Non-Parametric Inference and Tracking for Poisson Processes

CIP-DATA LIBRARY: Technische Universiteit Eindhoven

Non-Parametric Inference and Tracking for Poisson Processes / Serra, Paulo

A catalogue record is available from the Eindhoven University of Technology Library.
ISBN: 978-90-386-3439-5
2010 Mathematics Subject Classification: 62G05, 62G20, 62L20, 62M10, 62M05.
Subject headings: Nonparametric Inference, Sequential Methods, Inference from Stochastic Processes.

Printed by Wöhrmann Print Services, Zutphen, the Netherlands.

Non-Parametric Inference and Tracking for Poisson Processes

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de
Technische Universiteit Eindhoven, op gezag van de
rector magnificus, prof.dr.ir. C.J. van Duijn, voor een
commissie aangewezen door het College voor
Promoties in het openbaar te verdedigen op
maandag 23 september 2013 om 14.00 uur

door

Paulo Jorge de Andrade Serra

geboren te Lissabon, Portugal

Dit proefschrift is goedgekeurd door de promotiecommissie:

voorzitter: prof.dr. E.H.L. Aarts

promotor: prof.dr. J.H. van Zanten

copromotor: dr. E.N. Belitser

leden: prof.dr. A. van der Vaart (Universiteit Leiden)
prof.dr. A. Juditsky (Université Joseph Fourier)
prof.dr. C.C.P. Snijders
prof.dr. R.W. van der Hofstad

adviseur: dr. R.M. Pires da Silva Castro

Acknowledgments

THIS THESIS could not have been completed without the help and support – either direct or, indirect – of a large number of people. I would then like to mention some of them in acknowledgment of their contribution to this work having materialized into what you are now reading. First, some words for the family.

Para toda a família: obrigado. Obrigado, especialmente para os meus pais, que sempre me apoiaram, me darem liberdade, e para quem tenho uma dívida que nunca serei capaz de pagar. Para todos os amigos: obrigado, também. Obrigado por me fazem sentir bem vindo sempre que vos vejo, mesmo depois de todos estes anos e de me tratarem como se nunca me tenha ido embora. Obrigado em particular para a Ana pelos constantes esforços para me civilizar e para a Alexandra, Linda, Vanda e todos os outros por a ajudarem.

I would also like to thank all the colleagues I've had, and all the friends I've made over these last few years. Thanks in particular to Paul for being Paul and making sure I never have a boring moment, whether I want to or not. I'd also like to thank all the members[1] of our pub quizzing team, *Pony Farm*, for all the really fun Thursday nights and for teaching me that you can be awful at something and still enjoy it very much! Thank you all.

Living in Eindhoven was a very good experience and it also gave me a very friendly, relaxed and stimulating work environment. This nice environment is also, of course, due to the very many [2] wonderful officemates I had during these last four years. In particular I'd like to mention Florence, who was a constant presence in our shared EURANDOM office – either in physical form or, vicariously, through the massive mountain of articles that almost occluded her desk[3] – and whose cheerfulness always lightened the day. And what to say about Botond? Certainly that I could not have wished for a better officemate. Always enthusiastic, cheerful and dependable, an indispensable spice to both a fun day and a productive day. Thank you both, thank you all.

A word of appreciation goes out also to the members of the defense committee: Aad van der Vaart, Anatoli Juditsky, Chris Snijders, Remco van der Hofstad and last, but certainly not least, Rui Castro. Thank you for your willingness to be in the committee, the time you spent reading this thesis and for your helpful feedback.

I would, of course, also like to thank my supervisors for all their help and support over these last few years. To Harry, thank you for always being accessible and helpful, and for all the useful insight that you somehow always managed to convey in a very understandable manner. I'd also like to thank Eduard a lot for his massive support. It was clear to me the moment I met him for the first time, five years ago, to come up with a theme for my master thesis, that he was full of good ideas. But it was only through the years, that I also started to become aware and grow to appreciate the impressive thoroughness and preciseness of the way in which he works. This is something I look up to and hope one day be able to do as well. Thank you both for all your patience, your help and all the nice talks we had over the years.

Finally I would also like to thank Yoni Nazarathy and Avishai Mandelbaum for their help obtaining the data set which turned out to be an important part of Chapters 2 and 3, Mark Bebbington for providing us with the expressions for the test functions used in the simulation study in Chapter 2 and to all the staff at the TU/e and at EURANDOM for their friendliness and helpfulness.

[1] At least twenty nine persons, by my count, including regular participants, occasional participants, reluctant participants, special guests, innocent victims and clueless bystanders.

[2] "Very many" means eight; chronologically: Mitja, Botond, Jan-Pieter, Florence, René, Haralambie, Ervin and occasionally also Bartek.

[3] Thank you also, kind office desk designers; it is very impressive that Florence's desk could withstand that much volume of paper.

Non-Parametric Inference and Tracking for Poisson Processes

Non-parametric statistics provides a framework which is rich with tools that allow to make inference on high- or infinite-dimensional models under very weak assumptions on the underlying structure of the data. Poisson processes have a long-standing history as some of the most widely used processes, not just in Statistics, but also in fields such as communication, meteorology, seismology, hydrology, astronomy, biology, medicine, actuary sciences and queueing. This thesis is divided into two parts: Inference and Tracking. Most of the results in the first part apply specifically to Poisson processes. In the second part we work with more general models but some of the results there can also be applied to these processes.

Poisson processes are often used for modeling periodic time-varying phenomena. We study a semi-parametric estimator for the period of the cyclic intensity function of an inhomogeneous Poisson process. We make no parametric assumptions on the intensity function which is treated as an infinite dimensional nuisance parameter. A new family of estimators is proposed for the period of the intensity function; we address identifiability and consistency issues and present simulations which demonstrate good performance of the proposed estimation procedure. We compare our method to competing methods on synthetic data and apply it to a real data set from a call center.

The starting point for our next problem is also this call centre data. Having concluded that the data presents daily periodicity, we model it as Poisson process and propose methods to estimate its daily intensity function. More specifically, we use non-parametric Bayesian methods. We start by showing some estimates for the call center data obtained via MCMC for free knot B-spline based priors to which we will return later. The theoretical performance of these priors falls within the general approach that we apply for making Bayesian inference on Poisson processes. Our results cover the case when the process is observed in full but also the case when only a discretized version of it is observed which is quite important in applications. Under the assumption that the intensity function is an α-smooth function, these results imply that our B-spline based adaptive (i.e. without using knowledge of the smoothness α) prior delivers adaptive rates of contraction for estimating the parameter of the Poisson process. We present further results about using general stochastic process priors, endowing the scale of the intensity function with a prior and about priors on monotonous intensities.

Spline-based priors like the ones we have used for making inference on the call centre data are quite popular in Bayesian non-parametric statistics. Practitioners commonly use these priors, mostly due to their flexibility and to the ease with which one can design MCMC algorithms for them. The number and the location of the knots as well as the respective B-spline coefficients are endowed with a prior. Although the practice of using random knots is commonplace in applications, theoretical results on these priors were still missing in the literature. Under some mild, sufficient conditions, rates are attained for these adaptive priors. In doing so, we propose a theoretical framework that can be used to motivate certain design decisions when selecting priors that would otherwise be made arbitrarily. We present some numerical results for synthetic non-parametric regression data to illustrate

the advantages of working with random knots when estimating spatially inhomogeneous functions.

In the second part of this thesis we treat some tracking problems. Tracking refers to a framework where we sequentially collect data from a distribution indexed by a parameter that is slowly changing in time. This problem is of fundamental interest in sequential analysis and has many applications in signal processing, speech recognition, communication systems, neural physiology, ecology and econometrics. An important problem which we can fit into this framework is that of quantile regression. Paralleling regression, where we are interested in estimating the conditional mean of a response variable given some covariates, in quantile regression we want to estimate a conditional quantile. This alternative take on the problem of regression results in more robust estimates and if we estimate several quantiles simultaneously this gives us a quite complete description of the evolution of a regression function in time. Our results apply to a more general setting than quantile regression since we allow for conditional dependences between consecutive observations which is quite natural in the context of sequential sampling; also, we allow the level of the quantile being tracked to change in time. The algorithm we propose is recursive and can therefore be implemented in a straightforward and efficient way and we derive non-asymptotic, uniform bounds on its approximation error.

Finally, we consider also the problem of tracking a drifting multivariate parameter in a more general context: a time series is observed, each observation depending on a parameter which we allow to vary slowly. This constitutes a growing statistical model whose time-varying parameter we would like to track. Instead of assuming that we know the model completely, we only presume to have access to a so-called gain function that depends on the previous estimate and on a new observation. This gain function can be used, together with a step sequence, to update an arbitrary approximation of the drifting parameter. Applying this procedure sequentially results in a tracking algorithm. We derive a non-asymptotical, uniform error bound on the error of the tracking sequence and specify what this bound becomes for different assumptions on the variability of the drifting parameter. What constitutes a proper gain function depends on the parameter (or functional of the parameter) of interest and on the dependence of the model on the drifting parameter. We outline how gain functions can be constructed for general models and how they can eventually be modified to verify our assumptions. The problem of tracking drifting parameters in a d-dimensional autoregressive model is treated in detail, along with some simpler examples to illustrate the method. Some numerical results are also presented.

Notation

$\mathbb{N}$, $\mathbb{N}_0$, $\mathbb{Z}$, $\mathbb{Q}$	natural numbers, natural numbers including zero, integers, rational numbers
$\mathbb{R}$, $\mathbb{R}^+$, $\mathbb{R}_0^+$	real numbers, positive real numbers, non-negative real numbers
$\lesssim$, $\gtrsim$	less or equal (resp. larger or equal) up to a universal constant
$\ll$, $o(1)$; $\asymp$, $O(1)$	of a smaller order, of the same order
$[a]$, $\lfloor a \rfloor$, $\lceil a \rceil$	integer part of a, largest (resp. smallest) integer no larger (resp. smaller) than a
$a \pmod b$	a moludus b; remainder of the division of a by b
$a \wedge b$, $\min(a, b, \dots)$	minimum between a and b, minimum between $a, b, \dots$
$a \vee b$, $\max(a, b, \dots)$	maximum between a and b, maximum between $a, b, \dots$
$\mathbb{R}^d$	d-dimensional Euclidian space
$\mathbf{a}$, $\mathbf{a_d}$; $\mathbf{A}$, $\mathbf{A}_d$	vector in $\mathbb{R}^d$, random vector in $\mathbb{R}^d$
$\langle \mathbf{a}, \mathbf{b} \rangle$	inner product between $\mathbf{a}$ and $\mathbf{b}$
$\lVert \cdot \rVert_p$	l_p norm on $\mathbb{R}^d$ or L_p norm on L^2
$\mathbb{C}$	complex numbers
$[a_{i,j}]_{i,j}$	matrix with entries $a_{i,j}$
I, I_d, O, O_d, J, J_d	Identity of order d, zero matrix of order d, exchange matrix of order d
$\det(M)$, $\mathrm{tr}(M)$, $\mathrm{vect}(M)$	determinant of M, trace of M, column vector containing all entries of M
$\mathrm{diag}(\mathbf{a})$, $\mathrm{diag}(M)$	diagonal matrix containing entries of $\mathbf{a}$, vector with entries of main diagonal of M
$\mathscr{L}(M)$, $\mathscr{L}^\perp(M)$	linear space spanned by the columns of M, and its orthogonal complement
$\dim\big(M\big)$, $\dim\big(\mathscr{L}(M)\big)$	Dimension of the linear space spanned by the columns of M
$\lambda_{(i)}(M)$	i-th largest eigenvalue of M
$M > 0$, $M \geq 0$	M is positive (semi-)definite
$\lVert\!\lVert M \rVert\!\rVert_p$, $\lVert\!\lVert M \rVert\!\rVert_2$	operator norm induced by l_p, spectral norm (operator norm induced by l_2)
$\mathbb{1}\{\cdot\}$, $\mathbb{1}_{\{\cdot\}}$	indicator of $\{\cdot\}$
log, exp, e	natural logarithm, exponential function, Euler's number
$\mathcal{S}_{\mathbf{k}}^q$	space of all splines of order q with knots $\mathbf{k}$
$\big\{B_{\mathbf{k},1}^q, \dots, B_{\mathbf{k},j}^q\big\}$	B-spline basis for $\mathcal{S}_{\mathbf{k}}^q$
$s_{\mathbf{k},\boldsymbol{\theta}}$	spline with knots $\mathbf{k}$ and B-spline coefficients $\boldsymbol{\theta}$
$\mathrm{supp}\, f$	support of f
$[t_1, \dots, t_r]f$	r-th order divided difference of f over $t_1, \dots, t_r$
$f^{(0)}$, $f^{(r)}$	function f, r-th derivative of f
$\nabla_{\mathbf{a}} f$, $\boldsymbol{\partial \mathbf{f} / \partial \mathbf{a}}$	gradient of f with respect to $\mathbf{a}$, Jacobian of vector a valued $\mathbf{f}$ with respect to $\mathbf{a}$

$\mathcal{F}_\alpha$	space of functions with regularity α
$\mathcal{F}_\mathcal{A}$	family of spaces of function with regularity $\alpha \in \mathcal{A}$
$\mathcal{L}_\alpha$, $\mathcal{L}_\alpha(I)$, $\mathcal{L}_\alpha(L, I)$	Lipschitz space of function on I with Lipschitz constant L
$\mathcal{H}_\alpha$, $\mathcal{H}_\alpha(I)$, $\mathcal{H}_\alpha(L, I)$	Hölder space of function on I with Hölder constant L
$\mathcal{W}_\alpha$, $\mathcal{W}_\alpha(I)$, $\mathcal{W}_\alpha(L, I)$	Sobolev space of function on I with radius L
$N(\varepsilon, A, d)$	covering number of $A \subseteq \mathcal{A}$ using ε-balls according to the metric d
A^c, $\mathcal{A} \backslash A$	complement of $A \subseteq \mathcal{A}$
$\big(M, \mathcal{M}\big)$, $\big(M, \mathcal{M}, \mu\big)$	measurable space, measure/probability space
a.s., a.e.	(μ-) almost surely (μ-) almost every
P_θ, $P_\theta\big(\cdot \mid X\big)$	distribution, conditional distribution
$\mathbf{X}^{(n)}$, $\mathbf{X}_n$, $\mathbf{X}_{n,d}$	sample of size n, time series up to time n, d observations leading up to X_n
$\big(\mathfrak{F}_n, n \in \mathbb{N}\big)$	filtration
$\sigma(\mathbf{X}_n)$	σ-algebra generated by $\mathbf{X}_n$, natural filtration
$p_\theta^{(n)}\big(\mathbf{X}^{(n)}\big)$	likelihood of $\mathbf{X}^{(n)}$
$\mathcal{P} = \{P_\theta : \theta \in \Theta\}$	parametrized model
$X \sim P$	X distributed according to P
$\mathbb{E}[X \mid \mathfrak{F}]$, $\mathbb{E}[X \mid \mathbf{X}]$	conditional expectation
$\mathbb{P}_\theta$, $\mathbb{E}_\theta$, $\mathbb{V}_\theta$	probability, expectation, variance with respect to P_θ, resp.
$\mathbb{V}(\mathbf{X})$	variance-covariance matrix of the random vector $\mathbf{X}$
$o_P(1)$	converges to zero in probability
$O_P(1)$	bounded in probability
$\xrightarrow{\mathbb{P}}$, $\xrightarrow{a.s.}$	convergence in probability, almost sure convergence
R_n^W, R_n^p	L_p risk
$r_n^p(\Theta)$	minimax risk over Θ with respect to L_p loss
$M_n(\boldsymbol{\theta})$, $\Psi_n(\boldsymbol{\theta})$	criterion function (M- or Z-)
$m(\boldsymbol{\theta})$, $\psi(\boldsymbol{\theta})$	limiting criterion function (M- or Z-)
Rx, $R^n x$, $R^{-1}x$	operator R acting on x, operator R acting n times on x, pre-image of x via R
$\pi(\theta)$, $\Pi(\theta)$	density of prior on θ, prior measure on θ
$\pi(\theta \mid \mathbf{X}^{(n)})$, $\Pi(\theta \mid \mathbf{X}^{(n)})$	posterior density on θ, posterior measure on θ
$h(\theta, \vartheta)$, $h(p_\theta, p_\vartheta)$	Hellinger distance between p_θ and p_ϑ
$K(\theta, \vartheta)$, $K(p_\theta, p_\vartheta)$	Kullback-Leibler divergence between p_θ and p_ϑ
$V(\theta, \vartheta)$, $V(p_\theta, p_\vartheta)$	a Csiszár f-divergence between p_θ and p_ϑ
$B(\varepsilon, \theta)$	Kullback-Leibler ball centred at θ with radius ε
$A(\boldsymbol{\theta} \to \boldsymbol{\vartheta})$	MCMC acceptance probability for the move $\boldsymbol{\theta} \to \boldsymbol{\vartheta}$
φ_d, Φ	d-variate standard normal p.d.f., standard normal c.d.f.

Contents

1

Introduction

This chapter contains some introductory material for the remaining content of this thesis. We present a broad overview of some important concepts in Statistics providing details only when the material is relevant for what follows in the subsequent chapters. There will be some focus on basic elements of Mathematical Statistics with an emphasis on non-parametric estimation. M-estimation, stochastic approximation, non-parametric Bayesian statistics and adaptation will be treated in more detail, since they constitute the main content of the subsequent material. We also summarize the scope of the remaining chapters, including the models they apply to, the methods used, the results obtained and the applications of these results.

1.1 Non-Parametric Statistics and Inference

Statistics is a very diverse field which encompasses the study of quite a number of problems connected with extracting information from data. More specifically, we might want to summarize data, design experiments, model or infer upon certain phenomena, pick among competing alternatives, assess the uncertainty of a certain approximation, keep track of a varying quantity, forecast the evolution of a given phenomenon, etc. In mathematical statistics these tasks are performed on solid theoretical ground by using tools from different fields of Mathematics; primarily Probability Theory, but also Algebra, Analysis, Approximation Theory, Ergodic Theory, Functional Analysis, Information Theory, Measure Theory, Numerical Analysis and Stochastic Process Theory.

The central object of interest in Mathematical Statistics is the model and its relation with observed data. Models are collections of probability distributions which represent the different possible ways in which data may have been produced. These, in turn, are not observable, but data are, and carry information about their distribution and therefore about the model. Data can thus be used to perform what is arguably the most fundamental task in Statistics: inference; to learn from data. Inference is used to answer the following question: Given a model $\mathcal{P}$ and data produced by a fixed distribution from this model, say P_0, what can be said about P_0 based on the observed data? In this thesis we will be mostly concerned with inference. We will be working with certain fixed models, or sometimes with families of models, and we will use different techniques to approximate or estimate P_0, based on observed data, in a number of different settings.

Although in this thesis we will be mostly concerned with deriving theoretical results, we will, at some points, use the methods we develop on actual data. When working with real data – meaning data produced by some physical phenomenon – one might wonder how well a model could ever represent the mechanism producing such data. After all, no data, simulated or otherwise, is, to the best of anyone's knowledge, actually being produced by a probability measure. The data should instead be thought of as being produced by some unknown mechanism; the model should only be expected to act as a proxy for this mechanism and capture the main (probabilistic) features of the mechanism. The *true* distribution P_0 can be thought of as the best approximation in model $\mathcal{P}$ for this mechanism. To perform inference, we assume that the data were indeed generated according to some unknown "true" distribution $P_0 \in \mathcal{P}$ and then determine which element in $\mathcal{P}$ best plays this role. This justifies, to some extent, our use of *large*, infinite-dimensional models, to do what is known as *non-parametric* statistics. In the next section we start by formalizing the concept of *size* of a model.

1.2 Models and Parametrization

It should be intuitively clear that the choice of the model $\mathcal{P}$ is very important when making inference. If we work with a *large* model, containing many possible distributions, then we will have more flexibility when constructing estimators and will arguably be able to find a better approximation for the underlying mechanism generating the data. To make the concept of size of a model more precise, it is customary to *parametrize* the model. We then see the model $\mathcal{P}$ as the collection

$$\mathcal{P} = \{P_\theta : \theta \in \Theta\},$$

where we say that θ is the *parameter* of the model; it lives in the parameter set Θ and indexes the distributions P_θ in the model $\mathcal{P}$. A parametrization is simply a labelling of the distributions in the model and so it is not unique. We will, however, always make sure that parametrizations are such that if $P_\theta = P_{\theta'}$ then $\theta = \theta'$ (i.e. the map

$\theta \mapsto P_\theta$ is a bijection) in which case we say that the parameter θ is *identifiable*. It means that we can identify a distribution from the parameter that indexes it, making the terms *distribution* and *parameter* interchangeable, as well as the terms *model* and *parameter set*. The parameter set Θ will itself usually be a metric space (Θ, d) and the parametrization will be made in such a way that models whose parameters are *close* (according to the metric d) will be close themselves; cf. (1.22) for an example of a metric on $\mathcal{P}$.

If the parameter set Θ is such that, without loss of generality, $\Theta \subset \mathbb{R}^d$ for some $d \in \mathbb{N}$, then we call $\mathcal{P}$ a *parametric* model. If $\Theta \subset \mathcal{F}$ where $\mathcal{F}$ is an infinite-dimensional space (of, say, functions or sequences), then we call $\mathcal{P}$ a *non-parametric* model. In the mixed case, where $\Theta \subset \mathbb{R}^d \times \mathcal{F}$, we call $\mathcal{P}$ a *semi-parametric* model; sometimes we will also refer to a semi-parametric models as having a *parametric part* and a *non-parametric part* (here a subset of $\mathbb{R}^d$ and a subset of $\mathcal{F}$ respectively).

Typically we will be interested in estimating $\phi(\theta_0)$, some functional of the unknown parameter θ_0. This functional can be the parameter θ_0 itself, although one might also be interested in means, variances, modes or quantiles of P_{θ_0}. In non-parametric models, if θ_0 is a curve, we might also be interested in estimating a point of maximum or a derivative of θ_0, or θ_0 at a point. In the semi-parametric models, we will often only be interested in the parametric part of the model, in which case the non-parametric part will be referred to as a nuisance parameter.

We work with all three types of models in this thesis but we focus mostly on non-parametric models and methods. Roughly speaking, working in a non-parametric setting corresponds to making less assumptions about the underlying mechanism which generates the data. This results in a larger set of distributions being considered. Working with non-parametric models sometimes translates into methods that are, in a sense, simpler and more universal in their applicability, requiring less input from the user and being less dependent on the specific nature of the data. The difficulty in the use of non-parametric models is in establishing theoretical properties for the estimates; this follows from the fact that less assumptions are made about the data and that the models are allowed to be more versatile, and therefore more complex. Computational and numerical issues may also arise on occasion due to the use of high-dimensional objects. We walk then the line between making as few assumptions as possible – and in doing so increasing the applicability and flexibility of the method – and making enough assumptions – so that we still have tractable models and also enough structure to be able to establish theoretical properties and assure some level of precision.

We will defer our presentation of examples of estimation methods to Sections 1.8, 1.10 and 1.11 since these are already quite connected with the content of the remaining chapters. Before this, we focus on sampling, different types of data and some of the models we will be treating in later chapters.

1.3 Sampling

Consider a generic, parametrized model $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$. Each distribution P_θ is formally a probability measure on a common measurable space $(\Omega, \mathfrak{F})$, where Ω is the sample space and $\mathfrak{F}$ a σ-algebra of events on Ω. It is common to assume that the measure P_θ admits a density p_θ with respect to some common dominating measure, typically the Lebesgue measure μ, i.e., p_θ is the Radon-Nikodym derivative $dP_\theta/d\mu$. An observation is the outcome of a random variable X on a measurable space $(\mathcal{X}, \mathfrak{X})$, which maps Ω to $\mathcal{X}$ and which is $\mathfrak{F}$-measurable.

