1 in Doukhan and Neumann (2018). The corollary is valid under some conditions (Doukhan and Neumann, 2018, page 5), which for the most part the authors explicitly verified for our model. The only unchecked condition is their so called "geometric drift condition" which requires that there exist positive constants $a < \infty$ and $\kappa < 1$ such that almost surely

$$\mathbb{E}[\lambda_t \,|\, \lambda_{t-1}] \leq \kappa\, \lambda_{t-1} + a\,.$$

This condition is satisfied if we choose $\kappa = \ell$ and $a = M$ since

$$\begin{aligned}
\mathbb{E}[\lambda_t \,|\, \lambda_{t-1}] &= \mathbb{E}[m(\lambda_{t-1}, Y_{t-1}) \,|\, \lambda_{t-1}] \\
&\leq \mathbb{E}\Big[\,|m(\lambda_{t-1}, Y_{t-1}) - m(0, Y_{t-1})|\,\Big|\,\lambda_{t-1}\Big] + \mathbb{E}[m(0, Y_{t-1}) \,|\, \lambda_{t-1}] \\
&\leq \mathbb{E}[\ell\, \lambda_{t-1} | \, \lambda_{t-1}] + M \\
&= \ell\, \lambda_{t-1} + M \quad \text{a.s.}\,.
\end{aligned}$$

Thus, the just cited corollary can be applied. The weak convergence of marginal distributions $\mathbb{P}^{(\lambda_t, Y_t) | \, \lambda_0 = x} \rightsquigarrow \pi$ is a subsidiary result in the proof of Corollary 2.1 in Doukhan and Neumann (2018, page 19).

(ii) The existence of a two-sided nonparametric INGARCH(1,1) process follows from the existence of a stationary distribution (Doukhan and Neumann, 2018, page 5). The formal argument consists of an application of Kolmogorov's existence theorem (Shorack, 2017, page 104) to the family of distributions

$$\begin{aligned}
&P_{n,\ldots,n+k-1}(B) := \\
&\iint \ldots \int \mathbb{1}_B\big((\lambda_n, y_n), \ldots, (\lambda_{n+k-1}, y_{n+k-1})\big)\, P\big((\lambda_{n+k-2}, y_{n+k-2}); d(\lambda_{n+k-1}, y_{n+k-1})\big) \\
&\qquad\qquad \ldots\, P\big((\lambda_n, y_n); d(\lambda_{n+1}, y_{n+1})\big)\, \pi\big(d(\lambda_n, y_n)\big)
\end{aligned}$$

on $\mathscr{B}^k$, where $n \in \mathbb{Z}$ and $k \in \mathbb{N}_+$. The transition functions $P\big((\lambda, y); A\big)$ are defined as in Remark 2.1.2, and π is the stationary distribution from part (i).

Of course, $\sigma\{Y_s : s \leq t\}$ is contained in $\mathscr{F}_t$ since all Y_s with $s \leq t$ are measurable with respect to $\mathscr{F}_t$. For the opposite inclusion it is left to show that all λ_s up to time t are measurable with respect to $\sigma\{Y_s : s \leq t\}$. For the proof of that fact we find for every λ_s a function ϕ_s that is $(\sigma\{Y_t : s \leq t\} - \mathscr{B})$-measurable such that $\lambda_s = \phi_s$. Let $s \leq t$ be given, and consider the sequence of random variables $\big\{m^{[k]}(0, Y_{s-k}, \ldots, Y_s)\big\}_{k \in \mathbb{N}}$. We will show that for every $\omega \in \Omega$ this sequence has a

limit that equals $\lambda_s(\omega)$. We recall that the previously introduced substitution argument yielded

$$\left|\lambda_{s+1}(\omega) - m^{[k]}(0, Y_{s-k}, \dots, Y_s)(\omega)\right| \leq \ell^{k+1} \left|\lambda_{s-k}(\omega)\right| \leq \ell^{k+1} M .$$

Since by assumption $\ell < 1$, we conclude that for any $\varepsilon > 0$ and all $\omega \in \Omega$ there exists a $K(\varepsilon, \omega)$ such that for all $k \geq K$

$$\left|\lambda_{s+1}(\omega) - m^{[k]}(0, Y_{s-k}, \dots, Y_s)(\omega)\right| < \varepsilon .$$

Hence, for every $\omega \in \Omega$, the limit $\phi_s(\omega) := \lim_{k\to\infty} m^{[k]}(0, Y_{s-k}, \dots, Y_s)(\omega)$ exists and equals $\lambda_{s+1}(\omega)$. For every $k \in \mathbb{N}$, the values $\{m^{[k]}(0, Y_{s-k}, \dots, Y_s)(\omega) \colon \omega \in \Omega\}$ define a function that is a composition of $\sigma\{Y_s : s \leq t\}$-measurable functions and as such measurable. Hence, the limiting function λ_{s+1} is a $\sigma\{Y_s : s \leq t\}$-measurable random variable. □

The first part of the preceding lemma tells us that, regardless of the starting point, the process eventually approaches the stationary regime. We take this as a justification to suppose later that our observations are generated by a strictly stationary process. When we call a process stationary, we refer to strict stationarity.

We learn from the second part of Lemma 2.1.4 that it is indeed possible to neglect the intensity process in the estimation procedure without suffering any loss of information, at least in theory. In practice, however, it is never possible to reach infinitely far into the past of the count process. Any estimation procedure relying solely on the count data will therefore sustain a structural error from the resulting loss of information. Asymptotically, in the sense of the sample size growing to infinity, we have a chance to lose this error if the number of steps we look back grows to infinity as well.

2.2 Uniform mixing of the count process

To carry out a profound asymptotic analysis of any estimator whatsoever in the above introduced model, we need some understanding concerning the long time behavior of the underlying data generating process. The statistical problem of estimating a parameter, finite or infinite dimensional, is best understood for the case of independent and identically distributed (i.i.d.) data. Thus, it is in our vital interest to quantify how far away our time series is from being an i.i.d. sequence. As we already established stationarity of the data generating process, we are now concerned with a quantification of the degree of interdependence of the

process. Taking to indexes $n+k>n$, we want to establish how much the variable (λ_{n+k}, Y_{n+k}) depends on (λ_n, Y_n). We can immediately mitigate the question by recalling our goal to find an estimator that exclusively uses the count data $\{Y_t\}$. Hence, for the analysis of our estimator we will only need information about the dependence structure of the count process. Therefore, our question is: how much does the variable Y_{n+k} depend on Y_n? Or even more general, how do the count variables $\{Y_{n+k}, Y_{n+k+1}, \ldots\}$ depend on the past counts up to time n, $\{\ldots, Y_{n-1}, Y_n\}$?

A classical approach to quantify stochastic dependence is the notion of mixing coefficients between σ-fields. There are several distinct definitions of mixing coefficients, the first of which was suggested by Rosenblatt (1956). Later, Volkonskii and Rozanov (1959) and Ibragimov (1962) contributed further definitions. For a survey of the literature and an in-depth treatment of the field of mixing conditions, we refer to the treatises by Bradley (2007) and Doukhan (1994). For our purposes, the notion of uniform or ϕ-mixing is well suited. It was introduced by Ibragimov (1962) (Bradley, 2007, page 69). For this thesis uniform mixing shall be defined according to a characterization that was proved by Bradley (2007, page 89). In this context, a sub σ-field $\mathscr{E} \subset \mathscr{F}$ is called separable if it is countably generated, i.e. there exists a finite or countable sequence of sets $A_1, A_2, A_3, \ldots \in \mathscr{E}$ such that $\mathscr{E} = \sigma\{A_1, A_2, \ldots\}$ (Bradley, 2007, page 9).

DEFINITION 2.2.1. Let $\mathscr{A}$ and $\mathscr{E}$ be sub σ-fields of $\mathscr{F}$. Assume that $\mathscr{E}$ is separable and that there exists a regular conditional distribution $\mathbb{P}(\cdot|\mathscr{A})$ on $\mathscr{E}$. The uniform mixing coefficient of $\mathscr{A}$ and $\mathscr{E}$ is defined as

$$\phi(\mathscr{A}, \mathscr{E}) := \operatorname{ess\,sup}\Big\{ \sup\big\{ |\mathbb{P}(B|\mathscr{A}) - \mathbb{P}(B)| : B \in \mathscr{E} \big\} \Big\}.$$

In the sequel the σ-field $\mathscr{E}$ will be generated by a $\mathscr{B}^\infty$-measurable random sequence on $(\Omega, \mathscr{F}, \mathbb{P})$, i.e. there exists a random element $\boldsymbol{Y} : (\Omega, \mathscr{F}) \to (\mathbb{R}^\infty, \mathscr{B}^\infty)$ such that $\mathscr{E} = \sigma(\boldsymbol{Y})$. We show that $\mathscr{B}^\infty$ is separable and conclude thence that the same is true for $\mathscr{E}$. Recall that $\mathscr{B}^\infty = \sigma(\mathscr{C})$ with $\mathscr{C} := \big\{ B_k \times \mathbb{R}^\infty : B_k \in \mathscr{B}^k,\ k \in \mathbb{N}_+ \big\}$. We show that also $\mathscr{B}^\infty = \sigma(\mathscr{I}_Q)$ with

$$\mathscr{I}_Q := \big\{ (a_1, b_1) \times \ldots \times (a_k, b_k) \times \mathbb{R}^\infty : a_i, b_i \in \mathbb{Q};\ k \in \mathbb{N}_+ \big\}.$$

Similarly to the idea presented by Shiryayev (1984, page 144), we define for any $k \in \mathbb{N}_+$

$$\mathscr{C}_{k,Q} := \big\{ A \subset \mathbb{R}^k : \{ x \in \mathbb{R}^\infty : (x_1, \ldots, x_k) \in A \} \in \sigma(\mathscr{I}_Q) \big\}.$$

Note that the system $\mathscr{C}_{k,Q}$ is a σ-field over $\mathbb{R}^k$. For any $k \in \mathbb{N}_+$, the rectangles of the form $(a_1, b_1) \times \ldots \times (a_k, b_k)$ with rational endpoints are contained in $\mathscr{C}_{k,Q}$.

But the system of those rectangles generates $\mathscr{B}^k$. Hence, $\mathscr{B}^k \subset \sigma(\mathscr{C}_{k,Q}) = \mathscr{C}_{k,Q}$, which means that any set of the form $B_k \times \mathbb{R}^\infty$ with $B_k \in \mathscr{B}^k$ is contained in $\sigma(\mathscr{I}_Q)$. In other words, $\mathscr{C} \subset \sigma(\mathscr{I}_Q)$, and therefore $\mathscr{B}^\infty \subset \sigma(\mathscr{I}_Q) \subset \mathscr{B}^\infty$. Since the system $\mathscr{I}_Q$ is countable, we have shown that $\mathscr{B}^\infty$ is countably generated. Furthermore, $\sigma(\boldsymbol{Y}) = \boldsymbol{Y}^{-1}(\sigma(\mathscr{I}_Q)) = \sigma\big(\boldsymbol{Y}^{-1}(\mathscr{I}_Q)\big)$ (Shorack, 2017, page 24). Accordingly, the σ-field $\mathscr{E}$ is separable as well.

On the basis of the notion of mixing coefficients for σ-fields, we define mixing coefficients for the count process. The preceding paragraph ensures that all involved σ-algebras are separable.

DEFINITION 2.2.2. Suppose that $\{(\lambda_t, Y_t)\}_{t\in\mathbb{Z}}$ is a two-sided nonparametric INGARCH(1,1) process. For the corresponding count process $\{Y_t\}_{t\in\mathbb{Z}}$ and an integer n, let $\sigma\{Y_t\colon t \le n\} =: \mathscr{F}^Y_{-\infty,n}$ denote the σ-field that is generated by the stochastic process $\{Y_t\colon t\in\mathbb{Z},\ t\le n\}$. Furthermore, for an integer n and a natural number k, we refer with $\mathscr{F}^Y_{n+k,\infty} := \sigma\{Y_t\colon t \ge n+k\}$ to the σ-field generated by the process $\{Y_t\colon t\in\mathbb{Z},\ t\ge n+k\}$. Then the kth uniform mixing coefficient of the count process at time $n\in\mathbb{Z}$ is defined as

$$\phi(k,n) := \phi(\mathscr{F}^Y_{-\infty,n}, \mathscr{F}^Y_{n+k,\infty}).$$

Furthermore, we set $\phi(k) := \sup_{n\in\mathbb{Z}} \phi(k,n)$. The count process is called uniformly mixing if $\lim_{k\to\infty}\phi(k) = 0$.

LEMMA 2.2.3. *Suppose that $\{(\lambda_t, Y_t)\}_{t\in\mathbb{Z}}$ is a nonparametric INGARCH(1,1) process. For any $n\in\mathbb{Z}$ and $k\in\mathbb{N}$, let $\boldsymbol{Q}_{k,n}$ denote the regular conditional distribution $\boldsymbol{Q}_{k,n}(\omega,\cdot) = \mathbb{P}^{(Y_{n+k},Y_{n+k+1},\ldots)\,|\,\lambda_n(\omega),Y_n(\omega)}$, and let $P_{k,n} = \mathbb{P}^{(Y_{n+k},Y_{n+k+1},\ldots)}$. For two measures μ and ν on $\mathscr{A}$ let their total variation distance be given by $d_{TV}(\mu,\nu) := \sup_{A\in\mathscr{A}} |\mu(A) - \nu(A)|$. The kth uniform mixing coefficient $\phi(k,n)$ of the count process $\{Y_t\}$ at time $n\in\mathbb{Z}$ is given by*

$$\phi(k,n) = \operatorname{ess\,sup}\ d_{TV}(\boldsymbol{Q}_{k,n}(\omega,\cdot), P_{k,n}).$$

Proof. Recall the result of Lemma 2.1.4 (ii) which yielded that $\mathscr{F}^Y_{-\infty,n} = \mathscr{F}_n$. Combining this result with the strong Markov property, we can conclude that

$$\begin{aligned}
\phi(k,n) &= \operatorname{ess\,sup}\Big\{ \sup_{B\in\mathscr{F}^Y_{n+k,\infty}} \big|\mathbb{P}(B|\mathscr{F}^Y_{-\infty,n}) - \mathbb{P}(B)\big|\Big\} \\
&= \operatorname{ess\,sup}\Big\{ \sup_{C\in\mathscr{B}^\infty} \Big|\mathbb{P}\Big\{(Y_{n+k},Y_{n+k+1},\ldots)\in C|\mathscr{F}^Y_{-\infty,n}\Big\} \\
&\qquad\qquad -\mathbb{P}\{(Y_{n+k},Y_{n+k+1},\ldots)\in C\}\Big|\Big\} \\
&= \operatorname{ess\,sup}\Big\{ \sup_{C\in\mathscr{B}^\infty} \big|\mathbb{P}\{(Y_{n+k},Y_{n+k+1},\ldots)\in C|\mathscr{F}_n\} \\
&\qquad\qquad -\mathbb{P}\{(Y_{n+k},Y_{n+k+1},\ldots)\in C\}\big|\Big\}
\end{aligned}$$

$$= \operatorname{ess\,sup} \Big\{ \sup_{C \in \mathscr{B}^\infty} \big| \mathbb{P}\{(Y_{n+k}, Y_{n+k+1}, \ldots) \in C | \lambda_n, Y_n\} - \mathbb{P}\{(Y_{n+k}, Y_{n+k+1}, \ldots) \in C\} \big| \Big\}$$

$$= \operatorname{ess\,sup}\ d_{TV}(Q_{k,n}(\omega, \cdot), P_{k,n}). \tag{2.1}$$

This concludes the proof. □

REMARK 2.2.4. If $\{(\lambda_t, Y_t)\}_{t \in \mathbb{Z}}$ is a stationary version of the process, we obtain for any $k \in \mathbb{N}$ and $n \in \mathbb{Z}$ that

$$Q_{k,n} = \mathbb{P}^{\{Y_t\}_{t \geq n+k} | \lambda_n, Y_n} = \frac{d\mathbb{P}^{\{Y_t\}_{t \geq n+k}, Y_n, \lambda_n}}{d\mathbb{P}^{\lambda_n, Y_n}} = \frac{d\mathbb{P}^{\{Y_t\}_{t \geq k}, Y_0, \lambda_0}}{d\mathbb{P}^{\lambda_0, Y_0}} = Q_{k,0} \text{ (a.s.)}.$$

Consequently, the mixing coefficients do not depend on the time index n, and we obtain

$$\sup_{n \in \mathbb{Z}} \phi(k, n) = \operatorname{ess\,sup} d_{TV}(Q_{k,0}(\omega, \cdot), P_{k,0}).$$

The remaining part of this section is devoted to proving uniform mixing of the count process and a geometric decay of the mixing coefficients. The assumption of stationarity is not necessary. We will use the techniques of Neumann (2011) and Doukhan and Neumann (2018), who used a coupling to obtain estimates for the total variation distance between $Q_{k,n}$ and $P_{k,n}$. We introduce the coupling technique as defined by Lindvall (1992).

DEFINITION 2.2.5. Let μ and ν be two probability measures on a measurable space $(X, \mathscr{E})$. A coupling of μ and ν is a pair of random elements $(Z, \check{Z})$ that is defined on a common probability space (S, Σ, P) and takes values in $(X \times X, \mathscr{E} \otimes \mathscr{E})$ with marginal distributions $P^Z = \mu$ and $P^{\check{Z}} = \nu$.

Given two measures μ and ν on a measurable space, we can construct a coupling to bound the total variation distance between μ and ν. This fact is known as the fundamental coupling inequality (Lindvall, 1992, page 11).

LEMMA 2.2.6. *Let $(S, \Sigma, P, (Z, \check{Z}))$ be a coupling of two distributions μ and ν on a measurable space $(X, \mathscr{E})$. Then the following inequality holds:*

$$d_{TV}(\mu, \nu) \leq P\{Z \neq \check{Z}\}.$$

Proof. The proof is taken from Lindvall (1992, page 11). For an arbitrary measurable set $B \in \mathscr{E}$, we find

$$\begin{aligned} |\mu(B) - \nu(B)| &= |P\{Z \in B\} - P\{\check{Z} \in B\}| \\ &= |P\{Z \in B, \check{Z} \in B\} + P\{Z \in B, \check{Z} \notin B\} \end{aligned}$$

$$
\begin{aligned}
&\quad -P\{\check{Z} \in B\,, Z \notin B\} - P\{\check{Z} \in B\,, Z \in B\}| \\
&= |P\{Z \in B\,, \check{Z} \notin B\} - P\{\check{Z} \in B\,, Z \notin B\}| \\
&\leq P\{Z \neq \check{Z}\,, Z \in B\} + P\{Z \neq \check{Z}\,, Z \notin B\} \\
&= P\{Z \neq \check{Z}\}\,.
\end{aligned}
$$

The assertion follows by taking the supremum over all sets $B \in \mathscr{E}$. □

The coupling inequality offers a way to bound the total variation distance between the measures $P_{k,n}$ and $Q_{k,n}(\omega,\cdot)$. For every $n \in \mathbb{Z}$, $k \in \mathbb{N}$, and $\omega \in \Omega$, we try to construct a coupling $\big(S^{(k,n)}, \Sigma^{(k,n)}, P^{(k,n)}, (Z^{(k,n)}, \check{Z}^{(k,n)})\big)(\omega)$, where $Z^{(k,n)}$ has distribution $P_{k,n}$ and $\check{Z}^{(k,n)}$ has distribution $Q_{k,n}(\omega,\cdot)$. The two random elements have to be constructed in a way that ensures that

$$
\lim_{k\to\infty} \sup_{n\in\mathbb{Z}} \operatorname*{ess\,sup}_{\omega\in\Omega} \Big(P^{(k,n)}\big\{\check{Z}^{(k,n)} \neq Z^{(k,n)}\big\}\Big)(\omega) = 0\,.
$$

By the coupling inequality and Lemma 2.2.3, we could then conclude that the count process is uniformly mixing.

This goal will be achieved in the following way. Suppose that (S,Σ,P) is a sufficiently rich probability space. For every $\omega \in \Omega$, we define two families of processes, $\big\{\{Z_t^{(n)}\}_{t\in\mathbb{N}} \colon n \in \mathbb{Z}\big\}(\omega)$ and $\big\{\{\check{Z}_t^{(n)}\}_{t\in\mathbb{N}} \colon n \in \mathbb{Z}\big\}(\omega)$, on (S,Σ,P) such that $\{Z_t^{(n)}\}_{t\geq k}(\omega) \sim P_{k,n}$ and $\{\check{Z}_t^{(n)}\}_{t\geq k}(\omega) \sim Q_{k,n}(\omega,\cdot)$. Inevitably, $\{\check{Z}_t^{(n)}\}$ depends on $\omega \in \Omega$ due to the conditioning on (λ_n, Y_n). Since the constructions of the processes $\{\check{Z}_t^{(n)}\}$ and $\{Z_t^{(n)}\}$ will be heavily intertwined, the latter will inherit the dependence on ω from the first. Then we set

$$
\big(S^{(k,n)}, \Sigma^{(k,n)}, P^{(k,n)}, (Z^{(k,n)}, \check{Z}^{(k,n)})\big)(\omega) := \big(S, \Sigma, P, (\{Z_t^{(n)}\}_{t\geq k}(\omega), \{\check{Z}_t^{(n)}\}_{t\geq k}(\omega)\big)\,.
$$

If the processes are constructed such that

$$
P\big\{\exists\, t \in \mathbb{N} \text{ such that } \check{Z}_{k+t}^{(n)} \neq Z_{k+t}^{(n)}\big\} \overset{k\to\infty}{\longrightarrow} 0
$$

uniformly in $n \in \mathbb{Z}$ and $\omega \in \Omega$, we have found a sequence of couplings with the desired properties.

THEOREM 2.2.7. *Let $\{(\lambda_t, Y_t)\}_{t\in\mathbb{Z}}$ be a nonparametric INGARCH(1,1) process with link function m. Assume that m has range $[0,M]$ and satisfies the strong contractive condition (C^*). Then the count process $\{Y_t\}$ is uniformly mixing, and the mixing coefficients decrease with a geometric rate:*

$$
\phi(k) = \sup_{n\in\mathbb{Z}} \phi(k,n) \lesssim (L_1 + L_2)^k\,.
$$

Proof. Let (S,Σ,P) be a probability space, and let s denote a generic element of S. On $S\times\Omega$ we will construct two families of real valued mappings, $\{(s,\omega)\mapsto Z_t^{(n)}(s,\omega)\colon t\in\mathbb{N}, n\in\mathbb{Z}\}$ and $\{(s,\omega)\mapsto \check{Z}_t^{(n)}(s,\omega)\colon t\in\mathbb{N}, n\in\mathbb{Z}\}$, such that for fixed $\omega\in\Omega$ the mappings $s\mapsto\{Z_t^{(n)}(s,\omega)\}_{t\in\mathbb{N}}$ and $s\mapsto\{\check{Z}_t^{(n)}(s,\omega)\}_{t\in\mathbb{N}}$ are stochastic processes on (S,Σ,P) with distributions $\{Z_t^{(n)}(\cdot,\omega)\}_{t\geq k}\sim P_{k,n}$ and $\{\check{Z}_t^{(n)}(\cdot,\omega)\}_{t\geq k}\sim Q_{k,n}(\omega,\cdot)$ for any $k\in\mathbb{N}$. The dependence on ω is inherited from the regular conditional distribution $Q_{k,n}(\omega,\cdot)=\mathbb{P}^{(Y_{n+k},Y_{n+k+1},\ldots)|\lambda_n(\omega),Y_n(\omega)}$. Technically, all following constructions depend on this ω. However, for the core argument ω will always stay fixed. For example, we are interested in the ω-section $s\mapsto\check{Z}_t^{(n)}(s,\omega)$ of $(s,\omega)\mapsto\check{Z}_t^{(n)}(s,\omega)$ as a random variable on (S,Σ,P). In order to ease the notation, we stipulate for the rest of the proof that whenever a function $(s,\omega)\mapsto X(s,\omega)$ on $S\times\Omega$ is written without arguments, we refer to its ω-section $s\mapsto X(s,\omega)$. When we want to emphasize dependence on ω, we write $X(\cdot,\omega)$. Let us now pursue with the construction of the coupling.

Let $\omega\in\Omega$ be fixed, and suppose that (S,Σ,P) is sufficiently rich. Let $(\tilde{\lambda},\tilde{Y})$ be a bivariate random variable on (S,Σ,P) that assumes values in $\mathbb{R}\times\mathbb{N}$ and is distributed according to π, the stationary distribution established in Lemma 2.1.4. Additionally, we assume that on the same space there are three independent sequences of independent, uniformly over $[0,1]$ distributed random variables, $\{U_t\}$, $\{V_t\}$, and $\{W_t\}$. Using these sequences, we will construct two mappings, $(s,\omega)\mapsto\{\tilde{\lambda}_t^{\pi},\tilde{Y}_t^{\pi}\}_{t\in\mathbb{N}}(s,\omega)$ and $(s,\omega)\mapsto\{\tilde{\lambda}_t^{x_n},\tilde{Y}_t^{x_n}\}_{t\in\mathbb{N}}(s,\omega)$, such that their ω-sections, $\{\tilde{\lambda}_t^{\pi},\tilde{Y}_t^{\pi}\}_{t\in\mathbb{N}}$ and $\{\tilde{\lambda}_t^{x_n},\tilde{Y}_t^{x_n}\}_{t\in\mathbb{N}}$, are nonparametric INGARCH(1,1) processes on (S,Σ,P) with link function m. The notation is intended to symbolize that the first process is initiated according to the stationary distribution π and the second one according to a singular point measure δ_{x_n}. The corresponding point is taken to be $x_n:=(\lambda_n(\omega),Y_n(\omega))$, where λ_n and Y_n are the very same variables appearing in the definition of the regular conditional distribution $Q_{k,n}(\omega,\cdot)$. More specifically, we define $(\tilde{\lambda}_0^{x_n},\tilde{Y}_0^{x_n}):=x_n$, $(\tilde{\lambda}_0^{\pi},\tilde{Y}_0^{\pi}):=(\tilde{\lambda},\tilde{Y})$, and

$$\begin{aligned}\tilde{\lambda}_1^{\pi}&:=m(\tilde{\lambda},\tilde{Y}),\\ \tilde{\lambda}_1^{x_n}&:=m(\lambda_n(\omega),Y_n(\omega)).\end{aligned}$$

It is worth noting that at this stage $\tilde{\lambda}_1^{\pi}$ does not depend on ω, whilst $\tilde{\lambda}_1^{x_n}$ does exclusively depend on ω. The construction of the first count variables, $\tilde{Y}_1^{\pi}$ and $\tilde{Y}_1^{x_n}$, is a bit more involved. Let us fix the notation F_λ for the cumulative distribution function of a Poisson distribution with intensity λ. It is common knowledge that for any random variable $U\sim Unif\,[0,1]$ the element $F_\lambda^{-1}(U)$ has a Poisson distribution with parameter λ. Here, F^{-1} denotes the generalized inverse of a

right continuous non-decreasing function,

$$F^{-1}(u) := \inf\{x \in \mathbb{R} \colon F(x) \geq u\}.$$

Using this fact, we define a Poisson random variable with random intensity $|\tilde{\lambda}_1^{\pi} - \tilde{\lambda}_1^{x_n}|$ by

$$s \mapsto \Delta_1^{(n)}(s,\omega) := F^{-1}_{|\tilde{\lambda}_1^{x_n}(\omega) - \tilde{\lambda}_1^{\pi}(s)|}(U_1(s)) \sim Poiss(|\tilde{\lambda}_1^{x_n}(\omega) - \tilde{\lambda}_1^{\pi}(s)|).$$

It is our goal to construct $\tilde{Y}_1^{x_n}(s,\omega)$ and $\tilde{Y}_1^{\pi}(s,\omega)$ on $S \times \Omega$ in such a way that their ω-sections satisfy our model assumption, $P^{\tilde{Y}_1^{x_n}} = Poiss(\tilde{\lambda}_1^{x_n})$ and $P^{\tilde{Y}_1^{\pi}|\lambda_1^{\pi}} = Poiss(\tilde{\lambda}_1^{\pi})$. Additionally, we seek for all $s \in S$ the relation $|\tilde{Y}_1^{x_n}(s,\omega) - \tilde{Y}_1^{\pi}(s,\omega)| = \Delta_1^{(n)}(s,\omega)$. This specific construction ensures that the absolute difference of their ω-sections, $|\tilde{Y}_1^{x_n} - \tilde{Y}_1^{\pi}|$, has a Poisson distribution with intensity $|\tilde{\lambda}_1^{\pi} - \tilde{\lambda}_1^{x_n}|$. This fact turns out to be essential in our argument, but it does not hold for general Poisson variables. We reach this goal in the following way by using the auxiliary Poisson variables $\eta_1^{x_n}$ and η_1^{π}. These are defined by

$$\eta_1^{\pi} := F^{-1}_{\tilde{\lambda}_1^{\pi}}(W_1), \qquad \eta_1^{x_n} := F^{-1}_{\tilde{\lambda}_1^{x_n}}(V_1).$$

Now we set,

$$\tilde{Y}_1^{x_n}(\cdot,\omega) := \begin{cases} \eta_1^{x_n} & \text{, if } \tilde{\lambda}_1^{x_n} < \tilde{\lambda}_1^{\pi} \\ \Delta_1^{(n)} + \eta_1^{\pi} & \text{, if } \tilde{\lambda}_1^{x_n} \geq \tilde{\lambda}_1^{\pi} \end{cases}$$

$$\tilde{Y}_1^{\pi}(\cdot,\omega) := \begin{cases} \Delta_1^{(n)} + \eta_1^{x_n} & \text{, if } \tilde{\lambda}_1^{x_n} < \tilde{\lambda}_1^{\pi} \\ \eta_1^{\pi} & \text{, if } \tilde{\lambda}_1^{x_n} \geq \tilde{\lambda}_1^{\pi}, \end{cases}$$

which gives us the desired property. All following elements of the process are now constructed by applying the link function to the current pairs of intensities and counts in order to obtain the next generation of intensities,

$$\tilde{\lambda}_{t+1}^{x_n}(\cdot,\omega) := m(\tilde{\lambda}_t^{x_n}, \tilde{\lambda}_t^{x_n}), \qquad \tilde{\lambda}_{t+1}^{\pi}(\cdot,\omega) := m(\tilde{\lambda}_t^{\pi}, \tilde{Y}_t^{\pi}).$$

Thereafter we generate the count variables using the same procedure that was used to generate $\tilde{Y}_1^{x_n}$ and $\tilde{Y}_1^{\pi}$. For any $t \in \mathbb{N}$, assume that the intensities $\tilde{\lambda}_t^{x_n}$ and $\tilde{\lambda}_t^{\pi}$ are given. Then

$$\Delta_t^{(n)}(\cdot,\omega) := F^{-1}_{|\tilde{\lambda}_t^{x_n} - \tilde{\lambda}_t^{\pi}|}(U_t), \qquad \eta_t^{x_n} := F^{-1}_{\tilde{\lambda}_t^{x_n}}(V_t), \qquad \eta_t^{\pi} := F^{-1}_{\tilde{\lambda}_t^{\pi}}(W_t),$$

$$\tilde{Y}_t^{x_n}(\cdot,\omega) := \begin{cases} \eta_t^{x_n} & \text{, if } \tilde{\lambda}_t^{x_n} < \tilde{\lambda}_t^{\pi} \\ \Delta_t^{(n)} + \eta_t^{\pi} & \text{, if } \tilde{\lambda}_t^{x_n} \geq \tilde{\lambda}_t^{\pi}, \end{cases}$$

$$\tilde{Y}_t^{\pi}(\cdot,\omega) := \begin{cases} \Delta_t^{(n)} + \eta_t^{x_n} & \text{, if } \tilde{\lambda}_t^{x_n} < \tilde{\lambda}_t^{\pi} \\ \eta_t^{\pi} & \text{, if } \tilde{\lambda}_t^{x_n} \geq \tilde{\lambda}_t^{\pi}. \end{cases}$$

Let us verify that the ω-sections $\{(\tilde{\lambda}_t^{\pi}, \tilde{Y}_t^{\pi})\}_{t\in\mathbb{N}}$ and $\{(\tilde{\lambda}_t^{x_n}, \tilde{Y}_t^{x_n})\}_{t\in\mathbb{N}}$ obey the characteristic relation (IG) of nonparametric INGARCH(1,1) processes stipulated in Definition 2.1.1. The second equation of (IG) is obviously satisfied by both processes. Concerning the marginal distributions, we observe that

$$\begin{aligned} P\{\tilde{Y}_t^{\pi} = k \,|\, \tilde{\lambda}_t^{\pi} = r, \tilde{\lambda}_t^{x_n} = s\} &= \mathbb{1}_{\{r>s\}} P\{\Delta_t + \eta_t^{x_n} = k \,|\, \tilde{\lambda}_t^{\pi} = r, \tilde{\lambda}_t^{x_n} = s\} \\ &\quad + \mathbb{1}_{\{r\leq s\}} P\{\eta_t^{\pi} = k \,|\, \tilde{\lambda}_t^{\pi} = r, \tilde{\lambda}_t^{x_n} = s\} \\ &= \mathbb{1}_{\{r>s\}} P\{F_{r-s}^{-1}(U_t) + F_s^{-1}(V_t) = k\} \\ &\quad + \mathbb{1}_{\{r\leq s\}} P\{F_r^{-1}(W_t) = k\} \\ &= \frac{r^k}{k!} e^{-r} \text{ a.s. } (P), \end{aligned}$$

where we used that $E[h(X,Y)|Y=y] = Eh(X,y)$ (a.s. in P) for independent random variables X and Y and measurable functions h, and furthermore that $F_{r-s}^{-1}(U_t) + F_s^{-1}(V_t) \sim Poiss(r)$ by the independence of U_t and V_t and the general properties of the Poisson distribution. For any $k \in \mathbb{N}$ and $B \in \mathscr{B}$, we conclude that

$$\begin{aligned} P\{\tilde{Y}_t^{\pi} = k\,, \tilde{\lambda}_t^{\pi} \in B\} &= P\{\tilde{Y}_t^{\pi} = k\,, \tilde{\lambda}_t^{\pi} \in B\,, \tilde{\lambda}_t^{x_n} \in \mathbb{R}\} \\ &= \int_{B\times\mathbb{R}} P\{\tilde{Y}_t^{\pi} = k \,|\, \tilde{\lambda}_t^{\pi} = r, \tilde{\lambda}_t^{x_n} = s\}\, P^{(\tilde{\lambda}_t^{\pi}, \tilde{\lambda}_t^{x_n})}(d(r,s)) \\ &= \int_B \int_{\mathbb{R}} \mathbb{P}\{\tilde{Y}_t^{\pi} = k \,|\, \tilde{\lambda}_t^{\pi} = r, \tilde{\lambda}_t^{x_n} = s\}\, P^{\tilde{\lambda}_t^{x_n}|\tilde{\lambda}_t^{\pi}=r}(ds) P^{\tilde{\lambda}_t^{\pi}}(dr) \\ &= \int_B \frac{r^k}{k!} e^{-r} \int_{\mathbb{R}} P^{\tilde{\lambda}_t^{x_n}|\tilde{\lambda}_t^{\pi}=r}(ds)\, P^{\tilde{\lambda}_t^{\pi}}(dr) \\ &= \int_B \frac{r^k}{k!} e^{-r}\, P^{\tilde{\lambda}_t^{\pi}}(dr), \end{aligned}$$

which means that the relation $P^{\tilde{Y}_t^{\pi}|\tilde{\lambda}_t^{\pi}} = Poiss(\tilde{\lambda}_n^{\pi})$ holds almost surely in P. Analogously we obtain that $P^{\tilde{Y}_t^{x_n}|\tilde{\lambda}_t^{x_n}} = Poiss(\tilde{\lambda}_t^{x_n})$ almost surely in P.

The processes $\{(\tilde{\lambda}_t^{\pi}, \tilde{Y}_t^{\pi})\}_{t\in\mathbb{N}}$ and $\{(\tilde{\lambda}_t^{x_n}, \tilde{Y}_t^{x_n})\}_{t\in\mathbb{N}}$ are indeed nonparametric INGARCH(1,1) processes on (S,Σ,P) with link function m. The first process started randomly according to the stationary bivariate distribution π. Given the realization $x_n = (\lambda_n(\omega), Y_n(\omega))$ from the original process on $(\Omega,\mathscr{A},\mathbb{P})$, the process $\{(\tilde{\lambda}_t^{x_n}, \tilde{Y}_t^{x_n})\}_{t\in\mathbb{N}}$ started at the fixed point $x_n = (\lambda_n(\omega), Y_n(\omega))$. If we define

$$Z_t^{(n)}(s,\omega) := \tilde{Y}_t^{\pi}(s,\omega), \qquad \check{Z}_t^{(n)}(s,\omega) := \tilde{Y}_t^{x_n}(s,\omega),$$

for any $n \in \mathbb{Z}$ and $t \in \mathbb{N}$, we can conclude that $\{Z_t^{(n)}\}_{t\geq k} \sim P_{k,n}$ and $\{\check{Z}_t^{(n)}\}_{t\geq k} \sim Q_{k,n}(\omega,\cdot)$. This means that we have successfully constructed a coupling of the distributions $P_{k,n}$ and $Q_{k,n}(\omega,\cdot)$.

Recall that $\Delta_t^{(n)} = |\tilde{Y}_t^{\pi} - \tilde{Y}_t^{x_n}| = |Z_t^{(n)} - \check{Z}_t^{(n)}|$. We turn to the task of bounding the probabilities

$$\begin{aligned} P\{s\colon \{Z_t^{(n)}(s,\omega)\}_{t\geq k} \neq \{\check{Z}_t^{(n)}(s,\omega)\}_{t\geq k}\} &= P\{s\colon \exists t \in \mathbb{N} \text{ such that } \Delta_{k+t}^{(n)}(s,\omega) > 0\} \\ &= 1 - P\{s\colon \Delta_{k+t}^{(n)}(s,\omega) = 0, \text{ for all } t \in \mathbb{N}\}, \end{aligned}$$

uniformly in $n \in \mathbb{Z}$ and in ω over a set with $\mathbb{P}$-measure one. We will do this for $k \geq 2$. Any information about the distribution of $\Delta_t^{(n)}$ is obtained conditionally on the past. Let us therefore introduce the σ-fields

$$\Sigma_t^{(n)} := \Sigma_t^{(n)}(\omega) := \sigma\{\tilde{Y}_r^{\pi}, \tilde{\lambda}_r^{\pi}, \tilde{Y}_r^{x_n}, \tilde{\lambda}_r^{x_n} : r \leq t\}.$$

The fact that $\tilde{\lambda}_{t+1}^{(\cdot)} = m(\tilde{\lambda}_t^{(\cdot)}, \tilde{Y}_t^{(\cdot)})$ and the Markov property imply that

$$P^{\Delta_{t+1}^{(n)}|\Sigma_t^{(n)}} = P^{\Delta_{t+1}^{(n)}|\tilde{Y}_t^{\pi},\tilde{\lambda}_t^{\pi},\tilde{Y}_t^{x_n},\tilde{\lambda}_t^{x_n}} \sim Poiss\big(|\tilde{\lambda}_{t+1}^{x_n} - \tilde{\lambda}_{t+1}^{\pi}|\big), \text{ a.s. } (P).$$

This relation and the contractive property of m let us conclude that

$$\begin{aligned} \mathbb{1}_{\{\Delta_k^{(n)}=0\}} P\{\Delta_{k+1}^{(n)} = 0 \,|\, \Sigma_k^{(n)}\} &= \mathbb{1}_{\{\Delta_k^{(n)}=0\}} \exp\big(-\big|\tilde{\lambda}_{k+1}^{\pi} - \tilde{\lambda}_{k+1}^{x_n}\big|\big) \\ &\geq \mathbb{1}_{\{\Delta_k^{(n)}=0\}} \exp\big(-\big[L_1|\tilde{\lambda}_k^{\pi} - \tilde{\lambda}_k^{x_n}| + L_2\,\Delta_k^{(n)}\big]\big) \qquad (2.2) \\ &= \mathbb{1}_{\{\Delta_k^{(n)}=0\}} \exp\big(-L_1|\tilde{\lambda}_k^{\pi} - \tilde{\lambda}_k^{x_n}|\big) \end{aligned}$$

almost surely in P. Consequently,

$$\mathbb{E}\Big[\mathbb{1}_{\{\Delta_k^{(n)}=0\}} \mathbb{1}_{\{\Delta_{k+1}^{(n)}=0\}} \,\Big|\, \Sigma_k^{(n)}\Big] \geq \mathbb{1}_{\{\Delta_k^{(n)}=0\}} \exp\big(-L_1|\tilde{\lambda}_k^{\pi} - \tilde{\lambda}_k^{x_n}|\big) \text{ a.s. } (P) \qquad (2.3)$$

for any $k \in \mathbb{N}$. Furthermore, using the observation that $\Sigma_k^{(n)} \subset \Sigma_{k+1}^{(n)} \subset \ldots \subset \Sigma_{k+t-1}^{(n)}$ ($t \geq 2$) and the fact that $\mathbb{1}_{\{\Delta_{k+i}^{(n)}=0\}}$ is measurable with respect to all $\Sigma_{k+j}^{(n)}$ with $j \geq i$, a repeated application of inequality (2.3) yields

$$\begin{aligned} &E\Big[\prod_{i=0}^{t} \mathbb{1}_{\{\Delta_{k+i}^{(n)}=0\}} \,\Big|\, \Sigma_0^{(n)}\Big] \\ &= E\Big[E\big[\mathbb{1}_{\{\Delta_{k+t}^{(n)}=0\}} \mathbb{1}_{\{\Delta_{k+t-1}^{(n)}=0\}} \,\big|\, \Sigma_{k+t-1}^{(n)}\big] \prod_{i=0}^{t-2} \mathbb{1}_{\{\Delta_{k+i}^{(n)}=0\}} \,\Big|\, \Sigma_0^{(n)}\Big] \\ &\geq E\Big[\exp\big(-L_1|\tilde{\lambda}_{k+t-1}^{\pi} - \tilde{\lambda}_{k+t-1}^{x_n}|\big)\, \mathbb{1}_{\{\Delta_{k+t-1}^{(n)}=0\}} \prod_{i=0}^{t-2} \mathbb{1}_{\{\Delta_{k+i}^{(n)}=0\}} \,\Big|\, \Sigma_0^{(n)}\Big] \\ &= E\Big[\exp\big(-L_1|\tilde{\lambda}_{k+t-1}^{\pi} - \tilde{\lambda}_{k+t-1}^{x_n}|\big) \end{aligned}$$

$$E\big[\mathbb{1}_{\{\Delta^{(n)}_{k+t-1}=0\}}\mathbb{1}_{\{\Delta^{(n)}_{k+t-2}=0\}}\,\big|\,\Sigma^{(n)}_{k+t-2}\big]\prod_{i=0}^{t-3}\mathbb{1}_{\{\Delta^{(n)}_{k+i}=0\}}\,\Big|\,\Sigma^{(n)}_0\Big]$$

$$\geq E\Big[\exp\big(-(L_1^2+L_1)|\tilde\lambda^{\pi}_{k+t-2}-\tilde\lambda^{x_n}_{k+t-2}|\big)\mathbb{1}_{\{\Delta^{(n)}_{k+t-2}=0\}}\prod_{i=0}^{t-3}\mathbb{1}_{\{\Delta^{(n)}_{k+i}=0\}}\,\Big|\,\Sigma^{(n)}_0\Big]\ \text{ a.s. } (P).$$

To obtain the last inequality, we used additionally to (2.3) the fact that for any $s\in\{s\colon \Delta^{(n)}_{k+t-2}(s)=0\}$ the contraction property yields

$$\exp\big(-L_1|\tilde\lambda^{\pi}_{k+t-1}-\tilde\lambda^{x_n}_{k+t-1}|\big)\geq\exp\big(-L_1^2|\tilde\lambda^{\pi}_{k+t-2}-\tilde\lambda^{x_n}_{k+t-2}|\big),$$

as in (2.2). Repeating this argument and making use of Jensen's inequality, we see that

$$\begin{aligned}
E\left[\prod_{i=0}^{t}\mathbb{1}_{\{\Delta^{(n)}_{k+i}=0\}}\,\Big|\,\Sigma^{(n)}_0\right]&\geq E\left[\exp\Big(-\sum_{i=1}^{t}L_1^i|\tilde\lambda^{\pi}_k-\tilde\lambda^{x_n}_k|\Big)\mathbb{1}_{\{\Delta^{(n)}_k=0\}}\,\Big|\,\Sigma^{(n)}_0\right]\\
&=E\left[\exp\Big(-|\tilde\lambda^{\pi}_k-\tilde\lambda^{x_n}_k|\sum_{i=1}^{t}L_1^i\Big)E\big[\mathbb{1}_{\{\Delta^{(n)}_k=0\}}|\,\Sigma^{(n)}_{k-1}\big]\,\Big|\,\Sigma^{(n)}_0\right] \qquad (2.4)\\
&=E\Big[\exp\Big(-|\tilde\lambda^{\pi}_k-\tilde\lambda^{x_n}_k|\sum_{i=0}^{t}L_1^i\Big)\,\Big|\,\Sigma^{(n)}_0\Big]\\
&\geq\exp\Big(-E\big[|\tilde\lambda^{\pi}_k-\tilde\lambda^{x_n}_k|\,\big|\,\Sigma^{(n)}_0\big]\sum_{i=0}^{t}L_1^i\Big)\ \text{ a.s. } (P) \qquad (2.5)
\end{aligned}$$

if $k\geq 1$. In equation (2.4) we used that $\tilde\lambda^{(\cdot)}_k=m(\tilde\lambda^{(\cdot)}_{k-1},\tilde Y^{(\cdot)}_{k-1})$ is $\Sigma^{(n)}_{k-1}$-measurable. There is an intuitive illustration for the preceding estimate. Suppose the two count processes $\{\tilde Y^{\pi}_t\}$ and $\{\tilde Y^{x_n}_t\}$ have coincided at time k. The probability that they perpetually stay together is large if the corresponding intensities $\tilde\lambda^{\pi}_k$ and $\tilde\lambda^{x_n}_k$ at the time of coincidence of the counts are close together.

As it turns out, the intensity processes attract each other as time pasts, irrespectively of their initial values. This is a consequence of the full contraction property of m. From $\Sigma^{(n)}_0\subset\Sigma^{(n)}_1\subset\ldots\subset\Sigma^{(n)}_{k-2}$, we obtain for any $n\in\mathbb{Z}$ that

$$\begin{aligned}
&E\big[|\tilde\lambda^{\pi}_k-\tilde\lambda^{x_n}_k|\,\big|\,\Sigma^{(n)}_0\big]\\
&\leq E\Big[E\big[L_1|\tilde\lambda^{\pi}_{k-1}-\tilde\lambda^{x_n}_{k-1}|+L_2\Delta^{(n)}_{k-1}\,\big|\,\Sigma^{(n)}_{k-2}\big]\,\Big|\,\Sigma^{(n)}_0\Big]\\
&=E\Big[L_1|\tilde\lambda^{\pi}_{k-1}-\tilde\lambda^{x_n}_{k-1}|+L_2E\big[\Delta^{(n)}_{k-1}\,\big|\,\Sigma^{(n)}_{k-2}\big]\,\Big|\,\Sigma^{(n)}_0\Big]\\
&=E\big[(L_1+L_2)|\tilde\lambda^{\pi}_{k-1}-\tilde\lambda^{x_n}_{k-1}|\,\big|\,\Sigma^{(n)}_0\big]\\
&\ \ \vdots\\
&\leq(L_1+L_2)^k\operatorname{ess\,sup}|\tilde\lambda^{\pi}_0-\tilde\lambda^{x_n}_0|\\
&\leq(L_1+L_2)^kM\ \text{ a.s. } (P)
\end{aligned}$$

since $\sup_{n,\omega}\tilde\lambda^{x_n}_0=\sup_{n,\omega}\lambda_n(\omega)=\sup_{n,\omega}m(\lambda_{n-1}(\omega),Y_{n-1}(\omega))\leq M$ and $\tilde\lambda^{\pi}_0=m(\tilde\lambda,\tilde Y)\leq$

M. Hence, (2.5) yields that

$$\inf_{n\in\mathbb{Z}} E\left[\prod_{i=0}^{t} \mathbb{1}_{\{\Delta_{k+i}^{(n)}=0\}} \,\middle|\, \Sigma_0^{(n)}\right] \geq \exp\Big(-(L_1+L_2)^k M\, \frac{1-L_1^{t+1}}{1-L_1}\Big) \text{ a.s. } (P),$$

and this bound does not depend on ω. Putting everything together, we arrive at the following estimate that holds uniformly in $n\in\mathbb{Z}$ and $\omega\in\Omega$:

$$\begin{aligned} P\Big\{s\colon \Delta_{k+t}^{(n)}(s,\omega)=0, \text{ for all } t\in\mathbb{N}\Big\} &= E\Big[\lim_{t\to\infty}\prod_{i=0}^{t} \mathbb{1}_{\{s\colon \Delta_{k+i}^{(n)}(s,\omega)=0\}}\Big] \\ &= \lim_{t\to\infty} E\Big[\prod_{i=0}^{t} \mathbb{1}_{\{s\colon \Delta_{k+i}^{(n)}(s,\omega)=0\}}\Big] \qquad (2.6) \\ &\geq \lim_{t\to\infty} \exp\Big(-(L_1+L_2)^k M \frac{1-L_1^{t+1}}{1-L_1}\Big) \\ &\geq 1-(L_1+L_2)^k\, \frac{M}{1-L_1}. \end{aligned}$$

In equation (2.6) we used the fact that $\big\{\prod_{i=0}^{t} \mathbb{1}_{\{\Delta_{k+i}=0\}}\big\}_{t\geq 1}$ is a monotonically decreasing sequence with an integrable first element to apply the dominated convergence theorem in order to interchange limit and expectation. After passing to the limit $t\to\infty$, we exploited the relation $e^{-x}\geq 1-x$ for $x\geq 0$.

Now we use the characterization $\phi(k)=\sup_{n\in\mathbb{Z}} \operatorname{ess\,sup} d_{TV}(Q_{k,n},P_{k,n})$ from Lemma 2.2.3 and subsequently apply the coupling inequality with respect to the constructed coupling $\big(S,P,\Sigma,(\{Z_t^{(n)}\}_{t\geq k},\{\check{Z}_t^{(n)}\}_{t\geq k})\big)$. This yields,

$$\begin{aligned} \phi(k) &= \sup_{n\in\mathbb{Z}} \operatorname*{ess\,sup}_{\omega\in\Omega}\ d_{TV}(Q_{k,n},P_{k,n}) \\ &\leq \sup_{n\in\mathbb{Z}} \operatorname*{ess\,sup}_{\omega\in\Omega}\ P\big\{s\colon \{Z_t^{(n)}(s,\omega)\}_{t\geq k}\neq\{\check{Z}_t^{(n)}(s,\omega)\}_{t\geq k}\big\} \\ &= \sup_{n\in\mathbb{Z}} \operatorname*{ess\,sup}_{\omega\in\Omega}\ \big(1-P\big\{s\colon \Delta_{k+t}^{(n)}(s,\omega)=0, \text{ for all } t\in\mathbb{N}\big\}\big) \\ &\leq \frac{M}{1-L_1}(L_1+L_2)^k. \end{aligned}$$

The proof is complete. □

For a technical reason that will reveal itself later in this thesis, we are also interested in the mixing properties of a slightly different process.

DEFINITION 2.2.8. For a natural number t, we define the sequence of $\mathbb{R}^{t+3}$-valued random vectors $\{\boldsymbol{Y}_{n-t}^{n+1}\colon n\in\mathbb{Z}\}$ by the assignment $\boldsymbol{Y}_{n-t}^{n+1}:=(0,Y_{n-t},\ldots,Y_{n+1})$, and denote with $\phi^t(k,n):=\phi\big(\sigma\{\boldsymbol{Y}_{i-t}^{i+1}\colon i\leq n\},\sigma\{\boldsymbol{Y}_{i-t}^{i+1}\colon i\geq n+k\}\big)$ and $\phi^t(k)=\sup_n \phi^t(k,n)$ the corresponding kth mixing coefficients. As in Definition 2.2.2, all involved σ-fields are separable.

LEMMA 2.2.9. *Under the assumptions of Theorem 2.2.7, the process* $\{\boldsymbol{Y}_{n-t}^{n+1}\}_{n\in\mathbb{Z}}$ *with lag t is uniformly mixing, and the mixing coefficients* $\phi^t(k)$ *are geometrically decreasing:*

$$\phi^t(k) \lesssim (L_1+L_2)^{k-t}.$$

Proof. The proof works with the same methodology as the proof of the preceding theorem. Just observe that $\left(\tilde{Y}_{r-t}^{\pi},\ldots,\tilde{Y}_{r+1}^{\pi}\right) = \left(\tilde{Y}_{r-t}^{x_n},\ldots,\tilde{Y}_{r+1}^{x_n}\right)$ for all $r \geq k$ if and only if $\tilde{Y}_r^{\pi} = \tilde{Y}_r^{x_n}$ for all $r \geq k-t$, and conclude that

$$P\left\{(\check{Z}_{r-t},\ldots,\check{Z}_{r+1}) = (Z_{r-t},\ldots,Z_{r+1}),\ \text{for all } r \geq k\right\} = P\left\{\check{Z}_r = Z_r,\ \text{for all } r \geq k-t\right\}.$$

The proof then proceeds along the lines of the previous one. □

Chapter 3

Nonparametric inference on contractive link functions

3.1 Definition of the estimator

Before we suggest a particular choice for an estimator, we submit a slight yet significant simplification of the model in terms of additional shape restrictions that we impose on the true link function m. We introduce the new model parameter $B \in \mathbb{N}_+$ and assume that for any $\lambda \in [0,M]$ the function $y \mapsto m(\lambda, y)$ is constant for all $y \geq B-1$. Formally, the domain of estimation will be $[0,M] \times \mathbb{N}$, but apparently no additional effort at all is needed for the area beyond $B-1$. We introduce this alleviation of the theoretical investigation at the peril of the statistician who eventually wishes to apply our model, as we increase the risk of model misspecification. Assume B is chosen too small, i.e. the true function is constant only from a value $B^{(m)} > B$ onward. In this case our model lacks explanatory power in the area $[0,M] \times \{B, \dots, B^{(m)}\}$. If a fraction of observations falls into this area, we are confronted with a structural estimation error. At the end of this chapter, we will give some additional remarks on this issue.

DEFINITION 3.1.1. For fixed constants $M > 0$, $B \in \mathbb{N}_+$ and $0 < L_1, L_2 < 1$, let $\mathscr{G} = \mathscr{G}(M,B,L_1,L_2)$ be the class of all functions $g\colon [0,M] \times \mathbb{N} \to [0,M]$ that satisfy the strong contraction condition (C^*) and furthermore $g(\lambda, y) = g(\lambda, B-1)$ for all $y \geq B-1$ and all λ. Let $\{(\lambda_n, Y_n)\}_{n \in \mathbb{Z}}$ be a stationary version of a two-sided nonparametric INGARCH(1,1) process with link function $m \in \mathscr{G}$. We call this process the data generating process and use the notation $\mathscr{F}_n := \sigma\{(\lambda_k, Y_k)\colon k \leq n\}$.

The main benefit of the model restriction is that the estimation procedure in the new model is only with respect to the first component purely nonparametric,

because any function $m \in \mathscr{G}$ can now be viewed as a function

$$m : [0,M] \times \{0,\dots,B-1\} \to [0,M].$$

This offers the even simpler interpretation of m as a vector valued function,

$$m = (m_0,\dots,m_{B-1})' : [0,M]^B \to [0,M].$$

As a result of keeping B fixed, each component m_i of m can be estimated on the basis of a number of observations that is asymptotically bounded from below by n. In nonparametric regression models with first-degree smoothness and d-dimensional explanatory variables, the typical mini-max rate of convergence is $n^{-1/(2+d)}$. For the estimation of a function $m \in \mathscr{G}$, we hope to obtain a typical one-dimensional nonparametric rate of convergence $n^{-1/3}$ as opposed to the two-dimensional rate $n^{-1/4}$ that we expect for $B = \infty$.

Let us now come to the central part of this section. In order to estimate the link function m, we use an idea that was proposed by Meister and Kreiß (2016) in the context of a nonparametric GARCH(1,1) model. A reasonable approach would be to search for a function $g \in \mathscr{G}$ that minimizes the L_2 prediction error with respect to the distribution of the process,

$$\mathbb{E}\big(Y_{t+1} - g(\lambda_t, Y_t)\big)^2 \to \min_{g \in \mathscr{G}}.$$

The unique minimizer of that functional is the conditional expectation, for which we have the relation

$$\begin{aligned}\mathbb{E}\big[Y_{t+1} \,\big|\, \lambda_t, Y_t\big] &= \mathbb{E}\Big[\mathbb{E}\big[Y_{t+1} \,\big|\, \mathscr{F}_t\big] \,\Big|\, \lambda_t, Y_t\Big] \\ &= \mathbb{E}\big[\lambda_{t+1} \,\big|\, \lambda_t, Y_t\big] \\ &= m(\lambda_t, Y_t), \ \text{ a.s.}.\end{aligned}$$

From this equation we see that the procedure of minimizing the prediction error returns exactly the true link function. The problem with calculating the conditional expectation of Y_{t+1} given λ_t and Y_t is that we are by assumption not able to observe the intensity λ_t. Let us therefore try to project this conditional expectation onto the information that is available. Suppose we observe the count variables $Y_0,\dots,Y_t$. Since

$$\lambda_{t+1} = m(\lambda_t, Y_t) = m^{[t]}(\lambda_0, Y_0, \dots, Y_t),$$

we see that

$$\mathbb{E}\big[Y_{t+1}\,\big|\,\lambda_t, Y_t\big] = m^{[t]}(\lambda_0, Y_0, \ldots, Y_t), \quad \text{a.s.}$$

and conclude that the loss of information due the projection onto the information given by $Y_0, \ldots, Y_t$ is small: with probability one

$$\begin{aligned}
&\Big|\mathbb{E}\big[Y_{t+1}\,\big|\,\lambda_t, Y_t\big] - \mathbb{E}\Big[\mathbb{E}\big[Y_{t+1}\,\big|\,\lambda_t, Y_t\big]\,\Big|\,Y_0, \ldots, Y_t\Big]\\
&= \Big|m^{[t]}(\lambda_0, Y_0, \ldots, Y_t) - \mathbb{E}\big[m^{[t]}(\lambda_0, Y_0, \ldots, Y_t)\,\big|\,Y_0, \ldots, Y_t\big]\Big|\\
&= \Big|m^{[t]}(\lambda_0, Y_0, \ldots, Y_t) - m^{[t]}(0, Y_0, \ldots, Y_t)\\
&\qquad\qquad + \mathbb{E}\big[m^{[t]}(0, Y_0, \ldots, Y_t) - m^{[t]}(\lambda_0, Y_0, \ldots, Y_t)\,\big|\,Y_0, \ldots, Y_t\big]\Big|\\
&\le L_1^t|\lambda_0| + \mathbb{E}\big[|m^{[t]}(0, Y_0, \ldots, Y_t) - m^{[t]}(\lambda_0, Y_0, \ldots, Y_t)|\,\big|\,Y_0, \ldots, Y_t\big]\\
&\le 2M L_1^t.
\end{aligned}$$

Together with the fact that

$$\begin{aligned}
\mathbb{E}\big[Y_{t+1}\,\big|\,Y_0, \ldots, Y_t\big] &= \mathbb{E}\big[\mathbb{E}[Y_{t+1}|\mathscr{F}_t]\,\big|\,Y_0, \ldots, Y_t\big]\\
&= \mathbb{E}\big[\lambda_{t+1}\,\big|\,Y_0, \ldots, Y_t\big]\\
&= \mathbb{E}\big[m(\lambda_t, Y_t)\,\big|\,Y_0, \ldots, Y_t\big]\\
&= \mathbb{E}\Big[\mathbb{E}\big[Y_{t+1}\,\big|\,\lambda_t, Y_t\big]\,\Big|\,Y_0, \ldots, Y_t\Big], \quad \text{a.s.},
\end{aligned}$$

this means that if we approximate $\mathbb{E}[Y_{t+1}\,|\,Y_0, \ldots, Y_t]$ instead of $\mathbb{E}[Y_{t+1}|\lambda_t, Y_t]$, we are still close to the true function m, and the structural error due to the loss of information converges to zero with the rate L_1^t as $t \to \infty$. As the conditional expectation $\mathbb{E}[Y_{t+1}|Y_0, \ldots, Y_t]$ minimizes

$$\mathbb{E}\big(Y_{t+1} - h(Y_0, \ldots, Y_t)\big)^2 \tag{3.1}$$

over the class of all measurable functions $h\colon \mathbb{N}^{t+1} \to \mathbb{R}$, we approximate this minimizer by choosing $\hat{h}$ that minimizes an empirical approximation of (3.1). To construct this approximation, we need a further sample $Y_{t+1}, \ldots, Y_n$; $n \in \mathbb{N}_+$. Recalling the uniform mixing property of the lagged process $\{(Y_{n-t}, \ldots, Y_{n+1})\}_{n\in\mathbb{Z}}$, which implies ergodicity (Bradley, 2007, Remark 2.6 and Proposition 2.8 in combination with Proposition 3.11), a reasonable approximation of (3.1) might be

$$\frac{1}{n-t}\sum_{i=t}^{n-1}\big(Y_{i+1} - h(Y_{i-t}, \ldots, Y_i)\big)^2.$$

Since

$$\begin{aligned}
&\left|\mathbb{E}\left[Y_{t+1}\middle|Y_0,\dots,Y_t\right] - m^{[t]}(0,Y_0,\dots,Y_t)\right| \\
&\leq \mathbb{E}\left[\left|m^{[t]}(\lambda_0,Y_0,\dots,Y_t) - m^{[t]}(0,Y_0,\dots,Y_t)\right|\middle|Y_0,\dots,Y_t\right] \\
&= O(L_1^t),
\end{aligned}$$

we know that the function h that we intend to approximate is very close to $(0,y_0,\dots,y_t) \mapsto m^{[t]}(0,y_0,\dots,y_t)$ at the points of measurement. Hence, we may as well restrict the class of candidate functions to functions of that shape, i.e. $\left\{g^{[t]} : \ g \in \mathscr{G}\right\}$. This leads to the idea to choose $\hat{m} \in \mathscr{G}$ such that it minimizes

$$\frac{1}{n-t}\sum_{i=t}^{n-1}\left(Y_{i+1} - g^{[t]}(0,Y_{i-t},\dots,Y_i)\right)^2$$

among all functions $g \in \mathscr{G}$. This is almost the final version of the estimator. Two disturbing facts remain about the last idea. First, for indexes $i > t$ we could achieve higher accuracy choosing the iteration parameter i instead of t. This would also eliminate the second disturbance, namely the existence of a hyper-parameter t, which would demand a tuning procedure. We therefore propose the the estimator $\hat{m}_n$ that minimizes

$$Q_n(g) := \frac{1}{n}\sum_{i=0}^{n-1}\left(Y_{i+1} - g^{[i]}(0,Y_0,\dots,Y_i)\right)^2. \tag{3.2}$$

This estimator will be called the (theoretical) least squares estimator of m.

Before we proceed with a formal definition of the least squares estimator, we need to assure that there is a formally correct way to define it. Two conditions have to be checked. First, does the functional Q_n attain its infimum over the set $\mathscr{G}$? In other words, is there a function $\hat{m} \in \mathscr{G}$ such that for all $g \in \mathscr{G}$, $Q_n(\hat{m}_n) \leq Q_n(g)$? Second, is the resulting estimator $\hat{m}_n$ a random variable, i.e. is it a measurable function $\Omega \to \mathscr{G}$? Answering these questions will be the subject of the next two propositions. The preliminary definition will be used in the course of the next proof.

DEFINITION 3.1.2. A subset A of a metric space (X,d) is called totally bounded if for any $\varepsilon > 0$ there exists a covering of A by finitely many balls $B_d(x_i,\varepsilon)$ around centers $x_i \in A$ with radius ε measured in the metric d.

PROPOSITION 3.1.3. *For $n \in \mathbb{N}_+$, let $Y_0,\dots,Y_n$ be $n+1$ successive count variables of the data generating process. Let the functional Q_n be defined as in equation* (3.2)*. Then the infimum of Q_n over the function class $\mathscr{G}$ is attained.*

Proof. We know that a continuous functional attains its infimum over a compact metric space (Rudin, 1976, Theorem 4.16). Thus, we show that Q is a continuous function and $\mathscr{G}$ is a compact metric space. Both assertions are to be understood with respect to the norm $\|g\|_\infty := \sup_x |g(x)|$ on $\mathscr{G}$. The continuity is easily verified. We observe that,

$$\begin{aligned}
n\left|Q_n(g)-Q_n(h)\right| &= \Big|\sum_{i=0}^{n-1}\Big[\big(Y_{i+1}-g^{[i]}(0,Y_0,\ldots,Y_i)\big)^2-\big(Y_{i+1}-h^{[i]}(0,Y_0,\ldots,Y_i)\big)^2\Big]\Big| \\
&\le \sum_{i=0}^{n-1}\Big[2Y_{i+1}\left|g^{[i]}(0,Y_0,\ldots,Y_i)-h^{[i]}(0,Y_0,\ldots,Y_i)\right| \\
&\qquad +\left|\big(g^{[i]}(0,Y_0,\ldots,Y_i)\big)^2-\big(h^{[i]}(0,Y_0,\ldots,Y_i)\big)^2\right|\Big] \\
&\le \sum_{i=0}^{n-1}\Big[2Y_{i+1}\left|g^{[i]}(0,Y_0,\ldots,Y_i)-h^{[i]}(0,Y_0,\ldots,Y_i)\right| \\
&\qquad +\left|g^{[i]}(0,Y_0,\ldots,Y_i)+h^{[i]}(0,Y_0,\ldots,Y_i)\right| \\
&\qquad \left|g^{[i]}(0,Y_0,\ldots,Y_i)-h^{[i]}(0,Y_0,\ldots,Y_i)\right|\Big] \\
&\le \sum_{i=0}^{n-1}2(M+Y_{i+1})\left|g^{[i]}(0,Y_0,\ldots,Y_i)-h^{[i]}(0,Y_0,\ldots,Y_i)\right|,
\end{aligned}$$

and additionally, by the contractive property applied to the first component,

$$\begin{aligned}
&\left|g^{[i]}(0,Y_0,\ldots,Y_i)-h^{[i]}(0,Y_0,\ldots,Y_i)\right| \\
&=\left|g\big(g^{[i-1]}(0,Y_0,\ldots,Y_{i-1}),Y_i\big)-h\big(h^{[i-1]}(0,Y_0,\ldots,Y_{i-1}),Y_i\big)\right| \\
&\le\left|g\big(g^{[i-1]}(0,Y_0,\ldots,Y_{i-1}),Y_i\big)-g\big(h^{[i-1]}(0,Y_0,\ldots,Y_{i-1}),Y_i\big)\right| \\
&\quad+\left|g\big(h^{[i-1]}(0,Y_0,\ldots,Y_{i-1}),Y_i\big)-h\big(h^{[i-1]}(0,Y_0,\ldots,Y_{i-1}),Y_i\big)\right| \\
&\le L_1\left|g^{[i-1]}(0,Y_0,\ldots,Y_{i-1})-h^{[i-1]}(0,Y_0,\ldots,Y_{i-1})\right|+\|g-h\|_\infty.
\end{aligned}$$

This gives us by an induction argument that

$$\left|g^{[i]}(0,Y_0,\ldots,Y_i)-h^{[i]}(0,Y_0,\ldots,Y_i)\right| \le \|g-h\|_\infty \sum_{k=0}^{i} L_1^k, \tag{3.3}$$

and therefore

$$|Q_n(g)-Q_n(h)| \le \frac{2\sum_{i=0}^{n-1}(M+Y_{i+1})}{n(1-L_1)}\|g-h\|_\infty.$$

This proves continuity of Q.

Metric spaces are compact if and only if they are complete and totally bounded (Bass, 2013, Theorem 20.23). We proceed by showing completeness. To that end consider a fundamental sequence $\{g_m\}_{m\in\mathbb{N}} \subset (\mathscr{G}, \|\cdot\|_\infty)$. We have to show that this sequence has a limit in $(\mathscr{G}, \|\cdot\|_\infty)$. Select a sequence of natural numbers

$\{m(j)\}_{j\in\mathbb{N}} \subset \mathbb{N}$ such that $\|g_{m(j)} - g_{m(j+1)}\|_\infty < 2^{-j}$, which is possible since $\{g_m\}$ is a fundamental sequence. We show that there exists a bounded function $F_g = \sum_{j=1}^{\infty}\left(g_{m(j+1)} - g_{m(j)}\right)$ in $(\mathscr{G}, \|\cdot\|_\infty)$. This follows from the next consideration.

Let $\{f_n\}_{n\in\mathbb{N}}$ be a sequence of functions in $\mathscr{G}$ with the property that $\sum_{n=1}^{\infty}\|f\|_\infty < \infty$. We show that there exists a bounded function F such that $\|F - \sum_{n=1}^{N} f_n\|_\infty \to 0$, as $N \to \infty$. The function F is defined explicitly by the pointwise assignment $F(x,y) := \limsup_{N\to\infty}\sum_{n=1}^{N} f_n(x,y)$. Then

$$\sup_{x,y}|F(x,y)| = \sup_{x,y}\left|\limsup_{N\to\infty}\sum_{n=1}^{N} f_n(x,y)\right| \le \limsup_{N\to\infty}\sup_{x,y}\sum_{n=1}^{N}|f_n(x,y)| \le \sum_{n=1}^{\infty}\|f_n\|_\infty < \infty,$$

and moreover

$$\begin{aligned}\|F - \sum_{n=1}^{N} f_n\|_\infty &= \sup_{x,y}\left|\limsup_{m\to\infty}\sum_{n=1}^{m} f_n(x,y) - \sum_{n=1}^{N} f_n(x,y)\right| \\ &\le \limsup_{m\to\infty}\sum_{n=N}^{m}\sup_{x,y}|f_n(x,y)| \\ &= \sum_{n=N}^{\infty}\|f_n\|_\infty \to 0,\ (N\to\infty).\end{aligned}$$

Hence, $F = \sum_{n=1}^{\infty} f_n$ in uniform norm.

Apply this result to the sequence $\left\{g_{m(j+1)} - g_{m(j)}\right\}_{j\in\mathbb{N}}$ to conclude that there exists a function $F_g \in (\mathscr{G}, \|\cdot\|_\infty)$ with the property

$$\|F_g - (g_{m(N+1)} - g_{m(0)})\|_\infty = \|F_g - \sum_{j=0}^{N}(g_{m(j+1)} - g_{m(j)})\|_\infty \to 0\ (N\to\infty).$$

The function $g_\infty := F_g + g_{m(0)}$ is therefore the uniform limit of the sub sequence $\{g_{m(j)}\}_{j\in\mathbb{N}}$. This function satisfies the contractive property since for any $\varepsilon > 0$ there exists $j_\varepsilon \in \mathbb{N}$ such that for all $x,u \in [0,M]$ and all $y,v \in \{0,\dots,B-1\}$

$$\begin{aligned}|g_\infty(x,y) - g_\infty(u,v)| \le& |g_{m(j_\varepsilon)}(x,y) - g_\infty(x,y)| + |g_{m(j_\varepsilon)}(x,y) - g_{m(j_\varepsilon)}(u,v)| \\ &+ |g_{m(j_\varepsilon)}(u,v) - g_\infty(u,v)| \\ <& 2\varepsilon + L_1|x-u| + L_2|y-v|.\end{aligned}$$

This shows that $g_\infty \in \mathscr{G}$. Then g_∞ is also the limit of the whole sequence: let $k_\varepsilon := \inf\{n > m(j_\varepsilon)\colon \|g_r - g_s\| < \varepsilon \text{ for all } r,s > n\}$ and $j^* := \inf\{j \in \mathbb{N}\colon m(j) > k_\varepsilon\}$. Then for all $n \ge k_\varepsilon$,

$$\|g_\infty - g_n\|_\infty \le \|g_\infty - g_{m(j)}\|_\infty + \|g_{m(j)} - g_n\| < 2\varepsilon$$

for some suitable $j > j^*$ depending on ε. Therefore, every fundamental sequence in $(\mathscr{G}, \|\cdot\|_\infty)$ has a limit, and the space is complete.

It follows from the next lemma that $(\mathscr{G}, \|\cdot\|_\infty)$ is totally bounded. Together with the just established completeness, this implies that $\mathscr{G}$ is compact, which concludes the proof. □

The following paragraph has the purpose to examine the size of the function class $\mathscr{G}$ in terms of covering numbers.

DEFINITION 3.1.4. Suppose that (X,d) is a metric space and $\varepsilon > 0$ a real number. Let $B_d(x,r)$ denote the Ball around $x \in X$ having radius r with respect to the metric d. The number

$$N(\varepsilon,d,X) := \min\left\{N \in \mathbb{N} : \exists x_1,\ldots,x_N \in X \text{ such that } X \subset \bigcup_{i=1}^{N} B_d(x_i,\varepsilon)\right\}$$

is called covering number of (X,d) for the resolution level ε.

Recall that a metric space (X,d) is called totally bounded if for any $\varepsilon > 0$ there exists a finite set of points $\{x_1,\ldots,x_{k_\varepsilon}\} \subset X$ such that the union of the balls $B_d(x_i,\varepsilon)$ covers X. If we find a real valued function $N : [0,\infty) \to [0,\infty)$ such that $N(\varepsilon) = N(\varepsilon,d,X)$, we can immediately conclude that X is totally bounded. The underlying result to the next lemma is an old one by Kolmogorov and Tikhomirov (1993, page 93).