We mention first the framework of *independent sampling*. This is, by far, the most common sampling scheme

in Statistics, primarily due to the fact that the independence assumption considerably simplifies the theoretical treatment. In this setting, it is assumed that an observation X is distributed according to $P_0 = P_{\theta_0} \in \mathcal{P}$. An independent and identically distributed (i.i.d.) sample is a stochastic vector of observations $\mathbf{X}^{(n)} = \{X_1, X_2, \ldots, X_n\}$ which is distributed according to $P_0^{(n)}$, the n-fold product of the measure P_0 with itself, meaning the components X_i are mutually independent and distributed according to P_0. It is also possible that the sample is independent but not identically distributed (i.n.i.d.) in that each element X_i of the sample is distributed according to a (possibly different law) $P_{0,i} \in \mathcal{P}$. An example of this setup is that of regression with, say, fixed design, where we make observations of a function corrupted with noise at pre-specified points $\mathbf{t}^{(n)} = (t_1, \ldots, t_n)$ and then attempt to recuperate the underlying regression function based on these observations. We would then have that for a fixed (noiseless) regression function f_0, each observation X_i is distributed like $P_{\theta_{0,i}}$, with $\theta_{0,i} = \phi_i(f_0)$ where $\phi_i(f_0) = f_0(t_i)$, a functional of f_0.

There are also certain dependence assumptions one can make on the observations such that it is still fairly straightforward to make inference. For these, the data is typically assumed to be a time series, meaning that the observations $X_1, X_2, \ldots$ are seen as a sequence of random variables or vectors. The dependence structure is then expressed in terms of how the law of a certain observation depends on the observations that came before it. We speak then of the *memory* of the Stochastic process.

A Markov chain of order 1 (or with memory 1) is a discrete-time Stochastic process X_n such that $X_{n+1}|\mathbf{X}_n$ has the same (conditional) distribution as $X_{n+1}|X_n$, where $\mathbf{X}_n = (X_1, \ldots, X_n)$. Informally, given the *present*, the *future* of the chain and the *past* of the chain are independent. Further, if the distribution of $X_{n+1}|X_n$ does not depend on n, the chain is said to be time-homogeneous. If all the elements of the chain take values in the same space $\mathcal{X}$ (here called the state-space) then this means that our model is now parametrized by a parameter set $\Theta \times \mathcal{X}$ which corresponds to each observation $X_n|(X_{n-1} = x)$ being distributed according to a conditional probability measure $P_\theta(\cdot|x)$. The measure $P_\theta(\cdot|x)$ is also-called the *Markov kernel* of the chain and it is always required for the mapping $x \mapsto P_\theta(\cdot|x)$ to be $\mathfrak{F}$-measurable. These definitions can be generalized in the obvious fashion to Markov chains of order m.

An important quantity studied in the context of Markov chains is the *stationary distribution* of the chain. This is an invariant distribution for the chain in the sense that if an element of the chain is ever distributed according to this measure, then all subsequent states will have the same distribution. Given a Markov chain, one often wants to know if such a distribution exists and to determine it, or be interested in how long a chain which is started from an arbitrary state, takes to mix (read: to be distributed according to the stationary distribution). The stationary distribution of the chain is often studied using probabilistic methods but one can see it as the long-term behavior of a dynamical system where the initial measure for the chain evolves under the action of the transition kernel in the context of Ergodic Theory (cf. Section1.9).

A different type of memory structure which observations may have the Martingale property. A *filtration* of a measurable space $(\Omega, \mathfrak{F})$ is a growing sequence of σ-algebras $\mathfrak{F}_0 \subseteq \mathfrak{F}_1 \subseteq \mathfrak{F}_2 \subseteq \cdots \subseteq \mathfrak{F}$. A Stochastic process X_n is said to be adapted to a certain filtration $(\mathfrak{F}_k : k \in \mathbb{N}_0)$ if each X_n is $\mathfrak{F}_n$-measurable, $n \in \mathbb{N}$. One often considers the *natural filtration*, where for $n \in \mathbb{N}$, $\mathfrak{F}_n = \sigma(\mathbf{X}_n)$ is the σ-algebra generated by $\mathbf{X}_n$ (with $\mathfrak{F}_0$ the trivial σ-algebra). The discrete-time process X_n is called a Martingale with respect to the filtration $(\mathfrak{F}_k : k \in \mathbb{N}_0)$ if the process X_n is $(\mathfrak{F}_k : k \in \mathbb{N}_0)$-adapted, $\mathbb{E}\|X_n\|_1 < \infty$, and verifies the Martingale property: $\mathbb{E}[X_{n+1}|\mathfrak{F}_n] = X_n$. Informally, a filtration corresponds to the growing knowledge about a certain aspect of the process as time

progresses; the Martingale property states that knowledge of the past does not help to predict the future. In the same way, one speaks of a sub- or super-martingale if the martingale property is replaced with $\mathbb{E}[X_{n+1}|\mathfrak{F}_n] \geq X_n$ and $\mathbb{E}[X_{n+1}|\mathfrak{F}_n] \leq X_n$ respectively. A useful fact about Martingales which we will need in Chapters 5 and 6 is that Martingale increments (sometimes called Martingale difference sequence) $M_n = X_n - X_{n-1}$ (with $M_1 = X_1$) are uncorrelated.

We can also consider less structured, growing statistical models where each observation X_n, given $\mathbf{X}_{n-1}$, is distributed according to some conditional distribution depending on $\mathbf{X}_{n-1}$ and some vector of parameters $\boldsymbol{\theta}_n \in \Theta^n$. We will be referring to these as simply time series and both Markov chains and discrete-time Martingales will be particular types of time series.

In Chapters 5 and 6 we will be working directly with Markov chains and time series but we will not be concerned with stationary distributions, since they will in principle not exist in the setup we will be considering there. In Chapters 4 and 3 these processes will also make an appearance in the form of a Markov Chain Monte Carlo sampling algorithm (we will briefly summarise these sampling methods in Section 1.12). In these algorithms, the usual relation of the Markov chain with its stationary distribution is, in a sense, reversed. Given a measure from which we cannot sample directly, our aim is to construct a Markov chain which has this measure as stationary distribution. This will be used to indirectly obtain a sample distributed according to this measure: trajectories from such a chain can be used as approximate samples from its stationary measure, which is particularly useful in Bayesian non-parametric Statistics (cf. Section 1.11).

In the next section we provide some basic facts about Poisson processes which are a particular example of a continuous-time Markov processes and of a continuous-time sub-martingale. A description of continuous-time counterparts of Markov processes and Martingales can be found in [52], for example.

1.4 Poisson Point Processes

Poisson point processes are widely used statistical models in fields such as communication, meteorology, seismology, hydrology, astronomy, biology, medicine, actuary sciences, econometrics and queueing, to name but a few. They are point processes that can be used to model counts of random event. Their wide use can be attributed to several factors [89]. It seems to be a fairly accurate model in many applications which might be explained by the relatively mild qualitative conditions under which a point process is a Poisson process. These processes also have a simple structure and are commonly used both as a preliminary tool of study – eventually paving the way to the use of more sophisticated point processes – or as a basic component for constructing other stochastic processes whose sample paths are quite different from those of a Poisson process. Regardless, they are used extensively and have many interesting properties.

The literature on Poisson processes is vast so we collect here just a few fundamental results and definitions which are used in the remaining chapters. Poisson point processes are essentially random collections of points, with certain properties, usually on a d-dimensional Euclidian space, although they can be defined in more abstract spaces. The variable which indexes the process is usually referred to as *time* and/or *space*. The process N_t represents the number of times that a certain event has occurred by time t; say arrivals of calls at a call centre (see Chapters 2 and 3). One can also consider a time and a space variable and express the location of imperfections on a long strand of material under stress in time. With two space dimensions and one time dimension one can for example model rain droplets falling on a plot of land over a certain period of time (see the numerical example in Chapter 6).

A one-dimensional Poisson point process (cf. [54]) is usually simply referred to as a *Poisson process*. It is a continuous-time stochastic process N_t, indexed by time $t \geq 0$, non-decreasing, càdlàg and taking values in $\mathbb{N}_0$ such that $N_0 = 0$ (i.e. a continuous-time counting process). It has independent increments, and is parametrized by a non-negative function $\lambda(t)$, $t \geq 0$, called the *intensity function* of the process. When this function is constant, the process has stationary increments and is called *homogeneous*, otherwise it is called *inhomogeneous* (or non-homogeneous). Any increment of the process N_t is Poisson distributed, such that for $0 \leq a < b$, we have

$$\mathbb{P}\big(N_b - N_a = n\big) = \exp\Big(-\int_a^b \lambda(t)\,dt\Big)\frac{\Big(\int_a^b \lambda(t)\,dt\Big)^n}{n!}, \quad n \in \mathbb{N}_0.$$

Poisson point processes in $\mathbb{R}^d$ can be defined in a similar fashion and have analogous properties. We present here a simulation-based characterization that can be found in [91]. This characterization is adequate here since it defines Poisson processes in $\mathbb{R}^d$ in a way which is quite close to the procedure which is used to simulate them; we use simulated Poisson data for the numerical results in Chapters 2 and 6.

For $\mathcal{S} \subseteq \mathbb{R}^d$, a realization from a Poisson point process on $\mathcal{S}$, $\Xi = \{\Xi(S) : S \in \mathcal{S}\}$ is a random collection of points $\xi = (n, \{x_1, \ldots, x_n\})$; if $n = 0$ we have $(0, \varnothing)$. The event space of this process is then

$$\mathcal{E}(\mathcal{S}) = \big\{(0,\varnothing)\big\} \bigcup_{n=0}^{\infty} \big\{(n, \{x_1, \ldots, x_n\}) : x_i \in \mathcal{S}, i = 1, \ldots, n\big\}.$$

The process is characterized by a non-negative intensity function $\lambda : \mathcal{S} \mapsto \mathbb{R}^d$ for which it is assumed that

$$0 \leq \int_S \lambda(\mathbf{s})\,d\mathbf{s} < \infty,$$

for all bounded subsets $S \subseteq \mathcal{S}$. Note that we may have $\int_{\mathcal{S}} \lambda(\mathbf{s})\,d\mathbf{s} = \infty$.

Obtaining a realization ξ from $\Xi(S)$ on subsets $S \subseteq \mathcal{S}$, with intensity λ, a function on $S \subseteq \mathcal{S}$, can be represented as a two-step procedure. If $\int_S \lambda(\mathbf{s})\,d\mathbf{s} = 0$ then $\xi = (0, \varnothing)$. If $\int_S \lambda(\mathbf{s})\,d\mathbf{s} > 0$ then Ξ is obtained by first sampling a discrete random variable N with mass function

$$p_N(n) = \mathbb{P}(N = n) = \exp\Big(-\int_S \lambda(\mathbf{s})\,d\mathbf{s}\Big)\frac{\Big(\int_S \lambda(\mathbf{s})\,d\mathbf{s}\Big)^n}{n!}, \quad n \in \mathbb{N}_0.$$

If $n = 0$ then $\xi = (0, \varnothing)$ and the realization has been obtained. Otherwise, given $N = n > 0$, we sample n points according to i.i.d. continuous random variables X_i in S, with probability density function

$$p_X(x) = \frac{\lambda(x)}{\int_S \lambda(\mathbf{s})\,d\mathbf{s}}, \quad x \in S.$$

Thus, if the intensity λ is constant on S, then the density of the resulting homogeneous point process, given N is simply uniform on S. In the inhomogeneous case, given N, the density is proportional to the intensity λ.

This construction makes it clear how to draw a realization from $\Xi(S)$, call it $\xi = (n, \{x_1, \ldots, x_n\})$, $n \in \mathbb{N}_0$, $x_i \in S$ for $i = 1, \ldots, n$, via an acceptance-rejection algorithm [51]. We start by selecting an importance function[1]

[1] The function g must be positive and bounded on S and we must know a procedure to sample from g.

g and defining

$$M = \sup_{x \in S} \frac{p_X(x)}{g(x)}.$$

If S is a bounded set, then it is always possible to take g as the density of a uniform measure on S. To sample from $\Xi(S)$, we then perform the following procedure:

1. Take $\mathbf{x} = \varnothing$.
2. Sample n according to p_N.
3. If $n = 0$ then jump to step 7.
4. Generate a point x according to g and, independently, u according to a uniform measure on $[0, 1)$.
5. If $uMg(x) \leq p_X(x)$ then replace $\mathbf{x}$ with $\{\mathbf{x}, x\}$.
6. If $\mathbf{x}$ has less than n elements then return to step 4, otherwise continue to step 7.
7. Take $\xi = (n, \mathbf{x})$ as a realization of $\Xi(S)$.

Based on the representation given above, it is simple to see that the likelihood (with respect so an appropriate dominating product measure) at λ, given a realization ξ of Ξ, can be written as

$$\begin{aligned} p_\Xi(\xi) &= p_N(n) p_{\mathcal{X}|N}(x_1, \ldots, x_n \mid N = n) \\ &= \exp\Big(- \int_S \lambda(\mathbf{s})\, d\mathbf{s} \Big) \frac{\left(\int_S \lambda(\mathbf{s})\, d\mathbf{s} \right)^n}{n!} n! \prod_{i=1}^n \frac{\lambda(x_i)}{\int_S \lambda(\mathbf{s})\, d\mathbf{s}} \\ &= \exp\Big(- \int_S \lambda(\mathbf{s})\, d\mathbf{s} \Big) \prod_{i=1}^n \lambda(x_i). \end{aligned}$$

(The multiplicative factor $n!$ comes from the possible reorderings of the elements in $\{x_1, \ldots, x_n\}$.)

The particular case of the one-dimensional Poisson point process with intensity λ on $[0, t]$ occurs when we take $\mathcal{S} = \mathbb{R}^+$, $S = S_t = [0, t]$. If we denote $N = N_t = \Xi\big([0, t]\big)$, then the likelihood at λ becomes

$$p_\Xi(N) = \exp\Big(- \int_0^t \lambda(s)\, ds + \int_0^t \log\big(\lambda(s)\big)\, dN_s \Big).$$

If we refer to a Poisson process with constant intensity $\lambda(s) \equiv 1$, $s \in [0, t]$, as a *standard* Poisson process, then as the expression above suggests, we can write, for each $t \geq 0$, the density of N_t with respect to a standard Poisson process as

$$p_\lambda(N) = \exp\Big(- \int_0^t \big(\lambda(s) - 1\big)\, ds + \int_0^t \log\big(\lambda(s)\big)\, dN_s \Big).$$

(see for example [50].)

Another important feature of Poisson processes which can also be seen from an algorithmic perspective is *thinning* [91]. This procedure is used to reduce the number of events in a Poisson point process. Consider a Poisson point process Ξ on $\mathcal{S}$ with intensity λ and pick a function α such that for all $x \in \mathcal{S}$, $0 \leq \alpha(x) \leq 1$. The function $1 - \alpha(x)$, $x \in S$ represents the probability of removing from Ξ a point located at x. Given a realization

$\xi = (n, \{x_1, \dots, x_n\})$ of $\Xi(S)$, $S \subseteq \mathcal{S}$, the process of independent Bernoulli thinning corresponds to removing from ξ each x_i with probability $1 - \alpha(x_i)$, $i = 1, \dots, n$, independently of one another; the realization from the thinned point process is denoted $\xi_\alpha = (m, \{\varkappa_1, \dots, \varkappa_m\})$ and contains the points that were not removed. The index m corresponds to the number of points $\{\varkappa_1, \dots, \varkappa_m\} \subseteq \{x_1, \dots, x_n\}$. The realization $\xi_{1-\alpha}$ contains the removed points. Making use of the representation of Poisson point processes based on the acceptance-rejection method, one can see that ξ_α is actually a Poisson point process with intensity $\lambda_\alpha(x) = \alpha(x)\lambda(x)$. This also means that an inhomogeneous Poisson point process on S with intensity λ can simply be seen as a homogeneous Poisson point process on S thinned according to $\alpha(x) = \lambda(x)/\int_S \lambda(\mathbf{s})\, d\mathbf{s}$. All of this can be extended straightforwardly to allow thinning into more than two processes; in this case the procedure is usually called *coloring* [54].

Poisson point processes have another important property called *independent scattering*. The term independent scattering has been introduced into the literature quite recently in [69] but it conveys very well how the points in a realization of a Poisson point process arrange themselves independently of one another. The property can be described as follows. Consider a Poisson point process Ξ on $\mathcal{S} \subseteq \mathbb{R}^d$ with intensity λ and two disjoint subsets $A, B \subseteq \mathcal{S}$. Let Ξ_A and Ξ_B be obtained by restricting Ξ to A and B, respectively, in the sense that for any $S \subseteq \mathcal{S}$, given a realization ξ from $\Xi(S)$, ξ_A and ξ_B contain the points from ξ which are in $A \cap S$ and $B \cap S$, respectively. As the notation suggests, $\Xi_A = \{\Xi(S) : S \subseteq A\}$ and $\Xi_B = \{\Xi(S) : S \subseteq B\}$ are Poisson point processes on A and B since they are simply obtained from Ξ via thinning using $\alpha(x) = \mathbb{1}_A(x)$ and $\beta(x) = \mathbb{1}_B(x)$, respectively. With this in mind, it is straightforward to check that

$$p_{\Xi_{A\cup B}}(\xi) = p_{\Xi_A}(\xi_A)\, p_{\Xi_B}(\xi_B),$$

so that the two Poisson point processes are even independent. In passing we also see that by taking $B = A^c$, a thinned Poisson point process and the Poisson point process containing the remaining points are independent.

1.5 Consistency, Asymptotics and Non-Asymptotics

Let us return now to general, parametrized models. Once a model has been fixed and data collected, the goal of the inference procedure is to devise an estimate, i.e., a measurable function of the data, $\hat{\theta}_n = \vartheta(\mathbf{X}^{(n)})$; this estimate is meant to be an approximation for the unknown parameter θ_0 and therefore for the unknown distribution of the data, P_0. Assume that the parameter space Θ is endowed with a metric d. Being a function of the data, estimates are random and so any criterion to assess the *proximity* of an estimator and an estimate must be a probabilistic one. A basic requirement is that of *asymptotic consistency*. An estimator is *consistent* for θ_0 if,

$$\mathbb{P}_{\theta_0}\big(d(\hat{\theta}_n, \theta_0) > \varepsilon\big) \to 0, \text{ for every } \varepsilon > 0, \text{ or,} \tag{1.1}$$

$$\mathbb{P}_{\theta_0}\big(\lim_{n\to\infty} d(\hat{\theta}_n, \theta_0) = 0\big) = 1 \tag{1.2}$$

as n converges to infinity. We will sometimes write this as $\hat{\theta}_n \xrightarrow{P_0} \theta_0$ and $\hat{\theta}_n \xrightarrow{a.s.} \theta_0$ and call it *weak consistency* and *strong consistency*, respectively. Alternatively, one may also have consistency *in expectation* when it holds that

$$\mathbb{E}_{\theta_0} d(\hat{\theta}_n, \theta_0) \to 0, \tag{1.3}$$

as n converges to infinity. The expectation in the previous display is also known as a *risk function* of the estimator $\hat{\theta}_n$. (We will return to the risk of an estimator in the next section.)

Consistency as presented in (1.1) through (1.3) only provides us with a qualitative measure of the preciseness of an estimator; it states that the estimator is, with high probability, close to the unknown parameter as n grows. We are mostly interested, however, in *how close* the estimator is to the unknown parameter, typically as a function of the sample size n; we would like to quantify the estimation error. We speak then of an (asymptotic) rate of convergence r_n of an estimator $\hat{\theta}_n$ with respect to the metric d, if the rescaled sequence

$$r_n^{-1} d(\hat{\theta}_n, \theta_0) \tag{1.4}$$

is bounded, as n goes to infinity, either in probability, almost surely or in expectation; we will also say that $\hat{\theta}_n$ is r_n-consistent for θ_0.

The *asymptotics* of an estimator corresponds to its *large sample behavior*. This type of analysis is made mostly out of convenience since one can make use of several limiting results such as strong laws and central limit theorems in their analysis, as well as disregard certain approximation errors which vanish in the limit. It is then often possible to characterize quite precisely the asymptotic (limiting) probabilistic behavior of the estimator under study. Of particular interest is determining the optimal (read: smallest) sequence r_n such that (1.4) has a nontrivial (tight) distributional limit. The downside of asymptotics is that it is usually not known how large a sample should be, such that this limiting behavior comes into effect and such that it is in fact a good representative for the outcomes of the estimator. Nonetheless, determining the asymptotic behavior of an estimator constitutes a standard theoretical tool in Mathematical Statistics.

The alternative – or perhaps complementary – approach, is to establish *non-asymptotic* (meaning *finite sample*) results by either deriving the exact distribution of the estimator or attempting to compute exactly, or bound, quantities such as risk in (1.3), for each value of n. This is often a difficult task. Deriving the exact distribution of an estimator, in particular, can often only be done under very strict assumptions and computing risks such as the one in (1.3) is sometimes not informative since the quantity often depends on the unknown parameter θ_0 or may not vary monotonously. Nonetheless, making use of the structure of specific estimators (such as the recursive estimators we will present in Chapters 5 and 6), it is sometimes possible to derive non-asymptotic results. It will also be clear that in certain situation it will be possible to derive asymptotical results from these (stronger) finite sample results.

It is also of interest to know, for a given model, how properties such as consistency and rates of convergence, depend on the value of the parameter being estimated. Specifically, we would like to know how well a parameter can be estimated, uniformly over the model. This leads to the notion of *optimality* with respect to a certain criterion. We describe this in a bit more detail in the next section.

1.6 Minimax Risk and Minimax Rates

Risk functions play a very important role in Mathematical Statistics. They are used to define minimax risks, which are the most commonly used criteria for assessing the optimality of an estimator over a certain model. The inference procedure then boils down to the following: we pick a risk function and in doing so we implicitly define a minimax risk – the smallest risk that an estimator can achieve, uniformly over the model; we then design an

estimator which attains this minimax risk which we call an *optimal* estimator (with respect to that risk).

Risk function can be defined in a general context (cf. [49]) but we mention here just L_p-risk functions. These are perhaps the most commonly used risk functions when the parameter set Θ is a normed space, and are given by

$$R_n^p(\vartheta, \theta) = \mathbb{E}_\theta \|\vartheta(\mathbf{X}^{(n)}) - \theta\|_p^p, \tag{1.5}$$

where $1 \leq p \leq \infty$. The L_2-risk, in particular, plays a classical role in estimation theory. The so-called *bias-variance* decomposition, where the L_2-risk is written as the sum of a (squared) bias term and a variance term, respectively:

$$R_n^2(\vartheta, \theta) = \|\mathbb{E}_\theta \vartheta(\mathbf{X}^{(n)}) - \theta\|_2^2 + \mathbb{E}_\theta \|\vartheta(\mathbf{X}^{(n)}) - \mathbb{E}_\theta \vartheta(\mathbf{X}^{(n)})\|_2^2. \tag{1.6}$$

If the bias term is zero, or equivalently, if $\mathbb{E}_\theta \vartheta(\mathbf{X}^{(n)}) = \theta$, then the estimator is said to be *unbiased*; if this holds only in the limit, then the estimator is said to be *asymptotically unbiased*. Based on (1.6) one then sees that an estimator with small L_2-risk is concentrated around the "true" underlying parameter (small bias) and has low spread about its mean value (small variance) both of which are desirable properties. We would like to design estimators that have both low bias and low variance and consequently low risk. These are, however, to some extent, opposing tasks; reduction of the bias is usually achieved at the expense of increasing the complexity or variability of the estimator so that it may best fits the data, but this will often lead to an increase in the variance of said estimator in what is known as the *bias-variance trade-off*. For the rest of this section we will stick to the L_2-risk.

Returning to the general framework of estimation risk, it is instructive to look at the following quantity

$$r_n^p = r_n^p(\Theta) = \inf_{\vartheta} \sup_{\theta \in \Theta} R_n^p(\vartheta, \theta), \tag{1.7}$$

which is called the *minimax risk over* Θ [49]; the infimum is taken over all measurable functions of the data $\mathbf{X}^{(n)}$. An estimator ϑ_0 is then called a *minimax optimal estimator* over Θ, if asymptotically this function attains the minimax risk

$$\sup_{\theta \in \Theta} R_n^p(\vartheta_0, \theta) = r_n^p(\Theta)\big(1 + o(1)\big), \quad \text{as } n \to \infty, \tag{1.8}$$

although we will typically also be satisfied with near-optimal estimators ϑ_0' which satisfy

$$\sup_{\theta \in \Theta} R_n^p(\vartheta_0', \theta) \leq C_n\, r_n^p(\Theta)\,(1 + o(1)), \quad \text{as } n \to \infty, \tag{1.9}$$

for some bounded sequence C_n. We will also typically be satisfied if $C_n = C \log^c(n)$ for some constants $C \geq 1$ and $c \geq 0$. A minimax estimator matches then the risk of the "best" estimator at estimating the "worst" value for the unknown parameter.

For parametric models the minimax risk is typically, under some regularity conditions, of order $n^{-1/2}$; it is commonplace to refer to this rate as the *parametric rate* since it is the standard rate obtained for parametric models. Minimax risks over non-parametric models have been studied for many models since the first results in this context were derived in [90] for non-parametric regression with random design and density estimation. If

$\mathcal{F}_\alpha = C^\alpha(\mathbb{R}^d)$ denotes the space of all α-times continuously differentiable functions on $\mathbb{R}^d$, then the minimax risk of estimation of the m-th derivative ($m \in \mathbb{N}_0$) of a regression function or density, based on independent a sample of size n, is of the order $n^{-(\alpha-m)/(2\alpha+d)}$. This particular rate is often referred to as the *non-parametric rate* since it is of a larger order when compared to the parametric rate. In a sense this is the "price" to pay for considering larger models. Another family of models over which there is a typical value for the risk are models indexed by spaces of monotonous functions, where the risk is typically of order $n^{-1/3}$ (cf. [19, 102]).

The relation between the size of a model and the ease with which a parameter can be estimated should now be clearer, at least for the particular case where the parameter set $\mathcal{F}_\alpha$ is $C^\alpha(\mathbb{R}^d)$: since the spaces $\mathcal{F}_\alpha$ are nested, meaning that $\mathcal{F}_\alpha \subseteq \mathcal{F}_\beta$ whenever $\alpha \geq \beta$, the non-parametric rate (read: estimation error) drops, as the *smoothness* or *regularity* α of the model increases because the parameter space becomes smaller.

The choice to nest the parameter sets according to differentiability is a particular choice. There are other ways of "peeling" subsets of $C^0(I)$, the space of continuous functions on $I \subseteq \mathbb{R}$, which may be more convenient when working with specific types of non-parametric estimators. We can also simply consider the *space of Lipschitz functions* $\mathcal{L}_\alpha = \mathcal{L}_\alpha(L, I)$, $L > 0$ and $\alpha \in (0, 1]$, the space of all continuous functions f such that

$$|f(x) - f(y)| \leq L|x - y|^\alpha, \tag{1.10}$$

for all $x, y \in I$. One might also consider the *Hölder class* of functions $\mathcal{H}_\alpha = \mathcal{H}_\alpha(L, I)$, $L > 0$, which is the space of all $\alpha_0 = \lfloor \alpha \rfloor$ times differentiable function such that their α_0-th derivative satisfies the *Hölder condition*

$$|f^{(\alpha_0)}(x) - f^{(\alpha_0)}(y)| \leq L|x - y|^{\alpha - \alpha_0}, \tag{1.11}$$

for all $x, y \in I$. The *Sobolev class* of functions $\mathcal{W}_\alpha(L, I)$, $\alpha \in \mathbb{N}$, $L > 0$, is the space of all functions $f : I \mapsto \mathbb{R}$ such that $f^{(\alpha-1)}$ is absolutely continuous and

$$\int_I \left(f^{(\alpha)}(t)\right)^2 dt \leq L^2. \tag{1.12}$$

1.7 Adaptation

The typical situation for non-parametric models is that the minimax rate of estimation over a certain class Θ will depend, in terms of order, on some characteristic of the elements of Θ. To make this precise, we say that there exists some, *smoothness index* $\alpha \in \mathcal{A}$ and an order relation $\preceq$, such that for $\alpha, \beta \in \mathcal{A}$, $\alpha \preceq \beta$ implies $\Theta_\beta \subseteq \Theta_\alpha$. We then consider the problem of estimation over a family of models $\Theta = \Theta_\mathcal{A} = \bigcup_{\alpha \in \mathcal{A}} \Theta_\alpha$. Clearly, from the definition of minimax risk, if $\mathcal{A} \subseteq \mathcal{B}$ then $r_n^p(\Theta_\mathcal{A}) \leq r_n^p(\Theta_\mathcal{B})$, i.e., smoother models can, in a minimax sense, be estimated more accurately. In principle, we would then like to choose the smallest parameter set Θ which contains the true parameter of the data, and then we would like to consider an estimation problem over Θ only. This means that we would like to pick $\Theta = \Theta_\alpha$ such that for all β such that $\beta \preceq \alpha$, $P_0 \in \Theta_\beta$ and for all β such that $\beta \succ \alpha$, $P_0 \notin \Theta_\beta$.

This choice, however, is not feasible, in the sense that it relies on a-priori knowledge of a characteristic of the unknown θ_0. This is, in a sense, the cost for considering estimation problems over non-parametric families of models: the parameter set indexing the model becomes so large that the behavior of the estimation procedure quantitatively depends on the unknown parameter – a fixed estimator, tailored specifically to perform well for a *fixed* model, will no longer be able to effectively estimate an *arbitrary* model of the family.

The problem of *adaptation* [32, 75] then corresponds to designing estimators which do not rely on knowledge of the smoothness of the model but which nonetheless attain the rate $r_n^p(\Theta_\alpha)$, eventually up to a logarithmic factor. Such estimators are said to be *rate adaptive*. We will explain how adaptation can be performed in a Bayesian context in Section 1.11. In the next sections we will start outlining some techniques for constructing estimators.

1.8 M-estimation & Semi-Parametric Estimation

We turn our attention now to the construction of estimators, namely of M-estimators for parametric models and semi-parametric models (where the non-parametric part is treated as a nuisance parameter). In Chapter 2 we will use these results to produce an M-estimator for the period of an inhomogeneous Poisson process with a cyclical intensity function.

M-estimators represent a general class of statistics that are obtained as maximizers of a certain function of the data, called a *criterion function*. These estimators where first proposed in this form by Peter Huber in 1964 in the context of robust estimation and actually generalize many classical estimators. Robust estimation refers to the design of estimators which show "*insensitivity to small deviations from the assumptions*" [48].

In a classical M-estimation setup we are given a sample $\mathbf{X}^{(n)} = \{X_1, X_2, \dots, X_n\}$ from the distribution $P_{\theta_0} \in \mathcal{P} = \{P_\theta : \boldsymbol{\theta} \in \Theta \subset \mathbb{R}^d\}$, a parametric model taking values in $\mathcal{X} \subset \mathbb{R}^m$. We select a parametrized function of the sample $\rho : \mathcal{X} \times \Theta \mapsto \mathbb{R}$, giving rise to a criterion function which is defined as

$$M_n(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \rho(X_i, \boldsymbol{\theta}). \tag{1.13}$$

The M-estimator associated with this criterion function is the maximizer

$$\hat{\boldsymbol{\theta}}_n = \arg\max_{\boldsymbol{\theta} \in \Theta} M_n(\boldsymbol{\theta}). \tag{1.14}$$

Under some mild conditions, by the strong law of large numbers, for each $\boldsymbol{\theta} \in \Theta$,

$$M_n(\boldsymbol{\theta}) \xrightarrow{a.s.} M(\boldsymbol{\theta}) = \mathbb{E}\rho(X, \boldsymbol{\theta}),$$

as n tends to infinity. This particular form of the limiting criterion function motivates choosing the function ρ such that $M(\boldsymbol{\theta})$ has a maximum at the point $\boldsymbol{\theta} = \boldsymbol{\theta}_0$; if the criterion function is close to $M(\boldsymbol{\theta})$ and is maximal at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_n$, then $\hat{\boldsymbol{\theta}}_n$ should be close to $\boldsymbol{\theta}_0$.

Alternatively, for a vector valued function $\psi(x, \boldsymbol{\theta}) : \mathcal{X} \times \Theta \mapsto \mathbb{R}^d$, one might also consider the criterion

$$\Psi_n(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \psi(X_i, \boldsymbol{\theta}), \tag{1.15}$$

and define an M-estimator (in this case sometimes called a Z-estimator) as the solution $\hat{\boldsymbol{\theta}}_n \in \Theta$ to the system of equations

$$\Psi_n(\boldsymbol{\theta}) = \mathbf{0}. \tag{1.16}$$

The vector valued function $\psi(x, \boldsymbol{\theta})$ can be taken, for example, as the gradient, the vector of partial derivatives of

a function $\rho(x, \boldsymbol{\theta})$ with respect to each component of $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)$, if these exist.

The generality of this method can be seen by noting that several classical types of estimators can be obtained for specific choices of $\rho(x, \boldsymbol{\theta})$: if the data admits a density $f(x, \boldsymbol{\theta})$ and $\rho(x, \boldsymbol{\theta}) = \log f(x, \boldsymbol{\theta})$, the resulting M-estimator will be the maximum likelihood estimator; the M-estimator associated with the function $\rho(x, \boldsymbol{\theta}) = -\|x - g(\boldsymbol{\theta})\|_2^2$ where $g : \Theta \mapsto \mathbb{R}^m$ is a least squares estimator; if $m = 1$ and the first d moments of the data exist and can be expressed as $g_1(\boldsymbol{\theta}), \ldots, g_d(\boldsymbol{\theta})$ then the M-estimator corresponding to taking each component of ψ as $\psi_i(x, \boldsymbol{\theta}) = x^i - g_i(\boldsymbol{\theta})$, $i = 1, \ldots, d$, is a moment estimator; if for $\alpha \in (0, 1)$ we pick $\rho_\alpha(x, \theta) = -(\alpha - \mathbb{1}\{x \leq \theta\})(x - \theta)$ then the resulting M-estimator will be the α-th quantile of the data. (We will return to the problem of estimating quantiles in Chapter 5.) Besides generalizing these different estimators, we have in general some added flexibility in that we are allowed to choose the function ρ. A criterion function can then be picked specifically for each model or parameter of interest, but also, working with functions ρ which are truncated in some way may lead to the resulting M-estimators that have good robustness properties [48].

The following theorems make a statement about the asymptotical consistency of M-estimators and Z-estimators.

Theorem 1.1 (M-estimator consistency [96])
Let M_n be a sequence of random functions and let M be a fixed function of θ such that for every $\varepsilon > 0$,

$$\sup_{\boldsymbol{\theta} \in \Theta} |M_n(\boldsymbol{\theta}) - M(\boldsymbol{\theta})| \xrightarrow{P} 0$$
$$\sup_{\boldsymbol{\theta} : |\boldsymbol{\theta} - \boldsymbol{\theta}_0| \geq \varepsilon} M(\boldsymbol{\theta}) < M(\boldsymbol{\theta}_0).$$

Then any sequence of estimators $\hat{\boldsymbol{\theta}}_n$ such that $M_n(\hat{\boldsymbol{\theta}}_n) \geq M_n(\boldsymbol{\theta}_0) - o_p(1)$ converges in probability to $\boldsymbol{\theta}_0$.

Theorem 1.2 (Z-estimator consistency [96])
Let Ψ_n be a sequence of random vector-valued functions and let Ψ be a fixed vector-valued function of θ such that for every $\varepsilon > 0$,

$$\sup_{\boldsymbol{\theta} \in \Theta} \|\Psi_n(\boldsymbol{\theta}) - \Psi(\boldsymbol{\theta})\| \xrightarrow{P} 0$$
$$\inf_{\boldsymbol{\theta} : |\boldsymbol{\theta} - \boldsymbol{\theta}_0| \geq \varepsilon} \|\Psi(\boldsymbol{\theta})\| > 0 = \|\Psi(\boldsymbol{\theta}_0)\|.$$

Then any sequence of estimators $\hat{\boldsymbol{\theta}}_n$ such that $\Psi_n(\hat{\boldsymbol{\theta}}_n) = o_p(1)$ converges in probability to $\boldsymbol{\theta}_0$.

The requirement for the convergence of the criterion function to take place uniformly over the parameter set Θ is typically a stronger requirement than what is needed for the resulting M- or Z-estimator to be consistent but it leads to a straightforward proof of the results.

A couple of notes about the generality of Theorems 1.1 and 1.2. First note that these theorem do not necessarily require criterion functions to be of the form (1.13) or (1.15), respectively, and can be applied to any (measurable) sequence of functions. Note also that the conclusions of the theorems hold under the assumption that the M-

estimator (resp. Z-estimator) approximately maximises (resp. is a zero of) the criterion function; this typically gives some extra flexibility when designing computational implementations of these procedures.

In Chapter 2 we will use M-estimation theory to produce an estimator for the parametric part of a semi-parametric model. The criterion function we use cannot be written point-wisely as a mean of an i.i.d. sequence of random variables. The functions of θ that we are averaging to produce our criterion function will then not have the same mean. Instead, the mean of these functions will be a certain functional of the non-parametric part of the model. This adds some extra difficulty to checking the first condition of Theorem 1.1 so that we will need some results from ergodic theory in our proofs. For completeness, we present here a brief summary of some basic definitions and results from ergodic theory which we use in that chapter.

1.9 Brief Primer on Ergodic Theory

The adjective *ergodic* stems from the Greek words *ergon*, meaning work and *odos*, meaning path. This term was coined by Ludwig Boltzmann in the late 19th century while working on the kinetic theory of gases. Ergodic Theory has a transversal presence in many fields of Mathematics and in Statistical Mechanics and deals mainly with the long term evolution of the (average) behavior of dynamical systems. Ergodic Theorems give conditions under which statements about this behavior can be made and began appearing in the literature in the 1930's by the hand of John von Neumann and George David Birkhoff.

In this section we consider a probability space $(M, \mathcal{M}, \mu)$, where M is a compact metric space, μ a probability measure on the Borel σ-algebra on M and $\mathcal{M}$ is the completion of the Borel σ-algebra on M with respect to μ. Let R be a mapping from M onto itself. Given $x \in M$ we call the sequence $R^0x, Rx, R^2x, \ldots$, the trajectory of x (under the action R); R^k representing a k-fold composition of R with itself, R^0 the identity. This trajectory will typically correspond to states of a (discrete) dynamical system with initial state x evolving according to R. The *time average* of a μ-integrable function f over a trajectory is the mean

$$\frac{1}{n}\sum_{i=1}^{n} f(R^k x) \tag{1.17}$$

while its *space average* is the integral

$$\int_M f(x)\, d\mu. \tag{1.18}$$

One of the main questions of Ergodic Theory is to derive conditions under which the limit of the time average (1.17) exists and when it is equal to the space average (1.18). This equality usually takes place under mild conditions for μ-a.e. initial state x and is the content of the classical Birkhoff-Khinchin Theorem (Theorem 1.3).

Write now $RA = \{Rx : x \in A\}$ and $R^{-1}A = \{x \in M : Rx \in A\}$. We say that μ is invariant with respect to R if for any $A \in \mathcal{M}$ we have $\mu(R^{-1}A) = \mu(A)$. We can now state the Birkhoff-Khinchin Theorem.

Theorem 1.3 (Birkhoff Ergodic Theorem [100])
Consider a measure space $(M, \mathcal{M}, \mu)$ where $\mathcal{M}$ is a σ-algebra on M and μ is a probability measure on $(M, \mathcal{M})$,

invariant with respect to R. Then, if f is a μ-integrable function supported on M,

$$\frac{1}{n}\sum_{i=1}^{n} f(R^k x) \longrightarrow f^*(x),$$

as n goes to infinity, for some μ-integrable function f^* and μ-a.e. $x \in M$. Also $f^*(Rx) = f^*(x)$ for μ-a.e. x and $\int_M f^*(x)\,d\mu = \int_M f(x)\,d\mu$.

If in addition the mapping R is *ergodic* (see the definition bellow), then the function f^* is μ-a.e. constant and so $f^* \equiv \int_M f(x)\,d\mu$, μ-a.e.. With the extra assumption that R is ergodic, the Strong Law of Large Numbers can be shown to follow from this theorem (cf. [57]). In Chapter 2 we need, however, a uniform version of such a result and therefore we present here the (stronger) conditions under which this can be assured.

If R is invertible, continuous, with a continuous inverse and both RA and $R^{-1}A$ are measurable then R is called a *homeomorphism*. A measure μ, invariant with respect to R now verifies, for any $A \in \mathcal{M}$, $\mu(R^{-1}A) = \mu(A) = \mu(RA)$. A set $A \in \mathcal{M}$ is an *invariant set* if it satisfies $R^{-1}A = A$.

An invariant mapping R is said to be *ergodic* if each invariant set $A \in \mathcal{M}$ verifies either $\mu(A) = 0$ or $\mu(A) = 1$. It is known (cf. [22]) that for any continuous map R of a compact metric space onto itself, there exists a Borel probability measure μ which is invariant with respect to R; if that measure is unique then R is said to be *uniquely ergodic*.

We are now ready to state a uniform version of the Ergodic Theorem.

Theorem 1.4 (Uniform Ergodic Theorem [22])
Suppose R is a homeomorphism on a compact metric space M and μ a Borel probability measure, invariant with respect to R. If R is uniquely ergodic then for any continuous function f supported on M,

$$\frac{1}{n}\sum_{i=1}^{n} f(R^k x) \longrightarrow \int_M f(x)\,d\mu, \quad \text{as } n \to \infty,$$

uniformly over all $x \in M$.

An example of a setup where this theorem can be applied is the following. For $\tau > 0$ take $M = [0, \tau)$, let $\mathcal{M}$ be the Borel σ-algebra on M and let μ be the uniform measure on $[0, \tau)$ such that its density w.r.t. the Lebesgue measure is identically τ^{-1}. Consider now, for any $T > 0$, the mapping $R : [0, \tau) \mapsto [0, \tau)$ defined as $Rx = x + T \pmod{\tau}$. This mapping is called a *rotation of the circle* and is clearly a homeomorphism. Further, if T/τ is an irrational number, then μ is the unique invariant probability measure with respect to R which implies that R is uniquely ergodic (cf. [22], Chapter 3). Theorem 1.4 can then be applied to conclude that for any continuous, Lebesgue-integrable function f,

$$\frac{1}{n}\sum_{i=1}^{n} f(R^k x) \longrightarrow \frac{1}{\tau}\int_0^{\tau} f(x)\,dx,$$

as n goes to infinity, uniformly over all $x \in [0, \tau)$. We use this result in Chapter 2.

These rotations also play an important role in the generation of *pseudo-random numbers*. The value x is then

called a *seed* and for an appropriate choice of a value for T and τ, the resulting trajectory $(\tau^{-1}R^k x : k \in \mathbb{N}_0)$ can be used as an "approximation" for a sequence of uniformly distributed numbers on $[0,1)$. It is also possible to use mappings of the type $Rx = Tx \pmod{\tau}$. The two resulting procedures to generate pseudo-random number are usually called congruential generators [80]. In practice, selecting a good combination of seed, T and τ is quite important for the quality of the resulting sequence as random numbers and consequently for procedures in which they are used such as Monte Carlo simulation methods.

1.10 Stochastic Approximation & Tracking

Up to now our focus has been on estimation problems and the asymptotical properties of estimators. Typically, as the sample size grows, we are capable or producing more accurate estimates for the parameter indexing the distribution of the data. This is possible since as more data is collected it becomes evident that some values of the parameter are unlikely to have produced the data we have observed. This narrows down the likely parameters in the model to be producing the data, resulting in better estimates.

We can also consider a more general, non-static, *tracking* setup where it is natural to see the observations as a time series. We still collect data from a fixed model, but the parameter indexing the distribution of the data changes with time. Now there is no longer (necessarily) accumulation of information about some fixed, "true" parameter and the focus shifts from producing estimators to producing sequences that *track* the time-changing parameter.

The natural type of tracking algorithms are recursive ones: at each time point, we update our current approximation based on newly observed data and the structure of the model. The information which is implicitly being accumulated and which allows us to assure some level of precision for the algorithm will come from a) the fact that the model is assumed to have finite memory and b) the parameter indexing the distribution of the data is changing "slowly" such that, in some sense, past observations still contain relevant information about the current value of the time-changing parameter. The algorithms we propose in Chapter 5 and Chapter 6 can be seen as Stochastic Approximation algorithms and we develop techniques to study several different tracking problems.

Stochastic approximation algorithms form a generic class of stochastic procedures for optimising functions in a stochastic setting. They were introduced into the literature in the 1951 by Herbert Robbins and Sutton Monro [76] who were interested in the problem of finding the root of a function which is observed with noise. Below we recall the classical setting of Robbins and Monro.

Suppose that for some known value α in the range of a non-decreasing function M we want to solve the equation $M(x) = \alpha$, which is assumed to have a unique solution at $x = \theta$. The function M cannot be observed directly but instead we observe $Y(x)$ with distribution function $H(y|x)$ such that $\mathbb{E}Y(x) = M(x)$. The simplest example is

$$Y(x) = M(x) - \alpha + \xi,$$

where ξ is mean zero random noise. The naïve approach to this problem would be to attempt to use Newton's method and make successive observations according to

$$x_{n+1} = x_n - \frac{Y(x_n) - \alpha}{M'(x_n)}, \quad n \in \mathbb{N}_0,$$

for some initial x_0, assuming the derivative of M exists, is known and $M'(\theta) \neq 0$. As pointed out in [64], even under these generous assumptions, this would would result in

$$x_{n+1} = x_n - \frac{M(x_n) - \alpha}{M'(x_n)} - \frac{\xi_n}{M'(x_n)}.$$

This means that if x_n were to converge to θ such that $M(x_n) \to \alpha$ and $M'(x_n) \to M'(\theta)$ then this would imply that $\xi_n \to 0$ which is not a realistic assumption in many contexts (e.g., for i.i.d. errors ξ_n with positive variance).

Consider instead, for a positive sequence a_n, the following recursive algorithm as proposed by [76],

$$X_{n+1} = X_n + a_n\big(\alpha - Y(X_n)\big), \quad n \in \mathbb{N}_0, \tag{1.19}$$

for $n \in \mathbb{N}_0$ with X_0 arbitrary. The Markov Chain (1.19) constitutes the Robbins-Monro algorithm and the following result can be found in [76].

Theorem 1.5 (Robbins-Monro [76])
Let $M(x)$ be a non-decreasing function such that $M(\theta) = \alpha$ and such $M'(\theta)$ exists and is positive. Let $Y(x) \sim H(y|x)$ be a.s. uniformly bounded and such that $\mathbb{E}Y(x) = M(x)$. If a_n is a positive (non-increasing) sequence such that

$$\sum_{n=0}^{\infty} a_n = \infty \quad \text{and} \quad \sum_{n=0}^{\infty} a_n^2 < \infty,$$

then for the sequence (1.19) it holds that $\mathbb{E}(X_n - \theta)^2$ converges to zero as n goes to infinity.

For $a > 0$ the sequence $a_n = a\, n^{-1}$ verifies the conditions of the theorem. The function $G(\alpha, Y(X_n), X_n) = \alpha - Y(X_n)$ is called a gain function and has a simple interpretation in (1.19): if $Y(X_n)$ overshoots α then the gain will be negative causing X_n to decrease; if $Y_n(X_n)$ undershoots α then the gain will be positive and X_n gets increased. The decreasing sequence a_n dampens the changes of the tracking sequence enough to ensure convergence of the algorithm.

Different gain functions lead to different algorithms and a large number of problems can be tackled by selecting appropriate gain functions. Consider the problem of recursively estimating the maximum $M(\theta)$ of a function $M(x)$ with, as before, noisy observations $Y(X_n)$ such that $\mathbb{E}Y(x) = M(x)$. The following algorithm is known as the Keifer-Wolfowitz algorithm[53] and can be used for this purpose

$$X_{n+1} = X_n + a_n \frac{Y(X_n + b_n) - Y(X_n - b_n)}{b_n}, \quad n \in \mathbb{N}_0, \tag{1.20}$$

with X_0 arbitrary and where a_n and b_n are decreasing sequences.

Theorem 1.6 (Kiefer-Wolfowitz [53])
Let $M(x)$ be a convex function with a unique maximiser at $x = \theta$. Let $Y(x) \sim H(y|x)$ such that $\mathbb{E}Y(x) = M(x)$ and such that $\mathbb{E}\big(Y(x) - M(x)\big)^2$ is uniformly bounded. If a_n and b_n are positive sequences which converge to zero

as n goes to infinity such that

$$\sum_{n=0}^{\infty} a_n = \infty, \quad \sum_{n=0}^{\infty} a_n\, b_n < \infty \quad \text{and} \quad \sum_{n=0}^{\infty} \left(\frac{a_n}{b_n}\right)^2 < \infty,$$

then for the sequence (1.20) it holds that $\mathbb{E}(X_n - \theta)^2$ *converges to zero as n goes to infinity.*

The sequences $a_n = n^{-1}$ and $b_n = n^{-1/3}$, for example, verify the conditions of the theorem.