LEMMA 3.1.5. *For $B \in \mathbb{N}_+$ and $L > 0$, suppose that $\tilde{\mathscr{G}}$ is the class of functions defined by*

$$\tilde{\mathscr{G}} := \Big\{g = (g_0,\ldots,g_{B-1})' : [0,M]^B \to [0,M];\\ |g_i(x) - g_i(y)| \le L\,|x-y| \ \textit{ for all } x,y \in [0,M]\Big\}.$$

For the covering numbers of $\tilde{\mathscr{G}}$ endowed with the uniform norm $\|\cdot\|_\infty$ we have the bound

$$\log N(\varepsilon, \|\cdot\|_\infty, \tilde{\mathscr{G}}) \le 2MB(L \vee 1)/\varepsilon\,.$$

In particular, since $\mathscr{G} \subset \tilde{\mathscr{G}}$, the same bound holds for $\mathscr{G}$, and $(\mathscr{G}, \|\cdot\|_\infty)$ is totally bounded.

Proof. In a first step we approximate the one-dimensional functions $g_i : [0,M] \to [0,M]$. For a real number r let the symbol $\lceil r \rceil$ denote the smallest integer n such that $n \ge r$. We define $N := \lceil ML/\varepsilon \rceil$ and suppose that $\pi_N := \{x_0,\ldots,x_N\}$ with $0 = x_0 < \ldots < x_N = M$ is a partition of the interval $[0,M]$ such that

$$x_i - x_{i-1} = \begin{cases} \frac{\varepsilon}{L} & \text{if } i \in \{1,\ldots,N-1\},\\ M - (N-1)\frac{\varepsilon}{L} & \text{if } i = N\,. \end{cases}$$

Note that $x_N - x_{N-1} \le \varepsilon/L$ with equality occurring if and only if $N = ML/\varepsilon$. Therefore, $L\,|x_i - x_{i-1}| \le \varepsilon$ for all $i = 1,\dots,N$.

We define a set $\mathscr{S}_\varepsilon$ of functions $s\colon [0,M] \to [0,M]$ that have the following properties:

(1) $s(x_i) \in \{k\,\varepsilon\colon k = 0,1,\dots,\lfloor \frac{M}{\varepsilon} \rfloor\}$, for all $i = 0,\dots,N-1$;

(2) $|s(x_i) - s(x_{i-1})| = L\,|x_i - x_{i-1}|$ if $s(x_i) \neq s(x_{i-1})$;

(3) if $s(x_i) = s(x_{i-1})$, then $s(x_i) = s(x_{i-1}) \in \{\lfloor \frac{M}{\varepsilon} \rfloor\,\varepsilon\,, 0\}$;

(4) s interpolates the points $(x_i, s(x_i))$ linearly.

The selection of the partition π_N implies that, apart from the boundaries, $|s(x_i) - s(x_{i-1})| = \varepsilon$ if $i \in \{1,\dots,N-1\}$, and $|s(x_N) - s(x_{N-1})| \le \varepsilon$. It may be surprising that we mainly exclude functions with constant segments from the set $\mathscr{S}_\varepsilon$. However, the non-constant functions suffice to find reasonably good approximations as we shall see shortly.

The functions $s \in \mathscr{S}_\varepsilon$ can be identified with vectors $(b_0, b_1, \dots, b_N)$, where $b_0 \in \{0, \varepsilon, \dots, \lfloor M/\varepsilon \rfloor \varepsilon\}$ and $b_1,\dots,b_N \in \{-1,1\}$. Then the function s is given by

$$\begin{aligned} s(x_0) &= b_0\,, \\ s(x_i) &= \Big(\big(s(x_{i-1}) + b_i\,L\,|x_i - x_{i-1}|\big) \vee 0\Big) \wedge \Big\lfloor \frac{M}{\varepsilon} \Big\rfloor \varepsilon\,, \end{aligned}$$

for $i = 1,\dots,N$. The total number of such configurations is

$$\begin{aligned} &\#\big\{(b_0,b_1,\dots,b_N)\colon b_0 \in \{0,\varepsilon,2\varepsilon,\dots,\lfloor M/\varepsilon \rfloor\,\varepsilon\}\,,\, b_1,\dots,b_N \in \{-1,1\}\big\} \\ &\le 2^N\,(1 + M/\varepsilon) \\ &= \exp\big(\log(1 + \frac{M}{\varepsilon}) + N\log 2\big) \\ &\le \exp\big(\frac{M}{\varepsilon} + N\log 2\big) \\ &\le \exp\big(2M(L \vee 1)/\varepsilon\big)\,. \end{aligned}$$

We define $\mathscr{S}_\varepsilon^B := \{s = (s_0,\dots,s_{B-1})\colon s_i \in \mathscr{S}_\varepsilon \text{ for all } j = 0,\dots,B-1\}$. Repeating the previous procedure for every component s_i, $i \in \{0,\dots,B-1\}$, we obtain at most $e^{2MB(L\vee 1)/\varepsilon}$ functions in $\mathscr{S}_\varepsilon^B$.

We show that the set $\mathscr{S}_\varepsilon^B$ is an appropriate approximating set for $\tilde{\mathscr{G}}$. Let $g = (g_0,\dots,g_{B-1}) \in \tilde{\mathscr{G}}$ be arbitrary but fixed. The following argument holds for any $j \in \{0,\dots,B-1\}$. Certainly, there are points $y_{j,0} \in \{0,\varepsilon,\dots,\lfloor M/\varepsilon \rfloor \varepsilon\}$ such that $|y_{j,0} - g_j(x_0)| < \varepsilon$. For some $i \in \{1,\dots,N\}$, suppose that there exists a point $y_{j,i-1}$ such that $|y_{j,i-1} - g_j(x_{i-1})| \le \varepsilon$. Due to the Lipschitz property of g_j, the value $g_j(x_i)$ is contained in the interval $\big[g_j(x_{i-1}) - L|x_i - x_{i-1}|\,,\, g_j(x_{i-1}) + L|x_i - x_{i-1}|\big]$.

We define

$$y_{j,i} := \begin{cases} (y_{j,i-1} - L|x_i - x_{i-1}|) \vee 0 & \text{if } g_j(x_i) \le y_{j,i-1} \\ (y_{j,i-1} + L|x_i - x_{i-1}|) \wedge \lfloor \frac{M}{\varepsilon} \rfloor \varepsilon & \text{if } g_j(x_i) > y_{j,i-1} \end{cases}$$

and conclude that $|y_{j,i} - g_j(x_i)| \le \max\{|y_{j,i-1} - g_j(x_{i-1})|, |y_{j,i-1} - y_{j,i}|\} \le \varepsilon$.

The previous argument returns a set of points $\{y_{j,0}, \dots, y_{j,N} : j = 0, \dots, B-1\}$ that satisfy $y_{j,i} \in \{0, \varepsilon, \dots, \lfloor M/\varepsilon \rfloor \varepsilon\}$ and $|y_{j,i-1} - y_{j,i}| \in \{L|x_i - x_{i-1}|, 0\}$, taking 0 only at the boundaries, as well as $|y_{j,i} - g_j(x_i)| \le \varepsilon$, for all $i = 1, \dots, N$ and $j = 0, \dots, B-1$. Thus, there exists a function $s^{(g)} = (s_0^{(g)}, \dots, s_{B-1}^{(g)}) \in \mathscr{S}_\varepsilon^B$ with $s_j^{(g)}(x_i) = y_{j,i}$ and the property $|s_j^{(g)}(x_i) - g_j(x_i)| \le \varepsilon$ for all $i = 0, \dots, N$ and $j = 0, \dots, B-1$. Furthermore,

$$\max_j \sup_x |g_j(x) - s_j^{(g)}(x)| \le \varepsilon,$$

which follows from the Lipschitz property of g and the construction of $s^{(g)}$. An illustration of the approximation procedure is given in Figure 3.1.1.

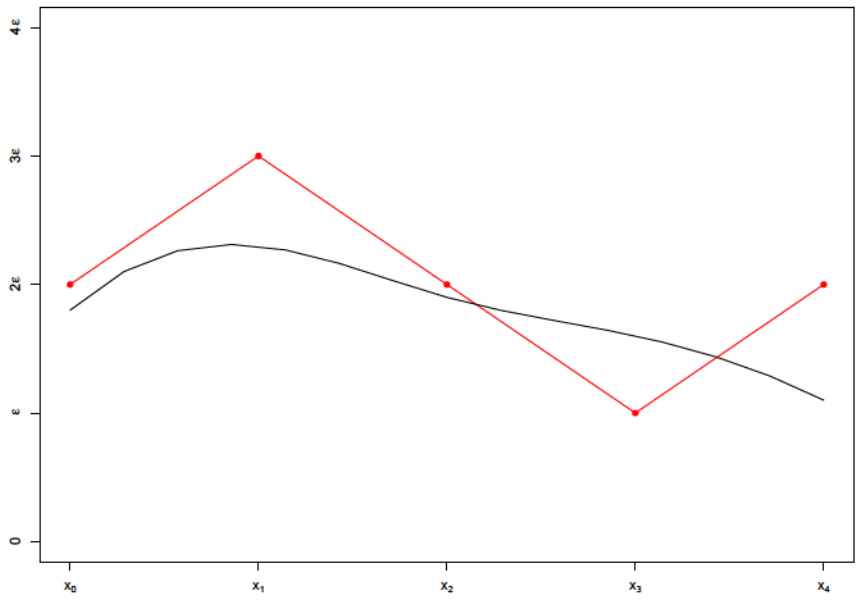


Figure 3.1.1: An illustration of a function g_j (black line) and the corresponding approximating function $s_j^{(g)}$ (red line).

Since g was arbitrary, we conclude that the balls $\{B_\infty(s,\varepsilon)\colon s \in \mathscr{S}_\varepsilon^B\}$ form an ε-cover of the class $\tilde{\mathscr{G}}$, with respect to the uniform norm. Therefore,

$$N(\varepsilon, \|\cdot\|_\infty, \tilde{\mathscr{G}}) \leq \#\mathscr{S}_\varepsilon^B \leq e^{2M(L\vee 1)B/\varepsilon}. \qquad \square$$

So far we have shown that the expression $\operatorname{arg\,min}_{g\in\mathscr{G}} Q(g)$ is meaningful in the sense that the set of functions at which the functional Q attains its minimum over $\mathscr{G}$ is not empty. Yet it is not clear whether this set contains more than one function. This leads to the second difficulty on the way towards a definition of an estimator: we have to find a way of selecting one function from the set of minimizers $\operatorname{arg\,min} Q$ based on the information given by the data.

Formally speaking, given realizations $y_0,\dots,y_n$ of the count process, we seek a measurable function $T(y_0,\dots,y_n)\colon \mathbb{R}^{n+1} \to \operatorname{arg\,min} Q$. This is not a trivial task since the functional Q also depends on the realizations of the data; a fact that is apparent from the definition of Q but was suppressed in the notation so far. Let us therefore change the notation in order to account for this fact, and write

$$Q_n(g) := Q_n(y_0,\dots,y_n;g) := \frac{1}{n}\sum_{i=0}^{n-1}\left(y_{i+1} - g^{[i]}(0,y_0,\dots,y_i)\right)^2. \tag{3.4}$$

Fortunately, we can recourse to established technical results from the field of set valued functions in order to resolve the matter.

LEMMA 3.1.6. *For $n \in \mathbb{N}_+$, let Q_n be defined as in equation* (3.4)*. There exists a $\mathscr{B}(\mathbb{R}^{n+1}) - \mathscr{B}(\mathscr{G})$ measurable function $T\colon \mathbb{R}^{n+1} \to \mathscr{G}$ such that*

$$T(y_0,\dots,y_n) \in \operatorname*{arg\,min}_{g\in\mathscr{G}} Q_n(y_0,\dots,y_n;g).$$

Proof. In a first step we prove that there exists a $(\mathscr{B}^{n+1} - \mathscr{B}^n)$-measurable function $\hat{\theta}(y_0,\dots,y_n)\colon \mathbb{R}^{n+1} \to \Theta \subset \mathbb{R}^n$ that minimizes

$$(y_0,\dots,y_n;\theta) \mapsto \sum_{i=0}^{n-1}(y_{i+1} - \theta(i))^2$$

over $\theta \in \Theta$. The set Θ is for our purposes best defined as

$$\Theta := \Big\{\theta = \big(\theta(0),\dots,\theta(n-1)\big)'\colon \text{there exists } g\in\mathscr{G} \text{ such that } g^{[i]}(0,y_0,\dots,y_i) = \theta(i) \text{ for all } i = 0,\dots,n-1\Big\}.$$

The existence of such a function $\hat{\theta}$ can be proven using the methodology of Jennrich (1969). The idea is quickly explained.

Assume that $\{f_\theta\}$ is a family of real valued measurable functions on a measurable space (S,Σ). Furthermore, suppose that for every $\omega \in S$ the mapping $\theta \mapsto f_\theta(\omega)$ is continuous. For $\omega \in S$ the infimum $a(\omega)$ of the set $\{f_\theta(\omega)\colon \theta \in \Theta\}$ can be realized by an approximating sequence, $\{f_{\theta_k(\omega)}\colon \theta_k(\omega) \in \Theta, k \in \mathbb{N}\}$, such that $a(\omega) = \lim_{k\to\infty} f_{\theta_k(\omega)}(\omega)$. We need the θ_k to be Σ-measurable. A simple way to find such a sequence is to construct an approximation of Θ by a sequence of finite subsets $\{\Theta_k\}_{k\in\mathbb{N}}$, $\Theta_k \subset \Theta_{k+1} \subset \Theta$ for all $k \in \mathbb{N}$, such that $\overline{\bigcup_{k=0}^{\infty} \Theta_k} = \Theta$. Then we choose $\theta_k(\omega)$ to minimize $\theta \mapsto f_\theta(\omega)$ over the finite set Θ_k. The function θ_k is measurable. To see this, let $\Theta_k = \{\vartheta_i\colon i = 1,\dots,N_k\}$ and observe

$$\{\omega\colon \theta_k(\omega) = \vartheta_i\} = \bigcap_{j=1}^{N_k} \{\omega\colon f_{\vartheta_i}(\omega) \le f_{\vartheta_j}(\omega)\} \in \Sigma\,.$$

Then we can show that $f_{\theta_k(\omega)}(\omega) \to a(\omega)$ as $k \to \infty$. Let $\varepsilon > 0$, $\omega \in S$, and $\theta \in \Theta$ be arbitrary but fixed. The function f is continuous in θ. Thus, there exists a $\delta(\theta,\varepsilon,\omega)$ such that for all $\theta^* \in \Theta$ with $\|\theta^* - \theta\| < \delta$ we can conclude that $|f_{\theta^*}(\omega) - f_\theta(\omega)| < \varepsilon$. Furthermore, since $\Theta_k \uparrow \Theta$, there exists an index $K(\delta,\theta,\varepsilon,\omega)$ such that for all $k \ge K$ there exists a $\vartheta(\theta,k,\varepsilon,\omega) \in \Theta_k$ such that $\|\theta - \vartheta\| < \delta$. Recall that $\theta_k(\omega) \in \Theta_k$ was chosen such that $f_{\theta_k(\omega)}(\omega) \le f_\vartheta(\omega)$ for all $\vartheta \in \Theta_k$. Therefore, we finally obtain

$$f_\theta(\omega) > f_{\vartheta(\theta,\varepsilon,\omega)}(\omega) - \varepsilon \ge f_{\theta_{k(\theta,\varepsilon,\omega)}(\omega)}(\omega) - \varepsilon \ge \inf_{k\in\mathbb{N}} f_{\theta_k(\omega)}(\omega) - \varepsilon\,.$$

Since $\theta \in \Theta$ was arbitrary, we conclude that $a(\omega) = \inf_\theta f_\theta(\omega) > \inf_k f_{\theta_k(\omega)}(\omega) - \varepsilon$. Since $\Theta_k \subset \Theta_{k+1}$ for any $k \in \mathbb{N}$, the sequence $\{f_{\theta_k(\omega)}(\omega)\}$ is non-increasing for any $\omega \in S$, whence it follows that

$$\lim_{k\to\infty} f_{\theta_k(\omega)}(\omega) = \inf_{k\in\mathbb{N}} f_{\theta_k(\omega)}(\omega) = a(\omega)\,.$$

Therefore, $\{\theta_k(\omega)\}_{k\in\mathbb{N}}$ is an appropriate sequence. If it had a convergent subsequence $\{\theta_{k(r)}(\omega)\}_{r\in\mathbb{N}}$, the corresponding limit $\theta_\infty(\omega) := \lim_{r\to\infty} \theta_{k(r)}(\omega)$ would serve our purpose since the limiting function $\omega \mapsto \theta_\infty(\omega)$ would be measurable and by continuity of f we could conclude that

$$a(\omega) = \lim_{r\to\infty} f_{\theta_{k(r)(\omega)}}(\omega) = f_{\theta_\infty(\omega)}(\omega)\,.$$

However, we have to be aware that the existence of the convergence sub-sequence $\{\theta_{k(r)}\}$ is in general not certain.

Let us apply this line of argument to our setting. Since $\mathbb{R}^n$ is separable, there exists a monotonically increasing sequence of finite subsets $\{\Theta_q\}_{q\in\mathbb{N}}$ such that $\Theta = \overline{\bigcup_{q=1}^{\infty} \Theta_q}$. Since Θ_q is finite, there exists a measurable function $\hat{\theta}_q(y_0,\dots,y_n)$

with values in $\mathbb{R}^n$ that satisfies

$$\sum_{i=0}^{n-1} \left(y_{i+1} - \hat{\theta}_q(y_0,\dots,y_n)[i]\right)^2 = \min_{\theta\in\Theta_q} \sum_{i=0}^{n-1} \left(y_{i+1} - \theta(i)\right)^2 .$$

We obtain a sequence $\{\hat{\theta}_q\}_{q\in\mathbb{N}}$ of n dimensional vectors $\hat{\theta}_q = \left(\hat{\theta}_q(0),\dots,\hat{\theta}_q(n-1)\right)'$. Now we have to find a sub-sequence that converges to an element in Θ. This leads to the notion of compactness. We show that Θ is compact.

Since $\mathscr{G}$ contains only functions that are bounded by M, the set Θ is a bounded subset of $\mathbb{R}^n$. In order to show compactness of Θ, we show that it is a closed set. Let $\{\theta_m\}_{m\in\mathbb{N}}$ be a sequence in Θ with $\lim_{m\to\infty}\theta_m = \theta_\infty$. By definition there exists a sequence of functions $\{g_m\}_{m\in\mathbb{N}} \subset \mathscr{G}$ such that

$$g_m^{[i]}(0, y_0,\dots,y_i) = \theta_m(i).$$

We have already shown that $(\mathscr{G}, \|\cdot\|)_\infty$ is compact, which implies the existence of a convergent sub sequence, $\{g_{r(j)}\}_{j\in\mathbb{N}} \subset \{g_m\}_{m\in\mathbb{N}}$. Denote by g_∞ the limit of this sub sequence, i.e. $\|g_\infty - g_{r(j)}\|_\infty \to 0$ as $j\to\infty$. We conclude by inequality (3.3) that

$$\begin{aligned} g_\infty^{[i]}(0, y_0,\dots,y_i) &= \lim_{j\to\infty} g_{r(j)}^{[i]}(0, y_0,\dots,y_i) \\ &= \lim_{j\to\infty} \theta_{r(j)}(i) \\ &= \theta_\infty(i), \end{aligned}$$

which implies that $\theta_\infty \in \Theta$. Therefore, Θ is also a closed sub set. Closed and bounded subsets of $\mathbb{R}^n$ are compact. Hence, Θ is compact.

Since Θ is compact, there exists a sub sequence $\{q(r)\}_{r\in\mathbb{N}} \subset \mathbb{N}$ such that the sequence of vectors $\left\{\left(\hat{\theta}_{q(r)}(0),\dots,\hat{\theta}_{q(r)}(n-1)\right)'\right\}_{r\in\mathbb{N}}$ converges in any norm on $\mathbb{R}^d$ to a limit $\left(\hat{\theta}_\infty(0),\dots,\hat{\theta}_\infty(n-1)\right)' \in \Theta$:

$$\lim_{r\to\infty} \left\| \begin{pmatrix} \hat{\theta}_{q_n(r)}(0) \\ \vdots \\ \hat{\theta}_{q_n(r)}(n-1) \end{pmatrix} - \begin{pmatrix} \hat{\theta}_\infty(0) \\ \vdots \\ \hat{\theta}_\infty(n-1) \end{pmatrix} \right\| = 0 .$$

For $\boldsymbol{y} := (y_1,\dots,y_n)$, we observe that the function $\theta \mapsto \sum_{i=0}^{n-1}(y_{i+1}-\theta(i))^2 = \|\boldsymbol{y}-\theta\|_2^2$ is continuous. The general argument at the beginning of the proof lets us conclude that

$$\begin{aligned} \sum_{i=0}^{n-1} \left(y_{i+1} - \hat{\theta}_\infty(i)\right)^2 &= \lim_{r\to\infty} \sum_{i=0}^{n-1} \left(y_{i+1} - \hat{\theta}_{q(r)}(i)\right)^2 \\ &= \lim_{r\to\infty} \min_{\theta\in\Theta_{q(r)}} \sum_{i=0}^{n-1} \left(y_{i+1} - \theta(i)\right)^2 \end{aligned}$$

$$= \inf_{\theta\in\Theta} \sum_{i=0}^{n-1} \big(y_{i+1} - \theta(i)\big)^2 .$$

This proves the existence of the measurable function $\hat{\theta}(y_0,\dots,y_n)$ that we were looking for.

It remains to be proven that we can select a function $g \in \mathscr{G}$ such that

$$\begin{pmatrix} \hat{\theta}_\infty(0) \\ \vdots \\ \hat{\theta}_\infty(n-1) \end{pmatrix} = \begin{pmatrix} g^{[0]}(0,y_0) \\ \vdots \\ g^{[n-1]}(0,y_0,\dots,y_{n-1}) \end{pmatrix}.$$

We prove the existence of a selection function $T(y_0,\dots,y_n)$ with values in the set $\hat{\mathscr{G}} \subset \mathscr{G}$ defined by

$$\hat{\mathscr{G}}(y_0,\dots,y_n) := \{g \in \mathscr{G} : g^{[i]}(0,y_0,\dots,y_i) = \hat{\theta}_\infty(i) \text{ for all } i = 0,\dots,n-1\}.$$

Our tool will be the Kuratowski-Ryll-Nardzewski selection theorem. We use the version that is stated in Aliprantis and Border (1994). For the readers convenience, we quote a formulation of the theorem. In the next definition, the term *correspondence* describes a set valued function.

DEFINITION 3.1.7. Let (S,Σ) be a measurable space and (X,d) a metric space. A correspondence $\phi : S \to 2^X$ is called weakly measurable if for any open subset $O \subset X$ the set

$$\phi^l(O) := \{x \in X : \phi(x) \cap O \neq \emptyset\}$$

is a Σ-measurable set.

THEOREM 3.1.8 (cf. Aliprantis and Border, 1994, pages 504–505). *Assume that the correspondence $(y_0,\dots,y_n) \mapsto \hat{\mathscr{G}}(y_0,\dots,y_n) \subset \mathscr{G}$ satisfies the following conditions:*

(1) $\hat{\mathscr{G}}$ is weakly measurable;

(2) for all $y \in \mathbb{R}^{n+1}$, the set $\hat{\mathscr{G}}(y)$ is nonempty and closed.

Then there exists a measurable function $T : \mathbb{R}^{n+1} \to \mathscr{G}$ such that for all $\boldsymbol{y} \in \mathbb{R}^{n+1}$ the relation $T(\boldsymbol{y}) \in \hat{\mathscr{G}}(\boldsymbol{y})$ holds.

In order to apply this theorem, we make use of a characterization of weak measurability that we also quote from Aliprantis and Border (1994).

DEFINITION 3.1.9 (Ibid., page 499). Let (S,Σ) be a measurable space, and let X and Y be topological spaces. A Carathéodory function is a function $f : S \times X \to Y$ satisfying the following conditions:

(1) for each $x \in X$, the function $f(\cdot, x)\colon S \to Y$ is $(\Sigma - \mathscr{B}(Y))$-measurable;

(2) for each $s \in S$, the function $f(s, \cdot)\colon X \to Y$ is continuous.

THEOREM 3.1.10 (Ibid., Theorem 14.78). *Let (S, Σ) be a measurable space and (X, d) a separable metric space, and let the correspondence $\phi\colon S \to 2^X$ be nonempty-valued. Define $\delta\colon S \times X \to \mathbb{R}$ by*

$$\delta(s, x) = d\big(x, \phi(s)\big) := \inf\big\{d(x, y)\colon y \in \phi(s)\big\}.$$

Then the correspondence ϕ is weakly measurable if and only if δ is a Carathéodory function.

Recall that in our case $S = \mathbb{R}^{n+1}$ and $(X, d) = (\mathscr{G}, \|\cdot\|_\infty)$. Employing Theorem 3.1.10 to verify the weak measurability of the correspondence $\hat{\mathscr{G}}$, we have to show that the function $\delta\colon \mathbb{R}^{n+1} \times \mathscr{G} \to \mathbb{R}$ given by

$$\delta\big((y_0, \ldots, y_n)', f\big) = \inf\big\{\|f - g\|_\infty\colon g \in \hat{\mathscr{G}}(y_0, \ldots, y_n)\big\}$$

is a Carathéodory function. Regarding the matter of continuity, let $f_1, f_2 \in \mathscr{G}$. For $i \in \{1, 2\}$, we can find functions $g_i \in \hat{\mathscr{G}}(y_0, \ldots, y_n)$ such that $\|f_i - g_i\|_\infty - \varepsilon < \delta\big((y_0, \ldots, y_n)', f_i\big)$. For $\boldsymbol{y} := (y_0, \ldots, y_n)'$, it follows that

$$\begin{aligned}
\delta(\boldsymbol{y}, f_1) - \delta(\boldsymbol{y}, f_2) &< \|f_1 - g_2\|_\infty - \big(\|f_2 - g_2\|_\infty - \varepsilon\big) \\
&\leq \|f_1 - f_2\|_\infty + \|f_2 - g_2\|_\infty - \|f_2 - g_2\|_\infty + \varepsilon \\
&= \|f_1 - f_2\|_\infty + \varepsilon.
\end{aligned}$$

Similarly, one shows that $\delta(\boldsymbol{y}, f_2) - \delta(\boldsymbol{y}, f_1) < \|f_1 - f_2\|_\infty + \varepsilon$, and we obtain for any $(y_0, \ldots, y_n)' \in \mathbb{R}^{n+1}$ and $\varepsilon > 0$ that

$$\big|\delta\big((y_0, \ldots, y_n)', f_1\big) - \delta\big((y_0, \ldots, y_n)', f_2\big)\big| < \|f_1 - f_2\|_\infty + \varepsilon.$$

Hence, the function $f \mapsto \delta(\boldsymbol{y}, f)$ is continuous.

It remains to be proven that $(y_0, \ldots, y_n) \mapsto \delta\big((y_0, \ldots, y_n)', f\big)$ is a $(\mathscr{B}^{n+1} - \mathscr{B})$-measurable function for every $f \in \mathscr{G}$. To that end, let $\mathscr{G}_n \subset \mathscr{G}$ be a countable dense subset. Such a set exists due to the Weierstrass approximation theorem (Bass, 2013, Theorem 20.41). Then we obtain

$$\begin{aligned}
&\Big\{\boldsymbol{y} \in \mathbb{R}^{n+1}\colon \inf_{g \in \hat{\mathscr{G}}(\boldsymbol{y})} \|f - g\|_\infty < u\Big\} \\
&= \Big\{\boldsymbol{y} \in \mathbb{R}^{n+1}\colon \exists g \in \hat{\mathscr{G}}(\boldsymbol{y}) \text{ such that } \|f - g\|_\infty < u\Big\} \\
&= \Big\{\boldsymbol{y} \in \mathbb{R}^{n+1}\colon \exists g \in \hat{\mathscr{G}}(\boldsymbol{y}) \cap \mathscr{G}_n \text{ such that } \|f - g\|_\infty < u\Big\}
\end{aligned}$$

$$= \bigcup_{g\in\mathscr{G}_n} \left\{ \boldsymbol{y}\in\mathbb{R}^{n+1} :\ g\in\hat{\mathscr{G}}(\boldsymbol{y}) \right\} \cap \left\{ \boldsymbol{y}\in\mathbb{R}^{n+1} : \|f-g\|_\infty < u \right\}$$
$$= \bigcup_{g\in\mathscr{G}_n} \bigcap_{i=0}^{n-1} \left\{ \boldsymbol{y}\in\mathbb{R}^{n+1} :\ g^{[i]}(0,y_0,\dots,y_i) = \hat{\theta}_\infty(i) \right\} \cap \left\{ \boldsymbol{y}\in\mathbb{R}^{n+1} : \|f-g\|_\infty < u \right\}.$$

The functions $g^{[i]}$ are measurable, which implies that the preimage of any singleton $\{\theta_\infty(i)\}$ is measurable. On the other hand, the set $\{y\colon \|f-g\| < u\}$ is either empty or equals $\mathbb{R}^{n+1}$. We conclude that for any $f\in\mathscr{G}$ the preimages of intervals $(-\infty,u)$ under the function $(y_0,\dots,y_n)\mapsto\delta\big((y_0,\dots,y_n)',f\big)$ are Borel sets. Hence, δ is a Caratéodory function. According to Theorem 3.1.10, the correspondence $\hat{\mathscr{G}}$ is weakly measurable.

It is left to check the second condition of Theorem 3.1.8. Let the sequence $\{g_m\}_{m\in\mathbb{N}}\subset\hat{\mathscr{G}}(y_0,\dots,y_n)$ converge to a limit, $\|g_m-g_\infty\|_\infty\to 0$ for some $g_\infty\in\mathscr{G}$. Using $\hat{\theta}_\infty(i)=g_m^{[i]}(0,y_0,\dots,y_i)$, we observe that

$$\begin{aligned}
&\left|g_\infty^{[i]}(0,y_0,\dots,y_i)-\hat{\theta}_\infty(i)\right| \\
&= \left|g_\infty^{[i]}(0,y_0,\dots,y_i)-g_m^{[i]}(0,y_0,\dots,y_i)\right| \\
&= \Big|g_\infty\left(g_\infty^{[i-1]}(0,y_0,\dots,y_{i-1}),y_i\right) \\
&\qquad\qquad -g_m\left(g_m^{[i-1]}(0,y_0,\dots,y_{i-1}),y_i\right)\Big| \\
&\le \Big|g_\infty\left(g_\infty^{[i-1]}(0,y_0,\dots,y_{i-1}),y_i\right) \\
&\qquad\qquad -g_m\left(g_\infty^{[i-1]}(0,y_0,\dots,y_{i-1}),y_i\right)\Big| \\
&\quad +\Big|g_m\left(g_\infty^{[i-1]}(0,y_0,\dots,y_{i-1}),y_i\right) \\
&\qquad\qquad -g_m\left(g_m^{[i-1]}(0,y_0,\dots,y_{i-1}),y_i\right)\Big| \\
&\le \|g_\infty-g_m\|_\infty + L\left|g_\infty^{[i-1]}(0,y_0,\dots,y_{i-1})-g_m^{[i-1]}(0,y_0,\dots,y_{i-1})\right| \\
&\ \vdots \\
&\le \|g_\infty-g_m\|_\infty \sum_{k=0}^{i} L^k \\
&\le \frac{\|g_\infty-g_m\|_\infty}{1-L}\to 0\ (m\to\infty),
\end{aligned}$$

which implies that for any $i\in\{0,\dots,n-1\}$

$$g_\infty^{[i]}(0,y_0,\dots,y_i)=\hat{\theta}_\infty(i).$$

We conclude that $g_\infty\in\hat{\mathscr{G}}(y_0,\dots,y_n)$ and thence that for all $(y_0,\dots,y_n)$ the set $\hat{\mathscr{G}}(y_0,\dots,y_n)$ is closed. The set is furthermore nonempty by the definition of Θ.

We are thus in a position to apply the Kuratowski-Ryll-Narzewski selection theorem, which concludes the proof. □

We have collected evidence that a minimizer of the functional Q_n can be well defined on the basis of our observations. The heuristic meaning of the selection function is merely that it takes the data driven information about the values of the estimator at a certain finite number of points and augments it to a whole function. Since the selection function is not unique, there are several possibilities to draw an estimation curve from our observations. This is quite similar to the classical isotonic least squares estimator (Barlow et al., 1972).

The object resulting from the following definition will be the subject of the subsequent asymptotic considerations.

DEFINITION 3.1.11. Let $n \in \mathbb{N}_+$, and suppose that $(y_0, \ldots, y_n) \mapsto T(y_0, \ldots, y_n)$ is a measurable selection function with values in the set of minimizers of the functional Q_n,

$$T(y_0, \ldots, y_n) \in \operatorname*{arg\,min}_{g \in \mathscr{G}} \sum_{i=0}^{n-1} \left(y_{i+1} - g^{[i]}(0, y_0, \ldots, y_i)\right)^2.$$

The least squares estimator $\hat{m}_n$ of m that is based on observations $Y_0(\omega), \ldots, Y_n(\omega)$ of $n+1$ consecutive count variables of the data generating process from Definition 2.1.1 is defined as

$$\hat{m}_n[Y_0, \ldots, Y_n] := T(Y_0, \ldots, Y_n).$$

By the prior results, the mapping $\hat{m}_n : \Omega \to \mathscr{G}$ is well defined and measurable.

3.2 Asymptotic error analysis of the estimator

3.2.1 Preliminary considerations and statement of the main theorem

Having defined an estimator for our problem, we need to evaluate its performance. We shall do this in terms of an asymptotic error analysis. To that end, we have to settle on the question which measure of error we lay as a foundation of our evaluation. It is sensible to put more weight on areas in the domain of m and $\hat{m}_n$ where more observation are made. Therefore, we decide to measure the estimation error in terms of the $L_2(\pi)$ loss function. Here, as above, π denotes the stationary distribution of the bivariate data generating process $\{(\lambda_t, Y_t)\}_{t \in \mathbb{Z}}$. We set

$$L(\hat{m}_n, m) := \int \left(\hat{m}_n[Y_0, \ldots, Y_n](x, y) - m(x, y)\right)^2 \pi(dx, dy).$$

Suppose that $\{(\lambda'_t, Y'_t)\}_{t\in\mathbb{Z}}$ is an independent copy of the data generating process in the stationary regime. We use this ghost process to derive an alternative expression of the loss $L(\hat{m}_n, m)$ in terms of conditional expectation:

$$\begin{aligned} L(\hat{m}_n, m) &= \mathbb{E}\Big[\big(g(\lambda'_0, Y'_0) - m(\lambda'_0, Y'_0)\big)^2 \,\Big|\, \hat{m}_n[Y_0, \ldots, Y_n] = g\Big], \text{ a.s.} \\ &=: \mathbb{E}_{|\hat{m}_n = g}\big[g(\lambda'_0, Y'_0) - m(\lambda'_0, Y'_0)\big]^2. \end{aligned}$$

Henceforth we will suppress the dependence $\hat{m}_n = \hat{m}_n[Y_0, \ldots, Y_n]$ in the notation at most occasions. However, we should be aware of this dependence as it implies that $L(\hat{m}_n, m)$ is a random variable. We want to analyze the rate of convergence $L(\hat{m}_n, m) \to 0$ as the number of observations tends to infinity. Before we proceed, we have to decide which of the probabilistic modes of convergence we want to work with. For our purposes, an examination of the sequence $\{L(\hat{m}_n, m)\}_{n\in\mathbb{N}}$ in terms of convergence in probability seems well suited.

DEFINITION 3.2.1. Let $(E, \mathscr{E}, P)$ be a probability space and $\{X_n\}_{n\in\mathbb{N}}$ a sequence of random variables $X_n : (E, \mathscr{E}) \to (\mathbb{R}, \mathscr{B})$. Let $\{r_n\}_{n\in\mathbb{N}}$ be a sequence of positive real numbers, $r_n \downarrow 0$. The sequence $\{X_n\}$ is of order $O_P(r_n)$ if the sequence $\{X_n/r_n\}$ is bounded in P, i.e. for any $\varepsilon > 0$, there exists a positive real number $K(\varepsilon)$ and an index $n_0(\varepsilon)$ such that

$$\sup_{n\geq n_0} P\Big\{\Big|\frac{X_n}{r_n}\Big| > K\Big\} < \varepsilon.$$

We use the notation $X_n = O_P(r_n)$.

Certainly, $X_n = O_P(r_n)$ for a positive sequence $r_n \downarrow 0$ means that $X_n \to 0$ in P-measure, since for any $\varepsilon > 0$ and any $\delta > 0$ there are $K(\varepsilon)$ and $n_0(\delta, \varepsilon)$ such that $\sup_{n\geq n_0} K r_n \leq \delta$, and for all $n \geq n_0$

$$P\{|X_n| > \delta\} \leq P\{|X_n| > r_n K\} < \varepsilon.$$

Hence, on our probability space $(\Omega, \mathscr{A}, \mathbb{P})$ the relation $L(\hat{m}_n, m) = O_{\mathbb{P}}(r_n)$ implies that $L(\hat{m}_n, m) \to 0$ in $\mathbb{P}$-measure. We will call such a sequence of estimators $\{\hat{m}_n\}_{n\in\mathbb{N}_+}$ consistent with rate $\sqrt{r_n}$. We emphasize that in order to prove consistency with a certain rate $\sqrt{r_n}$, we need to find a bound for exceedance probabilities of the form $\mathbb{P}\{L(\hat{m}_n, m) > \delta_n^2\}$, where $\{\delta_n\}$ is a sequence in $O(\sqrt{r_n})$.

THEOREM 3.2.2. *Let $\{(Y_i, \lambda_i)\}_{i\in\mathbb{Z}}$ be the bivariate data generating process from Definition 3.1.1. Assume that the parameters M, B, L_1, L_2 in the definition of the class $\mathscr{G}$ are fixed. For $n \in \mathbb{N}_+$, let $\hat{m}_n$ be the least squares estimator for m on the basis of observations $Y_0, \ldots, Y_n$ of the count process (cf. Definition 3.1.11). The sequence $\{\delta_n\}_{n\in\mathbb{N}}$ shall be given by $\delta(n) = n^{-1/3}\log(n)$. Then, for an independent*

copy λ_0', Y_0' *of* λ_0, Y_0 *it holds that*

$$\limsup_{n\to\infty} \mathbb{P}\left\{\mathbb{E}_{|\hat{m}_n=g}\left[m(\lambda_0', Y_0') - g(\lambda_0', Y_0')\right]^2 > \delta_n^2\right\} = 0\,.$$

In other words, the sequence of least squares estimators is consistent with rate $n^{-1/3}\log n$.

The proof of this claim will be the subject of the remaining section. The argument is long and technical. At this point we give an outline of the envisioned strategy. To get a glimpse on the basic thoughts guiding our procedure, we briefly retreat to the field of regular least squares estimation: let us for a moment assume that λ_t is an observable variable. Clearly, in this case the estimator would be defined as a measurable selection of

$$\hat{m}_n \in \arg\min_{g\in\mathscr{G}} \frac{1}{n}\sum_{i=0}^{n-1}\big(Y_{i+1} - g(\lambda_i, Y_i)\big)^2\,.$$

Hence, the fundamental principle of this estimator is that the empirical prediction error of the estimation $\hat{m}_n$ is lower than that of the true function m:

$$\frac{1}{n}\sum_{i=0}^{n-1}\big(Y_{i+1} - m(\lambda_i, Y_i)\big)^2 - \frac{1}{n}\sum_{i=0}^{n-1}\big(Y_{i+1} - \hat{m}_n(\lambda_i, Y_i)\big)^2 \geq 0\,. \tag{3.5}$$

The second important fact is a consequence of the model assumption

$$\mathbb{E}_{|\mathscr{F}_i} Y_{i+1} = \lambda_{i+1} = m(\lambda_i, Y_i)\,, \quad \text{a.s.}\,.$$

From this assumption we derive for any $g \in \mathscr{G}$ that

$$\begin{aligned}
&\mathbb{E}\Big[\big(Y_{i+1} - g(\lambda_i, Y_i)\big)^2 - \big(Y_{i+1} - m(\lambda_i, Y_i)\big)^2\Big] \\
&= \mathbb{E}\Big[\big(Y_{i+1} - m(\lambda_i, Y_i) + m(\lambda_i, Y_i) - g(\lambda_i, Y_i)\big)^2 - \big(Y_{i+1} - m(\lambda_i, Y_i)\big)^2\Big] \\
&= \mathbb{E}\big(g(\lambda_i, Y_i) - m(\lambda_i, Y_i)\big)^2 + 2\mathbb{E}\Big[\big(m(\lambda_i, Y_i) - g(\lambda_i, Y_i)\big)\,\mathbb{E}_{|\mathscr{F}_i}\big(Y_{i+1} - m(\lambda_i, Y_i)\big)\Big] \\
&= \mathbb{E}\big(m(\lambda_i, Y_i) - g(\lambda_i, Y_i)\big)^2\,.
\end{aligned} \tag{3.6}$$

Taken together, equations (3.5) and (3.6) give the following estimate of probabilities:

$$\begin{aligned}
&\mathbb{P}\Big\{L(\hat{m}_n, m) > \delta_n\Big\} \\
&\leq \mathbb{P}\Big\{\mathbb{E}_{|\hat{m}_n[Y_0,\ldots,Y_n]=g}\Big[\big(Y_{i+1}' - g(\lambda_i', Y_i')\big)^2 - \big(Y_{i+1}' - m(\lambda_i', Y_i')\big)^2\Big] > \delta_n\Big\} \\
&\leq \mathbb{P}\Big\{\frac{1}{n}\sum_{i=1}^{n}\Big(\big(Y_{i+1} - m(\lambda_i, Y_i)\big)^2 - \big(Y_{i+1} - \hat{m}_n(\lambda_i, Y_i)\big)^2 \\
&\qquad -\mathbb{E}_{|\hat{m}_n[Y_0,\ldots,Y_n]=g}\Big[\big(Y_{i+1}' - m(\lambda_i', Y_i')\big)^2 - \big(Y_{i+1}' - g(\lambda_i', Y_i')\big)^2\Big]\Big) > \delta_n\Big\}\,.
\end{aligned}$$

Since we do not have any anterior knowledge about the value of $\hat{m}_n[Y_0,\dots,Y_n]$, the usual strategy consists of finding a probabilistic bound for the sum

$$\frac{1}{n}\sum_{i=1}^{n}\Big(\big(Y_{i+1}-m(\lambda_i,Y_i)\big)^2-\big(Y_{i+1}-g(\lambda_i,Y_i)\big)^2 \\ -\mathbb{E}\Big[\big(Y'_{i+1}-m(\lambda'_i,Y'_i)\big)^2-\big(Y'_{i+1}-g(\lambda'_i,Y'_i)\big)^2\Big]\Big)$$

uniformly for all $g\in\mathscr{G}$. Note that the expectation is not conditioned on the realization of $\hat{m}_n[Y_0,\dots,Y_n]$ any more, and technically the ghost sample is not necessary in this expression. Finding such uniform bounds is the established strategy in the asymptotic examination of M-estimators, a good expositions of which can be found in the book of Van der Vaart (2000). Let us try to carry these ideas over to the case of unobserved intensities λ_i.

In our model the characteristic feature of the least squares estimator is the fact that it minimizes the empirical prediction error

$$g\mapsto\frac{1}{n}\sum_{i=0}^{n-1}\big(Y_{i+1}-g^{[i]}(0,Y_0,\dots,Y_i)\big)^2$$

over the class $\mathscr{G}$. In fact, we are equipped with an analogue of inequality (3.5):

$$\frac{1}{n}\sum_{i=0}^{n-1}\big(Y_{i+1}-m^{[i]}(0,Y_0,\dots,Y_i)\big)^2-\frac{1}{n}\sum_{i=0}^{n-1}\big(Y_{i+1}-\hat{m}_n^{[i]}(0,Y_0,\dots,Y_i)\big)^2\geq 0\,. \quad (3.7)$$

This is the reason why the random functional

$$f_i(g;0,Y_0,\dots,Y_{i+1}):=\big(Y_{i+1}-m^{[i]}(0,Y_0,\dots,Y_i)\big)^2-\big(Y_{i+1}-g^{[i]}(0,Y_0,\dots,Y_i)\big)^2$$

plays a very prominent role in the upcoming argumentation. We will establish in Lemma 3.2.4 an analogue of equation (3.6):

$$\mathbb{E}_{|\hat{m}_n=g}\big(-f_i(g;0,Y'_0,\dots,Y'_{i+1})\big)\geq C\,\mathbb{E}_{|\hat{m}_n=g}\big[m(\lambda'_0,Y'_0)-g(\lambda'_0,Y'_0)\big]^2-o(1)\ \text{ a.s.}$$

for some constant $C>0$. What is the meaning of this relation? Suppose that g was selected as an estimation of m and that the $L_2(\pi)$-difference between m and g is large in some sense. Then the above relation says that the expected prediction error of $g^{[i]}(0,Y_0,\dots,Y_i)$ is considerably larger than that of $m^{[i]}(0,Y_0,\dots,Y_i)$, and the difference between the two prediction errors,

$$\begin{aligned}&\mathbb{E}_{|\hat{m}_n=g}\big(-f_i(g;0,Y'_0,\dots,Y'_0)\big)\\ &=\mathbb{E}_{|\hat{m}_n=g}\Big[\big(Y_{i+1}-g^{[i]}(0,Y_0,\dots,Y_i)\big)^2-\big(Y_{i+1}-m^{[i]}(0,Y_0,\dots,Y_i)\big)^2\Big]\ \text{ a.s.}\,,\end{aligned}$$

is asymptotically at least as large as the distance between m and g. In total this

means that there exists a function $g \in \mathscr{G}$ such that the difference

$$\frac{1}{n}\sum_{i=t}^{n-1}\left(f_i(g;0,Y_0,\ldots,Y_{i+1}) - \mathbb{E}f_i(g;0,Y_0,\ldots,Y_{i+1})\right)$$

is large. If we were able to prove that the probability of this event is very small, we could conclude that the event of $\hat{m}_n$ being far away from m has even lower probability. This will be our line of argument.

3.2.2 Proof of Theorem 3.2.2

Recall that $\{(\lambda'_t, Y'_t)\}_{t\in\mathbb{Z}}$ is an independent copy of the data generating process. We start with two lemmas to establish the relation between the $L_2(\pi)$ risk and the expectation of the functional $f_t(g;0,Y_0,\ldots,Y_{t+1})$ that we mentioned above. In the following, all equations or inequalities for conditional expectations are understood to be almost surely with respect to the underlying probability measure.

LEMMA 3.2.3. *For a candidate function $g \in \mathscr{G}$ and a natural number t, the following inequality holds almost surely:*

$$\begin{aligned}&\mathbb{E}_{|\hat{m}_n=g}\left[\left(Y'_{t+1} - g^{[t]}(0,Y'_0,\ldots,Y'_t)\right)^2 - \left(Y'_{t+1} - m^{[t]}(0,Y'_0,\ldots,Y'_t)\right)^2\right]\\ &\geq \mathbb{E}_{|\hat{m}_n=g}\left[m^{[t]}(0,Y'_0,\ldots,Y'_t) - g^{[t]}(0,Y'_0,\ldots,Y'_t)\right]^2 - 3M^2L_1^t. \qquad (3.8)\end{aligned}$$

Proof. The following calculations hold with probability one. Define $\mathscr{F}'_t$ canonically to $\mathscr{F}_t$, and recall the facts $\mathbb{E}(Y'_{t+1}\,|\,\mathscr{F}'_t) = \lambda'_{t+1} = \mathrm{var}(Y'_{t+1}\,|\,\mathscr{F}'_t)$ and $\lambda'_{t+1} = m(\lambda'_t, Y'_t)$. Hence, for any $g \in \mathscr{G}$

$$\begin{aligned}&\mathbb{E}\left((Y'_{t+1} - \lambda'_{t+1})(\lambda'_{t+1} - g^{[t]}(0,Y'_0,\ldots,Y'_t))\right)\\ &= \mathbb{E}\left[\mathbb{E}_{|\mathscr{F}'_t}\left((Y'_{t+1} - \lambda'_{t+1})(m(\lambda'_t,Y'_t) - g^{[t]}(0,Y'_0,\ldots,Y'_t))\right)\right]\\ &= \mathbb{E}\left[(m(\lambda'_t,Y'_t) - g^{[t]}(0,Y'_0,\ldots,Y'_t)\mathbb{E}_{|\mathscr{F}'_t}\left((Y'_{t+1} - \lambda'_{t+1})\right)\right]\\ &= 0\,.\end{aligned}$$

This implies that

$$\begin{aligned}&\mathbb{E}_{|\hat{m}_n=g}\left[\left(Y'_{t+1} - g^{[t]}(0,Y'_0,\ldots,Y'_t)\right)^2\right]\\ &= \mathbb{E}_{|\hat{m}_n=g}\left[\left(Y'_{t+1} - \lambda'_{t+1}\right) + \left(\lambda'_{t+1} - g^{[t]}(0,Y'_0,\ldots,Y'_t)\right)\right]^2\\ &= \mathbb{E}\left(Y'_{t+1} - \lambda'_{t+1}\right)^2 + \mathbb{E}_{|\hat{m}_n=g}\left(\lambda'_{t+1} - g^{[t]}(0,Y'_0,\ldots,Y'_t)\right)^2.\end{aligned}$$

In particular,

$$\begin{aligned}&\mathbb{E}\left(\left(Y'_{t+1}-m^{[t]}(0,Y'_0,\dots,Y'_t)\right)^2\right)\\&=\mathbb{E}\left(Y'_{t+1}-\lambda'_{t+1}\right)^2+\mathbb{E}\left(\lambda'_{t+1}-m^{[t]}(0,Y'_0,\dots,Y'_t)\right)^2.\end{aligned}$$

In the following computation, we first plug in the previous two statements and subsequently use the established relation $\lambda'_{t+1}=m^{[t]}(\lambda'_0,Y'_0,\dots,Y'_t)$ and the contraction property of m. We conclude

$$\begin{aligned}&\mathbb{E}_{|\hat{m}_n=g}\Big[\left(Y'_{t+1}-g^{[t]}(0,Y'_0,\dots,Y'_t)\right)^2\\&\qquad\qquad-\left(Y'_{t+1}-m^{[t]}(0,Y'_0,\dots,Y'_t)\right)^2\Big]\\&=\mathbb{E}_{|\hat{m}_n=g}\left(\lambda'_{t+1}-g^{[t]}(0,Y'_0,\dots,Y'_t)\right)^2+\mathbb{E}\left(Y'_{t+1}-\lambda'_{t+1}\right)^2\\&\qquad-\mathbb{E}\left(Y'_{t+1}-\lambda'_{t+1}\right)^2-\mathbb{E}\left(\lambda'_{t+1}-m^{[t]}(0,Y'_0,\dots,Y'_t)\right)^2\\&=\mathbb{E}_{|\hat{m}_n=g}\left(m^{[t]}(\lambda'_0,Y'_0,\dots,Y'_t)-g^{[t]}(0,Y'_0,\dots,Y'_t)\right)^2\\&\qquad-\mathbb{E}\Big(\underbrace{m^{[t]}(\lambda'_0,Y'_0,\dots,Y'_t)-m^{[t]}(0,Y'_0,\dots,Y'_t)}_{|\cdot|\le L_1^t|\lambda'_0-0|}\Big)^2\\&\ge\mathbb{E}_{|\hat{m}_n=g}\left(m^{[t]}(\lambda'_0,Y'_0,\dots,Y'_t)-g^{[t]}(0,Y'_0,\dots,Y'_t)\right)^2-L_1^{2t}\|\lambda'_0-0\|_\infty^2\\&\ge\mathbb{E}_{|\hat{m}_n=g}\left(m^{[t]}(\lambda'_0,Y'_0,\dots,Y'_t)-g^{[t]}(0,Y'_0,\dots,Y'_t)\right)^2-M^2L_1^{2t}.\end{aligned}$$

This is almost the assertion. We proceed by showing

$$\begin{aligned}&\mathbb{E}_{|\hat{m}_n=g}\left[m^{[t]}(\lambda'_0,Y'_0,\dots,Y'_t)-g^{[t]}(0,Y'_0,\dots,Y'_t)\right]^2\\&\ge\mathbb{E}_{|\hat{m}_n=g}\left[m^{[t]}(0,Y'_0,\dots,Y'_t)-g^{[t]}(0,Y'_0,\dots,Y'_t)\right]^2-2M^2L_1^t.\end{aligned}\tag{3.9}$$

This is an immediate consequence of the contraction property of m. We observe

$$\begin{aligned}&\mathbb{E}_{|\hat{m}_n=g}\left(m^{[t]}(\lambda'_0,Y'_0,\dots,Y'_t)-g^{[t]}(0,Y'_0,\dots,Y'_t)\right)^2\\&=\mathbb{E}_{|\hat{m}_n=g}\Big[\left(m^{[t]}(\lambda'_0,Y'_0,\dots,Y'_t)-m^{[t]}(0,Y'_0,\dots,Y'_t)\right)\\&\qquad+\left(m^{[t]}(0,Y'_0,\dots,Y'_t)-g^{[t]}(0,Y'_0,\dots,Y'_t)\right)\Big]^2\\&=\mathbb{E}_{|\hat{m}_n=g}\underbrace{\left(m^{[t]}(\lambda'_0,Y'_0,\dots,Y'_t)-m^{[t]}(0,Y'_0,\dots,Y'_t)\right)^2}_{\ge0}\\&\quad+\mathbb{E}_{|\hat{m}_n=g}\left(m^{[t]}(0,Y'_0,\dots,Y'_t)-g^{[t]}(0,Y'_0,\dots,Y'_t)\right)^2\\&\quad+\mathbb{E}_{|\hat{m}_n=g}\Big[2\underbrace{\left(m^{[t]}(\lambda'_0,Y'_0,\dots,Y'_t)-m^{[t]}(0,Y'_0,\dots,Y'_t)\right)}_{|\cdot|\le ML_1^t}\end{aligned}$$

$$\cdot\left(m^{[t]}(0,Y_0',\dots,Y_t')-g^{[t]}(0,Y_0',\dots,Y_t')\right)\Big]$$
$$\geq \mathbb{E}_{|\hat{m}_n=g}\left(m^{[t]}(0,Y_0',\dots,Y_t')-g^{[t]}(0,Y_0',\dots,Y_t')\right)^2$$
$$-2ML_1^t \sup_{(y_0,\dots,y_t)\in\mathbb{R}^{t+1}}\left|m^{[t]}(0,y_0,\dots,y_t)-g^{[t]}(0,y_0,\dots,y_t)\right|$$
$$\geq \mathbb{E}_{|\hat{m}_n=g}\left(m^{[t]}(0,Y_0',\dots,Y_t')-g^{[t]}(0,Y_0',\dots,Y_t')\right)^2-2M^2L_1^t.$$

This proves inequality (3.9). Since $L_1^t > L_1^{2t}$, we obtain in summary

$$\mathbb{E}_{|\hat{m}_n=g}\left[\left(Y_{t+1}'-g^{[t]}(0,Y_0',\dots,Y_t')\right)^2-\left(Y_{t+1}'-m^{[t]}(0,Y_0',\dots,Y_t')\right)^2\right]$$
$$\geq \mathbb{E}_{|\hat{m}_n=g}\left[m^{[t]}(0,Y_0',\dots,Y_t')-g^{[t]}(0,Y_0',\dots,Y_t')\right]^2-\underbrace{\left(M^2L_1^{2t}+2M^2L_1^t\right)}_{\leq 3M^2L_1^t}. \qquad \square$$

LEMMA 3.2.4. *Suppose that $\Omega_0\in\mathscr{F}$ is the set of all $\omega\in\Omega$ such that for $g\in\mathscr{G}$, $\varepsilon\in(0,1)$ and some natural number t*

$$\varepsilon\left(\mathbb{E}_{|\hat{m}_n=g}\left[m(\lambda_0',Y_0')-g(\lambda_0',Y_0')\right]^2\right)^{1/2} > M\,L_1^t. \tag{3.10}$$

Then for almost all $\omega\in\Omega_0$ the following inequality holds:

$$\mathbb{E}_{|\hat{m}_n=g}\left[m^{[t]}(0,Y_0',\dots,Y_t')-g^{[t]}(0,Y_0',\dots,Y_t')\right]^2$$
$$> \frac{(1-\varepsilon)^2}{12}\,\mathbb{E}_{|\hat{m}_n=g}\left[m(\lambda_0',Y_0')-g(\lambda_0',Y_0')\right]^2-2M^2L_1^t \tag{3.11}$$

Proof. The following arguments hold almost surely. First of all, note that

$$\mathbb{E}_{|\hat{m}_n=g}\left[m(\lambda_0',Y_0')-g(\lambda_0',Y_0')\right]^2=\mathbb{E}_{|\hat{m}_n=g}\left[m(\lambda_{t+1}',Y_{t+1}')-g(\lambda_{t+1}',Y_{t+1}')\right]^2$$

by stationarity. With two successive applications of the triangle inequality, we insert $\mp g^{[t+1]}(\lambda_0',Y_0',\dots,Y_{t+1}')$ and $\mp g^{[t]}(\lambda_1',Y_1',\dots,Y_{t+1}')$ respectively:

$$\left[\mathbb{E}_{|\hat{m}_n=g}\left(m(\lambda_{t+1}',Y_{t+1}')-g(\lambda_{t+1}',Y_{t+1}')\right)^2\right]^{1/2}$$
$$\leq\left[\mathbb{E}_{|\hat{m}_n=g}\left(m(\lambda_{t+1}',Y_{t+1}')-g^{[t+1]}(\lambda_0',Y_0',\dots,Y_{t+1}')\right)^2\right]^{1/2}$$
$$+\left[\mathbb{E}_{|\hat{m}_n=g}\left(g^{[t+1]}(\lambda_0',Y_0',\dots,Y_{t+1}')-g(\lambda_{t+1}',Y_{t+1}')\right)^2\right]^{1/2}$$
$$\leq\left[\mathbb{E}_{|\hat{m}_n=g}\left(m(\lambda_{t+1}',Y_{t+1}')-g^{[t]}(\lambda_1',Y_1',\dots,Y_{t+1}')\right)^2\right]^{1/2} \tag{3.12}$$
$$+\Big[\mathbb{E}_{|\hat{m}_n=g}\Big(g^{[t]}(\lambda_1',Y_1',\dots,Y_{t+1}')$$
$$-g^{[t+1]}(\lambda_0',Y_0',\dots,Y_{t+1}')\Big)^2\Big]^{1/2} \tag{3.13}$$

$$+\left[\mathbb{E}_{|\hat{m}_n=g}\left(g^{[t+1]}(\lambda'_0,Y'_0,\dots,Y'_{t+1})-g(\lambda'_{t+1},Y'_{t+1})\right)^2\right]^{1/2}, \tag{3.14}$$

all three addends of which are treated separately. Regarding (3.13), note that $\hat{m}_n \in \mathscr{G}$ for any ω. Hence, the estimation has the contraction property, and for any $\omega \in \Omega$

$$\begin{aligned}
&\left|\hat{m}_n^{[t]}(\lambda'_1,Y'_1,\dots,Y'_{t+1})-\hat{m}_n^{[t+1]}(\lambda'_0,Y'_0,\dots,Y'_{t+1})\right| \\
&\le L_1\left|\hat{m}_n^{[t-1]}(\lambda'_1,Y'_1,\dots,Y'_t)-\hat{m}_n^{[t]}(\lambda'_0,Y'_0,\dots,Y'_t)\right| \\
&\le L_1^2\left|\hat{m}_n^{[t-2]}(\lambda'_1,Y'_1,\dots,Y'_{t-1})-\hat{m}_n^{[t-1]}(\lambda'_0,Y'_0,\dots,Y'_{t-1})\right| \\
&\quad\vdots \\
&\le L_1^t\left|\hat{m}_n(\lambda'_1,Y'_1)-\hat{m}_n^{[1]}(\lambda'_0,Y'_0,Y'_1)\right| \\
&\le ML_1^t.
\end{aligned}$$

To treat the term (3.14), we apply the contraction property of $\hat{m}_n$ once to conclude

$$\begin{aligned}
&\left|\hat{m}_n^{[t+1]}(\lambda'_0,Y'_0,\dots,Y'_{t+1})-\hat{m}_n(\lambda'_{t+1},Y'_{t+1})\right| \\
&=\left|\hat{m}_n\left(\hat{m}_n^{[t]}(\lambda'_0,Y'_0,\dots,Y'_t),Y'_{t+1}\right)-\hat{m}_n(\lambda'_{t+1},Y'_{t+1})\right| \\
&\le L_1\left|\hat{m}_n^{[t]}(\lambda'_0,Y'_0,\dots,Y'_t)-\lambda'_{t+1}\right| \\
&= L_1\left|\hat{m}_n^{[t]}(\lambda'_0,Y'_0,\dots,Y'_t)-m(\lambda'_t,Y'_t)\right|.
\end{aligned}$$

Plugging the previous two estimates into the lines (3.13) and (3.14), we obtain

$$\begin{aligned}
&\left[\mathbb{E}_{|\hat{m}_n=g}\left(m(\lambda'_{t+1},Y'_{t+1})-g(\lambda'_{t+1},Y'_{t+1})\right)^2\right]^{1/2} \\
&\le\left[\mathbb{E}_{|\hat{m}_n=g}\left(m(\lambda'_{t+1},Y'_{t+1})-g^{[t]}(\lambda'_1,Y'_1,\dots,Y'_{t+1})\right)^2\right]^{1/2} \\
&\quad+ML_1^t \\
&\quad+L_1\left[\mathbb{E}_{|\hat{m}_n=g}\left(g^{[t]}(\lambda'_0,Y'_0,\dots,Y'_t)-m(\lambda'_t,Y'_t)\right)^2\right]^{1/2}.
\end{aligned}$$

Now we want to use the stationarity of the process $\{(\lambda'_t,Y'_t)\}$. Since the data generating process is in the stationary regime, we conclude that

$$\begin{aligned}
&\left[\mathbb{E}_{|\hat{m}_n=g}\left(m(\lambda'_t,Y'_t)-g(\lambda'_t,Y'_t)\right)^2\right]^{1/2} \\
&\le\left[\mathbb{E}_{|\hat{m}_n=g}\left(m(\lambda'_t,Y'_t)-g^{[t]}(\lambda'_0,Y'_0,\dots,Y'_t)\right)^2\right]^{1/2} \\
&\quad+ML_1^t \\
&\quad+L_1\left[\mathbb{E}_{|\hat{m}_n=g}\left(g^{[t]}(\lambda'_0,Y'_0,\dots,Y'_t)-m(\lambda'_t,Y'_t)\right)^2\right]^{1/2}
\end{aligned}$$

$$= M L_1^t + (1 + L_1) \left[\mathbb{E}_{|\hat{m}_n = g} \left(m(\lambda'_t, Y'_t) - g^{[t]}(\lambda'_0, Y'_0, \ldots, Y'_t) \right)^2 \right]^{1/2}.$$

Using the condition

$$M L_1^t < \varepsilon \left[\mathbb{E}_{|\hat{m}_n = g} \left(m(\lambda'_t, Y'_t) - g(\lambda'_t, Y'_t) \right)^2 \right]^{1/2}$$

as well as the model assumption $L_1 < 1$, yields

$$\begin{aligned} (1 - \varepsilon) &\left[\mathbb{E}_{|\hat{m}_n = g} \left(m(\lambda'_t, Y'_t) - g(\lambda'_t, Y'_t) \right)^2 \right]^{1/2} \\ &< 2 \left[\mathbb{E}_{|\hat{m}_n = g} \left(m(\lambda'_t, Y'_t) - g^{[t]}(\lambda'_0, Y'_0, \ldots, Y'_t) \right)^2 \right]^{1/2}. \end{aligned} \tag{3.15}$$

Rewriting

$$m(\lambda'_t, Y'_t) = m^{[t]}(\lambda'_0, Y'_0, \ldots, Y'_t)$$

and inserting $\mp m^{[t]}(0, Y'_0, \ldots, Y'_t)$ and $\mp g^{[t]}(0, Y'_0, \ldots, Y'_t)$ respectively yields

$$\begin{aligned} &\left(m(\lambda'_t, Y'_t) - g^{[t]}(\lambda'_0, Y'_0, \ldots, Y'_t) \right)^2 \\ =&\left(m^{[t]}(\lambda'_0, Y'_0, \ldots, Y'_t) - g^{[t]}(\lambda'_0, Y'_0, \ldots, Y'_t) \right)^2 \\ =&\Big(m^{[t]}(\lambda'_0, Y'_0, \ldots, Y'_t) - m^{[t]}(0, Y'_0, \ldots, Y'_t) \\ &+ m^{[t]}(0, Y'_0, \ldots, Y'_t) - g^{[t]}(0, Y'_0, \ldots, Y'_t) \\ &+ g^{[t]}(0, Y'_0, \ldots, Y'_t) - g^{[t]}(\lambda'_0, Y'_0, \ldots, Y'_t) \Big)^2 \\ \leq& 3 \Big(\underbrace{m^{[t]}(\lambda'_0, Y'_0, \ldots, Y'_t) - m^{[t]}(0, Y'_0, \ldots, Y'_t)}_{|\cdot| \leq L_1^t M} \Big)^2 \\ &+ 3 \left(m^{[t]}(0, Y'_0, \ldots, Y'_t) - g^{[t]}(0, Y'_0, \ldots, Y'_t) \right)^2 \\ &+ 3 \Big(\underbrace{g^{[t]}(0, Y'_0, \ldots, Y'_t) - g^{[t]}(\lambda'_0, Y'_0, \ldots, Y'_t)}_{|\cdot| \leq L_1^t M} \Big)^2 \\ =& 6 M^2 L_1^{2t} + 3 \left(m^{[t]}(0, Y'_0, \ldots, Y'_t) - g^{[t]}(0, Y'_0, \ldots, Y'_t) \right)^2. \end{aligned}$$

Therefore, inequality (3.15) changes to

$$\begin{aligned} &\frac{(1 - \varepsilon)^2}{4} \mathbb{E}_{|\hat{m}_n = g} \left(m(\lambda'_t, Y'_t) - g(\lambda'_t, Y'_t) \right)^2 \\ &< 6 M^2 L_1^{2t} + 3 \, \mathbb{E}_{|\hat{m}_n = g} \left(m^{[t]}(0, Y'_0, \ldots, Y'_t) - g^{[t]}(0, Y'_0, \ldots, Y'_t) \right)^2. \end{aligned}$$

Finally, we obtain inequality (3.11) using the fact that $L_1^{2t} < L_1^t$:

$$\begin{aligned}&\frac{(1-\varepsilon)^2}{12}\,\mathbb{E}_{|\hat{m}_n=g}\left(m(\lambda_t',Y_t')-g(\lambda_t',Y_t')\right)^2-2M^2L_1^t\\&<\mathbb{E}_{|\hat{m}_n=g}\left(m^{[t]}(0,Y_0',\dots,Y_t')-g^{[t]}(0,Y_0',\dots,Y_t')\right)^2.\end{aligned}$$ □

The last computations required some lengthy displays. The responsible quantities will appear frequently during the course of this chapter. We introduce an abbreviating notation to facilitate the further proceedings.

DEFINITION 3.2.5. For integers $k \le l$, we define the vectors $\boldsymbol{Y}_k^l$ as

$$\boldsymbol{Y}_k^l := (0, Y_k, \dots, Y_l).$$

For an element $g \in \mathscr{G}$, the functional $f_t(g; \boldsymbol{Y}_{i-t}^{i+1})$ is defined as

$$\begin{aligned}f_t(g;\boldsymbol{Y}_{i-t}^{i+1}) &:= \left(Y_{i+1}-m^{[t]}(\boldsymbol{Y}_{i-t}^{i})\right)^2-\left(Y_{i+1}-g^{[t]}(\boldsymbol{Y}_{i-t}^{i})\right)^2\\&=\left(Y_{i+1}-m^{[t]}(0,Y_{i-t},\dots,Y_i)\right)^2\\&\qquad-\left(Y_{i+1}-g^{[t]}(0,Y_{i-t},\dots,Y_i)\right)^2.\end{aligned}$$

With respect to the ghost process $\{Y_t'\}_{t\in\mathbb{Z}}$, we stipulate the canonical notation $\boldsymbol{Y'}_k^l := (0, Y_k', \dots, Y_l')$.

To get acquainted with the new notation, we use it to summarize our current state of knowledge that we acquired after Lemma 3.2.3 and Lemma 3.2.4. Recall that $L_1 < 1$. This implies that $\lim_{t\to\infty} L_1^t = 0$. The combination of Lemma 3.2.3 and Lemma 3.2.4 enables us to draw the following conclusion. For $t \to \infty$ and almost all $\omega \in \Omega$ such that the condition (3.10) in Lemma 3.2.4 is satisfied, the expression

$$\mathbb{E}_{|\hat{m}_n=g}\left[\left(Y_{t+1}'-g^{[t]}(0,Y_0',\dots,Y_t')\right)^2-\left(Y_{t+1}'-m^{[t]}(0,Y_0',\dots,Y_t')\right)^2\right]$$

is an upper bound for the loss $L(\hat{m}_n, m)$, up to a negligible term. In the new notation, we have shown that

$$\mathbb{E}_{|\hat{m}_n=g}\left[-f_t(g;\boldsymbol{Y'}_{i-t}^{i+1})\right] \ge \frac{(1-\varepsilon)^2}{12}L(\hat{m}_n,m)-o(1). \tag{3.16}$$

Furthermore, inequality (3.7) can now be written as

$$\sum_{i=0}^{n-1} f_i(\hat{m}_n; \boldsymbol{Y}_0^{i+1}) \ge 0. \tag{3.17}$$

The next proposition gives an estimate of the difference between the normal-

ized sum $\frac{1}{n}\sum_{i=0}^{n-1} f_i(g;\boldsymbol{Y}_0^{i+1})$, where the number of iterations applied to g and m change with the index i, and the sum $\frac{1}{n-t}\sum_{i=t}^{n-1} f_t(g;\boldsymbol{Y}_{i-t}^{i+1})$, where the number of iterations stays fixed for each addend. For our purposes, we need this estimate to hold uniformly over the class $\mathscr{G}$. This will be the first of many instances in this proof where we deal with expressions of the form $\sup_{g\in\mathscr{G}_0} f_t(g;\boldsymbol{Y}_{i-t}^{i+1})$ for some subset $\mathscr{G}_0 \subset \mathscr{G}$ and some iteration index $0 \le t \le i$. Due diligence requires that we ensure measurability of these expressions before we proceed. At first glance, this is not obvious since the class $\mathscr{G}_0$ may in general be uncountable. However, the class $\mathscr{G}$ is separable and the functional $f_t(\cdot,\boldsymbol{Y}_{i-t}^{i+1})$ continuous, which allows to take the supremum over a countable subset.

LEMMA 3.2.6. *Let $T\colon \mathbb{R}\to\mathbb{R}$ be a continuous function. For any natural number $0 \le t \le i$ and any sub-class $\mathscr{G}_0 \subset \mathscr{G}$, the expression $\sup_{g\in\mathscr{G}_0} T(f_t(g;\boldsymbol{Y}_{i-t}^{i+1}))$ is $(\mathscr{F}-\mathscr{B})$ measurable.*

Proof. We show that the functional $g \mapsto f_t(g,\boldsymbol{Y}_{i-t}^{i+1})$ is continuous in $g \in (\mathscr{G}, \|\cdot\|_\infty)$. Let $g,h \in \mathscr{G}$. We observe

$$\begin{aligned}
&\left|f_t(g;\boldsymbol{Y}_{i-t}^{i+1}) - f_t(h;\boldsymbol{Y}_{i-t}^{i+1})\right| \\
&= \Big|\big(Y_{i+1} - m^{[t]}(\boldsymbol{Y}_{i-t}^{i})\big)^2 - \big(Y_{i+1} - g^{[t]}(\boldsymbol{Y}_{i-t}^{i})\big)^2 \\
&\qquad\qquad - \big(Y_{i+1} - m^{[t]}(\boldsymbol{Y}_{i-t}^{i})\big)^2 + \big(Y_{i+1} - h^{[t]}(\boldsymbol{Y}_{i-t}^{i})\big)^2\Big| \\
&= \Big|2Y_{i+1}\big(g^{[t]}(\boldsymbol{Y}_{i-t}^{i}) - h^{[t]}(\boldsymbol{Y}_{i-t}^{i})\big) + \big[h^{[t]}(\boldsymbol{Y}_{i-t}^{i})\big]^2 - \big[g^{[t]}(\boldsymbol{Y}_{i-t}^{i})\big]^2\Big| \\
&\le (2Y_{i+1} + 2M)\left|g^{[t]}(\boldsymbol{Y}_{i-t}^{i}) - h^{[t]}(\boldsymbol{Y}_{i-t}^{i})\right|.
\end{aligned}$$

We refer to the calculations preceding (3.3) to conclude that

$$\left|g^{[t]}(\boldsymbol{Y}_{i-t}^{i}) - h^{[t]}(\boldsymbol{Y}_{i-t}^{i})\right| \le \|g-h\|_\infty \sum_{i=0}^{\infty} L_1^i = \frac{\|g-h\|_\infty}{1-L_1}.$$

This proves continuity. Furthermore, $(\mathscr{G}, \|\cdot\|_\infty)$ is separable because it is totally bounded (Aliprantis and Border, 1994, Lemma 3.19). Therefore, there exists a countable dense subset $\tilde{\mathscr{G}} \subset \mathscr{G}$. We show that the set $\tilde{\mathscr{G}}$ can be chosen such that $\tilde{\mathscr{G}} \cap \mathscr{G}_0$ is dense in $\mathscr{G}_0$. Let $k \in \mathbb{N}$ be arbitrary. Since $\mathscr{G}$ is totally bounded, there exists a set $X_k = \{x^{(1,k)},\dots,x^{(n_k,k)}\} \subset \mathscr{G}$ such that the union of the balls $\bigcup_{i=1}^{n_k} B_\infty(x^{(i,k)},2^{-k})$ covers $\mathscr{G}$. For any $i \in \{1,\dots,n_k\}$, choose a representative $g_0^{(i,k)}$ from the set $B_\infty(x^{(i,k)},2^{-k}) \cap \mathscr{G}_0$. If this set is empty, simply take $g_0^{(i,k)} := x^{(i,k)}$. The set

$$\tilde{\mathscr{G}} := \bigcup_{k=1}^{\infty} \{x^{(1,k)}, g_0^{(1,k)}, \dots, x^{(n_k,k)}, g_0^{(n_k,k)}\}$$

is a countable dense subset of $\mathscr{G}$, and $\tilde{\mathscr{G}} \cap \mathscr{G}_0$ is dense in $\mathscr{G}_0$.