These algorithms have been extensively studied in the literature and the field of stochastic approximation has grown quite a lot since its inception (cf. [58] for a review). In Chapters 5 and 6 we propose and study a rather general stochastic approximation algorithm for the case where it is only assumed that the observations come from a time series.

1.11 Bayesian Statistics & Non-Parametric, Adaptive Estimation

Bayesian Statistics stems, in essence, from a different view on how data are generated from models. In the classical, *frequentist*, paradigm, data are assumed to be generated according to a fixed distribution in the model. This is intimately connected with the view of probabilities as limits of the frequency with which a certain event takes place. In the Bayesian paradigm, however, all probabilities are viewed as degrees of belief, making them intrinsically subjective. In our setup, where distributions in a model are indexed by parameters, this would make each distribution a conditional law on the data given the parameter. In a sense, this means that the parameters themselves are seen as random, in which case data would be generated from the model itself and not from a fixed distribution. Inference then reduces to computing the conditional distribution of the parameter, given the data, a task that can be performed via Bayes' formula which we reproduce below.

To make it precise, consider a model $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ with each distribution P_θ admitting a density p_θ. Assume that we have a vector of i.i.d. observations, $\mathbf{X}^{(n)}$ from the model $\mathcal{P}$. The Bayesian approach starts with the selection of a *prior* measure Π on Θ (and so on $\mathcal{P}$) which is then combined with the likelihood $p_\theta^{(n)}$ – seen now as the conditional density of the data, given the parameter – leading to a *posterior* distribution $\Pi(\cdot|\mathbf{X}^{(n)})$ via Bayes' formula:

$$\Pi\big(\mathcal{T}|\mathbf{X}^{(n)}\big) = \frac{\int_{\mathcal{T}} p_\theta^{(n)}\big(\mathbf{X}^{(n)}\big)\, d\Pi(\theta)}{\int_{\Theta} p_\theta^{(n)}\big(\mathbf{X}^{(n)}\big)\, d\Pi(\theta)} \tag{1.21}$$

for measurable sets $\mathcal{T} \subseteq \Theta$.

Our use of Bayes' formula in this thesis will be pragmatic and in line with [38], in that we will use (1.21) while still assuming that the data are being generated from a fixed model P_0. The prior measure encodes what is known about the parameter before data is observed, for example, its range. Once data is observed, the formula (1.21) prescribes how the prior should be updated to become a posterior measure. The posterior, being a conditional measure on the space Θ given the data $\mathbf{X}^{(n)}$, can in principle be used for inferential purposes, say, to produce *Bayesian point estimates*. Our treatment of Bayes' formula is therefore essentially algorithmic in that we see it as a means for making inference on the parameters; the prior will just be seen as a parameter of the algorithm.

If we are to produce estimators from a posterior measure, it is important to understand how such estimators behave probabilistically, as the sample size grows. Under the assumption that the data are distributed according

to some measure P_0, we intuitively expect the posterior measure to concentrate around laws which are in some sense close to P_0; if we were then to draw a realisation from the posterior, this realisation would ideally be close to θ_0 with high probability in much the same way that a realization from an estimator is expected to be close to the parameter to be estimated. This is indeed the case for parametric models under very mild assumptions on the prior (cf. [49]), but the same is not necessarily true for non-parametric models. For non-parametric models, the choice of the prior distribution plays a critical role in determining the asymptotic properties of the posterior distribution and therefore of any Bayesian estimators (cf. [29, 82]).

Before we can talk about statements concerning distances between data or between distributions, we must endow the model $\mathcal{P}$ with a metric. For two densities p_θ and p_ϑ with respect to some common dominating measure and $\theta, \vartheta \in \Theta$, define the (squared) Hellinger metric,

$$h^2(\theta, \vartheta) = 2\Big(1 - \mathbb{E}_\vartheta \sqrt{p_\theta(X)/p_\vartheta(X)}\Big). \tag{1.22}$$

Note that since we are working on a parametrized model, we ease the notation by writing $h^2(\theta, \vartheta)$ instead of $h^2(p_\theta, p_\vartheta)$; we return to this point later. In [38], sufficient conditions on a prior Π are provided, under which the corresponding posterior $\Pi(\cdot|\mathbf{X}^{(n)})$ verifies

$$\Pi\Big(\theta \in \Theta : h(\theta, \theta_0) \geq M\varepsilon_n | \mathbf{X}^{(n)}\Big) \to 0, \quad \text{as } n \to \infty, \tag{1.23}$$

for large enough $M > 0$, in P_0-probability, where h is the Hellinger metric (1.22); we say that the posterior contracts around θ_0 (or P_0) at a rate ε_n, with respect to the metric h. The interpretation of this convergence is that, as n grows to infinity, most of the mass in the (random) posterior measure will, in P_0 probability, be concentrated in a Hellinger ball of radius of order ε_n around the distribution P_0.

One can construct Bayesian point estimates with good frequentist properties from posteriors which verify (1.23). The posterior mean can be shown to be an ε_n-consistent estimator for θ_0 so long as the left-hand-side of (1.23) goes to zero fast enough as n goes to infinity (cf. [38]). If one considers a Hellinger ball of radius $M\varepsilon_n$, centered at θ, in the support of the posterior, then the value $\hat{\theta}$ for θ which (nearly) maximises the posterior mass in that ball is also an ε_n-consistent estimator for θ_0; details can be found in [65]).

In many situations it might be difficult to construct these point estimates directly since the posterior will not have an analytic expression. This happens whenever the integral in the denominator of the posterior measure (1.21) is not tractable, which often happens for arbitrary priors. When working with parametric models, this can sometimes be avoided by working with so-called conjugate priors – these are priors for which the resulting posterior has a closed form distribution. Generally, this is almost never the case when applying Bayes' formula to non-parametric models, and we usually only obtain a posterior as a multidimensional integral. This was a grave drawback to non-parametric Bayesian Statistics until the development of sampling techniques like *Markov Chain Monte Carlo*, abbreviated MCMC, from the work of [68] and [42]. We will present a brief summary of these algorithms in Section 1.12.

Sufficient conditions under which (1.23) can be shown to take place typically involve the Kullback-Leibler

divergence and another Csiszár f- divergence, respectively,

$$K(\theta, \vartheta) = -\mathbb{E}_\vartheta \log\big(p_\theta(X)/p_\vartheta(X)\big), \tag{1.24}$$
$$V(\theta, \vartheta) = \mathbb{E}_\vartheta \log^2\big(p_\theta(X)/p_\vartheta(X)\big), \tag{1.25}$$

and the corresponding *Kullback-Leibler ball* $B(\varepsilon, \theta_0) = \big\{\theta \in \Theta : K(\theta, \theta_0) \le \varepsilon^2, V(\theta, \theta_0) \le \varepsilon^2\big\}$. Let, on a metric space $(\mathcal{A}, d)$, the *covering number* $N(\varepsilon, A, d)$ be the minimal number of d-balls of radius $\varepsilon > 0$ needed to cover a subset $A \in \mathcal{A}$. The following theorem can be shown to hold.

Theorem 1.7 (Posterior contraction [37])
Suppose that for two positive sequences $\varepsilon_n \ge \bar{\varepsilon}_n$ such that $n\bar{\varepsilon}_n^2 > 1$ and $\varepsilon_n \to 0$ as $n \to \infty$, sets $\Theta_n \subseteq \Theta$ and constants $c_1, c_2, c_3, c_4 > 0$, the following conditions hold:

$$\log N\big(\varepsilon_n, \Theta_n, h\big) \le c_1 n\varepsilon_n^2, \tag{1.26}$$
$$\Pi(\Theta \backslash \Theta_n) \le c_2 e^{-(c_3+4)n\bar{\varepsilon}_n^2}, \tag{1.27}$$
$$\Pi(B(\bar{\varepsilon}_n, \theta_0)) \ge c_4 e^{-c_3 n\bar{\varepsilon}_n^2}. \tag{1.28}$$

Then, for large enough $M > 0$, $\Pi\big(\theta \in \Theta : h(\theta, \theta_0) \ge M\varepsilon_n | \mathbf{X}^{(n)}\big) \to 0$ as $n \to \infty$ in P_0-probability.

Note that if some semi-metric d verifies $d(\theta, \theta_0) \le h(\theta, \theta_0)$ for all $\theta \in \Theta$ such that $h(\theta, \theta_0)$ is small enough, then for sufficiently large n we have $\{\theta \in \Theta : d(\theta, \theta_0) \ge M\varepsilon_n\} \subseteq \{\theta \in \Theta : h(\theta, \theta_0) \ge M\varepsilon_n\}$ which implies that under the conditions of Theorem 1.7,

$$\Pi\big(\theta \in \Theta : d(\theta, \theta_0) \ge M\varepsilon_n | \mathbf{X}^{(n)}\big) \to 0, \quad \text{as } n \to \infty,$$

in P_0-probability, for M as in Theorem 1.7. Upper bounds on the metric h, and the discrepancies K and V, on the other hand, are useful to express the conditions of the theorem in terms of simpler metrics – such as l_p or L_p metrics, depending on the type of parameter set we work with – which makes verifying the conditions simpler. We elaborate on this in Chapters 3 and 4.

A few words about the conditions of Theorem 1.7. First note that these conditions do not, in principle, impose constraints on the dimension parameter set Θ other than requiring it to admit a subset Θ_n – sometimes called a *sieve* – with finite entropy as prescribed by the *entropy condition* (1.26). This means that this theorem may be applied to parametric[2] models as well as to non-parametric models.

The sieve Θ_n, or equivalently, the sub-model $\mathcal{P}_n = \{p_\theta : \theta \in \Theta_n\}$, can be seen as a sequence of *good approximations* for the true model P_0. The *remaining mass* condition (1.27) requires that the prior should be mostly supported on such a set Θ_n. Another lower bound on the complexity of the approximating models is enforced by the *prior mass condition* (1.28). Since θ_0 is an arbitrary point in Θ, one way to verify this condition is to pick a set Θ_n which is appropriately dense in Θ, such that any $\theta_0 \in \Theta$ $B(\bar{\varepsilon}_n, \theta_0)$ captures enough prior mass; this makes condition (1.26) more restrictive, however. Another possibility is to pick a larger sequence $\bar{\varepsilon}_n$ but this deteriorates

[2] For parametric models this theorem might deliver sub-optimal rates. There are slightly stronger conditions which remedy this (cf. [37]).

the rate. This condition also suggests that it might be reasonable to pick priors that are in some sense uniform over the parameter set Θ since the centre of the Kullback-Leibler ball is in principle arbitrary.

It is the interplay between these conditions that defines the range of rates which can be shown to be attainable for a given prior. Alternatively, for a given rate, this interplay gives sufficient conditions for a prior to deliver that rate. This result serves then both as a statement about the asymptotical behavior of a fixed prior but can also guide us in the construction of priors with good (frequentist) asymptotical properties.

There are several different priors that can be put on non-parametric models; we mention but a few that we will refer to again later on in this thesis. Perhaps the most straightforward prior we can consider on non-parametric models is to take a process whose trajectories live, almost surely, on the parameter set Θ and then take the law of such a process as a prior on Θ; these are called *Stochastic process priors*. Another possibility is to expand the functions on the space Θ on a particular basis (e.g. Fourier, splines, wavelets, Bernstein polynomials); we then consider either the full basis or a truncated basis (where only the first basis functions are taken) and put a prior on the coefficients of the chosen basis functions; these are usually reffered to as *random series priors*. It is also common to use *link functions*, smooth, increasing, bounded functions, which serve the purpose of mapping the range of a random process onto Θ. These can be used to enforce desired properties on the functions on the support of the prior, such as positivity and boundedness, without having to manipulate the process directly.

The fact that any information about the parameter set may, in principle, be incorporated into a prior distribution is certainly an attractive feature, mainly in applications. Another reason which may account for the popularity that Bayesian Statistics have gained in recent years, is the fact that it represents, in some sense, a unified approach to inference: any unknown parameters of the model are endowed with a prior and Bayes' formula produces a posterior which can be used to draw conclusions on the unknown parameters. This makes, at least conceptually, the construction of adaptive priors, for example, quite simple as we will see bellow. Obtaining point estimates also becomes straightforward and the problem of establishing asymptotics for such estimates can often be reduced, by use of consistency theorems such as the one above, to measure-theoretical considerations about the particular prior measure being used and to studying the approximation properties of the sieves Θ_n.

The problem of adaptive estimation, as seen in Section 1.7, consists of performing estimation over a large family of models of the type $\Theta_{\mathcal{A}} = \bigcup_{\alpha\in\mathcal{A}} \Theta_\alpha$ where the order of the minimax rate $r_n^p(\Theta_\alpha)$ for each model Θ_α is potentially different. If we knew that the data were distributed according to a parameter with regularity α in the model Θ_α, we could pick an appropriate prior Π_α and use the resulting posterior to derive an estimator which achieves the rate $r_n^p(\Theta_\alpha)$. For the examples seen in the previous paragraph, this choice would involve considering a specific number of basis function depending on α, enforcing the values of the coefficients on a basis to decay at a certain rate depending on α, or simply picking processes whose trajectories match the regularity of the functions in Θ_α. An adaptive estimator must however attain the rate $r_n^p(\Theta_\alpha)$ when $\theta_0 \in \Theta_\alpha$ *without* knowledge of the smoothness α; we also say it is *rate-adaptive*. If we know of a prior Π_α whose respective posterior $\Pi_\alpha(\cdot|\mathbf{X}^{(n)})$ contracts at a rate $r_n^p(\Theta_\alpha)$ when the data comes from a model in Θ_α, then we may treat α itself as a unknown, so-called *hyper-parameter* of the model, and endow it with a *hyper-prior*, in effect mixing the several priors Π_α. If the hyper-prior is chosen in an appropriate fashion, the resulting prior, which no longer depends on α, will result in a posterior which attains the rate $r_n^p(\Theta_\alpha)$. (In this case we refer to both the prior and the posterior as being adaptive).

1.12 Markov chain Monte Carlo Sampling

As mentioned in the previous section, it is common in non-parametric Bayes for the normalization constant of the posterior distribution (1.21) not to be known explicitly. In these cases one can use Markov Chain Monte Carlo (MCMC) samplers to obtain approximate samples from the posterior, which can then be used to compute Bayesian estimates. MCMC samplers are Markov chains constructed to have as stationary distribution a chosen *target distribution* which is known only up to a constant. These samplers are closely related to simulated annealing optimization procedures and acceptance-rejection methods [78]. The chain is started from an arbitrary state and evolves according to the following procedure. Suppose that the *current state* of the chain is $\boldsymbol{\theta}$, and that the target distribution admits a density f with a support whose elements all have the same dimension. The next state of the chain is determined in two steps: first, an arbitrary state, called a *proposal*, is generated from a fixed, yet arbitrary distribution $\boldsymbol{\vartheta} \sim g(\cdot|\boldsymbol{\theta})$, from which we can sample directly; secondly, this state is either *accepted* with probability

$$A(\boldsymbol{\theta} \to \boldsymbol{\vartheta}) = \min\left(1, \frac{f(\boldsymbol{\vartheta})g(\boldsymbol{\theta}|\boldsymbol{\vartheta})}{f(\boldsymbol{\theta})g(\boldsymbol{\vartheta}|\boldsymbol{\theta})}\right), \tag{1.29}$$

in which case the next state of the chain is $\boldsymbol{\vartheta}$, or rejected, in which case the next state of the chain is $\boldsymbol{\theta}$. It is straightforward to check that the transition kernel which we just described is, under very mild assumptions, reversible and has f as (unique) invariant, stationary density: it is sufficient (cf. [77]) to require the support of f (call it $\operatorname{supp} f$) to be connected and for the support of the proposals to cover the support of f, i.e.

$$\bigcup_{\boldsymbol{\theta} \in \operatorname{supp} f} \operatorname{supp} g(\cdot|\boldsymbol{\theta}) \supseteq \operatorname{supp} f.$$

It is clear from its definition that the probability (1.29) can be computed even if the density f is known only up to a constant; in fact, we only need to be able to compute the ratios $f(\boldsymbol{\vartheta})/g(\boldsymbol{\vartheta}|\boldsymbol{\theta})$ up to a constant that does not depend on $\boldsymbol{\theta}$. The Markov chain whose kernel we just described represents the classical formulation of the Metropolis–Hastings algorithm [42].

Using this mechanism, we can now sample from a posterior distributions, for example. One possibility is pick $g(\boldsymbol{\vartheta}|\boldsymbol{\theta}) = \Pi(\boldsymbol{\vartheta})$ in which case the ratio in (1.29) reduces to a likelihood ratio, namely,

$$A(\boldsymbol{\theta} \to \boldsymbol{\vartheta}) = \min\left(1, \frac{p_{\vartheta}^{(n)}(\mathbf{X}^{(n)})}{p_{\theta}^{(n)}(\mathbf{X}^{(n)})}\right).$$

This is a particular case of what is called an *independent Metropolis–Hastings sampler* since the proposals are independent of the current state of the chain. Another popular choice are the so-called *random walk Metropolis–Hastings* samplers: we take $\boldsymbol{\vartheta} = \boldsymbol{\theta} + \sigma\boldsymbol{\xi}$, where $\sigma > 0$, $\boldsymbol{\xi} \sim g(\cdot)$, a zero mean distribution, symmetrical around the origin. Proposed states are in this case perturbations of the current state and the acceptance probability becomes

$$A(\boldsymbol{\theta} \to \boldsymbol{\vartheta}) = \min\left(1, \frac{\Pi(\boldsymbol{\vartheta}|\mathbf{X}^{(n)})}{\Pi(\boldsymbol{\theta}|\mathbf{X}^{(n)})}\right).$$

The standard deviation σ becomes a parameter of the algorithm and typically needs to be picked according to the dimension of the support of the target distribution.

In this thesis, we also consider *adaptive priors* whose use we exemplify with both real and simulated data. As mentioned in Section 1.11, adaptation in the Bayesian setting is usually achieved by mixing different priors, each of which is optimal over a fixed model in the family. This means that the support of such a posterior is the union of spaces whose dimension might be different and the sampler outlined above cannot be used. For such situations we must employ so-called *reversible jump Markov chain Monte Carlo* samplers (cf. [40, 41]).

Reversible jump MCMC samplers generalize the Metropolis–Hastings samplers presented above in the sense that a) they allow for different *types of moves* for the chain, each corresponding to a different transition kernel and b) the kernels are defined in a more general way such that moves between states with different dimension are possible. The latter is the main point of interest of these samplers – their capability of producing samples from adaptive posteriors. To set up a reversible jump MCMC sampler we may construct a countable number of transition kernels. Generating the next element of the chain then has an extra step corresponding to selecting which transition kernel to use. We present here a specific sampler driven by the generation of random numbers which can be found in greater detail and generality in [41].

Suppose that the support of the target distribution is contained in the family of models $\Theta = \bigcup_{\alpha\in\mathcal{A}} \Theta_\alpha$; call a generic state of the chain $(\alpha, \boldsymbol{\theta}_\alpha)$ where the subscript α indicates the *model index*, such that $\boldsymbol{\theta}_\alpha$ belongs to Θ_α, which is a space with dimension d_α. We index each different type of move we wish to consider by $j \in \mathcal{J}$, a countable set. Next, we pick probabilities $p(j, \alpha, \boldsymbol{\theta}_\alpha)$ such that for each pair $(\alpha, \boldsymbol{\theta}_\alpha)$, $\sum_{j\in\mathcal{J}} p(j, \alpha, \boldsymbol{\theta}_\alpha) = 1$. These probabilities are parameters of the sampler. Given the current state of the chain, $(\alpha, \boldsymbol{\theta}_\alpha)$, the first step in producing the next state is to select which type of move $j \in \mathcal{J}$ will be performed. This is done at random according to $p(\cdot, \alpha, \boldsymbol{\theta}_\alpha)$. Once a move type, say j, has been selected, we pick the next state of the chain by constructing a random proposal $(\beta, \boldsymbol{\vartheta}_\beta)$ according to the j-th kernel – a proposal which, as before, will either be accepted or not. We construct this proposal by first selecting two numbers $r_\alpha, r_\beta \in \mathbb{N}_0$ such that $r_\alpha + d_\alpha = r_\beta + d_\beta$. We then generate two random vectors $\mathbf{u} \sim g_{j,r_\alpha}(\cdot)$ and $\mathbf{v} \sim h_{j,r_\beta}(\cdot)$ of dimension r_α and r_β, respectively, and take $(\beta, \boldsymbol{\vartheta}_\beta)$ such that $\boldsymbol{\vartheta}_\beta$ verifies $(\boldsymbol{\vartheta}_\beta, \mathbf{v}) = \rho_j(\boldsymbol{\theta}_\alpha, \mathbf{u})$, where $\rho_j(\cdot)$ is, for each $j \in \mathcal{J}$, a diffeomorphism. These extra "dummy" coordinates $\mathbf{u}$ and $\mathbf{v}$ are introduced to do what is known as *dimension matching*, allowing the chain to effectively jump between two "augmented" states $(\alpha, \boldsymbol{\theta}_\alpha, \mathbf{u})$ and $(\beta, \boldsymbol{\vartheta}_\beta, \mathbf{v})$ of equal dimension. Once the jump has been performed, these extra, auxiliary coordinates may be ignored. If again f represents the target density, this proposal is accepted with probability

$$A_j\big((\alpha, \boldsymbol{\theta}_\alpha) \to (\beta, \boldsymbol{\vartheta}_\beta)\big) = \min\left(1, \frac{f(\beta, \boldsymbol{\vartheta}_\beta)\, p(j, \beta, \boldsymbol{\vartheta}_\beta)\, h_{j,r_\beta}(\mathbf{v})}{f(\alpha, \boldsymbol{\theta}_\alpha)\, p(j, \alpha, \boldsymbol{\theta}_\alpha)\, g_{j,r_\alpha}(\mathbf{u})} \left|\frac{\partial(\boldsymbol{\vartheta}_\beta, \mathbf{v})}{\partial(\boldsymbol{\theta}_\alpha, \mathbf{u})}\right|\right). \tag{1.30}$$

If this proposal is accepted, then the chain jumps to the state $(\beta, \boldsymbol{\vartheta}_\beta)$, otherwise remaining at state $(\alpha, \boldsymbol{\theta}_\alpha)$. (The factor on the far right is the absolute value of the determinant of the Jacobian matrix of the mapping ρ_j). Note that (1.30) generalizes (1.29). This particular choice results in a Markov chain that is reversible and has f as stationary distribution. We will present two concrete samplers in Chapters 3 and 4.

1.13 Outline of this Thesis

Here we describe in a bit more detail the contributions contained in this thesis. Most of the content of the remaining chapters of this thesis is based on [7–12]; some extra material was added to Chapters 3, 5 and 6 which

is not in the mentioned references.

1.13.1 Period Estimation for Cyclical Inhomogeneous Poisson Process

In Chapter 2 we treat a semi-parametric estimation problem: given an inhomogeneous Poisson process with a periodic intensity function, estimate its period. We construct an M-estimator for the period; the intensity function is treated as an infinite-dimensional nuisance parameter. We establish the consistency of the estimator and in passing address the issue of the identifiability of the period, which for this problem cannot be taken for granted. In the literature, the issue of the identifiability of this parameter is mostly glossed over but we treat it here in detail based on a study of the Ergodic behavior of the expectation of the criterion function used to define the M-estimator. We close this chapter by presenting two numerical studies. The first is based on synthetic data. We look into the performance of the estimator by comparing it with already existing estimators in the literature. In the second study, we apply the estimator to a large, real data set of arrivals of calls to a bank's call center which we model as a Poisson process. This chapter is based on [11].

1.13.2 Bayesian Smoothing For Inhomogeneous Poisson Processes

In Chapter 3 we continue the study of the call centre data set used in Chapter 2. Having concluded that this data has daily periodicity, we apply Bayesian non-parametric methods to estimate the daily intensity of the data. This is done by using the adaptive, spline based priors that we present in Chapter 4. The estimates are obtained via a reversible jump Markov Chain Monte Carlo sampler designed specifically for these spline priors. The analysis of this dataset is used to illustrate the general theoretical results that we establish in this chapter. These results can be used to make non-parametric inference on the intensity function of a Poisson process. They are formulated in such a way that, when combined with already existing results about specific priors (such as the ones we prove in Chapter 4), they give upper bounds on the contraction rate of the resulting posteriors. We also obtain similar results for general stochastic process priors, for when link functions are used and for when a prior is put on the scale of the unknown intensity function. We exemplify the use of our results on two specific priors: the spline prior used in our numerical study and a prior on monotonous functions based on the Dirichlet process. Our results apply to situations when the full process is observed but also when only a discretized version of the process is observed. This chapter is an expanded version of [12].

1.13.3 Adaptive Priors based on Splines with Random Knots

In Chapter 4, we establish theoretical properties for priors constructed from splines with inhomogeneous knots such as the ones used in Chapter 3. More specifically, we consider priors obtained by endowing the number and locations of the knots of the spline, as well as its coefficients, with priors. These types of priors are widely used in practice but contraction rates for them had not yet been established. We give sufficient conditions on hierarchical, adaptive spline priors, such that the resulting posterior attains the minimax rate of contraction up to a logarithmic factor. These spline priors can be used to make inference on a wide range of models. We give examples of specific priors on the knots and coefficients for which our conditions hold and do a numerical study to illustrate the advantage of endowing the location of the knots with a prior when estimating spatially inhomogeneous functions. This chapter is based on [8].

1.13.4 Tracking of Conditional Quantiles

Chapter 5 contains results on quantile tracking. The problem of regression consists of estimating the conditional mean of a response variable given the values for predictors of this quantity. Quantile regression corresponds to estimating a conditional *quantile* (e.g., the median) of the response variable, rather than its mean, resulting in estimates that are more robust. By estimating quantiles of different levels simultaneously, one can obtain a comprehensive description of the response variable. In this chapter we assume data is collected sequentially, i.e., is a time series, and we also allow for conditional dependences between the distribution of each observation and the previous ones. This a more general setting than that of just quantile regression based on independent observations since we allow the quantile of interest to depend on the past of the time series. We propose a recursive algorithm for approximating a quantile of our choice (not necessarily of a fixed level). Our main result is the derivation of general non-asymptotical, uniform bounds on the approximation error of this *tracking* sequence. We specify these bounds for different variational setups for the quantile of interest, covering in particular the important case of non-parametric quantile regression. This chapter is based on material from [7] and [9].

1.13.5 Tracking of Drifting Parameters of a Time Series

We consider in Chapter 6 a general model, a time series with arbitrary memory that is allowed to depend on a time varying, multivariate parameter; we propose an algorithm to track this *drifting* parameter. It is not required for this growing statistical model to be known completely. Instead, we just assume that a *gain function* is available which can be used to update an arbitrary approximation for parameter of interest based on a new observation from the model and an appropriate *step sequence*. This gain function can be used to define a recursive algorithm for tracking the drifting parameter. Our main result is similar to that of Chapter 5 a non-asymptotical, uniform bound on the approximation error for this estimator. For different variational setups for the drifting parameter we specify how this error bound can be minimised by an appropriate choice of the step sequence of the algorithm. We show how gain functions can be constructed for specific models and how our results generalize several classical stochastic approximation algorithms in a setting where the parameter of interest is not static. Some concrete examples are treated and we study in more detail an autoregressive model with drifting parameters. We also present some numerical results. This chapter is based on [10].

Part I

Inference

2

Period Estimation for Cyclic Inhomogeneous Poisson Processes

Poisson processes are sometimes used to model periodic, time-varying phenomena. In this chapter we propose and study a semi-parametric estimator for the period of a cyclic intensity function of an inhomogeneous Poisson process. No parametric assumptions are made on the intensity function which is treated as an infinite-dimensional nuisance parameter. We propose a new family of estimators for the period of the intensity function, address identifiability and consistency issues and present simulations which demonstrate good performance of the proposed estimation procedure in practice. We compare our method to competing methods on synthetic data and apply it to a real data set from a call center.

2.1 Introduction

The Poisson process is a widely used model for point or count data in temporal and spatial settings, for example in areas such as communication, meteorology, seismology, hydrology, astronomy, biology, actuary sciences etc. Sometimes the occurrence of the events exhibits an intrinsic periodic behavior which is modeled by a periodic intensity function of the underlying Poisson process. In this paper we study the problem of estimating the period of the intensity function of an inhomogeneous Poisson process. We propose estimators, apply them to simulated and real data and study identifiability and consistency in an asymptotic setting where the observation time, i.e. the trajectory length of the inhomogeneous Poisson process, increases indefinitely.

Estimating the period is an interesting and important problem in itself but it is also an essential ingredient for non-parametric estimation of the intensity function. Existing methods for estimating the shape of the intensity function are based on the knowledge of the period and one needs a reliable estimator of the period prior to constructing a (non-parametric) estimator for the intensity function itself (e.g. [44–46, 62, 63]). The estimation problem can formally be stated as being a semi-parametric problem where we wish to estimate a one-dimensional parameter (the period) in the presence of an infinite-dimensional nuisance parameter (the intensity function).

The literature about the specific problem of estimating the period is not abundant. Vere-Jones used a spectral approach in [99] to create an estimator based on the maximum of the Bartlett periodogram; see [79]. A parametric assumption is made in [99] about the shape of the intensity function and the proposed estimator performs well only if the true intensity function has a shape similar to the assumed one, cf. [4].

In [67] and [43] the authors present a semi-parametric estimator. Here a trajectory of an inhomogeneous Poisson process is partitioned into intervals of some fixed length. The mean quadratic deviation between the number of events in each interval and the mean number of events taken over all intervals is considered as a function of the interval length. In [67] it is argued that this criterion function should have a local minimum at the point corresponding to the period which gives a method for constructing an estimator. In [4], however, some serious flaws of this estimator were discovered. There some modifications of the estimator from [67] and [43] were proposed that perform better in practice but no theoretical results are presented for it. An extensive comparative study of the above-mentioned methods (which is contained in Table 2.4.1 ahead), on simulated data from a range of test intensity functions can be found in [4]. Even the most robust estimator did not produce satisfactory results for all test functions simultaneously.

If one is to construct a more effective estimator, then it is important to understand why previous estimators perform poorly in certain situations. The estimators described above essentially work by detecting a value τ such that a fixed Poisson point process pattern emerges in the data when considering a partition of the observed trajectory into time intervals of length τ. While being natural, such an approach has its limitations. Firstly, a pattern will also emerge if a multiple of τ is considered. Secondly, the criterion function, based on whose minimization the estimator is obtained, is not only minimized at the point τ corresponding to the period but also at any value α such that the mean number of Poisson events over time intervals of length α is fixed, raising issues of identifiability. As seen in [4], it is quite simple to design intensity functions for which the corresponding criterion function has zeros at fractions of the period as well.

In this chapter, we present an alternative M-estimator for the period. To make our approach more flexible, we introduce an auxiliary parameter $T > 0$ into the method and partition the observation time into blocks of length T. For each block we then compare the number of events in the first θ time units and the last θ times

units. This will lead to a criterion function for which we can prove convergence to a limiting function whose zeros are related to the unknown period in an explicit way. The estimator is then defined as a near zero point of this criterion function, as it converges to zero at the multiples of the period. One needs to make a proper choice of the parameter T in the method to make sure the period is identifiable and the resulting estimator is consistent. We discuss this issue in more detail in the next section, where we also propose a couple of different practical approaches how to choose this parameter. The main idea is to exploit the fact that we can vary the auxiliary parameter T and study the behavior of different data-based functionals of the criterion function as functions of parameter T.

The paper is organized as follows. In Section 2.2 we describe our model, the assumptions and the estimators. Section 2.3 addresses identifiability and consistency, subsection 2.3.1 is dedicated to establishing the convergence of the criterion function, subsection 2.3.2 to identifiability of the period and subsection 2.3.3 concerns consistency. The last section, Section 2.4, contains numerical results. First we perform the simulation study from [4] for our estimator in subsection 2.4.1 and then we apply our estimator to a real data set in subsection 2.4.2. The first five columns of Table 2.4.1 contain the simulation study from [4] for certain test intensity functions and the last column shows the performance of our estimator for the same intensity functions.

2.2 Estimation Procedures

2.2.1 Preliminaries

We suppose that we observe a trajectory $(N_t : t \in [0, n))$ of an inhomogeneous Poisson process N with intensity function λ. In other words, N is a counting process with independent increments such that $N_0 = 0$ and for all $0 < a < b$, the increment $N_b - N_a$ has a Poisson distribution with parameter

$$\int_a^b \lambda(t)\, dt.$$

From now on we assume that the intensity function λ is continuous, bounded away from both 0 and infinity, and τ-periodic for some $\tau > 0$, i.e. $\lambda(\tau + t) = \lambda(t)$ for all $t \geq 0$. The number τ is assumed to be the minimal period, i.e. there exists no $\sigma < \tau$ such that λ is σ-periodic as well. Note that the assumptions exclude the case of constant λ, which is the case of a homogeneous Poisson process. Our goal is to estimate the parameter τ from the observations $(N_t : t \in [0, n))$, without knowledge about the shape of λ.

The first step in the construction of our estimator is to fix an auxiliary number $T > 0$ and to split up the observation time interval $[0, n)$ into $\lfloor n/T \rfloor$ intervals of equal length $[(i-1)T, iT)$, $i = 1, \ldots, \lfloor n/T \rfloor$, possibly disregarding the last smaller piece (here, as usual, $\lfloor a \rfloor = \max\{k \in \mathbb{Z} : k \leq a\}$). For each of these intervals, we are going to compare the number of events occurring in the first $\theta \leq T/2$ time units to the number of events occurring in the last θ time units. By the basic properties of the Poisson process, the expected difference of these two numbers is

$$\int_{(i-1)T}^{(i-1)T+\theta} \lambda(t)\, dt - \int_{iT-\theta}^{iT} \lambda(t)\, dt. \tag{2.1}$$

Now the periodicity of the function λ implies that integrals of λ over two different intervals coincide if the length of the intervals are both equal to the same multiple of the period τ, or if the intervals have equal length and they are a multiple of the period apart. These simple facts imply that if T is such that $(l-1)\tau < T < l\tau$ for some $l \in \mathbb{N}$,

then the quantity in (2.1) vanishes for θ equal to each of the values

$$0 < T - (l-1)\tau < \tau < T - (l-2)\tau < 2\tau < \ldots < T - (l-k)\tau < k\tau < \ldots. \tag{2.2}$$

We will prove below (see Theorem 2.2) that under appropriate assumptions on the auxiliary parameter T, these are in fact the *only* points at which the function

$$\psi_n(\theta, T) = \frac{1}{\lfloor n/T \rfloor} \sum_{i=1}^{\lfloor n/T \rfloor} \left(\int_{(i-1)T}^{(i-1)T+\theta} \lambda(t)\, dt - \int_{iT-\theta}^{iT} \lambda(t)\, dt \right)^2 \tag{2.3}$$

vanishes for all n.

In view of this, we base our inference on a criterion function that estimates the function ψ_n from the data. Specifically, we define the random function Ψ_n on $[0, T/2]$ as follows:

$$\Psi_n(\theta, T) = \frac{1}{\lfloor n/T \rfloor} \sum_{i=1}^{\lfloor n/T \rfloor} \left(N_i^-(\theta, T) - N_i^+(\theta, T) \right)^2 - N_i^-(\theta, T) - N_i^+(\theta, T),$$

where

$$N_i^-(\theta, T) = N_{(i-1)T+\theta} - N_{(i-1)T}$$

is the number of counts in the first θ time units of the interval $[(i-1)T, iT)$, and, similarly,

$$N_i^+(\theta, T) = N_{iT} - N_{iT-\theta}$$

is the number of counts in the last θ time units of that interval.

The proof of Theorem 2.1 below shows that Ψ_n indeed consistently estimates ψ_n, in the sense that, for fixed θ and T, $\Psi_n(\theta, T) - \psi_n(\theta, T)$ converges in probability to 0 as n goes to infinity.

2.2.2 Procedure using a-priori knowledge about the period

In this section we suppose that we know a-priori that $\tau \in [a, b]$, where $0 < a < b < 2a$. Then if we pick an auxiliary parameter value $T \in (b, 2a)$ we have that $\tau < T < 2\tau$ and $T - \tau > T - b > 0$. It will be shown below (see Theorem 2.2) that, under the technical condition that T/τ is irrational, the criterion function $\theta \mapsto \Psi_n(\theta, T)$ converges in probability, uniformly on $[T-b, T-a]$, to a smooth, nonnegative function ψ that has a unique zero at the point $T - \tau$. This observation motivates the definition of the estimator

$$\hat{\tau}_n = \hat{\tau}_n(T) = T - \hat{\theta}_n, \tag{2.4}$$

where $\hat{\theta}_n$ is a (near) minimizer of $\theta \mapsto |\Psi_n(\theta, T)|$ on $[T-b, T-a]$, i.e. a point such that

$$|\Psi_n(\hat{\theta}_n, T)| \leq \inf_{\theta \in [T-b, T-a]} |\Psi_n(\theta, T)| + o_P(1). \tag{2.5}$$

Such a point always exists since the function $\theta \mapsto \Psi_n(\theta, T)$ is almost surely a piece-wise constant function which changes values only at finitely many points. (The $o_P(1)$ term can give some added flexibility).

In Section 2.3, we prove that $\hat{\tau}_n$ is a consistent estimator for the period τ, provided the ratio T/τ is irrational. This condition arises when studying the limit of the function $\psi_n(\theta, T)$ defined by (2.3), as $n \to \infty$. The limit function $\psi(\theta, T)$ given below by (2.8) is found by relating ψ_n to the circle rotation $R : [0, \tau) \to [0, \tau)$, $R(x) = x + T \bmod \tau$. It is well known that the ergodic behavior of this map depends on whether or not T/τ is rational; see [22]. Although formally needed for the consistency proof, the condition that T/τ is irrational does not seem to be an issue in actual computations. Simulations indicate that any choice $T \in (b, 2a)$ yields a proper estimator for the period. Moreover, this issue can be avoided by randomizing the choice of T, for instance by taking it to be uniformly distributed on $(b, 2a)$, independent of the data $(N_t,\ t \in [0, n))$, cf. Corollary 2.1 ahead.

Implementation of the outlined estimation procedure is rather straightforward. Figure 2.2.1 illustrates the method for a simulated data example. We took $a = 5$, $b = 8$, and $T = 9$ and simulated a sample path of an inhomogeneous Poisson process on the time interval $[0, n]$ with $n = 100T = 900$ and intensity function

$$\lambda(t) = 5.5 + 5\cos(t), \tag{2.6}$$

which has period 2π. The corresponding criterion function $\theta \mapsto \Psi_n(\theta, T)$ is shown on the interval $[2, 4.5]$. The function is minimal at the point $\theta = 2.7260$, which yields the estimate $T - \theta = 6.2740$ for the period $\tau = 2\pi$ with an absolute error of 0.0092.

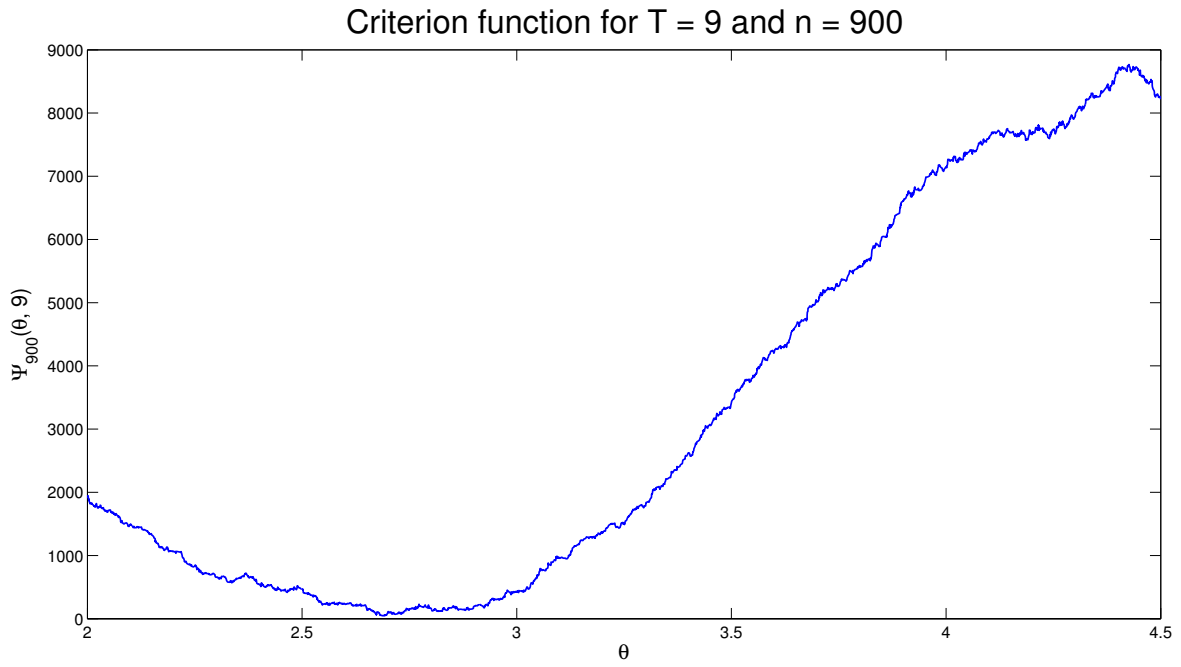

Figure 2.2.1: For each θ, the criterion function $\Psi_n(\theta, 9)$ is an average discrepancy between Poisson counts on specific intervals. By construction, if θ is close to $\theta_0 = T - \tau$ discrepancies are low since we are comparing i.i.d. Poisson counts. Asymptotically the criterion is minimized at θ_0 from which we can estimate τ since T is a parameter of our choice.

2.2.3 Procedure without a-priori knowledge about the period

In many applications it is not possible, or undesirable, to assume a-priori that the period τ belongs to some given compact interval. This leads to the problem of choosing the parameter T in the estimation procedure described above. In this section we propose alternative procedures which involve letting the auxiliary parameter T vary in a certain range. These procedures only involve choosing appropriate upper and/or lower bounds for the range of values for T being considered, which is less demanding. We propose two different approaches. Both exploit the fact that the function $\theta \mapsto |\Psi_n(\theta, T)|$ has approximate zeros at the points given by (2.2).

A graphical method

Notice that if $k\tau < T < (k+1)\tau$ for some $k \in \mathbb{N}$, then, according to (2.2), the first positive approximate zero of the function $|\Psi_n(\theta, T)|$ should be $T - k\tau$. If we now denote the first positive approximate zero of $\theta \mapsto |\Psi_n(\theta, T)|$ by $\hat{\theta}_T$, then plotting T against $T - \hat{\theta}_T$ should give a graph that approximately (for $T > \tau$) looks like a staircase function, with jumps at all multiples of the period, the jump height being always approximately equal to the period as well.

Concretely, we propose the following procedure:

1. Choose lower and upper bounds $0 < \delta < \bar{T}$ for the range of T's that will be considered.
2. For every $T \in [\delta, \bar{T}]$, define $\hat{\theta}_T = \min \arg\min_{\theta \in [\delta/2, T/2]} \left|\Psi_n(\theta, T)\right|$.
3. Plot the function $F_n : [\delta, \overline{T}] \to [0, \infty)$ defined by

$$F_n(T) = T - \hat{\theta}_T. \tag{2.7}$$

Among other things, we will prove in Section 2.3 that for every T such that T/τ is irrational and $k\tau < T < (k+1)\tau$ for $k \in \mathbb{N}$, $F_n(T)$ converges in probability to $k\tau$ (cf. Theorem 2.4). Hence for $\delta > 0$ small enough and $T > \tau$, the function F_n will asymptotically look like the step function $F(T) = \lfloor T/\tau \rfloor \tau$.

This provides a graphical method for estimating the period τ by reading off the jumps heights and locations from the graph of F_n. Figure 2.4.1 displays the graph of F_n for the real data set considered in Section 2.4.2. In practice one would compute $F_n(T)$ for all T on a grid on an interval interval $[\delta, \bar{T}]$.

If $T < \tau$, then the function $\theta \mapsto |\Psi_n(\theta, T_k)|$ has, at least in the limit, no zeros in the interval $[\delta/2, T_k/2]$. Therefore, for points T_k on the T-axis of Figure 2.4.1 for which $\theta \mapsto |\Psi_n(\theta, T_k)|$ is above a positive threshold on $[\delta/2, T_k/2]$, the point $(T_k, 0)$ is added to the graph instead of $(T_k, F_n(T_k))$. To determine the appropriate threshold, it is informative to look at the minimal values of the criterion function $m_n(T) = \min_{\theta \in [\delta/2, T/2]} |\Psi_n(\theta, T)|$ as well. In view of the results in Section 2.3, $m_n(T)$ is expected to be away from zero for $\delta < T < \tau$ as the limiting function $\theta \mapsto |\psi(\theta, T)|$ does not have zeros for $T \in (0, \tau)$. As soon as T exceeds τ, the minimal value $m_n(T)$ should drop close to zero.

The parameters δ and $\bar{T}$ should therefore be chosen in such a way that the first is sufficiently small and the second is sufficiently large to make sure the true value τ is between δ and $\bar{T}$. In fact, this can again be done by looking at the graphs of the functions F_n and m_n while varying the parameters δ and $\bar{T}$. As we have already mentioned, the function $m_n(T)$ changes from being distinctly positive to approximately zero at $T \approx \tau$ and stays

there from that point on, while the function $F_n(T)$ should have a staircase structure from that point so that the stair width equals to its height. If δ is too big (and $T > \tau$), we will see the function $m_n(T)$ close to zero immediately at $T = \delta$. If $\bar{T}$ is not big enough, we will not see distinct stairs in the plot of the function $F_n(T)$.

Simulations suggest that there is not much difference between the estimators of the period for different values of T as long as $T \in (\tau, 2\tau)$ and T is away from the endpoints τ and 2τ, the same holds for the multiples of the period.

Averaging over T

As we have already mentioned, our method is based on the fact that the criterion function $\Psi_n(\theta, T)$ has approximately the same zeros as the limiting function $\psi(\theta, T)$ defined by (2.8). The bivariate limiting function $(\theta, T) \mapsto \psi(\theta, T)$ will have zeros on the plane (θ, T) at all points on the lines defined by (2.2), i.e. $\theta = k\tau$ and $\theta = T - l\tau$ for $k, l \in \mathbb{N}$; see Figure 2.2.2 for the limiting criterion function corresponding to example (2.6). Notice that the common zeros of $\psi(\theta, T_1)$ and $\psi(\theta, T_2)$ for different T_1, T_2 will only be on the horizontal lines $\theta = k\tau$, $k \in \mathbb{N}$, unless T_1 and T_2 are separated by a multiple of the period τ.

The above observation brings us to the following idea. To avoid the issue of the choosing the parameter T, one can design another criterion function by averaging the criterion functions $\Psi_n(\theta, T)$ over a grid of values for T. Pick a sufficiently large $\bar{T}$ to make sure $\bar{T} > \tau$, take a grid $0 \le T_1 < T_2 < \ldots < T_m \le \bar{T}$ and consider a function $M_n(\theta) = \frac{1}{m}\sum_{i=1}^{m} |\Psi_n(\theta, T_i)|$. Theorem 2.1 ahead implies that if the T_i/τ are irrational (which can again be accomplished by randomization for instance), then M_n converges in probability to the function $M(\theta) = \frac{1}{m}\sum_{i=1}^{m} \psi(\theta, T_i)$. If at least one distance between the values from the grid $\{T_k,\ k = 1, \ldots, m\}$ is not a multiple of the period τ, then this limiting function has zeros only at multiples of τ. Therefore, the function $M_n(\theta)$ can be used as a new criterion function. The smallest (separated from zero by a positive value) argument of this function will provide an estimator for the period τ.