Assume now that $\sup_{g\in\mathscr{G}_0} T(f_t(g;\boldsymbol{Y}_{i-t}^{i+1})) \geq x$ for some $x\in\mathbb{R}$. Let $\varepsilon>0$ be arbitrary. There exists $g^*\in\mathscr{G}_0$ such that $T(f_t(g^*;\boldsymbol{Y}_{i-t}^{i+1}))>x-\frac{\varepsilon}{2}$. By continuity of $T\circ f_t(\cdot,\boldsymbol{Y}_{i-t}^{i-1})$ and the fact that $\tilde{\mathscr{G}}\cap\mathscr{G}_0$ is dense in $\mathscr{G}_0$, there exists $\tilde{g}\in\tilde{\mathscr{G}}\cap\mathscr{G}_0$ such that $|T(f_t(g^*;\boldsymbol{Y}_{i-t}^{i+1}))-T(f_t(\tilde{g};\boldsymbol{Y}_{i-t}^{i+1}))|<\frac{\varepsilon}{2}$. It follows that

$$\begin{aligned}\sup_{g\in\mathscr{G}_0\cap\tilde{\mathscr{G}}} T(f_t(g;\boldsymbol{Y}_{i-t}^{i+1})) &\geq T(f_t(\tilde{g};\boldsymbol{Y}_{i-t}^{i+1}))\\ &\geq T(f_t(g^*;\boldsymbol{Y}_{i-t}^{i+1}))-|T(f_t(g^*;\boldsymbol{Y}_{i-t}^{i+1}))-T(f_t(\tilde{g};\boldsymbol{Y}_{i-t}^{i+1}))|\\ &> x-\varepsilon.\end{aligned}$$

Therefore, we conclude that $\sup_{g\in\mathscr{G}_0\cap\tilde{\mathscr{G}}} T(f_t(g;\boldsymbol{Y}_{i-t}^{i+1}))\geq x$. The reverse implication follows from $\tilde{\mathscr{G}}\cap\mathscr{G}_0\subset\mathscr{G}_0$. We conclude that

$$\begin{aligned}\Big\{\sup_{g\in\mathscr{G}_0} T(f_t(g;\boldsymbol{Y}_{i-t}^{i+1}))\geq x\Big\} &= \Big\{\sup_{g\in\mathscr{G}_0\cap\tilde{\mathscr{G}}} T(f_t(g;\boldsymbol{Y}_{i-t}^{i+1}))\geq x\Big\}\\ &= \bigcap_{k=1}^{\infty}\bigcup_{g\in\tilde{\mathscr{G}}\cap\mathscr{G}_0}\Big\{T(f_t(g;\boldsymbol{Y}_{i-t}^{i+1}))>x-2^{-k}\Big\}\in\mathscr{F}.\end{aligned}$$ □

PROPOSITION 3.2.7. *With the notation from Definition 3.2.5, we have the following estimate:*

$$\mathbb{E}\sup_{g\in\mathscr{G}}\Big|\frac{1}{n}\sum_{i=0}^{n-1} f_i(g;\boldsymbol{Y}_0^{i+1})-\frac{1}{n-t}\sum_{i=t}^{n-1} f_t(g;\boldsymbol{Y}_{i-t}^{i+1})\Big|\lesssim\frac{t}{n}+L_1^t$$

Proof. First of all,

$$\begin{aligned}&\sup_{g\in\mathscr{G}}\Big|\frac{1}{n}\sum_{i=0}^{n-1} f_i(g;\boldsymbol{Y}_0^{i+1})-\frac{1}{n-t}\sum_{i=t}^{n-1} f_t(g;\boldsymbol{Y}_{i-t}^{i+1})\Big|\\ &\leq\frac{1}{n}\sum_{i=0}^{t-1}\sup_{g\in\mathscr{G}}\big|f_i(g;\boldsymbol{Y}_0^{i+1})\big|+\frac{1}{n-t}\sum_{i=t}^{n-1}\sup_{g\in\mathscr{G}}\Big|\frac{n-t}{n}f_i(g;\boldsymbol{Y}_0^{i+1})-f_t(g;\boldsymbol{Y}_{i-t}^{i+1})\Big|.\end{aligned}\tag{3.18}$$

We bound the expectation of the first sum on the right hand side of (3.18):

$$\begin{aligned}&\mathbb{E}\frac{1}{n}\sum_{i=0}^{t-1}\sup_{g\in\mathscr{G}}\big|f_i(g;\boldsymbol{Y}_0^{i+1})\big|\\ &=\mathbb{E}\frac{1}{n}\sum_{i=0}^{t-1}\sup_{g\in\mathscr{G}}\Big|\big(Y_{i+1}-m^{[i]}(\boldsymbol{Y}_0^i)\big)^2-\big(Y_{i+1}-g^{[i]}(\boldsymbol{Y}_0^i)\big)^2\Big|\\ &\leq\mathbb{E}\frac{1}{n}\sum_{i=0}^{t-1}\sup_{g\in\mathscr{G}}2\big(Y_{i+1}-g^{[i]}(\boldsymbol{Y}_0^i)\big)^2\\ &=\frac{2}{n}\sum_{i=0}^{t-1}\mathbb{E}\Big[\mathbb{E}_{|\mathscr{F}_i}\Big[\sup_{g\in\mathscr{G}}\big(Y_{i+1}-\mathbb{E}_{|\mathscr{F}_i}Y_{i+1}+\mathbb{E}_{|\mathscr{F}_i}Y_{i+1}-g^{[i]}(\boldsymbol{Y}_0^i)^2\Big]\Big]\\ &=\frac{2}{n}\sum_{i=0}^{t-1}\mathbb{E}\Big[\mathbb{E}_{|\mathscr{F}_i}\Big[\sup_{g\in\mathscr{G}}\Big(\big(Y_{i+1}-\mathbb{E}_{|\mathscr{F}_i}Y_{i+1}\big)^2+\big(\mathbb{E}_{|\mathscr{F}_i}Y_{i+1}-g^{[i]}(\boldsymbol{Y}_0^i)\big)^2\end{aligned}$$

$$
\begin{aligned}
&\qquad\qquad +2\big(Y_{i+1}-\mathbb{E}_{|\mathscr{F}_i}Y_{i+1}\big)\big(\mathbb{E}_{|\mathscr{F}_i}Y_{i+1}-g^{[i]}(\boldsymbol{Y}_0^i)\big)\Big)\Big]\Big]\\
&\le \frac{2}{n}\sum_{i=0}^{t-1}\mathbb{E}\Big[\mathbb{E}_{|\mathscr{F}_i}\Big[\big(Y_{i+1}-\mathbb{E}_{|\mathscr{F}_i}Y_{i+1}\big)^2+\sup_{g\in\mathscr{G}}\big(\mathbb{E}_{|\mathscr{F}_i}Y_{i+1}-g^{[i]}(\boldsymbol{Y}_0^i)\big)^2\\
&\qquad\qquad +2\sup_{g\in\mathscr{G}}\Big\{\big(Y_{i+1}-\mathbb{E}_{|\mathscr{F}_i}Y_{i+1}\big)\big(\mathbb{E}_{|\mathscr{F}_i}Y_{i+1}-g^{[i]}(\boldsymbol{Y}_0^i)\big)\Big\}\Big]\Big]\\
&= \frac{2}{n}\sum_{i=0}^{t-1}\mathbb{E}\Big[\underbrace{\mathrm{var}_{|\mathscr{F}_i}(Y_{i+1})}_{=\lambda_i\le M}+\mathbb{E}_{|\mathscr{F}_i}\sup_{g\in\mathscr{G}}\underbrace{\big(\lambda_{i+1}-g^{[i]}(\boldsymbol{Y}_0^i)\big)^2}_{\le M^2}\\
&\qquad\qquad +2\sup_{g\in\mathscr{G}}\Big[\big(Y_{i+1}-\lambda_{i+1}\big)\big(\lambda_{i+1}-g^{[i]}(\boldsymbol{Y}_0^i)\big)\Big]\Big].
\end{aligned}
$$

Furthermore,

$$
\begin{aligned}
&\mathbb{E}_{|\mathscr{F}_i}\Big(\sup_{g\in\mathscr{G}}\Big[\big(Y_{i+1}-\lambda_{i+1}\big)\big(\lambda_{i+1}-g^{[i]}(\boldsymbol{Y}_0^i)\big)\Big]\Big)\\
&= \mathbb{E}_{|\mathscr{F}_i}\Big(\big(Y_{i+1}-\lambda_{i+1}\big)\Big[\mathbb{1}_{\{Y_{i+1}-\lambda_{i+1}\ge 0\}}\sup_{g\in\mathscr{G}}\underbrace{\big(\lambda_{i+1}-g^{[i]}(\boldsymbol{Y}_0^i)\big)}_{\le M}\\
&\qquad\qquad +\mathbb{1}_{\{Y_{i+1}-\lambda_{i+1}<0\}}\inf_{g\in\mathscr{G}}\underbrace{\big(\lambda_{i+1}-g^{[i]}(\boldsymbol{Y}_0^i)\big)}_{\ge -M}\Big]\Big)\\
&\le M\,\mathbb{E}_{|\mathscr{F}_i}\Big[\big(Y_{i+1}-\lambda_{i+1}\big)\big(\mathbb{1}_{\{Y_{i+1}-\lambda_{i+1}\ge 0\}}-\mathbb{1}_{\{Y_{i+1}-\lambda_{i+1}<0\}}\big)\Big]\\
&= M\,\mathbb{E}_{|\mathscr{F}_i}\big|Y_{i+1}-\lambda_{i+1}\big|\\
&\le M\Big[\mathbb{E}_{|\mathscr{F}_i}\big(Y_{i+1}-\lambda_{i+1}\big)^2\Big]^{1/2}\\
&= M\sqrt{\mathrm{var}_{|\mathscr{F}_i}(Y_{i+1})}\\
&\le M^{3/2}.
\end{aligned}
$$

Thus, we can bound the expectation of the first sum on the right hand side of (3.18) by a multiple of t/n:

$$
\mathbb{E}\frac{1}{n}\sum_{i=0}^{t-1}\sup_{g\in\mathscr{G}}\big|f_i(g;\boldsymbol{Y}_0^{i+1})\big|\le 2(M+M^2+2M^{3/2})\frac{t}{n}.
$$

In order to bound the second sum, we use that for any g and $i>t$

$$
\begin{aligned}
\Big|\big(g^{[i]}(\boldsymbol{Y}_0^i)\big)^2-\big(g^{[t]}(\boldsymbol{Y}_{i-t}^i)\big)^2\Big| &\le 2M\big|g^{[i]}(\boldsymbol{Y}_0^i)-g^{[t]}(\boldsymbol{Y}_{i-t}^i)\big|\\
&= 2M\Big|g^{[t]}\big(g^{[i-t]}(\boldsymbol{Y}_0^{i-t-1}),Y_{i-t},\dots,Y_i\big)\\
&\qquad\qquad -g^{[t]}\big(0,Y_{i-t},\dots,Y_i\big)\Big|\\
&\le 2M\,L_1^t\,\big|g^{[i-t]}(\boldsymbol{Y}_0^{i-t-1})\big|\\
&\le 2M^2L_1^t.
\end{aligned}
$$

We can conclude,

$$
\begin{aligned}
&\mathbb{E}\sup_{g\in\mathscr{G}}\Big|\frac{n-t}{n}\big(Y_{i+1}-g^{[i]}(\boldsymbol{Y}_0^i)\big)^2-\big(Y_{i+1}-g^{[t]}(\boldsymbol{Y}_{i-t}^i)\big)^2\Big|\\
&=\mathbb{E}\sup_{g\in\mathscr{G}}\Big|\Big(\frac{n-t}{n}-1\Big)Y_{i+1}^2+\frac{n-t}{n}\underbrace{\Big(\big(g^{[i]}(\boldsymbol{Y}_0^i)\big)^2-\big(g^{[t]}(\boldsymbol{Y}_{i-t}^i)\big)^2\Big)}_{|\cdot|\le 2M^2L_1^t}+\big(\frac{n-t}{n}-1\big)\big(g^{[t]}(\boldsymbol{Y}_{i-t}^i)\big)^2\\
&\qquad -2Y_{i+1}\Big(\frac{n-t}{n}\underbrace{\big(g^{[i]}(\boldsymbol{Y}_0^i)-g^{[t]}(\boldsymbol{Y}_{i-t}^i)\big)}_{|\cdot|\le ML_1^t}+\big(\frac{n-t}{n}-1\big)g^{[t]}(\boldsymbol{Y}_{i-t}^i)\Big)\Big|\\
&\le\mathbb{E}\Big[\frac{t}{n}Y_{i+1}^2+\frac{n-t}{n}2M^2L_1^t+\frac{t}{n}M^2+2Y_{i+1}\Big(\frac{n-t}{n}ML_1^t+\frac{t}{n}M\Big)\Big].
\end{aligned}
$$

By $\mathbb{E}(Y_{i+1}^2\,|\,\mathscr{F}_i)=\operatorname{var}(Y_{i+1}\,|\,\mathscr{F}_i)+\big(\mathbb{E}(Y_{i+1}\,|\,\mathscr{F}_i)\big)^2\le M+M^2$ (a.s.), the last line is bounded by

$$
\frac{t}{n}\big[M+4M^2+2M^3\big]+L_1^t\big[\big(4M^2+2M^3\big)\frac{n-t}{n}\big]\lesssim\frac{t}{n}+L_1^t.
$$

Since $m\in\mathscr{G}$, we can use the estimate

$$
\begin{aligned}
&\Big|\frac{n-t}{n}\big(Y_{i+1}-m^{[i]}(\boldsymbol{Y}_0^i)\big)^2-\big(Y_{i+1}-m^{[t]}(\boldsymbol{Y}_{i-t}^i)\big)^2\Big|\\
&\le\sup_{g\in\mathscr{G}}\Big|\frac{n-t}{n}\big(Y_{i+1}-g^{[i]}(\boldsymbol{Y}_0^i)\big)^2-\big(Y_{i+1}-g^{[t]}(\boldsymbol{Y}_{i-t}^i)\big)^2\Big|.
\end{aligned}
$$

Thus, we obtain the desired estimate in virtue of

$$
\begin{aligned}
&\mathbb{E}\sup_{g\in\mathscr{G}}\Big|\frac{n-t}{n}f_i(g;\boldsymbol{Y}_0^{i+1})-f_t(g;\boldsymbol{Y}_{i-t}^{i+1})\Big|\\
&=\mathbb{E}\sup_{g\in\mathscr{G}}\Big|\frac{n-t}{n}\big(Y_{i+1}-m^{[i]}(\boldsymbol{Y}_0^i)\big)^2-\frac{n-t}{n}\big(Y_{i+1}-g^{[i]}(\boldsymbol{Y}_0^i)\big)^2\\
&\qquad\qquad-\big(Y_{i+1}-m^{[t]}(\boldsymbol{Y}_{i-t}^i)\big)^2+\big(Y_{i+1}-g^{[t]}(\boldsymbol{Y}_{i-t}^i)\big)^2\Big|\\
&\le\mathbb{E}\sup_{g\in\mathscr{G}}\Big[\Big|\frac{n-t}{n}\big(Y_{i+1}-g^{[i]}(\boldsymbol{Y}_0^i)\big)^2-\big(Y_{i+1}-g^{[t]}(\boldsymbol{Y}_{i-t}^i)\big)^2\Big|\\
&\qquad\qquad+\Big|\frac{n-t}{n}\big(Y_{i+1}-m^{[i]}(\boldsymbol{Y}_0^i)\big)^2-\big(Y_{i+1}-m^{[t]}(\boldsymbol{Y}_{i-t}^i)\big)^2\Big|\Big]\\
&\le 2\,\mathbb{E}\sup_{g\in\mathscr{G}}\Big|\frac{n-t}{n}\big(Y_{i+1}-g^{[i]}(\boldsymbol{Y}_0^i)\big)^2-\big(Y_{i+1}-g^{[t]}(\boldsymbol{Y}_{i-t}^i)\big)^2\Big|\\
&\lesssim\frac{t}{n}+L_1^t.
\end{aligned}
$$

□

We are prepared to formulate the first essential lemma in the asymptotic analysis of the least squares estimator. It combines inequalities (3.16) and (3.17) to derive a first estimate on the exceedance probability of the risk $L(\hat{m}_n,m)$.

LEMMA 3.2.8. *Let $\delta = \delta(n) = n^{-1/3}\log n$ be as in Theorem 3.2.2 and $t = t(n) = -\frac{2}{3\log L_1}\log n$. Then there exists a positive constant $\gamma > 0$ such that for almost all $n \in \mathbb{N}$*

$$\begin{aligned}
&\mathbb{P}\Big\{\mathbb{E}_{|\hat{m}_n=g}\big[m(\lambda_0', Y_0') - g(\lambda_0', Y_0')\big]^2 > \delta^2\Big\} \\
&\le \mathbb{P}\bigcup_{k=0}^{\infty}\Bigg\{\sup_{\substack{g\in\mathscr{G}:\\ \mathbb{E}[m^{[t]}(\boldsymbol{Y}_0^t)-g^{[t]}(\boldsymbol{Y}_0^t)]^2\le 2^{2k+2}\gamma\delta^2}} \frac{1}{n-t}\sum_{i=t}^{n-1}\big(f_t(g;\boldsymbol{Y}_{i-t}^{i+1}) - \mathbb{E}f_t(g;\boldsymbol{Y}_{i-t}^{i+1})\big) > 2^{2k-2}\gamma\delta^2\Bigg\} \\
&\qquad + \mathbb{P}\Big\{\Delta_n > \frac{\gamma\delta^2}{4}\Big\},
\end{aligned}$$

where $\mathbb{E}\Delta_n \lesssim \frac{t}{n} + L_1^t$.

Proof. We see that $L_1^t = e^{-\log(L_1)\frac{2}{3\log(L_1)}\log(n)} = n^{-2/3}$ and conclude that for some $\varepsilon \in (0,1)$ the relation $\varepsilon\delta > ML_1^t$ holds for almost all $n \in \mathbb{N}$. Hence, there exists a number n_0 such that on the set

$$\Omega_0 := \Big\{\omega \in \Omega\colon \mathbb{E}_{|\hat{m}_n=g}\big[m(\lambda_0', Y_0') - g(\lambda_0', Y_0')\big]^2 > \delta^2\Big\}$$

the condition of Lemma 3.2.4 is satisfied for all $n \in \mathbb{N}$ with $n > n_0$. As a consequence, there exists a number $n_1 \ge n_0$ such that for almost all $\omega \in \Omega_0$ we obtain the following two relations for all $n \in \mathbb{N}$ with $n > n_1$. As a consequence of Lemma 3.2.4,

$$\begin{aligned}
&\mathbb{E}_{|\hat{m}_n=g}\Big[m^{[t]}(0, Y_0', \dots, Y_t') - g^{[t]}(0, Y_0', \dots, Y_t')\Big]^2 \\
&\ge \frac{(1-\varepsilon)^2}{12}\mathbb{E}_{|\hat{m}_n=g}\big[m(\lambda_0', Y_0') - g(\lambda_0', Y_0')\big]^2 - 2M^2L_1^t \\
&> \frac{(1-\varepsilon)^2}{12}\delta^2 - \underbrace{2M^2L_1^t}_{=o(\delta^2)} \\
&\ge \underbrace{\frac{(1-\varepsilon)^2}{24}}_{:=\gamma}\delta^2,
\end{aligned} \tag{3.19}$$

since $\limsup_{n\to\infty} 2M^2L_1^t/\delta^2 = \limsup_{n\to\infty} 2M(\log n)^{-2} = 0$, and therefore $2M^2L_1^t < \gamma\delta^2$ for almost all $n \in \mathbb{N}$. By Lemma 3.2.3 and the fact that $\limsup_{n\to\infty} 3M^2L_1^t/\delta^2 < \gamma/2$, we obtain the second important relation:

$$\begin{aligned}
&\mathbb{E}_{|\hat{m}_n=g}\Big[\overbrace{\Big(Y_{t+1}' - g^{[t]}(0, Y_0', \dots, Y_t')\Big)^2 - \Big(Y_{t+1}' - m^{[t]}(0, Y_0', \dots, Y_t')\Big)^2}^{=-f_t(g;\boldsymbol{Y'}_0^{t+1})}\Big] \\
&\quad \ge \mathbb{E}_{|\hat{m}_n=g}\Big[m^{[t]}(0, Y_0', \dots, Y_t') - g^{[t]}(0, Y_0', \dots, Y_t')\Big]^2 - 3M^2L_1^t \\
&\quad > \mathbb{E}_{|\hat{m}_n=g}\Big[m^{[t]}(0, Y_0', \dots, Y_t') - g^{[t]}(0, Y_0', \dots, Y_t')\Big]^2 - \frac{\gamma}{2}\delta^2
\end{aligned} \tag{3.20}$$

for almost all $n \in \mathbb{N}$. By the preceding two ω-wise relations, we obtain for $n > n_1$ that Ω_0 is, up to a null set, contained in the set of all ω such that the relations (3.19) and (3.20) hold. Using the abbreviating notation $(0, Y_0', \ldots, Y_t') =: \boldsymbol{Y'}_0^t$ from Definition 3.2.5, the monotonicity of the probability measure $\mathbb{P}$ lets us conclude that

$$\begin{aligned}
&\mathbb{P}\left\{\mathbb{E}_{|\hat{m}_n=g}\left[m(\lambda_0', Y_0') - g(\lambda_0', Y_0')\right]^2 > \delta^2\right\} \\
&\leq \mathbb{P}\left\{\mathbb{E}_{|\hat{m}_n=g}\left[m^{[t]}(\boldsymbol{Y'}_0^t) - g^{[t]}(\boldsymbol{Y'}_0^t)\right]^2 > \gamma\delta^2;\right. && \text{by (3.19)} \\
&\qquad \left.\mathbb{E}_{|\hat{m}_n=g}\left[-f_t(g; \boldsymbol{Y'}_0^{t+1})\right] > \mathbb{E}_{|\hat{m}_n=g}\left[m^{[t]}(\boldsymbol{Y'}_0^t) - g^{[t]}(\boldsymbol{Y'}_0^t)\right]^2 - \frac{\gamma}{2}\delta^2\right\}, && \text{by (3.20)}
\end{aligned}$$

for $n \in \mathbb{N}$ with $n > n_1$. By the definition of the least squares estimator, we have the relation

$$\sum_{i=0}^{n-1} f_i(\hat{m}_n; \boldsymbol{Y}_0^{i+1}) = \sum_{i=0}^{n-1}\left(Y_{i+1} - m^{[i]}(\boldsymbol{Y}_0^i)\right)^2 - \sum_{i=0}^{n-1}\left(Y_{i+1} - \hat{m}_n^{[i]}(\boldsymbol{Y}_0^i)\right)^2 \geq 0. \quad (3.21)$$

Note that in this relation we have to use the original sample $Y_0, \ldots, Y_n$ as opposed to the ghost sample $Y_0', \ldots, Y_n'$ since the least squares estimator is based on observations of the original count process. As a consequence of relation (3.21),

$$\begin{aligned}
&\mathbb{P}\left\{\mathbb{E}_{|\hat{m}_n=g}\left[m^{[t]}(\boldsymbol{Y'}_0^t) - g^{[t]}(\boldsymbol{Y'}_0^t)\right]^2 > \gamma\delta^2;\right. \\
&\qquad \left.\mathbb{E}_{|\hat{m}_n=g}\left[-f_t(g; \boldsymbol{Y'}_0^{t+1})\right] > \mathbb{E}_{|\hat{m}_n=g}\left[m^{[t]}(\boldsymbol{Y'}_0^t) - g^{[t]}(\boldsymbol{Y'}_0^t)\right]^2 - \frac{\gamma}{2}\delta^2\right\} \\
&\leq \mathbb{P}\left\{\mathbb{E}_{|\hat{m}_n=g}\left[m^{[t]}(\boldsymbol{Y'}_0^t) - g^{[t]}(\boldsymbol{Y'}_0^t)\right]^2 > \gamma\delta^2;\right. \\
&\underbrace{\frac{1}{n}\sum_{i=0}^{n-1} f_i(\hat{m}_n; \boldsymbol{Y}_0^{i+1})}_{\geq 0, \text{ by (3.21)}} - \mathbb{E}_{|\hat{m}_n=g} f_t(g; \boldsymbol{Y'}_0^{t+1}) > \left.\mathbb{E}_{|\hat{m}_n=g}\left[m^{[t]}(\boldsymbol{Y'}_0^t) - g^{[t]}(\boldsymbol{Y'}_0^t)\right]^2 - \frac{\gamma}{2}\delta^2\right\}.
\end{aligned}$$

So far, the event inside the probability concerns the behavior of the specific function $\hat{m}_n \in \mathscr{G}$. Of course, any event describing that a specific $g_0 \in \mathscr{G}$ has a certain property is contained in the event that *there exists* a function $g \in \mathscr{G}$ that has the described property. The argument becomes clearer if we use the fact that by independence of $\{Y_i\}$ and $\{Y_i'\}$

$$\mathbb{E}_{|\hat{m}_n=g}\left[m^{[t]}(\boldsymbol{Y'}_0^t) - g^{[t]}(\boldsymbol{Y'}_0^t)\right]^2 = \int \left[m^{[t]}(y) - \hat{m}_n^{[t]}(y)\right]^2 \mathbb{P}^{\boldsymbol{Y'}_0^t}(dy) \text{ a.s.}$$

and

$$\mathbb{E}_{|\hat{m}_n=g} f_t(g; \boldsymbol{Y'}_0^{t+1}) = \int f_t(\hat{m}_n; y)\, \mathbb{P}^{\boldsymbol{Y'}_0^{t+1}}(dy) \text{ a.s.}.$$

Call $\Omega_1 \subset \Omega$ the set of ω for which the two relations

$$\int \left[m^{[t]}(y) - \hat{m}_n^{[t]}(y) \right]^2 \mathbb{P}^{\boldsymbol{Y'}_0^t}(dy) > \gamma\delta^2$$

and

$$\frac{1}{n}\sum_{i=0}^{n-1} f_i(\hat{m}_n; \boldsymbol{Y}_0^{i+1}) - \int f_t(\hat{m}_n; y)\, \mathbb{P}^{\boldsymbol{Y'}_0^{t+1}}(dy) > \int \left[m^{[t]}(y) - \hat{m}_n^{[t]}(y) \right]^2 \mathbb{P}^{\boldsymbol{Y'}_0^t}(dy) - \frac{\gamma}{2}\delta^2$$

hold. Because $\hat{m}_n$ is always in $\mathscr{G}$, the set Ω_1 is contained in the set of all ω for which *there exists* some function $g \in \mathscr{G}$ such that

$$\int \left[m^{[t]}(y) - g^{[t]}(y) \right]^2 \mathbb{P}^{\boldsymbol{Y'}_0^t}(dy) > \gamma\delta^2$$

and

$$\frac{1}{n}\sum_{i=0}^{n-1} f_i(g; \boldsymbol{Y}_0^{i+1}) - \int f_t(g; y)\, \mathbb{P}^{\boldsymbol{Y'}_0^{t+1}}(dy) > \int \left[m^{[t]}(y) - g^{[t]}(y) \right]^2 \mathbb{P}^{\boldsymbol{Y'}_0^t}(dy) - \frac{\gamma}{2}\delta^2.$$

Note that the integrals $\int \left[m^{[t]}(y) - g^{[t]}(y) \right]^2 \mathbb{P}^{\boldsymbol{Y'}_0^t}(dy)$ and $\int f_t(g; y)\, \mathbb{P}^{\boldsymbol{Y'}_0^{t+1}}(dy)$ do not dependent any more on the original sample $Y_0, \ldots, Y_n$. This allows us to write them almost surely as unconditional expectations. Moreover, we can use that $\{Y_i\}$ and $\{Y_i'\}$ have the same distribution to take these unconditional expectations with respect to $\{Y_i\}$ as opposed to $\{Y_i'\}$:

$$\int \left[m^{[t]}(y) - g^{[t]}(y) \right]^2 \mathbb{P}^{\boldsymbol{Y'}_0^t}(dy) = \mathbb{E}\left[m^{[t]}(\boldsymbol{Y}_0^t) - g^{[t]}(\boldsymbol{Y}_0^t) \right]^2$$
$$\int f_t(g; y)\, \mathbb{P}^{\boldsymbol{Y'}_0^{t+1}}(dy) = \mathbb{E} f_t(g; \boldsymbol{Y}_0^{t+1}).$$

We obtain the bound

$$\begin{aligned}
&\mathbb{P}\Big\{ \mathbb{E}_{|\hat{m}_n = g}\left[m^{[t]}(\boldsymbol{Y'}_0^t) - g^{[t]}(\boldsymbol{Y'}_0^t) \right]^2 > \gamma\delta^2; \\
&\frac{1}{n}\sum_{i=0}^{n-1} f_i(\hat{m}_n; \boldsymbol{Y'}_0^{i+1}) - \mathbb{E}_{|\hat{m}_n = g} f_t(g; \boldsymbol{Y'}_0^{t+1}) > \mathbb{E}_{|\hat{m}_n = g}\left[m^{[t]}(\boldsymbol{Y'}_0^t) - g^{[t]}(\boldsymbol{Y'}_0^t) \right]^2 - \frac{\gamma}{2}\delta^2 \Big\} \\
&\leq \mathbb{P}\Big\{ \exists g \in \mathscr{G} \colon \mathbb{E}\left[m^{[t]}(\boldsymbol{Y}_0^t) - g^{[t]}(\boldsymbol{Y}_0^t) \right]^2 > \gamma\delta^2; \qquad (3.22) \\
&\frac{1}{n}\sum_{i=0}^{n-1} f_i(g; \boldsymbol{Y}_0^{i+1}) - \mathbb{E} f_t(g; \boldsymbol{Y}_0^{t+1}) > \mathbb{E}\left[m^{[t]}(\boldsymbol{Y}_0^t) - g^{[t]}(\boldsymbol{Y}_0^t) \right]^2 - \frac{\gamma}{2}\delta^2 \Big\}.
\end{aligned}$$

Now we use the fact

$$\left\{\mathbb{E}\left[m^{[t]}(\boldsymbol{Y}_0^t)-g^{[t]}(\boldsymbol{Y}_0^t)\right]^2>\delta^2\gamma\right\}$$
$$=\bigcup_{k=0}^{\infty}\left\{2^{2k+2}\delta^2\gamma\geq\mathbb{E}\left[m^{[t]}(\boldsymbol{Y}_0^t)-g^{[t]}(\boldsymbol{Y}_0^t)\right]^2>2^{2k}\delta^2\gamma\right\}$$

to write (3.22) further as

$$\mathbb{P}\bigcup_{k=0}^{\infty}\Bigg\{\exists g\in\mathscr{G}:\ 2^{2k+2}\gamma\delta^2\geq\mathbb{E}\left[m^{[t]}(\boldsymbol{Y}_0^t)-g^{[t]}(\boldsymbol{Y}_0^t)\right]^2>2^{2k}\gamma\delta^2;$$
$$\frac{1}{n}\sum_{i=0}^{n-1}f_i(g;\boldsymbol{Y}_0^{i+1})-\mathbb{E}f_t(g;\boldsymbol{Y}_0^{t+1})>\underbrace{\mathbb{E}\left[m^{[t]}(\boldsymbol{Y}_0^t)-g^{[t]}(\boldsymbol{Y}_0^t)\right]^2}_{>2^{2k}\gamma\delta^2\text{ in this event}}-\frac{\gamma}{2}\delta^2\Bigg\}$$
$$\leq\mathbb{P}\bigcup_{k=0}^{\infty}\Bigg\{\exists g\in\mathscr{G}:\ 2^{2k+2}\gamma\delta^2\geq\mathbb{E}\left[m^{[t]}(\boldsymbol{Y}_0^t)-g^{[t]}(\boldsymbol{Y}_0^t)\right]^2;$$
$$\frac{1}{n}\sum_{i=0}^{n-1}f_i(g;\boldsymbol{Y}_0^{i+1})-\mathbb{E}f_t(g;\boldsymbol{Y}_0^{t+1})>2^{2k-1}\gamma\delta^2\Bigg\}$$
$$\leq\mathbb{P}\bigcup_{k=0}^{\infty}\Bigg\{\sup_{\substack{g\in\mathscr{G}:\\ \mathbb{E}[m^{[t]}(\boldsymbol{Y}_0^t)-g^{[t]}(\boldsymbol{Y}_0^t)]^2\leq 2^{2k+2}\gamma\delta^2}}\frac{1}{n}\sum_{i=0}^{n-1}f_i(g;\boldsymbol{Y}_0^{i+1})-\mathbb{E}f_t(g;\boldsymbol{Y}_0^{t+1})>2^{2k-1}\gamma\delta^2\Bigg\}.$$

This is almost the statement of the lemma. We just have to substitute the term $\frac{1}{n}\sum_{i=0}^{n-1}f_i(g;\boldsymbol{Y}_0^{i+1})$ by $\frac{1}{n-t}\sum_{i=t}^{n-1}f_t(g;\boldsymbol{Y}_{i-t}^{i+1})$, which we do with the help of Proposition 3.2.7. We invoke the triangle inequality for probabilities and the fact that by stationarity $\mathbb{E}f_t(g;\boldsymbol{Y}_0^{t+1})=\mathbb{E}f_t(g;\boldsymbol{Y}_{i-t}^{i+1})$ for $i\geq t$ to conclude

$$\mathbb{P}\bigcup_{k=0}^{\infty}\Bigg\{\sup_{\substack{g\in\mathscr{G}:\\ \mathbb{E}[m^{[t]}(\boldsymbol{Y}_0^t)-g^{[t]}(\boldsymbol{Y}_0^t)]^2\leq 2^{2k+2}\gamma\delta^2}}\frac{1}{n}\sum_{i=0}^{n-1}f_i(g;\boldsymbol{Y}_0^{i+1})-\mathbb{E}f_t(g;\boldsymbol{Y}_0^{t+1})>2^{2k-1}\gamma\delta^2\Bigg\}$$
$$\leq\mathbb{P}\bigcup_{k=0}^{\infty}\Bigg\{\sup_{\substack{g\in\mathscr{G}:\\ \mathbb{E}[m^{[t]}(\boldsymbol{Y}_0^t)-g^{[t]}(\boldsymbol{Y}_0^t)]^2\leq 2^{2k+2}\gamma\delta^2}}\frac{1}{n-t}\sum_{i=t}^{n-1}\left(f_t(g;\boldsymbol{Y}_{i-t}^{i+1})-\mathbb{E}f_t(g;\boldsymbol{Y}_{i-t}^{i+1})\right)>2^{2k-2}\gamma\delta^2\Bigg\}$$
$$+\mathbb{P}\Bigg\{\underbrace{\sup_{g\in\mathscr{G}}\left|\frac{1}{n}\sum_{i=0}^{n-1}f_i(g;\boldsymbol{Y}_0^{i+1})-\frac{1}{n-t}\sum_{i=t}^{n-t}f_t(g;\boldsymbol{Y}_{i-t}^{i+1})\right|}_{=:\Delta_n}>\frac{\gamma\delta^2}{4}\Bigg\}$$

This yields the assertion in view of Proposition 3.2.7. □

The last lemma provides the insight that the crucial step in proving closeness of $\hat{m}_n$ to m consists of finding a uniform bound for the trajectories of the stochastic process

$$\left\{\frac{1}{n-t}\sum_{i=t}^{n-1}\left(f_t(g;\boldsymbol{Y}_{i-t}^{i+1})-\mathbb{E}f_t(g;\boldsymbol{Y}_{i-t}^{i+1})\right)\right\}_{g\in\mathscr{G}}$$

in the balls

$$\mathscr{G}_k := \left\{ g \in \mathscr{G} \colon \mathbb{E}\big[m^{[t]}(\boldsymbol{Y}_0^t) - g^{[t]}(\boldsymbol{Y}_0^t)\big]^2 \leq 2^{2k+2}\gamma\delta^2 \right\}.$$

This fact reflects the underlying idea outlined at the beginning of the section. Suppose that the trajectories of the process are smooth in the balls $\mathscr{G}_k$ in such a way that their fluctuations around zero are reasonably small at the periphery of the balls $\mathscr{G}_k$, and decrease further as we move from the periphery towards the center. Then the approximation of $\mathbb{E}f(g;\boldsymbol{Y}_0^{t+1})$ by its empirical analogue is sufficiently accurate to distinguish an element $g \in \mathscr{G}$ from the true function m, even if g is close to m. In this case it is a justified belief that $\hat{m}_n$ is close to m.

The rest of the section is devoted to bound the oscillations of the above process in probability. This is a classical object of interest in the theory of empirical processes. The available tools from the standard theory work well for processes based on i.i.d. samples of random variables. In order to use these tools, we once more employ the coupling method. The aim is to find a coupling $(S,\Sigma,P,(V',V^*))$ such that V' has the same distribution as $\{\boldsymbol{Y}_{i-t}^{i+1}\}_{i\in\mathbb{Z}}$, and the process V^* is a sequence of q-dependent random variables. Once we have established such a coupling, we can in a first step replace $\mathbb{P}^{\{\boldsymbol{Y}_{i-t}^{i+1}\}_{i\in\mathbb{N}}}$ by $P^{\{V_i'\}_{i\in\mathbb{N}}}$, and further bound $P^{\{V_i'\}_{i\in\mathbb{N}}}$ by $P^{V^*} + d_{TV}(P^{\{V_i'\}_{i\in\mathbb{N}}}, P^{\{V_i^*\}_{i\in\mathbb{N}}})$. The merit of this procedure is that we end up dealing with independent blocks of random variables allowing us to apply the tools from classical empirical process theory. The total variation error term will be negligible due to the mixing property of the process $\{\boldsymbol{Y}_{i-t}^{i+1}\}_{i\in\mathbb{Z}}$.

The described procedure is an established strategy in the asymptotic analysis of absolutely regular processes (Doukhan, 1994; Doukhan et al., 1995). Doukhan (1994, page 36) employed it to prove an exponential tail inequality for sums of absolutely regular random sequences. As uniform mixing implies absolute regularity (Doukhan, 1994, page 4), we may adapt the idea to our setting.

LEMMA 3.2.9. *Let $\{\boldsymbol{Y}_{i-t}^{i+1} \colon i \in \mathbb{Z}\}$ be the sequence of random variables on $(\Omega,\mathscr{F},\mathbb{P})$ that is defined by $\boldsymbol{Y}_{i-t}^{i+1} = (0, Y_{i-t},\ldots,Y_{i+1})$. For $q \in \mathbb{N}_+$, there exists a probability space (S,Σ,P), and on this space two processes $V' = \{V_i'\}_{i\in\mathbb{Z}}$ and $V^* = \{V_{t+i}^*\}_{i\in\mathbb{N}}$ with values in $\mathbb{R}^{t+3}$,*

$$\begin{aligned} V_i' &= \big(0, Y_{i-t}', \ldots, Y_{i+1}'\big) \\ V_i^* &= \big(0, Y_{i-t}^*, \ldots, Y_{i+1}^*\big), \end{aligned}$$

such that the following statements hold.

(i) The process V' has the same distribution as $\{\boldsymbol{Y}_{i-t}^{i+1}\}$, i.e. $P^{V'} = \mathbb{P}^{\{\boldsymbol{Y}_{i-t}^{i+1}\}}$.

(ii) The process V^* is q-dependent, i.e. the block sequences,

$$\left\{\boldsymbol{V}^*_{2j} = \left(V^*_{t+2jq}, V^*_{t+2jq+1}, \ldots, V^*_{t+2jq+q-1}\right);\ j \in \mathbb{N}\right\}$$
$$\left\{\boldsymbol{V}^*_{2j+1} = \left(V^*_{t+(2j+1)q}, V^*_{t+(2j+1)q+1}, \ldots, V^*_{t+(2j+1)q+q-1}\right);\ j \in \mathbb{N}\right\}$$

are i.i.d., respectively.

(iii) $P^{\left(V'_{jq+1},\ldots,V'_{jq+q}\right)} = P^{\left(V^*_{jq+1},\ldots,V^*_{jq+q}\right)}$ *for any j.*

Furthermore, let $\phi^t(n)$ be the nth coefficient of uniform mixing corresponding to the process $\{\boldsymbol{Y}^{i+1}_{i-t}\}$. Then

$$P\left\{\exists\, i \in \{t,\ldots,n-1\}\colon V'_i \neq V^*_i\right\} \leq \frac{n-t}{q}\phi^t(q).$$

Proof. Apply Lemma A.1.2. □

Note that we applied the coupling to $\{\boldsymbol{Y}^{i+1}_{i-t}\}$ and not to the original count process $\{Y_i\}$, which means that the Y^*_j only formally describe the coordinates of the V^*_i. In order to operate with analogues of $\boldsymbol{Y}^i_{i-t}$ on the coupling space (S,Σ,P), we need to introduce the variables $Z^*_i = (0, Y^*_{i-t}, \ldots, Y^*_i)$ that are to be understood canonically to $\boldsymbol{Y}^i_{i-t} = (0, Y_{i-t}, \ldots, Y_i)$. Both Y^*_i and Z^*_i are formally defined as projections of V^*_t. By measurability of theses projections, we conclude that $P^{Y^*_i} = \mathbb{P}^{Y_i}$ as well as $P^{Z^*_i} = \mathbb{P}^{\boldsymbol{Y}^i_{i-t}}$.

In the following proceedings, we have to distinguish strictly between the two probability spaces $(\Omega,\mathscr{F},\mathbb{P})$ and (S,Σ,P). On the first space, the original data generating process is defined. The second one is a mere technical construction to facilitate the dependence structure in virtue of the previous coupling lemma and thereby enabling further probabilistic bounds. Whenever a line is concerned with the original process $\{Y_i\}$, we have to use the measure $\mathbb{P}$ and the corresponding expectation operator $\mathbb{E}$. On the other hand, when we deal with the auxiliary process $\{V^*_i\}$, we must use P and the corresponding expectation E.

The next corollary shows how we use Lemma 3.2.9 to translate the bound from Lemma 3.2.8 in a bound involving P and $\{V^*_i\}$.

COROLLARY 3.2.10. *Let E denote the expectation with respect to P. Then,*

$$\begin{aligned}
&\mathbb{P}\left\{\mathbb{E}_{|\hat{m}_n = g}\left[m(\lambda'_0, Y'_0) - g(\lambda'_0, Y'_0)\right]^2 > \delta^2\right\} \\
&\quad\leq \sum_{k=0}^{\infty} P\left\{\sup_{\substack{g\in\mathscr{G} \\ E[m^{[t]}(Z^*_t) - g^{[t]}(Z^*_t)]^2 \leq 2^{2k+2}\gamma\delta^2}} \frac{1}{n-t}\sum_{i=t}^{n-1}\left(f_t(g;V^*_i) - Ef_t(g;V^*_i)\right) > 2^{2k-2}\gamma\delta^2\right\} \\
&\qquad + \frac{n-t}{q}\phi^t(q) + \mathbb{P}\left\{\Delta_n > \frac{\gamma\delta}{4}\right\}.
\end{aligned}$$

Proof. First of all, apply Lemma 3.2.8. As Z_t^* and $\boldsymbol{Y}_0^t$ are equal in law, it is justified to substitute $\mathbb{E}[m^{[t]}(\boldsymbol{Y}_0^t) - g^{[t]}(\boldsymbol{Y}_0^t)]^2$ by $E[m^{[t]}(Z_t^*) - g^{[t]}(Z_t^*)]^2$. Since $\mathbb{P}^{\{\boldsymbol{Y}_{i-t}^{i+1}\}} = P^{\{V_i'\}}$, we can change $\mathbb{P}$ to P if we change $\boldsymbol{Y}_{i-t}^{i+1}$ to V_i'. The rest follows by the ω wise substitution of V_i' by V_i^* on the Σ-measurable set $A_n := \{\omega \in S : V_i'(\omega) = V_i^*(\omega) \text{ for all } i = t, \dots, n\}$.

$$
\begin{aligned}
&P \bigcup_{k=0}^{\infty} \Bigg\{ \sup_{\substack{g \in \mathscr{G} \\ E[m^{[t]}(Z_t^*) - g^{[t]}(Z_t^*)]^2 \le 2^{2k+2}\gamma\delta^2}} \frac{1}{n-t} \sum_{i=t}^{n-1} \left(f_t(g; V_i') - E f_t(g; V_i') \right) > 2^{2k-2}\gamma\delta^2 \Bigg\} \\
&\le P\Bigg(A_n \cap \bigcup_{k=0}^{\infty} \Bigg\{ \sup_{\substack{g \in \mathscr{G} \\ E[m^{[t]}(Z_t^*) - g^{[t]}(Z_t^*)]^2 \le 2^{2k+2}\gamma\delta^2}} \frac{1}{n-t} \sum_{i=t}^{n-1} \left(f_t(g; V_i^*) - E f_t(g; V_i^*) \right) \\
&\qquad\qquad > 2^{2k-2}\gamma\delta^2 \Bigg\} \Bigg) + P(A_n^c)
\end{aligned}
$$

A bound for $P(A_n^c)$ is given by the Coupling Lemma 3.2.9. The claim follows by an application of Bonferroni's union bound. □

Henceforth, we want to make heavy use of the fact that the block sequences $\{\boldsymbol{V}_{2j+1}^*\}_{j \in \mathbb{N}}$ and $\{\boldsymbol{V}_{2j}^*\}_{j \in \mathbb{N}}$ are independent and identically distributed. We introduce the variables X_j^* to describe the partial sums

$$
\begin{aligned}
X_{2j}^*(g) &:= \frac{1}{q} \sum_{i=0}^{q-1} \left(f_t(g; V_{t+2jq+i}^*) - E f_t(g; V_{t+2jq+i}^*) \right) \\
X_{2j+1}^*(g) &:= \frac{1}{q} \sum_{i=0}^{q-1} \left(f_t(g; V_{t+(2j+1)q+i}^*) - E f_t(g; V_{t+(2j+1)q+i}^*) \right),
\end{aligned}
$$

and make the following rearrangements. We want to write the total sum

$$
\frac{1}{n-t} \sum_{i=t}^{n-1} \left(f_t(g; V_i^*) - E f_t(g; V_i^*) \right) \tag{3.23}
$$

as a sum of N blocks X_{2j}^* with even indexes and N blocks X_{2j+1}^* with uneven indexes. Since we want the numbers of "even" and "uneven" blocks to be equal, and the total number of addends, $n-t$, is not necessarily a multiple of the supposed block length, we are faced with a remainder term of asymptotically negligible size.

We formally define the number $N = N(n,t,q)$ as follows. The crucial parameter for the definition is the total number of summands, $n-t$, divided by the supposed block length q. Set

$$
N := \begin{cases} \frac{1}{2} \left\lfloor \frac{n-t}{q} \right\rfloor & \text{if } \left\lfloor \frac{n-t}{q} \right\rfloor \text{ is even,} \\ \frac{1}{2} \left(\left\lfloor \frac{n-t}{q} \right\rfloor - 1 \right) & \text{if } \left\lfloor \frac{n-t}{q} \right\rfloor \text{ is odd.} \end{cases}
$$

Now, we can find the following expression for the sum (3.23) as a sum of even and uneven blocks:

$$\begin{aligned}
&\frac{1}{n-t}\sum_{i=t}^{n-1}\left(f_t(g;V_i^*)-Ef_t(g;V_i^*)\right)\\
&=\frac{1}{n-t}\sum_{i=0}^{n-1-t}\left(f_t(g;V_{t+i}^*)-Ef_t(g;V_{t+i}^*)\right)\\
&=\frac{1}{n-t}\sum_{i=0}^{2Nq-1}\left(f_t(g;V_{t+i}^*)-Ef_t(g;V_{t+i}^*)\right)+R_n(g)\\
&=\frac{q}{n-t}\sum_{j=0}^{2N-1}\frac{1}{q}\sum_{i=0}^{q-1}\left(f_t(g;V_{t+jq+i}^*)-Ef_t(g;V_{t+jq+i}^*)\right)+R_n(g)\\
&=\frac{q}{n-t}\sum_{j=0}^{N-1}X_{2j}^*(g)+\frac{q}{n-t}\sum_{j=0}^{N-1}X_{2j+1}^*(g)+R_n(g). \qquad (3.24)
\end{aligned}$$

In the above display, the remainder term R_n is the partial sum of the addends $f_t(g;V_i^*)-Ef_t(g;V_i^*)$ with indexes in the range of $i=t+2Nq,\dots,n-1$:

$$R_n(g)=\frac{1}{n-t}\sum_{i=2Nq}^{n-1-t}\left(f_t(g;V_{t+i}^*)-Ef_t(g;V_{t+i}^*)\right). \qquad (3.25)$$

The number of addends in the remainder term does not exceed $2q$.

We remark that now all terms but the remainder are sums of centered i.i.d. random variables. Using the property of the process $\{V_i^*\}$ to be q-dependent, we can establish a variance bound for the sums of independent blocks. This will be necessary later.

LEMMA 3.2.11. *Assume that the quantities δ and t are given by $\delta(n)=n^{-1/3}\log n$ and $t(n)=-\frac{2}{3\log L_1}\log n>0$. Let the length of the blocks $X_{2j}^*(g)$ also depend on the sample size in such a way that $q(n)\asymp t(n)$. Recall the constant γ introduced in the proof of Lemma 3.2.8. Then the following statements hold:*

$$\sup_{E\left[m^{[t]}(Z_t^*)-g^{[t]}(Z_t^*)\right]^2\le 2^{2k+2}\gamma\delta^2}\operatorname{var}\left(\frac{q}{n-t}\sum_{j=0}^{N-1}X_{2j}^*(g)\right)\le(M^2+2M^{3/2}+M)2^{2k+4}\gamma\delta^2\frac{q}{n-t}.$$

And as a consequence, there exists a number n_0 such that

$$\sup_{E\left[m^{[t]}(Z_t^*)-g^{[t]}(Z_t^*)\right]^2\le 2^{2k+2}\gamma\delta^2}\operatorname{var}\left(\frac{q}{n-t}\sum_{j=0}^{N-1}X_{2j}^*(g)\right)\le\left(2^{2k-4}\gamma\delta^2\right)^2/8$$

for all $k\in\mathbb{N}$ and all $n>n_0$.

Proof. By $N < \frac{n-t}{q}$, the fact that $E X_{2j}^*(g) = 0$ for all j, and independence and identical distribution of the blocks $\{X_{2j}^*(g)\}$, we obtain

$$\begin{aligned} E\left[\frac{q}{n-t}\sum_{j=0}^{N-1} X_{2j}^*(g)\right]^2 &= \frac{q^2}{(n-t)^2}\sum_{j_1,j_2=0}^{N-1} E X_{2j_1}^*(g) X_{2j_2}^*(g) \\ &= \frac{q^2}{(n-t)^2} N\, E\left[X_0^*(g)\right]^2 \\ &\le \frac{q}{n-t} E\left[X_0^*(g)\right]^2 \\ &= \frac{q}{n-t}\frac{1}{q^2}\sum_{i_1,i_2=0}^{q-1} \operatorname{cov}(f_t(g;V_{i_1+t}^*), f_t(g;V_{i_2+t}^*)). \end{aligned}$$

We use the stationarity of the process V^* and the Cauchy Schwartz Inequality to derive for $i_1, i_2 = 0, 1, \ldots, q-1$ the general bound

$$\begin{aligned} &\operatorname{cov}(f_t(g;V_{i_1+t}^*), f_t(g;V_{i_2+t}^*)) \\ &= E\big(f_t(g;V_{i_1+t}^*) f_t(g;V_{i_2+t}^*)\big) - E\big(f_t(g;V_{t+i_1}^*)\big) E\big(f_t(g;V_{t+i_2}^*)\big) \\ &\le \sqrt{E\big(f_t(g;V_{i_1+t}^*)\big)^2}\sqrt{E\big(f_t(g;V_{i_2+t}^*)\big)^2} - \Big(E\big(f_t(g;V_t^*)\big)\Big)^2 \\ &\le E\big(f_t(g;V_t^*)\big)^2. \end{aligned}$$

This means

$$E\left[\frac{q}{n-t}\sum_{j=0}^{N-1} X_{2j}^*(g)\right]^2 \le \frac{q}{n-t} E\big(f_t(g;V_t^*)\big)^2. \tag{3.26}$$

Let us recall that $Z_t^* = \big(0, Y_0^*, \ldots, Y_t^*\big)$ is measurable with respect to the σ-field $\mathscr{F}_t^* := \sigma\{Y_s^* : s \le t\}$. Therefore,

$$\begin{aligned} &E_{|\mathscr{F}_t^*} \sup_{\alpha\in[0,M]} \Big[\big(Y_{t+1}^* - \alpha\big)\big(m^{[t]}(Z_t^*) - g^{[t]}(Z_t^*)\big)\Big]^2 \\ &= E_{|\mathscr{F}_t^*}\Big[\big(m^{[t]}(Z_t^*) - g^{[t]}(Z_t^*)\big)^2 \sup_{\alpha\in[0,M]}\big(Y_{t+1}^* - \alpha\big)^2\Big] \\ &= \big(m^{[t]}(Z_t^*) - g^{[t]}(Z_t^*)\big)^2 E_{|\mathscr{F}_t^*}\Big[\sup_{\alpha\in[0,M]}\big(Y_{t+1}^* - E_{|\mathscr{F}_t^*}Y_{t+1}^* + E_{|\mathscr{F}_t^*}Y_{t+1}^* - \alpha\big)^2\Big] \\ &\le \big(m^{[t]}(Z_t^*) - g^{[t]}(Z_t^*)\big)^2\Big[E_{|\mathscr{F}_t^*}\Big(Y_{t+1}^* - E_{|\mathscr{F}_t^*}Y_{t+1}^*\Big)^2 + \sup_{\alpha\in[0,M]}\big(\underbrace{E_{|\mathscr{F}_t^*}Y_{t+1}^* - \alpha}_{|\cdot|\le M}\big)^2 \\ &\qquad\qquad +2\sup_{\alpha\in[0,M]}\Big(\big|E_{|\mathscr{F}_t^*}Y_{t+1}^* - \alpha\big|\Big)\Big(E_{|\mathscr{F}_t^*}\big|Y_{t+1}^* - E_{|\mathscr{F}_t^*}Y_{t+1}^*\big|\Big)\Big] \\ &\le \big(m^{[t]}(Z_t^*) - g^{[t]}(Z_t^*)\big)^2\Big[\operatorname{var}_{|\mathscr{F}_t^*}\big(Y_{t+1}^*\big) + M^2 + 2M\sqrt{\operatorname{var}_{|\mathscr{F}_t^*}\big(Y_{t+1}^*\big)}\Big] \\ &\le (M^2 + 2M^{3/2} + M)\big(m^{[t]}(Z_t^*) - g^{[t]}(Z_t^*)\big)^2 \ \text{ a.s.}. \end{aligned}$$

Taking expectations yields

$$E \sup_{\alpha\in[0,M]} \Big[\big(Y^*_{t+1}-\alpha\big)\big(m^{[t]}(Z^*_t)-g^{[t]}(Z^*_t)\big)\Big]^2 \\ \le \big(M^2+2M^{3/2}+M\big)E\big(m^{[t]}(Z^*_t)-g^{[t]}(Z^*_t)\big)^2. \qquad (3.27)$$

Therefore,

$$\begin{aligned} E\big(f_t(g;V^*_t)\big)^2 &= E\Big[\big(Y^*_{t+1}-m^{[t]}(Z^*_t)\big)^2-\big(Y^*_{t+1}-g^{[t]}(Z^*_t)\big)^2\Big]^2 \\ &= E\Big[\big(m^{[t]}(Z^*_t)\big)^2-\big(g^{[t]}(Z^*_t)\big)^2-2Y^*_{t+1}\big(m^{[t]}(Z^*_t)-g^{[t]}(Z^*_t)\big)\Big]^2 \\ &\le E\Big[\big(m^{[t]}(Z^*_t)+g^{[t]}(Z^*_t)\big)\big(m^{[t]}(Z^*_t)-g^{[t]}(Z^*_t)\big) \\ &\qquad -2Y^*_{t+1}\big(m^{[t]}(Z^*_t)-g^{[t]}(Z^*_t)\big)\Big]^2 \\ &= 4\,E\left[\Big(Y^*_{t+1}-\frac{m^{[t]}(Z^*_t)+g^{[t]}(Z^*_t)}{2}\Big)\big(m^{[t]}(Z^*_t)-g^{[t]}(Z^*_t)\big)\right]^2 \\ &\le 4\,E\sup_{\alpha\in[0,M]}\Big[\big(Y^*_{t+1}-\alpha\big)\big(m^{[t]}(Z^*_t)-g^{[t]}(Z^*_t)\big)\Big]^2 \qquad (3.28) \\ &\le 4\big(M^2+2M^{3/2}+M\big)E\big(m^{[t]}(Z^*_t)-g^{[t]}(Z^*_t)\big)^2. \end{aligned}$$

In line (3.28) we used the fact that $\frac{1}{2}(m^{[t]}(Z^*_t)+g^{[t]}(Z^*_t))\in[0,M]$, followed by an application of inequality (3.27). In conclusion, we find the bound

$$\begin{aligned} &\sup_{E[m^{[t]}(Z^*_t)-g^{[t]}(Z^*_t)]^2\le 2^{2k+2}\gamma\delta^2} \operatorname{var}\left(\frac{q}{n-t}\sum_{i=0}^{N}X^*_{2j}(g)\right) \\ =\ &\sup_{E[m^{[t]}(Z^*_t)-g^{[t]}(Z^*_t)]^2\le 2^{2k+2}\gamma\delta^2} E\Big(\frac{q}{n-t}\sum_{i=0}^{N}X^*_{2j}(g)\Big)^2 \\ \le\ &\sup_{E[m^{[t]}(Z^*_t)-g^{[t]}(Z^*_t)]^2\le 2^{2k+2}\gamma\delta^2} \frac{q}{n-t}E\big(f_t(g;V^*_t)\big)^2 \\ \le\ &\frac{q}{n-t}\,4(M^2+2M^{3/2}+M)2^{2k+2}\gamma\delta^2. \qquad (3.29) \end{aligned}$$

We have proven the first statement of the lemma. The second one is a simple consequence. Since $q(n)\asymp\log n$, it is evident that

$$\limsup_{n\to\infty}\frac{q}{(n-t)\delta^2}=0.$$

We conclude that

$$\sup_{E[m^{[t]}(Z^*_t)-g^{[t]}(Z^*_t)]^2\le 2^{2k+2}\gamma\delta^2} \operatorname{var}\left(\frac{q}{n-t}\sum_{i=0}^{N}X^*_{2j}(g)\right)\le 2^{2k-8}\gamma^2\delta^4/8 \qquad (3.30)$$

for all k and almost all n if the conditions in the formulation of the lemma are satisfied. This is an even tighter result than we need. Multiplying the bound

(3.30) by $2^{2k} \geq 1$ gives the formulated result. □

The technical step incorporated by the next lemma is often called symmetrization. It is an established tool in the asymptotic analysis of empirical processes (Van der Vaart and Wellner, 1996; Giné and Nickl, 2016) and has been successfully applied in the analysis of least squares estimators for nonparametric regression (Györfi et al., 2002). Roughly speaking, we add some randomness to the sum of blocks in order to facilitate further computations. The lemma is an immediate consequence of the symmetrization Lemma A.3.1, which is the source of the technical main condition.

LEMMA 3.2.12. *Let $\{\varepsilon_i\}_{i\in\mathbb{N}}$ be an i.i.d. sequence of Rademacher random variables on (S,Σ,P), i.e. $P\{\varepsilon_i = 1\} = P\{\varepsilon_i = -1\} = \frac{1}{2}$. Assume that the sequence $\{\varepsilon_i\}$ is independent of the process $\{V_i^*\}$ and that the quantities δ and t are defined as in Lemma 3.2.11. Then there exists a number n_0 such that for any $k \in \mathbb{N}$ and $n > n_0$*

$$\begin{aligned}
&P\Bigg\{ \sup_{\substack{g\in\mathscr{G} \\ E[m^{[t]}(Z_t^*)-g^{[t]}(Z_t^*)]^2\leq 2^{2k+2}\gamma\delta^2}} \frac{1}{n-t}\sum_{i=t}^{n-1}\big(f_t(g;V_i^*) - Ef_t(g;V_i^*)\big) > 2^{2k-2}\gamma\delta^2 \Bigg\} \\
&\leq 8P\Bigg\{ \sup_{\substack{g\in\mathscr{G} \\ E[m^{[t]}(Z_t^*)-g^{[t]}(Z_t^*)]^2\leq 2^{2k+2}\gamma\delta^2}} \frac{q}{n-t}\sum_{j=0}^{N-1}\varepsilon_j X_{2j}^*(g) > 2^{2k-6}\gamma\delta^2 \Bigg\} \\
&\quad + P\Bigg\{ \sup_{\substack{g\in\mathscr{G} \\ E[m^{[t]}(Z_t^*)-g^{[t]}(Z_t^*)]^2\leq 2^{2k+2}\gamma\delta^2}} R_n(g) > 2^{2k-3}\gamma\delta^2 \Bigg\}
\end{aligned}$$

Proof. First, we use the introduced variables $X_j^*(g)$ and equation (3.24). Since

$$P^{\sum_{j=0}^{N-1} X_{2j}^*(g)} = P^{\sum_{j=0}^{N-1} X_{2j+1}^*(g)},$$

we obtain after two applications of the triangle inequality for probabilities the relation

$$\begin{aligned}
&P\Bigg\{ \sup_{\substack{g\in\mathscr{G} \\ E[m^{[t]}(Z_t^*)-g^{[t]}(Z_t^*)]^2\leq 2^{2k+2}\gamma\delta^2}} \frac{1}{n-t}\sum_{i=t}^{n-1}\big(f_t(g;V_i^*) - Ef_t(g;V_i^*)\big) > 2^{2k-2}\gamma\delta^2 \Bigg\} \\
&\quad \leq 2P\Bigg\{ \sup_{\substack{g\in\mathscr{G} \\ E[m^{[t]}(Z_t^*)-g^{[t]}(Z_t^*)]^2\leq 2^{2k+2}\gamma\delta^2}} \frac{q}{n-t}\sum_{j=0}^{N-1} X_{2j}^*(g) > 2^{2k-4}\gamma\delta^2 \Bigg\} \\
&\qquad + P\Bigg\{ \sup_{\substack{g\in\mathscr{G} \\ E[m^{[t]}(Z_t^*)-g^{[t]}(Z_t^*)]^2\leq 2^{2k+2}\gamma\delta^2}} R_n(g) > 2^{2k-3}\gamma\delta^2 \Bigg\}.
\end{aligned}$$

In order to apply the Symmetrization Lemma A.3.1 on the first probability, we

have to check the variance condition (A.2), which reads in our case

$$\sup_{E\left[m^{[t]}(Z_t^*)-g^{[t]}(Z_t^*)\right]^2\le 2^{2k+2}\gamma\delta^2} \operatorname{var}\left(\frac{q}{n-t}\sum_{j=0}^{N-1} X_{2j}^*(g)\right) \le \left(2^{2k-4}\gamma\delta^2\right)^2/8.$$

But this is exactly the second statement of Lemma 3.2.11. Thus, the variance condition is satisfied for all $n > n_0$, for some $n_0 \in \mathbb{N}$, and all $k \in \mathbb{N}$. Applying the Symmetrization Lemma, we learn that there exists an independent copy $\{X_{2j}^{**}(g)\}$ of $\{X_{2j}^*(g)\}$ on (S,Σ,P) such that

$$\begin{aligned} &P\left\{\sup_{\substack{g\in\mathscr{G}:\\ E\left[m^{[t]}(Z_t^*)-g^{[t]}(Z_t^*)\right]^2\le 2^{2k+2}\gamma\delta^2}} \frac{q}{n-t}\sum_{j=0}^{N-1} X_{2j}^*(g) > 2^{2k-4}\gamma\delta^2\right\} \\ &\le 2P\left\{\sup_{\substack{g\in\mathscr{G}:\\ E\left[m^{[t]}(Z_t^*)-g^{[t]}(Z_t^*)\right]^2\le 2^{2k+2}\gamma\delta^2}} \frac{q}{n-t}\sum_{j=0}^{N-1}\left[X_{2j}^*(g)-X_{2j}^{**}(g)\right] > 2^{2k-5}\gamma\delta^2\right\}. \end{aligned}$$

Now we come to the point where we add some extra randomness in form of the sequence of Rademacher variables. To that end, we recall the following fact. Let $\{\xi_i\}_{i\in\mathbb{N}}$ be a sequence of independent random variables on (S,Σ,P) such that for all i and for all $A \in \mathscr{B}$ we have $P\{\xi_i \in A\} = P\{-\xi_i \in A\}$. Then for any $\theta = (\theta_1,\dots,\theta_m)' \in \{-1,1\}^m$ and $A_1,\dots,A_m \in \mathscr{B}$, we have

$$P\bigcap_{i=1}^m \{\theta_i\xi_i \in A_i\} = \prod_{i=1}^m P\{\theta_i\xi_i \in A_i\} = \prod_{i=1}^m P\{\xi_i \in A_i\} = P\bigcap_{i=1}^m \{\xi_i \in A_i\}.$$

Now, for the given i.i.d. sequence of Rademacher random variables $\{\varepsilon_i\}$, which is assumed to be independent of $\{\xi_i\}$, we observe

$$\begin{aligned} &P\{(\varepsilon_1\xi_1,\dots,\varepsilon_m\xi_m)\in A_1\times\ldots\times A_m\} \\ &= \sum_{\theta\in\{-1,1\}^m} P\{(\theta_1\xi_1,\dots,\theta_m\xi_m)\in A_1\times\ldots\times A_m;\ (\varepsilon_1,\dots,\varepsilon_m)=(\theta_1,\dots,\theta_m)\} \\ &= \sum_{\theta\in\{-1,1\}^m} P\{(\theta_1\xi_1,\dots,\theta_m\xi_m)\in A_1\times\ldots\times A_m\}P\{(\varepsilon_1,\dots,\varepsilon_m)=(\theta_1,\dots,\theta_m)\} \\ &= P\{(\xi_1,\dots,\xi_m)\in A_1\times\ldots\times A_m\}. \end{aligned}$$

Since $\left\{X_{2i}^*(g)-X_{2i}^{**}(g)\right\}$ is a sequence of independent symmetric random variables, the above argument shows that an independent random sign change in each addend does not change the distribution of the whole sum. Taken into account the fact that

$$P^{\sum_{j=0}^{N-1}\varepsilon_j X_{2j}^*} = P^{-\sum_{j=0}^{N-1}\varepsilon_j X_{2j}^{**}},$$

we conclude once again with the triangle inequality

$$
\begin{aligned}
&P\Big\{ \sup_{\substack{g\in\mathscr{G}:\\ E\left[m^{[t]}(Z_t^*)-g^{[t]}(Z_t^*)\right]^2\le 2^{2k+2}\gamma\delta^2}} \frac{q}{n-t}\sum_{j=0}^{N-1}\left(X_{2j}^*(g)-X_{2j}^{**}(g)\right) > 2^{2k-5}\gamma\delta^2\Big\} \\
&= P\Big\{ \sup_{\substack{g\in\mathscr{G}:\\ E\left[m^{[t]}(Z_t^*)-g^{[t]}(Z_t^*)\right]^2\le 2^{2k+2}\gamma\delta^2}} \frac{q}{n-t}\sum_{j=0}^{N-1}\varepsilon_j\left(X_{2j}^*(g)-X_{2j}^{**}(g)\right) > 2^{2k-5}\gamma\delta^2\Big\} \\
&\le 2P\Big\{ \sup_{\substack{g\in\mathscr{G}:\\ E\left[m^{[t]}(Z_t^*)-g^{[t]}(Z_t^*)\right]^2\le 2^{2k+2}\gamma\delta^2}} \frac{q}{n-t}\sum_{j=0}^{N-1}\varepsilon_j X_{2j}^*(g) > 2^{2k-6}\gamma\delta^2\Big\}. \qquad \square
\end{aligned}
$$

We will see in a short while that the remainder term R_n is of small order. As a result, we are endowed with a considerably easier modification of the original problem. This is the case since the object of interest is now a sum of independent and identically distributed random variables. In other words, the object

$$\Big\{\frac{q}{n-t}\sum_{j=0}^{N-1}\varepsilon_j X_{2j}^*(g)\Big\}_{g\in\mathscr{G}}$$

is a classical empirical process, save for the standardizing factor. In order to bound its supremum over subsets of $\mathscr{G}$, we employ the so called chaining technique. It is a well established tool in empirical process literature, and standard references include Van der Vaart and Wellner (1996), Van de Geer (2000), and Giné and Nickl (2016). For a good exposition in the context of nonparametric regression see also Györfi et al. (2002).

We do not describe this method in full generality. Instead, we immediately adapt the abstract idea as presented in the just mentioned references to our specific problem at hand. Before we proceed to the chaining argument, we shall prove three auxiliary propositions. The first one gives a bound of the maximum of observed count variables.

PROPOSITION 3.2.13. *The maximum of n observations of the count process is of stochastic order $O_{\mathbb{P}}(\log n)$. In fact, for almost all n*

$$\mathbb{P}\Big\{\max_{1\le i\le n} Y_i > 2(k+1)\log n\Big\} \le e^M\, n^{-(k+1)}$$

Proof. We use Lemma A.4.5 and the bound $|\lambda_i(\omega)| \le M$ to obtain the following chain of inequalities for almost all n:

$$
\begin{aligned}
&\mathbb{P}\Big\{\max_{1\le i\le n} Y_i > 2(k+2)\log n\Big\} \\
&\le \sum_{i=1}^{n}\mathbb{E}\Big[\mathbb{P}\big\{Y_i-\lambda_i > 2(k+2)\log(n)-\lambda_i \;\big|\; \lambda_i\big\}\Big]
\end{aligned}
$$

$$
\begin{aligned}
&\leq \sum_{i=1}^{n} \mathbb{E}\Big[\exp\Big(- \frac{\big(2(k+2)\log(n) - \lambda_i\big)^2}{2\lambda_i + \frac{2}{3}(2(k+2)\log(n) - \lambda_i)} \Big) \Big] \\
&\leq \sum_{i=1}^{n} \mathbb{E}\Big[\exp\Big(- \frac{\big(2(k+2)\log(n)\big)^2 + \lambda_i^2 - 4(k+2)\log(n)\lambda_i}{4(k+2)\log(n)} \Big) \Big] \\
&\leq \sum_{i=1}^{n} \mathbb{E}\Big[\exp\Big(- (k+2)\log(n) + \lambda_i \Big) \Big] \\
&\leq \exp\Big(\log(n) - (k+2)\log(n) + M \Big) \\
&= e^M n^{-(k+1)}.
\end{aligned}
$$

□

PROPOSITION 3.2.14. *Let $\mathscr{G}_\varepsilon$ be an ε-cover of $\mathscr{G}$ with respect to the uniform norm, i.e. for all $g \in \mathscr{G}$ there exists a $g^\varepsilon \in \mathscr{G}_\varepsilon$ such that $\|g - g^\varepsilon\|_\infty \leq \varepsilon$. Then, for $V_i^* = (0, Y_{i-t}^*, \ldots, Y_{i+1}^*)$ we have the following bound:*

$$
\big| f_t(g; V_i^*) - f_t(g^\varepsilon; V_i^*) \big| \leq \frac{2(Y_{i+1}^* + M)}{1 - L_1} \varepsilon .
$$

Proof. The proof is essentially the same as for the continuity of the functional Q_n, which we verified in the proof of Proposition 3.1.3. We observe that

$$
\begin{aligned}
&\big| f_t(g; V_i^*) - f_t(g^\varepsilon; V_i^*) \big| \\
&= \Big| \big(Y_{i+1}^* - m^{[t]}(Z_i^*)\big)^2 - \big(Y_{i+1}^* - g^{[t]}(Z_i^*)\big)^2 - \big(Y_{i+1}^* - m^{[t]}(Z_i^*)\big)^2 + \big(Y_{i+1}^* - (g^\varepsilon)^{[t]}(Z_i^*)\big)^2 \Big| \\
&= \Big| 2Y_{i+1}^* \big(g^{[t]}(Z_i^*) - (g^\varepsilon)^{[t]}(Z_i^*) \big) + \big[(g^\varepsilon)^{[t]}(Z_i^*) \big]^2 - \big[g^{[t]}(Z_i^*) \big]^2 \Big| \\
&= \Big| 2Y_{i+1}^* \big(g^{[t]}(Z_i^*) - (g^\varepsilon)^{[t]}(Z_i^*) \big) + \big[(g^\varepsilon)^{[t]}(Z_i^*) + g^{[t]}(Z_i^*) \big] \big[(g^\varepsilon)^{[t]}(Z_i^*) - g^{[t]}(Z_i^*) \big] \Big| \\
&\leq \big[2Y_{i+1}^* + \big(g^{[t]}(Z_i^*) + (g^\varepsilon)^{[t]}(Z_i^*) \big) \big] \big| g^{[t]}(Z_i^*) - (g^\varepsilon)^{[t]}(Z_i^*) \big| \\
&\leq (2Y_{i+1}^* + 2M) \big| g^{[t]}(Z_i^*) - (g^\varepsilon)^{[t]}(Z_i^*) \big| .
\end{aligned}
$$

We refer to the calculations preceding (3.3) to conclude that

$$
\big| g^{[t]}(Z_i^*) - (g^\varepsilon)^{[t]}(Z_i^*) \big| \leq \varepsilon \sum_{i=0}^{\infty} L_1^i = \frac{\varepsilon}{1 - L_1} .
$$

□

COROLLARY 3.2.15. *Let $q = q(n)$ depend on the sample size n. The expectation of the remainder term from Lemma 3.2.12, $E \sup_{g \in \mathscr{G}} R_n(g)$, is of order $O(q/(n-t))$.*

Proof. We already mentioned just below equation (3.25) that the number of addends in the remainder term does not exceed $2q$. Moreover, since $f_t(m; V_i^*) = 0$, we can conclude with the help of Proposition 3.2.14 that

$$
\begin{aligned}
\big| f_t(g; V_i^*) \big| &= \big| f_t(g; V_i^*) - f_t(m; V_i^*) \big| \\
&\leq \frac{2\|m - g\|_\infty}{1 - L_1} (Y_{i+1}^* + M)
\end{aligned}
$$

$$\leq \frac{2M}{1-L_1}(Y^*_{i+1}+M).$$

We can therefore conclude,

$$\begin{aligned}
E\big[\sup_{g\in\mathscr{G}}|R_n(g)|\big] &= E\Big[\sup_{g\in\mathscr{G}}\frac{1}{n-t}\Big|\sum_{i=2Nq}^{n-1}\big(f_t(g;V^*_{t+i})-Ef_t(g;V^*_{t+i})\big)\Big|\Big]\\
&\leq E\Big[\sup_{g\in\mathscr{G}}\frac{1}{n-t}\sum_{i=2Nq}^{n-1}\big|f_t(g;V^*_{t+i})\big|+E\big|f_t(g;V^*_{t+i})\big|\Big]\\
&\leq \frac{1}{n-t}\sum_{i=2Nq}^{n-1}\frac{2M}{1-L_1}\,2E\big(Y^*_{t+i+1}+M\big)\\
&\leq \frac{q}{n-t}\frac{16M^2}{1-L_1}.
\end{aligned}$$

□

Note that in Proposition 3.2.14 the bound on the difference between $g^{[t]}$ and $(g^\varepsilon)^{[t]}$ only depends on the distance $\|g-g^\varepsilon\|_\infty$, and it is independent of the sample size n. This will be essential for the development of the chaining argument. In order to estimate the supremum of the empirical process over the set

$$\mathscr{G}_k=\Big\{g\in\mathscr{G}: E\big[m^{[t]}(Z^*_t)-g^{[t]}(Z^*_t)\big]^2\leq 2^{2k+2}\gamma\delta^2\Big\}, \tag{3.31}$$

we seek a sequence of finite sets $\{\mathscr{G}_k^{(s)}\}_{s\in\mathbb{N}}$ that constitute $\|\cdot\|_\infty$-ball coverings $\big\{B(g_{s,k};r_s)\colon g_{s,k}\in\mathscr{G}_k^{(s)}\big\}$ of $\mathscr{G}_k$, with radii $\{r_s\}$ decreasing in s. Specifically, for $s\in\mathbb{N}$, let $\mathscr{G}_k^{(s)}\subset\mathscr{G}_k$ be a set such that for any $g\in\mathscr{G}_k$ there exists $g_{s,k}\in\mathscr{G}_k^{(s)}$ such that

$$\|g-g_{s,k}\|_\infty\leq 2^{-s}2^{k+1}\sqrt{\gamma}\delta\,.$$

The next proposition states that these sets $\mathscr{G}_k^{(s)}$ are well defined. Moreover, we have a bound on the number of elements of these sets.

PROPOSITION 3.2.16. *For any $k\in\mathbb{N}$, let the subset $\mathscr{G}_k\subset\mathscr{G}$ be defined as in* (3.31)*. Then, for any $s\in\mathbb{N}$, there exists a set $\mathscr{G}_k^{(s)}\subset\mathscr{G}_k$ with at most $e^{2MB2^{s-k}/(\sqrt{\gamma}\delta)}$ elements and a selection function $\pi_{s,k}\colon\mathscr{G}_k\to\mathscr{G}_k^{(s)}$ such that for any $g\in\mathscr{G}_k$*

$$\|g-\pi_{s,k}g\|_\infty\leq 2^{-s}2^{k+1}\sqrt{\gamma}\delta\,.$$

For $g\in\mathscr{G}_k$ we fix the notation $\pi_{s,k}g=:g_{s,k}$.

Proof. Let $\epsilon>0$ be arbitrary. We have seen in Lemma 3.1.5 that it takes at most

$$N:=N\big(\epsilon,\mathscr{G},\|\cdot\|_\infty\big)\leq e^{2MB/\epsilon}$$

balls to cover the whole class $\mathscr{G}$ with $\|\cdot\|_\infty$-balls of radius ϵ. Now let $\mathscr{G}'\subset\mathscr{G}$ be an arbitrary subset and assume that the elements $\{h_1,\dots,h_N\}\subset\mathscr{G}$ constitute a

covering of $\mathscr{G}$ with such balls. We want to find a set $\{h'_1,\dots,h'_N\}\subset\mathscr{G}'$ that constitutes a covering of $\mathscr{G}'$ with balls of radius 2ϵ. Since $\{h_1,\dots,h_N\}\subset\mathscr{G}$ constitutes an ϵ-ball covering of $\mathscr{G}$, we can select a minimal subset $\{h_{i_1},\dots,h_{i_{N'}}\}\subset\{h_1,\dots,h_N\}$ with $1\le i_1<\ldots<i_{N'}\le N$ that constitutes an ϵ-ball covering of $\mathscr{G}'$. If all $h_{i_j}\in\mathscr{G}'$, everything is proven. In this case, the set $\{h_{i_1},\dots,h_{i_{N'}}\}\subset\mathscr{G}'$ constitutes also a 2ϵ-ball covering of $\mathscr{G}'$. Otherwise, assume that $h_{i_j}\in\mathscr{G}\setminus\mathscr{G}'$. Since the subset $\{h_{i_1},\dots,h_{i_{N'}}\}$ is without loss of generality assumed to be minimal, the set $B(h_{i_j},\varepsilon)\cap\mathscr{G}'$ is non-empty. Now pick an arbitrary $h'_{i_j}\in B(h_{i_j},\epsilon)\cap\mathscr{G}'$, and observe that $B(h_{i_j},\varepsilon)\cap\mathscr{G}'\subset B(h'_{i_j},2\epsilon)$. We can carry out this procedure for every $h_{i_j}\notin\mathscr{G}'$ and replace this element with the obtained h'_{i_j}. All h_{i_j} that are elements of $\mathscr{G}'$ in the first place are simply relabeled h'_{i_j}. Hence, there exists a set $\{h'_{i_1},\dots,h'_{i_{N'}}\}\subset\mathscr{G}'$ that constitutes a cover of $\mathscr{G}'$ with balls of radius 2ϵ. This means that

$$N(2\epsilon,\mathscr{G}',\|\cdot\|_\infty)\le N'\le N=N(\epsilon,\mathscr{G},\|\cdot\|_\infty)\le e^{2MB/\epsilon}=e^{4MB/(2\epsilon)}$$

for any $k\in\mathbb{N}$. Since this consideration was independent of the specification of $\mathscr{G}'$, we conclude that for all k

$$N(\epsilon,\mathscr{G}_k,\|\cdot\|_\infty)\le e^{4MB/\epsilon}.$$

Plugging in $\epsilon=2^{1+k-s}\sqrt{\gamma}\delta$ gives

$$N\big(2^{-s}2^{k+1}\sqrt{\gamma}\delta,\mathscr{G}_k,\|\cdot\|_\infty\big)\le e^{4MB/(2^{1+k-s}\sqrt{\gamma}\delta)}=e^{2MB2^{s-k}/(\sqrt{\gamma}\delta)}.$$

This means that there exists a set $\mathscr{G}_k^{(s)}\subset\mathscr{G}_k$ with at most $e^{2MB2^{s-k}/(\sqrt{\gamma}\delta)}$ elements such that for any $g\in\mathscr{G}$ there exists an element $h\in\mathscr{G}_k^{(s)}$ with $\|g-h\|_\infty<2^{-s}2^{k+1}\sqrt{\gamma}\delta$.

Let $g\in\mathscr{G}_k$ be arbitrary. Since $\mathscr{G}_k^{(s)}$ is finite, the set

$$\begin{aligned}\Pi_{s,k}(g)&:=\operatorname*{arg\,min}_{h\in\mathscr{G}_k^{(s)}}\|g-h\|_\infty\\&=\big\{h'\in\mathscr{G}_k^{(s)}\colon\|g-h'\|_\infty\le\|g-h\|\ \text{ for all } h\in\mathscr{G}_k^{(s)}\big\}\end{aligned}$$

is not empty. Choose a representative from the finite set $\Pi_{s,k}(g)$ and call it $g_{s,k}$. □

An immediate consequence of the last proposition is that any $f_t(g;\,\cdot)$ can be displayed as a (point-wise) telescope sum:

$$f_t(g;\,\cdot)=f_t(g;\,\cdot)-f_t(g_{\check{S},k};\,\cdot)+f_t(g_{0,k};\,\cdot)+\sum_{s=0}^{\check{S}-1}\big(f_t(g_{s+1,k};\,\cdot)-f_t(g_{s,k};\,\cdot)\big).$$

Recalling the definition of the variables $X^*_{2j}(g)$, we can translate the previous

telescope sum to

$$\begin{aligned}
X^*_{2j}(g) &= \frac{1}{q}\sum_{i=0}^{q-1}\big(f_t(g;V^*_{t+2jq+i}) - E\,f_t(g;V^*_{t+2jq+i})\big)\\
&= \frac{1}{q}\sum_{i=0}^{q-1}\Bigg(\Big[f_t(g;V^*_{t+2jq+i}) - f_t(g_{\check{S},k};V^*_{t+2jq+i}) + f_t(g_{0,k};V^*_{t+2jq+i})\\
&\qquad\qquad + \sum_{s=0}^{\check{S}-1}\big(f_t(g_{s+1,k};V^*_{t+2jq+i}) - f_t(g_{s,k};V^*_{t+2jq+i})\big)\Big]\\
&\quad -E\Big[f_t(g;V^*_{t+2jq+i}) - f_t(g_{\check{S},k};(V^*_{t+2jq+i}) + f_t(g_{0,k};(V^*_{t+2jq+i})\\
&\qquad\qquad + \sum_{s=0}^{\check{S}-1}\big(f_t(g_{s+1,k};V^*_{t+2jq+i}) - f_t(g_{s,k};V^*_{t+2jq+i})\big)\Big]\Bigg)\\
&= X^*_{2j}(g) - X^*_{2j}(g_{\check{S},k}) + X^*_{2j}(g_{0,k}) + \sum_{s=0}^{\check{S}-1}\big(X^*_{2j}(g_{s+1,k}) - X^*_{2j}(g_{s,k})\big) \qquad (3.32)
\end{aligned}$$

The following lemma contains the actual chaining argument which essentially consists of an application of equation (3.32). The addends in the telescope sum are considered links in a chain of random variables approximating $X^*_{2j}(g)$, hence the name of the chaining argument. It is the last big step in the proof of Theorem 3.2.2.

LEMMA 3.2.17. *Suppose that the quantities δ and t are defined as in Lemma 3.2.11. Moreover, assume that the quantity q fulfills $q(n) \asymp t(n)$. In accordance with Proposition 3.2.16, for $g \in \mathscr{G}$ and*

$$\check{S}(n) := \min\Big\{s \in \mathbb{N}\colon \frac{4}{1-L_1}2^{-s}\sqrt{\gamma}\delta \le 2^{-6}\gamma\delta^2/(15M)\Big\},$$

the functions $g_{0,k},\dots,g_{\check{S},k}$ shall be given such that

$$\|g - g_{s,k}\|_\infty \le 2^{-s}2^{k+1}\sqrt{\gamma}\delta\,.$$

Recall that the symbol $\mathscr{G}_k$ refers to the set

$$\mathscr{G}_k = \Big\{g \in \mathscr{G}\colon E\big[m^{[t]}(Z^*_t) - g^{[t]}(Z^*_t)\big]^2 \le 2^{2k+2}\gamma\delta^2\Big\}.$$

Then there exists a positive constant C and a natural number n_0 such that for all $n \ge n_0$ and all $k \in \mathbb{N}$

$$\begin{aligned}
&P\Big\{\sup_{g\in\mathscr{G}_k}\frac{q}{n-t}\sum_{j=0}^{N-1}\varepsilon_j X^*_{2j}(g) > 2^{2k-6}\gamma\delta^2\Big\}\\
&\qquad \lesssim 2^{-2k}\frac{\log n}{n} + n^{-(k+1)} + \exp\Big(-C\,n^{1/3}\,2^k\Big) + 2^{-k}(\log n)^{-1}.
\end{aligned}$$

Proof. We observe that for every $n \in \mathbb{N}$ the maximal index $\check{S} = \check{S}(n)$ is given by

$$\check{S} = \min\left\{ s \in \mathbb{N} \colon \frac{2}{1-L_1} 2^{-s} 2^{k+1} \sqrt{\gamma}\delta \le 2^{k-6}\gamma\delta^2/(15M) \right\}.$$

Using the telescope representation of $X^*_{2j}(g)$ and the triangle inequality for probabilities, we obtain

$$\begin{aligned}
&P\left\{ \sup_{g\in\mathscr{G}_k} \frac{q}{n-t} \sum_{j=0}^{N-1} \varepsilon_j X^*_{2j}(g) > 2^{2k-6}\gamma\delta^2 \right\} \\
&= P\left\{ \sup_{g\in\mathscr{G}_k} \frac{q}{n-t} \sum_{j=0}^{N-1} \varepsilon_j \Big[X^*_{2j}(g) - X^*_{2j}(g_{\check{S},k}) + X^*_{2j}(g_{0,k}) \right. \\
&\qquad \left. + \sum_{s=0}^{\check{S}-1} \big(X^*_{2j}(g_{s+1,k}) - X^*_{2j}(g_{s,k}) \big) \Big] > 2^{2k-6}\gamma\delta^2 \right\} \\
&\le P\left\{ \sup_{g\in\mathscr{G}_k} \frac{q}{n-t} \sum_{j=0}^{N-1} \varepsilon_j \big[X^*_{2j}(g) - X^*_{2j}(g_{\check{S},k}) \big] > 2^{2k-6}\gamma\delta^2/3 \right\} \\
&\qquad + P\left\{ \sup_{g\in\mathscr{G}_k} \frac{q}{n-t} \sum_{j=0}^{N-1} \varepsilon_j X^*_{2j}(g_{0,k}) > 2^{2k-6}\gamma\delta^2/3 \right\} \\
&\qquad + P\left\{ \sup_{g\in\mathscr{G}_k} \sum_{s=0}^{\check{S}-1} \frac{q}{n-t} \sum_{j=0}^{N-1} \varepsilon_j \big(X^*_{2j}(g_{s+1,k}) - X^*_{2j}(g_{s,k}) \big) > 2^{2k-6}\gamma\delta^2/3 \right\} \\
&=: P_1 + P_2 + P_3 .
\end{aligned}$$

We treat each of the three terms respectively. The first term vanishes due to the definition of $\check{S}$. The index was chosen such that the approximation of g by $g_{\check{S},k}$ is very accurate. Using the definition of $X^*_{2j}(g)$, Proposition 3.2.14, and the definition of the sequence $\{g_{s,k}\}_{s=0,\dots,\check{S}}$, we observe

$$\begin{aligned}
\big| X^*_{2j}(g) - X^*_{2j}(g_{\check{S},k}) \big| &= \frac{1}{q} \Big| \sum_{i=0}^{q-1} \Big(f_t(g; V^*_{t+2jq+i}) - f_t(g_{\check{S},k}; V^*_{t+2jq+i}) \\
&\qquad - E\left[f_t(g; V^*_{t+2jq+i}) - f_t(g_{\check{S},k}; V^*_{t+2jq+i}) \right] \Big) \Big| \\
&\le \frac{1}{q} \sum_{i=0}^{q-1} \Big(\underbrace{\big| f_t(g; V^*_{t+2jq+i}) - f_t(g_{\check{S},k}; V^*_{t+2jq+i}) \big|}_{\le 2\|g-g_{\check{S},k}\|_\infty (Y^*_{t+2jq+i+1}+M)/(1-L_1)} \\
&\qquad + \underbrace{E\big| f_t(g; V^*_{t+2jq+i}) - f_t(g_{\check{S},k}; V^*_{t+2jq+i}) \big|}_{\le 2\|g-g_{\check{S},k}\|_\infty (EY^*_{t+2jq+i+1}+M)/(1-L_1)} \Big) \\
&\le \frac{1}{q} \sum_{i=0}^{q-1} \Big(\frac{2(Y^*_{t+2jq+i+1}+M)}{1-L_1} 2^{-\check{S}} 2^{k+1} \sqrt{\gamma}\delta \\
&\qquad + \frac{4M)}{1-L_1} 2^{-\check{S}} 2^{k+1} \sqrt{\gamma}\delta \Big) \\
&= \frac{1}{q} \sum_{i=0}^{q-1} (Y^*_{t+2jq+i+1} + 3M) \frac{2}{1-L_1} 2^{-\check{S}} 2^{k+1} \sqrt{\gamma}\delta
\end{aligned}$$

$$\leq \frac{1}{q}\sum_{i=0}^{q-1}(Y^*_{t+2jq+i+1}+3M)2^{k-6}\gamma\delta^2/(15M).$$

In the last estimate, we used the fact that $\frac{2}{1-L_1}2^{-\check{S}}2^{k+1}\sqrt{\gamma}\delta \leq 2^{k-6}\gamma\delta^2/(15M)$, which follows from the definition of $\check{S}$. We conclude,

$$\begin{aligned}
&P\Big\{\sup_{g\in\mathscr{G}_k}\frac{q}{n-t}\Big|\sum_{j=0}^{N-1}\varepsilon_j[X^*_{2j}(g)-X^*_{2j}(g_{\check{S},k})]\Big|>2^{2k-6}\gamma\delta^2/3\Big\}\\
\leq& P\Big\{\frac{1}{n-t}\Big|\sum_{i=t}^{n-1}(Y^*_{i+1}+3M)2^{k-6}\gamma\delta^2/(15M)\Big|>2^{2k-6}\gamma\delta^2/3\Big\}\\
\leq& P\Big\{\frac{1}{n-t}\Big|\sum_{i=t}^{n-1}Y^*_{i+1}\Big|>2M2^k\Big\}\\
\leq& P\Big\{\frac{1}{n-t}\Big|\sum_{i=t}^{n-1}(Y^*_{i+1}-EY^*_{i+1})\Big|>M2^k\Big\}\\
\leq& \frac{2^{-2k}}{M^2(n-t)^2}\mathrm{var}\big(\sum_{i=t}^{n-1}Y^*_i\big),
\end{aligned}$$

for any k. We recall that the process $\{Y^*_i\}$ is q-dependent and stationary and conclude that

$$\begin{aligned}
\mathrm{var}\big(\sum_{i=0}^{n-1}Y^*_i\big)&=\sum_{i,j=0}^{n-1}\mathrm{cov}\big(Y^*_i,Y^*_j\big)\\
&=\sum_{\substack{0\leq i,j\leq n-1\\ |i-j|\leq q}}\big(E(Y^*_iY^*_j)-EY^*_iEY^*_j\big)\\
&\leq 2\sum_{r=0}^{q-1}\sum_{i=1}^{n-r}\Big(\sqrt{EY^{*\,2}_i}\sqrt{EY^{*\,2}_{i+r}}-EY^*_iEY^*_{i+r}\Big)\\
&\leq 2nq\Big(E\big[Y^{*\,2}_0\big]-\big[EY^*_0\big]^2\Big)\\
&\leq 2Mnq.
\end{aligned}$$

This proves that there exists a constant $C_0>0$ and a number $n^{(P_1)}$ such that for all $n\in\mathbb{N}$ with $n>n^{(P_1)}$ and all $k\in\mathbb{N}$

$$P_1=P\Big\{\sup_{g\in\mathscr{G}_k}\frac{q}{n-t}\Big|\sum_{j=0}^{N-1}\varepsilon_j[X^*_{2j}(g)-X^*_{2j}(g_{\check{S},k})]\Big|>2^{2k-6}\gamma\delta^2/3\Big\}\leq C_02^{-2k}\,\frac{q}{n-t}.$$

We proceed by addressing the second term, P_2. First of all, since for any $g\in\mathscr{G}_k$ the first approximation $g_{0,k}=\pi_{0,k}g$ is selected from the finite set $\mathscr{G}^{(0)}_k$, we may reduce the supremum of all possible values of $\pi_{0,k}g$ for g ranging in $\mathscr{G}_k$ to a maximum of all elements $h_{k,0}\in\mathscr{G}^{(0)}_k$:

$$P\Big\{\sup_{g\in\mathscr{G}_k}\frac{q}{n-t}\sum_{j=0}^{N-1}\varepsilon_jX^*_{2j}(g_{0,k})>2^{2k-6}\gamma\delta^2/3\Big\}$$

$$= P\Big\{ \max_{h_{0,k}\in\mathscr{G}_k^{(0)}} \frac{q}{n-t} \sum_{j=0}^{N-1} \varepsilon_j X_{2j}^*(h_{0,k}) > 2^{2k-6}\gamma\delta^2/3 \Big\}.$$

This exceedance probability will be bounded with the help of Bernstein's inequality (Lemma A.4.3). A necessary condition for the application of the inequality in the form of Lemma A.4.3 is that all random variables in the sum are bounded. To that end, we introduce the Σ-measurable events

$$A_{k,2j} = A_{k,2j}(n) := \{\omega \in S : \max_{i=0,\dots,q-1} Y_{t+2jq+i+1}^* \le 2(k+2)\log n\}$$

By Proposition 3.2.13, we conclude that $P\Big(\bigcup_{j=0}^{N-1} A_{k,2j}^c\Big) \le e^M n^{-(k+1)}$. This implies

$$\begin{aligned}
&P\Big\{ \max_{h_{0,k}\in\mathscr{G}_k^{(0)}} \frac{q}{n-t} \sum_{j=0}^{N-1} \varepsilon_j X_{2j}^*(h_{0,k}) > 2^{2k-6}\gamma\delta^2/3 \Big\} \\
&\le e^M n^{-(k+1)} + P\Big\{ \max_{h_{0,k}\in\mathscr{G}_k^{(0)}} \frac{q}{n-t} \sum_{j=0}^{N-1} \varepsilon_j X_{2j}^*(h_{0,k}) \mathbb{1}_{A_{k,2j}} > 2^{2k-6}\gamma\delta^2/3 \Big\} \\
&= e^M n^{-(k+1)} + P \bigcup_{h_{0,k}\in\mathscr{G}_k^{(0)}} \Big\{ \frac{q}{n-t} \sum_{j=0}^{N-1} \varepsilon_j X_{2j}^*(h_{0,k}) \mathbb{1}_{A_{k,2j}} > 2^{2k-6}\gamma\delta^2/3 \Big\} \\
&\le e^M n^{-(k+1)} + \sum_{h_{0,k}\in\mathscr{G}_k^{(0)}} P\Big\{ \frac{q}{n-t} \sum_{j=0}^{N-1} \varepsilon_j X_{2j}^*(h_{0,k}) \mathbb{1}_{A_{k,2j}} > 2^{2k-6}\gamma\delta^2/3 \Big\}.
\end{aligned}$$

Now all involved variables are bounded. In order to apply Bernstein's inequality, we need bounds on the variance of the sum $\sum_{j=0}^{N-1} \varepsilon_j X_{2j}^*(h_{0,k}) \mathbb{1}_{A_{k,2j}}$, and a bound on the absolute values of the addends $\varepsilon_j X_{2j}^*(h_{0,k}) \mathbb{1}_{A_{k,2j}}$. Furthermore, the addends have to be centered. For the variance bound recall that the sequences $\{\varepsilon_j\}$ and $\{Y_i^*\}$ are independent. Hence, $\{\varepsilon_j X_{2j}^*(h_{0,k}) \mathbb{1}_{A_{k,2j}}\}_{j\in\mathbb{N}}$ is a sequence of i.i.d. random variables. Moreover, we observe $E\big[\varepsilon_j X_{2j}^*(h_{0,k}) \mathbb{1}_{A_{k,2j}}\big] = E\varepsilon_j E\big[X_{2j}^*(h_{0,k}) \mathbb{1}_{A_{k,2j}}\big] = 0$. Thus, for any $h_{0,k} \in \mathscr{G}_k^{(0)} \subset \mathscr{G}_k$, we can invoke the first statement of Lemma 3.2.11 to conclude that there exists a number n^* such that for all $n > n^*$ and for all $k \in \mathbb{N}$

$$\begin{aligned}
\operatorname{var}\Big(\sum_{j=0}^{N-1} \varepsilon_j X_{2j}^*(h_{0,k}) \mathbb{1}_{A_{k,2j}} \Big) &= E\Big(\sum_{j=0}^{N-1} \varepsilon_j X_{2j}^*(h_{0,k}) \mathbb{1}_{A_{k,2j}} \Big)^2 \\
&= \sum_{j=0}^{N-1} E\big(X_{2j}^*(h_{0,k}) \mathbb{1}_{A_{k,2j}} \big)^2 \\
&\le \sum_{j=0}^{N-1} E\big(X_{2j}^*(h_{0,k}) \big)^2 \\
&= \frac{(n-t)^2}{q^2} E\Big(\frac{q}{n-t} \sum_{j=0}^{N-1} X_{2j}^*(h_{0,k}) \Big)^2 \\
&= \frac{(n-t)^2}{q^2} \operatorname{var}\Big(\frac{q}{n-t} \sum_{j=0}^{N-1} X_{2j}^*(h_{0,k}) \Big)
\end{aligned}$$

$$
\begin{aligned}
&\leq \frac{(n-t)^2}{q^2}(M^2+2M^{3/2}+M)2^{2k+4}\gamma\delta^2\frac{q}{n-t}\\
&= C_1\frac{(n-t)}{q}2^{2k}\delta^2\\
&:= \sigma_n^2 .
\end{aligned}
$$

In the second to last line, we introduced the positive constant $C_1 = 16(M^2 + 2M^{3/2} + M)\gamma$. Let us now bound the absolute values of the addends. First, we remark that $f_t(m;V^*_{t+2jq+i}) = 0$, and we infer with the use of Proposition 3.2.14 that

$$
\begin{aligned}
\left|f_t(h_{0,k};V^*_{t+2jq+i})\right| &= \left|f_t(h_{0,k};V^*_{t+2jq+i}) - f_t(m;V^*_{t+2jq+i})\right|\\
&\leq \frac{2\|h_{0,k}-m\|_\infty}{1-L_1}(M+Y^*_{t+2jq+i+1})\\
&\leq \frac{2M}{1-L_1}(M+Y^*_{t+2jq+i+1}).
\end{aligned}
$$

In virtue of the definition of the events $A_{k,2j}$, we obtain

$$
\begin{aligned}
\left|\varepsilon_j X^*_{2j}(h_{0,k})\mathbb{1}_{A_{k,2j}}\right| &= \left|X^*_{2j}(h_{0,k})\mathbb{1}_{A_{k,2j}}\right|\\
&= \mathbb{1}_{A_{k,2j}}\frac{1}{q}\Big|\sum_{i=0}^{q-1}\left[f_t(h_{0,k};V^*_{t+2jq+i}) - Ef_t(h_{0,k};V^*_{t+2jq+i})\right]\Big|\\
&\leq \mathbb{1}_{A_{k,2j}}\frac{1}{q}\sum_{i=0}^{q-1}\Big(\left|f_t(h_{0,k};V^*_{t+2jq+i})\right| + E\left|f_t(h_{0,k};V^*_{t+2jq+i})\right|\Big)\\
&\leq \mathbb{1}_{A_{k,2j}}\frac{1}{q}\sum_{i=0}^{q-1}\frac{2M}{1-L_1}\left[(Y^*_{t+2jq+i+1}+M)+E(Y^*_{t+2jq+i+1}+M)\right]\\
&\leq \frac{1}{q}\sum_{i=0}^{q-1}\frac{2M}{1-L_1}(Y^*_{t+2jq+i+1}+3M)\mathbb{1}_{A_{k,2j}}\\
&\leq \frac{2M}{1-L_1}(2(k+2)\log(n)+3M)\\
&\leq C_2(k+1)\log n\\
&=: b_n
\end{aligned}
$$

with $C_2 = \frac{10M}{1-L_1}$, for all $n \in \mathbb{N}$ with $n \geq e^{3M}$ and all $k \in \mathbb{N}$. We are ready to apply Bernstein's inequality. To resemble the notation in Lemma A.4.3, we introduce the variables

$$
\eta_j := \varepsilon_j X^*_{2j}(h_{0,k})\mathbb{1}_{A_{k,2j}} \qquad\qquad x_n := \tfrac{n-t}{q}2^{2k-6}\gamma\delta^2/3
$$

and obtain the display

$$
P\Big\{\sum_{j=0}^{N-1}\varepsilon_j X^*_{2j}(h_{0,k})\mathbb{1}_{A_{k,2j}} > \frac{n-t}{q}2^{2k-6}\gamma\delta^2/3\Big\} = P\Big\{\sum_{j=0}^{N-1}\eta_j > x_n\Big\}.
$$

We have shown that the random variables η_j are independent and centered. They take values in the interval $[-b_n, b_n]$, and $\operatorname{var}(\eta_0+\ldots+\eta_{N-1}) \le \sigma_n^2$. We can therefore apply Bernstein's inequality in the form of Lemma A.4.3, and conclude that

$$P\Big\{\sum_{j=0}^{N-1}\eta_j > x_n\Big\} \le \exp\Big(-\frac{1}{2}\frac{x^2}{\sigma_n^2 + x_n b_n/3}\Big). \tag{3.33}$$

In this case, $\sigma_n^2 \asymp 2^{2k}\frac{n-t}{q}\delta^2$ is dominated by $x_n\, b_n \asymp 2^{2k}\frac{n-t}{q}\delta^2\,(k+1)\log n$ since

$$\limsup_{n\to\infty}\sup_{k\in\mathbb{N}}\frac{\sigma_n^2}{x_n b_n} \le \limsup_{n\to\infty}\sup_{k\in\mathbb{N}}\frac{C}{k+1}\frac{1}{\log n} \le C\limsup_{n\to\infty}(\log n)^{-1} = 0$$

with a positive constant C. Hence, there exists a number n^{**}, independent of k, such that $\sigma_n^2 \le \frac{2}{3}x_n b_n$ for all $n > n^{**}$ and all $k \in \mathbb{N}$. Consequently, under the assumptions $\delta_n = n^{-1/3}\log n$ and $t(n) \asymp q(n) \asymp \log n$, we obtain for the exponent in (3.33)

$$\begin{aligned}\liminf_{n\to\infty}\frac{1}{2}\frac{x_n^2}{\sigma_n^2 + x_n b_n/3}\cdot n^{-1/3} &\ge \liminf_{n\to\infty}\frac{x_n}{2\,b_n}\cdot n^{-1/3}\\ &= \frac{2^{-6}\gamma}{6C_2}\,\frac{2^{2k}}{k+1}\liminf_{n\to\infty}\frac{n-t}{q}\frac{\delta^2}{\log n}\cdot n^{-1/3}\\ &\ge C_3\,2^k\end{aligned}$$

for some positive constant C_3. We conclude that there exists a number $n_0 \ge \max\{e^{3M}, n^*, n^{**}\}$ such that $\frac{1}{2}\frac{x^2}{\sigma_n^2 + x_n b_n/3} \ge C_3\,2^k\,n^{1/3}$ for all $n \ge n_0$ and all $k \in \mathbb{N}$. Thus, for all k and all $n \ge n_0$

$$P\Big\{\sum_{j=0}^{N-1}\varepsilon_j X_{2j}^*(h_{0,k})\mathbb{1}_{A_{k,2j}} > \frac{n-t}{q}2^{2k-6}\gamma\delta^2/3\Big\} \le \exp\Big(-C_3\,2^k n^{1/3}\Big).$$

Note, that n_0 does not depend on k. The previous bound is independent of the specific function $h_{0,k} \in \mathscr{G}_k^{(0)}$. Thus, for any $n \ge n_0$

$$\begin{aligned}&\sum_{h_{0,k}\in\mathscr{G}_k^{(0)}} P\Big\{\frac{q}{n-t}\sum_{j=0}^{N-1}\varepsilon_j X_{2j}^*(h_{0,k})\mathbb{1}_{A_{k,2j}} > 2^{2k-6}\gamma\delta^2/3\Big\}\\ &\le \#\mathscr{G}_k^{(0)}\ \exp\big(-C_3\,2^k n^{1/3}\big)\\ &= \exp\Big(\log\big(\#\mathscr{G}_k^{(0)}\big) - C_3\,2^k\,n^{1/3}\Big)\\ &\le \exp\Big(\frac{C_4}{\delta} - C_3\,2^k\,n^{1/3}\Big)\\ &\le \exp\Big(2^k\big(\frac{C_4}{\delta} - C_3 n^{1/3}\big)\Big).\end{aligned}$$

In the last estimate we used the fact that for $C_4 := 2MB/\sqrt{\gamma}$ we have

$$\sup_k \log\big(\#\mathscr{G}_k^{(0)}\big) \le \sup_k C_4\, 2^{-k}/\delta \le C_4/\delta\,, \tag{3.34}$$

which follows from the bounds in Proposition 3.2.16 with $s = 0$. Subsequently, we observe that $\lim_{n\to\infty} n^{-1/3}\delta^{-1} = 0$ and conclude that there exists a number $n_1 \in \mathbb{N}$ such that $C_4\delta^{-1} - C_3 n^{1/3} \le -\frac{C_3}{2}\, n^{1/3}$ for all $n \ge n_1$. Hence, for all $n \ge n_0 \vee n_1$ and all $k \in \mathbb{N}$

$$\sum_{h_{0,k}\in\mathscr{G}_k^{(0)}} P\bigg\{\frac{q}{n-t}\sum_{j=0}^{N-1}\varepsilon_j X_{2j}^*(h_{0,k})\mathbb{1}_{A_{k,2j}} > 2^{2k-6}\gamma\delta^2/3\bigg\} \le \exp\big(-\frac{C_3}{2}\,2^k n^{1/3}\big).$$

In conclusion, there exists a natural number $n^{(P_2)} \ge n_0 \vee n_1$ such that for all $n \ge n^{(P_2)}$ and all $k \in \mathbb{N}$ the term P_2 is bounded by

$$P\bigg\{\sup_{g\in\mathscr{G}_k}\frac{q}{n-t}\sum_{j=0}^{N-1}\varepsilon_j X_{2j}^*(g_{0,k}) > 2^{2k-6}\gamma\delta^2/3\bigg\} < e^M\, n^{-(k+1)} + \exp\Big(-C\,2^k\, n^{1/3}\Big),$$

with some positive constant C.

It is left to find a bound for the third term, P_3. For that sake we define the sets $\mathscr{M}_{s,k}$ by

$$\mathscr{M}_{s,k} := \Big\{(g_1,g_2)\colon g_1 \in \mathscr{G}_k^{(s)},\, g_2 \in \mathscr{G}_k^{(s+1)},\, \|g_1 - g_2\|_\infty \le 2^{-s}2^{k+2}\sqrt{\gamma}\delta\Big\}.$$

Let $g \in \mathscr{G}_k$. Then for the images of g under the mappings $\pi_{s,k}\colon \mathscr{G}_k \to \mathscr{G}_k^{(s)}$ and $\pi_{s+1,k}\colon \mathscr{G}_k \to \mathscr{G}_k^{(s+1)}$ it holds

$$\begin{aligned}\|g - g_{s,k}\|_\infty &\le 2^{-s}2^{k+1}\sqrt{\gamma}\delta\,,\\ \|g - g_{s+1,k}\|_\infty &\le 2^{-(s+1)}2^{k+1}\sqrt{\gamma}\delta\,,\end{aligned}$$

respectively. The triangle inequality implies

$$\|g_{s,k} - g_{s+1,k}\|_\infty \le \|g_{s,k} - g\|_\infty + \|g - g_{s+1,k}\|_\infty \le 2^{-s}2^{k+2}\sqrt{\gamma}\delta$$

and therefore $(g_{s,k}, g_{s+1,k}) \in \mathscr{M}_{s,k}$. We conclude,

$$\begin{aligned}&\sup_{g\in\mathscr{G}_k}\sum_{s=0}^{\check{S}-1}\frac{q}{n-t}\Big|\sum_{j=0}^{N-1}\varepsilon_j\big(X_{2j}^*(g_{s+1,k}) - X_{2j}^*(g_{s,k})\big)\Big|\\ &= \max_{(g_1,g_2)\in\mathscr{M}_{k,s}}\sum_{s=0}^{\check{S}-1}\frac{q}{n-t}\Big|\sum_{j=0}^{N-1}\varepsilon_j\big(X_{2j}^*(g_2) - X_{2j}^*(g_1)\big)\Big|\\ &\le \sum_{s=0}^{\check{S}-1}\bigg[\max_{(g_1,g_2)\in\mathscr{M}_{k,s}}\frac{q}{n-t}\Big|\sum_{j=0}^{N-1}\varepsilon_j\big(X_{2j}^*(g_2) - X_{2j}^*(g_1)\big)\Big|\bigg].\end{aligned} \tag{3.35}$$

We observe, again by Proposition 3.2.14,

$$\begin{aligned}\left|\left(X^*_{2j}(g_2)-X^*_{2j}(g_1)\right)\right| &\le \frac{1}{q}\sum_{i=0}^{q-1}\left|f_t(g_1;V^*_{t+2jq+i})-f_t(g_2;V^*_{t+2jq+i})\right| \\ &\quad+\frac{1}{q}\sum_{i=0}^{q-1}E\left|f_t(g_1;V^*_{t+2jq+i})-f_t(g_2;V^*_{t+2jq+i})\right| \\ &\le \frac{2\|g_1-g_2\|_\infty}{1-L_1}\frac{1}{q}\sum_{i=0}^{q-1}\left((Y^*_{t+2jq+i+1}+M)+E(Y^*_{t+2jq+i+1}+M)\right) \\ &\le \frac{2\|g_1-g_2\|_\infty}{1-L_1}\frac{1}{q}\sum_{i=0}^{q-1}(Y^*_{t+2jq+i+1}+3M).\end{aligned}$$

By the fact that the variables ε_j are independent and satisfy $E\varepsilon_j=0$, we can apply Hoeffding's inequality (Corollary A.4.2) conditionally on $\left(V^*_t,\dots,V^*_{n-1}\right)'$ to obtain for $(g_1,g_2)\in\mathscr{M}_{k,s}$ the bound

$$\begin{aligned}&P\left\{\frac{q}{n-t}\left|\sum_{j=0}^{N-1}\varepsilon_j\left(X^*_{2j}(g_2)-X^*_{2j}(g_1)\right)\right|>x\,\middle|\,V^*_t,\dots,V^*_{n-1}\right\} \\ &\le 2\exp\left(-\frac{x^2\,\frac{(n-t)^2}{q^2}}{2\sum_{j=0}^{N-1}\left(X^*_{2j}(g_2)-X^*_{2j}(g_1)\right)^2}\right) \\ &\le 2\exp\left(-\frac{x^2\,\frac{(n-t)^2}{q^2}}{2\sum_{j=0}^{N-1}\left[\frac{2\|g_1-g_2\|_\infty}{(1-L_1)}\frac{1}{q}\sum_{i=0}^{q-1}\left(Y^*_{t+2jq+i+1}+3M\right)\right]^2}\right) \\ &= 2\exp\left(-\frac{x^2\,\frac{(n-t)^2}{q^2}}{2\sum_{j=0}^{N-1}\left[\frac{2\cdot2^{-s}2^{k+2}\sqrt{\gamma}\delta}{(1-L_1)}\frac{1}{q}\sum_{i=0}^{q-1}\left(Y^*_{t+2jq+i+1}+3M\right)\right]^2}\right) \\ &= 2\exp\left(-\frac{x^2}{2\sum_{j=0}^{N-1}\left[\frac{2^{-s}2^{k+3}\sqrt{\gamma}\delta}{(n-t)(1-L_1)}\sum_{i=0}^{q-1}\left(Y^*_{t+2jq+i+1}+3M\right)\right]^2}\right).\end{aligned}$$

We apply the integrated Bernstein inequality (Lemma A.4.4) with $a=0$ and

$$b=\sum_{j=0}^{N-1}\left[\frac{2^{-s}2^{k+3}\sqrt{\gamma}\delta}{(n-t)(1-L_1)}\sum_{i=0}^{q-1}\left(Y^*_{t+2jq+i+1}+3M\right)\right]^2.$$

Conditionally on $V^*_t,\dots,V^*_{n-1}$, this yields almost surely in P that

$$\begin{aligned}&E\left[\max_{(g_1,g_2)\in\mathscr{M}_{k,s}}\frac{q}{n-t}\left|\sum_{j=0}^{N-1}\varepsilon_j\left(X^*_{2j}(g_2)-X^*_{2j}(g_1)\right)\right|\,\middle|\,V^*_t,\dots,V^*_{n-1}\right] \\ &\le C\sqrt{\log\#\mathscr{M}_{k,s}}\sqrt{\sum_{j=0}^{N-1}\left[\frac{2^{-s}2^{k+3}\sqrt{\gamma}\delta}{(n-t)(1-L_1)}\sum_{i=0}^{q-1}\left(Y^*_{t+2jq+i+1}+3M\right)\right]^2} \qquad (3.36)\end{aligned}$$

for some positive constant C. Furthermore, we use the known bound $E_{|\mathscr{F}^*_i}Y^{*2}_{i+1}=$

$\mathrm{var}_{|\mathscr{F}_i^*} Y_{i+1}^* + (E_{|\mathscr{F}_i^*} Y_{i+1}^*)^2 \le M + M^2$ and the triangle inequality for the $L_2(P)$ norm to conclude

$$\begin{aligned}\left(E\left[\sum_{i=0}^{q-1}\left(Y_{t+2jq+i+1}^* + 3M\right)\right]^2\right)^{1/2} &\le \sum_{i=0}^{q-1}\left(E\left[Y_{t+2jq+i+1}^* + 3M\right]^2\right)^{1/2} \\ &= q\, E\left(Y_0^* + 3M\right)^2 \\ &= q\left[EY_0^{*\,2} + 6M\, EY_0^* + 9M^2\right] \\ &\le q\left[(M+M^2) + 15M^2\right] \\ &\le 16(M^2+M)q \\ &=: \sqrt{C_3}\, q\,.\end{aligned}$$

This means,

$$E\left[\sum_{i=0}^{q-1}\left(Y_{t+2jq+i+1}^* + 3M\right)\right]^2 \le C_3\, q^2\,. \tag{3.37}$$

Together with the last bound as well as the established fact $N < \frac{n-t}{q}$, taking expectations in inequality (3.36) yields

$$\begin{aligned}&E\left[E\left[\max_{(g_1,g_2)\in\mathscr{M}_{k,s}} \frac{q}{n-t}\left|\sum_{j=0}^{N-1}\varepsilon_j\left(X_{2j}^*(g_2) - X_{2j}^*(g_1)\right)\right| \middle| V_t^*,\ldots,V_{n-1}^*\right]\right] \\ &\le C\sqrt{\log \#\mathscr{M}_{k,s}}\, E\sqrt{\sum_{j=0}^{N-1}\left[\frac{2^{-s}2^{k+3}\sqrt{\gamma}\delta}{(n-t)(1-L_1)}\sum_{i=0}^{q-1}\left(Y_{t+2jq+i+1}^* + 3M\right)\right]^2} \\ &\le C\sqrt{\log \#\mathscr{M}_{k,s}}\sqrt{E\sum_{j=0}^{N-1}\left[\frac{2^{-s}2^{k+3}\sqrt{\gamma}\delta}{(n-t)(1-L_1)}\sum_{i=0}^{q-1}\left(Y_{t+2jq+i+1}^* + 3M\right)\right]^2} \\ &= C\sqrt{\log \#\mathscr{M}_{k,s}}\sqrt{\frac{64\gamma}{(1-L_1)^2}\sum_{j=0}^{N-1}\left[\frac{2^{-s+k}\delta}{(n-t)}\right]^2 E\left[\sum_{i=0}^{q-1}\left(Y_{t+2jq+i+1}^* + 3M\right)\right]^2} \\ &\le C\sqrt{\log \#\mathscr{M}_{k,s}}\sqrt{\frac{64\gamma}{(1-L_1)^2}\sum_{j=0}^{N-1}\left[\frac{2^{-s+k}\delta}{(n-t)}\right]^2 C_3 q^2} \qquad \textit{by (3.37)} \\ &= \underbrace{C\frac{\sqrt{64C_3\gamma}}{1-L_1}}_{=:C_5}\sqrt{\log \#\mathscr{M}_{k,s}}\,\sqrt{N}\frac{2^{k-s}\delta q}{(n-t)} \\ &\le C_5\sqrt{\log \#\mathscr{M}_{k,s}}\frac{2^{k-s}\delta\sqrt{q}}{\sqrt{n-t}} \qquad (3.38)\end{aligned}$$

Concerning the cardinality of the set $\mathscr{M}_{k,s}$, we observe that

$$\#\mathscr{M}_{k,s} \le \#\mathscr{G}_k^{(s)}\#\mathscr{G}_k^{(s+1)} \le e^{2MB(2^{s-k}+2^{s+1-k})/(\sqrt{\gamma}\delta)} \le e^{2MB\,2^{s+2-k}/(\sqrt{\gamma}\delta)},$$

or, with $C_6 := \sqrt{\frac{8MB}{\sqrt{\gamma}}}$,

$$\sqrt{\log \#\mathscr{M}_{k,s}} \le C_6\, 2^{s/2} 2^{-k/2} \delta^{-1/2}. \tag{3.39}$$

Applying the Markov inequality, we finally arrive at a bound for P_3:

$$\begin{aligned}
&P\Big\{ \sup_{g\in\mathscr{G}_k} \sum_{s=0}^{\check{S}-1} \frac{q}{n-t} \Big| \sum_{j=0}^{N-1} \varepsilon_j \big(X^*_{2j}(g_{s+1,k}) - X^*_{2j}(g_{s,k}) \big) \Big| > 2^{2k-6} \gamma \delta^2/3 \Big\} \\
&\le \frac{3\cdot 2^6}{\gamma} 2^{-2k} \delta^{-2}\, E\Big[\sup_{g\in\mathscr{G}_k} \sum_{s=0}^{\check{S}-1} \frac{q}{n-t} \Big| \sum_{j=0}^{N-1} \varepsilon_j \big(X^*_{2j}(g_{s+1,k}) - X^*_{2j}(g_{s,k}) \big) \Big| \Big] \\
&\le \frac{3\cdot 2^6}{\gamma} 2^{-2k} \delta^{-2} \sum_{s=0}^{\check{S}-1} E\Big[\max_{(g_1,g_2)\in\mathscr{M}_{k,s}} \frac{q}{n-t} \Big| \sum_{j=0}^{N-1} \varepsilon_j \big(X^*_{2j}(g_2) - X^*_{2j}(g_1) \big) \Big| \Big] \qquad \textit{by (3.35)} \\
&\le \frac{3\cdot 2^6 C_5}{\gamma}\, 2^{-2k} \delta^{-2} \sum_{s=0}^{\infty} \sqrt{\log \#\mathscr{M}_{k,s}}\, \frac{2^{k-s}\delta\sqrt{q}}{\sqrt{n-t}} \qquad \textit{by (3.38)} \\
&\le \underbrace{\frac{3\cdot 2^6 C_5}{\gamma} C_6}_{=:C_7}\, 2^{-2k} \delta^{-2} \sum_{s=0}^{\infty} 2^{s/2} 2^{-k/2} \delta^{-1/2} \frac{2^{k-s}\delta\sqrt{q}}{\sqrt{n-t}} \qquad \textit{by (3.39)} \\
&= C_7\, 2^{-3k/2} \underbrace{\delta^{-3/2} (n-t)^{-1/2} q^{1/2}}_{\asymp (\log n)^{-1}} \underbrace{\sum_{s=0}^{\infty} 2^{-s/2}}_{<\infty}.
\end{aligned}$$

There exists a constant $C > 0$ and a number $n_3 \in \mathbb{N}$ such that

$$C_7\, \delta^{-3/2} (n-t)^{-1/2} q^{1/2} \sum_{s=0}^{\infty} 2^{-s/2} \le C\, (\log n)^{-1}.$$

for all $n \ge n_3$. Of course, $2^{-3k/2} \le 2^{-k}$ for all $k \in \mathbb{N}$. Thus, for $n \ge n^{(P_3)} = n_3$ and all $k \in \mathbb{N}$ we obtain the following bound for P_3:

$$P\Big\{ \sup_{g\in\mathscr{G}_k} \sum_{s=0}^{\check{S}-1} \frac{q}{n-t} \Big| \sum_{j=0}^{N-1} \varepsilon_j \big(X^*_{2j}(g_{s+1,k}) - X^*_{2j}(g_{s,k}) \big) \Big| > 2^{2k-6} \gamma \delta^2/3 \Big\} \le C\, 2^{-k} (\log n)^{-1}.$$

To summarize, we have found that

$$P_1 + P_2 + P_3 \lesssim 2^{-2k} \frac{q}{n-t} + n^{-(k+1)} + \exp\Big(-C\, 2^k\, n^{1/3} \Big) + 2^{-k} (\log n)^{-1}$$

for all $n \ge \max\{ n^{(P_1)}, n^{(P_2)}, n^{(P_3)} \}$ and all $k \in \mathbb{N}$. This concludes the proof. □

Conclusion of the proof of Theorem 3.2.2. We can combine the established auxiliary results to obtain the final chain of inequalities. The first two inequalities are obtained by applying Corollary 3.2.10 and Lemma 3.2.12 respectively. The third step is a simple application of Markov's inequality on the exceedance probability of the remainder terms $R_n(g)$ and Δ_n. The final bound is obtained by Lemma

3.2.17, Corollary 3.2.15 to bound $E(R_n(g))$, and Lemma 2.2.9 to bound the mixing coefficients $\phi^t(q)$. The bound $\mathbb{E}\Delta_n \lesssim \frac{t}{n} + L_1^t$ is taken from Lemma 3.2.8. Thus, we obtain the following inequalities for all but finitely many n and all $k \in \mathbb{N}$:

$$
\begin{aligned}
&\mathbb{P}\Big\{\mathbb{E}_{|\hat{m}_n = g}\big[m(\lambda_0', Y_0') - g(\lambda_0', Y_0')\big]^2 > \delta^2\Big\} \\
&\leq \sum_{k=0}^{\infty} P\Bigg\{ \sup_{\substack{g \in \mathscr{G} \\ E[m^{[t]}(Z_t^*) - g^{[t]}(Z_t^*)]^2 \leq 2^{2k+2}\gamma\delta^2}} \frac{1}{n-t} \sum_{i=t}^{n-1} \big(f_t(g; V_i^*) - E f_t(g; V_i^*)\big) > 2^{2k-2}\gamma\delta^2 \Bigg\} \\
&\qquad + \frac{n-t}{q}\phi^t(q) + \mathbb{P}\Big\{\Delta_n > \frac{\gamma\delta^2}{4}\Big\} \\
&\leq \sum_{k=0}^{\infty} \Bigg[8P\Bigg\{ \sup_{\substack{g \in \mathscr{G} \\ E[m^{[t]}(Z_t^*) - g^{[t]}(Z_t^*)]^2 \leq 2^{2k+2}\gamma\delta^2}} \frac{q}{n-t} \sum_{j=0}^{N-1} \varepsilon_j X_{2j}^*(g) > 2^{2k-6}\gamma\delta^2 \Bigg\} \\
&\qquad + P\Bigg\{ \sup_{\substack{g \in \mathscr{G} \\ E[m^{[t]}(Z_t^*) - g^{[t]}(Z_t^*)]^2 \leq 2^{2k+2}\gamma\delta^2}} R_n(g) > 2^{2k-3}\gamma\delta^2 \Bigg\} \Bigg] \\
&\qquad + \frac{n-t}{q}\phi^t(q) + \mathbb{P}\Big\{\Delta_n > \frac{\gamma\delta^2}{4}\Big\} \\
&\leq \sum_{k=0}^{\infty} \Bigg[8P\Bigg\{ \sup_{\substack{g \in \mathscr{G} \\ E[m^{[t]}(Z_t^*) - g^{[t]}(Z_t^*)]^2 \leq 2^{2k+2}\gamma\delta^2}} \frac{q}{n-t} \sum_{j=0}^{N-1} \varepsilon_j X_{2j}^*(g) > 2^{2k-6}\gamma\delta^2 \Bigg\} \\
&\qquad + \frac{8}{\gamma} \cdot 2^{-2k} \delta^{-2} E\Big(\sup_{g \in \mathscr{G}} R_n(g)\Big) \Bigg] \\
&\qquad + \frac{n-t}{q}\phi^t(q) + \frac{4}{\gamma}\delta^{-2}\mathbb{E}\Delta_n \\
&\lesssim \sum_{k=0}^{\infty} \Big[2^{-2k}\frac{\log n}{n} + n^{-k}n^{-1} + \exp\Big(-C\, n^{1/3}\, 2^k\Big) + 2^{-k}(\log n)^{-1} \\
&\qquad + 2^{-2k}\delta^{-2}\frac{q}{n-t} \Big] \\
&\qquad + \frac{n-t}{q}(L_1 + L_2)^{q-t} + \delta^{-2}\big(\frac{t}{n} + L_1^t\big).
\end{aligned}
$$

We argue that this quantity is in $o(1)$ as $n \to \infty$. To that end, recall that $t(n) = -\frac{2}{3\log L_1}\log n$ and $q(n) \asymp t(n)$ by the assumptions in Lemma 3.2.8 and Lemma 3.2.17 respectively. These facts imply that $\delta^{-2}\frac{q}{n-t} \asymp n^{-1/3}(\log n)^{-1}$, from which it can be seen that for all $k \in \mathbb{N}$

$$
\lim_{n\to\infty} \Big[2^{-2k}\frac{\log n}{n} + n^{-k}n^{-1} + \exp\Big(-C\, n^{1/3}\, 2^k\Big) + 2^{-k}(\log n)^{-1} + 2^{-2k}\delta^{-2}\frac{q}{n-t} \Big] = 0\,.
$$

Since there exists an absolute summable sequence $\{\eta_k\}_{k\in\mathbb{N}} \subset \mathbb{R}$ such that for any k

$$
\sup_{n\geq 2} \Big| 2^{-2k}\frac{\log n}{n} + n^{-k}n^{-1} + \exp\Big(-C\, n^{1/3}\, 2^k\Big) + 2^{-k}(\log n)^{-1} + 2^{-2k}\delta^{-2}\frac{q}{n-t} \Big| \leq \eta_k\,,
$$

we conclude that

$$\sum_{k=0}^{\infty}\Big[2^{-2k}\frac{\log n}{n}+n^{-k}n^{-1}+\exp\Big(-C\,n^{1/3}\,2^{k}\Big)+2^{-k}(\log n)^{-1}+2^{-2k}\frac{q}{n-t}\Big]=o(1)$$

as $n\to\infty$. If we furthermore specify the quantity q as $q(n)=-\big(\frac{1}{\log(L_1+L_2)}+\frac{2}{3\log L_1}\big)\log n$, we obtain

$$\frac{n-t}{q}(L_1+L_2)^{q-t}+\delta^{-2}\big(\frac{t}{n}+L_1^t\big)\asymp(\log n)^{-1}+n^{-1/3}(\log n)^{-1}+(\log n)^{-2}$$

implying that

$$\frac{n-t}{q}(L_1+L_2)^{q-t}+\delta^{-2}\big(\frac{t}{n}+L_1^t\big)=o(1).$$

In total, we have shown that

$$\limsup_{n\to\infty}\ \mathbb{P}\Big\{\mathbb{E}_{|\hat{m}_n=g}\big[m(\lambda_0',Y_0')-g(\lambda_0',Y_0')\big]^2>\delta^2\Big\}=0.$$

The proof of Theorem 3.2.2 is now complete. □

3.2.3 Growing classes of candidate functions

At last we want to address the implications of the cutoff threshold B. Recall that we considered all functions $g\in\mathscr{G}$ constant with respect to the count component from B onward. This contained the class of candidate functions to a size that allowed the rate $n^{-1/3}\log n$. If we forwent this cutoff, i.e. setting $B=\infty$, we would blow up the class $\mathscr{G}$ dramatically. The consequence would be much larger covering numbers and supposedly a rate of convergence not faster that $n^{-1/4}$. However, from the model selection point of view this advantage is worthless if we have assigned B too small a value. This would lead to model misspecification, as we indicated at the beginning of the chapter. We need to find a compromise between a broad model and a fast rate of convergence.

Let m denote the true link function. We stick to the assumption that there exists a number B^* such that $y\mapsto m(\lambda,y)$ is constant for all $y\geq B^*-1$. Since m is unknown so is B^*, and we face the challenge to specify a number B such that $m\in\mathscr{G}(M,B,L_1,L_2)$. We suggest to circumvent this challenge with the following idea. Assume that $\{B_n\}_{n\in\mathbb{N}}\subset\mathbb{N}_+$ is a sequence with $B_n\leq B_{n+1}$ for all n, and $B_n\to\infty$. With a slight abuse of notation, we define the classes

$$\mathscr{G}_\infty:=\mathscr{G}(M,\infty,L_1,L_2),$$
$$\mathscr{G}_n:=\mathscr{G}(M,B_n,L_1,L_2),$$

$$\mathscr{G}^* := \mathscr{G}(M, B^*, L_1, L_2).$$

There exists a number $n^* \in \mathbb{N}$ such that $B_{n^*-1} \leq B^* \leq B_{n^*}$ and therefore

$$\mathscr{G}_0 \subset \ldots \mathscr{G}_{n^*-1} \subset \mathscr{G}^* \subset \mathscr{G}_{n^*} \subset \ldots \subset \mathscr{G}_\infty.$$

Note that for any $g \in \bigcup_{n=0}^{\infty} \mathscr{G}_n$ the strong contractive condition holds by assumption. We define a sequence of modified least squares estimators.

DEFINITION 3.2.18. Let $\{(\lambda_t, Y_t)\}_{t\in\mathbb{Z}}$ be a stationary version of a two-sided nonparametric INGARCH(1,1) process with link function $m \in \mathscr{G}^*$. The estimator $\tilde{m}_n[Y_0, \ldots, Y_n] := T_n(Y_0, \ldots, Y_n)$ of m on the basis of $n+1$ successive observations of the count process is given by the measurable selection functional $(y_0, \ldots, y_n) \mapsto T_n(y_0, \ldots, y_n)$ with

$$T_n(y_0, \ldots, y_n) \in \arg\min_{g\in\mathscr{G}_n} \sum_{i=0}^{n-1} \left(y_{i+1} - g^{[i]}(0, y_0, \ldots, y_i)\right)^2.$$

The definition is meaningful in the sense that for any n it is possible to find a selection functional T_n with the desired properties. This can be seen after an inspection of the proofs of Proposition 3.1.3 and Lemma 3.1.6. The proofs of these statements required the set of candidate functions to be totally bounded. Furthermore, we assumed that the contractive condition is valid for any candidate function. Both conditions are satisfied by the sets $\mathscr{G}_n$ for any n. Let us briefly discuss the asymptotic behavior of the sequence $\{\tilde{m}_n\}_{n\in\mathbb{N}_+}$.

PROPOSITION 3.2.19. *Suppose that $m \in \mathscr{G}(M, B^*, L_1, L_2)$, and let $\hat{m}_n$ denote the least squares estimator of m chosen from the correctly specified set of candidate functions $\mathscr{G}(M, B^*, L_1, L_2)$ on the basis of Definition 3.1.11. Let the non-decreasing sequence $\{B_n\}_{n\in\mathbb{N}} \subset \mathbb{N}_+$ satisfy $B_n \leq B_0\sqrt{\log n}$. Then the sequence of estimators $\{\tilde{m}_n\}_{n\in\mathbb{N}}$ chosen accordingly to Definition 3.2.18 from the growing classes of candidate functions $\{\mathscr{G}_n\}_{n\in\mathbb{N}}$, respectively, attains the same rate of convergence as $\{\hat{m}_n\}_{n\in\mathbb{N}}$, i.e. $L(\tilde{m}_n, m) = O_{\mathbb{P}}(n^{-2/3}(\log n)^2)$.*

Sketch of a proof. We will try to copy the approach taken in the investigation of $\hat{m}_n$. Essentially we inspect all results for the effect of replacing $\mathscr{G}$ with $\mathscr{G}_n$ as the set of candidate functions.

Lemma 3.2.3 and Lemma 3.2.4 hold for all $g \in \mathscr{G}_n$, even if $\mathscr{G}^* \subsetneq \mathscr{G}_n$. The reason is that the only requirement on g that were necessary in the proofs are the strong contractive condition (C^*). By assumption, this condition is satisfied by any $g \in \mathscr{G}_n$, for any $n \in \mathbb{N}$. Hence, Lemma 3.2.3 and Lemma 3.2.4 hold if we substitute $\mathscr{G}$ with $\mathscr{G}_n$. The same is true for Lemma 3.2.6 in which we discussed measurability of suprema over subsets $\mathscr{G}_0 \subset \mathscr{G}$. The proof required continuity

of the functional $g \mapsto f_t(g, \boldsymbol{Y}_{i-t}^{i+1})$ and separability of $\mathscr{G}$. Both arguments can be upheld with $\mathscr{G}_n$ in place of $\mathscr{G}$. Hence, the proposition remains valid for any n if we consider sub-classes $\mathscr{G}_0 \subset \mathscr{G}_n$.

Proposition 3.2.7 is unaffected by the change from $\mathscr{G}$ to $\mathscr{G}_n$ because the contractive property stays valid. Lemma 3.2.8 relied on the fact that the true function m is an element of the set of candidate functions. This is true for the candidate set $\mathscr{G}_n$ if $n \geq n^*$. Thus, for all $n \geq n*$ the lemma remains valid if we substitute the supremum over $\mathscr{G}$ with a supremum taken over $\mathscr{G}_n$. The coupling, the variance bound, and the symmetrization argument (i.e. Lemma 3.2.9 and Corollary 3.2.10, Lemma 3.2.11, and Lemma 3.2.12, respectively) remain unaffected by a substitution of $\mathscr{G}$ with $\mathscr{G}_n$. Proposition 3.2.14 uses the contractive condition which is granted with the use of $\mathscr{G}_n$ as well. Corollary 3.2.15 is a direct consequence of the former and is valid under the same conditions.

Proposition 3.2.16 is valid for $\mathscr{G}_n$ replacing $\mathscr{G}$ if we substitute B with B_n. Thus, the number of elements needed to cover the classes $\mathscr{G}_n$ with balls of radius $2^{-s+k+1}\sqrt{\gamma}\delta$ is bounded by $e^{2MB_n 2^{s-k}/(\gamma\delta)}$. Proposition 3.2.13 can be left unaltered.

In Lemma 3.2.17 the bound has to be corrected by the rate of B_n. Recall that $B_n = O(\sqrt{\log n})$. The bounds for P_1 and P_2 can be upheld. Note that the quantity C_4 introduced in line (3.34) is now $O(\sqrt{\log n})$. However, the final bound for P_2 relies on the fact that there exists a number n_1 such that

$$C_4 \delta^{-1} n^{-1/3} - C_3 \leq -\frac{C_3}{2}$$

for all $n \geq n_1$. This is still valid for $C_4 = O(\sqrt{\log n})$ because $C_4 \delta_n^{-1} n^{-1/3} = o(1)$.

The bound for P_3 has to be slightly corrected. The constant C_6 from line (3.39) would now be of order $O((\log n)^{1/4})$, and consequently $C_7 = O((\log n)^{1/4})$ as well. Thus, the term P_3 can be bounded by $C 2^{-k} (\log n)^{-3/4}$, where $C > 0$ is some positive constant. This is still good enough to obtain the rate $n^{-1/3} \log n$ for the sequence of estimators $\{\tilde{m}_n\}$ because the sum $P_1 + P_2 + P_3$ from Lemma 3.2.17 is still in $o(1)$ for any $k \in \mathbb{N}$, as $n \to \infty$. □

We are thus equipped with a tool to deal with the unknown parameter B^*. As opposed to work with a fixed value B, choosing a least squares estimator $\hat{m}_n \in \mathscr{G}(M, B, L_1, L_2)$, and risking to work with too restrictive a model, we can now choose the least squares estimator $\tilde{m}_n \in \mathscr{G}(M, B_n, L_1, L_2)$. If the sample size is sufficiently large, such that $B_n \geq B^*$, we know that $m \in \mathscr{G}_n$. Thus, asymptotically, the alteration of the model by introducing the boundaries B_n does not result in an elevated risk of model misspecification. We have to bear in mind, however, that a larger parameter B_n induces a larger class of candidate functions, which makes it harder to find an actual least squares estimation in these classes.

3.2.4 Conclusion and final remarks

This chapter was concerned with the problem of estimating the link function in a nonparametric INGARCH(1,1) model with hidden intensities under the strong contractive assumption. We proposed a least squares estimator that selects an estimation as a minimizer of the sum of squares functional $\sum_{i=0}^{n-1}\left(Y_{i+1}-g^{[i]}(0,Y_0,\dots,Y_i)\right)^2$ over the function class of candidate functions $\mathscr{G}$ or, in case of growing classes, $\mathscr{G}_n$. We assured that this estimator is well defined and examined its performance in terms of the rate of convergence of its $L_2(\pi)$-risk. As the main result of this chapter, Theorem 3.2.2 states that the $L_2(\pi)$-risk is in $O_{\mathbb{P}}(n^{-2/3}(\log n)^2)$.

The core arguments in the proof of this claim were the coupling in Lemma 3.2.9 and the chaining approximation technique in Lemma 3.2.17. Both arguments relied to some extent on the assumption of full contractivity. The coupling used the uniform mixing property of the data generating process. This property was derived in Chapter 1 under the assumption that the link function m is a full contraction. However, there is room for a relaxation of the contractive assumption with respect to the coupling. Berbee's coupling lemma, which was the basis of our argument, can as well be applied to processes that are absolutely regular instead of uniformly mixing. In a recent publication, Doukhan and Neumann (2018) found a way to prove absolute regularity of the count process under the semi-contractive condition (C_*). This means that the coupling should work under the semi-contractive condition. On the contrary, the chaining argument fails if we drop the assumption that all candidate functions $g \in \mathscr{G}$ satisfy (C_*).

Chapter 4

Practical nonparametric inference on semi-contractive link functions

So far, we have proven that a theoretically attainable realization of the least squares estimator from Definition 3.1.11 is consistent with the rate $\delta_n = n^{-1/3}\log n$. However, even under the premise that this rate is nearly optimal, there are several shortcomings of the previous approach.

First, the estimator was defined as the solution of an optimization over a huge class of functions. Finding the solution to such a problem is a computationally unfeasible task. We desire an estimator that is easier to obtain.

Second, in the preceding chapter we assumed the constants L_1 and L_2 to be fixed. The quality of an approximation of the theoretical estimator would depend heavily on our ability to guess these parameters. Of course we know that the constants are not larger than one. But apart from this, virtually anything can happen. Choosing a class $\mathscr{G}$ of candidate functions with too small specifications of L_1 and L_2 would introduce a structural estimation error that is asymptotically not negligible. On the other hand, choosing larger constants than necessary would waste computational resources as the complexity of the underlying optimization problem soars. It would be better to have an estimator that does not necessarily depend on the specification of the true constants L_1 and L_2.

Third, we have seen that the contraction property for the second component (i.e. $L_2 < 1 - L_1$) is only necessary to obtain uniform mixing of the count process and to deliver the chaining argument. If we had an estimator the asymptotic analysis of which does not rely on these tools, we could drop the assumption of full contractivity and thus broaden our model.

The next section shows how we can achieve these goals. Restricting the set of candidate functions to a finite grid allows us to abandon the chaining approximation technique in the asymptotic analysis. Furthermore, if the grid is

fine enough, we shall see that the assumption of contractivity can be relinquished entirely with respect to the candidate functions and partially with respect to the true function. Considering convergence in terms of the empirical mean square error enables us to use martingale techniques to get hold on the dependencies of the data. We can therefore abandon all considerations related to the notion of mixing and most of the model assumptions associated to the derivation of this property.

4.1 Estimation on a finite grid of functions

4.1.1 Preliminaries

In the following definition we fix the set of assumptions that we impose throughout this chapter. The notation will differ slightly from the last chapter since it is better suited to clarify the underlying ideas in the asymptotic analysis.

DEFINITION 4.1.1. (a) For fixed constants $M > 0$, $B \in \mathbb{N}_+$, and $0 \le \ell < 1$, let the function $m\colon [0,M] \times \mathbb{N} \to [0,M]$ satisfy the semi-contractive condition (C_*) with constant ℓ, and assume furthermore that $m(\lambda, y) = m(\lambda, B-1)$ for all $y \ge B-1$ and all $\lambda \in [0,M]$. Let the data generating process $\{(\lambda_t, Y_t)\}_{t\in\mathbb{Z}}$ be given by a stationary version of a two-sided nonparametric INGARCH(1,1) process with link function m.

(b) Let $\{B_n\}_{n\in\mathbb{N}_+} \subset \mathbb{N}_+$ be a non-decreasing unbounded sequence of natural numbers such that $B_n \le B_0 \log n$ for all n, and let $L \in [\ell, \infty)$ be fixed. For $n \in \mathbb{N}_+$, the set $\mathscr{G}(M, B_n, L)$ of possible candidate functions is defined as the set of all functions $g\colon [0,M] \times \mathbb{N} \to [0,M]$ that satisfy

$$\sup_y |g(\lambda_1, y) - g(\lambda_2, y)| \le L|\lambda_1 - \lambda_2| \tag{L}$$

and furthermore $g(\lambda, y) = g(\lambda, B_n - 1)$ for all $y \ge B_n - 1$ and all λ.

(c) For $n \in \mathbb{N}_+$ and a function $g \in \mathscr{G}(M, B_n, L)$, we define the processes $\{\lambda_i^g\}_{i\in\mathbb{N}_+}$ by $\lambda_{i+1}^g := g^{[i]}(0, Y_0, \ldots, Y_i)$.

(d) For $n \in \mathbb{N}_+$, let $\mathscr{G}_n \subset \mathscr{G}(M, B_n, L)$ be a finite subset of candidate functions. Let $m_n \in \mathscr{G}_n$ be an element in $\mathscr{G}_n$ such that

$$\|m - m_n\|_\infty = \min_{g\in\mathscr{G}_n} \|m - g\|_\infty,$$

i.e. m_n is a best approximation of m among all elements in $\mathscr{G}_n$. The sequence $\{\rho_n\}_{n\in\mathbb{N}_+} \subset \mathbb{R}$ shall be given by $\rho_n := \|m - m_n\|_\infty$. The estimator $\hat{m}_n[Y_0,\dots,Y_n]$ on the basis of observations of $Y_0,\dots,Y_n$ is defined as

$$\hat{m}_n[Y_0,\dots,Y_n] := \arg\min_{g\in\mathscr{G}_n} \sum_{i=0}^{n-1} \left(Y_{i+1} - g^{[i]}(0,Y_0,\dots,Y_n)\right)^2 = \arg\min_{g\in\mathscr{G}_n} \sum_{i=1}^{n} \left(Y_i - \lambda_i^g\right)^2$$

We call $\hat{m}_n$ the *approximate least squares estimator* on the basis of the approximating set $\mathscr{G}_n$.

The function classes $\mathscr{G}(M,B_n,L)$ of candidate functions differ from the classes $\mathscr{G}$ and $\mathscr{G}(M,B_n,L_1,L_2)$ that we considered in the previous chapter. Here we impose considerably less smoothness on the candidate functions: in the first argument the contractive condition is weakened to a simple Lipschitz condition, and in the second argument no restrictions are imposed at all. We remark that certainly $m \in \mathscr{G}(M,B_n,L)$ if n is sufficiently large. This is the case if $B_n \geq B$. But even then it is not true in general that m is an element of $\mathscr{G}_n$. We chose the name 'approximate least squares' since the estimator minimizes the sum of squares over the subclass $\mathscr{G}_n$ as opposed to the whole class $\mathscr{G}(M,B_n,L)$.

4.1.2 Asymptotic analysis

As in the previous chapter, we investigate the asymptotic properties of $\hat{m}_n$ in terms of convergence in probability with the sample size tending to infinity. We differ, however, from the previous chapter in the measure of distance between m and $\hat{m}_n$. As opposed to investigating the $L_2(\pi)$ risk, we want to consider the asymptotic properties of the empirical mean square error (MSE),

$$\frac{1}{n}\sum_{i=2}^{n} \left(m(\lambda_{i-1},Y_{i-1}) - \hat{m}_n(\lambda_{i-1},Y_{i-1})\right)^2.$$

We begin the investigation with two preliminary lemmas.