This method can for example be used on a portion of the trajectory of the inhomogeneous Poisson process, to obtain a preliminary estimate for the period which in turn allows the identification of an interval $[a, b]$ containing the period and such that $0 < a < b < 2a$, as it is required in the construction of the estimator (2.4). Once such an interval is chosen, one can use the rest of the data for the construction of the estimator (2.4).

Figure 2.2.3 shows the criterion M_n for the same simulated data example as in Section 2.2.2, where we took $\bar{T} = 75$, ($n = 900$ as before) and the uniform grid $T_k = k\bar{T}/m$, $k = 1, \ldots, m$, with $m = 100$. The function is plotted on the interval $[0, \bar{T}/4]$. We obtain in this case 6.2815 as preliminary estimator for τ.

2.3 Identifiability and Consistency

2.3.1 Uniform convergence of the criterion function

In this section we establish the uniform convergence of the criterion function $\theta \mapsto \Psi_n(\theta, T)$ to the limit function $\theta \mapsto \psi(\theta, T)$ defined by

$$\psi(\theta, T) = \frac{1}{\tau}\int_0^{\tau} \left(\int_t^{t+\theta} \lambda(s)\, ds - \int_{t+T-\theta}^{t+T} \lambda(s)\, ds \right)^2 dt. \tag{2.8}$$

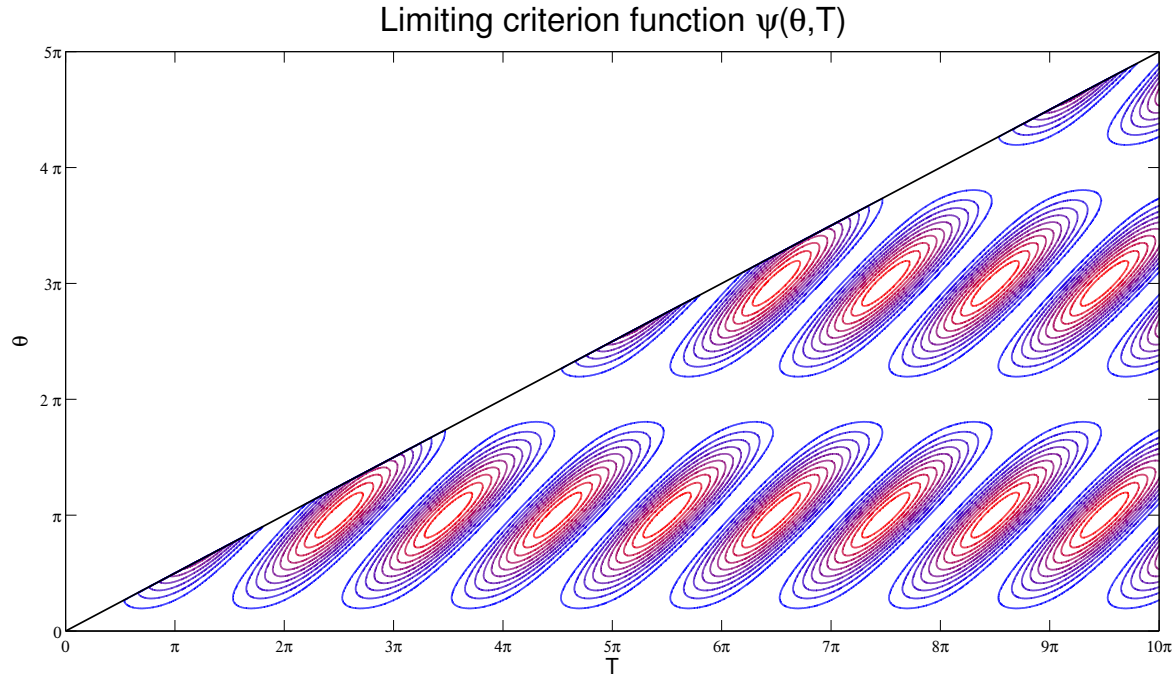

Figure 2.2.2: Contour plot of the function $(\theta, T) \mapsto \psi(\theta, T)$ corresponding to $\lambda(t) = 5.5 + 5\cos(t)$ on the set $\theta \leq T/2$. This function is null over the horizontal lines $\theta = k\tau$ and the oblique lines $\theta = T - k\tau$.

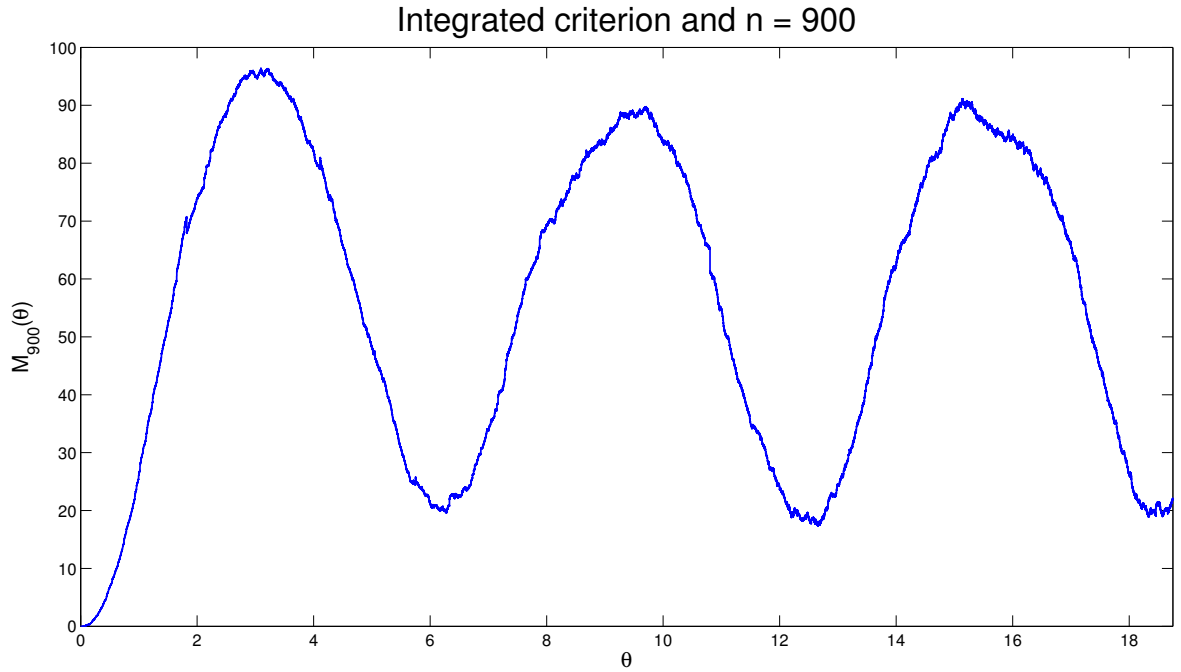

Figure 2.2.3: Example of a criterion function $M_n(\theta)$, which approximates the integrals of the function $\psi(\theta, T)$ depicted in Figure 2.2.2 over horizontal lines, i.e. $M_n(\theta) \approx \int \Psi_n(\theta, T)\, dT$.

Theorem 2.1 (Uniform convergence of the criterion function)
For all $T > 0$ such that T/τ is irrational,

$$\sup_{\theta\in[0,T/2]} \left|\Psi_n(\theta, T) - \psi(\theta, T)\right| \xrightarrow{P} 0$$

as $n \to \infty$.

Proof: We write $\Psi_n = \Psi_{1,n} - \Psi_{2,n}$, where

$$\Psi_{1,n}(\theta, T) = \frac{1}{\lfloor n/T \rfloor} \sum_{i=1}^{\lfloor n/T \rfloor} \left(\left(N_i^-(\theta, T)\right)^2 + \left(N_i^+(\theta, T)\right)^2 \right),$$
$$\Psi_{2,n}(\theta, T) = \frac{1}{\lfloor n/T \rfloor} \sum_{i=1}^{\lfloor n/T \rfloor} \left(2N_i^-(\theta, T)N_i^+(\theta, T) + N_i^-(\theta, T) + N_i^+(\theta, T) \right).$$

Note that both the $\theta \mapsto \Psi_{j,n}(\theta, T)$ are non-decreasing, càdlàg functions. For the limiting function ψ we have a similar decomposition $\psi = \psi_1 - \psi_2$, where

$$\begin{aligned}
\psi_1(\theta, T) = {} & \frac{1}{\tau}\int_0^\tau \Big(\int_t^{t+\theta} \lambda(s)\, ds \Big)^2 dt + \frac{1}{\tau}\int_0^\tau \Big(\int_t^{t+\theta} \lambda(s)\, ds \Big) dt \\
& + \frac{1}{\tau}\int_0^\tau \Big(\int_{t+T-\theta}^{t+T} \lambda(s)\, ds \Big)^2 dt + \frac{1}{\tau}\int_0^\tau \Big(\int_{t+T-\theta}^{t+T} \lambda(s)\, ds \Big) dt, \\
\psi_2(\theta, T) = {} & \frac{2}{\tau}\int_0^\tau \Big(\int_t^{t+\theta} \lambda(s)\, ds \int_{t+T-\theta}^{t+T} \lambda(s)\, ds \Big) dt \\
& + \frac{1}{\tau}\int_0^\tau \Big(\int_t^{t+\theta} \lambda(s)\, ds \Big) dt + \frac{1}{\tau}\int_0^\tau \Big(\int_{t+T-\theta}^{t+T} \lambda(s)\, ds \Big) dt.
\end{aligned}$$

The functions $\theta \mapsto \psi_j(\theta, T)$ are non-decreasing as well as differentiable.

Suppose we establish that, for $j = 1, 2$ and every fixed $\theta \in [0, T/2]$,

$$\Psi_{j,n}(\theta, T) \xrightarrow{P} \psi_j(\theta, T), \qquad \text{as} \quad n \to \infty. \tag{2.9}$$

Then the theorem would follow. Indeed, the point-wise convergence in probability, together with almost sure monotonicity (also càdlàg) and monotonicity and boundedness of the limit $\psi_j(\theta, T)$, imply uniform convergence in θ. We skip the details of the proof of this fact since it can be proved exactly in the same way as the Glivenko-Cantelli Theorem with the only differences that the (distribution) functions in the Glivenko-Cantelli Theorem are bounded by 1 and the convergence is in almost sure sense.

Thus, it remains to show (2.9). In what follows, we give the details for $j = 1$, the other case is completely analogous. By construction, the random variables $N_i^-(\theta, T)$ are independent Poisson variables with parameters

$$\int_{(i-1)T}^{(i-1)T+\theta} \lambda(t)\, dt$$

and the $N_i^+(\theta, T)$ are independent Poisson variables with parameters

$$\int_{iT-\theta}^{iT} \lambda(t)\, dt.$$

It follows that $\Psi_{1,n}(\theta, T)$ has expectation

$$\psi_{1,n}(\theta, T) = \frac{1}{\lfloor n/T \rfloor} \sum_{i=1}^{\lfloor n/T \rfloor} \Bigg(\Big(\int_{(i-1)T}^{(i-1)T+\theta} \lambda(t)\, dt \Big)^2 + \int_{(i-1)T}^{(i-1)T+\theta} \lambda(t)\, dt + \Big(\int_{iT-\theta}^{iT} \lambda(t)\, dt \Big)^2 + \int_{iT-\theta}^{iT} \lambda(t)\, dt \Bigg).$$

We have

$$\Psi_{1,n}(\theta, T) - \psi_{1,n}(\theta, T) = \frac{1}{\lfloor n/T \rfloor} \sum_{i=1}^{\lfloor n/T \rfloor} \Big(\big(N_i^-(\theta, T) \big)^2 - \mathbb{E}\big(N_i^-(\theta, T) \big)^2 \Big) + \frac{1}{\lfloor n/T \rfloor} \sum_{i=1}^{\lfloor n/T \rfloor} \Big(\big(N_i^+(\theta, T) \big)^2 - \mathbb{E}\big(N_i^+(\theta, T) \big)^2 \Big)$$

and the centered variables appearing in each sum are independent and have a variance that is bounded by a fixed polynomial function (of degree four) of

$$\int_{(i-1)T}^{(i-1)T+\theta} \lambda(t)\, dt \quad \text{and} \quad \int_{iT-\theta}^{iT} \lambda(t)\, dt.$$

Since $\theta \leq T/2$ and λ is bounded and periodic, the latter quantities are uniformly bounded. By Chebychev's inequality, it now follows that

$$\Psi_{1,n}(\theta, T) - \psi_{1,n}(\theta, T) \xrightarrow{P} 0$$

as $n \to \infty$.

To complete the proof (for $j = 1$) it remains to show that, for any fixed $\theta \in [0, T/2]$, $\psi_{1,n}(\theta, T) \to \psi_1(\theta, T)$ as n goes to infinity. Define the map $R : [0, \tau) \to [0, \tau)$ by $R(t) = (T + t) \bmod \tau$. Denoting iterations of R by R^i, i.e. R^0 is the identity, $R^1 = R$, $R^2 = R \circ R$ etc., we can then, by the periodicity of λ, write

$$\psi_{1,n}(\theta, T) = \frac{1}{\lfloor n/T \rfloor} \sum_{i=1}^{\lfloor n/T \rfloor} h(R^{i-1}(0), \theta),$$

where

$$h(t, \theta) = \Big(\int_t^{t+\theta} \lambda(s)\, ds \Big)^2 + \int_t^{t+\theta} \lambda(s)\, ds + \Big(\int_{t+T-\theta}^{t+T} \lambda(s)\, ds \Big)^2 + \int_{t+T-\theta}^{t+T} \lambda(s)\, ds.$$

It is well known (see for example [22]) that the condition that T/τ is irrational implies that the circle rotation R is uniformly ergodic, and that its unique invariant measure on $[0, \tau)$ is given by normalized Lebesgue measure on $[0, \tau)$. Therefore, by the uniform ergodic theorem ([22]),

$$\psi_{1,n}(\theta, T) \to \frac{1}{\tau} \int_0^\tau h(t, \theta)\, dt = \psi_1(\theta, T)$$

as $n \to \infty$. This completes the proof. □

Remark 2.1 *In the case where T/τ is not irrational the sequence of functions ψ_n still has a limit. If m is the least common multiple of T and τ, then the rotation $R^i(0)$ will have m/T elements that repeat themselves sequentially. The limit is the mean of the functions $h(R^0(0), \theta)$, $h(R^1(0), \theta)$, ..., $h(R^{m/T-1}(0), \theta)$.*

2.3.2 Identifiability of the period

In this section we address the identifiability issue. The following theorem establishes that the points (2.2) are the only zeros of the limiting function $\psi(\theta, T)$ defined by (2.8).

Theorem 2.2 (Zeros of the limiting criterion function)
Suppose T/τ is irrational. Then all zeros of the function $\theta \mapsto \psi(\theta, T)$, $\theta \in (0, T)$, are the points θ such that either θ is a multiple of τ, or $T - \theta$ is a multiple of τ.

Proof: As mentioned already in Section 2.2.1, from the periodicity of λ it is easy to see that the function $\theta \mapsto \psi(\theta, T)$ has zeros in the points (2.2). In the remainder of the proof we show that there are no other zeros.

Let $\theta \in (0, T)$ be such that $\psi(\theta, T) = 0$. Then

$$\int_t^{t+\theta} \lambda(s)\, ds = \int_{t+T-\theta}^{t+T} \lambda(s)\, ds \tag{2.10}$$

for all $t \in [0, \tau]$, and hence, by periodicity, for all $t \geq 0$. If we successively take $t = 0, T - \theta, 2(T - \theta), \ldots$ in (2.10), we see that

$$\int_0^{\theta} \lambda(s)\, ds = \int_{k(T-\theta)}^{k(T-\theta)+\theta} \lambda(s)\, ds \tag{2.11}$$

for every $k \in \mathbb{N}$. Observe that (2.10) implies that we also have

$$\int_t^{t+T-\theta} \lambda(s)\, ds = \int_{t+\theta}^{t+T} \lambda(s)\, ds$$

for all $t \geq 0$. Successively taking $t = 0, \theta, 2\theta, \ldots$ gives

$$\int_0^{T-\theta} \lambda(s)\, ds = \int_{k\theta}^{T-\theta+k\theta} \lambda(s)\, ds \tag{2.12}$$

for every $k \in \mathbb{N}$.

Now define the maps $R_1, R_2 : [0, \tau) \to [0, \tau)$ by putting $R_1(t) = t + (T - \theta) \bmod \tau$ and $R_2(t) = t + \theta \bmod \tau$. Then by periodicity, relations (2.11) and (2.12) imply that

$$\int_0^{\theta} \lambda(s)\, ds = \int_{R_1^k(0)}^{R_1^k(0)+\theta} \lambda(s)\, ds, \qquad \int_0^{T-\theta} \lambda(s)\, ds = \int_{R_2^k(0)}^{R_2^k(0)+T-\theta} \lambda(s)\, ds$$

for all $k \in \mathbb{N}$. Since T/τ is irrational, either $(T - \theta)/\tau$ or θ/τ is irrational. Hence, by a well-known result from ergodic theory [22], one of the orbits $\{R_i^k(0) : k \in \mathbb{N}\}$ is dense in $[0, \tau)$. To complete the proof, we note that if

the relation

$$\int_0^a \lambda(s)\,ds = \int_t^{t+a} \lambda(s)\,ds$$

holds true for all t in a set that is dense in $[0,\tau)$, then a is a multiple of the period. Indeed, the function $g(t) = \int_t^{t+\theta} \lambda(s)\,ds$ is continuous which, together with the periodicity of λ, implies that the relation $\int_0^a \lambda(s)\,ds = \int_t^{t+a} \lambda(s)\,ds$ is in fact satisfied for all $t \geq 0$. We can differentiate both sides of this relation with respect to t to obtain that $\lambda(t+a) = \lambda(t)$ for all $t \geq 0$. □

2.3.3 Consistency

Using Theorems 2.1 and 2.2 and the standard theory for M-estimators, it is now straightforward to derive consistency of the estimator (2.4).

Theorem 2.3 (Consistency of the M-estimator)
Let the estimator $\hat{\tau}_n(T)$ be defined by (2.4) with $\tau \in [a,b]$, $0 < a < b < 2a$, $T \in (b,2a)$ and T/τ irrational, then $\hat{\tau}_n \xrightarrow{P} \tau$ as $n \to \infty$.

Proof: In view of Theorem 2.2, we are in a position to apply Corollary 3.2.3 of [98]. The theorem follows. □

The technical condition that T/τ should be irrational can be handled by randomizing the choice of parameter T. Take T to be uniformly distributed on $(b,2a)$, independent of observed process $(N_t,\ t \in [0,n])$, and denote by A the event that T/τ is irrational. Then the preceding theorem implies that $\mathbb{P}(|\hat{\tau}_n - \tau| > \varepsilon \mid T)1_A \to 0$, almost surely. Since $\mathbb{P}(A) = 1$, by taking the expectation of this relation, we derive the following corollary.

Corollary 2.1 *Let the estimator $\hat{\tau}_n(T)$ be defined by (2.4) with $\tau \in [a,b]$, $0 < a < b < 2a$, $T \in (b,2a)$ and let T be uniformly distributed on $(b,2a)$, independent of observed process $(N_t,\ t \in [0,n])$. Then $\hat{\tau}_n \xrightarrow{P} \tau$ as $n \to \infty$.*

In Section 2.2.3, we outlined a graphical procedure based on the random function F_n defined by (2.7). The following theorem describes the asymptotic behavior of this function.

Theorem 2.4 (Consistency of the step function)
Suppose that T/τ is irrational and $k\tau < T < (k+1)\tau$ for $k \in \mathbb{N}$. Then

$$F_n(T) \xrightarrow{P} k\tau$$

as $n \to \infty$.

Proof: Let $k \in \mathbb{N}$ be such that $k\tau < T < (k+1)\tau$. It follows from Theorem 2.2 that the function $\theta \mapsto \psi(\theta, T)$ has finitely many zeros in the interval $[\delta/2, T/2]$, the smallest one being $T - k\tau$. Now partition $[\delta/2, T/2]$ into finitely many intervals T_j such that each interval contains exactly one zero of the function in its interior. Define

auxiliary "estimators"

$$\hat{\theta}_{T,j} = \arg\min_{\theta \in T_j} \left|\Psi_n(\theta, T)\right|$$

Notice that $\hat{\theta}_T = \min_j \hat{\theta}_{T,j}$ and by Theorem 2.1 we obtain that $\sup_{\theta \in T_j} \left|\Psi_n(\theta, T) - \psi(\theta, T)\right| \xrightarrow{P} 0$ as $n \to \infty$. Since by construction the nonnegative, continuous function $\psi(\cdot, T)$ has a unique zero in the interior of T_j, it follows by Corollary 3.2.3 of [98] that for every j, $\hat{\theta}_{T,j}$ converges in probability to the unique zeros of $\psi(\cdot, T)$ in T_j. But then the estimator $\hat{\theta}_T$ converges in probability to the smallest positive zero of $\psi(\cdot, T)$, which is $T - k\tau$. This implies that $F_n(T) \to k\tau$ in probability. □

2.4 Experimental Results

2.4.1 Simulation study

To evaluate and compare the performance of our estimator, we run the same tests as in [4]. In that paper, the authors generate inhomogeneous Poisson process trajectories for a collection of periodic intensity functions. A number of test intensity functions is considered, varying shape, number of peaks, relative frequencies and relative amplitudes of the harmonics which make up the test intensity function. We refer to [4] for the plots of the various test intensity functions. The period is then estimated on the basis of generated trajectories and the percentage of estimates that fall within a ±10% range relative to the true value of the period τ (chosen to be $\tau = 50$ by the authors) is taken as a measurement of the accuracy of considered estimation procedures. Most of the experiments have 1000 observations in approximately 20 cycles of length 50. Besides, the authors implement the sensitivity analysis with respect to the number of observations per cycle and the number of cycles observed.

Table 2.4.1 below reports the accuracy measurements (the percentage of estimates that fall within a ±10% range relative to the true value of the period τ) for different test function and different estimators. The rows in the table refer to different test functions and the columns to different estimators. Table 2.4.1 extends the table given in [4]: we added the results for our estimator $\hat{\tau}_n$ defined by (2.4) as the last column in the table.

The first column corresponds to Vere-Jones' parametric estimator from [99] based on periodogram. It is a parametric estimator and its performance is not satisfactory if the true intensity function is not close to the parametric family of functions considered in [99]. The next column summarizes the results for an M-estimator $\hat{\tau}_{n,\max}$ which is based on a certain modification of the estimator $\hat{\tau}_{n,\min}$ given in [67], its definition can also be found in [4]. This estimator clearly fails to perform well either when few periods are observed or when the intensity function is not unimodal. This could be related to identifiability issues. Columns three and four contain two "smoothed" versions of the estimator $\hat{\tau}_{n,\max}$ described in [4]. They perform slightly better than $\hat{\tau}_{n,\max}$ but still unsatisfactory for a number of test functions, especially for harmonic functions with multiple peaks per cycle (3b, 3c, 3d), or with varying relative amplitude (4a), or with varying number of points per cycle (7b, 7c, 7d). [4] omit the performance numbers for the estimator $\hat{\tau}_{n,\min}$ as, with the exception of test 7d, none of the estimates were within 10% of the true period.

According to [4], the most robust estimator among the ones they studied (periodogram, $\hat{\tau}_{n,\max}$, $\hat{\tau}_{n,2}$, and $\hat{\tau}_{n,3}$) is $\hat{\tau}_{n,2}$. But even the estimator $\hat{\tau}_{n,2}$ did not produce satisfactory results for all test function simultaneously. One heuristic explanation of this is that the criterion functions on which those estimators are based may have other minima different from the multiples of the period.