LEMMA 4.1.2. *Recall the definitions of the estimator $\hat{m}_n$ and the process $\{\lambda_i^{\hat{m}_n}\}$ from Definition 4.1.1. Then*

$$\frac{1}{n}\sum_{i=2}^{n} \left(m(\lambda_{i-1},Y_{i-1}) - \hat{m}_n(\lambda_{i-1},Y_{i-1})\right)^2 \leq 4(1+L^2)\,\frac{1}{n}\sum_{i=1}^{n} \left(\lambda_i - \lambda_i^{\hat{m}_n}\right)^2$$

for all $\omega \in \Omega$.

Proof. For $i = 2, \dots, n$ it holds that

$$\begin{aligned}
&\Big(m(\lambda_{i-1}, Y_{i-1}) - \hat{m}_n(\lambda_{i-1}, Y_{i-1})\Big)^2 \\
&= \Big(m(\lambda_{i-1}, Y_{i-1}) - \hat{m}_n(\lambda_{i-1}^{\hat{m}_n}, Y_{i-1}) + \hat{m}_n(\lambda_{i-1}^{\hat{m}_n}, Y_{i-1}) - \hat{m}_n(\lambda_{i-1}, Y_{i-1})\Big)^2 \\
&\le 2\Big(m(\lambda_{i-1}, Y_{i-1}) - \hat{m}_n(\lambda_{i-1}^{\hat{m}_n}, Y_{i-1})\Big)^2 + 2\Big(\hat{m}_n(\lambda_{i-1}^{\hat{m}_n}, Y_{i-1}) - \hat{m}_n(\lambda_{i-1}, Y_{i-1})\Big)^2 \\
&\le 2\Big(m(\lambda_{i-1}, Y_{i-1}) - \hat{m}_n(\lambda_{i-1}^{\hat{m}_n}, Y_{i-1})\Big)^2 + 2L^2\Big(\lambda_{i-1}^{\hat{m}_n} - \lambda_{i-1}\Big)^2 \\
&= 2\Big(\lambda_i - \lambda_i^{\hat{m}_n}\Big)^2 + 2L^2\Big(\lambda_{i-1}^{\hat{m}_n} - \lambda_{i-1}\Big)^2 \\
&\le 2(1+L^2)\Big[\big(\lambda_i - \lambda_i^{\hat{m}_n}\big)^2 + \big(\lambda_{i-1}^{\hat{m}_n} - \lambda_{i-1}\big)^2\Big].
\end{aligned}$$

Therefore, we obtain the following estimate:

$$\begin{aligned}
&\frac{1}{n}\sum_{i=2}^{n}\Big(m(\lambda_{i-1}, Y_{i-1}) - \hat{m}_n(\lambda_{i-1}, Y_{i-1})\Big)^2 \\
&\le \frac{2(1+L^2)}{n}\Big[\sum_{i=2}^{n}\big(\lambda_i - \lambda_i^{\hat{m}_n}\big)^2 + \sum_{i=2}^{n}\big(\lambda_{i-1} - \lambda_{i-1}^{\hat{m}_n}\big)^2\Big] \\
&\le 4(1+L^2)\frac{1}{n}\sum_{i=1}^{n}\big(\lambda_i - \lambda_i^{\hat{m}_n}\big)^2. \qquad \square
\end{aligned}$$

LEMMA 4.1.3. *There exists a sequence $\{t_n\}$ with $0 < t_n < n$ such that for all $\omega \in \Omega$ the following relation holds:*

$$\begin{aligned}
&\frac{2}{n}\sum_{i=1}^{n}(Y_i - \lambda_i)\Big[\big(\lambda_i - \lambda_i^{m_n}\big) - \big(\lambda_i - \lambda_i^{\hat{m}_n}\big)\Big] \\
&\qquad \ge \frac{1}{n}\sum_{i=1}^{n}\big(\lambda_i - \lambda_i^{\hat{m}_n}\big)^2 - \left[\frac{1+M}{1-\ell}\right]^2\Big(\frac{t_n}{n} + \big(\rho_n + \ell^{t_n}\big)^2\Big).
\end{aligned}$$

Proof. In virtue of the assumptions on m and m_n, we conclude that

$$\begin{aligned}
\big|\lambda_i - \lambda_i^{m_n}\big| &= \big|m(\lambda_{i-1}, Y_{i-1}) - m_n(\lambda_{i-1}^{m_n}, Y_{i-1})\big| \\
&\le \big|m(\lambda_{i-1}, Y_{i-1}) - m(\lambda_{i-1}^{m_n}, Y_{i-1})\big| \\
&\qquad + \big|m(\lambda_{i-1}^{m_n}, Y_{i-1}) - m_n(\lambda_{i-1}^{m_n}, Y_{i-1})\big| \\
&\le \ell\,\big|\lambda_{i-1} - \lambda_{i-1}^{m_n}\big| + \underbrace{\|m - m_n\|_\infty}_{=\rho_n} \\
&\le \ell\Big(\ell\,\big|\lambda_{i-2} - \lambda_{i-2}^{m_n}\big| + \rho_n\Big) + \rho_n \\
&\;\;\vdots \\
&\le \ell^j\,\big|\lambda_{i-j} - \lambda_{i-j}^{m_n}\big| + \sum_{k=0}^{j-1}\ell^k\rho_n \\
&\;\;\vdots
\end{aligned}$$

$$
\begin{aligned}
&\leq \ell^i \left|\lambda_0 - \lambda_0^{m_n}\right| + \sum_{k=0}^{i-1} \ell^k \rho_n \\
&\leq M\,\ell^i + \frac{\rho_n}{1-\ell} \\
&\leq \frac{1+M}{1-\ell}\big(\rho_n + \ell^i\big)
\end{aligned}
$$

for all $i \in \{1,\dots,n\}$. It follows that

$$
\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\big(\lambda_i - \lambda_i^{m_n}\big)^2 &= \frac{1}{n}\sum_{i=1}^{t_n}\big(\lambda_i - \lambda_i^{m_n}\big)^2 + \frac{1}{n}\sum_{i=t_n+1}^{n}\big(\lambda_i - \lambda_i^{m_n}\big)^2 \\
&\leq M^2\frac{t_n}{n} + \frac{n-t_n}{n}\left[\frac{1+M}{1-\ell}\right]^2\big(\rho_n + \ell^{t_n}\big)^2 \\
&\leq \left[\frac{1+M}{1-\ell}\right]^2\Big(\frac{t_n}{n} + \big(\rho_n + \ell^{t_n}\big)^2\Big). \qquad (4.1)
\end{aligned}
$$

The next inequality is basically an immediate consequence of the definition of $\hat{m}_n$ combined with the fact that $m_n \in \mathscr{G}_n$.

$$
\begin{aligned}
0 &\leq \frac{1}{n}\sum_{i=1}^{n}\big(Y_i - \lambda_i^{m_n}\big)^2 - \frac{1}{n}\sum_{i=1}^{n}\big(Y_i - \lambda_i^{\hat{m}_n}\big)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\big(Y_i - \lambda_i + \lambda_i - \lambda_i^{m_n}\big)^2 - \frac{1}{n}\sum_{i=1}^{n}\big(Y_i - \lambda_i + \lambda_i - \lambda_i^{\hat{m}_n}\big)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\big(Y_i - \lambda_i\big)^2 + \frac{1}{n}\sum_{i=1}^{n}\big(\lambda_i - \lambda_i^{m_n}\big)^2 + \frac{2}{n}\sum_{i=1}^{n}\big(Y_i - \lambda_i\big)\big(\lambda_i - \lambda_i^{m_n}\big) \\
&\quad - \frac{1}{n}\sum_{i=1}^{n}\big(Y_i - \lambda_i\big)^2 - \frac{1}{n}\sum_{i=1}^{n}\big(\lambda_i - \lambda_i^{\hat{m}_n}\big)^2 - \frac{2}{n}\sum_{i=1}^{n}\big(Y_i - \lambda_i\big)\big(\lambda_i - \lambda_i^{\hat{m}_n}\big) \\
&= \frac{2}{n}\sum_{i=1}^{n}(Y_i - \lambda_i)\Big[\big(\lambda_i - \lambda_i^{m_n}\big) - \big(\lambda_i - \lambda_i^{\hat{m}_n}\big)\Big] + \frac{1}{n}\sum_{i=1}^{n}\big(\lambda_i - \lambda_i^{m_n}\big)^2 \\
&\qquad - \frac{1}{n}\sum_{i=1}^{n}\big(\lambda_i - \lambda_i^{\hat{m}_n}\big)^2.
\end{aligned}
$$

Hence,

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\big(\lambda_i - \lambda_i^{\hat{m}_n}\big)^2 \\
&\leq \frac{2}{n}\sum_{i=1}^{n}(Y_i - \lambda_i)\Big[\big(\lambda_i - \lambda_i^{m_n}\big) - \big(\lambda_i - \lambda_i^{\hat{m}_n}\big)\Big] + \frac{1}{n}\sum_{i=1}^{n}\big(\lambda_i - \lambda_i^{m_n}\big)^2 \\
&\leq \frac{2}{n}\sum_{i=1}^{n}(Y_i - \lambda_i)\Big[\big(\lambda_i - \lambda_i^{m_n}\big) - \big(\lambda_i - \lambda_i^{\hat{m}_n}\big)\Big] + \left[\frac{1+M}{1-\ell}\right]^2\Big(\frac{t_n}{n} + \big(\rho_n + \ell^{t_n}\big)^2\Big). \qquad \square
\end{aligned}
$$

Let us shortly pause to think about the consequences of Lemma 4.1.2 and Lemma 4.1.3. Assume for a moment that $\lambda_i = \lambda_i^{m_n}$ for all $i \in \mathbb{N}$. In this case the statement of Lemma 4.1.2 combined with the second part of the proof of Lemma

4.1.3 would have yielded that

$$\frac{1}{n}\sum_{i=2}^{n}\Big(m(\lambda_{i-1},Y_{i-1})-\hat{m}_n(\lambda_{i-1},Y_{i-1})\Big)^2 \lesssim \frac{1}{n}\sum_{i=1}^{n}\big(\lambda_i-\lambda_i^{\hat{m}_n}\big)^2$$
$$\leq \frac{1}{n}\sum_{i=1}^{n}(Y_i-\lambda_i)\big(\lambda_i^{\hat{m}_n}-\lambda_i\big). \qquad (4.2)$$

This is the essence of both lemmas. The difference between inequality (4.2) and Lemma 4.1.3 is merely attributed to the fact that $|\lambda_i-\lambda_i^{m_n}|>0$. However, due to the contraction property, the iteration scheme ensures that $|\lambda_i-\lambda_i^{m_n}|$ is of negligible size if i is large enough. The quantity t_n serves as a threshold value: for $i\geq t_n$ we have $\lambda_i\approx\lambda_i^{m_n}$, otherwise their difference may be substantial. The error term in Lemma 4.1.3 reflects the fact that we have to balance the requirements $t_n\to\infty$ and $\frac{t_n}{n}\to 0$ to approximately uphold (4.2).

In order to find a bound for the MSE of the approximate least squares estimator, it suffices to control the magnitude of the linear term in line (4.2). The subject of the following pages is an examination of this linear term. Using its martingale structure, we will derive a bound in probability. To get prepared for the core argument, we first invoke a peeling argument that is similar to the approach of Lemma 3.2.8.

LEMMA 4.1.4. *Assume that the sequences $\{\delta_n\}$, $\{t_n\}$, $\{\rho_n\}$, as well as the constant ℓ satisfy the condition*

$$\limsup_{n\to\infty}\frac{\frac{t_n}{n}+\big(\rho_n+\ell^{t_n}\big)^2}{\delta_n^2}=0.$$

Then there exists a constant $\gamma>0$ and a number n_0 such that

$$\mathbb{P}\bigg\{\frac{1}{n}\sum_{i=2}^{n}\Big(m(\lambda_{i-1},Y_{i-1})-\hat{m}_n(\lambda_{i-1},Y_{i-1})\Big)^2>\delta_n^2\bigg\}$$
$$\leq\sum_{k=1}^{\infty}\mathbb{P}\bigg\{\exists g\in\mathscr{G}_n\colon 2^{k-1}\gamma\delta_n^2<\frac{1}{n}\sum_{i=1}^{n}\big(\lambda_i-\lambda_i^g\big)^2\leq 2^k\gamma\delta_n^2\ ;$$
$$\frac{2}{n}\sum_{i=1}^{n}(Y_i-\lambda_i)\Big[(\lambda_i-\lambda_i^{m_n})-(\lambda_i-\lambda_i^g)\Big]>2^{k-1}\frac{\gamma}{2}\delta_n^2\bigg\}$$

for all $n\in\mathbb{N}$ with $n>n_0$.

Proof. In view of Lemma 4.1.2, we see that for $\gamma:=\frac{1}{4(1+L^2)}$ the following chain of inequalities holds:

$$\mathbb{P}\bigg\{\frac{1}{n}\sum_{i=2}^{n}\Big(m(\lambda_{i-1},Y_{i-1})-\hat{m}_n(\lambda_{i-1},Y_{i-1})\Big)^2>\delta_n^2\bigg\}$$
$$\leq\mathbb{P}\bigg\{\frac{1}{n}\sum_{i=1}^{n}\big(\lambda_i-\lambda_i^{\hat{m}_n}\big)^2>\gamma\delta_n^2\bigg\}$$

$$\leq \mathbb{P}\bigcup_{k=1}^{\infty}\left\{2^{k-1}\gamma\delta_n^2 < \frac{1}{n}\sum_{i=1}^{n}\left(\lambda_i - \lambda_i^{\hat{m}_n}\right)^2 \leq 2^k\gamma\delta_n^2\right\}$$
$$\leq \sum_{k=1}^{\infty}\mathbb{P}\left\{2^{k-1}\gamma\delta_n^2 < \frac{1}{n}\sum_{i=1}^{n}\left(\lambda_i - \lambda_i^{\hat{m}_n}\right)^2 \leq 2^k\gamma\delta_n^2\right\}.$$

From the assumption in the formulation of the lemma, we conclude that there exists a number $n_0 \in \mathbb{N}$ such that for all $n > n_0$ and all $k \in \mathbb{N}$

$$\left[\frac{1+M}{1-\ell}\right]^2\left(\frac{t_n}{n} + \left(\rho_n + \ell^{t_n}\right)^2\right) \leq 2^{k-2}\gamma\delta_n^2.$$

Using the result of Lemma 4.1.3 and the fact that $\hat{m}_n \in \mathscr{G}_n$, we can continue the above chain of inequalities in the following way: for all $n > n_0$

$$\begin{aligned}
&\sum_{k=1}^{\infty}\mathbb{P}\left\{2^{k-1}\gamma\delta_n^2 < \frac{1}{n}\sum_{i=1}^{n}\left(\lambda_i - \lambda_i^{\hat{m}_n}\right)^2 \leq 2^k\gamma\delta_n^2\right\}\\
&= \sum_{k=1}^{\infty}\mathbb{P}\Bigg\{2^{k-1}\gamma\delta_n^2 < \frac{1}{n}\sum_{i=1}^{n}\left(\lambda_i - \lambda_i^{\hat{m}_n}\right)^2 \leq 2^k\gamma\delta_n^2\ ;\\
&\qquad \frac{2}{n}\sum_{i=1}^{n}(Y_i - \lambda_i)\left[\left(\lambda_i - \lambda_i^{m_n}\right) - \left(\lambda_i - \lambda_i^{\hat{m}_n}\right)\right]\\
&\qquad\quad \geq \underbrace{\frac{1}{n}\sum_{i=1}^{n}\left(\lambda_i - \lambda_i^{\hat{m}_n}\right)^2}_{>2^{k-1}\gamma\delta_n^2} - \underbrace{\left[\frac{1+M}{1-\ell}\right]^2\left(\frac{t_n}{n} + \left(\rho_n + \ell^{t_n}\right)^2\right)}_{\leq \frac{1}{2}2^{k-1}\gamma\delta_n^2}\Bigg\}\\
&\leq \sum_{k=1}^{\infty}\mathbb{P}\Bigg\{2^{k-1}\gamma\delta_n^2 < \frac{1}{n}\sum_{i=1}^{n}\left(\lambda_i - \lambda_i^{\hat{m}_n}\right)^2 \leq 2^k\gamma\delta_n^2\ ;\\
&\qquad \frac{2}{n}\sum_{i=1}^{n}(Y_i - \lambda_i)\left[\left(\lambda_i - \lambda_i^{m_n}\right) - \left(\lambda_i - \lambda_i^{\hat{m}_n}\right)\right] > \frac{1}{2}2^{k-1}\gamma\delta_n^2\Bigg\}.
\end{aligned}$$

□

DEFINITION 4.1.5. For every $g \in \mathscr{G}(M, B_n, L)$, we define the process $\{X_i(g)\}_{i\in\mathbb{N}_+}$ by

$$X_i(g) := (Y_i - \lambda_i)\left[\left(\lambda_i - \lambda_i^{m_n}\right) - \left(\lambda_i - \lambda_i^{g}\right)\right].$$

LEMMA 4.1.6. *For every function $g \in \mathscr{G}(M, B_n, L)$, the process $\{X_i(g)\}$ is a square integrable martingale difference with respect to the natural filtration $\{\mathscr{F}_n\}$ generated by the bivariate data generating process. Consequently, the processes $\{M_n(g)\}_{n\in\mathbb{N}}$ with $M_0(g) = 0$ and*

$$M_n(g) := \sum_{i=1}^{n} X_i(g)$$

are square integrable $\{\mathscr{F}_n\}$-martingales.

Proof. We remark that the variables λ_i, $\lambda_i^{m_n}$, and λ_i^g are measurable with respect to $\mathscr{F}_{i-1}$. Hence,

$$\begin{aligned}\mathbb{E}\big[X_i(g)\,|\,\mathscr{F}_{i-1}\big] &= \mathbb{E}\big[(Y_i-\lambda_i)\big[(\lambda_i-\lambda_i^{m_n})-(\lambda_i-\lambda_i^g)\big]\,|\,\mathscr{F}_{i-1}\big]\\ &= \big[(\lambda_i-\lambda_i^{m_n})-(\lambda_i-\lambda_i^g)\big]\;\mathbb{E}\big[(Y_i-\lambda_i)\,|\,\mathscr{F}_{i-1}\big]\\ &= 0,\ \text{a.s. .}\end{aligned}$$

Clearly, $X_i(g)$ is $\mathscr{F}_i$-measurable, and $\mathbb{E}|X_i(g)|^2<\infty$. The immediate consequence is that $\{M_n(g)\}$ is adapted to the filtration $\{\mathscr{F}_n\}$, and

$$\mathbb{E}[M_n(g)\,|\,\mathscr{F}_{n-1}] = \mathbb{E}[M_{n-1}(g)+X_n(g)\,|\,\mathscr{F}_{n-1}] = M_{n-1}(g),\ \text{a.s. .}$$

Furthermore, $M_n(g)$ is square integrable because the triangle inequality implies that

$$\big(\mathbb{E}M_n^2(g)\big)^{1/2} \le \sum_{i=1}^n \big(\mathbb{E}X_i^2(g)\big)^{1/2} < \infty$$

for all n. Hence, $\{M_n(g)\}$ is a square integrable martingale with respect to the filtration $\{\mathscr{F}_n\}$. □

At this point it may become clear why we need to substitute all events describing $\hat{m}_n$ with events making uniform statements over $\mathscr{G}_n$. If we had not done this, we would be confronted with the process $\{M_n(\hat{m}_n)\}$. This process is not a martingale since $\lambda_n^{\hat{m}_n} = \lambda_n^{\hat{m}_n[Y_0,\ldots,Y_n]}$ depends on the full information up to Y_n and is thus not measurable with respect to $\mathscr{F}_{n-1}$. Therefore, $\mathbb{E}[X_n(\hat{m}_n)\,|\,\mathscr{F}_{n-1}]$ is not in general zero.

Our main tool in the subsequent analysis will be an exponential tail bound for the martingales $\{M_n(g)\}$. We use a result of Dzhaparidze and van Zanten (2001). In order to apply their result, we have to control the following quantity that measures the magnitude of $M_n(g)$.

DEFINITION 4.1.7. Let $\{a_n\}$ be a sequence of positive numbers. For a function $g\in\mathscr{G}(M,B_n,L)$, the random variable $H_n^{a_n}(g)$ is defined as

$$H_n^{a_n}(g) := \sum_{i=1}^n X_i^2(g)\,\mathbb{1}_{\{|X_i(g)|>a_n\}} + \sum_{i=1}^n \mathbb{E}\big[X_i^2(g)\,|\,\mathscr{F}_{i-1}\big].$$

The second term in the definition of $H_n^{a_n}(g)$ is the sum of conditional variances. Finding a way to bound it will be crucial for our purpose. The first term, however, will be dominated with high probability by the variance term if the cutoff threshold a_n is large enough and the increments $X_i(g)$ satisfy a suitable moment condition. We remark that the resemblance in the notation between the objects of Definition

4.1.7 and Theorem A.4.7 is intended. It is justified by the following observations. By definition of $M_n(g)$, the jump process of increments is given by $\Delta M_n(g) = X_n(g)$. Furthermore, the predictable quadratic variation of the process $M(g)$ is given by $\langle M(g)\rangle_0 = 0$ and

$$\begin{aligned}\langle M(g)\rangle_n &= \sum_{i=1}^{n}\Big(\mathbb{E}\big[M_i^2(g)\,|\,\mathscr{F}_{i-1}\big] - M_{i-1}^2(g)\Big)\\ &= \sum_{i=1}^{n}\mathbb{E}\Big[2M_{i-1}(g)X_i(g) + X_i^2(g)\,\Big|\,\mathscr{F}_{i-1}\Big]\\ &= \sum_{i=1}^{n}\mathbb{E}\big[X_i^2(g)\,|\,\mathscr{F}_{i-1}\big],\ \text{a.s.}\end{aligned}$$

for $n \geq 1$. This follows from the Doob-Meyer decomposition (Karatzas and Shreve, 1991, page 21).

LEMMA 4.1.8. *For the predictable quadratic variation of the Martingale $M(g)$, we have the following bound:*

$$\sum_{i=1}^{n}\mathbb{E}\big[X_i^2(g)\,|\,\mathscr{F}_{i-1}\big] \leq \frac{2(1+M)^3}{(1-\ell)^2}\Big[t_n + n\big(\rho_n + \ell^{t_n}\big)^2 + \sum_{i=1}^{n}\big(\lambda_i - \lambda_i^g\big)^2\Big],\ \ a.s..$$

Proof. We compute straightforwardly that with probability one

$$\begin{aligned}\sum_{i=1}^{n}\mathbb{E}\big[X_i^2(g)\,|\,\mathscr{F}_{i-1}\big] &= \sum_{i=1}^{n}\mathbb{E}\Big[(Y_i-\lambda_i)^2\big[(\lambda_i - \lambda_i^{m_n}) - (\lambda_i - \lambda_i^g)\big]^2\,|\,\mathscr{F}_{i-1}\Big]\\ &= \sum_{i=1}^{n}\big[(\lambda_i - \lambda_i^{m_n}) - (\lambda_i - \lambda_i^g)\big]^2\mathbb{E}\big[(Y_i-\lambda_i)^2\,|\,\mathscr{F}_{i-1}\big]\\ &= \sum_{i=1}^{n}\lambda_i\big[(\lambda_i - \lambda_i^{m_n}) - (\lambda_i - \lambda_i^g)\big]^2\\ &\leq 2M\Big[\sum_{i=1}^{n}(\lambda_i - \lambda_i^{m_n})^2 + \sum_{i=1}^{n}(\lambda_i - \lambda_i^g)^2\Big]\\ &\leq 2M\Big[n\Big[\frac{1+M}{1-\ell}\Big]^2\Big(\frac{t_n}{n} + \big(\rho_n + \ell^{t_n}\big)^2\Big) + \sum_{i=1}^{n}\big(\lambda_i - \lambda_i^g\big)^2\Big]\\ &\leq \frac{2(1+M)^3}{(1-\ell)^2}\Big[t_n + n\big(\rho_n + \ell^{t_n}\big)^2 + \sum_{i=1}^{n}\big(\lambda_i - \lambda_i^g\big)^2\Big].\end{aligned}$$

In the next-to-last line, we used the inequality (4.1) to bound $\sum_{i=1}^{n}\big(\lambda_i - \lambda_i^{m_n}\big)^2$. This is the desired result. □

LEMMA 4.1.9. *Let $a_{n,k} = 2(M^2+2M)(k+1)\log n$. Then for all $n \in \mathbb{N}$ with $n > e$, and all $k \in \mathbb{N}$,*

$$\mathbb{P}\Big\{\exists g \in \mathscr{G}_n\colon \sum_{i=1}^{n}X_i^2(g)\mathbb{1}_{\{|X_i(g)|>a_{n,k}\}} \neq 0\Big\} \leq e^M n^{-(k+1)}.$$

Proof. Certainly, for all natural numbers $n > e$ and all $k \in \mathbb{N}$

$$(k+1)\log(n)\frac{2M^2+4M}{M} \geq 2M\,.$$

The following chain of inequalities holds for all $n > e$:

$$\begin{aligned}
&\mathbb{P}\Big\{\exists\, g \in \mathscr{G}_n\colon \sum_{i=1}^{n} X_i^2(g)\,\mathbb{1}_{\{|X_i(g)|>a_{n,k}\}} \neq 0\Big\}\\
&\leq \mathbb{P}\Big\{\exists\, g \in \mathscr{G}_n\colon \max_{i\leq n}|X_i(g)| > a_{n,k}\Big\}\\
&= \mathbb{P}\Big\{\exists\, g \in \mathscr{G}_n\colon \max_{i\leq n}|Y_i-\lambda_i|\,|\lambda_i-\lambda_i^{m_n}-\lambda_i+\lambda_i^{g}| > a_{n,k}\Big\}\\
&\leq \mathbb{P}\Big\{\exists\, g \in \mathscr{G}_n\colon \max_{i\leq n}(Y_i+\lambda_i)\,|\lambda_i^{g}-\lambda_i^{m_n}| > a_{n,k}\Big\}\\
&\leq \mathbb{P}\Big\{\max_{i\leq n}(Y_i+M)\,M > a_{n,k}\Big\}\\
&\leq \mathbb{P}\Big\{\max_{i\leq n} Y_i > \frac{2M^2+4M}{M}(k+1)\log(n) - M\Big\}\\
&\leq \mathbb{P}\Big\{\max_{i\leq n} Y_i > \underbrace{\frac{2M^2+4M}{2M}}_{=2+M>2}(k+1)\log(n)\Big\}\\
&\leq \mathbb{P}\Big\{\max_{i\leq n} Y_i > 2(k+1)\log(n)\Big\}\,.
\end{aligned}$$

Now the assertion follows in view of Proposition 3.2.13. □

COROLLARY 4.1.10. *Assume that*

$$\limsup_{n\to\infty}\frac{t_n + n\big(\rho_n+\ell^{t_n}\big)^2}{n\delta_n^2} = 0\,.$$

Then there exists a number n_0 such that for all $n \in \mathbb{N}$ with $n > n_0$ and all $k \in \mathbb{N}$,

$$\sum_{k=1}^{\infty}\mathbb{P}\Big\{\exists\, g\in\mathscr{G}_n\colon 2^{k-1}\gamma\delta_n^2 < \frac{1}{n}\sum_{i=1}^{n}\big(\lambda_i-\lambda_i^g\big)^2 \leq 2^k\gamma\delta_n^2\ ;\ H_n^{a_{n,k}}(g) > \frac{8(1+M)^3}{(1-\ell)^2}2^k\gamma n\delta_n^2\Big\} \leq \frac{e^M}{n}\,.$$

Proof. Using

$$H_n^{a_{n,k}}(g) := \sum_{i=1}^{n} X_i^2(g)\,\mathbb{1}_{\{|X_i(g)|>a_{n,k}\}} + \sum_{i=1}^{n}\mathbb{E}\big[X_i^2(g)\,|\,\mathscr{F}_{i-1}\big]$$

and the triangle inequality for probabilities, we observe that

$$\begin{aligned}
&\sum_{k=1}^{\infty}\mathbb{P}\Big\{\exists\, g\in\mathscr{G}_n\colon 2^{k-1}\gamma\delta_n^2 < \frac{1}{n}\sum_{i=1}^{n}\big(\lambda_i-\lambda_i^g\big)^2 \leq 2^k\gamma\delta_n^2\ ;\ H_n^{a_{n,k}}(g) > \frac{8(1+M)^3}{(1-\ell)^2}2^k\gamma n\delta_n^2\Big\}\\
&\leq \sum_{k=1}^{\infty}\mathbb{P}\Big\{\exists\, g\in\mathscr{G}_n\colon 2^{k-1}\gamma\delta_n^2 < \frac{1}{n}\sum_{i=1}^{n}\big(\lambda_i-\lambda_i^g\big)^2 \leq 2^k\gamma\delta_n^2\ ;
\end{aligned}$$

$$\sum_{i=1}^{n} X_i^2(g)\mathbb{1}_{\{|X_i^2(g)|>a_{n,k}\}} > \frac{4(1+M)^3}{(1-\ell)^2}2^k\gamma n\delta_n^2\Big\}$$
$$+\sum_{k=1}^{\infty}\mathbb{P}\Big\{\exists g\in\mathscr{G}_n\colon 2^{k-1}\gamma\delta_n^2 < \frac{1}{n}\sum_{i=1}^{n}\big(\lambda_i-\lambda_i^g\big)^2 \le 2^k\gamma\delta_n^2\ ;$$
$$\sum_{i=1}^{n}\mathbb{E}[X_i^2(g)\,|\,\mathscr{F}_{i-1}] > \frac{4(1+M)^3}{(1-\ell)^2}2^k\gamma n\delta_n^2\Big\}$$
$$\le \sum_{k=1}^{\infty}\mathbb{P}\Big\{\exists g\in\mathscr{G}_n\colon \sum_{i=1}^{n} X_i^2(g)\mathbb{1}_{\{|X_i^2(g)|>a_{n,k}\}} > \frac{4(1+M)^3}{(1-\ell)^2}2^k\gamma n\delta_n^2\Big\}$$
$$+\sum_{k=1}^{\infty}\mathbb{P}\Big\{\exists g\in\mathscr{G}_n\colon 2^{k-1}\gamma\delta_n^2 < \frac{1}{n}\sum_{i=1}^{n}\big(\lambda_i-\lambda_i^g\big)^2 \le 2^k\gamma\delta_n^2\ ;$$
$$\sum_{i=1}^{n}\mathbb{E}[X_i^2(g)\,|\,\mathscr{F}_{i-1}] > \frac{4(1+M)^3}{(1-\ell)^2}2^k\gamma n\delta_n^2\Big\}.$$

Due to the formulated condition, there exists a number n_0 such that for all $n\in\mathbb{N}$ with $n>n_0$ and all $k\in\mathbb{N}$

$$t_n + n\big(\rho_n+\ell^{t_n}\big)^2 \le \frac{\gamma}{2}n\delta_n^2\,.$$

For any $k\in\mathbb{N}$, all $n\in\mathbb{N}$ with $n>n_0$, and all $g\in\mathscr{G}_n$, we conclude in view of the previous fact and Lemma 4.1.8, that for almost all ω in the sets $\{\omega\in\Omega\colon 2^k\gamma n\delta_n^2 \ge \sum_{i=1}^n\big(\lambda_i-\lambda_i^g\big)^2 > 2^{k-1}\gamma n\delta_n^2\}$ the relation

$$\sum_{i=1}^{n}\mathbb{E}\big[X_i^2(g)\,|\,\mathscr{F}_{i-1}\big] \le \frac{2(1+M)^3}{(1-\ell)^2}\Big[\underbrace{t_n+n\big(\rho_n+\ell^{t_n}\big)^2}_{\le\frac{\gamma}{2}n\delta_n^2} + \underbrace{\sum_{i=1}^{n}\big(\lambda_i-\lambda_i^g\big)^2}_{\ge\frac{\gamma}{2}n\delta_n^2}\Big]$$
$$\le \frac{4(1+M)^3}{(1-\ell)^2}\sum_{i=1}^{n}\big(\lambda_i-\lambda_i^g\big)^2$$
$$\le \frac{4(1+M)^3}{(1-\ell)^2}\,2^k\gamma n\delta_n^2\,.$$

The set $\mathscr{G}_n$ is finite. Therefore,

$$\sup_{n>n_0}\sum_{k=1}^{\infty}\mathbb{P}\Big\{\exists g\in\mathscr{G}_n\colon 2^{k-1}\gamma n\delta_n^2 < \sum_{i=1}^{n}\big(\lambda_i-\lambda_i^g\big)^2 \le 2^k\gamma n\delta_n^2\ ;$$
$$\sum_{i=1}^{n}\mathbb{E}[X_i^2(g)\,|\,\mathscr{F}_{i-1}] > \frac{4(1+M)^3}{(1-\ell)^2}2^k\gamma n\delta_n^2\Big\} = 0\,.$$

The other sum will be treated with an application of Lemma 4.1.9. According to Lemma 4.1.9,

$$\sum_{k=1}^{\infty}\mathbb{P}\Big\{\exists g\in\mathscr{G}_n\colon \sum_{i=1}^{n} X_i^2(g)\mathbb{1}_{\{|X_i^2(g)|>a_{n,k}\}} > \frac{4(1+M)^3}{(1-\ell)^2}2^k\gamma n\delta_n^2\Big\}$$
$$\le \frac{e^M}{n}\sum_{k=1}^{\infty} n^{-k}$$

$$= \frac{e^M}{n}\Big(\frac{n}{n-1}-1\Big),$$

for all $n > e$. Thus, the claim is correct for all $n > \max\{n_0, e\}$ and all $k \in \mathbb{N}$. □

LEMMA 4.1.11. *We suppose again that the condition*

$$\limsup_{n\to\infty} \frac{t_n + n\big(\rho_n + \ell^{t_n}\big)^2}{n\delta_n^2} = 0$$

holds. Then there exists a positive constant C and a number n_0 such that for all $n \in \mathbb{N}$ with $n > n_0$

$$\sum_{k=1}^{\infty} \mathbb{P}\Big\{\exists g \in \mathscr{G}_n : 2^{k-1}\gamma\delta_n^2 < \frac{1}{n}\sum_{i=1}^{n}\big(\lambda_i - \lambda_i^g\big)^2 \le 2^k\gamma\delta_n^2\ ; \\ M_n(g) > 2^{k-2}\frac{\gamma}{2}n\delta_n^2\Big\} \\ \le \frac{e^M}{n} + 4\ \#\mathscr{G}_n\, e^{-C\frac{n\delta_n^2}{\log n}}.$$

Proof. First of all, we use that

$$\Omega = \Big\{H_n^{a_{n,k}}(g) > \frac{8(1+M)^3}{(1-\ell)^2}2^k\gamma n\delta_n^2\Big\} \cup \Big\{H_n^{a_{n,k}}(g) \le \frac{8(1+M)^3}{(1-\ell)^2}2^k\gamma n\delta_n^2\Big\}$$

is a disjoint partition and use the additivity of $\mathbb{P}$. Then, in view of Corollary 4.1.10, there exists a number n_0 such that for all $n > n_0$

$$\sum_{k=1}^{\infty} \mathbb{P}\Big\{\exists g \in \mathscr{G}_n : 2^{k-1}\gamma\delta_n^2 < \frac{1}{n}\sum_{i=1}^{n}\big(\lambda_i - \lambda_i^g\big)^2 \le 2^k\gamma\delta_n^2\ ; \\ M_n(g) > 2^{k-2}\frac{\gamma}{2}n\delta_n^2\Big\}$$

$$= \sum_{k=1}^{\infty} \mathbb{P}\Big\{\exists g \in \mathscr{G}_n : 2^{k-1}\gamma\delta_n^2 < \frac{1}{n}\sum_{i=1}^{n}\big(\lambda_i - \lambda_i^g\big)^2 \le 2^k\gamma\delta_n^2\ ; \\ H_n^{a_{n,k}}(g) > \frac{8(1+M)^3}{(1-\ell)^2}2^k\gamma n\delta_n^2; \\ M_n(g) > 2^{k-2}\frac{\gamma}{2}n\delta_n^2\Big\}$$

$$+ \sum_{k=1}^{\infty} \mathbb{P}\Big\{\exists g \in \mathscr{G}_n : 2^{k-1}\gamma\delta_n^2 < \frac{1}{n}\sum_{i=1}^{n}\big(\lambda_i - \lambda_i^g\big)^2 \le 2^k\gamma\delta_n^2\ ; \\ H_n^{a_{n,k}}(g) \le \frac{8(1+M)^3}{(1-\ell)^2}2^k\gamma n\delta_n^2; \\ M_n(g) > 2^{k-2}\frac{\gamma}{2}n\delta_n^2\Big\}$$

$$\le \sum_{k=1}^{\infty} \mathbb{P}\Big\{\exists g \in \mathscr{G}_n : 2^{k-1}\gamma\delta_n^2 < \frac{1}{n}\sum_{i=1}^{n}\big(\lambda_i - \lambda_i^g\big)^2 \le 2^k\gamma\delta_n^2\ ;$$

$$
\begin{aligned}
&\qquad\qquad H_n^{a_{n,k}}(g) > \frac{8(1+M)^3}{(1-\ell)^2} 2^k \gamma n \delta_n^2 \Big\} \\
&\quad + \sum_{k=1}^{\infty} \mathbb{P}\Big\{ \exists g \in \mathscr{G}_n : M_n(g) > 2^{k-2} \frac{\gamma}{2} n\delta_n^2 \ ; \ H_n^{a_{n,k}}(g) \le \frac{8(1+M)^3}{(1-\ell)^2} 2^k \gamma n \delta_n^2 \Big\} \\
&\le \frac{e^M}{n} \\
&\quad + \sum_{k=1}^{\infty} \mathbb{P}\Big\{ \exists g \in \mathscr{G}_n : M_n(g) > 2^{k-2} \frac{\gamma}{2} n\delta_n^2 \ ; \ H_n^{a_{n,k}}(g) \le \frac{8(1+M)^3}{(1-\ell)^2} 2^k \gamma n \delta_n^2 \Big\}.
\end{aligned}
$$

The probabilities in the last line will be estimated using the exponential tail bound for martingales by Dzhaparidze and van Zanten (2001). We have stated their theorem in the Appendix in Lemma A.4.7. We introduce the abbreviations $C_1 := 8(1+M)^3/(1-\ell)^2$ and $C_2 := 2(M^2+2M)$. Then, for

$$
\begin{aligned}
a_{n,k} &= C_2 \ (k+1) \log n \,, \\
z_{n,k} &= 2^{k-2} \frac{\gamma}{2} \, n \, \delta_n^2 \,, \\
L_{n,k} &= C_1 \ 2^k \gamma \, n \, \delta_n^2 \,,
\end{aligned}
$$

we obtain according to Lemma A.4.7

$$
\begin{aligned}
&\sum_{k=1}^{\infty} \mathbb{P}\Big\{ \exists g \in \mathscr{G}_n : M_n(g) > 2^{k-2} \frac{\gamma}{2} n\delta_n^2 \ ; \ H_n^{a_{n,k}}(g) \le \frac{8(1+M)^3}{(1-\ell)^2} 2^k \gamma n \delta_n^2 \Big\} \\
&= \sum_{k=1}^{\infty} \mathbb{P}\Big\{ \exists g \in \mathscr{G}_n : M_n(g) > z_{n,k} \ ; \ H_n^{a_{n,k}}(g) \le L_{n,k} \Big\} \\
&\le \sum_{k=1}^{\infty} \sum_{g \in \mathscr{G}_n} \mathbb{P}\Big\{ M_n(g) > z_{n,k} \ ; \ H_n^{a_{n,k}}(g) \le L_{n,k} \Big\} \\
&\le \sum_{k=1}^{\infty} \sum_{g \in \mathscr{G}_n} 2 \exp\Big(- \frac{1}{2} \frac{z_{n,k}^2}{L_{n,k}} \ \psi\Big(\frac{a_{n,k} \, z_{n,k}}{L_{n,k}} \Big) \Big) \\
&= \#\mathscr{G}_n \sum_{k=1}^{\infty} 2 \exp\Big(- \frac{1}{2} \frac{2^{2k-4} \big(\frac{\gamma}{2}\big)^2 (n\delta_n^2)^2}{C_1 2^k \gamma n \delta_n^2} \ \psi\Big(\frac{C_2 (k+1) \log(n) 2^{k-2} \frac{\gamma}{2} n\delta_n^2}{C_1 2^k \gamma n \delta_n^2} \Big) \Big) \\
&= 2 \, \#\mathscr{G}_n \sum_{k=1}^{\infty} \exp\Big(- \frac{\gamma}{2^8 C_1} \, n\delta_n^2 \, 2^{k+1} \ \psi\Big(\frac{C_2}{8 C_1} (k+1) \log n \Big) \Big).
\end{aligned}
$$

We use the fact that $\psi(x) \ge 1/(1+\frac{x}{3})$ for $x \ge -1$. Choose a number $n_1 \ge e^{24 C_2/C_1}$. For all $n \in \mathbb{N}$ with $n > n_1$

$$
\frac{C_2}{24 C_1} \log n \ge 1 \,,
$$

and we conclude that

$$
\psi\Big(\frac{C_2}{8 C_1} (k+1) \log n \Big) \ge \frac{1}{1 + \frac{C_2}{24 C_1} (k+1) \log n} \ge \frac{12 C_1}{C_2 (k+1) \log n}
$$

for all $n > n_1$ and all $k \in \mathbb{N}$. Therefore, using the notation $C := 2^{-6}\gamma/C_2$ and the fact that $2^k \geq k^2/3$ for all k, we obtain

$$\begin{aligned}
&2\#\mathscr{G}_n \sum_{k=1}^{\infty} \exp\Big(-\frac{\gamma}{2^8 C_1} 2^{k+1} n\delta_n^2 \; \psi\big(\frac{C_2}{8C_1}(k+1)\log n\big)\Big) \\
&\leq 2\#\mathscr{G}_n \sum_{k=1}^{\infty} \exp\Big(-\frac{12\gamma}{2^8 C_2} \frac{n\delta_n^2}{\log n} \frac{2^{k+1}}{(k+1)}\Big) \\
&\leq 2\#\mathscr{G}_n \sum_{k=1}^{\infty} \exp\Big(-C\,\frac{n\delta_n^2}{\log n}(k+1)\Big) \\
&< 2\#\mathscr{G}_n \frac{e^{-C\frac{n\delta_n^2}{\log n}}}{1-e^{-C\frac{n\delta_n^2}{\log n}}}
\end{aligned}$$

for all $n > \max\{n_0, n_1\}$. Furthermore, $\liminf_{n\to\infty}\Big(1-e^{-C\frac{n\delta_n^2}{\log n}}\Big) > \frac{1}{2}$, which means that there exists a number n_2 such that for all $n \in \mathbb{N}$ with $n > n_2$

$$2\#\mathscr{G}_n \frac{e^{-C\frac{n\delta_n^2}{\log n}}}{1-e^{-C\frac{n\delta_n^2}{\log n}}} \leq 4\#\mathscr{G}_n\, e^{-C\frac{n\delta_n^2}{\log n}}.$$

In summary, we have proven that

$$\begin{aligned}
&\sum_{k=1}^{\infty} \mathbb{P}\Big\{\exists g \in \mathscr{G}_n : 2^{k-1}\gamma\delta_n^2 < \frac{1}{n}\sum_{i=1}^{n}\big(\lambda_i - \lambda_i^g\big)^2 \leq 2^k\gamma\delta_n^2 \;; \\
&\qquad\qquad M_n(g) > 2^{k-2}\frac{\gamma}{2} n\delta_n^2\Big\} \\
&\leq \frac{e^M}{n} + 4\;\#\mathscr{G}_n\, e^{-C\frac{n\delta_n^2}{\log n}}
\end{aligned}$$

for all $n > \max\{n_0, n_1, n_2\}$. □

We are now in the position to state and prove the main theorem about the rate of convergence of the approximate least squares estimator.

THEOREM 4.1.12. *In the setting of Definition 4.1.1, we have the following result. Assume that $c_\rho \cdot n^{-1/3} \leq \rho_n \leq C_\rho \cdot n^{-1/3}$ for some constants $c_\rho, C_\rho > 0$. Then there exist finite subsets $\{\mathscr{G}_n \subset \mathscr{G}(M, B_n, L) \colon n \in \mathbb{N},\ n \geq 2\}$ such that $\min_{g_n \in \mathscr{G}_n} \|m - g_n\| = \rho_n$ and $\#\mathscr{G}_n \leq e^{\kappa n^{1/3}\log n}$ for some positive constant κ and all but finitely many $n \in \mathbb{N}$. Suppose that $\{\mathscr{G}_n\}$ is any such sequence of subsets. Let $\hat{m}_n$ denote the approximate least squares estimator of m on the basis of the approximating set $\mathscr{G}_n$. With respect to the empirical mean square error, the sequence of estimators $\{\hat{m}_n\}$ is consistent with rate $n^{-1/3}\log n$. In other words, for any $\varepsilon > 0$ there exists a constant $K(\varepsilon)$ such that*

$$\limsup_{n\to\infty} \mathbb{P}\Big\{\frac{1}{n}\sum_{i=2}^{n}\big(m(\lambda_{i-1}, Y_{i-1}) - \hat{m}_n(\lambda_{i-1}, Y_{i-1})\big)^2 > K \cdot n^{-2/3}(\log n)^2\Big\} < \varepsilon.$$

Proof. We have seen in the proof of Lemma 3.1.5 that we can find a set $\mathscr{G}_n$ containing at most $e^{2MB_n(L\vee 1)/\rho_n} \le e^{\frac{2MB_0(L\vee 1)}{c_\rho} n^{1/3}\log n}$ elements such that the balls $\{B_\infty(g,\rho)\colon g\in\mathscr{G}_n\}$ cover $\mathscr{G}(M,B_n,L)$. Let $n^* := \max\{n\in\mathbb{N}\colon B_n \le B\}$. If $n > n^*$, the true link function m is contained in $\mathscr{G}(M,B_n,L)$. Thus, there exists a constant $\kappa > 0$ such that for any $n > n^*$ there exists a set $\mathscr{G}_n$ with $\#\mathscr{G}_n \le e^{\kappa\cdot n^{1/3}\log n}$ such that the approximation condition $\min_{g_n\in\mathscr{G}_n}\|m-g_n\|_\infty = \rho_n$ is satisfied.

Let $\varepsilon > 0$ be arbitrary. We define $\delta_n := \sqrt{K}\, n^{-1/3}\log n$, where the constant K is chosen according to an inequality that will appear later in this proof. Choosing $\{t_n\}$ such that $t_n = o(n\delta_n^2)$ and $\ell^{t_n} = o(\delta_n^2)$ (e.g. $t_n = -\frac{2}{3\log\ell}\log n$), we see that the condition

$$\limsup_{n\to\infty} \frac{t_n + n\left(\rho_n + \ell^{t_n}\right)^2}{n\delta_n^2} = 0$$

is satisfied. Therefore, we may apply Lemma 4.1.4 and plug in the abbreviation

$$M_n(g) = \sum_{i=1}^n (Y_i - \lambda_i)\left[(\lambda_i - \lambda_i^{m_n}) - (\lambda_i - \lambda_i^g)\right].$$

Subsequently we apply Lemma 4.1.11. Taken these two steps together, we conclude that there exist a constant $C > 0$ and a number n_0 such that for all $n\in\mathbb{N}$ with $n > n_0$

$$\begin{aligned}
&\mathbb{P}\left\{\frac{1}{n}\sum_{i=2}^n \left(m(\lambda_{i-1},Y_{i-1}) - \hat{m}_n(\lambda_{i-1},Y_{i-1})\right)^2 > \delta_n^2\right\} \\
&\le \sum_{k=1}^\infty \mathbb{P}\Bigg\{\exists g\in\mathscr{G}_n\colon 2^{k-1}\gamma\delta_n^2 < \frac{1}{n}\sum_{i=1}^n \left(\lambda_i - \lambda_i^g\right)^2 \le 2^k\gamma\delta_n^2\ ; \\
&\qquad\qquad M_n(g) > 2^{k-2}\frac{\gamma}{2}n\delta_n^2\Bigg\} \\
&\le \frac{e^M}{n} + 4\ \#\mathscr{G}_n e^{-C\frac{n\delta_n^2}{\log n}}.
\end{aligned}$$

We choose the constant $K = K(C,\kappa)$ such that $(\kappa - CK) < -1$. There exists a number $n_1(\varepsilon) > n^*$ such that for all $n > n_1$

$$\begin{aligned}
\#\mathscr{G}_n\, e^{-C\frac{n\delta_n^2}{\log n}} &\le \exp\left(\kappa\, n^{1/3}\log(n) - CK\, n^{1/3}\log n\right) \\
&= \exp\left(n^{1/3}\log(n)\,(\kappa - CK)\right) \\
&< \varepsilon.
\end{aligned}$$

Thus, for $n > \max\{n_0, n_1\}$ and the choice $K = K(\kappa,C)$, the assertion is true. In fact, K is independent of ε, which means that we have proven a stronger statement than necessary. □

4.1.3 Discussion: extending the result to the $L_2(\pi)$ risk

We have proven that the empirical MSE of the approximate least squares estimator converges in probability to zero with the rate $n^{-2/3}(\log n)^2$. Having established this result, we are interested in the question whether there is a way to deduce a similar behavior of the $L_2(\pi)$ risk of the estimator. One possible way to answer this question is to examine the difference between empirical MSE and $L_2(\pi)$-risk,

$$\Big[\mathbb{E}_{|\hat{m}_n[Y_0,\ldots,Y_n]=g}\big(m(\lambda_0',Y_0')-g(\lambda_0',Y_0')\big)^2 \\ -\frac{1}{n}\sum_{i=2}^{n}\big(m(\lambda_{i-1},Y_{i-1})-\hat{m}_n[Y_0,\ldots,Y_n](\lambda_{i-1},Y_{i-1})\big)^2\Big].$$

In order to deal with the randomness of $\hat{m}_n$, we bound this difference by a maximum over the class $\mathscr{G}_n$,

$$\max_{g\in\mathscr{G}_n}\Big[\mathbb{E}\big(m(\lambda_0,Y_0)-g(\lambda_{i-1},Y_{i-1})\big)^2-\frac{1}{n}\sum_{i=2}^{n}\big(m(\lambda_{i-1},Y_{i-1})-g(\lambda_{i-1},Y_{i-1})\big)^2\Big].$$

What are the chances to bound this quantity in probability? In view of the size of the class $\mathscr{G}_n$, we need an exponential tail bound for the differences

$$\Big[\mathbb{E}\big(m(\lambda_0,Y_0)-g(\lambda_0,Y_0)\big)^2-\frac{1}{n}\sum_{i=2}^{n}\big(m(\lambda_{i-1},Y_{i-1})-g(\lambda_{i-1},Y_{i-1})\big)^2\Big]$$

for every g. Our favorite tools, exponential tail bounds for martingales or mixing sequences, are unfortunately not applicable because neither is the above sum a martingale in n nor is the bivariate INGARCH(1,1) process mixing. However, there are results available that assure absolute regularity of the lagged count process $\{\boldsymbol{Y}_{n-t}^n\colon n\in\mathbb{Z}\}$, with $\boldsymbol{Y}_{n-t}^n:=(0,Y_{n-t},\ldots,Y_n)$, under the conditions of Definition 4.1.1. The notion of absolute regularity was introduced by Volkonskii and Rozanov (1959) who attributed it to Kolmogorov (Bradley, 2007, page 66). We use the characterization that we found in Bradley (2007, page 89).

DEFINITION 4.1.13. Let $\mathscr{A}$ and $\mathscr{E}$ be sub σ-fields of $\mathscr{F}$. Suppose that $\mathscr{E}$ is separable and that there exists a regular conditional distribution $\mathbb{P}(\cdot\,|\,\mathscr{A})$ on $\mathscr{E}$. The coefficient of absolute regularity between $\mathscr{A}$ and $\mathscr{E}$ is given by

$$\beta(\mathscr{A},\mathscr{E}):=\mathbb{E}\big[\sup\big\{|\mathbb{P}(B\,|\,\mathscr{A})-\mathbb{P}(B)|\colon B\in\mathscr{E}\big\}\big].$$

For integers $n\in\mathbb{Z}$ and $k\in\mathbb{N}$, the absolute regularity coefficients of the count process $\{Y_t\}_{t\in\mathbb{Z}}$ are given by

$$\beta(k,n):=\beta(\mathscr{F}_{-\infty,n}^{Y},\mathscr{F}_{n+k,\infty}^{Y}),$$

$$\beta(k) := \sup_{n\in\mathbb{Z}} \beta(k,n),$$

where $\mathscr{F}^Y_{-\infty,n} = \sigma\{Y_t\colon t \le n\}$ and $\mathscr{F}^Y_{n+k,\infty} = \sigma\{Y_t\colon t \ge n+k\}$. The process $\{Y_t\}$ is called absolutely regular if $\lim_{k\to\infty}\beta(k) = 0$. The absolute regularity coefficients of the lagged count process $\{\boldsymbol{Y}^n_{n-t}\colon n\in\mathbb{Z}\}$ are denoted

$$\begin{aligned}\beta^t(k,n) &:= \beta\big(\sigma\{\boldsymbol{Y}^i_{i-t}\colon i\le n\}, \sigma\{\boldsymbol{Y}^i_{i-t}\colon i\ge n+k\}\big),\\ \beta^t(k) &:= \sup_{n\in\mathbb{Z}}\beta^t(k,n).\end{aligned}$$

Since all involved σ-fields are separable, the coefficients of absolute regularity of the count process and the lagged count process, respectively, are well defined. The main result of Doukhan and Neumann (2018) suggest that under the semi-contractive condition the lagged count process is absolute regular with mixing coefficients $\beta^t(k) \lesssim \varrho^{\sqrt{k-t}}$ for some $0<\varrho<1$. This can be seen by combining the proof of our Lemma 2.2.9 with the proof of Theorem 2.1 in the paper by Doukhan and Neumann (2018) and the statement of their Proposition 2.1 (ibid.). We can therefore use the Bernstein inequality for absolute regular processes (Doukhan, 1994, page 36) and obtain the necessary exponential bound to establish a bound of

$$\begin{aligned}\sup_{g\in\mathscr{G}_n}\mathbb{E}\Big[m^{[t_n]}(0,Y_0,\dots,Y_{t_n}) - g^{[t_n]}(0,Y_0,\dots,Y_{t_n})\Big]^2 \\ -\frac{1}{n-t_n}\sum_{i=t_n}^{n-1}\big(m^{[t_n]}(0,Y_{i-t_n},\dots,Y_i) - g^{[t_n]}(0,Y_{i-t_n},\dots,Y_i)\big)^2.\end{aligned} \tag{4.3}$$

The result of the next lemma will come into effect after we apply a peeling argument to the expression in (4.3), which we postpone to Corollary 4.1.15.

LEMMA 4.1.14. *Let $t = t_n \asymp \log n$. Let $\{(\lambda_t, Y_t)\}_{t\in\mathbb{Z}}$ be the data generating process from Definition 4.1.1. The process $\{\boldsymbol{Y}^n_{n-t}\colon n\in\mathbb{Z}\}$ is absolutely regular with coefficients $\beta^t(k)\le\beta_0\,\varrho^{\sqrt{k-t}}$ for some positive constants $\varrho<1$ and $\beta_0<\infty$. Let $\mathscr{G}_n$ be a subset of $\mathscr{G}(M,B_n,L)$ with the same properties specified in Theorem 4.1.12. There exists a number n_0 and positive constants C_1,C_2,C_3 such that for $\delta_n := C_1 n^{-1/3}(\log n)^{3/2}$ the following inequality holds for all $k\in\mathbb{N}_+$:*

$$\begin{aligned}\sup_{n>n_0}\mathbb{P}\Bigg\{ \max_{\substack{g\in\mathscr{G}_n:\\ \mathbb{E}\left[m^{[t]}(\boldsymbol{Y}^t_0)-g^{[t]}(\boldsymbol{Y}^t_0)\right]^2\le 2^{k+1}\delta_n^2}} \mathbb{E}\Big[\big(m^{[t]}(\boldsymbol{Y}^t_0) - g^{[t]}(\boldsymbol{Y}^t_0)\big)^2\Big] \\ -\frac{1}{n-t}\sum_{i=t}^{n-1}\big[m^{[t]}(\boldsymbol{Y}^i_{i-t}) - g^{[t]}(\boldsymbol{Y}^i_{i-t})\big]^2 > 2^{k-1}\delta_n^2\Bigg\} \\ \le C_2\Big(n^{-2k}(\log n)^{-2} + e^{-C_3\,\frac{2^k}{k^2}n^{1/3}\log n}\Big).\end{aligned}$$

Proof. For the mixing property of the stationary two-sided count process under the semi-contractive condition, we refer to Doukhan and Neumann (2018, Theorem 2.1). Similarly to the argument in Lemma 2.2.9, the coupling result of Doukhan and Neumann (2018) can be extended to the lagged count process. Thence we obtain the bounds for the absolute regularity coefficients of the lagged count process.

Let $g \in \mathscr{G}_n$ be chosen such that $\mathbb{E}\big[m^{[t]}(\boldsymbol{Y}_0^t) - g^{[t]}(\boldsymbol{Y}_0^t)\big]^2 \leq 2^{k+1}\delta_n^2$. We define $X_i(g)$ by

$$\begin{aligned} -X_i(g) &:= \Big(\big[m^{[t]}(\boldsymbol{Y}_{i-t}^i) - g^{[t]}(\boldsymbol{Y}_{i-t}^i)\big]^2 - \mathbb{E}\big[m^{[t]}(\boldsymbol{Y}_0^t) - g^{[t]}(\boldsymbol{Y}_0^t)\big]^2\Big) \\ &= \Big(\big[m^{[t]}(\boldsymbol{Y}_{i-t}^i) - g^{[t]}(\boldsymbol{Y}_{i-t}^i)\big]^2 - \mathbb{E}\big[m^{[t]}(\boldsymbol{Y}_{i-t}^i) - g^{[t]}(\boldsymbol{Y}_{i-t}^i)\big]^2\Big). \end{aligned}$$

Note that the process $\{X_i(g)\colon i \in \mathbb{Z}\}$ is stationary. We want to adapt the proof of Bernstein's inequality for absolute regular processes (Doukhan, 1994, page 36) to bound $\frac{1}{n-t}\sum_{i=t}^{n-1} X_i(g)$ in probability. For $k \in \mathbb{N}_+$, we introduce the variables

$$q_k = q_{k,n} = \left\lceil t + \left(\frac{(2k+1)\log n}{|\log \varrho|}\right)^2 \right\rceil. \tag{4.4}$$

Note that there exists a constant $c_t > 0$ and a number $n^{(t)} \in \mathbb{N}_+$ such that $t \leq c_t \log n$ for $n > n^{(t)}$. Thus,

$$\underbrace{\frac{1}{|\log \varrho|^2}}_{=:c_q}(k\log n)^2 \leq q_k \leq \underbrace{\big[c_t + \frac{9}{|\log \varrho|^2} + 1\big]}_{=:C_q}(k\log n)^2$$

for all natural numbers $n > \max\{n^{(t)}, e\}$ and all $k \in \mathbb{N}_+$. Furthermore, q_k satisfies $\varrho^{\sqrt{q_k - t}} \leq n^{-(2k+1)}$ for $n \in \mathbb{N}_+$. The variable q_k will determine the length of the blocks in the following coupling argument. It should be emphasized that here, in contrast to the proof of Theorem 3.2.2, the block length depends on an additional index k. The reason for this difference will become apparent after the proof of this lemma.

As in Chapter 3, we invoke Berbee's coupling to obtain i.i.d. sequences of blocks of length q_k. The technical argument is identical to the one that we delivered in Lemma 3.2.9: there exists a coupling $(S, \Sigma, P, (V', V^*))$, where $V' = \{V_i'\}_{i\in\mathbb{Z}}$ and $\{V_i^*\}_{i\in\mathbb{Z}}$ are vector-valued processes with the following properties.

(1) The processes $\{V_i'\}$ and $\{\boldsymbol{Y}_{i-t}^i\}$ are identically distributed.
(2) For each index $j \in \mathbb{N}$, the distributions of the blocks $\big(V'_{t+jq_k}, \ldots, V'_{t+jq_k+q_k-1}\big)$ and $\big(V^*_{t+jq_k}, \ldots, V^*_{t+jq_k+q_k-1}\big)$ are identical.
(3) The blocks $\big\{\big(V^*_{t+jq_k}, \ldots, V^*_{t+jq_k+q_k-1}\big)\colon j \text{ is even}\big\}$ form a sequence of i.i.d. variables, and so do the blocks $\big\{\big(V^*_{t+jq_k}, \ldots, V^*_{t+jq_k+q_k-1}\big)\colon j \text{ is uneven}\big\}$.

(4) For any $j \in \mathbb{N}$,

$$P\Big\{\big(V^*_{t+jq_k},\dots,V^*_{t+jq_k+q_k-1}\big) \neq \big(V'_{t+jq_k+1},\dots,V'_{t+jq_k+q_k}\big) \text{ for some } 0 \le j < \frac{n-t}{q_k}\Big\}$$
$$\le \frac{n-t}{q_k}\beta^t(q_k).$$

We introduce the variables $X^*_i(g)$ that are defined canonically to $X_i(g)$ by

$$-X^*_i(g) := \Big(\big[m^{[t]}(V^*_i) - g^{[t]}(V^*_i)\big]^2 - E\big[m^{[t]}(V^*_i) - g^{[t]}(V^*_i)\big]^2\Big).$$

Note that the new sequence $\{X^*_i(g)\}$ satisfies for all $j \in \mathbb{N}$ the relation

$$P^{\left(X^*_{t+2jq_k+1}(g),\dots,X^*_{t+2j+q_k}(g)\right)} = \mathbb{P}^{\left(X_{t+2jq_k+1}(g),\dots,X_{t+2j+q_k}(g)\right)} = \mathbb{P}^{\left(X_{t+1}(g),\dots,X_{t+q_k}(g)\right)}.$$

We have the estimate

$$\mathbb{P}\Big\{\max_{\substack{g\in\mathscr{G}_n:\\ \mathbb{E}\left[m^{[t]}(\boldsymbol{Y}^t_0)-g^{[t]}(\boldsymbol{Y}^t_0)\right]^2\le 2^{k+1}\delta_n^2}} \frac{1}{n-t}\sum_{i=t}^{n-1} X_i(g) > 2^{k-1}\delta_n^2\Big\}$$
$$\le \frac{n-t}{q_k}\beta_t(q_k) + P\Big\{\max_{\substack{g\in\mathscr{G}_n:\\ \mathbb{E}\left[m^{[t]}(\boldsymbol{Y}^t_0)-g^{[t]}(\boldsymbol{Y}^t_0)\right]^2\le 2^{k+1}\delta_n^2}} \frac{1}{n-t}\sum_{i=t}^{n-1} X^*_i(g) > 2^{k-1}\delta_n^2\Big\}$$
$$\le \frac{n-t}{q_k}\beta_t(q_k) + \sum_{\substack{g\in\mathscr{G}_n:\\ \mathbb{E}\left[m^{[t]}(\boldsymbol{Y}^t_0)-g^{[t]}(\boldsymbol{Y}^t_0)\right]^2\le 2^{k+1}\delta_n^2}} P\Big\{\frac{1}{n-t}\sum_{i=t}^{n-1} X^*_i(g) > 2^{k-1}\delta_n^2\Big\}.$$

Applying the same strategy as in the computations leading to (3.24), we arrange the sum $\frac{1}{n-t}\sum_{i=t}^{n-1} X^*_i(g)$ into sums of even and uneven blocks, respectively, and a remainder term. Defining $N_k := \frac{1}{2}\left\lfloor \frac{n-t}{q_k}\right\rfloor$ if $\left\lfloor \frac{n-t}{q_k}\right\rfloor$ is even, or $N_k := \frac{1}{2}\left(\left\lfloor \frac{n-t}{q_k}\right\rfloor - 1\right)$ if $\left\lfloor \frac{n-t}{q_k}\right\rfloor$ is odd, as in the paragraph preceding (3.24), we have

$$\frac{1}{n-t}\sum_{i=t}^{n-1} X^*_i(g) = \frac{1}{n-t}\sum_{j=0}^{N_k-1}\big(X^*_{t+(2j+1)q} + \dots + X^*_{t+(2j+1)q_k+q_k-1}\big)$$
$$+\frac{1}{n-t}\sum_{j=0}^{N_k-1}\big(X^*_{t+2jq_k} + \dots + X^*_{t+2jq_k+q_k-1}\big)$$
$$+\frac{1}{n-t}\sum_{i=2N_k}^{n-1} X^*_{t+i}(g).$$

The sum in the remainder term

$$R_{n,k} := \frac{1}{n-t}\sum_{i=2N_k}^{n-1} X^*_{t+i}(g)$$

contains no more than $2q_k$ addends, which are bounded by M^2. The term $R_{n,k}$ is therefore bounded by $2M^2\frac{q_k}{n-t} \leq 2M^2\frac{C_q(k\log n)^2}{n-t}$ for $n > n^{(q)}$, which means that

$$\limsup_{n\to\infty}\sup_k \frac{R_{n,k}}{2^k\delta_n^2} \leq \limsup_{n\to\infty}\sup_k 2M^2C_q\,\frac{k^2}{2^k}\frac{(\log n)^2}{(n-t)\delta_n^2} = 0\,.$$

Thus, there exists a number $n_0 \in \mathbb{N}$ such that for all $n > n_0$

$$\sup_k P\left\{R_{n,k}(g) > 2^{k-2}\delta_n^2\right\} = 0\,.$$

And we obtain in resemblance to the approach in Lemma 3.2.12 that

$$\begin{aligned} &P\left\{\frac{1}{n-t}\sum_{i=t}^{n-1} X_i^*(g) > 2^{k-1}\delta_n^2\right\} \\ &\qquad \leq 4P\left\{\frac{1}{n-t}\sum_{j=0}^{N_k-1}\left(X^*_{t+2jq_k}(g)+\ldots+X^*_{t+2jq_k+q_k-1}(g)\right) > 2^{k-3}\delta_n^2\right\} \end{aligned}$$

for all $k \in \mathbb{N}$ and all $n > n_0$. In order to apply the Bernstein inequality for sums of independent and bounded random variables, we check the following conditions.

(1) $E\,X_i^*(g) = 0$ for all natural numbers $i \geq t$;
(2) there exists a $\sigma_{n,k}^2 > 0$ such that for all $j \in \mathbb{N}$

$$E\left[X^*_{t+2jq_k}(g)+\ldots+X^*_{t+2jq_k+q_k-1}(g)\right]^2 \leq \sigma_{n,k}^2;$$

(3) there exists a $b_{n,k}$ such that for all $j \in \mathbb{N}$

$$\left|X^*_{t+2jq_k}(g)+\ldots+X^*_{t+2jq_k+q_k-1}(g)\right| \leq b_{n,k}\,.$$

Condition (1) is obviously satisfied, as is condition (3). As a bounding constant we can choose $b_{n,k} = q_k M^2$. To verify condition (2), we apply the triangle inequality and obtain

$$\begin{aligned} \left(E[X^*_{t+2jq_k}(g)+\ldots+X^*_{t+2jq_k+q_k-1}(g)]^2\right)^{1/2} &= \left(\mathbb{E}[X_t(g)+\ldots+X_{t+q_k-1}(g)]^2\right)^{1/2} \\ &\leq \left(\mathbb{E}[X_t^2(g)]\right)^{1/2}+\ldots+\left(\mathbb{E}[X^2_{t+q_k-1}(g)]\right)^{1/2} \\ &= q_k\left(\mathbb{E}[X_t^2(g)]\right)^{1/2}. \end{aligned}$$

Furthermore,

$$\begin{aligned} \mathbb{E}X_t^2(g) &= \mathbb{E}\left[m^{[t]}(\boldsymbol{Y}_0^t) - g^{[t]}(\boldsymbol{Y}_0^t)\right]^4 - \left(\mathbb{E}\left[m^{[t]}(\boldsymbol{Y}_0^t) - g^{[t]}(\boldsymbol{Y}_0^t)\right]^2\right)^2 \\ &\leq M^2\mathbb{E}\left[m^{[t]}(\boldsymbol{Y}_0^t) - g^{[t]}(\boldsymbol{Y}_0^t)\right]^2 \\ &\leq M^2\,2^{k+1}\delta_n^2\,, \end{aligned}$$

where we used in the last line that g is chosen from the set

$$\left\{g \in \mathscr{G}_n : \mathbb{E}\big[m^{[t]}(\boldsymbol{Y}_0^t) - g^{[t]}(\boldsymbol{Y}_0^t)\big]^2 \le 2^{k+1}\delta_n^2\right\}.$$

Therefore,

$$\sup_{j\in\mathbb{N}} E[X^*_{t+2jq_k}(g) + \ldots + X^*_{t+2jq_k+q_k-1}(g)]^2 \le M^2 q_k^2 2^{k+1}\delta_n^2 =: \sigma^2_{n,k}.$$

For illustrative purposes, we define for $g \in \mathscr{G}_n$ the variables

$$\begin{aligned}\eta_{j,k}(g) &:= X^*_{t+2jq_k+1}(g) + \ldots + X^*_{t+2jq_l+q_k}(g)\\ x_{k,n} &:= (n-t)2^{k-3}\delta_n^2\\ v_{n,k} &:= N_k\,\sigma^2_{n,k}.\end{aligned}$$

The random variables $\eta_{0,k}(g), \ldots, \eta_{N_k-1,k}(g)$ are independent, centered, and bounded. Thus, we can apply Bernstein's inequality in the version of Lemma A.4.3. For any $g \in \mathscr{G}_n$

$$\begin{aligned}&P\left\{\frac{1}{n-t}\sum_{j=0}^{N_k-1}\left(X^*_{t+2jq_k}(g) + \ldots + X^*_{t+2jq_k+q_k-1}(g)\right) > 2^{k-3}\delta_n^2\right\}\\ &= P\left\{\sum_{j=0}^{N_k-1}\eta_{j,k}(g) > x_{n,k}\right\}\\ &\le \exp\left(-\frac{1}{2}\frac{x^2_{n,k}}{v_{n,k} + b_{n,k}\,x_{n,k}/3}\right).\end{aligned}$$

From the definition of N_k, it is apparent that N_k is a uniform (in k) asymptotic (as $n \to \infty$) upper bound for $(n-t)/q_k$. To make this argument explicit, we observe

$$\sup_k\left|\frac{(n-t)/q_k}{N_k}\right| \le \sup_k\left|\frac{(n-t)/q_k}{\frac{1}{2}\left(\frac{n-t}{q_k} - 2\right)}\right| \le \sup_k \frac{1}{\left|\frac{1}{2} - \frac{q_k}{n-t}\right|}.$$

Recall that $c_q(k\log n)^2 \le q_k \le C_q(k\log n)^2$, which means that there exists a $k_0 \in \mathbb{N}_+$ such that

$$\limsup_{n\to\infty}\sup_k \frac{(n-t)/q_k}{N_k} \le \limsup_{n\to\infty}\frac{1}{\left|\frac{1}{2} - \frac{q_{k_0}}{n-t}\right|} < \infty.$$

We conclude that $x_{n,k}\,b_{n,k}$ is asymptotically bounded from above by $v_{n,k}$, uniformly in k: there exists a constant $c_1 > 0$ and a number $n_1 \in \mathbb{N}$ such that for all $n > n_1$

and all k

$$\frac{x_{n,k}b_{n,k}}{v_{n,k}} = \frac{2^{k-3}(n-t)\delta_n^2 q_k M^2}{N_k q_k^2 M^2 2^{k+1}\delta_n^2} = \underbrace{\frac{\frac{n-t}{q_k}}{N_k}}_{=O(1)} 2^{-4} \leq 3\,c_1\,.$$

Recall that $q_k \leq C_q(k\log n)^2$ for $n > n^{(q)}$, and that $t \leq c_t \log n$ for $n > n^{(t)}$. We conclude for the whole exponent that there exists a positive constant c_2 and number $n_2 \in \mathbb{N}$ such that for all $n > n_2$ and all $k \in \mathbb{N}_+$

$$\begin{aligned}
\frac{1}{2}\frac{x_{n,k}^2}{v_{n,k}+x_{n,k}b_{n,k}/3} &\geq \frac{x_{n,k}^2}{2(1+c_1)v_{n,k}} \\
&= \frac{1}{2(1+c_1)}\frac{(n-t)^2}{N_k q_k^2}\,\frac{2^{2k-6}\delta_n^4}{M^2 2^{k+1}\delta_n^2} \\
&= \frac{2^{-6}}{4(1+c_1)M^2}\frac{(n-t)/q_k}{N_k}\frac{(n-t)}{q_k}\delta_n^2 2^k \\
&\geq c_2\,\frac{2^k}{k^2}\,\frac{n}{(\log n)^2}\delta_n^2\,.
\end{aligned}$$

Therefore, for all $n > n_2$ and all $k \in \mathbb{N}_+$ we have the bound

$$\begin{aligned}
&P\Big\{\frac{1}{n-t}\sum_{j=0}^{N_k-1}\big(X^*_{t+2jq_k+1}(g)+\ldots+X^*_{t+2jq_k+q_k}(g)\big) > 2^{k-3}\delta_n^2\Big\} \\
&\leq \exp\big(-c_2\,\frac{2^k}{k^2}\,\frac{n}{(\log n)^2}\delta_n^2\big)\,.
\end{aligned}$$

According to Theorem 4.1.12, we can assume without loss of generality that $\#\mathscr{G}_n \leq e^{\kappa n^{1/3}\log n}$ for all $n > n_2$. Furthermore, $2^k > k^2/3$ for all $k \geq 0$. Hence, we obtain for all $n > n_2$ and all $k \in \mathbb{N}_+$

$$\begin{aligned}
&\mathbb{P}\Big\{\max_{\substack{g\in\mathscr{G}_n:\\ \mathbb{E}\left[m^{[t]}(\boldsymbol{Y}^i_{i-t})-g^{[t]}(\boldsymbol{Y}^i_{i-t})\right]^2\leq 2^{k+1}\delta_n^2}} \frac{1}{n-t}\sum_{i=t}^{n-1} X_i(g) > 2^{k-1}\delta_n\Big\} \\
&\leq \frac{n-t}{q_k}\beta^t(q_k) + \sum_{\substack{g\in\mathscr{G}_n:\\ \mathbb{E}\left[m^{[t]}(\boldsymbol{Y}^i_{i-t})-g^{[t]}(\boldsymbol{Y}^i_{i-t})\right]^2\leq 2^{k+1}\delta_n^2}} 4\,P\Big\{\sum_{j=0}^{N_q-1}\eta_{j,k}(g) > x_{k,n}\Big\} \\
&\leq \frac{n-t}{q_k}\beta^t(q_k) + 4\,\#\mathscr{G}_n \exp\big(-c_2\,\frac{2^k}{k^2}\,\frac{n}{(\log n)^2}\delta_n^2\big) \\
&\leq \frac{n-t}{q_k}\beta^t(q_k) + 4\,\exp\Big(\kappa n^{1/3}\log n - c_2\,\frac{2^k}{k^2}\,\frac{n}{(\log n)^2}\delta_n^2\Big) \\
&< \frac{n-t}{q_k}\beta^t(q_k) + 4\,\exp\Big(3\kappa\frac{2^k}{k^2}\,n^{1/3}\log n - c_2\,\frac{2^k}{k^2}\,\frac{n}{(\log n)^2}\delta_n^2\Big)\,.
\end{aligned}$$

With $\delta_n^2 := \frac{6\kappa}{c_2} n^{-2/3} (\log n)^3$ we obtain

$$\mathbb{P}\Big\{ \max_{\substack{g\in\mathscr{G}_n:\\ \mathbb{E}\left[m^{[t]}(\boldsymbol{Y}_{i-t}^i)-g^{[t]}(\boldsymbol{Y}_{i-t}^i)\right]^2 \le 2^{k+1}\delta_n^2}} \frac{1}{n-t}\sum_{i=t}^{n-1} X_i(g) > 2^{k-1}\delta_n \Big\}$$
$$\le \frac{n-t}{q_k}\beta^t(q_k) + 4\,\exp\big(-3\kappa\,\frac{2^k}{k^2} n^{1/3}\log n\big)$$

for $n > n_2$ and all $k \in \mathbb{N}_+$. From the facts that $\beta^t(q_k) \le \beta_0 \varrho^{\sqrt{q_k - t}} \le \beta_0 n^{-(2k+1)}$ and $q_k \ge c_q (k\log n)^2$, we conclude that there exists a number $n_3 \in \mathbb{N}$ and positive constants c_3, c_3' such that for all $n > n_3$ and all $k \in \mathbb{N}_+$

$$\frac{n-t}{q_{n,k}}\beta^t(q_{n,k}) \le c_3 \frac{n-t}{(k\log n)^2} n^{-(2k+1)}$$
$$\le c_3' \frac{n^{-2k}}{(\log n)^2}.$$

If we define $C_1 := \sqrt{\frac{6\kappa}{c_2}}$, $C_2 := 4 + c_3'$, and $C_3 := 3\kappa$, we get the claim of the Lemma for all $n \ge \max\{n_0, \dots, n_3\}$ and all $k \in \mathbb{N}_+$. □

What is the merit of the previous result? As we indicated above, we use it to establish a connection between the modified empirical MSE of certain estimators and their modified $L_2(\pi)$ risk. For these estimators we will intentionally use the notation $\tilde{m}_n$ to emphasize the fact that the approximate least squares estimator $\hat{m}_n$ does not necessarily satisfy the condition of the following Corollary 4.1.15. The reason is that without the assumption of full contractivity of the estimation, it seems hardly possible to establish a link between the empirical MSE of Theorem 4.1.12,

$$\frac{1}{n}\sum_{i=2}^{n} \big(m(\lambda_{i-1}, Y_{i-1}) - \hat{m}(\lambda_{i-1}, Y_{i-1})\big)^2,$$

and the modified empirical MSE,

$$\frac{1}{n-t_n}\sum_{i=t_n}^{n-1} \big(m^{[t_n]}(0, Y_{i-t_n}, \dots, Y_i) - \tilde{m}_n^{[t_n]}(0, Y_{i-t_n}, \dots, Y_i)\big)^2,$$

that shapes the condition of Corollary 4.1.15.

COROLLARY 4.1.15. *Let $\tilde{m}_n[Y_0, \dots, Y_n]$ be an estimator with values in $\mathscr{G}_n$ and the property that*

$$\frac{1}{n-t_n}\sum_{i=t_n}^{n-1} \big(m^{[t_n]}(0, Y_{i-t_n}, \dots, Y_i) - \tilde{m}_n^{[t_n]}(0, Y_{i-t_n}, \dots, Y_i)\big)^2 = O_{\mathbb{P}}(n^{-2/3}(\log n)^3)$$

for a sequence $\{t_n\}$ with $t_n \asymp \log n$. Then, for an independent copy $\{(\lambda_i', Y_i')\}_{i\in\mathbb{Z}}$ of the data generating process, the estimator fulfills also

$$\mathbb{E}_{|\tilde{m}_n[Y_0,\dots,Y_n]=g}\left[m^{[t_n]}(0,Y_0',\dots,Y_{t_n}') - g^{[t_n]}(0,Y_0',\dots,Y_{t_n}')\right]^2 = O_{\mathbb{P}}(n^{-2/3}(\log n)^3)$$

Proof. The proof consists of the following chain of inequalities and an application of the previous Lemma. Let $\varepsilon > 0$ be arbitrary but fixed, and let $C_1 > 0$ and $n_0 \in \mathbb{N}$ be the same quantities as in the previous lemma. We set $\delta_n := C_1 n^{-1/3}(\log n)^{3/2}$ and observe that

$$\begin{aligned}
&\mathbb{P}\left\{\mathbb{E}_{|\tilde{m}_n=g}\left[m^{[t_n]}(\boldsymbol{Y}'^{t_n}_0) - g^{[t_n]}(\boldsymbol{Y}'^{t_n}_0)\right]^2 > 2^l\delta_n^2\right\}\\
&= \mathbb{P}\bigcup_{k=l}^{\infty}\left\{2^{k+1}\delta_n^2 \geq \mathbb{E}_{|\tilde{m}_n=g}\left[m^{[t_n]}(\boldsymbol{Y}'^{t_n}_0) - g^{[t_n]}(\boldsymbol{Y}'^{t_n}_0)\right]^2 > 2^k\delta_n^2\right\}\\
&= \mathbb{P}\bigcup_{k=l}^{\infty}\Bigg(\Bigg\{\mathbb{E}_{|\tilde{m}_n=g}\left[m^{[t_n]}(\boldsymbol{Y}'^{t_n}_0) - g^{[t_n]}(\boldsymbol{Y}'^{t_n}_0)\right]^2\\
&\qquad - \frac{1}{n-t_n}\sum_{i=t_n}^{n-1}\left(m^{[t_n]}(\boldsymbol{Y}^i_{i-t_n}) - \tilde{m}_n^{[t_n]}(\boldsymbol{Y}^i_{i-t_n})\right)^2\\
&\qquad\qquad + \frac{1}{n-t_n}\sum_{i=t_n}^{n-1}\left(m^{[t_n]}(\boldsymbol{Y}^i_{i-t_n}) - \tilde{m}_n^{[t_n]}(\boldsymbol{Y}^i_{i-t_n})\right)^2 > 2^k\delta_n^2\Bigg\}\\
&\qquad\qquad \cap\left\{\mathbb{E}_{|\tilde{m}_n=g}\left[m^{[t_n]}(\boldsymbol{Y}'^{t_n}_0) - g^{[t_n]}(\boldsymbol{Y}'^{t_n}_0)\right]^2 \leq 2^{k+1}\delta_n^2\right\}\Bigg)\\
&\leq \mathbb{P}\bigcup_{k=l}^{\infty}\Bigg(\Bigg\{\mathbb{E}_{|\tilde{m}_n=g}\left[m^{[t_n]}(\boldsymbol{Y}'^{t_n}_0) - g^{[t_n]}(\boldsymbol{Y}'^{t_n}_0)\right]^2\\
&\qquad - \frac{1}{n-t_n}\sum_{i=t_n}^{n-1}\left(m^{[t_n]}(\boldsymbol{Y}^i_{i-t_n}) - \tilde{m}_n^{[t_n]}(\boldsymbol{Y}^i_{i-t_n})\right)^2 > 2^{k-1}\delta_n^2\Bigg\}\\
&\qquad\qquad \cap\left\{\mathbb{E}_{|\tilde{m}_n=g}\left[m^{[t_n]}(\boldsymbol{Y}'^{t_n}_0) - g^{[t_n]}(\boldsymbol{Y}'^{t_n}_0)\right]^2 \leq 2^{k+1}\delta_n^2\right\}\Bigg)\\
&\quad + \mathbb{P}\bigcup_{k=l}^{\infty}\Bigg(\Bigg\{\frac{1}{n-t_n}\sum_{i=t_n}^{n-1}\left(m^{[t_n]}(\boldsymbol{Y}^i_{i-t_n}) - \tilde{m}_n^{[t_n]}(\boldsymbol{Y}^i_{i-t_n})\right)^2 > 2^{k-1}\delta_n^2\Bigg\}\\
&\qquad\qquad \cap\left\{\mathbb{E}_{|\tilde{m}_n=g}\left[m^{[t_n]}(\boldsymbol{Y}'^{t_n}_0) - g^{[t_n]}(\boldsymbol{Y}'^{t_n}_0)\right]^2 \leq 2^{k+1}\delta_n^2\right\}\Bigg)\\
&\leq \mathbb{P}\bigcup_{k=l}^{\infty}\Bigg(\Bigg\{\mathbb{E}_{|\tilde{m}_n=g}\left[m^{[t_n]}(\boldsymbol{Y}'^{t_n}_0) - g^{[t_n]}(\boldsymbol{Y}'^{t_n}_0)\right]^2\\
&\qquad - \frac{1}{n-t_n}\sum_{i=t_n}^{n-1}\left(m^{[t_n]}(\boldsymbol{Y}^i_{i-t_n}) - \tilde{m}_n^{[t_n]}(\boldsymbol{Y}^i_{i-t_n})\right)^2 > 2^{k-1}\delta_n^2\Bigg\}\\
&\qquad\qquad \cap\left\{\mathbb{E}_{|\tilde{m}_n=g}\left[m^{[t_n]}(\boldsymbol{Y}'^{t_n}_0) - g^{[t_n]}(\boldsymbol{Y}'^{t_n}_0)\right]^2 \leq 2^{k+1}\delta_n^2\right\}\Bigg)\\
&\quad + \mathbb{P}\bigcup_{k=l}^{\infty}\left(\left\{\frac{1}{n-t_n}\sum_{i=t_n}^{n-1}\left(m^{[t_n]}(\boldsymbol{Y}^i_{i-t_n}) - \tilde{m}_n^{[t_n]}(\boldsymbol{Y}^i_{i-t_n})\right)^2 > 2^{k-1}\delta_n^2\right\}\right)
\end{aligned}$$

$$
\begin{aligned}
&\leq \sum_{k=l}^{\infty} \mathbb{P}\Bigg\{ \max_{\substack{g\in\mathscr{G}_n:\\ \mathbb{E}\left[m^{[t]}(\boldsymbol{Y}_0^{t_n})-g^{[t]}(\boldsymbol{Y}_0^{t_n})\right]^2\leq 2^{k+1}\delta_n^2}} \mathbb{E}\Big[\big(m^{[t_n]}(\boldsymbol{Y}_0^{t_n})-g^{[t_n]}(\boldsymbol{Y}_0^{t_n})\big)^2\Big] \\
&\qquad\qquad -\frac{1}{n-t_n}\sum_{i=t}^{n-1}\big[m^{[t_n]}(\boldsymbol{Y}_{i-t_n}^{i})-g^{[t_n]}(\boldsymbol{Y}_{i-t_n}^{i})\big]^2 > 2^{k-1}\delta_n\Bigg\} \\
&\quad \underbrace{+\mathbb{P}\Big\{\frac{1}{n-t_n}\sum_{i=t_n}^{n-1}\big(m^{[t_n]}(\boldsymbol{Y}_{i-t_n}^{i})-\tilde{m}_n^{[t_n]}(\boldsymbol{Y}_{i-t_n}^{i})\big)^2 > 2^{l-1}\delta_n^2\Big\}}_{\to 0\ (l\to\infty)} \\
&\leq C_2 \underbrace{\sum_{k=l}^{\infty}\big(n^{-2k}(\log n)^{-2}+e^{-C_3\frac{2^k}{k^2}n^{1/3}\log n}\big)}_{\to 0\ (l\to\infty)}+\frac{\varepsilon}{2} \\
&\leq \varepsilon
\end{aligned}
$$

for all $l > l_0(\varepsilon)$ if $l_0(\varepsilon)$ is chosen sufficiently large, and all $n > n_0$. □

It appears to be a reasonable guess that the modified approximate least squares estimator defined by

$$
\tilde{m}_n := \arg\min_{g\in\mathscr{G}_n}\sum_{i=t_n}^{n-1}\big(Y_{i+1}-g^{[t_n]}(0,Y_{i-t_n},\dots,Y_i)\big)^2 \tag{4.5}
$$

satisfies the condition of Corollary 4.1.15. The strategy for a proof of that claim would closely resemble the proof of Theorem 4.1.12. However, since no new ideas would emerge from a detailed display of the argument, we decide to leave it at the conjecture.

Instead of the original rate $n^{-2/3}(\log n)^2$, we were only able to show that the rate $n^{-2/3}(\log n)^3$ could be established for the $L_2(\pi)$-risk of $\tilde{m}_n$ if the estimator satisfies the condition of Corollary 4.1.15. The additional logarithmic factor can be attributed to the fact that the absolute regularity coefficients of the lagged count process decrease only sub-geometrically. This has the effect that the quantities q_k in the proof of Lemma 4.1.14 increase like $O(\log n)^2$. With geometrically decreasing mixing coefficients, it would suffice to choose these quantities to be of the smaller order $O(\log n)$. This effect can be considered the price to pay for relinquishing the strong contractive condition in Definition 4.1.1.

Alternatively, choosing the thresholds $\{B_n\}$ such that $B_n = O(1)$ would bring us back to the old rate because this choice would spare us the logarithmic factor in the bound of $\log \#\mathscr{G}_n$. However, a bounded sequence of thresholds $\{B_n\}$ would deprive us of the comfortable certainty that our set of candidate functions contains the true link function if only the sample size is sufficiently large. The potential drawbacks of this situation has been discussed in Subsection 3.2.3. A strategy offering this prospect does not seem advisable.

4.1.4 Conclusion

In this section we have proposed a slightly different estimation approach compared to the perspective taken in Chapter 3. Instead of minimizing a least squares functional over the whole class of admissible functions, we search for a solution in a finite subset.

We were able to prove a rate of convergence for the approximate least squares estimator from Definition 4.1.1. Our main result states that the empirical mean square error of the estimator has the order $O_{\mathbb{P}}\big(n^{-2/3}(\log n)^2\big)$. The essential tool for the proof of this result was an exponential tail bound for martingales by Dzhaparidze and van Zanten (2001).

With limited success we contemplated on the possibility to derive a rate for the $L_2(\pi)$-risk on the bases of the result from Theorem 4.1.12. We were not able to adapt the martingale based approach to this endeavor. Hence, we had to employ again techniques that are based on the notion of mixing. As a consequence, we ended up with the condition of Corollary 4.1.15, which unfortunately does not match the result of Theorem 4.1.12 for the approximate least squares estimator. Thus, the conditions under which we could prove that a rate for the empirical MSE implies a similar rate for the $L_2(\pi)$-risk are not suitable for the approximate least squares estimator. However, we believe that this implication might be valid for the slightly modified estimator of equation (4.5).

4.2 The least squares spline estimator

4.2.1 Splines as the approximating class

The last section has revealed the main condition that a possible choice of the approximating grids $\{\mathscr{G}_n\}$ has to satisfy: we have to ensure that the sequence of fineness parameters $\{\rho_n\}$ is asymptotically equivalent to $n^{-1/3}$, and that the number of elements in $\mathscr{G}_n$ is bounded by $e^{\kappa \cdot n^{1/3} \log n}$ for some positive constant κ. Thus, in order to asses the suitability of a proposed set $\mathscr{G}_n$, we need results from approximation theory concerning the distance of $\mathscr{G}$ to $\mathscr{G}_n$ in the uniform metric. A popular and well understood approximation technique for which such results are available is the approximation of continuous functions by splines (de Boor, 1978; Dierckx, 1995; Lyche and Mørken, 2008). We define splines of degree k according to the definition stated in Powell (1981, page 29).

DEFINITION 4.2.1. A function $s\colon [0,M] \to \mathbb{R}$ is called a piece wise polynomial of degree k on the interval $[0,M]$ if $s \in C([0,M])$ and there exist points $0 = \xi_0 < \xi_1 < \ldots < \xi_l = M$ such that s is a polynomial of degree at most k on $[\xi_{i-1}, \xi_i]$ for every $i = 1,\ldots,l$. If additionally $s \in C^{(k-1)}([0,M])$, we call s a spline of degree k with knots $\{\xi_i,\ i = 1,\ldots,l\}$. The set of all such splines will be denoted by $\mathscr{S}(k;\xi_0,\xi_1,\ldots,\xi_l)$.

The next result is well established. We mainly worked with the remarks of Powell (1981, page 29). Another good reference is Lyche and Mørken (2008).

LEMMA 4.2.2. *The spline space $\mathscr{S}(k;\xi_0,\ldots,\xi_l)$ is a linear space with dimension $k+l$.*

Proof. A linear combination of piece wise polynomials is again a piece wise polynomials. The continuous differentiability carries over to linear combinations as well. Hence, $\mathscr{S}(k;\xi_0,\ldots,\xi_l)$ is a linear space.

For $i = 0,\ldots,l-1$ and $j = 0,\ldots,k$, the functions $\mu_{i,j}$ with

$$\mu_{i,j}(x) = \begin{cases} (x-\xi_i)^j \mathbb{1}_{[\xi_i,\xi_{i+1})}(x) & \text{for } i \in \{0,\ldots,l-2\}, \\ (x-\xi_i)^j \mathbb{1}_{[\xi_i,\xi_{i+1}]}(x) & \text{for } i = l-1 \end{cases}$$

are linearly independent and span the space of piece wise polynomials of degree at most k. Therefore, the dimension of the space of piece wise polynomials is $(k+1)\cdot l$.

Consider a piece wise polynomial $s(x) = \sum_{i=0}^{l-1} \sum_{j=0}^{k} \alpha_{i,j} \mu_{i,j}(x)$. All derivatives of s exist except at the knots. At $x \in [0,M] \setminus \{\xi_0,\ldots,\xi_l\}$, the derivative of order q is given by

$$(\partial_x^q s)(x) = \sum_{i=0}^{l-1} \sum_{j=q}^{k} \alpha_{i,j} \frac{j!}{(j-q)!} (x-\xi_i)^{j-q} \mathbb{1}_{[\xi_i,\xi_{i+1})}(x).$$

Since the limits $\lim_{x\to\xi_i}(\partial_x^q s)(x)$ exist and are finite, the derivative $(\partial_x^q s)(x)$ exists at every point $x \in [0,M]$ if and only if $\lim_{x\uparrow\xi_i}(\partial_x^q s)(x) = \lim_{x\downarrow\xi_i}(\partial_x^q s)(x)$ for all $i \in \{1,\dots,l-1\}$. This claim can be easily verified. To prove sufficiency, let $a := \lim_{x\to\xi_i}(\partial_x^q s)(x)$ and note that for any $\varepsilon > 0$ there exists a $\delta(\varepsilon) > 0$ such that $a-\varepsilon < (\partial_x^q s)(x) < a+\varepsilon$ for all $x \in [\xi_i-\delta,\xi_i+\delta]\setminus\{\xi_i\}$. Hence, $a-\varepsilon \le \frac{(\partial_x^{q-1}s)(x)-(\partial_x^{q-1}s)(y)}{x-y} \le a+\varepsilon$ for all $\xi_i-\delta < y < x < \xi_i+\delta$ (Königsberger, 2004, page 164), which means that $\partial_x^-(\partial_x^{q-1}s)(\xi_i)$ and $\partial_x^+(\partial_x^{q-1}s)(\xi_i)$ exist and are equal to a. To verify necessity, we invoke the fact that a derivative cannot have simple discontinuities (Rudin, 1976, page 109).

Therefore, $s \in C^{(k-1)}([0,M])$ if and only if the constraints

$$\sum_{j=q}^{k} \alpha_{i,j} \frac{j!}{(j-q)!}(\xi_{i+1}-\xi_i)^{j-q} = q!\,\alpha_{i+1,q}$$

for $q = 0,\dots,k-1$ and $i = 0,\dots,l-2$ are satisfied. In other words, a piece wise polynomial $s = \sum_{i=0}^{l-1}\sum_{j=0}^{k}\alpha_{i,j}\mu_{i,j}$ is an element of $\mathscr{S}(k;\xi_0,\dots,\xi_l)$ if and only if it solves

$$\begin{pmatrix}\boldsymbol{A^{(i)}} & \boldsymbol{B}\end{pmatrix}\begin{pmatrix}\alpha_{i,0}\\ \vdots \\ \alpha_{i,k}\\ \alpha_{i+1,0}\\ \vdots\\ \alpha_{i+1,k}\end{pmatrix} = \begin{pmatrix}0\\ \vdots\\ 0\end{pmatrix}$$

for all $i \in \{0,\dots,l-2\}$, where the $(k+1)\times(k+1)$ matrices $\boldsymbol{A^{(i)}}$ and $\boldsymbol{B}$ are given by

$$\boldsymbol{A^{(i)}} := \begin{pmatrix} A_{0,0}^{(i)} & A_{0,1}^{(i)} & \dots & A_{0,k-1}^{(i)} & A_{0,k}^{(i)}\\ 0 & A_{1,1}^{(i)} & \dots & A_{1,k-1}^{(i)} & A_{1,k}^{(i)}\\ \vdots & \ddots & \ddots & \vdots & \vdots\\ 0 & \dots & 0 & A_{k-1,k-1}^{(i)} & A_{k-1,k}^{(i)}\\ 0 & \dots & 0 & 0 & 0\end{pmatrix},\ A_{q,j}^{(i)} := \binom{j}{q}(\xi_{i+1}-\xi_i)^{j-q},$$

and

$$\boldsymbol{B} := \begin{pmatrix} -1 & 0 & 0 & \dots & 0\\ 0 & -1 & 0 & \dots & 0\\ \vdots & \ddots & \ddots & \ddots & \vdots\\ 0 & \dots & 0 & -1 & 0\\ 0 & \dots & 0 & 0 & 0\end{pmatrix}$$

and have rank k. Hence, with respect to the basis $\{\mu_{i,j} : i = 0,\dots,l-1; j = 0,\dots,k\}$, the set $\mathscr{S}(k;\xi_0,\dots,\xi_l)$ is equal to the kernel of

$$\boldsymbol{A} := \begin{pmatrix} \boldsymbol{A^{(0)}} & \boldsymbol{B} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{A^{(1)}} & \boldsymbol{B} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \mathbf{0} \\ \mathbf{0} & \dots & \mathbf{0} & \boldsymbol{A^{(l-2)}} & \boldsymbol{B} \\ \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{pmatrix}.$$

Recall that the space of piece wise polynomials has dimension $(k+1)l$. Since the matrix $\boldsymbol{A}$ has rank $(l-1)k$, its kernel has the dimension $(k+1)l-(l-1)k = k+l$. We conclude that $\mathscr{S}(k;\xi_0,\dots,\xi_l)$ is a linear space with dimension $k+l$.

Alternatively, consider the basis of truncated power functions

$$\left\{(x-\xi_i)^j \mathbb{1}_{(\xi_i,\xi_l]} : i = 0,\dots,l-1;\ j = 0,\dots,k\right\}.$$

Since $(x-\xi_j)_+$ is not differentiable in x at $x = \xi_j$, all truncated power functions $(x-\xi_i)_+^j$, for which $j \in \{0,1,\dots,k-1\}$ and $i \in \{1,\dots,l-1\}$, are not in $C^{(k-1)}([0,M])$ because $\partial_x^-\left(\partial_x^{j-1}(x-\xi_i)_+^j\right)$ and $\partial_x^+\left(\partial_x^{j-1}(x-\xi_i)_+^j\right)$ at $x = \xi_i$ exist but they do not coincide. For the same reason, any linear combination of truncated power functions that involve some $(x-\xi_i)_+^j$ for $i > 0$ and $j < k$ is not in $C^{(k-1)}([0,M])$. Therefore, any spline of order k is a linear combination of the remaining of $k+l$ truncated power functions. □

We have already learned about two different bases of the spline space. Yet, they are not the most practical ones. Next, we want to introduce the alternative basis of B-splines. Again, we follow the presentation of Powell (1981, pages 229 and 241).

DEFINITION 4.2.3. For a subset $\{\xi_p,\dots,\xi_{p+k+1}\} \subset \{\xi_{-k},\dots,\xi_{l+k}\}$ of $k+1$ successive points, the B-spline with knots $\xi_p,\dots,\xi_{p+k+1}$ and order k is defined as

$$B_{p,k}(x) := \sum_{j=p}^{p+k+1} \Big[\prod_{\substack{i=p \\ i\neq j}}^{p+k+1} \frac{1}{(\xi_i - \xi_j)} \Big] (x-\xi_j)_+^k. \tag{4.6}$$

The normalized B-splines are given by

$$N_{p,k}(x) := (\xi_{p+k+1} - \xi_p) B_{p,k}(x). \tag{4.7}$$

LEMMA 4.2.4. *Consider a set of $l+2k+1$ points in ascending order,*

$$\xi_{-k} < \dots < \xi_0 < \dots < \xi_l < \dots < \xi_{l+k},$$

with $\xi_0 = 0$ *and* $\xi_l = M$*. For* $p \in \{-k, \dots, l-1\}$*, let the B-spline* $B_{p,k}$ *be given by* (4.6)*. The set of B-splines restricted to the interval* $[0, M]$*,*

$$\left\{ B_{p,k}\big|_{[0,M]} \colon p \in \{-k, \dots, l-1\} \right\},$$

is a basis of the vector space $\mathcal{S}(k; \xi_0, \dots, \xi_l)$*. Of course, the same is true for the set of normalized B-splines restricted to* $[0, M]$*.*

Proof. Cf. Lemma A.5.1. □

Lemma 4.2.2 tells us that the space of splines with a certain degree k and number of knots $l+1$ is a linear space with dimension $l+k$. However, the crucial question remains to be answered yet. How accurately can a continuous function be approximated by a spline of given degree k? It seems intuitive that the quality of the approximation grows with the dimension of the approximating space. We will see that this intuition is quite arguable. The next lemma and its proof are modeled on the exposition of Powell (1981, Theorem 20.2).

LEMMA 4.2.5. *Let* $f \in C[0, M]$ *be a Lipschitz continuous function with Lipschitz constant* L*. Then there exists a spline function* $s \in \mathcal{S}(k; \xi_0, \dots, \xi_l)$ *such that*

$$\|f - s\|_\infty \leq (k+1) L \max_{i \in \{0, \dots, l-1\}} |\xi_{i+1} - \xi_i|.$$

Moreover, if the knots are equidistant, then s *satisfies the same Lipschitz condition as* f*.*

Proof. We consider the absolute difference between f and a spline s at a point $x \in [0, M]$. The spline s can be written as a linear combination of the normalized B-splines,

$$s(x) = \sum_{p=-k}^{l-1} \alpha_p N_{p,k}(x).$$

The spline s that we use to approximate f shall be given by the coefficients $\alpha_p := f(\xi_p)$. Then, by the partition of unity property of the normalized B-splines on $[0, M]$ (Lemma A.5.1 (vi)), we obtain

$$|f(x) - s(x)| = \left| \sum_{p=-k}^{l-1} \big(f(x) - \alpha_p\big) N_{p,k}(x) \right|.$$

Assume without loss of generality that $x \in [\xi_i, \xi_{i+1}]$ and $i \in \{0, \dots, l-1\}$. Lemma A.5.1 (i) and the continuity of B-splines tells us that in this case $N_{p,k}(x) = 0$ if

$p \le i-k-1$ or $p \ge i+1$. Then the sum in the last display boils down to

$$\begin{aligned}\Big|\sum_{p=i-k}^{i}\big(f(x)-\alpha_p\big)N_{p,k}(x)\Big| &\le \max_{i-k\le p\le i}|f(x)-\alpha_p|\sum_{p=i-k}^{i}N_{p,k}(x)\\ &= \max_{i-k\le p\le i}|f(x)-\alpha_p|\sum_{p=-k}^{l-1}N_{p,k}(x)\\ &= \max_{i-k\le p\le i}|f(x)-\alpha_p|. \qquad (4.8)\end{aligned}$$

Recall that $\alpha_p = f(\xi_p)$. We assume without loss of generality that the knots outside the interval $[0,M]$ can be chosen such that their distance to each other is smaller than the maximal distance between knots inside the interval. Hence,

$$\begin{aligned}|f(x)-s(x)| &\le \max_{i-k\le p\le i}|f(x)-f(\xi_p)|\\ &\le L\max_{i-k\le p\le i}|x-\xi_p|\\ &\le L\max_{i}|\xi_{i+1}-\xi_{i-k}|\\ &\le L(k+1)\max_{p}|\xi_p-\xi_{p-1}|.\end{aligned}$$

It remains to be shown that s has the Lipschitz property in case of equidistant knots. For that sake we compute the derivative of s on $[\xi_0,\xi_l]$ in between the knots. With the formula of Lemma A.5.1 and the fact that the B-splines are non-negative, we see that

$$\begin{aligned}|\partial_x s(x)| &= \Big|\sum_{p=-k}^{l-1} f(\xi_p)(\xi_{p+k+1}-\xi_p)\,\partial_x B_{p,k}(x)\Big|\\ &= \Big|\sum_{p=-k}^{l-1} k f(\xi_p)\big(B_{p,k-1}(x)-B_{p+1,k-1}(x)\big)\Big|\\ &= \Big|k\,f(\xi_{-k})\underbrace{B_{-k,k-1}(x)}_{=0 \text{ if } x\ge\xi_0}\\ &\qquad + \sum_{p=-k+1}^{l-1}\Big[k\big(f(\xi_p)-f(\xi_{p-1})\big)B_{p,k-1}(x)\Big]\\ &\qquad\qquad -k f(\xi_{l-1})\underbrace{B_{l,k-1}(x)}_{=0 \text{ if } x\le\xi_l}\Big|\\ &= \Big|\sum_{p=-k+1}^{l-1}\frac{f(\xi_p)-f(\xi_{p-1})}{\xi_p-\xi_{p-1}}\,k(\xi_p-\xi_{p-1})B_{p,k-1}(x)\Big|\\ &\le L\sum_{p=-(k-1)}^{l-1}(\xi_{p+k}-\xi_p)B_{p,k-1}(x)\\ &= L.\end{aligned}$$

□

An immediate corollary of the last lemma is that in the case of equidistant knots the distance between the class of Lipschitz functions on $[0,M]$ and

$\mathscr{S}(k;\xi_0,\ldots,\xi_m)$ intersected with that same class is at most

$$\sup_f \inf_s \|f-s\|_\infty \leq (k+1)L \max_i |\xi_{i-1}-\xi_i|.$$

The only way to increase the dimension of $\mathscr{S}$ while keeping k fixed is the insertion of new knots. If we assume that the knots are evenly distributed among the interval $[0,M]$, increasing the number of knots has the consequence that the size of the partition in terms of the maximal distance of two neighboring knots decreases. Under the condition of fixed degree of smoothness k and balanced knot distribution, we see now that increasing the dimension of the approximating spline space indeed lowers the maximal approximation error.

So far, we have investigated the properties of the finite dimensional set $\mathscr{S}(k;\xi_0,\ldots,\xi_l)$ and concluded, given a reasonable knot sequence, that it is appropriate for the approximation of Lipschitz functions. Yet, to realize the ideas of the approximate least squares estimator, we need an approximating set of finitely many elements. We can obtain such a set by restricting the coefficients $\{\alpha_p\}_{p=-k}^{l-1}$ to a finite set A_n.

DEFINITION 4.2.6. Let $k \geq 1$ be a natural number and

$$\xi_{-k} < \ldots < \xi_0 = 0 < \xi_1 < \ldots < \xi_l = M < \xi_{l+1} < \ldots < \xi_{l+k}$$

a sequence of equidistant knots. Suppose that A_n is a set of real numbers. Let $\mathscr{S}(k;\xi_{-k},\ldots,\xi_{l+k};A_n)$ denote the linear combinations of the normalized B-splines $\{N_{p,k}|_{[0,M]}\}_{p=-k}^{l-1}$ with coefficients α_p from the set A_n.

For the set $\mathscr{S}(k;\xi_{-k},\ldots,\xi_{l+k};A_n)$ we have a statement about the quality of approximations of bounded Lipschitz functions, which is similar to Lemma 4.2.5.

LEMMA 4.2.7. *Let $k \geq 1$ be an integer and $\{\xi_p\}_{p=-k}^{l+k}$ be a knot sequence as in the previous definition. Let the finite set A_n be such that*

$$\sup_{x\in[0,M]} \min_{\lambda\in A_n} |\lambda - x| \leq \max_p |\xi_p - \xi_{p-1}|.$$

Let $f:[0,M]\to[0,M]$ be a Lipschitz continuous function with Lipschitz constant L. Then there exists a spline function $s \in \mathscr{S}(k;\xi_{-k},\ldots,\xi_{l-k};A_n)$ such that

$$\|f-s\|_\infty \leq (Lk+L+1) \max_{p\in\{0,\ldots,l-1\}} |\xi_{p+1}-\xi_p|.$$

Proof. Copy the proof of Lemma 4.2.5 until line (4.8). Here we make the slight modification defining

$$\alpha_p := \arg\min_{\lambda \in A_n} |\lambda - f(\xi_p)|.$$

By assumption, it follows that

$$|\alpha_p - f(\xi_p)| \le \max_p |\xi_p - \xi_{p-1}|.$$

We conclude that

$$\begin{aligned} |f(x) - \alpha_p| &= |f(x) - f(\xi_p) + f(\xi_p) - \alpha_p| \\ &\le |f(x) - f(\xi_p)| + \max_p |\xi_p - \xi_{p-1}| \end{aligned}$$

and

$$\begin{aligned} |f(x) - s(x)| &\le \max_i \max_{\substack{i-k \le p \le i \\ x \in [\xi_i, \xi_{i+1}]}} |f(x) - \alpha_p| \\ &\le \max_p |\xi_p - \xi_{p-1}| + \max_i \max_{\substack{i-k \le p \le i \\ x \in [\xi_i, \xi_{i+1}]}} |f(\xi_p) - f(x)| \\ &\le \max_p |\xi_p - \xi_{p-1}| + L \max_i |\xi_{i-k} - \xi_{i+1}| \\ &\le \max_p |\xi_p - \xi_{p-1}| + L(k+1) \max_p |\xi_p - \xi_{p-1}|. \end{aligned}$$

□

In the next definition we use the preceding results on splines to define a specific approximation grid $\mathscr{G}_n$ from which we select the approximate least squares estimator. The grid will consist of functions $s(\lambda, y) \in \mathscr{G}$ such that for all y the univariate functions $\lambda \mapsto s(\lambda, y)$ are splines of order k. It is convenient to choose $k = 2$ since this ensures continuous differentiability with respect to λ.

DEFINITION 4.2.8. Let the equidistant knot sequence $\{\xi_p\}_{p=-2}^{l_n+2}$ be given by

$$\xi_{-2} < \xi_{-1} < \xi_0 = 0 < \xi_1 \ldots < \xi_{l_n} = M < \xi_{l_n+1} < \xi_{l_n+2}.$$

The knot distance is denoted by $\Delta_n := \xi_{i+1} - \xi_i$ for $i \in \{-2, \ldots, l_n + 1\}$. Let A_n be an equidistant partition of $[0, M]$ with $\frac{l_n}{2} + 1 \le \# A_n \le C\, l_n$ points; $A_n = (\alpha_0, \alpha_1, \ldots, \alpha_{K_n})$ in ascending order with $\alpha_0 = 0$ and $\alpha_{K_n} = M$. The sequence $\{\Delta_n\}$ is required to satisfy $\Delta_n \asymp n^{-1/3}$. Note that then $l_n = \frac{M}{\Delta_n} \asymp n^{1/3}$. Let the class $\mathscr{G}(M, B_n, L$ be as in Definition 4.1.1. The grid $\mathscr{G}_n$ is then defined as

$$\mathscr{G}_n := \left\{ s \in \mathscr{G}(M, B_n, L) \colon s(\cdot, y) \in \mathscr{S}(2; \xi_{-2}, \ldots, \xi_{l_n+2}; A_n) \text{ for all } y \in \{0, \ldots, B_n - 1\} \right\}.$$

The approximate least squares estimator on the basis of the approximating set $\mathscr{G}_n$

in the sense of Definition 4.1.1 is called *least squares spline estimator*.

LEMMA 4.2.9. *The least squares spline estimator is consistent with rate* $n^{-1/3}\log n$.

Proof. We quickly verify that the set $\mathscr{G}_n$ satisfies the conditions of Theorem 4.1.12. The distance of any point $x \in [0,M]$ to its closest element in A_n is at most $\frac{M}{2K_n} = \frac{M}{2(\#A_n-1)} \leq \Delta_n$. This means that the condition "$\sup_{x\in[0,M]} \min_{\lambda\in A_n} |\lambda - x| \leq \max_p |\xi_p - \xi_{p-1}|$" from Lemma 4.2.7 is satisfied, and we can conclude that for any $g^* \in \mathscr{G}(M,B_n,L)$ there exist functions $s_0^*,\ldots,s_{B_n-1}^* \in \mathscr{S}(2;\xi_{-2},\ldots,\xi_{l_n+2};A_n)$ such that for all $y \in \{0,\ldots,B_n-1\}$

$$\sup_{\lambda\in[0,M]} |s_y^*(\lambda) - g^*(\lambda,y)| \leq (2L+L+1)\Delta_n\,.$$

Defining $s^* \in \mathscr{G}_n$ by $s(\lambda,y) := s_y^*(\lambda)$, we see that

$$\begin{aligned}
\min_{s\in\mathscr{G}_n} \|g^* - s\|_\infty &= \min_{s\in\mathscr{G}_n} \max_{y\in\{0,\ldots,B_n-1\}} \sup_{\lambda\in[0,M]} \left|s(\lambda,y) - g^*(\lambda,y)\right| \\
&\leq \max_{y\in\{0,\ldots,B_n-1\}} \sup_{\lambda\in[0,M]} \left|s^*(\lambda,y) - g^*(\lambda,y)\right| \\
&= \max_{y\in\{0,\ldots,B_n-1\}} \sup_{\lambda\in[0,M]} \left|s_y^*(\lambda) - g^*(\lambda,y)\right| \\
&\leq (2L+L+1)\Delta_n \asymp n^{-1/3}\,.
\end{aligned}$$

Since g^* was arbitrary, the relations holds uniformly for all $g^* \in \mathscr{G}$, which includes $g^* = m$. This means in the notation of Theorem 4.1.12 that ρ_n has the right order. The number of elements in $\mathscr{G}_n$ is bounded by

$$\#\mathscr{G}_n = (\#A_n)^{B_n(l_n+2)} = e^{B_n(l_n+2)\log \#A_n} \leq e^{\kappa\cdot n^{1/3}\log n}$$

for some positive constant κ. Thus, Theorem 4.1.12 is valid for the least squares spline estimator, and the claim follows. □

The rate is conjectured to be optimal up to the logarithmic term. With the next lemma we provide the recipe to compute the least squares spline estimator.

LEMMA 4.2.10. *Let the* $(l_n+2)B_n$*-dimensional vector* $\boldsymbol{\alpha}$ *be given by*

$$\boldsymbol{\alpha} := \big(\alpha_{-2}(y),\ldots,\alpha_{l_n-1}(y)\big)_{y=0,\ldots,B_n-1} \in A_n^{(l_n+2)B_n}\,,$$

and let the mappings $S_n(\boldsymbol{\alpha})\colon A_n^{(l_n+2)B_n} \to \mathscr{G}_n$ *and* $Q(\boldsymbol{\alpha},\boldsymbol{y})\colon A_n^{(l_n+2)B_n}\times\mathbb{N}^{n+1} \to [0,\infty)$*, and the constraint function* $C = \big(C_{y,0}\,, C_{y,1}\big)_{y=0,\ldots,B_n-1}\colon A_n^{(l_n+2)B_n} \to \mathbb{R}^{2B_n}$ *be given by*

$$S_n\boldsymbol{\alpha}[\lambda,y] := \sum_{p=-2}^{l-1} \alpha_p(y) N_{p,2}(\lambda),$$

$$Q(\boldsymbol{\alpha},\boldsymbol{y}) := \sum_{i=0}^{n-1} \Big(y_{i+1} - \big(S_n \boldsymbol{\alpha}\big)^{[i]}[0, y_0, \ldots, y_i] \Big)^2$$

$$C_{y,i}(\boldsymbol{\alpha}) := \sup_{\lambda \in [0,M]} \sum_{p=-1}^{l_n - 1} \Big((-1)^i \frac{\alpha_p(y) - \alpha_{p-1}(y)}{\Delta_n} - L \Big) N_{p,1}(\lambda), \quad i \in \{0,1\}.$$

The mapping S_n maps the parameter vector $\boldsymbol{\alpha}$ to the corresponding spline function. The functional Q is the sum of squares functional for this spline function and the observations $\boldsymbol{y} = (y_0, \ldots, y_n)$.

(i) Let $\boldsymbol{\alpha}_n^* \in A_n^{(l_n+2)B_n}$ be a solution of the restricted optimization problem

$$\begin{aligned} &\operatorname*{arg\,min}_{\boldsymbol{\alpha}} \big\{ Q(\boldsymbol{\alpha}, \boldsymbol{Y}) \colon \boldsymbol{\alpha} \in A_n^{(l_n+2)B_n} \big\} \\ &\text{subject to } C(\boldsymbol{\alpha}) \le 0, \end{aligned} \tag{4.9}$$

where the vector $\boldsymbol{Y} = (Y_0, \ldots, Y_n)$ is given by $n+1$ successive count variables of the data generating process. The least squares spline estimator is given by

$$\hat{m}_n = S_n(\boldsymbol{\alpha}_n^*).$$

(ii) Let the underlying knot sequence $\{\xi_i\}_{i=-2}^{2+l_n}$ be equidistant with $l_n \le L$. Then the least squares spline estimator is given by $\hat{m}_n = S_n(\boldsymbol{\alpha}_n^*)$, where

$$\boldsymbol{\alpha}_n^* = \operatorname*{arg\,min}_{\boldsymbol{\alpha}} \big\{ Q(\boldsymbol{\alpha}, Y) \colon \boldsymbol{\alpha} \in A_n^{(l_n+2)B_n} \big\}$$

is the unconstrained minimizer of $\boldsymbol{\alpha} \mapsto Q(\boldsymbol{\alpha}, \boldsymbol{Y})$.

Proof. (i) Certainly, every element in $\mathscr{G}_n$ corresponds to a unique element $\boldsymbol{\alpha} \in A_n^{(l_n+2)B_n}$ that satisfies $S_n(\boldsymbol{\alpha}) \in \mathscr{G}(M, B_n, L)$. Therefore, the least squares spline estimator is uniquely determined by the minimizer of the functional $\boldsymbol{\alpha} \mapsto Q(\boldsymbol{\alpha}, \boldsymbol{Y})$ under the restriction that $S_n(\boldsymbol{\alpha}) \in \mathscr{G}(M, B_n, L)$. An element $\boldsymbol{\alpha} \in A_n^{(l_n+2)B_n}$ is called feasible if $S_n(\boldsymbol{\alpha}) \in \mathscr{G}(M, B_n, L)$. The necessary and sufficient condition for a vector $\boldsymbol{\alpha}$ to be feasible is that $S_n(\boldsymbol{\alpha})$ satisfies the smoothness condition (L) from Definition 4.1.1 (b). This means that we have to check the condition

$$\big| S_n \boldsymbol{\alpha}[\lambda_1, y] - S_n \boldsymbol{\alpha}[\lambda_2, y] \big| \le L \, |\lambda_1 - \lambda_2|$$

for all y. Since the functions $\lambda \mapsto S_n \boldsymbol{\alpha}[\lambda, y]$ are by assumption splines of order 2, we know that the first derivatives exist everywhere. Hence, the smoothness condition is fulfilled if and only if for all $y \in \{0, \ldots, B_n - 1\}$ the derivative with respect to λ is at most L:

$$\max_y \sup_\lambda \big| \partial_\lambda (S_n \boldsymbol{\alpha})[\lambda, y] \big| \le L. \tag{4.10}$$

The formula for the derivatives of spline functions,

$$\partial_x\Big(\sum_{p=-k}^{l-1}\alpha_p N_{p,k}(x)\Big)=\sum_{p=-(k-1)}^{l-1}k\,\frac{\alpha_p-\alpha_{p-1}}{\xi_{p+k}-\xi_p}N_{p,k-1}(x),$$

can be found in Lemma A.5.1 (iv). We are in the case of equidistant knots with distance Δ_n and spline degree $k=2$. This fact and the partition of unity property, $\sum_{p=-1}^{l-1}N_{p,1}\equiv 1$, yield

$$\begin{aligned}&\sup_{\lambda\in[0,M]}\Big((-1)^i\,\partial_\lambda(S_n\boldsymbol{\alpha})[\lambda,y]-L\Big)\\&=\sup_{\lambda\in[0,M]}\Big(\sum_{p=-1}^{l-1}(-1)^i\,\frac{\alpha_p(y)-\alpha_{p-1}(y)}{\Delta_n}N_{p,1}(\lambda)-L\sum_{p=-1}^{l-1}N_{p,1}(\lambda)\Big)\\&=C_{y,i}(\boldsymbol{\alpha}).\end{aligned}$$

The necessary bound (4.10) holds for $\boldsymbol{\alpha}$ if and only if $C_{y,i}(\boldsymbol{\alpha})\le 0$ for all $i\in\{-1,1\}$ and all $y\in\{0,\dots,B_n-1\}$. This proves the first part.

(ii) By definition, all coefficients $\alpha_p(y)$ are selected from the set $A_n\subset[0,M]$. The distance between the equidistant knots is $\Delta_n=M/l_n$. Therefore,

$$|a_p(y)-a_{p-1}(y)|\le M=l_n\,\Delta_n\le L\,\Delta_n$$

for all $y=0,\dots,B_n-1$. Note that the function $\lambda\mapsto\partial_\lambda(S_n\boldsymbol{\alpha})[\lambda,y]$ is a spline of first degree. We invoke Lemma A.5.1 (v) and the last estimate to conclude that for any y the absolute value $\big|\partial_\lambda(S_n\boldsymbol{\alpha})[\lambda,y]\big|$ is bounded by

$$\max_p\,2\,\frac{|\alpha_p(y)-\alpha_{p-1}(y)|}{\xi_{p+2}-\xi_p}=\max_p\frac{|\alpha_p(y)-\alpha_{p-1}(y)|}{\Delta_n}\le L.$$

Thus, the constraint (4.10) is always satisfied. □

In order to obtain an estimation, we have to solve the high dimensional constrained optimization problem (4.9). If we attempt to solve this optimization problem, we have to come up with an algorithm that is designed to find an approximate global maximum point of a high dimensional real valued function over a very fine grid. This calls for algorithms from the field of integer programming. An alternative approach could be to treat the variables α_p as continuous variables. In this case we would have to solve a so called global optimization problem for an objective function without indication of convexity. The acquired solution will be close to the solution of the original discrete problem if the grid A_n is very fine. In Appendix A.6 we gathered some additional information about the computational complexity of these problems and some techniques that are available to solve them.

Once we have figured out a way to solve Problem (4.9), we face the challenge to choose the hyper-parameters that arise from the definition of the estimator. The central objects that have to be determined prior to an application of our model are the set $\mathscr{G}(M,B_n,L)$ and the approximating finite subsets $\mathscr{G}_n$. The defining parameters for $\mathscr{G}$ are the domain boundaries M and B_n, and the Lipschitz constant L; for the finite subsets $\mathscr{G}_n$, we additionally have to specify l_n and A_n, the number of knots and the set of coefficients for the splines respectively. To avoid model misspecification, the constants M, B_n, and L must not be chosen too small: if the true function m is not contained in $\mathscr{G}(M,B_n,L)$, we have obviously no chance to select a realization of a consistent estimator.

Having observed the data, it has to be made plausible that the given choices of the constants are indeed large enough. Even though B_n grows with the sample size, we do not know whether our specific sample is sufficiently large to guarantee that $B_n \geq B$. In the case that $B_n < B$, difficulties may arise if some observations exceed B_n. However, if this does not happen, a possible model misspecification due to $B_n < B$ will not be effective. In this case, we may safely accept the proposed value for B_n.

Determining adequacy of the choice of M (i.e. the boundary of the domain of estimation with respect to the intensity variable) requires a more involved procedure since the intensities are not observed. However, the value M should be large enough to explain the observed counts reasonably well. Assume that we have used prior information to guess an upper bound M for the intensities. Formally, we could consider at each time t the family of distributions $\mathscr{P}_t := \{\mathbb{P}^{Y_t|\lambda_t=\lambda} : \lambda \in \mathbb{R}_+\}$ and test the hypothesis

$$H_0 : \lambda \geq M \qquad vs. \qquad H_1 : \lambda < M .$$

For a given level of significance α_t, a test $T_t(Y_t)$ that satisfies

$$\sup_{\lambda \geq M} \beta_{T_t}(\lambda) \leq \alpha_t \tag{4.11}$$

for the power function $\beta_{T_t}(\lambda) := \mathbb{E}[T_t(Y_t) \,|\, \lambda_t = \lambda]$ would be given by a randomized test rejecting H_0 with the probability

$$T_t(Y_t) = \begin{cases} 1 & \text{if } Y_t < c_{\alpha_t} \\ \gamma_{\alpha_t} & \text{if } Y_t = c_{\alpha_t} \\ 0 & \text{if } Y_t > c_{\alpha_t} . \end{cases}$$

The critical value c_{α_t} is given by $c_{\alpha_t} = \inf\{c : \mathbb{P}\{Y_t \leq c \,|\, \lambda_t = M\} \geq \alpha_t\}$, and $\gamma_{\alpha_t} = \frac{\alpha_t - \mathbb{P}\{Y_t < c_{\alpha_t} | \lambda_t = M\}}{\mathbb{P}\{Y_t = c_{\alpha_t} | \lambda_t = M\}}$. The family $\mathscr{P}_t$ is by assumption the family of Poisson distribu-

tions with intensities $\lambda > 0$. This family has a monotone likelihood ratio, which implies that the power function $\lambda \mapsto \beta_{T_t}(\lambda)$ is monotonically non-increasing (Shao, 2003, Lemma 6.3), and the property (4.11) follows from $\beta_{T_t}(M) = \alpha_t$. If at all time instances the null hypothesis can be rejected, we have reason to believe that all intensities λ_t have fallen into the interval $[0, M]$. Since the power function β_{T_t} is non-increasing, the test T_t has larger power in regions that are more distant from the interval $[M, \infty)$. Thus, in order to prevent failure in rejecting H_0 for values $\lambda < M$, it seems advisable to avoid narrow choices of M.

Regarding the evaluation of the choice for L, we can safely accept any choice $L \geq 1$ because this guarantees $\ell < L$.

So far, the described ad-hoc procedures give hints as to whether we should dismiss a particular choice for M, B_n, or L as too small. With regard to excessively large values of the constants, however, they are insensitive. From a statistical point of view, extreme choices of model parameters should be avoided. Even thought the asymptotic result holds regardless of the constants' magnitudes, large constants may deteriorate the estimators performance. The proof of Theorem 4.1.12 suggests that larger classes $\mathscr{G}_n$ lead to larger bounds for the MSE in probability. The quantities M, B_n, L, A_n, and l_n determine the size of the sets $\mathscr{G}_n$. Larger values of the constants lead to larger sets of candidate functions, from which we expect larger estimation errors. In analogy to non-parametric smoothing techniques, we could term this phenomenon *over-fitting*. Its terminological counterpart, *over-smoothing*, is expected to occur with too small choices of M, B_n, L, A_n, and l_n. In order to prevent the occurrence of these phenomena, it seems advisable to apply an adaptive procedure of hyper-parameter selection.

An illustrative example of such an adaptive inference procedure is the data driven choice of the bandwidth for kernel regression or density estimators. Several methods have been proposed to select the bandwidth in a data driven way. A classical approach is the use of leave-one-out cross-validation for kernel-based estimators in i.i.d. settings (Rudemo, 1982; Stone, 1984; Hardle and Marron, 1985; Györfi et al., 2002; Tsybakov, 2008). Adapting the presentation of Györfi et al. (2002, page 112), we give a short motivation of the method, and we briefly mention the modifications which may be appropriate to adapt this method to our setting. The basic i.i.d. setting is as follows.

Suppose that the standard regression model $\mathbb{E}(Y \mid X = x) = m(x)$ explains the data $D_n = \{(X_i, Y_i) \colon i = 1, \ldots, n\}$. We consider a finite set of possible parameters $H = \{h_1, \ldots, h_q\}$ such that for every $h \in H$ there is a corresponding estimator $\hat{m}_{n,h}$ of m (e.g. the bandwidth of a kernel regression estimator). Our goal is to select the best deterministic choice $\bar{h}_n$, defined by

$$\mathbb{E} \int \left(\hat{m}_{n,\bar{h}_n} - m \right)^2 dP^X = \min_{h \in H} \mathbb{E} \int \left(\hat{m}_{n,h} - m \right)^2 dP^X ,$$

in a data driven way. The idea of cross-validation is to divide the sample in training and validation samples. In the case of leave-one-out, we would define for $i = 1,\dots,n$ the ith training sample as

$$D_{-i} = \left\{(X_j, Y_j)\colon j \in \{1,\dots,n\} \setminus \{i\}\right\}.$$

Accordingly, $\hat{m}_{n,h}[D_{-i}]$ denotes the estimator based on the ith training sample with parameter $h \in H$. The best deterministic choice of h can be estimated by choosing,

$$h^* := \arg\min_{h\in H} \frac{1}{n}\sum_{i=1}^{n} \left[\left(\hat{m}_{n,h}[D_{-i}]\right)(X_i) - Y_i\right]^2.$$

Of course, this approach has to be modified to work in a setting with dependent data. The intuitive reason is that the training samples D_{-i} and their corresponding validation singletons $\{(X_i, Y_i)\}$ are not independent any more, which may lead to over-adaption to the specific sample and hence a generally bad estimator for $\bar{h}_n$. A possible modification was suggested for instance by Györfi et al. (1989) and Burman et al. (1994). It reflects the idea to leave a gap in the training samples that grows with the sample size, and to take the validation sample out of these gaps. For a sufficiently large gap, this has the effect that dependencies between training and validation sample are small. The ith training sample now contains less variables,

$$D_{-i,b_n} = \left\{(X_1, Y_1),\dots,(X_{i-b_n}, Y_{i-b_n}), (X_{i+b_n}, Y_{i+b_n}),\dots,(X_n, Y_n)\right\},$$

and h is now chosen as the minimizer of

$$\frac{1}{n}\sum_{i=1}^{n} \left[\left(\hat{m}_{n,h}[D_{-i,b_n}]\right)(X_i) - Y_i\right]^2.$$

If we wanted to apply the cross validation technique in our model, we would have to encounter the fact that we are unable to evaluate $\hat{m}_{n,h}[D_{-i,b_n}](\lambda_i)$ because λ_i is a hidden variable. Hence, an iteration procedure has to be applied again. The modified cross validation functional to be minimized is then

$$\frac{1}{n}\sum_{i=1}^{n} \left[\left(\hat{m}_{n,h}[D_{-i,b_n}]\right)^{[i]}(0, Y_0,\dots,Y_{i-1}) - Y_i\right]^2.$$

Intuitively, it appears natural to let b_n tend to infinity while the fraction b_n/n should tend to zero, which had been remarked by Burman et al. (1994) referring to Györfi et al. (1989) in the context of kernel regression estimators.

At this point, we want to give the computational costs of adaptive parameter selection some consideration. As we cannot make use of Lemma 4.2.10 (ii), each in-

stance of determining a least squares spline estimation of m requires the solution of a constrained optimization problem. After an inquiry of methods in nonlinear constrained optimization, which is to a limited extend presented in Appendix A.6, we find that the standard approach to such a problem is the use of penalty based methods. These methods consist of a sequence of unconstrained optimization problems with gradually increasing penalty terms that penalize unfeasible solutions (cf. Definition A.6.8). Thus, every solution of the high-dimensional constrained optimization problem require a number of solutions of high-dimensional unconstrained problems. Due to the large number of variables, these problems themselves are highly complex, which is also discussed in Appendix A.6. Thus, the cross validation scheme lets the computational costs soar.

If we have a very high preference for containing the computational effort of the procedure, we might ignore possible over-fitting effects and restrict our efforts to avoiding too small choices for the constants: we would accept any sufficiently large choice that guarantees $m \in \mathscr{G}$. For the boundary values M and B_n, we would apply the ad hoc testing procedures that we presented previously. For the choice of L and l_n, we might adopt the strategy to choose a value as small as possible for $l_n \geq 1$ that is just about tolerable from a statistical point of view. Subsequently set $L = l_n$. This can be motivated as follows. In order to save computing power, the number of knots should be chosen as small as possible because it determines the dimension of the optimization problem associated to the least squares spline estimation. Very large choices of l_n mean that $\#A_n$ is large, and we come close to the complexity of a high dimensional global optimization problem, which is computationally extremely challenging (Appendix A.6). A statistically tolerable choice is one that does not contradict the requirement $\Delta_n \asymp n^{-1/3}$. The choice $L = l_n$ is motivated by Lemma 4.2.10 (ii). It is the smallest possible choice to ensure that we only need to solve an unconstrained optimization problem as opposed to a constrained one. This saves computational costs as well.

In view of limited computational resources, we pursued this strategy in our simulation study that we present in the next section.

4.2.2 A simulation study

Now that we have presented an approach to actually compute estimations of the regression function m from the data, it is in order to demonstrate the feasibility of this approach. We will present a simulation study in which we use the ideas of the fourth chapter to calculate estimations of the intensity function of a simulated count processes.

We briefly describe the basic procedure. We will generate a count process based on an intensity function of linear type with trigonometric perturbations

similar to the example Meister and Kreiß (2016) chose to illustrate their results. We declare a domain $[0,M]\times\{0,\dots,B-1\}$ and a specification of the parameters a,b,c,d,ν that determine the link function

$$m(\lambda,y)=\left(a+b\,\lambda+c\,(y\wedge(B-1))+d\left[\sin\left(\frac{2\pi}{\nu}\,\lambda\right)+\cos\left(\frac{2\pi}{\nu}\,(y\wedge(B-1))\right)\right]\right)\wedge M\,.$$

Next, we simulate the count process starting with $\lambda_1:=1$ and a random number $y_1\sim Poiss(\lambda_1)$. Then we proceed following the model equation. Given $(\lambda_{i-1},y_{i-1})\in[0,M]\times\mathbb{N}$, we calculate $m(\lambda_{i-1},y_{i-1})=:\lambda_i$ and generate a random variable $y_i\sim Poiss(\lambda_i)$. Since we want to work with a process in its stationary regime, we define a burn-in period of length n_0 and throw away all (x_i,y_i) with $i\le n_0$.

It is our aim to infer the true function m from the simulated process while we pretend not to have any information about the intensities $\{\lambda_i\}$. We use the least squares spline estimator to tackle this inference problem and assess the quality of our estimation using the empirical mean square error (MSE),

$$\frac{1}{n-n_0}\sum_{i=n_0+1}^{n}\left(m(x_i,y_i)-\hat{m}_n(x_i,y_i)\right)^2,$$

where n_0 is the length of the burn-in period.

In order to cap the size of the optimization problem, we wanted to restrict the domain with respect to the count variable to $\{0,1,2,3,4,5\}$. The parameter M was set to $M:=2$, which allows a reasonable approximation with 15 knots for every possible count value. This means that a total number of 72 variables need to be optimized. To avoid an overly homogeneous count process, the intensity function has to provide enough steepness. For our simulation, we selected the parameters $a=b=c=0.3$, $d=-0.1$ and $\nu=2$.

We estimated m on the domain $[0,2]\times\{0,\dots,5\}$ with the least squares splines estimator from Definition 4.2.8. The defining constants were set to $M=2,B_n=6$, and $L=l_n=10$; the underlying knot sequence was $\{-0.4,-0.2,0.0,\dots,2.0,2.2,2.4\}$. The set A_n was chosen to be an equidistant partition of the interval $[0,2]$ in 20 parts. According to Lemma 4.2.10 (ii), we were spared the handling of a constrained optimization problem to calculate the estimation. To solve the resulting optimization problem, we used the Genetic Algorithm (GA) in the form of the MATLAB® (version R2018a) function `ga()`. The main motivation for the choice of the algorithm was accessibility. The function `ga()` is an implementation of a global optimization algorithm that is able to process integer constraints.

In order to detect a possible over-fitting effect caused by too many candidate functions, we experimented with a larger specification of the set A_n by setting $A_n=[0,M]$. The resulting global optimization problem was solved using the GA, and additionally we used Particle Swarm Optimization (PSO) and Simulated

Annealing (SA), which are implemented in MATLAB® R2018a in form of the functions `particleswarm()` and `simulannealbnd()`, respectively. Again, the main considerations in the choice for these algorithm were practicability. The mentioned algorithms are easy to access and produce fairly well results in high-dimensional global optimization problems. A short description of the used algorithms can be found in Appendix A.6.

In total, we conducted 32 experiments treating two different data sets with four different approaches. The first 16 Experiments were carried out with a sample size of $n = 1050$ with burn-in period from index 1 to $n_0 = 50$ such that the effective sample consisted of 1000 realizations of the bivariate process. Experiments 1 – 4 were carried out using the GA and the set $A_n = \{0, 0.1, \ldots, 1.9, 2.0\}$. Experiments 5 – 16 are a succession of four applications of GA, PSO and SA, respectively, with $A_n = [0, M]$. In Experiments 17–32, we used an independent sample of $n = 150$ realizations with burn-in period from 1 to $n_0 = 50$. We conducted the experiments according to the same pattern as in Experiments 1–16. The repeated application of one algorithm to the same data set has the purpose to illustrate the inherent randomness of the used optimization procedures; a specific estimation is in general not reproducible from the same data set, and two independent runs yield two different estimations. All estimations were evaluated using the empirical mean square error. To get an idea about the scales, we computed the mean of all intensities, $\overline{\lambda}_n = \frac{1}{n}\sum_{i=1}^{n} \lambda_i$, as a reference value. The results are reported in detail in Table A.9.1 and Figures A.8.2–A.8.9. The experiments were carried out on an Intel® Core™ i7-2600 3.40 GHz machine with 16.0 GB RAM, using MATLAB® version R2018a.

Using the script shown in Listing A.7.1, data sets one and two were generated as independent realizations of 1050 and 150 elements of the bivariate process, respectively. The first 50 realizations were used as a burn in to approach the stationary regime. Illustrations of the data sets are given in Figure A.8.1. Estimations were computed using the scripts in Listing A.7.2. The results suggest that the strategy to use a discrete set A_n of coefficients and the Genetic Algorithm as a solver is preferable to the use of a non-discrete set A_n and global optimization. This impression is particularly clear in the experiments with sample size 1000. With smaller sample sizes the difference between the performances are less indicative. It is not clear whether the difference in the quality of the estimations with different specifications of A_n could be attributed to the over-fitting effect that was discussed earlier or simply to differences in the performances of the used algorithms. The different performance of GA-discrete and GA-global for the larger sample size hint at the presence of an over-fitting effect.

Lastly we want to emphasize that this simulation study has a purely illustrative character. Its primary purpose is to show that the estimation of a semi-

contractive link function, using the least squares spline estimator, is feasible. Numerical experiments that can serve as a solid base for conjectures about finite sample error bounds of the prescribed methods would require a lot more repetitions.

Chapter 5

Discussion

In this thesis we presented two main results. In a nonparametric integer-valued GARCH model for count data with hidden intensities, we proved the existence of an entirely nonparametric least squares estimator and demonstrated that it attains a rate of convergence that we suspect to be optimal up to logarithmic terms. The essential assumption in the model was a contractive condition which is a Lipschitz condition with constant smaller than one. We proposed two different approaches to least squares estimation that require different extents to which the contractive condition has to be satisfied. Furthermore, the approaches are distinguished by the sets of candidate functions from which the estimations are selected.

In the first approach taken in this thesis, we specified a class of functions $\mathscr{G}(M,B,L_1,L_2)$ that was assumed to contain the true function m. At the same time, the class $\mathscr{G}(M,B,L_1,L_2)$ served as the set of candidate functions, which means that the estimator minimizes the least squares functional over the class $\mathscr{G}(M,B,L_1,L_2)$. The methods we employed to prove asymptotic bounds for the $L_2(\pi)$-risk required a strong form of the contractive assumption. The proof relied heavily on a chaining argument which required the strong contractive condition to be valid for all functions in $\mathscr{G}(M,B,L_1,L_2)$. Furthermore, the assumption of full contractivity of m reduced the necessary effort to prove uniform mixing of the count process. We proposed a generalization of this first approach by letting the constants B grow gently with the sample size.

In the second approach, we restricted the set of candidate functions to a finite grid of functions. The size of the grid had to depend on the sample size in order to ensure that the accuracy of the estimator grows with the sample size. The circumstance that the estimation was selected from finitely many candidate functions allowed us to omit the chaining procedure, and we were able to prove

bounds for the empirical MSE of the so called approximate least squares estimator, using only the assumption that the true link functions has the semi-contractive property. For the set of candidate functions, we are then free to relax the conditions to Lipschitz continuity in the first argument with constants not necessarily smaller than one. While the first approach relied on exponential inequalities for mixing sequences, the second approach was based on martingale techniques. In both settings we were able show that the respective estimator attains the rate of convergence $n^{-1/3}$ up to a logarithmic term.

The rate reflects the smoothness of the link function and the dimension of its domain. In terms of mini-max theory, the optimal rate of convergence for nonparametric estimators with degree of smoothness β and domain dimension d is typically of order $n^{-\beta/(2\beta+d)}$. In their examination of a nonparametric GARCH(1,1) model, which is closely related to our count model, Meister and Kreiß (2016) proved the following lower bound for the rate of convergence. Assume that the true function m belongs to a Hölder class of monotone functions with smoothness parameter β over a domain with dimension $d = 2$ (ibid., page 3014). Then

$$\inf_{\{\hat{m}_n\}} \liminf_{n\to\infty} n^{2\beta/(2\beta+2)} \sup_{m} \mathbb{E} \int \left(m - \hat{m}_n\right)^2 d\pi > 0,$$

where π denotes the stationary distribution of the data generating process. In our case, the assumption that the functions are constant in the count variable from a threshold value B onward reduces the problem to a parametric one in the second component. As a consequence, nonparametric smoothing is only necessary with respect to the intensity variable, and the effective dimension is $d = 1$. The contractive property is a tighter version of Lipschitz continuity. Lipschitz functions are absolutely continuous and therefore differentiable almost everywhere. Of course, in this case the derivative is bounded almost everywhere by the Lischitz constant (Bass, 2013, page 128). Consequently, for any $y_0 \in \{0, \dots, B-1\}$,

$$\int_{[0,M]} \left|\partial_\lambda m(\lambda, y_0)\right|^2 d\lambda < \infty.$$

Thence we conclude that $m(\cdot, y)$ is an element of the Sobolev class of functions with smoothness parameter $\beta = 1$ (Tsybakov, 2008, page 13). Hence, in view of the classical results in nonparametric statistics, we suspect the optimal rate in our model to be of order $n^{-1/3}$. If this is true, the rates that we provide in Theorems 3.2.2 and 4.1.12 are optimal only up to a sub-polynomial term.

We shall briefly discuss the origin of the disturbing logarithmic terms that separate us from an optimal rate. A close inspection of the proofs uncovers several pitfalls that may be responsible for these terms. Let us start discussing the approach leading to Theorem 3.2.2. We examine the structure of the random

quantities that play the central role in the analysis. Recall that due to the iteration procedure, the essential random quantity driving our estimator are functions of the form

$$\big(Y_{i+1} - g^{[t]}(Y_{i-t},\dots,Y_i)\big)^2. \tag{5.1}$$

The parameter t indicates how many iterations we apply to approximate the unobservable intensities $\lambda_i \approx m^{[t]}(0,Y_{i-t},\dots,Y_i)$. We have investigated the mixing properties of these variables. They are measurable transformations of the lagged count process $\{\boldsymbol{Y}_{i-t}^{i+1}\}$. Hence, the dependence structure is driven by the mixing properties of this process. For the mixing coefficients we have acquired the bound

$$\phi^t(r) \le \min\{1, C L^{r-t}\} \tag{5.2}$$

for some positive constant C. The ultimate step in our argumentation was the application of Bernstein's inequality to sums of variables of the form (5.1). Here the first difficulties arise when we try to find bounds for the variance of sums of square integrable transformations $T(V_i) := T(\boldsymbol{Y}_{i-t}^{i+1})$ of the lagged process. Using the covariance inequalities for uniformly mixing process (Lemma A.2.1) and the bound (5.2), we obtain

$$\begin{aligned}
\operatorname{var}\Big(\sum_{i=t}^{n-1} T(V_i)\Big) &= \sum_{i_1,i_2=t}^{n-1} \operatorname{cov}(\varphi(V_{i_1}), T(V_{i_2})) \\
&\le 2 \sum_{r=0}^{n-1-t} \sum_{i_1=t}^{n-1-r} \operatorname{cov}(T(V_{i_1}), T(V_{i_1+r})) \\
&\le 4 \sum_{r=0}^{n-1-t} \sum_{i_1=t}^{n-1-r} \sqrt{\phi^t(r)}\, \mathbb{E}T^2(V_0) \\
&\lesssim n \underbrace{\sum_{r=0}^{\infty} \sqrt{\min\{1, C L^{r-t}\}}}_{O(t)} \\
&\lesssim n\, t\,.
\end{aligned}$$

The order t is owed to the absence of a better bound for indexes with small time gaps such that the mixing coefficients are not small enough. Recall that the iteration parameter t has to grow with the sample size, $t = t(n) \asymp \log n$. This means that

$$\operatorname{var}\Big(\sum_{i=t}^{n} T(V_i)\Big) = O(n \log n)\,.$$

This extra logarithmic factor inevitably carries over to any application of an exponential tail bound.

Another pitfall with the application of exponential inequalities in this setting is that they require the random variables under consideration to be bounded. As the count process is conditionally Poisson distributed, the variables (5.1) are unbounded. To remedy this deficiency, we need to introduce a cutoff threshold for the variables, which has to grow with n. This is another source for disturbing logarithmic terms. Last, we remark that the blocking technique, which was used to obtain i.i.d. sequences blocks of length q via a coupling, contributes a disturbance in form of the block length q as can be inspected in the proof of Lemma 3.2.17.

The martingale based approach leading to Theorem 4.1.12 avoids the technical problems related to the use of mixing properties. We are thus spared from the difficulties that arise from the variance bounds and blocking procedure which were necessary in the proof of Theorem 3.2.2. However, the fact remains that we deal with unbounded variables. This persists to be a problem. The exponential tail bound for martingales with unbounded increments (Lemma 4.1.11), which is well suited for this situation, requires a cutoff procedure as well. This cutoff is hidden in the definition of the term H_t^a, and it is responsible for the extra logarithmic factor in the result of Theorem 4.1.12. We suggest that a model with bounded count variables would allow for a result in the shape of Theorem 4.1.12 but without the disturbing factor.

We conclude the discussion of the theoretical results with a remark on the involved constants. A correct application of the proposed model would include a preceding specification of the model parameters that define domain and smoothness of the true function and the candidate functions. To avoid model misspecification, a sufficiently large choice of these constants is in order. Apart from this conditions, we have seen that the rate of convergence is unaffected by the specific choice of the constants. However, we strongly suspect that large constants deteriorate the estimator's performance in terms of finite sample error bounds.

In Chapter 4 we have proposed and discussed a possible way to approximately calculate a nonparametric least squares estimator from count data. The resulting least squares spline estimator can still be interpreted as a nonparametric estimator. Even if we construct the estimator from a finite sub-class, no parametric assumption on the true link function are imposed, which is the key feature of nonparametric inference. We experimentally approximated the least squares spline estimator from the count data of a simulated bivariate data generating process using different heuristic optimization algorithms. The outcomes differed in quality, but all estimations had reasonably small errors.

Computing a nonparametric estimation comes at the expense to solve computationally hard optimization problems and therefore requires more resources than a parametric model. The relative comfort of a model with less severe structural assumptions that is associated with the choice of a nonparametric approach has

to be weighted against two main disadvantages in form of a nonparametric rate of convergence, as opposed to the rate $n^{-1/2}$ in a correctly specified parametric model, and a much higher demand for computing power.