Intensity function	Periodogram	$\hat{\tau}_{n,\max}$	$\hat{\tau}_{n,2}$	$\hat{\tau}_{n,3}$	$\hat{\tau}_n$
1a Cosine	100	100	100	100	98
1b Square	100	98	98	96	96
1c Sawtooth	100	100	100	100	100
2a 2 Steps	100	98	98	96	96
2b 3 Steps	100	90	100	94	88
2c 4 Steps	100	94	96	84	86
3a 1 Pk/Cycle	100	100	100	100	98
3b 2 Pk/Cycle	0	0	96	88	100
3c 3 Pk/Cycle	0	4	22	32	100
3d 4 Pk/Cycle	0	0	46	46	100
4a $A_1/A_2 = \exp(0.5)$	0	4	22	32	100
4b $A_1/A_2 = \exp(1.5)$	0	50	96	98	100
4c $A_1/A_2 = \exp(2.5)$	0	44	100	100	100
5a min $\lambda = 0$	100	100	100	100	100
5b min $\lambda = 0.5$	100	100	100	100	98
5c min $\lambda = 0.75$	90	92	98	88	78
6a 25 Pts/Cycle	0	6	20	24	100
6b 50 Pts/Cycle	0	4	22	32	100
6c 100 Pts/Cycle	0	30	64	68	100
7a 40 Cycles	0	38	78	78	100
7b 20 Cycles	0	4	22	32	100
7c 10 Cycles	0	8	12	20	100
7d 5 Cycles	0	8	8	10	100

Table 2.4.1: Accuracy (percentage within ±10%) of estimator ($\tau = 50$).

Our criterion function is more strict in the sense that the results of Section 2.3 guarantee that, under mild conditions on the parameter T, the multiples of the period are essentially the only zeros of the criterion function. For our estimator $\hat{\tau}_n$, we took T to be a uniformly generated number in a small neighborhood of 65 for all test intensity functions λ and use the same performance criterion (the percentage of estimates that fall within a ±10% range relative to the true value of the period) as in the above mentioned paper for the sake of comparison.

The last column of Table 1 shows that our estimator performs well for all test intensity functions including the problematic ones, 3b, 3c, 3d, 4a, 7b, 7c and 7d. Whenever the estimator fails to outperform the competing estimators, it does so by a small margin and always provides comparable results – in these cases (1b, 2a, 2b, 2c, 5b, 5c) the intensities have a lot of self-similarity which would account for the fact that our estimator, which explores discrepancies between non-congruent regions of the intensity function, drops in performance. These intensities are, however, parametric and in cases close to being constant and therefore fall outside our setup anyway.

2.4.2 Real data example

In this section we apply our methodology to a real data set. This data was obtained from the S.E.E. Center (http://ie.technion.ac.il/Labs/Serveng/) of the Faculty of Industrial Engineering and Management, Technion in Haifa, Israel. It consists of counts for calls arriving at a bank's 24 hour a day call center in the United States of America. Overall there are records of more than 43 million events between the 26th May 2001 and 26th of October

2003. These events are recorded in 30 second intervals with an average of 32 calls per minute.

The main appeal of using this particular data is that it comes from a situation where one could argue that the ”true” period is in some sense known to be 24 hours and so it offers a rare opportunity to actually quantify if our estimator comes close to the ”true” value of the period or not; this is usually not possible unless the data is simulated. Indeed, an extensive empirical study (cf. [18]) made on a similar data set for data collected within the year of 1999 indicates that the intensity with which these calls are received has daily periodicity. For our data set the plot of the “step function” F_n given in Figure 2.4.1 confirms this finding.

For a more accurate estimate, we consider as given that $\tau \in [20, 30] = [a, b]$. We then take $T \in [b, 2a] = [30, 40]$ as one and a half day or 36 hours. For these choices, Figure 2.4.2 presents the criterion function on $[6, 16] = [T - b, \min(T - a, T/2)]$.

The point at which the value of the criterion function is closest to zero is found to be $\theta = 11.3667$ hours, resulting in an estimate $\hat{\tau} = T - \theta = 24.6333$ hours or 24 hours and 38 minutes.

The estimator performs quite adequately even though for this real data some of our model assumptions can be challenged. Indeed, by looking at the counts one can see that the call center receives a higher number of calls during holidays. Also, due to technical difficulties, the call center was shut down for a few hours throughout the time considered in our data set. Strictly speaking, these two facts violate the assumptions that the intensity function is periodic and that the intensity function is bounded away from zero. Still, our method is in this case robust enough to produce a reasonable estimate of the period.

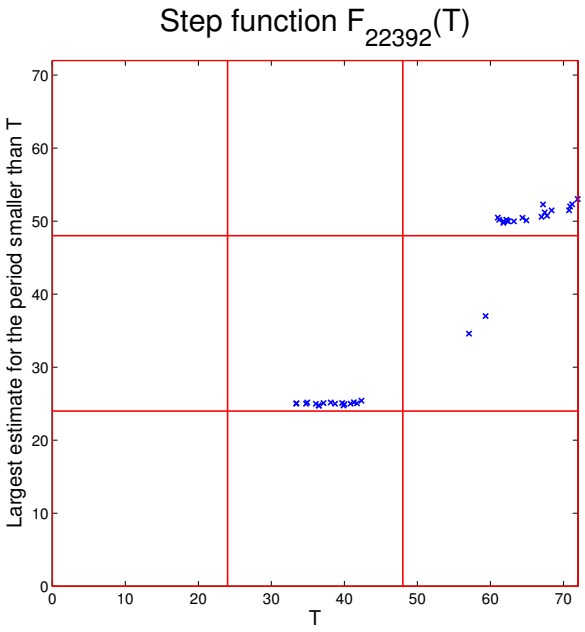

Figure 2.4.1: Step function $F_n(T)$ for the call center data. For each T, the smallest minimizer of $\psi(\cdot, T)$ on $(0, T/2]$ is $\theta_T = T - \lfloor T/\tau \rfloor \tau$. We plot T versus $T - \hat{\theta}_T$, which is an approximation for the step function $T \mapsto \lfloor T/\tau \rfloor \tau$. This provides a graphical tool for selecting a reasonable value for T. The picture suggests considering values for T roughly between 30 and 40. For comparison, we add a grid of lines separated by 24 units, the presumed period.

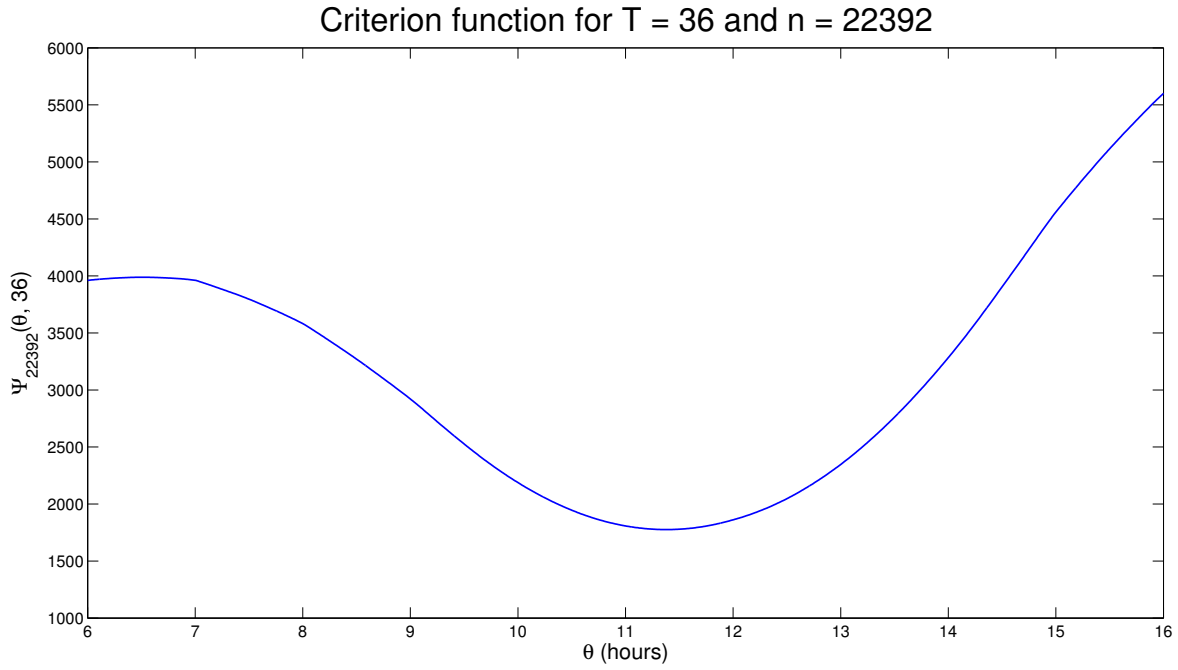

Figure 2.4.2: Criterion function $\Psi_n(\theta, 36)$ for the call center data set. The function has a clear minimizer at $\theta = 11.3667$ which corresponds to an estimate $\hat{\tau} = 24.6333$, or 24 hours and 38 minutes.

3

Bayesian Smoothing for Inhomogeneous Poisson Processes

We apply nonparametric Bayesian methods to study the problem of estimating the intensity function of an inhomogeneous Poisson process. We show that a certain spline prior which we will discuss in depth in the next chapter is computationally feasible and enjoys desirable theoretical optimality properties. We illustrate its practical use by analysing the call centre count data from Chapter 2. Theoretically we derive a new general theorem on contraction rates for posteriors in the setting of intensity function estimation. Practical choices that have to be made in the construction of our concrete prior, such as choosing the priors on the number and the locations of the spline knots, are based on these theoretical findings. We show that when properly constructed, our prior yields a rate-optimal procedure that automatically adapts to the regularity of the unknown intensity function.

3.1 Introduction

In this chapter we focus on inhomogenous Poisson processes on the line with periodic intensity functions, which are models for count data in settings with a natural periodicity.

Nonparametric Bayesian methods, which are used more and more in many different statistical settings, have so far only been used on a limited scale to analyze such models. From the applied perspective they can be attractive for making inference about intensity functions, for the same reasons as they are appealing in other situations. Estimating the intensity essentially requires some sort of smoothing of the count data and a nonparametric Bayesian approach can provide a natural way for achieving this. Using hierarchical priors we can automatically achieve a data-driven selection of the degree of smoothing. Moreover, Bayesian methods provide a way to quantify the uncertainty about the intensity using the spread of the posterior distribution. A typical implementation provides a computational algorithm that can generate a large number of (approximate) draws from the posterior. From this it is usually straightforward to construct numerical credible bands or sets.

The relatively small number of papers using nonparametric Bayesian methodology for intensity function smoothing have explored various possible prior distributions on intensities. An early reference is [70], who consider log-Gaussian priors. Other papers employing Gaussian process priors, combined with suitable link functions, include [1] and [73]. Kernel mixtures priors are considered in [56]. See also the related paper [30], in which count data is analysed using spline-based priors.

The cited papers show that nonparametric Bayesian inference for inhomogenous Poisson processes can give satisfactory results in various applications. On the theoretical side however the existing literature provides no performance guarantees in the form of consistency theorems or related results. It is by now well known that nonparametric Bayes methods may suffer from inconsistency, even when seemingly reasonable priors are used (e.g. [29]). The purpose of this chapter is therefore to present a Bayesian approach to nonparametric intensity smoothing that is both computationally feasible and at the same time theoretically underpinned by results on consistency and related issues like convergence rates and adaptation to smoothness. Such theoretical results have in the last decade been obtained for various statistical settings, including density estimation, regression, classification, drift estimation for diffusions, etc. (see e.g. [35] for an overview of some of these results). Until now, intensity estimation for inhomogenous Poisson processes has remained largely unexplored.

As motivation and starting point for treating this problem, we consider the analysis of the call center count data from Chapter 2. In the previous chapter we analysed this data and concluded that it presented daily periodicity. The same type of data were analyzed by frequentist methods in the paper [18]. We revisit the problem using a nonparametric Bayesian method employing a spline-based prior on the unknown intensity function. In addition to a single estimator of the intensity, this method provides credible bounds indicating the degree of uncertainty. In Section 3.3 we study theoretical properties of our procedure, namely consistency, posterior contraction rates and adaptation to smoothness. The results show that we have set up our procedure in such a way that we obtain consistent, rate-optimal estimation of the intensity and that the method adapts automatically to the unknown smoothness of the intensity curve, up to the level of the order of the splines that are used.

3.2 Analysis Of Call Center Data

3.2.1 Data and Statistical Model

The approach we propose and study is motivated by the wish to analyze the dataset consisting of counts of telephone calls arriving at a certain call center already considered in Chapter 2. We consider the records for the period from November 1, 2001 until December 31, 2001, covering a total of about 2.8 million incoming phone calls. These events are recorded in 30 second intervals with an average of about 32 calls per minute. The raw data are plotted in Figure 3.2.1.

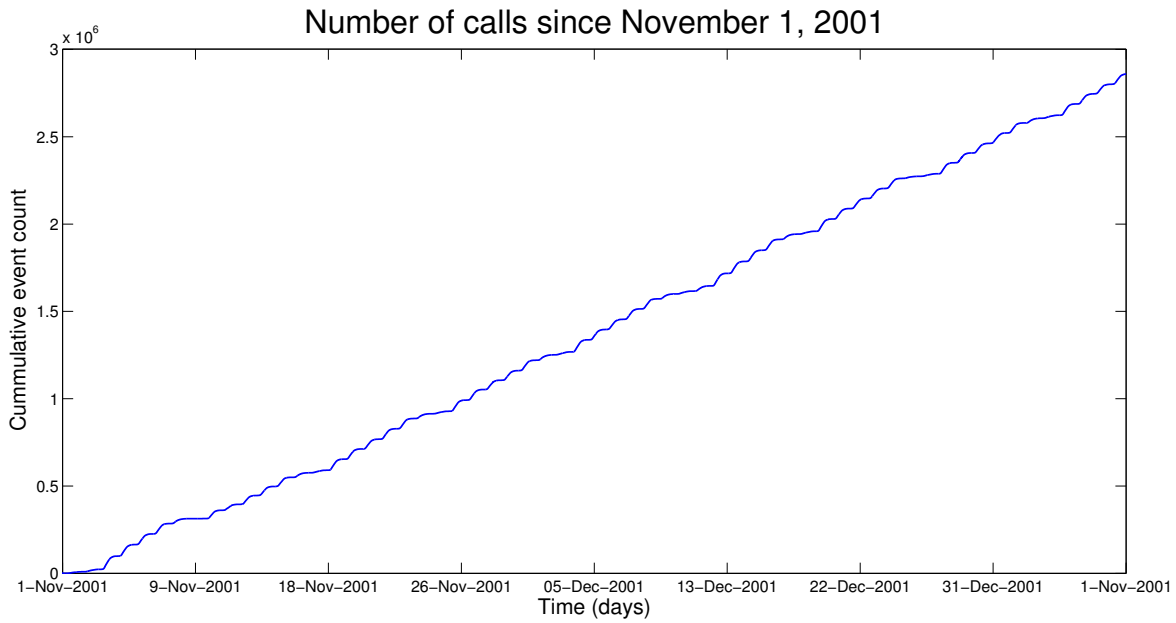

Figure 3.2.1: Number of incoming phone calls between November 1, 2001 and December 31, 2001.

We model the full count data as the realization of an inhomogenous Poisson process N with an intensity function λ that is periodic, the period being 24 hours; we have already seen in Chapter 2 that this is a reasonable assumption. This Poisson assumption is also natural and is investigated in some detail in [18], who could not find significant evidence to the contrary in a similar dataset (same kind of data, but over a different time period).

Let n be the number of days for which we have data ($n = 61$) and let τ be the period (24 hours). Then the full call arrival counting process is given by $N = (N_t : t \in [0, n\tau])$, where N_t is the number of calls arriving in the time interval $[0, t]$. The Poisson assumption means that for every $0 \le s \le t$, the number of arrivals $N_t - N_s$ is independent of the history $(N_u : u \le s)$ up till time s and that it has a Poisson distribution with mean

$$\int_s^t \lambda(u)\, du.$$

We will assume throughout that λ is at least a continuous function. The periodicity assumption then means that λ is a τ-periodic function, i.e. $\lambda(t + \tau) = \lambda(t)$ for all $t \ge 0$. For $i = 1, \ldots, n$ we define the counting process

$N^{(i)} = (N_t^{(i)} : t \in [0, \tau])$ by

$$N_t^{(i)} = N_{(i-1)\tau+t} - N_{(i-1)\tau}, \qquad t \in [0, \tau],$$

i.e. $N^{(i)}$ counts the number of arrivals during day i. Note that by the independence of the increments of the process N, the processes $N^{(i)}$ are independent inhomogenous Poisson processes which have the restriction of λ to $[0, \tau]$ as intensity function.

Our goal is to make inference about this function. Note that we do not observe the full process N. We only observe it at discrete times, namely every 30 seconds. On average about 16 calls arrive in a 30 second time interval, so we really only see aggregated counts. Let Δ be the time between observations (30 seconds) and let $m = \tau/\Delta$ be the number of counts per day that we have in our dataset ($m = 2880$ in our case). Then for every $i = 1, \dots, n$ and $j = 1, \dots, m$, the number of arrivals

$$C_{ij} = N_{j\Delta}^{(i)} - N_{(j-1)\Delta}^{(i)} \tag{3.1}$$

in the jth time interval on day i has a Poisson distribution with parameter

$$\lambda_j = \int_{(j-1)\Delta}^{j\Delta} \lambda(t)\, dt. \tag{3.2}$$

We denote the total available count data by $C^n = (C_{ij} : i = 1, \dots, n, j = 1, \dots, m)$. It follows that the likelihood is given by

$$p_\lambda(C^n) = \prod_{i=1}^{n} \prod_{j=1}^{m} \frac{\lambda_j^{C_{ij}} e^{-\lambda_j}}{C_{ij}!}. \tag{3.3}$$

In the following section we describe the prior we place on the intensity function λ.

3.2.2 Prior on the Intensity Function

There are different possible choices of priors for the function λ. A number of earlier considered options were already mentioned in the introduction (Gaussian processes, kernel mixtures, splines). Our objective is to propose a procedure which is computationally manageable on the one hand and with theoretical performance guarantees on the other. Still, there will conceivably be more than one sensible choice meeting these requirements. For our numerical study, we restrict ourselves to the investigation of the spline-based prior which was studied in Chapter 4.

This free-knot spline prior is similar to priors considered earlier in different contexts (see for instance [28, 30, 88], or, more recently [85] and the references therein). Such priors have proven to be numerically attractive and capable of capturing abrupt changes in functions of interest. This last point is relevant for our particular application, since we expect fluctuations during the day due to the varying activity of businesses over the day. Recently, several theoretical results were derived for spline-based priors in various settings as well (e.g. [38], [39], [25], [7]). We will show in the next section that the procedure that we construct and implement has several desirable theoretical properties.

Background information on splines can be found, for example, in [24] or [81]. We briefly summarize the notation and terminology. (For more detailed description see Chapter 4.) A function is called a spline of order $q \in \mathbb{N}$, with respect to a certain partition of its support, if it is $q - 2$ times continuously differentiable and when restricted to each interval in the partition, it coincides with a polynomial of degree at most $q - 1$.

Now consider $q \geq 2$. For any $j \in \mathbb{N}$, such that $j \geq q$ let $\mathcal{K}_j = \{(k_1, \ldots, k_{j-q}) \in (0,\tau)^{j-q} : 0 < k_1 < \cdots < k_{j-q} < 1\}$. We will refer to a vector $\mathbf{k} \in \mathcal{K}_j$ as a sequence of *inner knots*. A vector $\mathbf{k} \in \mathcal{K}_j$ induces the partition $\{[k_0, k_1), [k_1, k_2), \ldots, [k_{j-q}, k_{j-q+1}]\}$ of $[0,\tau]$, with $k_0 = 0$ and $k_{j-q+1} = \tau$. For $\mathbf{k} \in \mathcal{K}_j$, we denote by $\mathcal{S}^q_{\mathbf{k}}$ the linear space of splines of order q on $[0,\tau]$ with simple knots $\mathbf{k}$ (see the definition of simple knots in, e.g., [81]). This space has dimension j and admits a basis of B-splines $\{B^q_{\mathbf{k},1}, \ldots, B^q_{\mathbf{k},j}\}$. The construction of $\{B^q_{\mathbf{k},1}, \ldots, B^q_{\mathbf{k},j}\}$ involves the knots $k_{-q+1}, \ldots, k_{-1}, k_0, k_1, \ldots, k_{j-q}, k_{j-q+1}, k_{j-q+2}, \ldots, k_j$, with arbitrary extra knots $k_{-q+1} \leq \cdots \leq k_{-1} \leq k_0 = 0$ and $\tau = k_{j-q+1} \leq k_{j-q+2} \leq \cdots \leq k_j$. Usually one takes $k_{-q+1} = \cdots = k_{-1} = k_0 = 0$ and $\tau = k_{j-q+1} = \cdots = k_j$, and we adopt this choice as well. For $\mathbf{k} \in \mathcal{K}_j$ and $\boldsymbol{\theta} \in \mathbb{R}^j$ we denote by $s_{\mathbf{k},\boldsymbol{\theta}}$ the spline in $\mathcal{S}^q_{\mathbf{k}}$ that has coefficient vector $\boldsymbol{\theta}$ relative to the basis $\{B^q_{\mathbf{k},1}, \ldots, B^q_{\mathbf{k},j}\}$, i.e.,

$$s_{\mathbf{k},\boldsymbol{\theta}}(t) = \sum_{i=1}^{j} \theta_i B^q_{\mathbf{k},j}(t), \qquad t \in [0,\tau].$$

To define the prior Π on λ that we use here we first fix the order $q \geq 2$ of the splines that we use (cubic splines are popular and correspond to the choice $q = 4$) and the minimum and maximum intensities $0 \leq M_1 < M_2$. Then a draw from the prior Π is constructed as follows:

1. (Number of B-splines): Draw $J \geq q$ from a shifted Poisson distribution with mean μ.

2. (Location of the knots): Given $J = j$, construct a regular $1/j^2$-spaced grid in $(0,\tau)$. Then uniformly at random, choose $j - q$ grid elements (without replacement) to form a sequence of inner knots $\mathbf{k}$.

3. (B-spline coefficients): Also given $J = j$, and independent of the previous step, draw a vector $\boldsymbol{\theta}$ of j independent, uniform $U[M_1, M_2]$-distributed B-spline coefficients.

4. (Random spline): Finally, construct the random spline $s_{\mathbf{k},\boldsymbol{\theta}}$ of order q corresponding to the inner knots $\mathbf{k}$ and with B-spline coefficient vector $\boldsymbol{\theta}$.

The specific choices made in the construction of the prior, like the Poisson distribution on J, choosing the knots uniformly at random from a grid, etc., are motivated by the optimality theory that we derive in Section 3.3. The theory shows that there is some more flexibility, but for choices too far from the ones proposed above the performance guarantees brake down. Technically the prior on λ is the measure Π on the space $C[0,\tau]$ of continuous functions on $[0,\tau]$ given by the law, or distribution of the random spline $s_{\mathbf{k},\boldsymbol{\theta}}$ described above. The splines in $\mathcal{S}^q_{\mathbf{k}}$ are $q-2$ times continuously differentiable, hence in this sense the choice of q determines the regularity of the prior. We will see in the next section that it also determines the maximal degree of smoothness of the true underlying intensity to which our procedure can adapt. In applications like the one we are interested in here, a sensible choice of the parameters M_1 and M_2 will typically be suggested by the average number of counts per time unit in the data. In Section 3.3 we present other possibilities, such as putting a prior on these parameters as well, in order to let their value be determined by the data automatically. This is possible, but will come at an additional computational cost. The construction of the grid in step 2. is non-standard compared to other spline-based priors proposed in the literature. It is motivated by our results from Chapter 4 and will allow us to derive desirable theoretical properties in the next section.

3.2.3 Posterior inference

For the data described in Section 3.2.1, with likelihood (3.3), and the spline prior Π described in Section 3.2.2, we implemented an MCMC procedure to sample from the corresponding posterior distribution of the intensity function λ of interest. The sampler we propose here is similar to the one seen in Chapter 4. The minimal and maximal intensity parameters M_1 and M_2 were set to 200 and 20000, respectively; these values were motivated by the range of the data (time is measured in hours). We took the order q of the splines equal to 4.