Appendix A

Supplementary material

A.1 Berbee's coupling

The next Lemma is taken from Doukhan et al. (1995). It is originally attributed to Berbee (1979).

LEMMA A.1.1 (Berbee's Lemma; Doukhan et al., 1995, page 406). *On a probability space $(E,\mathscr{E},P)$, let X and Y be two random variables taking theirs values in Borel spaces S_1 and S_2 respectively, and let U be a random variable with uniform distribution over $[0,1]$, independent of (X,Y). Then there exists a random variable $Y^* = f(X,Y,U)$, where f is a measurable function from $S_1 \times S_2 \times [0,1]$ into S_2 such that:*

1. *Y^* is independent of X and has the same distribution as Y;*
2. *$P\{Y^* \neq Y\} = \beta(\sigma(X),\sigma(Y))$.*

The next Lemma is adapted from Proposition 2 in Doukhan et al. (1995, page 407) which is stated there without a proof.

LEMMA A.1.2. *Let $q \in \mathbb{N}_+$ be a positive integer. Suppose that a two-sided sequence $\{V_t\}_{t\in\mathbb{Z}}$ of random vectors V_t with values is $(\mathbb{R}^d,\mathscr{B}^d)$ is defined on a probability space $(E,\mathscr{E},P)$, and let $\beta(q,t)$ and $\phi(q,t)$ denote the coefficients of absolute regularity and uniform mixing, respectively, between the sub σ-fields $\sigma\{V_s : s \leq t\}$ and $\sigma\{V_s : s \geq t+q\}$. As in Definitions 2.2.2 and 4.1.13, let $\beta(q) = \sup_t \beta(q,t)$ and $\phi(q) = \sup_t \phi(q,t)$. Suppose that the probability space $(E,\mathscr{E},P)$ is sufficiently rich such that there exists a sequence $\{U_t\}_{t\in\mathbb{N}_+}$ of independent, uniformly over $[0,1]$ distributed random variables that is independent of $\{V_t\}_{t\in\mathbb{Z}}$. Then there exists a one-sided random sequence $\{V_t^*\}_{t\in\mathbb{N}}$ on $(E,\mathscr{E},P)$ with the following properties:*

1. $P^{\left(V^*_{ql},V^*_{ql+1},\dots,V^*_{ql+q-1}\right)} = P^{\left(V_{ql},V_{ql+1},\dots,V_{ql+q-1}\right)}$ *for all* $l = 0,1,2,\dots$;

2. *The even blocks* $\left\{\left(V^*_{ql},V^*_{ql+1},\dots,V^*_{ql+q-1}\right)\colon l = 0,2,4,\dots\right\}$ *are mutually independent, as are the uneven blocks* $\left\{\left(V^*_{ql},V^*_{ql+1},\dots,V^*_{ql+q-1}\right)\colon l = 1,3,5\dots\right\}$;

3. $P\left\{\exists\, i = 0,\dots,n-1\colon V_i \neq V^*_i\right\} \leq \frac{n}{q}\beta(q) \leq \frac{n}{q}\phi(q)$, *for* $n \geq q$.

Proof. In the course of the proof, we will use the following two facts about the coefficients of absolute regularity. First, let $\mathscr{A}_0, \mathscr{A}, \mathscr{F}_0$, and $\mathscr{F}$ be sub σ-fields of $\mathscr{E}$ such that $\mathscr{A}_0 \subset \mathscr{A}$ and $\mathscr{F}_0 \subset \mathscr{F}$. Then $\beta(\mathscr{A}_0,\mathscr{F}_0) \leq \beta(\mathscr{A},\mathscr{F})$ (Bradley, 2007, page 68).

Second, suppose that $\mathscr{F}$, $\mathscr{G}$, $\mathscr{C}$, and $\mathscr{D}$ are sub σ-fields of $\mathscr{E}$, and the σ-fields $\sigma(\mathscr{F}\cup\mathscr{G})$ and $\sigma(\mathscr{C}\cup\mathscr{D})$ are independent. Then

$$\beta\big(\sigma(\mathscr{F}\cup\mathscr{C}),\sigma(\mathscr{G}\cup\mathscr{D})\big) \leq \beta(\mathscr{F},\mathscr{G}) + \beta(\mathscr{C},\mathscr{D}) - \beta(\mathscr{F},\mathscr{G})\cdot\beta(\mathscr{C},\mathscr{D})$$

(Bradley, 2007, Lemma 6.4 (b), page 194). If $\mathscr{D} = \{\emptyset, E\}$, then $\sigma(\mathscr{G}\cup\mathscr{D}) = \mathscr{G}$, $\beta(\mathscr{C},\mathscr{D}) = 0$, and the inequality reduces to

$$\beta\big(\sigma(\mathscr{F}\cup\mathscr{C}),\mathscr{G}\big) \leq \beta(\mathscr{F},\mathscr{G}). \tag{A.1}$$

Let us proceed with the main part of the proof. We split the sequence $(V_0,V_1,V_2,V_3,\dots)$ in blocks of length q,

$$\begin{aligned}\xi_i &:= \left(V_{2iq},\dots,V_{(2i+1)q-1}\right)\\ \eta_i &:= \left(V_{(2i+1)q},\dots,V_{2(i+1)q-1}\right),\end{aligned}$$

$i \in \mathbb{N}$. Then $\{V_i\}_{i\in\mathbb{N}} = \{\xi_i,\eta_i\}_{i\in\mathbb{N}}$. We shall construct a sequence $\{\xi^*_i\}_{i\in\mathbb{N}}$ of independent random variables that satisfy $P^{\xi^*_i} = P^{\xi_i}$ and $P\{\xi_i \neq \xi^*_i\} \leq \beta(q)$. Define $\xi^*_0 := \xi_0$. Note that $(\mathbb{R}^{d\cdot q},\mathscr{B}^{d\cdot q})$ is a Borel space (Bradley, 2007, page 11). According to Berbee's Lemma, there exists a measurable function such that the variable $\xi^*_1 = f_1(\xi_0,\xi_1,U_1)$ satisfies the following conditions:

(1) ξ^*_1 is independent of $\xi^*_0 = \xi_0$;
(2) $P^{\xi^*_1} = P^{\xi_1}$;
(3) $P\left\{\xi^*_1 \neq \xi_1\right\} = \beta\big(\sigma\{\xi_0\},\sigma\{\xi_1\}\big)$.

Since $\sigma\{\xi_1\} \subset \sigma\{V_t\colon t \geq 2q\}$, $\sigma\{\xi_0\} \subset \sigma\{V_t\colon t \leq q-1\}$, and the sequence $\{\beta(q)\}_{q\in\mathbb{N}_+}$ is non-increasing, it follows that $\beta\big(\sigma\{\beta_0\},\sigma\{\xi_1\}\big) \leq \beta(q+1,q-1) \leq \beta(q+1) \leq \beta(q)$. Therefore, $P\left\{\xi^*_1 \neq \xi_1\right\} \leq \beta(q)$.

In the ith step we do the following. Suppose we already have a vector $(\xi^*_0,\dots,\xi^*_{i-1})$ of independent random variables such that for $j \in \{1,\dots,i-1\}$ the variable ξ^*_j is independent of $(\xi^*_0,\dots,\xi^*_{j-1})$; $P^{\xi^*_j} = P^{\xi_j}$; and $\xi_j = f_j(\xi^*_0,\dots\xi^*_{j-1},\xi_j,U_j)$

for some measurable function f_j. Again, $(\mathbb{R}^{d\cdot q\cdot i}, \mathscr{B}^{d\cdot q\cdot i})$ and $(\mathbb{R}^{d\cdot q}, \mathscr{B}^{d\cdot q})$ are Borel spaces. According to Berbee's Lemma, there exists a random variable $\xi_i^* = f_i(\xi_0^*, \ldots, \xi_{i-1}^*, \xi_i, U_i)$ with the following properties:

(1) ξ_i^* is independent of $(\xi_0^*, \ldots, \xi_{i-1}^*)$;
(2) $P^{\xi_i^*} = P^{\xi_i}$;
(3) $P\{\xi_i^* \neq \xi_i\} = \beta\big(\sigma\{\xi_0^*, \ldots, \xi_{i-1}^*\}, \sigma\{\xi_i\}\big)$.

The construction of $(\xi_0^*, \ldots, \xi_{i-1}^*)$ implies that there exists a vector-valued, measurable function F_i such that $(\xi_0^*, \ldots, \xi_{i-1}^*) = F_j(\xi_0, \ldots, \xi_{i-1}, U_1, \ldots, U_{i-1})$. Hence, $\sigma\{\xi_0^*, \ldots, \xi_{i-1}^*\} \subset \sigma\{\xi_0, \ldots, \xi_{i-1}, U_1, \ldots, U_{i-1}\}$. Since the σ-fields $\sigma\{U_1, \ldots, U_{i-1}\}$ and $\sigma\{\xi_0, \ldots, \xi_i\}$ are independent, we can apply inequality A.1 with $\mathscr{F} = \sigma\{\xi_0, \ldots, \xi_{i-1}\}$, $\mathscr{C} = \sigma\{U_1, \ldots, U_{i-1}\}$, and $\mathscr{G} = \sigma\{\xi_i\}$ to conclude that

$$\begin{aligned}
\beta\big(\sigma\{\xi_0^*, \ldots, \xi_{i-1}^*\}, \sigma\{\xi_i\}\big) &\leq \beta(\sigma\{\xi_0, \ldots, \xi_{i-1}, U_1, \ldots, U_{i-1}\}, \sigma\{\xi_i\}) \\
&\leq \beta(\sigma\{\xi_0, \ldots, \xi_{i-1}\}, \sigma\{\xi_i\}) \\
&\leq \beta(\sigma\{V_t : t \leq 2iq - (q+1)\}, \sigma\{V_t : t \geq 2iq\}) \\
&\leq \beta(q+1, 2iq - (q+1)) \\
&\leq \beta(q).
\end{aligned}$$

Thence we deduce that $P\{\xi_i^* \neq \xi_i\} \leq \beta(q)$.

The sequence $\{\eta_i^*\}_{i\in\mathbb{N}}$ is constructed analogously. We finally obtain a sequence $(\xi_1^*, \eta_1^*, \xi_2^*, \eta_2^*, \ldots)$ such that $P^{\xi_i} = P^{\xi_i^*}$; $P^{\eta_i} = P^{\eta_i^*}$; and all ξ_i are mutually independent, as are the η_i. Moreover, $P\{\xi_i \neq \xi_i^*\} \leq \beta(q)$ and $P\{\eta_i \neq \eta_i^*\} \leq \beta(q)$. The proof is almost complete. We only remark that

$$\begin{aligned}
&P\{\exists\, i = 0, \ldots, n-1 : V_i \neq V_i^*\} \\
&\leq P\left\{\exists\, i = 0, \ldots, \frac{n+q}{2q} - 1 : \xi_i \neq \xi_i^* \text{ or } \eta_i \neq \eta_i^*\right\} \\
&\leq \frac{n}{q}\beta(q)
\end{aligned}$$

if $n \geq q$. For a uniformly mixing sequence, we obtain the bound with $\phi(q)$ since any uniformly mixing sequence is absolutely regular as well, with $\beta(q) \leq \phi(q)$ (Bradley, 2007, page 76). □

A.2 Covariance bounds for mixing processes

LEMMA A.2.1 (Doukhan, 1994, page 9). *Let $(E,\mathscr{E},P)$ be a probability space and $\mathscr{F}$ and $\mathscr{B}$ be two sub σ-fields of $\mathscr{E}$. Let X and Y be two random variables that are measurable with respect to $\mathscr{F}$ and $\mathscr{B}$ respectively. Then,*

$$|\mathrm{cov}(X,Y)| \le 2\left(\phi(\mathscr{F},\mathscr{B})\right)^{1/p}\left(EX^p\right)^{1/p}\left(EY^q\right)^{1/q},$$

for any $p,q \ge 1$ with $\frac{1}{p}+\frac{1}{q}=1$. Furthermore,

$$|\mathrm{cov}(X,Y)| \le 8\left(\alpha(\mathscr{F},\mathscr{B})\right)^{1/r}\left(EX^p\right)^{1/p}\left(EY^q\right)^{1/q},$$

for any $p,q,r \ge 1$ such that $\frac{1}{r}+\frac{1}{p}+\frac{1}{q}=1$.

A.3 A symmetrization lemma

LEMMA A.3.1 (Giné and Nickl, 2016, page 131). *On a probability space $(E,\mathscr{E},P)$, let $Y_1,\dots,Y_n$ be $\mathbb{R}^d$-valued random variables and $Y_1',\dots,Y_n'$ independent copies. Let $\mathscr{G}$ be a class of continuous functions $g\colon \mathbb{R}^d \to \mathbb{R}$ such that $\int |g(Y)|^2\,dP < \infty$. Then, for $t > s > 0$,*

$$P\left\{\sup_{g\in\mathscr{G}}\left|\sum_{i=1}^n g(Y_i)\right| > t\right\} \le \frac{P\left\{\sup_{g\in\mathscr{G}}\left|\sum_{i=1}^n [g(Y_i)-g(Y_i')]\right| > t-s\right\}}{1-\sup_{g\in\mathscr{G}} P\left\{\left|\sum_{i=1}^n g(Y_i)\right| > s\right\}}.$$

If additionally

$$\sup_{g\in\mathscr{G}} \mathrm{var}\Big(\sum_{i=1}^n g(Y_i)\Big) \le \frac{t^2}{8}, \tag{A.2}$$

then

$$P\left\{\sup_{g\in\mathscr{G}}\left|\sum_{i=1}^n g(Y_i)\right| > t\right\} \le 2P\left\{\sup_{g\in\mathscr{G}}\left|\sum_{i=1}^n [g(Y_i)-g(Y_i')]\right| > t/2\right\}.$$

Proof. Measurability of the supremum is secured by the fact that $\mathscr{G}$ is separable with respect to the uniform norm; consult the proof of Lemma 3.2.6 for the details of this argument.

Let $g^* \in \mathscr{G}$ such that $\left|\sum_{i=1}^n g^*(Y_i)\right| > t$ if such an element exists. Otherwise let g^* be any function from $\mathscr{G}$. Note that g^* depends on $Y_1,\dots,Y_n$. Hence, it is a $\mathscr{G}$-valued random element. Denote $\boldsymbol{Y} := (Y_1,\dots,Y_n)$. Then,

$$P\left\{\sup_{g\in\mathscr{G}}\left|\sum_{i=1}^n g(Y_i) - \sum_{i=1}^n g(Y_i')\right| > t-s\right\}$$

$$
\begin{aligned}
&\geq P\Big\{\big|\sum_{i=1}^n g^*(Y_i) - \sum_{i=1}^n g^*(Y_i')\big| > t-s\Big\} \\
&\geq \int P\Big\{\big|\sum_{i=1}^n g^*(Y_i) - \sum_{i=1}^n g^*(Y_i')\big| > t-s \Big| \boldsymbol{Y} = \boldsymbol{y}\Big\}\, P^{\boldsymbol{Y}}(d\boldsymbol{y}) \\
&\geq \int_{\{|\sum_{i=1}^n g^*(y_i)|>t\}} P\Big\{\big|\sum_{i=1}^n g^*(Y_i) - \sum_{i=1}^n g^*(Y_i')\big| > t-s \Big| \boldsymbol{Y} = \boldsymbol{y}\Big\}\, P^{\boldsymbol{Y}}(d\boldsymbol{y}) \\
&\geq \int_{\{|\sum_{i=1}^n g^*(y_i)|>t\}} P\Big\{\big|\sum_{i=1}^n g^*(Y_i')\big| \leq s \Big| \boldsymbol{Y} = \boldsymbol{y}\Big\}\, P^{\boldsymbol{Y}}(d\boldsymbol{y}) \\
&\geq P\Big\{\big|\sum_{i=1}^n g^*(Y_i)\big| > t\Big\} \inf_{g\in\mathscr{G}} P\Big\{\big|\sum_{i=1}^n g(Y_i')\big| \leq s\Big\} \\
&= P\Big\{\sup_{g\in\mathscr{G}}\big|\sum_{i=1}^n g(Y_i)\big| > t\Big\} \inf_{g\in\mathscr{G}} P\Big\{\big|\sum_{i=1}^n g(Y_i')\big| \leq s\Big\} \\
&= P\Big\{\sup_{g\in\mathscr{G}}\big|\sum_{i=1}^n g(Y_i)\big| > t\Big\}\Big(1 - \sup_{g\in\mathscr{G}} P\Big\{\big|\sum_{i=1}^n g(Y_i')\big| > s\Big\}\Big) \\
&\geq P\Big\{\sup_{g\in\mathscr{G}}\big|\sum_{i=1}^n g(Y_i)\big| > t\Big\}\Big(1 - \sup_{g\in\mathscr{G}} \frac{\operatorname{var}\sum_{i=1}^n g(Y_i')}{s^2}\Big).
\end{aligned}
$$

This proves the first assertion, and setting $s = \frac{t}{2}$ yields the second one. □

A.4 Tail bounds

All results in this section are to be understood with respect to a probability space $(E, \mathscr{E}, P)$ and, if necessary, a filtration $\{\mathscr{E}_n\}_{n\in\mathbb{N}}$ of sub σ-fields of $\mathscr{E}$.

LEMMA A.4.1 (Hoeffding's inequality; Giné and Nickl, 2016, page 114). *For $n \in \mathbb{N}_+$, let $X_1, \ldots, X_n$ be independent random variables with $EX_i = 0$, taking values in $[a_i, b_i]$ with $-\infty < a_i < 0 \leq b_i < \infty$, for all $i \in \{1, \ldots, n\}$. Then, for all $t \geq 0$,*

$$
P\Big\{\big|\sum_{i=1}^n X_i\big| > t\Big\} \leq 2e^{-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}}
$$

COROLLARY A.4.2. *Let $\varepsilon_1, \ldots, \varepsilon_n$ be independent Rademacher variables, i.e. $P\{\varepsilon_i = 1\} = \frac{1}{2} = P\{\varepsilon_i = -1\}$, and let α_i be real numbers. It holds that*

$$
P\Big\{\frac{1}{n}\big|\sum_{i=1}^n \alpha_i \varepsilon_i\big| > t\Big\} \leq 2\exp\Big(-\frac{(nt)^2}{2\sum_{i=1}^n \alpha_i^2}\Big)
$$

LEMMA A.4.3 (Bernstein's inequality; Giné and Nickl, 2016, page 118). *For $n \in \mathbb{N}_+$, suppose that the random variables $\eta_1, \ldots, \eta_n$ are independent, have bounded ranges $[-b, b]$, and satisfy $E\eta_i = 0$ for all $i \in \{1, \ldots, n\}$. For non-negative real numbers x,*

$$P\{\eta_1 + \ldots + \eta_n > x\} \leq e^{-\frac{1}{2}\frac{x^2}{\sigma_n^2 + bx/3}},$$

where $\sigma_n^2 \geq \operatorname{var}(\eta_1 + \ldots + \eta_n)$.

The proof of the following integrated version of Bernstein's inequality is adapted from Doukhan et al. (1995, page 408).

LEMMA A.4.4 (Bernstein-type bound). *Let $X_1, \ldots, X_m$ be random variables that satisfy the tail bound*

$$P\{|X_i| > x\} \leq 2e^{-\frac{1}{2}\frac{x^2}{b + ax}},$$

for all non-negative real numbers x and fixed constants $a, b \geq 0$. Then there exists a universal constant $C > 0$ such that

$$E\Big[\max_{1 \leq i \leq m} |X_i|\Big] \leq C\left(\sqrt{b \log m} + a \log m\right).$$

Proof.

$$\begin{aligned}
E\Big[\max_{1 \leq i \leq m} |X_i|\Big] &= \int_0^\infty P\Big\{\max_{i \leq m} |X_i| > x\Big\}\, dx \\
&= \int_0^\infty 1 \wedge P\Big\{\max_{i \leq m} |X_i| > x\Big\}\, dx \\
&\leq \int_0^\infty 1 \wedge \sum_{i=1}^m P\Big\{|X_i| > x\Big\}\, dx \\
&\leq \int_0^\infty 1 \wedge \sum_{i=1}^m 2e^{-\frac{1}{2}\frac{x^2}{b + ax}}\, dx \\
&\leq \int_0^\infty 1 \wedge \sum_{i=1}^m 2e^{-\frac{1}{2}\frac{x^2}{2(b \vee ax)}}\, dx \\
&\leq \int_0^\infty 1 \wedge \sum_{i=1}^m 2\Big(e^{-\frac{1}{4}\frac{x^2}{b}} + e^{-\frac{1}{4}\frac{x}{a}}\Big)\, dx \\
&\leq \int_0^\infty 1 \wedge \left(2m\, e^{-\frac{x^2}{4b}}\right) dx + \int_0^\infty 1 \wedge \left(2m\, e^{-\frac{x}{4a}}\right) dx \\
&= \int_0^\infty e^{-\left(\frac{x^2}{4b} - \log(2m)\right)^+}\, dx + \int_0^\infty e^{-\left(\frac{x}{4a} - \log(2m)\right)^+}\, dx \\
&\leq \sqrt{4b \log(2m)} + \int_0^\infty e^{-\frac{x^2}{4b}}\, dx + 4a \log(2m) + \int_0^\infty e^{-x/(4a)}\, dx \\
&\leq \sqrt{4b \log(2m)} + \sqrt{b\pi} + 4a \log(2m) + 4a\,.
\end{aligned}$$

□

LEMMA A.4.5. *Let $Y \sim Poiss(\lambda)$. For non-negative real numbers x,*

$$P\{Y - \lambda > x\} \leq e^{-\frac{x^2}{2\lambda + 2x/3}}.$$

Proof. The moment generating function of a centered Poisson random variable is given by

$$Ee^{s(Y-\lambda)} = e^{\lambda(e^s - 1 - s)}.$$

Therefore, according to Giné and Nickl (2016, page 116), the claim follows. □

DEFINITION A.4.6. Let $M = \{M_n\}_{n\in\mathbb{N}}$ be a square integrable martingale with respect to the filtration $\{\mathscr{E}_n\}$. The *predictable quadratic variation* $\langle M\rangle$ is defined as the, up to indistinguishability, unique increasing process $\{\langle M\rangle_n\}_{n\in\mathbb{N}}$ such that $\langle M\rangle_0 = 0$ a.s. and

(1) for every $n > 0$ the random variable $\langle M\rangle_n$ is $\mathscr{E}_{n-1}$-measurable;

(2) the process $\{m_n\}_{n\in\mathbb{N}}$, given by $m_n = M_n^2 - \langle M\rangle_n$, is a martingale with respect to the filtration $\{\mathscr{E}_n\}$.

The predictable quadratic variation is well defined according to the Doob-Meyer decomposition (Karatzas and Shreve, 1991, page 21). The jump process $\Delta M = \{\Delta M_n\}_{n\in\mathbb{N}}$ is defined by the increments $\Delta M_n = M_n - M_{n-1}$ if $n > 0$, and $\Delta M_0 = M_0$ otherwise.

LEMMA A.4.7 (Dzhaparidze and van Zanten, 2001, page 110). *Let $\{M_n, \mathscr{E}_n\}_{n\in\mathbb{N}}$ be a square integrable martingale, and for $a > 0$ let the process $\{H_n^a\}_{n\in\mathbb{N}}$ be defined by*

$$H_n^a := \langle M\rangle_n + \sum_{k=0}^{n} (\Delta M_k)^2 \mathbb{1}_{\{|\Delta M_k| > a\}}.$$

Then for every finite stopping time τ, and any real numbers $a, z, L > 0$,

$$P\Big\{\max_{n\leq\tau} |M_n| > z\,; H_\tau^a \leq L\Big\} \leq 2\exp\Big(-\frac{1}{2}\frac{z^2}{L}\psi\big(az/L\big)\Big),$$

where the function ψ is given by

$$\psi(x) = \frac{2}{x^2}\int_0^x \log(1+y)\, dy.$$

It satisfies $\psi(x) \geq \frac{1}{1+x/3}$ for $x \geq -1$.

REMARK A.4.8. The original version of the previous lemma was formulated for continuous time martingales $\{M_t, \mathscr{E}_t\}_{t\geq 0}$ that have right-continuous trajectories

with left-hand limits. However, defining $M_t = \sum_{n=0}^{\infty} M_n \mathbb{1}_{[n,n+1)}(t)$, the result can be adapted to the case of discrete time.

A.5 Some properties of splines

In the next lemma, we collect the main features of B-splines for our purposes. Our sources are Powell (1981), Dierckx (1995), Györfi et al. (2002), and Lyche and Mørken (2008).

LEMMA A.5.1. *Let $\boldsymbol{\xi}$ be a set of $l+2k+1$ points in ascending order,*

$$\xi_{-k} < \ldots < \xi_0 < \ldots < \xi_l < \ldots < \xi_{l+k}\,,$$

with $\xi_0 = a$ and $\xi_l = b$. For $p \in \{-k,\ldots,l-1\}$, let the B-spline $B_{p,k}$ be given by (4.6)*. Then the following statements hold:*

(i) The B-splines are non-negative, and $\operatorname{supp} B_{p,k} = [\xi_p, \xi_{p+k+1}]$.

(ii) The following recurrence relation holds:

$$B_{p,k}(x) = \frac{(x-\xi_p)B_{p,k-1}(x) + (\xi_{p+k+1}-x)B_{p+1,k-1}(x)}{(\xi_{p+k+1}-\xi_p)}\,.$$

(iii) The set of B-splines restricted to the interval $[a,b]$,

$$\left\{B_{p,k}\big|_{[a,b]} \colon p \in \{-k,\ldots,l-1\}\right\},$$

form a basis of the vector space $\mathscr{S}(k;\xi_0,\ldots,\xi_l)$.

(iv) The derivative of the B-splines are given by

$$\partial_x B_{p,k}(x) = k\,\frac{B_{p,k-1}(x) - B_{p+1,k-1}(x)}{\xi_{p+k+1}-\xi_p}\,.$$

Consequently, $\partial_x N_{p,k}(x) = k\frac{N_{p,k-1}}{\xi_{p+k}-\xi_p} - k\frac{N_{p+1,k-1}}{\xi_{p+1+k}-\xi_{p+1}}$, and for $x \in [0,M]$

$$\partial_x \sum_{p=-k}^{l-1} \alpha_p N_{p,k}(x) = \sum_{p=-k+1}^{l-1} k\,\frac{\alpha_p - \alpha_{p-1}}{\xi_{p+k}-\xi_p} N_{p,k-1}(x)\,.$$

(v) Let $s = \sum_{p=-k}^{l-1} \alpha_p N_{p,k}(x)$ be an element of $\mathscr{S}(k;\xi_0,\ldots,\xi_l)$. Then the maximal value and the minimal value of s are bounded by the maximal and minimal coefficient, respectively:

$$\min_p \alpha_p \le s(x) \le \max_p \alpha_p\,.$$

(vi) The B-splines of order $k \geq 1$, normalized by the factor $(\xi_{p+k+1} - \xi_p)$, form a partition of unity on $[0, M]$, i.e.

$$\sum_{p=-k}^{l-1} (\xi_{p+k+1} - \xi_p) B_{p,k}(x) = 1 \quad \textit{for all } x \in [0, M].$$

Proof. (i) The first assertion follows from Theorem 19.1 of Powell (1981, page 230). For the second claim, we follow the argument outlined by Powell (1981, page 229). From the definition it follows immediately that $B_{p,k}(x) = 0$ for all $x < \xi_p$. We shall show that additionally

$$\sum_{i=p}^{p+k+1} d_i (x - \xi_i)_+^k = 0$$

for all $x \in [\xi_{p+k+1}, b]$, with $d_i = \prod_{\substack{j=p \\ j \neq i}}^{p+k+1} \frac{1}{(\xi_j - \xi_i)}$. For that purpose, recall that the Lagrangian interpolation polynomial of degree $k+1$ for interpolating a continuous function f at the $k+2$ points $\xi_p, \ldots, \xi_{p+k+1}$ is given by

$$Pf[x] := \sum_{i=p}^{p+k+1} f(\xi_i) \prod_{\substack{j=p \\ j \neq i}}^{p+k+1} \frac{(x - \xi_j)}{(\xi_i - \xi_j)},$$

and that Pf is unique among all polynomials with degree $k+1$. Therefore, $Pf = f$ if f itself is such a polynomial. In particular, for any $n \in \{0, \ldots, k+1\}$, we have $Px^n = x^n$. Whence we infer that

$$\begin{aligned}
x^n &= \sum_{i=p}^{p+k+1} (\xi_i)^n \prod_{\substack{j=p \\ j \neq i}}^{p+k+1} \frac{(x - \xi_j)}{(\xi_i - \xi_j)} \\
&= \sum_{i=p}^{p+k+1} \frac{(\xi_i)^n}{\prod_{\substack{j=p \\ j \neq i}}^{p+k+1} (\xi_i - \xi_j)} \prod_{\substack{j=p \\ j \neq i}}^{p+k+1} (x - \xi_j) \\
&= \sum_{i=p}^{p+k+1} \frac{(\xi_i)^n}{\prod_{\substack{j=p \\ j \neq i}}^{p+k+1} (\xi_i - \xi_j)} \sum_{q=0}^{k+1} x^{k+1-q} \underbrace{\sum_{\substack{\{r_1, \ldots, r_q\} \\ \subset \{p, \ldots, p+k+1\} \setminus \{i\}}} (-1)^q \xi_{r_1} \ldots \xi_{r_q}}_{\Sigma_\emptyset := 1} \\
&= \sum_{q=0}^{n} x^{k+1-q} \sum_{i=p}^{p+k+1} \frac{(\xi_i)^n}{\prod_{\substack{j=p \\ j \neq i}}^{p+k+1} (\xi_i - \xi_j)} \sum_{\substack{\{r_1, \ldots, r_q\} \\ \subset \{p, \ldots, p+k+1\} \setminus \{i\}}} (-1)^q \xi_{j_1} \ldots \xi_{i_q} \\
&=: \sum_{q=0}^{n} x^{k+1-q} \, d'_{k+1-q,n}
\end{aligned}$$

Comparing the coefficients on both sides, we see that d'_{k+1-q} must be equal to zero

for all $q \neq (k+1)-n$. For $n < k+1$ and $q = 0$, this means that

$$0 = d'_{k+1,n} = \sum_{i=p}^{p+k+1} \frac{(\xi_i)^n}{\prod_{\substack{j=p \\ j\neq i}}^{p+k+1}(\xi_i - \xi_j)} = \sum_{i=p}^{p+k+1} (-1)^{k+1} d_i \, (\xi_i)^n$$

For $x > \xi_{p+k+1}$, we obtain

$$\begin{aligned}(-1)^{k+1} \sum_{i=p}^{p+k+1} d_i \, (x-\xi_i)_+^k &= (-1)^{k+1} \sum_{i=p}^{p+k+1} d_i \, (x-\xi_i)^k \\ &= \sum_{n=0}^{k} \binom{k}{n} x^{k-n} \sum_{i=p}^{p+k+1} (-1)^{k+1} d_i \, (\xi_i)^n \\ &= 0\,.\end{aligned}$$

This gives the second claim.

(ii) We follow Powell (1981, page 234). Rewrite the formula for the B-spline as

$$B_{p,k}(x) = \sum_{j=p}^{p+k+1} \Delta_{p,k}^{(j)}(x),$$

where the incremental functions $\Delta_{p,k}^{(j)}$ have the form,

$$\Delta_{p,k}^{(j)}(x) = \begin{cases} 0 & \text{if } x \leq \xi_j \\ \frac{(x-\xi_j)^k}{\prod_{\substack{i=p \\ i\neq j}}^{p+k+1}(\xi_i - \xi_j)} & \text{if } x \geq \xi_j\,. \end{cases}$$

Then, for $x \geq \xi_p$, the right hand side of the iteration formula reads

$$\begin{aligned}&\frac{(x-\xi_p)\sum_{j=p}^{p+k} \Delta_{p,k-1}(j)(x) + (\xi_{p+k+1}-x)\sum_{j=p+1}^{p+k+1} \Delta_{p+1,k-1}^{(j)}(x)}{\xi_{p+k+1}-\xi_p} \\ &= \frac{x-\xi_p}{\xi_{p+k+1}-\xi_p} \Delta_{p,k-1}^{(p)}(x) \\ &\quad + \sum_{j=p+1}^{p+k} \frac{(x-\xi_p)\Delta_{p,k-1}^{(j)} + (\xi_{p+k+1}-x)\Delta_{p+1,k-1}^{(j)}(x)}{\xi_{p+k+1}-\xi_p} \\ &\quad + \frac{\xi_{p+k+1}-x}{\xi_{p+k+1}-\xi_p} \Delta_{p+1,k-1}^{(p+k+1)}(x) \\ &= \frac{(x-\xi_p)^k}{\xi_{p+k+1}-\xi_p} \prod_{\substack{i=p \\ i\neq p}}^{p+k} \frac{1}{\xi_i - \xi_p} \\ &\quad + \sum_{j=p+1}^{p+k} \mathbb{1}_{\{x \geq \xi_j\}} \frac{(x-\xi_p)\frac{(x-\xi_j)^{k-1}}{\prod_{\substack{i=p \\ i\neq j}}^{p+k}(\xi_i-\xi_j)} + (\xi_{p+k+1}-x)\frac{(x-\xi_j)^{k-1}}{\prod_{\substack{i=p+1 \\ i\neq j}}^{p+k+1}(\xi_i-\xi_j)}}{\xi_{p+k+1}-\xi_p}\end{aligned}$$

$$
\begin{aligned}
&+\mathbb{1}_{\{x\geq\xi_{p+k+1}\}}\frac{(x-\xi_{p+k+1})^k}{\xi_p-\xi_{p+k+1}}\prod_{\substack{i=p+1\\ i\neq p+k+1}}^{p+k+1}\frac{1}{\xi_i-\xi_{p+k+1}}\\
&=\frac{(x-\xi_p)^k}{\prod_{\substack{i=p\\ i\neq p}}^{p+k+1}(\xi_i-\xi_p)}\\
&\quad+\sum_{j=p+1}^{p+k}\mathbb{1}_{\{x\geq\xi_j\}}\frac{(x-\xi_j)^k}{\prod_{\substack{i=p\\ i\neq j}}^{p+k+1}(\xi_i-\xi_j)}\cdot\frac{(x-\xi_p)(\xi_{p+k+1}-\xi_j)-(x-\xi_{p+k+1})(\xi_p-\xi_j)}{(\xi_{p+k+1}-\xi_p)(x-\xi_j)}\\
&\quad+\mathbb{1}_{\{x\geq\xi_{k+p+1}\}}\frac{(x-\xi_{p+k+1})^k}{\prod_{\substack{i=p\\ i\neq p+k+1}}^{p+k+1}(\xi_i-\xi_{p+k+1})}\\
&=\sum_{i=p}^{p+k+1}\Delta_{p,k}^{(j)}(x)
\end{aligned}
$$

The last equation follows from

$$
\begin{aligned}
&(x-\xi_p)(\xi_{p+k+1}-\xi_j)-(x-\xi_{p+k+1})(\xi_p-\xi_j)\\
&=(x-\xi_j)(\xi_{p+k+1}-\xi_j)+(\xi_j-\xi_p)(\xi_{p+k+1}-\xi_j)\\
&\quad-(x-\xi_j)(\xi_p-\xi_j)-(\xi_j-\xi_{p+k+1})(\xi_p-\xi_j)\\
&=(\xi_{p+k+1}-\xi_p)(x-\xi_j),
\end{aligned}
$$

which means that

$$
\frac{(x-\xi_p)(\xi_{p+k+1}-\xi_j)-(x-\xi_{p+k+1})(\xi_p-\xi_j)}{(\xi_{p+k+1}-\xi_p)(x-\xi_j)}=1.
$$

This proves the iteration formula.

(iii) Confer Powell (1981, page 232). The set $\{B_{p,k}\}_{p=-k}^{l-1}$ has the right number of elements. In order to show that this set is indeed a basis of $\mathscr{S}(k;\xi_0,\ldots,\xi_l)$, we need to establish its linear independence. Let

$$
s(x)=\sum_{p=-k}^{l-1}\alpha_p B_{p,k}(x)
$$

be a linear combination such that $s(x)=0$ for all $x\in[\xi_0,\xi_l]$. Furthermore, we know by the definition of B-splines that $s(x)=0$ for $x\leq\xi_{-k}$. Since $s\in C^{(k-1)}$ and on any sub interval $[\xi_j,\xi_{j+1}]$ it is a polynomial of degree k, we conclude that the $k-1$ fold derivative $\partial_x^{k-1}s$ is piece wise linear. That together with $s\equiv 0$ outside (ξ_{-k},ξ_0) implies that $\partial_x^{k-1}s$ does not change signs on $[\xi_{-k},\xi_{-k+1}]$ and $[\xi_{-1},\xi_0]$, respectively. Therefore $\partial_x^{k-1}s$ has less than $k-2$ changes of sign on the interval (ξ_{-k},ξ_0). Hence, $\partial_x^{k-2}s$ has less than $k-2$ local extreme points in (ξ_{-k},ξ_0). On the other hand, $\partial_x^{k-2}s\equiv 0$ outside (ξ_{-k},ξ_0) and has therefore at most $k-3$ changes in

sign. Inductively, $\partial_x s$ has no change in sign and s no extreme point in the interval (ξ_{-k}, ξ_0). It must therefore be constantly zero on $[\xi_{-k}, \xi_l]$.

Now assume that there exist integers $p_1 < \ldots < p_n$ such that $\alpha_{p_i} \neq 0$ but $\alpha_p = 0$ for any $p \notin \{p_1, \ldots, p_n\}$. Since on $[\xi_{p_1}, \xi_{p_1+1}]$ the only non zero B-spline from the set $\{B_{p_i,k}\}_{i=1}^n$ is $B_{p_1,k}$, we obtain for $x \in [\xi_{p_1}, \xi_{p_1+1}]$

$$s(x) = \alpha_{p_1} B_{p_1,k}(x) \neq 0\,.$$

This is a contradiction. Therefore, all α_p must be zero and the set $\{B_{p,k}\}_{p=-k}^{l-1}$ is linearly independent.

(iv) For the proof we refer to Györfi et al. (2002, page 265).

(v) The claim follows easily from the partition of unity property:

$$\sum_{p=-k}^{l-1} \alpha_p \, N_{p,k}(x) \leq \max_p \alpha_p \sum_{p=-k}^{l-1} N_{p,k}(x) = \max_p \alpha_p\,.$$

The lower bound is proven analogously.

(vi) Cf. Powell (1981, page 242). For $k = 0$, the normalized B-splines are piece wise constant functions:

$$\begin{aligned}(\xi_{p+1} - \xi_p) B_{p,0} &= \frac{(\xi_{p+1} - \xi_p)}{(\xi_{p+1} - \xi_p)} \mathbb{1}_{(\xi_p, \infty)}(x) + \frac{(\xi_{p+1} - \xi_p)}{(\xi_p - \xi_{p+1})} \mathbb{1}_{(\xi_{p+1}, \infty)}(x) \\ &= \mathbb{1}_{(\xi_p, \xi_{p+1}]}(x)\,.\end{aligned}$$

By the recurrence formula, the first order B-splines are piece wise linear:

$$(\xi_{p+2} - \xi_p) B_{p,1}(x) = \begin{cases} \frac{x - \xi_p}{\xi_{p+1} - \xi_p} & \text{if } x \in (\xi_p, \xi_{p+1}] \\ 1 - \frac{x - \xi_{p+1}}{\xi_{p+2} - \xi_{p+1}} & \text{if } x \in (\xi_{p+1}, \xi_{p+2}] \\ 0 & \text{else}\,. \end{cases}$$

Thus, the only B-splines of order $k = 1$ that overlap at $x \in [\xi_i, \xi_{i+1}]$ are $B_{i-1,1}$ and $B_{i,1}$. This yields for $x \in [\xi_i, \xi_{i+1}]$ that

$$\begin{aligned}\sum_{p=-1}^{l-1} (\xi_{p+2} - \xi_p) B_{p,1}(x) &= (\xi_{i+1} - \xi_{i-1}) B_{i-1,1}(x) + (\xi_{i+2} - \xi_i) B_{i,1} \\ &= \frac{\xi_{i+1} - x}{\xi_{i+1} - \xi_i} + \frac{x - \xi_i}{\xi_{i+1} - \xi_i} \\ &= 1\,.\end{aligned}$$

Now assume that the statement is true for $k - 1$, i.e.

$$\sum_{p=-k+1}^{l-1} (\xi_{p+k} - \xi_p) B_{p,k-1}(x) = 1$$

Note that the function $B_{-k,k-1}$ is identically zero outside the interval (ξ_{-k},ξ_0) and the same is true for $B_{l,k-1}$ outside the interval (ξ_l,ξ_{l+k}). Again by the recurrence formula, we obtain for $x\in[\xi_0,\xi_l]$

$$
\begin{aligned}
&\sum_{p=-k}^{l-1}(\xi_{p+k+1}-\xi_p)B_{p,k}(x)\\
&=\sum_{p=-k}^{l-1}\frac{x-\xi_p}{\xi_{p+k}-\xi_p}(\xi_{p+k}-\xi_p)B_{p,k-1}(x)+\frac{\xi_{p+k+1}-x}{\xi_{p+k+1}-\xi_{p+1}}(\xi_{p+k+1}-\xi_{p+1})B_{p+1,k-1}(x)\\
&=\frac{x-\xi_{-k}}{\xi_0-\xi_{-k}}(\xi_0-\xi_{-k})B_{-k,k-1}(x)\\
&\quad+\sum_{p=-k+1}^{l-1}\frac{x-\xi_p}{\xi_{p+k}-\xi_p}(\xi_{p+k}-\xi_p)B_{p,k-1}(x)\\
&\quad+\sum_{p=-k+1}^{l-1}\frac{\xi_{p+k}-x}{\xi_{p+k}-\xi_p}(\xi_{p+k}-\xi_p)B_{p,k-1}(x)\\
&\quad+\frac{\xi_{l+k}-\xi_l}{\xi_{l+k}-\xi_l}(\xi_{l+k}-\xi_l)B_{l,k-1}(x)\\
&=\sum_{p=-k+1}^{l-1}(\xi_{p+k}-\xi_p)B_{p,k-1}(x)\\
&=1.
\end{aligned}
$$

The last equality follows from the induction hypothesis. □

A.6 A glance at global and combinatorical optimization

A.6.1 Unconstrained global optimization: information based complexity and worst case analysis

We want to present some well known results about the complexity analysis of global optimization algorithms. For that sake, let us introduce the basic notions used in the theory of information based complexity. We use the notation of Traub et al. (1988). Let $F_0\subset F_1$ be a symmetric and convex subset of a space of functions. Given some function $f\in F_0$, our aim is to approximate a characteristic quantity of this function. Formally, this characteristic quantity is given by a parameter operator $S\colon F_0\to F_2$ taking values in a normed space F_2 which we call the parameter space. We want to approximate $S(f)$, given some incomplete information about the function f.

We shall suppose that this information is given by the value of the function (or its derivatives) at certain points. It is delivered by an information operator N,

$$N(f)=N_f(f)=\left[L_1(f),L_{2,f}(f;y_1),\dots,L_{n(f),f}(f;y_1,\dots,y_{n-1})\right],$$

with $y_i = L_{i,f}(f; y_1, \ldots, y_{i-1})$. The operators $L_{i,f}(\,\cdot\,; y_1, \ldots, y_{i-1})$ are assumed to be linear. Note that the ith information operator depends on f in that it includes the knowledge about $y_1, \ldots, y_{i-1}$ in the choice of L_i. Therefore, an information N with such a structure is called adaptive or sequential information. Adaptive information means that the point of evaluation of f (or ∂f) is chosen according to the previous evaluations. In contrast, non adaptive information means that the evaluation points are set in advance, following a predetermined design. An instance of a non-adaptive information operator is N_{f^*} for some fixed function f^*. This means that the function f^* determines the schedule of the evaluation points. In this case,

$$N_{f^*}(f) = \left[L_1(f), L_{2,f^*}(f; y_1^*), \ldots, L_{n,f^*}(f; y_1^*, \ldots, y_{n-1}^*)\right],$$

with $y_i^* = L_{i,f^*}(f^*; y_1^*, \ldots, y_{i-1}^*)$. Hence, we can write $N_{f^*}(f) = \left[L_1^*(f), \ldots, L_n^*(f)\right]$. From this form it is visible that N_{f^*} does not adapt to the information given by f.

In order to approximate $S(f)$ on the basis of the available information, we use algorithms. They are formally introduced as mappings, $\phi\colon N(F_0) \to F_2$, from the set of attainable information into the parameter space. In analogy to the mini-max approach in nonparametric statistics, the radius and diameter of information are defined as

$$\begin{aligned} r(N; S, F_0) &= \inf_{\phi \in \Phi(N)} \sup_{f \in F_0} \|S(f) - \phi(N(f))\|_{F_2}, \\ d(N; S, F_0) &= \sup_{f \in F_0} \sup_{\substack{g \in F_0 \\ N(f) = N(g)}} \|S(f) - S(g)\|_{F_2} \end{aligned}$$

respectively, where $\Phi(N)$ denotes the class of algorithms using the information N. The quantity $e(\phi, N; S, F_0) := \sup_{f \in F_0} \|S(f) - \phi(N(f))\|_{F_2}$ is called the error of the algorithm with respect to S and F_0. We give a short proposition to get acquainted with the terms. It contains a basic relation between radius and diameter of information (Wasilkowski, 1984, inequality (2.16)).

PROPOSITION A.6.1. *The radius and diameter of information satisfy the relation*

$$\frac{1}{2} d(N; S, F_0) \le r(N; S, F_0) \le d(N; S, F_0).$$

Proof. We start with the second inequality.

$$\begin{aligned} \sup_{f \in F_0} \sup_{\substack{g \in F_0 \\ N(f) = N(g)}} \|S(f) - S(g)\|_{F_2} &= \sup_{y \in N(F_0)} \sup_{\substack{g \in F_0 : \\ N(g) = y}} \sup_{\substack{f \in F_0 : \\ N(f) = y}} \|S(f) - S(g)\|_{F_2} \\ &\ge \sup_{y \in N(F_0)} \inf_{\substack{g \in F_0 : \\ N(g) = y}} \sup_{\substack{f \in F_0 : \\ N(f) = y}} \|S(f) - S(g)\|_{F_2} \end{aligned}$$

For every $y \in N(F_0)$ and any $\varepsilon > 0$, there exists an element $s^*(y,\varepsilon) \in S\big(N^{-1}\{y\}\big) \subset F_2$ in the parameter space such that

$$\inf_{\substack{g\in F_0:\\ N(g)=y}} \sup_{\substack{f\in F_0:\\ N(f)=y}} \|S(f)-S(g)\|_{F_2} > \sup_{\substack{f\in F_0:\\ N(f)=y}} \|S(f)-s^*\|_{F_2} - \varepsilon\,.$$

Defining the mapping $\phi_\varepsilon^* : N(F_0) \to F_2$ by $\phi_\varepsilon^*(y) := s^*(y,\varepsilon)$, we obtain

$$\begin{aligned}
\sup_{f\in F_0} \sup_{\substack{g\in F_0\\ N(f)=N(g)}} \|S(f)-S(g)\|_{F_2} &> \sup_{y\in N(F_0)} \sup_{\substack{f\in F_0:\\ N(f)=y}} \|S(f)-\phi_\varepsilon^*(y)\|_{F_2} - \varepsilon \\
&= \sup_{y\in N(F_0)} \sup_{\substack{f\in F_0:\\ N(f)=y}} \|S(f)-\phi_\varepsilon^*(N(f))\|_{F_2} - \varepsilon \\
&= \sup_{f\in F_0} \|S(f)-\phi_\varepsilon^*(N(f))\|_{F_2} - \varepsilon \\
&\geq \inf_{\phi\in\Phi(N)} \sup_{f\in F_0} \|S(f)-\phi(N(f))\|_{F_2} - \varepsilon\,.
\end{aligned}$$

Since $\varepsilon > 0$ was arbitrary, the second inequality follows. The first one follows from the fact that for any algorithm $\phi \in \Phi(N)$

$$\begin{aligned}
&\sup_{f\in F_0} \sup_{\substack{g\in F_0\\ N(f)=N(g)}} \|S(f)-S(g)\|_{F_2} \\
&= \sup_{y\in N(F_0)} \sup_{\substack{g\in F_0:\\ N(g)=y}} \sup_{\substack{f\in F_0:\\ N(f)=y}} \|S(f)-S(g)\|_{F_2} \\
&\leq \sup_{y\in N(F_0)} \sup_{\substack{g\in F_0:\\ N(g)=y}} \sup_{\substack{f\in F_0:\\ N(f)=y}} \|S(f)-\phi(y)\|_{F_2} + \|S(g)-\phi(y)\|_{F_2} \\
&\leq 2 \sup_{f\in F_0} \|S(f)-\phi(N(f))\|_{F_2}\,.
\end{aligned}$$

Taking the infimum on both sides and noting that the left hand side does not depend on ϕ yields the desired property. □

We are interested in the comparison of global optimization and function approximation, from a worst case point of view. To formalize these problems, we define $F_1 := C([0,1]^d)$ to be the space of continuous functions from $[0,1]^d$ to $\mathbb{R}$, equipped with the $\|\cdot\|_\infty$-norm. In the case of function approximation, the interesting parameter $S(f)$ of a continuous function $f \in F_1$ is the function itself. The approximation problem can therefore be formulated in the above setting with $F_2 = F_1$ and $S = I$, the identity operator. In global optimization, the interesting parameters are real numbers. Thus, the formulation of the global optimization problem is given by $F_2 = \mathbb{R}$ and $S(f) := S_m(f) := \max_{x\in[0,1]^d} f(x)$. In order to find a minimum, just maximize $-f$.

The next lemma, which we adapt from Wasilkowski (1984), examines the complexity of global optimization. The key idea in the proof is that a global

optimization algorithm has to be able to distinguish the zero function from an arbitrary function with small support. Furthermore, the global optimization problem is not substantially easier than the approximation problem with non-adaptive information.

LEMMA A.6.2. *Let $F_0 \subset F_1$ be a subset of the class of continuous functions on $[0,1]^d$. Let $f^* \equiv 0$ be the function that is constantly zero. For any information operator N, suppose there exists a non-trivial, non-negative perturbation function $h \in F_0$ such that $N_{f^*}(h) = 0$. In other words, the functions f^* and $f^* + h$ are not distinguishable by means of the information delivered by N_{f^*}. Then,*

$$r(N; S_m, F_0) \geq \frac{1}{2}\|h\|_\infty .$$

If F_0 is symmetric and convex, it follows that $r(N; S_m, F_0) \geq \frac{1}{4} r(N_{f^*}; I, F_0)$.

Proof. We follow the presentation of Wasilkowski (1984). Let $h \in F_0$ be such that $N_{f^*}(h) = 0$. It follows that $N_{f^*+h}(f^* + h) = N_{f^*}(f^* + h)$, i.e. the information does not adapt to h, and N cannot distinguish h from f^*:

$$N(f^*) = N_{f^*}(f^*) = N_{f^*}(f^*) + N_{f^*}(h) = N_{f^*}(f^* + h) = N_{f^*+h}(f^* + h) = N(f^* + h).$$

Then, by $S_m(f^*) = 0$,

$$\begin{aligned} 2 \sup_{f \in F_0} \left| S_m(f) - \phi(N(f)) \right| &\geq \left| S_m(f^*) - \phi(N(f^*)) \right| + \left| S_m(f^* + h) - \phi(N(f^* + h)) \right| \\ &= \left| S_m(f^*) - \phi(N(f^*)) \right| + \left| S_m(f^* + h) - \phi(N(f^*)) \right| \\ &\geq \left| S_m(f^*) + S_m(f^* + h) \right| \\ &= \max_{x \in [0,1]^d} h(x). \end{aligned}$$

It follows that $r(N; S_m, F_0) \geq \frac{1}{2}\|h\|_\infty$.

Suppose F_0 is symmetric and convex. We find for $f, g \in F_0$ that $\frac{1}{2}f - \frac{1}{2}g \in F_0$, and therefore

$$\begin{aligned} \frac{1}{2} d(N_{f^*}; I, F_0) &= \sup \left\{ \frac{1}{2}\|f - g\|_\infty : f, g \in F_0, N_{f^*}(f - g) = 0 \right\} \\ &\leq \sup \left\{ \|h\|_\infty : h \in F_0, \ N_{f^*}(h) = 0 \right\} \\ &\leq 2\, r(N; S_m, F_0). \end{aligned}$$

We have shown that,

$$r(N; S_m, F_0) \geq \frac{1}{4} d(N_{f^*}; I, F_0) \geq \frac{1}{4} r(N_{f^*}; I, F_0).$$

This is the desired result. □

We show that there exist functions $h \in F_0$ such that $N_{f^*}(h) = 0$. For that purpose, recall that the ε-covering number $N(\varepsilon, [0,1]^d, \|\cdot\|_\infty)$ of the d-dimensional unit cube is defined as the minimal number of points $x_1, \ldots, x_N$ such that the balls $B_\infty(x_i, \varepsilon)$ with respect to the uniform norm cover the unit cube. Let $\#N_{f^*}$ be the cardinality of the information N_{f^*}, i.e. the number of the a priori determined evaluation points. Chose ε_h such that

$$\varepsilon_h = \frac{1}{2} \sup \left\{ 1 > \varepsilon > 0 \colon N(\varepsilon, \|\cdot\|_\infty, [0,1]^d) > \#N_{f^*} \right\}. \tag{A.3}$$

Let then h be a function with the desired smoothness, i.e. $h \in F_0$, such that $\operatorname{supp} h \subset B_\infty(x_0, \varepsilon_h)$ for some point $x_0 \in [0,1]^d$. Due to the choice of ε_h, we can choose x_0 in order to ensure that no evaluation point defined by the information operator N_{f^*} lies in the support of h. Therefore, f^* and $f^* + h$ are not distinguishable by means of the information operator N_{f^*}, and $N_{f^*+h}(f^* + h) = N_{f^*}(f^* + h)$ as well as $N_{f^*}(h) = 0$.

Now we state a lemma concerning the error in the problem of global optimization of a function belonging to a Hölder class. In the following, the Hölder class $C_d^{k,\alpha}$ over the unit cube $[0,1]^d$ is defined as

$$C_d^{k,\alpha} := \left\{ f \colon D \to \mathbb{R};\ |\partial^{(l)} f(x) - \partial^{(l)} f(y)| \le \max_{i=1,\ldots,d} |x_i - y_i|^\alpha, |l| = k \right\},$$

where l is a multi-index.

The lemma states that the worst case error in global optimization over the d-dimensional unit cube suffers from the curse of dimensions, i.e. the amount of information needed to ensure the worst case error to be less then ε grows like ε^{-d}. In the next lemma, we present results that we found in publications by Novak (1988) and Vavasis (1991). Note that the methodology is essentially the same as in the proof of minimax lower bounds for nonparametric estimators as presented for instance in Tsybakov (2008).

LEMMA A.6.3. *Let $F_0 = C_d^{k,\alpha}$, and let N be any information operator with at most $n-1$ function evaluations. Then there exists a constant $c > 0$ such that*

$$r(N; S_m, F_0) \ge c \cdot n^{-\frac{k+\alpha}{d}}.$$

Proof. The proof starts giving a lower bound for the covering number of the d-dimensional unit cube with respect to the norm $\|\cdot\|_\infty$. Let $D(\varepsilon, \|\cdot\|_\infty, [0,1]^d)$ be the maximal number of points $\{x_1, \ldots, x_m\} \subset [0,1]^d$ such that, $\|x_i - x_j\|_\infty > \varepsilon$. This number is called *ε-packing number* of the unit cube. An equidistant grid with mesh size 2ε gives an ε-packing of the unit cube. Therefore, the packing number for the unit cube with respect to the uniform norm is at least $(2\varepsilon)^{-d}$. The covering

number $N(\varepsilon/2, \|\cdot\|_\infty, [0,1]^d)$ is bounded from below by $D(\varepsilon, \|\cdot\|_\infty, [0,1]^d)$ (Van der Vaart and Wellner, 1996, page 98). Therefore, $N(\varepsilon, \|\cdot\|_\infty, [0,1]^d) \geq (4\varepsilon)^{-d}$. We have assumed that $\#N_{f^*} \leq n-1$. This means that we can estimate the covering number from below by $N\big((n/4)^{-1/d}, \|\cdot\|_\infty, [0,1]^d\big) \geq n > \#N_{f^*}$. For ε_h from equation (A.3), we conclude that $\varepsilon_h \geq \frac{1}{2}(n/4)^{-1/d}$.

We construct the function h. Let $x_0 \in [\varepsilon_h, 1-\varepsilon_h]^d$ be some point with the property that the ball $B_\infty(x_0, \varepsilon_h)$ does not contain any evaluation point given by N_{f^*}. We choose a function $g\colon \mathbb{R}^d \to \mathbb{R}$ with $\operatorname{supp} g \subset [-1,1]^d$ and $g(0) = \|g\|_\infty = 1$ that satisfies the Hölder condition $|\partial^{(l)} g(x) - \partial^{(l)} g(y)| \leq \max_i |x_i - y_i|^\alpha$ for $|l| = k$, and define

$$h(x) := \varepsilon_h^{k+\alpha}\, g\left(\frac{x - x_0}{\varepsilon_h}\right).$$

If l is a multi-index with $|l| = k$, we obtain

$$\left|\partial^{(l)} h(x) - \partial^{(l)} h(y)\right| \leq \varepsilon_h^{k+\alpha-k} \max_i \left|\frac{x_i - y_i}{\varepsilon_h}\right|^\alpha = \max_i |x_i - y_i|^\alpha.$$

Hence, $h \in C_d^{k,\alpha}$. Moreover, $\|h\|_\infty = \varepsilon_h^{k+\alpha} \|g\|_\infty \geq c\, n^{-(k+\alpha)/d}$ for some positive constant c. Recall that the function h was chosen such that the information operator N cannot distinguish $f^* \equiv 0$ from $f^* + h$. In virtue of Lemma A.6.2, we obtain the assertion. □

The next result shows that there is no way to circumvent the curse of dimensions by means of randomized algorithms. A common reference for this result are Nemirovsky and Yudin (1983). However, for the proof they refer to one of their papers (Nemirovsky and Yudin, 1978) which is written in Russian. We think that there are more accessible references for a similar insight; e.g. Novak (1988) and Bull (2011). We use a result from the latter reference.

For the further proceedings, we need to introduce the notion of a randomized algorithm. Let $(\Omega', \mathscr{A}', P)$ be a probability space. On this space we consider a random information operator N and a random algorithm ϕ. We define the estimated point of global maximum after n evaluations of the objective function f or its derivative as $x_n^*(\omega) := \phi(\omega, N_f(\omega))$. We assume that the available information N_f consists of evaluations of f or its derivative at the points $x_1, \ldots, x_n$. We suppose that the evaluation points are random variables, i.e. they are $(\mathscr{A}' - \mathscr{B}^d)$-measurable. The estimated point of global maximum, x_n^*, is assumed to be an element of $\{x_1, \ldots, x_n\}$. Again, we presuppose measurability.

LEMMA A.6.4 (Bull, 2011). *There exists a constant $C > 0$ such that for any random algorithm*

$$\sup_{f \in C_d^{k,\alpha}} \int_{\Omega'} \left| f(x_n^*(\omega)) - \max f \right| P(d\omega) \geq C \cdot n^{-\frac{k+\alpha}{d}}.$$

Proof. The main difference to the ideas of the deterministic case is that we do not only need one perturbation function h. If we have at least one more function than evaluation points, almost surely no evaluation point falls inside the support of one perturbation function.

As in the previous proof, let $g\colon \mathbb{R}^d \to \mathbb{R}$ with $\operatorname{supp} g \subset [-1,1]^d$ and $g(0) = \max g = 1$ satisfy the Hölder condition. Then we define functions $h_i \in C^{k,\alpha}([0,1]^d)$ by,

$$h_{i,\varepsilon}(x) := \varepsilon^{k+\alpha} g\left(\frac{x - x_{i,\varepsilon}}{\varepsilon}\right),$$

where the number $\varepsilon > 0$ and the points $\{x_{i,\varepsilon}\colon i = 1,\dots,n+1\} \subset [\varepsilon, 1-\varepsilon]^d$ are supposed to ensure that the functions $\{h_{i,\varepsilon}\colon i = 1,\dots,n+1\}$ have disjoint supports. This means that ε must be chosen such that the ε-packing number of the d-dimensional unit cube is not smaller than $n+1$. We shall choose $\varepsilon := 2^{-(1+d)/d} n^{-1/d}$. Then, $D(\varepsilon, \|\cdot\|_\infty, [0,1]^d) \geq (2\varepsilon)^{-d} \geq n+1$.

Again, let f^* be the zero function. We define the random variable N to be the number of $h_{j,\varepsilon}$ such that $h_{j,\varepsilon}(x_i) = 0$ for all i:

$$N := \sum_{j=1}^{n+1} \mathbb{1}_{\{h_{j,\varepsilon}(x_i)=0 \text{ for all } i\}}.$$

Define the random variable j^* by

$$j^*(\omega) = \begin{cases} \min\{j = 1,\dots,n+1\colon h_{j,\varepsilon}(x_i) = 0 \text{ for all } x_i\} & \text{if } N(\omega) > 0 \\ 0 & \text{if } N(\omega) = 0. \end{cases}$$

This variable is indeed measurable. Since all functions $h_{i,\varepsilon}$ have disjoint supports and the total number of these functions is $n+1$, it follows that the event that for any $h_{j,\varepsilon}$ there are some $x_i \in \operatorname{supp} h_{j,\varepsilon}$ is the impossible event. Hence, $(j^*)^{-1}\{0\} = \emptyset \in \mathscr{A}'$. For $k = 1,\dots,n$

$$\{\omega\colon j^*(\omega) = k\} = \bigcap_{j=1}^{k-1} \bigcup_{i=1}^{n} \{\omega\colon h_j(x_i(\omega)) > 0\} \cap \bigcap_{i=1}^{n} \{\omega\colon h_k(x_i(\omega)) = 0\} \in \mathscr{A}'.$$

Since x_n^* is picked from $\{x_1, \ldots, x_n\}$, we have $h_{j^*,\varepsilon}(x_n^*) = 0$ and conclude

$$\begin{aligned}
&\sup_f \int |f(x_n^*) - \max f| \, dP \\
&\geq \frac{1}{2}\Big[\int_{\{N>0\}} |f^*(x_n^*) - \max f^*| \, dP + \int_{\{N>0\}} |h_{j^*,\varepsilon}(x_n^*) - \max h_{j^*,\varepsilon}| \, dP\Big] \\
&\geq \frac{1}{2}\int_{\{N>0\}} \min_j |\max h_{j,\varepsilon}| \, dP \\
&= \frac{1}{2}\varepsilon^{k+\alpha} P\{N>0\} \\
&= 2^{-(k+\alpha)(1+2d)/d} n^{-(k+\alpha)/d} P\{N>0\}\,.
\end{aligned}$$

We have seen that $N > 0$ for all ω. Hence, $P\{N > 0\} = 1$, and the statement of the lemma follows. □

We learn from the results of the worst case analysis in global optimization that for any algorithm, random or deterministic, there is a pathological function such that the algorithm does not outperform the naive grid search approach in optimizing this function. In the grid search approach, we construct an equidistant grid of mesh size ε over the domain $[0,1]^d$, evaluate the objective function at each node, and return the maximum of these values as the approximate maximum of the function. If the objective function f satisfies a Lipschitz condition with Lipschitz constant L, we are guaranteed to approximate the maximal value of the function with accuracy $L\varepsilon$. However, this approach needs ε^{-d} function evaluations, which quickly amounts to an unaffordable computing time even if d grows modestly.

Of course, there may be functions for which grid search is outperformed. But from the current point of view there is no general rigorous argument as to whether one particular algorithm is preferable to another. For a specific problem at hand, several algorithms have to be exhibited and an individual choice has to be made. We want to give a short review of the result of our search for global optimization strategies. This is not intended to be an exhaustive comparative study on global optimization strategies, which would lie beyond the scope of this thesis. Instead, we want to give an impression of the variety of competing methods that have been developed in this field.

A.6.2 Deterministic strategies: branch-and-bound

Most deterministic approaches to find a global extreme point of a function consist of constructing a sequence of partitions of the function domain. In the literature these strategies are sometimes referred to as “partitioning techniques” (Rios and Sahinidis, 2013). In contrast to the naive grid search approach, available

information about the function is employed to drive the partitioning process. In that way one can hope for a little more efficiency. On the partitions, an approximation of the objective function is constructed employing information about the values of the objective function f at certain evaluation points. This approximation, in turn, is used to estimate the minimum (or maximum) of the objective function. Strategies differ in the specific partitioning strategy, i.e. which subset of the current partition should be selected for further division, and in the choice of the approximation of the objective function. The essence of the approach is perfectly depicted by the following two steps:

> 1. *Branching: A set of solutions ... can be partitioned into mutually exclusive sets.*
> 2. *Lower bounding: An algorithm is available for calculating a lower bound on the cost [i.e. value of the objective function] of any solution in a given subset,*
>
> (Papadimitriou and Steiglitz, 1998, page 438)

The basic branch-and-bound algorithm for combinatorial optimization follows this principle. It is displayed below in Algorithm 1.

In the case of global optimization with non-discrete variables, many algorithms follow similar ideas. One example for the extension to the case of continuous variables is the DIRECT algorithm that is designed to find the minimum of a Lipschitz continuous function over the d-dimensional unit cube. The DIRECT algorithm is a partitioning scheme first proposed by Jones et al. (1998). It resulted from advancing the underlying ideas of Piyavskii's/Schubert's algorithm which is used for minimizing one-dimensional lipschitzian functions (Shubert, 1972; Piyavskii, 1972). The Piyavskii/Shubert algorithm uses an underestimation of saw-tooth functions based on a partition of the objective function's domain; for a more detailed exposition cf. Hansen and Jaumard (1995) or Gablonsky (2001).

The name DIRECT stands for DIvide RECTangles and refers to the partitioning nature of the method. The domain $[0,1]^d$ is partitioned into sub rectangles, and the objective functions is evaluated at the centers of these rectangles. We will briefly explain the partitioning procedure following the outline of Gablonsky (2001), pages 34–35. We distinguish two cases. The current region I can either be a hyper cube or a hyper rectangle. Let us begin with the first case. Denote by c the center of the hyper cube I, and by l the length of its edges. We evaluate the objective function at two points along each axis. Let e_i be the ith vector of the Euclidean basis. We collect values,

$$y_{i,+} = f\left(c + \frac{l}{3} e_i\right)$$

$$y_{i,-} = f\left(c - \frac{l}{3} e_i\right)$$
$$y_i := \min\{y_{i,+}, y_{i,-}\}.$$

Let $(i_1, \ldots, i_d)$ be an ordering of the coordinates such that $y_{i_1} \leq \ldots \leq y_{i_d}$. We divide the cube perpendicular to the direction i_1 into three rectangles of equal size. The rectangle containing the center will be divided perpendicular to i_2, again into three parts of equal size. We repeat this procedure until we have divided the cube with respect to all coordinates.

In case the current region I is a hyper rectangle, we identify the maximum length l among all edges of I. We collect all edges with this length in a set L and proceed with the same partitioning scheme as for hyper cubes, applying it to all edges in L. Now that we have established a branching scheme, we are left with the decision which rectangles are up for further division. For that purpose, DIRECT identifies the set of all hyper rectangles that fulfill the property of being "potentially optimal". We refer to Gablonsky (2001) for a definition and discussion of that term. The two decisive parameters in this instance are the values of the objective function at the center of a hyper rectangle and the potential rate of decrease of f by dividing a certain rectangle (Rios and Sahinidis, 2013). These two parameters are handled in a way to balance local and global search for a minimum. The DIRECT algorithm can be displayed as in Algorithm 2 (Gablonsky, 2001, p. 38). An implementation `Direct.m` for MATLAB® is available (Finkel, 2004).

A more involved branch-and-bound method to minimize a function is described by Liberti (2004). We depict the basic idea of a "spatial Branch-and-Bound algorithm" (ibid., page 106) in Algorithm 2. An essential part of the approach is the construction of convex relaxations Φ_R of the objective function over sub-regions R of its domain. For an example of the construction of such a convex relaxation, we refer to the α-BB algorithm. This algorithms dates back to the work of Androulakis et al. (1995) and was thoroughly analyzed in the papers Adjiman et al. (1998b) and Adjiman et al. (1998a). In the work of Liberti (2004) it is seen as an instance of a spatial branch-and-bound algorithm. It is assumed that the objective function f is twice continuously differentiable. Suppose that the current region is a sub interval $R = [l, u]$. Then the function

$$\Phi_{\alpha,R}(x) := f(x) + \alpha \sum_{i=1}^{d} (l_i - x_i)(u_i - x_i)$$

is a convex underestimator of f with respect to R if and only if

$$\alpha \geq \max\left\{0, -\frac{1}{2} \min_{x \in R} \lambda_{\min} H_f(x)\right\},$$

Eichfelder et al. (2016). Here, $\lambda_{\min} H_f(x)$ denotes the smallest eigenvalue of the hessian of f at the point x. For details, we refer to Eichfelder et al. (2016) who cite Maranas and Floudas (1994). We find a minimum point x_R^* of $\Phi_{\alpha,R}$ using standard methods from convex optimization, e.g. sequential quadratic programming (Geiger and Kanzow, 2002). For the upper bound, we could use $u_R := f(x_R^*)$ (Eichfelder et al., 2016). For the α-BB algorithm, convergence results and rates of convergence are available (Eichfelder et al., 2016). Of course, the convergence deteriorates with increasing dimension.

There have been proposed several other branch and bound algorithms in the literature. For a first overview, one can refer to Liberti (2004) or Hansen and Jaumard (1995).

```
begin
    activeset := {0}; (comment: "0" is the original problem)
    U := ∞ ;
    x* := ∞;
    currentbest := anything ;
    while activeset is not empty do
        choose a set K ∈ activeset;
        remove K from activeset;
        generate the children C_1 ..., C_{n_k} of K;
        for i = 1,..., n_k do
            calculate the corresponding points x_i for
            which l_i := f(x_i) are lower bounds of {f(x): C_i};
            if l_i ≥ U then
                kill child C_i
            end
            else if child i is a complete solution then
                U := l_i;
                currentbest := child C_i;
                x* := x_i;
            end
            else
                add child C_i to activeset
            end
        end
    end
    return U and x*;
end
```

Algorithm 1: Basic branch-and-bound algorithm, cf. Papadimitriou and Steiglitz (1998, page 438).

```
begin
    c* := c_1; (c_1 is the center of the unit cube)
    f_min := f(c*);
    t = 0; (number of iterations)
    m = 0; (number of function evaluations)
    initialize m_max, t_max ∈ ℕ (maximal number of iterations evaluations)
    while t ≤ t_max and m ≤ m_max do
        S* := the set of potentially optimal rectangles;
        for R ∈ S* do
            evaluate f at the centers of the new rectangles;
            apply the division rule to R;
            update f_min;
            update m;
            update c*;
        end
        update S*;
        update t;
    end
    return f_min and c*
end
```

Algorithm 2: DIRECT algorithm, cf. Gablonsky (2001, page 38).

begin
  initialize *activeset*; (list of regions comprising the entire set of variable ranges)
  initialize $\varepsilon > 0$; (convergence tolerance)
  $x^* := \infty$;
  $U := \infty$; (current output for the global minimum)
  for $R \in$ *activeset* **do**
    $l_R := -\infty$; (lower bound on $\min_{x \in R} f(x)$)
  end
  while *activeset* $\neq \emptyset$ **do**
    choose a region $R \in$ *activeset*;
    activeset = *activeset* $\setminus \{R\}$;
    generate a convex relaxation Φ_R of the objective function over the region R and compute
      $l_R := \min_{x \in R} \Phi_R(x)$;
    if $l_R \leq U$ **then**
      find some $x_R \in R$ such that $u_R := f(x_R)$ is an upper bound for the global minimum;
      if $u_R < U$ **then**
        set $U := u_R$ and $x^* := x_R$;
        delete all regions r from *activeset* for which $l_r > u$;
        if $u_R - l_R < \varepsilon$ **then**
          accept u_R as $\min_{x \in R} f(x)$;
        end
        else
          Apply a branching rule to split R into disjoint sub-regions $S_1, \ldots, S_N$;
          Add the sub-regions to *activeset*;
          $l_{S_i} := l_R$ $(i = 1, \ldots, N)$;
        end
      end
    end
  end
  return U and x^*;
end

Algorithm 3: Spatial branch-and-bound algorithm for global optimization according to Liberti (2004, page 106).

begin
  generate $X_1 \sim Unif[0,1]^d$;
  initialize a grid $\{L_k\}_{k \in \mathbb{N}}$; (from which L is estimated)
  initialize $n \in \mathbb{N}$ (maximum number of iterations)
  initialize $t := 1$; (number of iterations)
  initialize $\hat{L} := 0$; (initial estimate for L)
  evaluate $f(X_1)$;
  while $t < n$ **do**
    generate $B_{t+1} \sim Bin(1, p)$;
    $\mathscr{X}_{\hat{L},t} := \{x \in [0,1]^d : \min_{i=1,\ldots,t} f(X_i) + \hat{L} \|x - X_i\|_2 \geq \max_{i=1,\ldots,t} f(X_i)\}$;
    if $B_{t+1} = 1$ **then**
      generate $X_{t+1} \sim Unif\,[0,1]^d$;
    end
    else
      generate $X_{t+1} \sim Unif\,(\mathscr{X}_{\hat{L},t})$;
    end
    evaluate $f(X_{t+1})$;
    $t = t + 1$;
    update $\hat{L} := \inf\left\{L_k : \max_{i \neq j} \frac{|f(X_i) - f(X_j)|}{\|X_i - X_j\|_2} \leq L_k\right\}$;
  end
  return X^*, $f(X^*)$ with $f(X^*) = \max_{t \leq n} f(X_t)$;
end

Algorithm 4: AdaLIPO algorithm, cf. Malherbe and Vayatis (2017).

A.6.3 Random search

THE (ADA)LIPO ALGORITHM

We present a recently proposed example of a random search algorithm. The LIPO algorithm and its extension adaLIPO have been introduced in the article of Malherbe and Vayatis (2017). LIPO is a global random search method for finding the maximum of a lipschitzian function with known Lipschitz constant L. Its extension to functions with unknown lipschitz constant is called Adaptive LIPO, or AdaLIPO. The LIPO algorithm works by generating in each step a sample point from a uniform distribution over $[0,1]^d$, and evaluating f at that point if a decision rule is fulfilled: given sample points $X_1,\dots,X_t$ and a new point $X_{t+1} \sim Unif[0,1]^d$, evaluate $f(X_{t+1})$ if

$$\min_{i\leq t}\left[f(X_i)+L\,\|X_{t+1}-X_i\|_2\right] \geq \max_{i\leq t} f(X_i). \tag{LIPO}$$

After the current number t of iterations has reached a previously determined maximum n, the output is defined by $\hat{X}_n := \arg\max_{i\leq n} f(X_i)$.

The AdaLIPO algorithm (Algorithm 4) works with an estimate of the Lipschitz constant which is updated in each iteration. The parameter p represents the trade off between pure random exploration of new areas in the domain (in the case $B_{t+1}=1$) and exploitation of known function values by applying a LIPO step. In their experimental studies, Malherbe and Vayatis (2017) used $p=0.1$. The quantity $\max f - \max_{i\leq n} f(X_i)$ is proven to be of order $O_p(n^{-1/d})$ (ibid.).

BAYESIAN OPTIMIZATION: EXPECTED IMPROVEMENT

The main motivation for the so called Bayesian approach to global optimization is that the worst case point of view may be too pessimistic. We strongly suspect that there are functions for which a sensible algorithm outperforms the simple grid search approach. And it might as well be the case that these functions are the ones that would most likely appear in a real world problem. This idea suggests to assess an algorithm based on the average performance over a class of functions rather than on the worst possible performance. It is well known that some characteristics of a problem may change if we adopt the average case point of view. For instance, in global optimization adaptive information does not yield better rates of convergence than non-adaptive information, in the worst case setting (Novak, 1988; Wasilkowski, 1984); in the average case analysis, however, it was shown by Calvin (1997) that adaptation indeed improves the rates.

In order to formulate the average case approach, we have to amend the setting of random algorithms by a distribution over the set of objective functions. We

mainly reproduce the conception of Bull (2011). We are in the setting of random information and random algorithms. Suppose now that not only the information $\{x_n\}_{n\in\mathbb{N}}$ and the estimated minimum point x_n^* is random but also that f is a random function. It will be assumed that the algorithm at time n does not depend on information about f that is yet unknown. This can be formulated in the restriction that x_n is independent of f, conditionally on the σ-field

$$\mathscr{F}_{n-1} := \sigma\big(x_i, f(x_i)\colon\, i = 1,\dots,n-1\big).$$

Assume that $x_n^* \in \{x_1,\dots,x_n\}$ is the estimated minimum point after n function evaluations. We are interested in the average error

$$\int_{\Omega'} |f(\omega, x_n^*(\omega)) - \min f(\omega)|\, P(d\omega).$$

The available type of information consists of an evaluation point x and the function value $f(x)$ at that point. At step i of the algorithm, given the information $z_i = (x_i, f(x_i)$, a decision function d_i chooses the next evaluation point $x_{i+1} = d_i(z_i)$. The sequence $\{x_i(d)\}_{i=1}^N$ of evaluation points that we obtain depends heavily on the strategy d. Our goal is to find the strategy d_b such that

$$E|f(x_N^*(d_b)) - \min f| = \min_d E|f(x_N^*(d)) - \min f|.$$

Such a strategy is called the Bayesian strategy, cf. Mockus (1989). Here, E denotes the expectation with respect to the prior distribution P. Using the iterative conditioning

$$E\big[E\big[\dots E\big[E\big[|f(x_N^*(d_b)) - \min f|\,\big|z_{N-1}\big]\,\big|z_{N-2}\big]\dots\big|z_{i+1}\big]\,\big|z_1\dots,z_i\big],$$

the Bayesian strategy can be formulated as the solution of a recurrent system of dynamic programming (Mockus, 1989). The last display may be seen as the formal expression for being at time i and looking ahead to time N, i.e. $N-i$ steps ahead. A sensible approximation of this strategy would be to look merely one step ahead, i.e. $N = i+1$. Being at time n, we would choose x_{n+1} to minimize

$$E\big[|f(x_{n+1}^*) - \min f|\,\big|z_1,\dots,z_n\big]\,. \tag{A.4}$$

Choosing the evaluation points in this way is called the strategy of "expected improvement" (Bull, 2011; Brochu et al., 2010) or "one step approximation" (Mockus, 1989). In order to carry out this procedure, we have to calculate the posterior distribution $Law(f(x)|\,z_1,\dots,z_n)$ for which we need to specify the prior probability measure P. It is a common assumption that f is a stationary Gaussian process under

P, with expectation $Ef \equiv \mu$ and covariance function $\operatorname{cov}[f(x), f(y)] = \sigma^2 K_\theta(x-y)$, where K_θ is called the correlation kernel which is described by smoothness parameters ν and α (Bull, 2011, pages 5-6). Using that all finite dimensional distributions $P^{(f(t_1),\ldots,f(t_n))}$ of f are multivariate normal distributions, one can derive explicit expressions for the posterior and hence also for the expression

$$\mathrm{EI}_n(x_{n+1}, P) := E\left[\left(f(x_n^*) - f(x_{n+1})\right)_+ \middle| z_1, \ldots, z_n\right].$$

A strategy that chooses x_{n+1} to maximize EI_n minimizes the quantity (A.4) (Bull, 2011, page 5). It has been remarked by Bull (2011) that the unknown parameters μ, σ, θ can be estimated using a maximum likelihood approach suggested by Jones et al. (1998). It is one of the main results of Bull (2011) that the following expected improvement scheme attains near optimal rates (up to a logarithmic factor) in the worst case setting.

DEFINITION A.6.5 (Expected improvement strategy; Bull, 2011, Definition 4). An optimization algorithm is called $\mathrm{EI}(P, \varepsilon)$ strategy if for a given $0 < \varepsilon < 1$:

1. it chooses initial design points $x_1, \ldots, x_k$ independently of f;
2. with probability $1 - \varepsilon$, it chooses a design point x_{n+1} $(n \geq k)$ to maximize $\mathrm{EI}_n(x_{n+1}, P)$ or,
3. with probability ε, it chooses x_{n+1} $(n \geq k)$ uniformly at random from $[0,1]^d$.

The parameter ε has to be chosen in advance. It "controls the trade-off between global and local search" (Bull, 2011, page 14). For fixed parameter θ, the worst case error is considered over a reproducing-kernel Hilbert space $\mathscr{H}_\theta$ that is equivalent to the Sobolev Hilbert space $H^{\nu+d/2}([0,1]^d)$ (Bull, 2011, page 10). In this setting, the error of the algorithm using the $\mathrm{EI}(P, \varepsilon)$ strategy to minimize a function $f \in \mathscr{H}_\theta$ converges to zero with a rate that is asymptotically bounded from above by $(n/\log(n))^{-\nu/d}(\log n)^\alpha$,

$$\sup_{\|f\|_{\mathscr{H}_\theta} \leq R} E_f^{\mathrm{EI}}\left[f(x_n^*) - \min f\right] = O\left((n/\log n)^{-\nu/d}(\log n)^\alpha\right)$$

for any $R > 0$ (Bull, 2011, Theorem 5), where ν and α are the smoothness parameter of the covariance kernel K_θ (Bull, 2011, page 6). The results can be extended to the case where the parameter θ is estimated.

This result demonstrates that the Bayesian approach is not inferior to other algorithms, from the worst case perspective. However, due to its foundations, there are grounds for expecting a better average performance. An expected improvement algorithm is implemented in the MATLAB®® function `bayesopt()`. In the literature it is insinuated that Bayesian optimization works best in a low dimensional setting, typically with 10 variables or less.

A.6.4 Heuristic methods

In this section we depict the heuristic optimization methods that we used in our simulation study. They originate in analogies to physical or biological processes. The underlying idea is that physical or biological systems often evolve into very stable states in terms of energy or fitness, for instance. We try to understand such a process as a search over the space of all possible states of a system which terminates in an optimal or nearly optimal state. The aim in all these heuristic methods is to imitate such a natural system by a random search optimization algorithm.

Simulated Annealing

Simulated Annealing (SA) is a minimization algorithm. It tries to resemble the process of cooling liquid or solid matter. The foundation of this approach is the observation that the molecules of a liquid or solid piece of matter align themselves in very a stable structure if the piece is cooled sufficiently slow. Hence, this cooling procedure corresponds to the process of finding the state of minimum energy for the system. In statistical physics the Boltzmann distribution is used to describe a system in contact with a heat reservoir. Such a model is often called the *canonical ensemble* (Herman, 2005). The probability for a canonical ensemble to be in a state of energy E_i is described by a discrete probability distribution over the space of all energy states. This distribution is given by

$$P\{E_i\} = \frac{1}{Z_n} e^{-\frac{E_i}{k_b T}}.$$

Here, Z_n is a normalizing constant, T is the temperature of the system and K_b is the Boltzmann constant of statistical physics (Herman, 2005; Spall, 2003). The distribution P is often called Boltzmann distribution.

Assume now that f is our objective function which we intend to minimize. Imagine a canonical ensemble with the energy states $f(x)$. Then the energy states have the distribution proportional to

$$p_T(x) = e^{-f(x)/(k_b T)}.$$

If we now generate a random sample from these distributions while lowering the temperature T sufficiently slow, we expect to arrive eventually in the state of minimum energy which corresponds to a global minimum of f.

To be more specific, we state a rigorous result of Romeijn and Smith (1994) (Boender and Romeijn, 1995, page 838). Denote by P_T the probability distribution with density $c\, p_T(x)$, where $x \in \mathbb{R}^d$, and c is the normalizing constant. Then for all

$\varepsilon > 0$,

$$\lim_{T \downarrow 0} P_T\{x \colon f(x) < \min f + \varepsilon\} = 1\,. \tag{A.5}$$

Thus, we have to employ a method which enables us to simulate a sample from the distributions P_T. Such a method is given by the Metropolis-Hastings algorithm which goes back to the work of Metropolis et al. (1953) and Hastings (1970). It generates a Markov chain $\{X_n^T\}_{n\in\mathbb{N}}$ such that $P^{X_n^T} \rightsquigarrow P_T$. For a detailed exposition of the Metropolis algorithm and its convergence, we refer to Fishman (1996, pages 384–388).

The Basic idea of simulated annealing is to use the Metropolis-Hastings algorithm to generate samples $\{X_n^T\}_{n\in\mathbb{N}}$ for different temperatures T decreasing to zero. We combine the convergence $P^{X_n^T} \rightsquigarrow P_T$ and equation (A.5) to obtain

$$\lim_{T \downarrow 0} \lim_{n \to \infty} P\{f(X_n^T) < \min f + \varepsilon\} = 1$$

(Boender and Romeijn, 1995, page 841). In practice we would let $T = T_n$ grow slowly with n. The basic SA algorithm can be found for instance in Boender and Romeijn (1995, page 841). A theorem of Bélisle (1992) (Boender and Romeijn, 1995, Theorem 5) ensures convergence to a global minimum point in probability under rather general conditions.

The general popularity of Simulated Annealing in various fields is backed by a vast amount of numerical studies reporting good results (Spall, 2003). An extensive review of the distribution of SA methods among practitioners is given by Suman and Kumar (2006). The SA algorithm is implemented in form of the MATLAB® function `simulannealbnd()`.

Genetic Algorithms

Genetic algorithms try to copy the principles of evolutionary biology to find the global maximum of a function. The model of evolution is used in life sciences to explain the development of species. Its main principle is the survival of the fittest. If the fitness of a state x is expressed by a fitness functions f, the fittest state corresponds to a point that maximizes the fitness function. We give a short description of the principles of evolutionary algorithms following the presentation of Spall (2003).

The genetic algorithm (GA) is a so called population-based method. This means that in each iteration not only one candidate point is evaluated but a whole population $\{x_1, \ldots, x_N\}$, with a previously determined population size N, is under consideration. The x_i are called chromosomes. Spall (2003) sketches

the underlying steps of a genetic algorithm. In order to give the reader some structural idea about the procedure, we will reproduce this sketch relying on the intuitive meaning of the used terminology. Afterwards we give proper definitions.

DEFINITION A.6.6 (Core GA Steps; cf. Spall, 2003, page 246).

1. Define a non-negative fitness function f to be maximized over $[0,1]$.
2. *Initialization.* Randomly generate an initial population $\{x_1,\ldots,x_N\}$ of N chromosomes and evaluate the fitness function $f(x_i)$.
3. *Parent Selection.* Select a set of parent chromosomes according to their fitness. The parents are then used to produce a next generation of chromosomes.
4. *Recombination / Cross-over.* For each pair of parents, a random variable $C \sim Bin(1,p)$ decides whether the two offspring chromosomes are created by a cross-over procedure, in which gene sequences of the parents are interchanged ($C = 1$), or by exactly copying the parents ($C = 0$).
5. *Replacement and Mutation.* Replace the parent generation by the offspring generation. For each offspring chromosome mutate the individual genes with low probability.
6. *Fitness evaluation.* Evaluate the fitness of the new generation. If some termination criterion is fulfilled, stop the algorithm. Otherwise go to step 2.

This scheme illustrates that the central idea of genetic algorithms is to imitate a natural evolution process by accounting for the main evolutionary factors *recombination, mutation* and *selection.* Let us give meaning to these expressions.

A vector $x = (\xi_1,\ldots,\xi_d) \in [0,1]^d$ is called chromosome, its components are the genes. Each gene is written in binary code. Assume that we want an accuracy of 3 positions after the decimal point. We choose the number of bits b such that

$$b = \inf\{b \in \mathbb{N}\colon 10^3 \leq \sum_{i=0}^{b-1} 2^i\}.$$

Since $2^0 + \ldots + 2^{b-1} = 2^b - 1$, this is the smallest integer b such that $b \geq \log_2(1001)$. Every $\xi \in [0,1]$ will then be encoded as the coefficients of the binary representation $\lfloor \xi(2^b-1)\rfloor = \sum_{k=0}^{b-1} a_k 2^k$. In the same way, a bit sequence $(a_0,\ldots,a_{b-1}) \in \{0,1\}^b$ can be translated in to the real number $(a_0 2^0 + a_1 2^1 + \ldots + a_{b-1}2^{b-1})/(2^b-1)$ (Spall, 2003, page 239). Therefore, a chromosome will be represented in a bit sequence of length $B = d \cdot b$. The fitness of a chromosome x is given by the value of the fitness function f at that point.

We shortly describe one possible way of *parent selection*. This approach is called "fitness proportionate selection" (Spall, 2003, page 243). Let $(x_1,\ldots,x_N)$ be an enumeration of the current generation of chromosomes. For each $k = 1,\ldots,N$

define the cumulative fitness values by,

$$S_f(x_k) = \sum_{i=1}^{k} f(x_i).$$

Let $(U_1,\ldots,U_N)$ be a vector of i.i.d. random variables, $U_j \sim Unif\,[0,S_f(x_N)]$. A sample $\{P_1,\ldots,P_N\}$ of parent chromosomes is now drawn with replacement from the set $\{x_1,\ldots,x_N\}$ using the following mechanism. For each $j \in \{1,\ldots,N\}$, define

$$k_j = \min\left\{k \in \{1,\ldots,N\} \colon S_f(x_k) \geq U_j\right\}.$$

Then set $P_j := x_{k_j}$. The above definition is correct since S_f is increasing due to the non-negativity of f. The selection mechanism ensures that the chromosomes with higher fitness values are more likely to be chosen as parents.

For the *cross-over* procedure (Spall, 2003, page 244), the set of parents are grouped in pairs, (P_i,P_{i+1}) with $i = 1,3,\ldots,N-1$ (assume that the population size is an even number). Now we generate the offspring of these pairs of parents. If the Bernoulli variable C takes the value zero, the parents are cloned. In this case the offspring is defined as $(\tilde{x}_i,\tilde{x}_{i+1}) := (P_i,P_{i+1})$. Otherwise choose a random partition of the set $\{0,\ldots,B-1\}$ in q slices,

$$\{0,\ldots,i_1|i_2,\ldots,i_3|\ldots\ldots\ldots|i_k,\ldots,i_{k+1}|\ldots\ldots\ldots|i_{q-1},\ldots,i_q\},$$

where $i_q = B-1$. Then the bit sequences $P_i = (a_0,\ldots,a_{B-1})$ and $P_{i+1} = (b_0,\ldots,b_{B-1})$ can be displayed as

$$\begin{aligned}
&\left(a_0,\ldots,a_{i_1}|a_{i_2},\ldots,a_{i_3}|\ldots\ldots\ldots|a_{i_k},\ldots,a_{i_{k+1}}|\ldots\ldots\ldots|a_{i_{q-1}},\ldots,a_{B-1}\right),\\
&\left(b_0,\ldots,b_{i_1}|b_{i_2},\ldots,b_{i_3}|\ldots\ldots\ldots|b_{i_k},\ldots,b_{i_{k+1}}|\ldots\ldots\ldots|b_{i_{q-1}},\ldots,b_{B-1}\right).
\end{aligned}$$

Now the two offspring chromosomes are defined as the cross-over of P_i and P_{i+1} in the sense that

$$\begin{aligned}
\tilde{x}_i &= \left(a_0,\ldots,a_{i_1}|b_{i_2},\ldots,b_{i_3}|a_{i_4},\ldots,a_{i_5}|\ldots\ldots\ldots\right),\\
\tilde{x}_{i+1} &= \left(b_0,\ldots,b_{i_1}|a_{i_2},\ldots,a_{i_3}|b_{i_4},\ldots,b_{i_5}|\ldots\ldots\ldots\right).
\end{aligned}$$

The notion of *mutation* is quickly explained. Consider a chromosome in bit encoding, $x = (a_0,\ldots,a_{B-1}) \in \{0,1\}^B$. For every bit a_i, we define a random variable $M_i \sim Bin(1,\theta)$ which decides whether the bit is mutated, i.e. replaced with $1-a_i$.

After the mutation phase, the new bit has the form

$$\tilde{a}_i = \begin{cases} a_i & \text{if } M_i = 0 \\ 1 - a_i & \text{if } M_i = 1\,. \end{cases}$$

After the offspring was generated via the crossover method and completed the mutation phase, we obtain the new generation.

For a discussion on convergence of the algorithm and a performance comparison to other optimization methods, we refer to Spall (2003, pages 268 – 275) and the references therein. A genetic algorithm for global function minimization is implemented in the MATLAB® function `ga()`.

Particle Swarm Algorithms

In order to maximize a given function, the Particle Swarm Algorithm (PSA) is designed to mimic the dynamics of social groups. It closely resembles the principles of swarm intelligence, which was pointed out by Kennedy and Eberhart (1995) who introduced the algorithm and refer for these principles to Millionas (1994). We will give a short description of the algorithm, mostly citing the original paper of Kennedy and Eberhart (1995) and a recent literature survey by Bonyadi and Michalewicz (2017). For a short introduction to the subject, one can also refer to Weise (2009) and the references therein.

Similarly to the Genetic Algorithm, the Particle Swarm Algorithm is a population based method. The population can be imagined as points (particles) forming a swarm searching through the domain $[0,1]^d$ for the maximum of a specified function. At each time every particle has a location and a velocity vector containing the direction and speed of its movement. These vectors are updated in each iteration according to a certain rule. We describe the original version of Kennedy and Eberhart (1995) which is referred to as "Original Particle Swarm Optimization" by Bonyadi and Michalewicz (2017).

Assume that we are at iteration t, and consider the ith particle. The necessary information is carried by the characteristic triplet (x_t^i, v_t^i, p_t^i). The first component is the location of the particle; the second component is the velocity vector, i.e. direction and absolute value of movement; and the last component is the best position that the particle has visited up to time t (Bonyadi and Michalewicz, 2017). The original update rule of Kennedy and Eberhart (1995) in the exposition of Bonyadi and Michalewicz (2017) reads as follows. For each dimension $j \in 1,\dots,d$ generate two random numbers $U_t^i(j), W_t^i(j) \sim Unif[0,1]$. Define the jth component

of the velocity vector v_{t+1}^i as

$$v_{t+1}^i(j) := v_t^i(j) + \phi_1 U_t^i(j)\left(p_t^i - x_t^i\right) + \phi_2 W_t^i(j)\left(g_t - x_t^i\right).$$

Here, g_t denotes the best position over all particles in the population at time t, and ϕ_1, ϕ_2 are fixed positive numbers which are called *cognitive* and *social weights* respectively (Bonyadi and Michalewicz, 2017). After the velocity has been updated, the new position of the particle is

$$x_{t+1}^i = x_t^i + v_{t+1}^i.$$

This is the original version of the PSO, but several modifications of it have been proposed in the literature. For instance, one could modify the velocity update in such a way that each particle considers only the information available from its immediate surroundings instead of the whole swarm.

The PSA has become a very popular tool in global optimization for the reason that it is easily adapted for different applications and that it "delivers reasonable results for many different applications" (Bonyadi and Michalewicz, 2017). We refer to the same paper for an exposition of theoretical results on the asymptotic properties of the PSA. Although there have been published some results in this direction, the algorithm's popularity cannot be backed entirely by rigorous theoretical analysis. An implementation of the PSA for global function minimization is available in MATLAB® in form of the function `particleswarm()`.

A.6.5 From constrained to unconstrained global optimization

The global optimization algorithms presented in the last section were designed to solve unconstrained problems. We shall discuss a method how to deal with possible constraints. Our aim is to reformulate a constrained problem such that we are allowed to apply methods from unconstrained optimization.

Suppose that $\mathscr{X}$ is a compact subset of $\mathbb{R}^n$, the functions $f\colon \mathscr{X} \to \mathbb{R}$, $g\colon \mathscr{X} \to \mathbb{R}^m$ and $h\colon \mathscr{X} \to \mathbb{R}^p$ are continuously differentiable, and that the so called set of feasibility,

$$\mathfrak{F} := \{x \in \mathscr{X} : g(x) \leq 0,\ h(x) = 0\},$$

is nonempty. The nonlinear optimization problem of interest is given by

$$\begin{aligned} &\text{minimize } f(x) \\ &\text{subject to } g(x) \leq 0,\ h(x) = 0. \end{aligned} \tag{C}$$

Since many accessible optimization algorithms are designed to search for points of local or global minima over the whole domain of the objective function, it would be convenient to have an alternative formulation of (C) without constraints. A possible approach to this goal is to penalize minimizers of the objective function that are not in the set of feasible solutions. The degree to which an unfeasible solution will be penalized is determined by a so called penalty parameter ε. We are therefore looking for a function $F_\varepsilon(x)$ that assigns roughly the same values as $f(x)$ to feasible points $x \in \mathfrak{F}$ but penalizes unfeasible points $x \notin \mathfrak{F}$ according to the penalty parameter ε. Such a function will be called penalty function. Under certain circumstances, we call F an exact penalty function.

DEFINITION A.6.7. The function $F\colon \mathscr{X} \times [0,\infty) \to \mathbb{R}$ is called an exact penalty function for the problem (C) if there exists a critical value $\varepsilon^* > 0$ such that, for any penalty parameter $0 < \varepsilon \leq \varepsilon^*$, any solution of (C) is a solution to

$$\text{minimize } F(x,\varepsilon),\ x \in \mathscr{X}^o, \tag{U}$$

and conversely, if for some $\varepsilon* > 0$ and all $\varepsilon \in (0,\varepsilon^*]$, any solution of (U) is a solution of (C). The symbol $\mathscr{X}^o$ denotes the interior of the set $\mathscr{X}$.

As we mentioned before, this approach to constrained optimization is a classical one, and there is a huge amount of literature covering this field. Our main reference in this regard is Di Pillo and Grippo (1989). The above definition corresponds to their definition of a weakly exact penalty function (ibid.). In the same paper, they give sufficient conditions to ensure that a penalty function is exact with respect to a given nonlinear program. However, we are not interested in the generality of their abstract theory. We only want to establish exactness of the specific class of l_q-penalty functions in order to demonstrate a possible approach to the solution of our nonlinear constrained problem (4.9).

DEFINITION A.6.8. For $1 \leq q < \infty$, the l_p penalty function for the problem (C) is defined as

$$J_q(x,\varepsilon) := f(x) + \frac{1}{\varepsilon}\left[\sum_{i=1}^{m}(g_i(x))_+^q + \sum_{j=1}^{p}|h_j(x)|^q\right]^{1/q}.$$

Collecting the values $\{(g_i(x))_+\}_{i=1}^m$ and $\{|h_j(x)|\}_{j=1}^p$ in a vector $P(x)$, we may write,

$$J_q(x,\varepsilon) = f(x) + \frac{1}{\varepsilon}\|P(x)\|_q,$$

where $\|\cdot\|_q$ denotes the q-norm on $\mathbb{R}^{m+p}$.

The term $\|P(x)\|_q$ can be seen as a penalty for those x that are not in the set of feasible solutions. Note that for any point $x \in \mathscr{X}$ we have $P(x) = 0$, and therefore

$J_q(x) = f(x)$, if and only if $x \in \mathfrak{F}$. The proof of exactness of the l_q penalties will move alongside the ideas of Di Pillo and Grippo (1989). In order to follow their argument, we need the to define what we mean by the Mangasarian-Fromowitz constraint qualification.

DEFINITION A.6.9. The Mangasarian-Fromowitz constraint qualification (MFCQ) holds at $x \in \mathbb{R}^n$ if the set $\{\nabla h_j(x)\colon j = 1,\dots,p\}$ is linearly independent and there exists a vector $z \in \mathbb{R}^n$ such that for all $j = 1,\dots,p$

$$\langle \nabla h_j(x), z\rangle = 0$$

and for all i with $g_i(x) = 0$

$$\langle \nabla g_i(x), z\rangle < 0\,.$$

LEMMA A.6.10 (Di Pillo and Grippo, 1989, Proposition 4). *Let $x_0 \in \mathfrak{F}$ be a feasible point such that the MFSQ holds at x_0. Then, there exist positive numbers ε_0 and σ_0 depending on x_0 such that, for all $\varepsilon \in (0,\varepsilon_0]$, if x_ε is a critical point of $J_q(\cdot,\varepsilon)$, i.e. the directional derivatives $\partial_d J_q$ satisfy*

$$\partial_d J_q(x_\varepsilon,\varepsilon) \geq 0 \;\; \textit{for all directions } d \in \mathbb{R}^n,$$

and furthermore $\|x_0 - x_\varepsilon\| \leq \sigma_0$, then x_ε is also a feasible point, i.e. $x_\varepsilon \in \mathfrak{F}$.

For a proof of this lemma we refer to Di Pillo and Grippo (1988). The following theorem states that under reasonable assumptions the l_q penalty functions are exact.

THEOREM A.6.11 (Di Pillo and Grippo, 1989, Theorem 4). *Assume that the MFCQ holds at every global solution to (C) and that every global solution belongs to $\mathscr{X}^o$. Then the function J_q is an exact penalty function for the problem (C).*

Proof. The proof proceeds alongside the arguments in the proof of Theorem 1 in Di Pillo and Grippo (1989). We start with the more difficult part to show that for all positive ε up to some threshold value ε^* every global minimizer of $J_q(\cdot,\varepsilon)$ is a solution to (C). We assume the contrary, i.e. for every natural number k there exists an $\varepsilon_k \in (0,\frac{1}{k}]$ such that there exists a minimizer x_k of $J_q(\cdot,\varepsilon_k)$ that is not a solution to (C). We collect all those values ε_k and x_k in the sequences $\{\varepsilon_k\}_{k\in\mathbb{N}}$ and $\{x_k\}_{k\in\mathbb{N}}$. Assume furthermore that $x_{\min}$ is a solution to (C). Then $x_{\min} \in \mathfrak{F}$, and for any k

$$f(x_{\min}) = J_q(x_{\min},\varepsilon_k) \geq J_q(x_k,\varepsilon_k)\,.$$

This means that the sequence $\{J_q(x_k,\varepsilon_k)\}_{k\in\mathbb{N}}$ is bounded: $\limsup_k J_p(x_k,\varepsilon_k) \leq f(x_{\min}) < \infty$. Since $\frac{1}{\varepsilon_k} \to \infty$, it follows that $\limsup_k \|P(x_k)\|_q = 0$. The sequence $\{x_k\}$ is contained in the compact subset $\mathscr{X}$ and thus has a convergent sub sequence $\{x_{k(l)}\}_{l\in\mathbb{N}}$ with $\lim_l x_{k(l)} =: x_\infty \in \mathscr{X}$. By continuity of P and the q-norm, we have

$$\|P(x_\infty)\|_q = \lim_{l\to\infty} \|P(x_{k(l)})\|_q = 0\,,$$

which tells us that x_∞ is a feasible point. By continuity of f, we have

$$\begin{aligned} f(x_{\min}) &\geq \sup_{k\in\mathbb{N}} J(x_k,\varepsilon_k) \\ &\geq \limsup_{l\to\infty} J_q(x_{k(l)},\varepsilon_{k(l)}) \\ &\geq \limsup_{l\to\infty} f(x_{k(l)}) \\ &= f(x_\infty)\,. \end{aligned}$$

Hence, x_∞ is a global solution to (C). By assumption, the MFCQ holds at x_∞, and $x_\infty \in \mathscr{X}^o$. Since x_∞ is a feasible point where the MFCQ holds, we can apply Lemma A.6.10 to the critical points x_k of $J_q(\cdot,\varepsilon_k)$. There exist positive constants ε_0 and σ_0 such that for any $\varepsilon_k \leq \varepsilon_0$ with $\|x_\infty - x_k\| \leq \sigma_0$ we can conclude that $x_k \in \mathfrak{F}$. By choosing $L \in \mathbb{N}$ large enough, we know from $x_{k(L)} \to x_\infty$ that $\|x_{k(L)} - x_\infty\|$ can be made arbitrarily small and in particular smaller than σ_0. Therefore, this $x_{k(L)}$ is a feasible point, and we conclude

$$f(x_{\min}) = J_q(x_{\min},\varepsilon_{k(L)}) \geq J_q(x_{k(L)},\varepsilon_{k(L)}) = f(x_{k(L)})\,.$$

Hence, $x_{k(L)}$ is a feasible point and $f(x_{k(L)})$ is a global minimum of f. This yields a contradiction to the original assumption. We conclude that there exists a number $\varepsilon^* > 0$ such that for all $0 < \varepsilon \leq \varepsilon^*$ every global minimum point of $J_p(\cdot,\varepsilon)$ is a feasible solution to (C).

It is left to show that any feasible minimum point $x_{\min}$ of f is also a minimum point of $J_q(\cdot,\varepsilon)$ for all $\varepsilon \in (0,\varepsilon^*]$. For any $x \in \mathbb{R}^n$,

$$J_q(x,\varepsilon) \geq f(x) \geq f(x_{\min}) = J_q(x_{\min},\varepsilon)\,.$$

This concludes the proof. □

We want to apply this result to the calculation of the least squares spline estimator in the case that $A_n = [0,M]$, which means that $\mathscr{X} = [0,1]^{(l_n+2)B}$. Recall the constraint functions $C_{y,i}$ from Lemma 4.2.10. We have to check the MFCQ for these functions. To simplify matters, we mitigate the constraints a little by substituting the supremum over $\lambda \in [0,M]$ with a maximum over the partition $0 = \xi_0 < \ldots < \xi_l = M$. If the mesh size, i.e. the knot distance, is small enough, we

will be content with a solution that satisfies

$$C_{y,i,\xi_j}(\boldsymbol{\alpha}) := \sum_{p=-1}^{l-1} \left((-1)^i \frac{\alpha_p(y) - \alpha_{p-1}(y)}{\Delta} - L \right) N_{p,1}(\xi_j) \leq 0$$

for all $y \in \{0, \ldots, B-1\}$, $i \in \{0,1\}$, and $\xi_j \in \{\xi_0, \ldots, \xi_{l-1}\}$. Let $I_0(\boldsymbol{\alpha}_0)$ be the set of all tuples (y,i,ξ_j) such that the constraint C_{y,i,ξ_j} is active at the point $\boldsymbol{\alpha}_0$:

$$I_0 := \left\{ (y,i,\xi_j) \colon C_{y,i,\xi_j}(\boldsymbol{\alpha}_0) = 0 \right\}.$$

For indexes $(y^*, i^*, \xi_j^*) \in I_0(\boldsymbol{\alpha}_0)$, we consider the gradients $\nabla_{\boldsymbol{\alpha}} C_{y,i,\xi_j}(\boldsymbol{\alpha}_0)$ at some point $\boldsymbol{\alpha}_0 \in A_n$. The partial derivatives are given by

$$\begin{aligned} \partial_{\alpha_p(y)} C_{y^*,i^*,\xi_j^*}(\boldsymbol{\alpha}_0) &= \begin{cases} \frac{(-1)^{i^*}}{\Delta} \left(N_{p,1}(\xi_j) - N_{p+1,1}(\xi_j) \right) & \text{if } y = y^* \\ 0 & \text{else.} \end{cases} \\ &= \begin{cases} \frac{(-1)^{i^*}}{\Delta} & \text{if } y = y^* \text{ and } j = p+1 \\ \frac{(-1)^{1-i^*}}{\Delta} & \text{if } y = y^* \text{ and } j = p+2 \\ 0 & \text{else.} \end{cases} \end{aligned}$$

Hence, the gradient $\nabla_{\boldsymbol{\alpha}} C_{y^*,i^*,\xi_j^*}(\boldsymbol{\alpha}_0)$ has the form

$$\frac{(-1)^{i^*}}{\Delta} (0, \ldots, 0, -1, 1, 0, \ldots, 0).$$

The only triple $(\tilde{y}, \tilde{i}, \tilde{\xi}_j)$ such that $\nabla_{\boldsymbol{\alpha}} C_{\tilde{y},\tilde{i},\tilde{\xi}_j}(\boldsymbol{\alpha}_0)$ is a multiple of $\nabla_{\boldsymbol{\alpha}} C_{y^*,i^*,\xi_j^*}(\boldsymbol{\alpha}_0)$ would be $(\tilde{y}, \tilde{i}, \tilde{\xi}_j) = (y^*, 1-i^*, \xi_j^*)$. Suppose that $(y^*, i^*, \xi_j^*) \in I_0$. This is equivalent to the statement that

$$(-1)^{i^*} \sum_{p=-1}^{l-1} \frac{\alpha_p(y^*) - \alpha_{p-1}(y^*)}{\Delta} N_{p,1}(\xi_j^*) = L \sum_{p=-1}^{l-1} N_{p,1}(\xi_j^*) > 0,$$

and we conclude that in this case $(y^*, 1-i^*, \xi_j^*) \notin I_0$. From these facts, we infer that for every $\boldsymbol{\alpha}_0 \in A_n$ the set $\left\{ \nabla_{\boldsymbol{\alpha}} C_{y,i,\xi_j}(\boldsymbol{\alpha}_0) \colon (y,i,\xi_j) \in I_0(\boldsymbol{\alpha}_0) \right\}$ is linearly independent. This property is called Linear Independence Constraint Qualification (LICQ). From the validity of the LICQ at the point α_0, we can conclude that the MFCQ hold at this point as well (Di Pillo and Grippo, 1989). Granted the assumption that a global solution lies in the interior of $\mathscr{X}$, we could apply the theorems of Di Pillo and Grippo (1989) to solve the problem (4.9) for $A_n = [0, M]$ and discretized constraints. If the knot distance Δ is small, the set of disretized constraints approximate the original constraint reasonably well.

On the grounds of the above theorem, we can use an iterative scheme defined by a sequence of unconstrained problems with successively increased penalties

to approximately solve our initial problem. Such schemes are widely used to solve constrained nonlinear programs. In Geiger and Kanzow (2002) an iterative procedure called "multiplier penalty method" is proposed using an augmented lagrangian penalty function. Further instances are given in Joines and Houck (1994), Wah et al. (2007) and Chen and Chen (2010) to name but a few. We will use an iteration scheme similar to the one proposed by Chen and Chen (2010).

DEFINITION A.6.12 (Penalty scheme; Chen and Chen, 2010, page 51). Assume that we are in the ith iteration and z_i is the minimizer of $J_1(x, \varepsilon_i)$. Then, setting $\alpha_i = \frac{1}{\varepsilon_i}$, we update the penalty by

$$\alpha_{i+1} \leftarrow \alpha_i + \rho_i \|P(z_i)\|_1 .$$

The vector P is to be understood in accordance with definition A.6.8. The number ρ_i is updated by

$$\rho_{i+1} \leftarrow \rho_i \delta ,$$

where $\delta > 1$. We stop if the current solution z_i is feasible.

Using such an iterative penalty based approach, we have to solve a nonlinear program in each iteration. Given that the variable of interest is high dimensional (we talk about dimensions ≥ 50), we conclude from the previous complexity analysis that this is extremely demanding in terms of computational resources.

A.7 Listings

A.7.1 Generating the sample

The following script generates 1000 realizations (λ_t, Y_t) of the bivariate process in the approximate stationary regime. First, the auxiliary functions `function_true()` and `Count_Process()` are defined.

```
function z = function_true (a,b,c,d,x,y)
y = floor(y);
z = a + b.*x + c.* min(y,5) + d.*(sin((pi).*x)
      + cos((pi).* min(y,5)));
z = min(2,z);
end


function Z = Count_Process(a,b,c,d,n,m)
% the first m simulations are thrown away
D = zeros(n,2);
D(1,1) = 1;
D(2,1) = poissrnd(D(1,1));
for i = 2:n
    x = function_true(a,b,c,d,D(i-1,1),D(i-1,2));
    y = poissrnd(x);
    D(i,1) = x;
    D(i,2) = y;
end
Z = D(m+1:n,:);
end
```

```

%%% sample size
n = 1050;
m = 50;   % length of burn-in period

%%% parameters for the link function
a = 0.3;
b = 0.3;
c = 0.3;
d = -0.1;

%%% The sample
D = Count_Process(a,b,c,d,n,m);
Y = D(:,2);
X = D(:,1);
```

A.7.2 Estimation

We define the auxiliary function `B_spline_approx(vec,x,y)` that creates a function $S_n(\texttt{vec})\colon [0,2]^{72} \to [0,2]$, with the values of the 72-dimensional vector `vec` as coefficients, and evaluates it at the point (x,y).

```
function M = B_spline_approx(vec,x,y)
Mat = vec2mat(vec,12);
i = y + 1;
i = min(1,6);
knots = -0.4:.2:2.4;
coef = Mat(i,:);
M = fnval(spmak(knots,coef),x);
end
```

The function `objective_fun` calculates the value $Q(\texttt{vec},\boldsymbol{Y})$ for a given vector of coefficients, `vec`, and a vector of counts, $\boldsymbol{Y}$; cf. Lemma 4.2.10.

```
function z = objective_fun(vec,Y)
vec = 0.1*vec; % Only if discrete GA is used!
Xhat = Y;
Xhat(1) = mean(Y);
for i = 2:(length(Y))
   Xhat(i) = B_spline_approx(vec,Xhat(i-1),Y(i-1));
end
diff = zeros(length(Y),1);
for i = 2:(length(Y))
    diff(i) = (Y(i)-Xhat(i))^2;
end
z = sum(diff);
end
```

In the following script, the least squares spline estimator is calculated using the genetic algorithm with $A_n = \{0.0, 0.1, \ldots, 2.0\}$.

```
f = @(B,Y)objective_fun(B,Y);
fun = @(B)f(B,Y);
A0 = ones(72,1);
lb = zeros(72,1);
ub = ones(72,1);

IntCon = 1:72; %(Integer Constraints)
options = optimoptions('ga','ConstraintTolerance',1e-10);
fhat = 0.1*ga(fun, 72 ,[],[],[],[],lb,20*ub,[],IntCon,options);
```

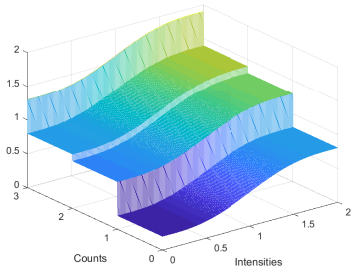
If we use GA, PSO, or SA, with $A_n = [0, M]$, we substitute the last line by

```
fhat = ga(fun, 72 ,[],[],[],[],lb,2*ub,[],[],options); %GA
fhat = particleswarm(fun,72,lb,2*ub); %PSO
fhat = simulannealbnd(fun,A0,lb,2*ub); %SA
```

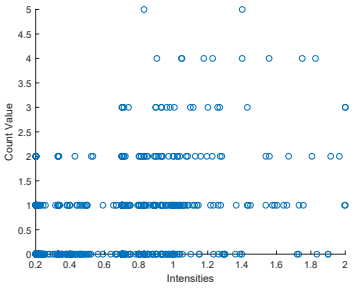
respectively.

A.8 Figures

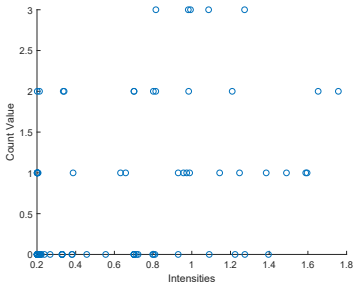
Underlying data for the computer experiments



(a) 3D plot of the true intensity function.



(b) 1000 realizations of the pairs (λ_t, Y_t)



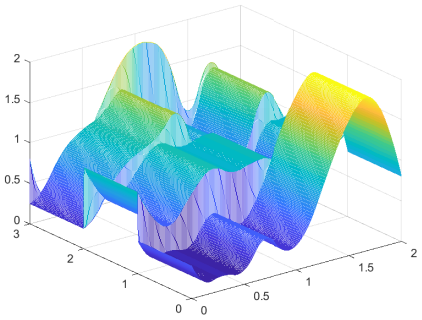
(c) 100 realizations of the pairs (λ_t, Y_t)

Figure A.8.1: Underlying data for the numerical experiments. In (a) the true link function is shown; (b) shows the 1000 realizations of the count process that were used for estimation in experiments 1–12; (c) shows the 100 realizations of the count process that were used for estimation in experiments 13–24. The Data in (b) and (c) were generated independently of each other.

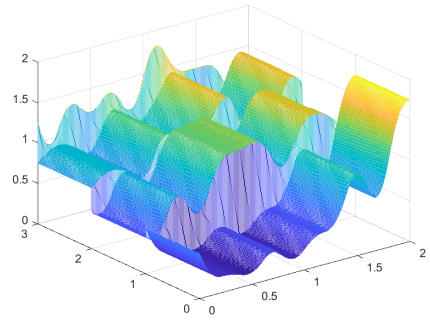
Estimations with sample size $n = 1000$

We show estimations of the link function shown in Fig. A.8.1a, on the basis of the data shown in Fig. A.8.1b. The least squares spline estimator was approximated using the GA (Figures A.8.2 and A.8.5), PSO (Figure A.8.3), and SA (Figure A.8.4). The mean square errors are displayed in Table A.9.1.

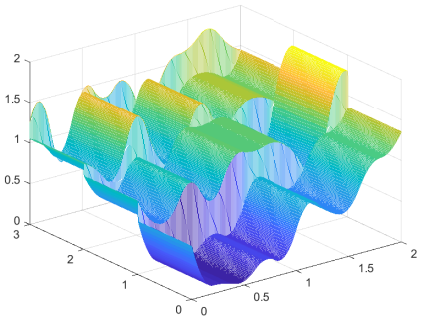
Figure A.8.2: Estimation with GA, $A_n = \{0, 0.2, \ldots, 1.8, 2.0\}$, $n = 1000$.



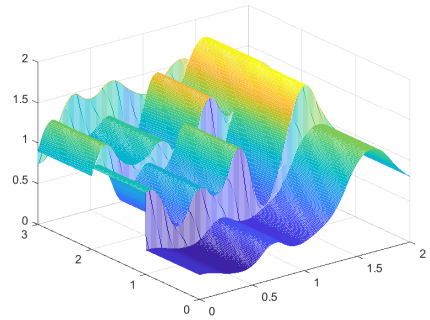
(a) Estimation in experiment no. 1



(b) Estimation in experiment no. 2

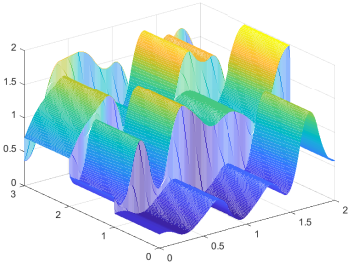


(c) Estimation in experiment no. 3

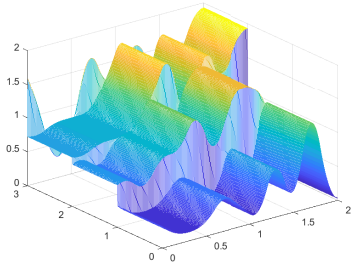


(d) Estimation in experiment no. 4

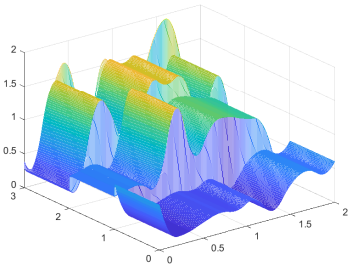
Figure A.8.3: Estimation with PSO, $A_n = [0, M]$, $n = 1000$.



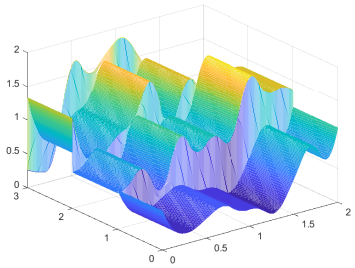
(a) Estimation in experiment no. 5



(b) Estimation in experiment no. 6

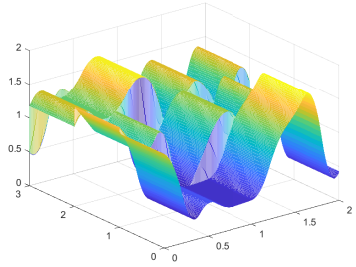


(c) Estimation in experiment no. 7

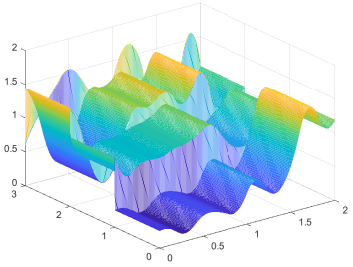


(d) Estimation in experiment no. 8

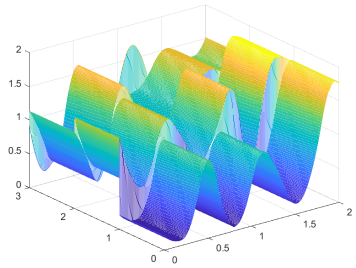
Figure A.8.4: Estimation with SA, $A_n = [0, M]$, $n = 1000$.



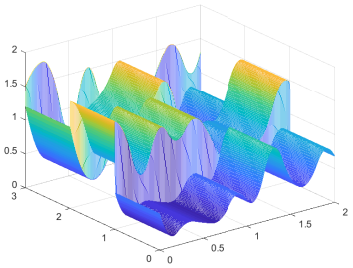
(a) Estimation in experiment no. 9



(b) Estimation in experiment no. 10

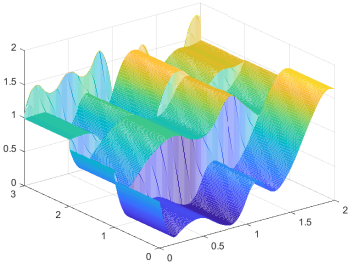


(c) Estimation in experiment no. 11

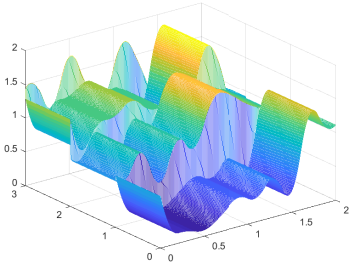


(d) Estimation in experiment no. 12

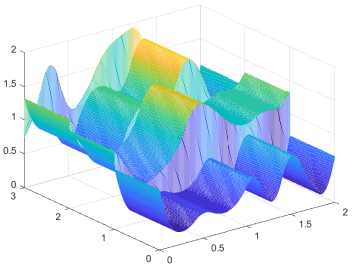
Figure A.8.5: Estimation with GA, $A_n = [0, M]$, $n = 1000$.



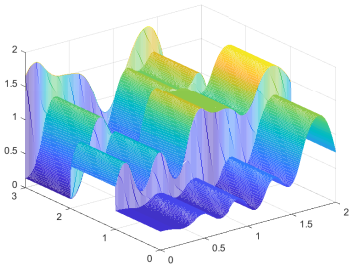
(a) Estimation in experiment no. 13



(b) Estimation in experiment no. 14



(c) Estimation in experiment no. 15

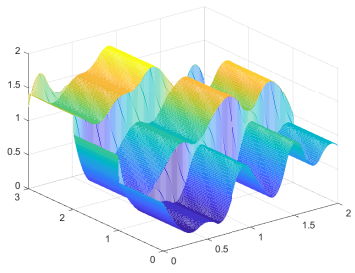


(d) Estimation in experiment no. 16

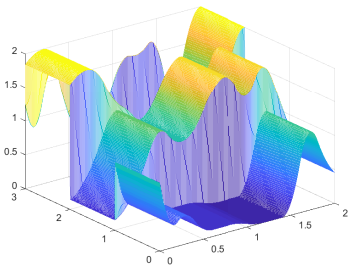
Estimations with sample size $n = 100$

We show estimations of the link function shown in Fig. A.8.1a, on the basis of the data shown in Fig. A.8.1b. The least squares spline estimator was approximated using the GA (Figures A.8.6 and A.8.9), PSO (Figure A.8.7), and SA (Figure A.8.8). The mean square errors are displayed in Table A.9.2.

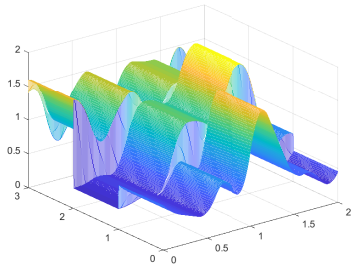
Figure A.8.6: Estimation with GA, $A_n = \{0, 0.2, \ldots, 1.8, 2.0\}$, $n = 100$.



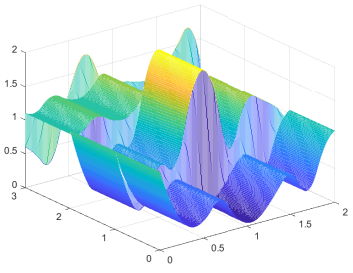
(a) Estimation in experiment no. 17



(b) Estimation in experiment no. 18

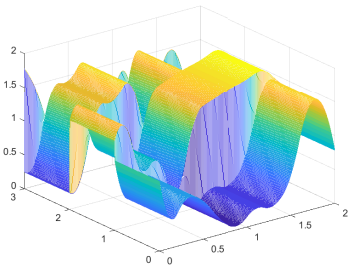


(c) Estimation in experiment no. 19

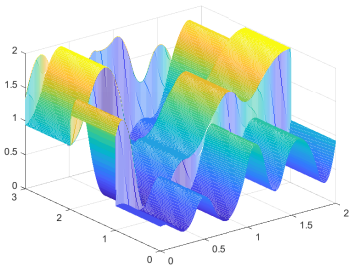


(d) Estimation in experiment no. 20

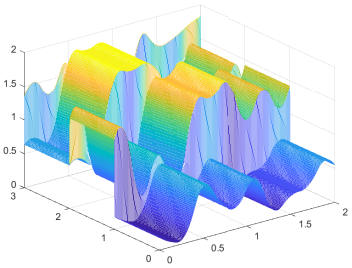
Figure A.8.7: Estimation with PSO, $A_n = [0, M]$, $n = 100$.



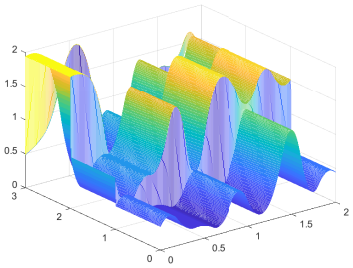
(a) Estimation in experiment no. 21



(b) Estimation in experiment no. 22

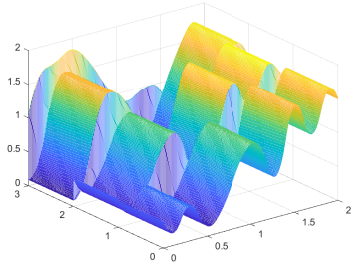


(c) Estimation in experiment no. 23

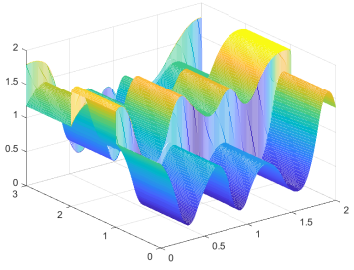


(d) Estimation in experiment no. 24

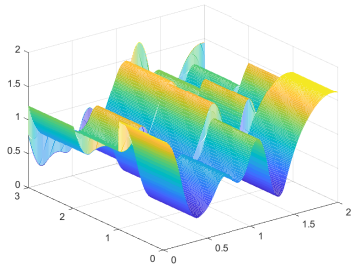
Figure A.8.8: Estimation with SA, $A_n = [0, M]$, $n = 100$.



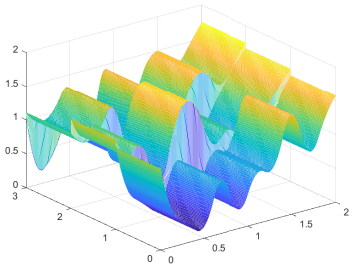
(a) Estimation in experiment no. 25



(b) Estimation in experiment no. 26

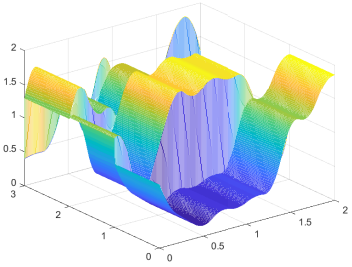


(c) Estimation in experiment no. 27

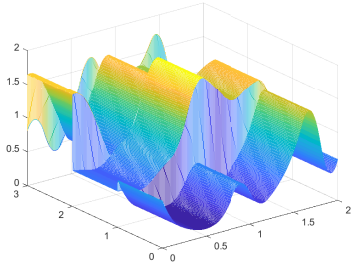


(d) Estimation in experiment no. 28

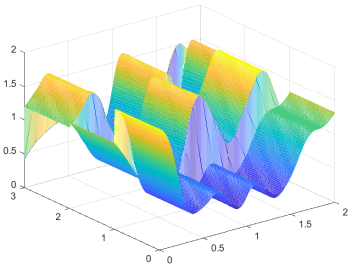
Figure A.8.9: Estimation with GA, $A_n = [0, M]$, $n = 100$.



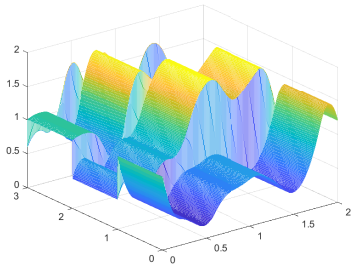
(a) Estimation in experiment no. 29



(b) Estimation in experiment no. 30



(c) Estimation in experiment no. 31



(d) Estimation in experiment no. 32

A.9 Tables

In the tables the suffix '-global' in the name of an algorithm refers to the fact that we used the set $A_n = [0, M]$ in this case. In contrast, the term 'GA-discrete' describes the usage of the genetic algorithm with $A_n = \{0.0, 0.1, \ldots, 1.9, 2.0\}$.

Table A.9.1: Experimental results for the sample size $n = 1000$.

Experiment No.	Algorithm	MSE
1	GA-discrete	0.0383
2		0.0277
3		0.0278
4		0.0237
5	PSO-global	0.0733
6		0.0364
7		0.0707
8		0.0320
9	SA-global	0.1874
10		0.0352
11		0.0763
12		0.0500
13	GA-global	0.0355
14		0.0331
15		0.0351
16		0.0420
Control Reference	$\overline{\lambda}_n$	0.1497

Table A.9.2: Experimental results for the sample size $n = 100$

Experiment No.	Algorithm	MSE
17	GA-discrete	0.0625
18		0.1473
19		0.1257
20		0.1133
21	PSO-global	0.1575
22		0.0825
23		0.1537
24		0.1745
25	SA-global	0.1136
26		0.1605
27		0.3121
28		0.0857
29	GA-global	0.1154
30		0.0793
31		0.3699
32		0.1053
Control Reference	$\overline{\lambda}_n$	0.1855

Bibliography

Adjiman, C., Androulakis, I., and Floudas, C. (1998a). A global optimization method, αbb, for general twice-differentiable constrained nlps—ii. implementation and computational results. *Computers & Chemical Engineering*, 22(9):1159 – 1179.

Adjiman, C., Dallwig, S., Floudas, C., and Neumaier, A. (1998b). A global optimization method, αbb, for general twice-differentiable constrained nlps — i. theoretical advances. *Computers & Chemical Engineering*, 22(9):1137 – 1158.

Aliprantis, C. and Border, K. (1994). *Infinite Dimensional Analysis: A Hitchhiker's Guide*. Springer-Verlag, Berlin Heidelberg.

Androulakis, I. P., Maranas, C. D., and Floudas, C. A. (1995). αbb: A global optimization method for general constrained nonconvex problems. *J. Global Optimization*, 7:337–363.

Barlow, R., Bartholomew, D., Bremner, J., and Brunk, H. (1972). *Statistical Inference Under Order Restrictions: The Theory and Application of Isotonic Regression*. J. Wiley, Chichester, UK.

Bass, R. (2013). *Real Analysis for Graduate Students*. CreateSpace Independent Publishing Platform.

Bélisle, C. J. P. (1992). Convergence theorems for a class of simulated annealing algorithms on $\mathbb{R}^d$. *Journal of Applied Probability*, 29(4):885–895.

Berbee, H. C. P. (1979). *Random Walks with stationary increments and renewal theory*. Math. Cent. Tracts, Amsterdam.

Boender, C. G. E. and Romeijn, H. E. (1995). Stochastic methods. In Pardalos, P. M. and Horst, R., editors, *Handbook of Global Optimization*, volume 2 of *Nonconvex Optimization and its Applications*, pages 829–869. Kluwer Academic Publishers, Dordrecht.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3):307 – 327.

Bonyadi, M. R. and Michalewicz, Z. (2017). Particle swarm optimization for single objective continuous space problems: A review. *Evolutionary Computation*, 25(1):1–54. PMID: 26953883.

Bradley, R. (2007). *Introduction to Strong Mixing Conditions, Volume 1*. Kendrick Press, Heber City, UT.

Brochu, E., Cora, V. M., and de Freitas, N. (2010). A tutorial on bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. arXiv:1012.2599v1.

Bull, A. D. (2011). Convergence rates of efficient global optimization algorithms. *The Journal of Machine Learning Research*, 12:2879–2904.

Burman, P., Chow, E., and Nolan, D. (1994). A cross-validatory method for dependent data. *Biometrika*, 81(2):351–358.

Calvin, J. M. (1997). Average performance of a class of adaptive algorithms for global optimization. *The Annals of Applied Probability*, 7(3):711–730.

Chen, Y. and Chen, M. (2010). Extended duality for nonlinear programming. *Computational Optimization and Applications*, 47(1):33–59.

Cox, D. R. (1981). Statistical analysis of time series: Some recent developments [with discussion and reply]. *Scandinavian Journal of Statistics*, 8(2):93–115.

Davis, R. A., Dunsmuir, W. T. M., and Streett, S. B. (2003). Observation-driven models for poisson counts. *Biometrika*, 90(4):777–790.

de Boor, C. (1978). *A Practical Guide to Splines*. Springer-Verlag, New York, NY.

Di Pillo, G. and Grippo, L. (1988). On the exactness of a class of non-differentiable penalty functions. *J. Optim. Theory Appl.*, 57(3):399 – 410.

Di Pillo, G. and Grippo, L. (1989). Exact penalty functions in constrained optimization. *SIAM J. Control Optim.*, 27(6):1333–1360.

Dierckx, P. (1995). *Curve and Surface Fitting with Splines*. Clarendon Press, Oxford, UK.

Doukhan, P. (1994). *Mixing: Properties and Examples*. Springer-Verlag, New York, NY.

Doukhan, P., Massart, P., and Rio, E. (1995). Invariance principles for absolutely regular empirical processes. *Annales de'l H.P., section B*, 31(2):393 – 427.

Doukhan, P. and Neumann, M. H. (2018). Absolute regularity of semi-contractive garch-type processes. arXiv:1711.04282v3. Forthcoming in *Journal of Applied Probability*.

Durbin, J. and Koopman, S. J. (1997). Monte carlo maximum likelihood estimation for non-gaussian state space models. *Biometrika*, 84(3):669–684.

Durbin, J. and Koopman, S. J. (2000). Time series analysis of non-gaussian observations based on state space models from both classical and bayesian perspectives. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 62(1):3–56.

Dzhaparidze, K. and van Zanten, J. H. (2001). On bernstein-type inequalities for martingales. *Stochastic Processes and their Applications*, 93:109–117.

Eichfelder, G., Gerlach, T., and Sumi, S. (2016). A modification of the α bb method for box-constrained optimization and an application to inverse kinematics. *EURO Journal on Computational Optimization*, 4(1):93–121.

Enders, W. (1995). *Applied econometric time series*. Wiley, Hoboken, NJ.

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica*, 50(4):987–1007.

Ferland, R., Latour, A., and Oraichi, D. (2006). Integer-valued garch process. *Journal of Time Series Analysis*, 27(6):923–942.

Finkel, D. E. (2004). Direct-a global optimization algorithm. `http://www4.ncsu.edu/~ctk/Finkel_Direct/`. accessed: March 29, 2018.

Fishman, G. (1996). *Monte Carlo: Concepts, Algorithms and Applications*. Springer-Verlag, New York, NY.

Fokianos, K., Rahbek, A., and Tjøstheim, D. (2009). Poisson autoregression. *Journal of the American Statistical Association*, 104(488):1430–1439.

Fokianos, K. and Tjøstheim, D. (2012). Nonlinear Poisson autoregression. *Annals of the Institute of Statistical Mathematics*, 64(6):1205–1225.

Gablonsky, J. M. (2001). *Modifications of the DIRECT Algorithm*. PhD thesis, North Caroline State University, Raleigh, NC.

Geiger, C. and Kanzow, C. (2002). *Theorie und Numerik restringierter Optimierungsprobleme*. Springer-Verlag, Berlin, Heidelberg.

Giné, E. and Nickl, R. (2016). *Mathematical Foundations of Infinite-Dimensional Statistical Models*. Cambridge University Press, New York, NY.

Györfi, L., Härdle, W., Sarda, P., and Vieu, P. (1989). *Nonparametric Curve Estimation from Time Series*. Springer-Verlag, New York, NY.

Györfi, L., Krzyzak, A., Kohler, M., and Walk, H. (2002). *A Distribution-Free Theory of Nonparametric Regression*. Springer-Verlag, New York, NY.

Hansen, P. and Jaumard, B. (1995). Lipschitz optimization. In Horst, R. and Pardalos, P. M., editors, *Handbook of Global Optimization*, volume 2 of *Nonconvex Optimization and its Applications*, pages 407–487. Kluwer Academic Publishers, Dordrecht.

Hardle, W. and Marron, J. S. (1985). Optimal bandwidth selection in nonparametric regression function estimation. *Ann. Statist.*, 13(4):1465–1481.

Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109.

Heinen, A. (2003). Modelling time series count data: An autoregressive conditional poisson model. `https://mpra.ub.uni-muenchen.de/id/eprint/8113`, access date: February 8, 2019.

Herman, C. (2005). *Statistical Physics Including Applications to Condensed Matter*. Springer-Verlag, New York, NY.

Ibragimov, I. A. (1962). Some limit theorems for stationary processes. *Theory of Probability & Its Applications*, 7(4):349–382.

Jennrich, R. I. (1969). Asymptotic properties of non-linear least squares estimators. *Ann. Math. Statist.*, 40(2):633–643.

Joines, J. A. and Houck, C. R. (1994). On the use of non-stationary penalty functions to solve nonlinear constrained optimization problems with ga's. In *Proceedings of the First IEEE Conference on Evolutionary Computation. IEEE World Congress on Computational Intelligence*, pages 579–584 vol.2.

Jones, D. R., Perttunen, C. D., and Stuckman, B. E. (1998). Efficient global optimization of expensive black-box functions. *J. Global Optim.*, 13(4).

Jung, R. C. and Tremayne, A. R. (2011). Useful models for time series of counts or simply wrong ones? *AStA Advances in Statistical Analysis*, 95(1):59–91.

Karatzas, I. and Shreve, S. E. (1991). *Brownian Motion and Stochastic Calculus*. Springer-Verlag, New York, NY.

Kedem, B. and Fokianos, K. (2002). *Regression Models for Time Series Analysis*. Wiley, Hoboken, NJ.

Kennedy, J. and Eberhart, R. (1995). Particle swarm optimization. In *Proceedings of the IEEE International Conference on Neural Networks, 1995*, volume 4, pages 1942–1948, Perth, WA.

Kolmogorov, A. N. and Tikhomirov, V. M. (1993). ε-entropy and ε-capacity of sets in function spaces. In Shiriyayev, A. N., editor, *Selected Works of A. N. Kologorov*, volume III, pages 86–170. Kluwer Academic Publishers, Dordrecht.

Königsberger, K. (2004). *Analysis 1*. Springer-Verlag, Berlin.

Liberti, L. S. (2004). *Reformulation and Convex Relaxation Techniques for Global Optimization*. PhD thesis, Imperial College London.

Lindvall, T. (1992). *Lectures on the Coupling Method*. Wiley, New York, NY.

Lyche, T. and Mørken, K. (2008). Spline methods draft. `http://www.uio.no/studier/emner/matnat/ifi/INF-MAT5340/v10/undervisningsmateriale/book.pdf`. accessed February 15, 2018.

Malherbe, C. and Vayatis, N. (2017). Global optimization of Lipschitz functions. In Precup, D. and Teh, Y. W., editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 2314–2323, Sydney, AU. PMLR.

Maranas, C. D. and Floudas, C. A. (1994). Global minimum potential energy conformations of small molecules. *Journal of Global Optimization*, 4(2):135–170.

Meister, A. and Kreiß, J.-P. (2016). Statistical inference for nonparametric garch models. *Stochastic Processes and their Applications*, 126(10):3009 – 3040.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of State Calculations by Fast Computing Machines. *J. Chem. Phys.*, 21:1087–1092.

Meyn, S. and Tweedie, R. (2009). *Markov Chains and Stochastic Stability*. Cambridge University Press, Cambridge, UK.

Millionas, M. (1994). Swarms phase transitions and collective intelligence. In Langton, C. G., editor, *Artificial Life III*. Addison Wesley, Reading, MA.

Mockus, J. (1989). *Bayesian Approach to Global Optimization*. Mathematics and its Applications (Sov. Ser.). Kluwer Academic Publishers, Dordrecht.

Nemirovsky, A. and Yudin, D. (1978). Efficiency of randomization of control. In *Problems of random search*. Zinate, Riga. In Russian.

Nemirovsky, A. and Yudin, D. (1983). *Problem Complexity and Method Efficiency in Optimization*. Wiley, Chichester, UK.

Neumann, M. H. (2011). Absolute regularity and ergodicity of poisson count processes. *Bernoulli*, 17(4):1268–1284.

Novak, E. (1988). *Deterministic and Stochastic Error Bounds in Numerical Analysis*. Number 1349. Springer-Verlag, Berlin, Heidelberg.

Papadimitriou, C. H. and Steiglitz, K. (1998). *Combinatorial Optimization: Algorithms and Complexity*. Dover Publications, Minneola, NY.

Piyavskii, S. (1972). An algorithm for finding the absolute extremum of a function. *USSR Computational Mathematics and Mathematical Physics*, 12(4):57 – 67.

Powell, M. (1981). *Approximation Theory and Methods*. Cambridge University Presss, Cambridge, UK.

Rios, L. M. and Sahinidis, N. V. (2013). Derivative-free optimization: a review of algorithms and comparison of software implementations. *J. Glob. Optim.*, 56:1247–1293.

Robert Koch-Institut (2019). SurvStat@RKI 2.0. `https://survstat.rki.de`, access date: February 06, 2019.

Romeijn, H. E. and Smith, R. L. (1994). Simulated annealing for constrained global optimization. *Journal of Global Optimization*, 5(2):101–126.

Rosenblatt, M. (1956). A central limit theorem and a strong mixing condition. *Proceedings of the National Academy of Sciences of the United States of America*, 42(1):43–47.

Rudemo, M. (1982). Empirical choice of histograms and kernel density estimators. *Scandinavian Journal of Statistics*, 9(2):65–78.

Rudin, W. (1976). *Principles of Mathematical Analysis*. McGraw-Hill, New York, NY.

Rydberg, T. H. and Shephard, N. (2000). BIN Models for Trade-by-Trade Data. Modelling the Number of Trades in a Fixed Interval of Time. Econometric Society World Congress 2000 Contributed Papers 0740, Econometric Society.

Shao, J. (2003). *Mathematical Statistics*. Springer-Verlag, New York, NY.

Shephard, N. (1996). Statistical aspects of arch and stochastic volatility. In Cox, D., Hinkley, D., and Barndorff-Nielsen, O., editors, *Time Series Models: In econometrics, finance and other fields*. Taylor & Francis, London, UK.

Shiryayev, A. (1984). *Probability*. Springer, New York, NY.

Shorack, G. (2017). *Probability for Statisticians*. Springer-Verlag, New York, NY.

Shubert, B. O. (1972). A sequential method seeking the global maximum of a function. *SIAM Journal on Numerical Analysis*, 9(3):379–388.

Sim, T., Douc, R., and Roueff, F. (2016). General-order obersvation-driven models. `https://hal.archives-ouvertes.fr/hal-01383554`, access date: February 8, 2019. working paper or preprint.

Spall, J. (2003). *Introduction to stochastic search and optimization: estimation, simulation, and control*. Wiley, Hoboken, NJ.

Stone, C. J. (1984). An asymptotically optimal window selection rule for kernel density estimates. *Ann. Statist.*, 12(4):1285–1297.

Suman, B. and Kumar, P. (2006). A survey of simulated annealing as a tool for single and multiobjective optimization. *Journal of the Operational Research Society*, 57(10):1143–1160.

Traub, J. F., Wasilkowski, G. W., and Wozniakowski, H. (1988). *Information-based complexity*. Acad. Press, Boston, MA.

Tsybakov, A. (2008). *Introduction to Nonparametric Estimation*. Springer-Verlag, New York, NY.

Van de Geer, S. (2000). *Empirical Processes in M-Estimation*. Cambridge University Press, Cambridge, UK.

Van der Vaart, A. (2000). *Asymptotic Statistics*. Cambridge University Press, New York, NY.

Van der Vaart, A. W. and Wellner, J. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer-Verlag, New York, NY.

Vavasis, S. A. (1991). *Nonlinear Optimization: Complexity Issues*. Oxford University Press, New York, NY.

Volkonskii, V. A. and Rozanov, Y. A. (1959). Some limit theorems for random functions. i. *Theory of Probability & Its Applications*, 4(2):178–197.

Wah, B. W., Chen, Y., and Wang, T. (2007). Simulated annealing with asymptotic convergence for nonlinear constrained optimization. *Journal of Global Optimization*, 39(1):1–37.

Wasilkowski, G. W. (1984). Some nonlinear problems are as easy as the approximation problem. *Com. Math. Appl.*, 10(4/5):351–363.

Weise, T. (2009). Global optimization algorithms – theory and application. `http://www.it-weise.de/projects/book.pdf`. accessed: April 5, 2018.

Ehrenwörtliche Erklärung

Ich erkläre ehrenwörtlich,

dass mir die Promotionsordnung der Fakultät für Mathematik und Informatik der Friedrich-Schiller-Universität bekannt ist;

dass ich die vorliegende Dissertation selbst angefertigt habe, keine Textabschnitte eines Dritten oder eigener Prüfungsarbeiten ohne Kennzeichnung übernommen und alle von mir benutzten Hilfsmittel, persönlichen Mitteilungen und Quellen in dieser Arbeit angegeben habe;

dass ich die Hilfe einer Promotionsberaterin oder eines Promotionsberaters nicht in Anspruch genommen habe und dass Dritte weder unmittelbar noch mittelbar geldwerte Leistungen von mir für Arbeiten erhalten haben, die im Zusammenhang mit dem Inhalt der vorliegenden Dissertation stehen;

dass ich die vorliegende Dissertation noch nicht als Prüfungsarbeit für eine staatliche oder andere wissenschaftliche Prüfung eingereicht habe;

dass ich weder eine gleiche noch eine in wesentlichen Teilen ähnliche Abhandlung bei einer anderen Hochschule als Dissertation eingereicht habe.

Jena, der 30. Oktober 2019 Maximilian Wechsung