Since our prior is quite similar to the ones used previously in for instance [30] or [85] in regression or hazard rate estimation settings, our computational algorithm is a rather straightforward adaptation of existing methods. A generic state of the chain is a $(2J-q+1)$-dimensional vector $\big(j, \mathbf{k}, \boldsymbol{\theta}\big)$ where $j \in \mathbb{N}$, $j \geq q$ is the model index, $\mathbf{k} = \mathbf{k}_j \in (0,\tau)^{j-q}$ is a vector of inner knots and $\boldsymbol{\theta} = \boldsymbol{\theta}_j \in \mathbb{R}^j$ is a vector of B-spline coordinates. Together, these index a spline $s_{\mathbf{k},\boldsymbol{\theta}} = s^q_{\mathbf{k}_j,\boldsymbol{\theta}_j} \in S^q_{\mathbf{k}_j}$. We will abbreviate the corresponding posterior density by $\pi\big(j, \mathbf{k}, \boldsymbol{\theta} \mid C^n\big)$. Since the splines involved are easy to evaluate and integrate we can compute the likelihood, and then the posterior, up to the normalization constant, efficiently and without any approximations being needed.

We consider four different types of moves for the MCMC chain, namely: a) perturbing the coefficients $\boldsymbol{\theta}$, b) moving the location of one knot in $\mathbf{k}$, c) birth of a new knot and d) death of an existing knot. Each of these moves is proposed, independently and respectively, with probabilities p_a, p_b, $p_c(j)$ and $p_d(j)$ where for each $j \geq q$, $p_a + p_b + p_c(j) + p_d(j) = 1$. In fact, we start by picking $0 < p_a + p_b < 1$ as parameters of the algorithm; if μ is the mean of the prior on J, then we take $p_c(q) = 1 - p_a - p_b$, $p_d(q) = 0$ and, for $j > q$, $p_{c,j} = (1-p_a-p_b)2^{-(j-q)/(\mu-q)}$ and $p_{d,j} = (1-p_a-p_b)(1-2^{-(j-q)/(\mu-q)})$. This choice results in $p_{c,j} = p_{d,j}$ if $j = \mu$, $p_{c,j} > p_{d,j}$ if $j < \mu$, $p_{c,j} < p_{d,j}$ if $j < \mu$.

When perturbing the coefficients we perform simple (Gaussian) random walk MCMC steps; the standard deviation of the random walk was chosen such that we obtained an acceptance rate of roughly 23% for this type of move, as prescribed in [34]. Let φ_j be the joint density of j i.i.d. standard normal random variables. Our proposals correspond to a move $\big(j, \mathbf{k}, \boldsymbol{\theta}\big) \to \big(j, \mathbf{k}, \boldsymbol{\vartheta}\big)$, $\boldsymbol{\vartheta} = \boldsymbol{\theta} + \sigma\mathbf{u}$, which we accept with probability $\min\big(A\big((j, \mathbf{k}, \boldsymbol{\theta}) \to (j, \mathbf{k}, \boldsymbol{\vartheta})\big), 1\big)$, with

$$A\Big(\big(j, \mathbf{k}, \boldsymbol{\theta}\big) \to \big(j, \mathbf{k}, \boldsymbol{\vartheta}\big)\Big) = \frac{\pi\big(j, \mathbf{k}, \boldsymbol{\theta} + \sigma\mathbf{u} \mid C^n\big)\, p_a \varphi_j(-\sigma\mathbf{u})}{\pi\big(j, \mathbf{k}, \boldsymbol{\theta} \mid C^n\big)\, p_a \varphi_j(\sigma\mathbf{u})} = \frac{\pi\big(j, \mathbf{k}, \boldsymbol{\vartheta} \mid C^n\big)}{\pi\big(j, \mathbf{k}, \boldsymbol{\theta} \mid C^n\big)}.$$

Moving a knot is also straightforward; one of the current $j-q$ knots, say k_i, is picked uniformly at random among those in $\mathbf{k}$ and we propose to change its location depending on how many of its neighboring position on the j^{-2}-spaced grid are free – we say that two knots k, k' are neighbors if $|k - k'| \leq j^{-2}$. This means that we propose a move $\big(j, \mathbf{k}, \boldsymbol{\theta}\big) \to \big(j, \boldsymbol{\kappa}, \boldsymbol{\theta}\big)$ where $\mathbf{k}$ and $\boldsymbol{\kappa}$ differ only at the i-th position: if k_i has two free neighboring positions, then it moves to either of them with equal probability $c_i = c_i(k_{i-1}, k_i, k_{i+1}) = 1/2$; if k_i only has one free neighboring position, then, with equal probability $c_i = 1/2$, it either moves to this free position or it does not move at all; if k_i has no free neighboring positions then it does not move, with probability $c_i = 1$. These particular choices assure the reversibility of the moves. We accept such a proposal with probability $\min\big(A\big((j, \mathbf{k}, \boldsymbol{\theta}) \to (j, \boldsymbol{\kappa}, \boldsymbol{\theta})\big), 1\big)$ for

$$A\Big(\big(j, \mathbf{k}, \boldsymbol{\theta}\big) \to \big(j, \boldsymbol{\kappa}, \boldsymbol{\theta}\big)\Big) = \frac{\pi\big(j, (k_1, \ldots, k'_i, \ldots, k_{j-q}), \boldsymbol{\theta} \mid C^n\big)\, p_b\, (j-q)^{-1}\, c_i}{\pi\big(j, (k_1, \ldots, k_i, \ldots, k_{j-q}), \boldsymbol{\theta} \mid C^n\big)\, p_b\, (j-q)^{-1}\, c_i} = \frac{\pi\big(j, \boldsymbol{\kappa}, \boldsymbol{\theta} \mid C^n\big)}{\pi\big(j, \mathbf{k}, \boldsymbol{\theta} \mid C^n\big)}.$$

Birth moves and death moves, where a new knot is respectively added and removed, are reverse moves of one

another and so we will outline only how to perform the birth move. We propose a move $\big(j, \mathbf{k}, \boldsymbol{\theta}\big) \to \big(j+1, \boldsymbol{\kappa}, \boldsymbol{\vartheta}\big)$ where we add a new knot to the vector $\mathbf{k}$ and a new coefficient to the vector $\boldsymbol{\theta}$. In doing so, a new B-spline is introduced to the B-spline basis and a new B-spline coefficient is generated. The new knot vector $\boldsymbol{\kappa}$ contains all knots from $\mathbf{k}$ rounded to the closest grid point on a $(j+1)^{-2}$ spaced grid with the extra knot then picked uniformly at random among the remaining free positions; call it $k_{i-1} < k' < k_i$. Note that this construction does not prevent two knots in $\boldsymbol{\kappa}$ from occupying the same position; such knot vectors have posterior probability 0, though, so that the probability of moving to such a state is zero. The coefficients on this basis are then picked as $\boldsymbol{\vartheta} = f(\boldsymbol{\theta}, u) = (\theta_1, \ldots, \theta_{m-1}, u, \theta_m, \ldots, \theta_j)$ where f is linear and invertible, and u is a random seed, a normally distributed random number with mean $\eta(\boldsymbol{\theta})$, to be picked later, and variance 1. The new knot will belong to the support of q B-splines, namely the i-th through $(i+q-1)$-th B-splines and we pick the index m in $\{i, \ldots, i+q\}$ depending on the knot's position within the interval $[k_{i-1}, k_i]$; namely $m = i + \lfloor (q+1)(k' - k_{i-1})/(k_i - k_{i-1}) \rfloor$. The mean of the random seed u will be picked as a weighted mean of the coefficients $\boldsymbol{\theta}$, namely, $\eta(\boldsymbol{\theta}) = \sum_{i=1}^{m-1} w_i \theta_i + \sum_{i=m}^{j} w_{i-1} \theta_i$, where the weights w_i are normalized and

$$w_i \propto \int_0^\tau B_m^{\boldsymbol{\kappa}}(t) B_i^{\mathbf{k}}(t)\, dt, \quad i = 1, \ldots, j+1.$$

With probability $\min\big(A\big((j, \mathbf{k}, \boldsymbol{\theta}) \to (j+1, \boldsymbol{\kappa}, \boldsymbol{\vartheta})\big), 1\big)$ we make the move $\big(j, \mathbf{k}, \boldsymbol{\theta}\big) \to \big(j+1, \boldsymbol{\kappa}, \boldsymbol{\vartheta}\big)$, with $\boldsymbol{\vartheta} = f(\boldsymbol{\theta}, u)$ and $\boldsymbol{\kappa} = (k_1, \ldots, k_{i-1}, k', k_i, \ldots, k_{j-q})$, where

$$A\Big(\big(j, \mathbf{k}, \boldsymbol{\theta}\big) \to \big(j+1, \boldsymbol{\kappa}, \boldsymbol{\vartheta}\big)\Big) = \frac{\pi\big(j+1, \boldsymbol{\kappa}, \boldsymbol{\vartheta} \mid C^n\big)\, p_{d,j+1}\, (j-q+1)^{-1}}{\pi\big(j, \mathbf{k}, \boldsymbol{\theta} \mid C^n\big)\, p_{c,j}\, (j^2 - j + q)^{-1} \varphi_1(u)} |J_f|$$

where $|J_f|$ is the Jacobian of the linear mapping described before.

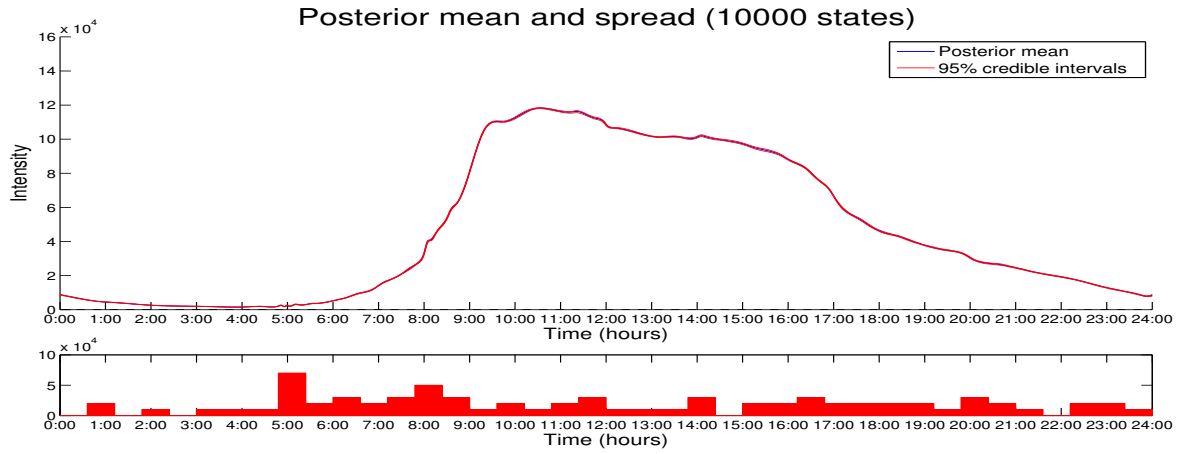

Figure 3.2.2: Top panel: posterior distribution of the intensity function λ based on the thinned data. (Blue: posterior mean, red: point-wise 95% credible intervals). Lower panel: posterior distribution of the knot locations (Histogram).

Figure 3.2.2 summarizes the outcome of the analysis. In the top panel it shows the posterior and 95% point-wise credible intervals, based on 10,000 draws from posterior. The lower panel shows a histogram for the locations of the knots corresponding to the samples from the chain used to generate the top panel. Note that as expected,

relatively many knots are placed in periods in which there are relatively many fluctuations in the intensity.

Due to the large event rate (almost 3 million counts in total), the credible bands are very narrow. To illustrate the dependence on the amount of data we ran the analysis again with a thinned dataset. We thinned the data by randomly removing counts, retaining about 1, 000 counts. The same analysis then leads to the posterior plot given in Figure 3.2.3. In this case, the uncertainty in the posterior distribution becomes clearly visible.

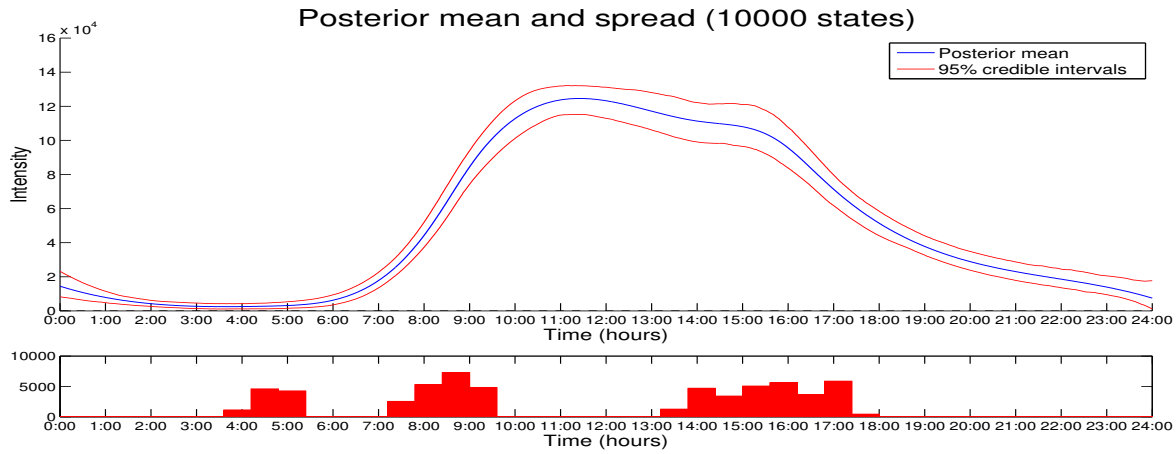

Figure 3.2.3: Top panel: posterior distribution of the intensity function λ based on the data. (Blue: posterior mean, red: point-wise 95% credible intervals). Lower panel: posterior distribution of the knot locations (Histogram).

We find that the prior that we defined in Section 3.2.2 is a computationally feasible choice for nonparametric Bayesian intensity smoothing in the context of this kind of periodic count data. In the next section we analyze its fundamental theoretical performance. See in particular Theorem 3.5 in Section 3.3.2.

3.3 Theoretical results

3.3.1 Contraction rates for general priors

We derive our theoretical results for the particular prior that we used in the Section 3.2 from general rate of contraction results that we present in this section. These are in the spirit of the general theorems about convergence rates of nonparametric Bayes procedures that were first developed for density estimation ([38]) and later for various other statistical settings, see for instance [94], [36], [74]. Here we complement this literature with general rate results regarding intensity estimation for inhomogenous Poisson processes. These results are not only applicable to the spline priors we consider in this chapter, but may also be used to analyze contraction rates of other priors. Moreover, we formulate the theorems not just for the case that we have discrete observations of aggregated data, as in our data example, but also for the case that the full counting process is observed.

The setting is as in Section 3.2.1. We fix a period $\tau > 0$. In the full observations case we assume that for $n \in \mathbb{N}$, we observe an inhomogeneous Poisson process $N^n = (N_t^n : t \in [0, n\tau])$ up to time $n\tau$, with a τ-periodic intensity function λ. Equivalently, we can say that we observe n independent inhomogeneous Poisson processes $N^{(1)}, \ldots, N^{(n)}$, indexed by $[0, \tau]$, and with a common intensity function λ, which is a positive, integrable function

on $[0, \tau]$. It is well known that the law of N under the intensity function λ is equivalent to the law of a standard Poisson process and that the corresponding likelihood is given by

$$p_\lambda(N^n) = e^{-\int_0^{n\tau}(\lambda(t)-1)\,dt+\int_0^{n\tau}\log\left(\lambda(t)\right)dN_t^n}$$

(see for instance [50]). We consider prior distributions that charge strictly positive, continuous functions. Given such a prior Π_n on λ (which we allow to depend on n) we can then compute the corresponding posterior distribution $\Pi_n(\cdot \mid N^n)$ by Bayes' formula

$$\Pi_n(\lambda \in B \mid N^n) = \frac{\int_B p_\lambda(N^n)\,\Pi_n(d\lambda)}{\int p_\lambda(N^n)\,\Pi_n(d\lambda)}.$$

Formally we can view the prior Π_n as a measure on the space $\Lambda \subset C[0, \tau]$ of all continuous, strictly positive functions on $[0, \tau]$, endowed with its Borel σ-field. If we endow Λ with the uniform norm, then the likelihood is a continuous function on Λ. Hence, the posterior is a well-defined measure on Λ.

The following theorem considers the frequentist setting in which the data are assumed to be generated by an unknown, "true" intensity function λ_0. It gives conditions on the prior Π_n under which the posterior $\Pi_n(\cdot \mid N^n)$ contracts around the true λ_0 at a certain rate as the number of observed periods tends to infinity. The assumptions and conclusions of the theorem are formulated in terms of various distances on the intensity functions. For a continuous function f on $[0, \tau]$ we define the norms $\|f\|_2$ and $\|f\|_\infty$ as usual by

$$\|f\|_2^2 = \int_0^\tau f^2(t)\,dt, \qquad \|f\|_\infty = \sup_{t\in[0,\tau]} |f(t)|.$$

For a set of positive continuous functions $\mathcal{F}$ we write $\mathcal{F}^c$ for its complement and $\sqrt{\mathcal{F}} = \{\sqrt{f} : f \in \mathcal{F}\}$. For $\varepsilon > 0$ and a norm $\|\cdot\|$ on $\mathcal{F}$, let $N(\varepsilon, \mathcal{F}, \|\cdot\|)$ be the minimal number of balls of $\|\cdot\|$-radius ε needed to cover $\mathcal{F}$.

Theorem 3.1 (Contraction rate for full observations)
Assume that λ_0 is bounded away from 0. Suppose that for positive sequences $\tilde{\varepsilon}_n, \bar{\varepsilon}_n \to 0$ such that $n(\tilde{\varepsilon}_n \wedge \bar{\varepsilon}_n)^2 \to \infty$ as $n \to \infty$, and constants $c_1, c_2 > 0$ it holds that for all $c_3 > 1$, there exist subsets $\Lambda_n \subset \Lambda$ and a constant $c_4 > 0$ such that

$$\Pi_n(\lambda : \|\lambda - \lambda_0\|_\infty \le \tilde{\varepsilon}_n) \ge c_1 e^{-c_2 n\tilde{\varepsilon}_n^2}, \tag{3.4}$$

$$\Pi_n(\Lambda_n^c) \le e^{-c_3 n\tilde{\varepsilon}_n^2}, \tag{3.5}$$

$$\log N(\bar{\varepsilon}_n, \sqrt{\Lambda_n}, \|\cdot\|_2) \le c_4 n\bar{\varepsilon}_n^2. \tag{3.6}$$

Then for $\varepsilon_n = \tilde{\varepsilon}_n \vee \bar{\varepsilon}_n$ and all sufficiently large $M > 0$,

$$\mathbb{E}_{\lambda_0}\Pi_n(\lambda \in \Lambda : \|\sqrt{\lambda} - \sqrt{\lambda_0}\|_2 \ge M\varepsilon_n \mid N^n) \to 0 \tag{3.7}$$

as $n \to \infty$.

The proof of this theorem is given in Section 3.4.1. The assumptions of the theorem parallel those of similar theorems obtained earlier for other settings including density estimation, regression, and classification. The first condition (3.4), the so-called *prior mass condition*, requires that the prior puts sufficient mass near the truth. Conditions (3.5)–(3.6) together require that most of the prior mass, quantified in the sense of the *remaining mass condition* (3.5), is concentrated on *sieves* Λ_n which are "small" in the sense of metric entropy, quantified by the *entropy condition* (3.6).

The proof of the theorem shows that conditions (3.5)–(3.6) can in fact be slightly weakened, at the cost of using more complicated distance measures on the intensities. The conditions in the theorem are easier to work with when studying concrete priors and are expected to give sharp results in many cases. We note that if under the prior all intensities are bounded away from 0, then the set $\sqrt{\Lambda_n}$ in (3.6) may be replaced by Λ_n. Moreover, if all intensities are uniformly bounded by a common constant under the prior, then the square-root norm $\|\sqrt{\cdot}\|_2$ in (3.7) may be replaced by the L^2-norm $\|\cdot\|_2$ itself. In the next section we verify the conditions of the theorem for the spline priors used in Section 3.2.1.

In the discrete observations case we only observe, for some $m \in \mathbb{N}$ and $\Delta = \tau/m$, aggregated counts C_{ij} for $i = 1, \ldots, n$ and $j = 1, \ldots, m$, given by (3.1). As before, we summarize these data using the notation $C^n = (C_{ij} : i = 1, \ldots, n, j = 1, \ldots, m)$. As explained in Section 3.2.1 the likelihood is in that case given by (3.3), where the λ_j's are defined as in (3.2). Consequently, the discrete-observations posterior is given by

$$\Pi_n(\lambda \in B \mid C^n) = \frac{\int_B p_\lambda(C^n)\,\Pi_n(d\lambda)}{\int p_\lambda(C^n)\,\Pi_n(d\lambda)}.$$

In this case it is clear that we can not consistently identify the whole intensity function λ from the data, but only the integrals $\lambda_1, \ldots, \lambda_m$. In the following theorem, which deals with the convergence of the posterior distribution in the case of discrete observations, we therefore measure the convergence using a semi-metric that identifies intensity functions with the same integrals over time intervals in which we make observations. For $\lambda, \lambda' \in \Lambda$, we define the semi-metric ρ by setting

$$\rho^2(\lambda, \lambda') = \sum_{j=1}^{m} \left(\sqrt{\lambda_j} - \sqrt{\lambda'_j}\right)^2 = \sum_{j=1}^{m} \left(\left[\int_{(j-1)\Delta}^{j\Delta} \lambda(t)\,dt \right]^{1/2} - \left[\int_{(j-1)\Delta}^{j\Delta} \lambda'(t)\,dt \right]^{1/2} \right)^2.$$

The theorem has exactly the same assumptions on the prior as Theorem 3.1 above, but gives a contraction rate relative to the distance ρ. The proof of the theorem is given in Section 3.4.2.

Theorem 3.2 (Contraction rate for discrete observations)
Assume that λ_0 is bounded away from 0. Suppose that for positive sequences $\tilde{\varepsilon}_n, \bar{\varepsilon}_n \to 0$ such that $n(\tilde{\varepsilon}_n \wedge \bar{\varepsilon}_n)^2 \to \infty$ as $n \to \infty$, and constants $c_1, c_2 > 0$ it holds that for all $c_3 > 1$, there exist subsets $\Lambda_n \subset \Lambda$ and a constant $c_4 > 0$ such that (3.4)–(3.6) hold. Then for $\varepsilon_n = \tilde{\varepsilon}_n \vee \bar{\varepsilon}_n$ and all sufficiently large $M > 0$,

$$\mathbb{E}_{\lambda_0} \Pi_n(\lambda \in \Lambda : \rho(\lambda, \lambda_0) \geq M\varepsilon_n \mid C^n) \to 0$$

as $n \to \infty$.

The requirement that Π_n must charge only positive functions ensures that such a prior may be seen as a measure on Λ, the space where the "true" intensity λ_0 lives. This can be easily enforced in our spline prior by endowing the coefficients of the B-splines with priors that put mass only on the positive reals. It is often the case though, that priors on non-parametric spaces are taken as the law of a stochastic process. If we would like to use a process whose trajectories are not necessarily positive, then we can apply to the process a so-called link function to map their range. We can then extend the previous results to general stochastic process priors defined as the law of $\Psi(W)$ for some strictly increasing link function $\Psi : \mathbb{R} \to (0, \infty)$ and a stochastic process $W = (W_t : t \in [0, \tau])$. Let W_n be a stochastic process with square integrable sample paths. In the literature there are various results for stochastic process priors that assert or imply that under certain conditions, for sequences $\tilde{\varepsilon}_n, \bar{\varepsilon}_n \to 0$ such that $n(\tilde{\varepsilon}_n \wedge \bar{\varepsilon}_n) \to \infty$, constants $c_1, c_2 > 0$ and $w_0 \in L^2[0, \tau]$, it holds that for every $c_3 > 1$ there exist measurable sets $B_n \subset L^2[0, \tau]$ and a constant $c_4 > 0$ such that

$$\mathbb{P}(\|W_n - w_0\|_\infty \leq \tilde{\varepsilon}_n) \geq c_1 e^{-c_2 n \tilde{\varepsilon}_n^2}, \tag{3.8}$$

$$\mathbb{P}(W_n \notin B_n) \leq e^{-c_3 n \tilde{\varepsilon}_n^2}, \tag{3.9}$$

$$\log N(\bar{\varepsilon}_n, B_n, \|\cdot\|_2) \leq c_4 n \bar{\varepsilon}_n^2. \tag{3.10}$$

The following theorem, whose proof can be found in Section 3.4.3, is formulated in such a way that it directly links Theorems 3.1 and 3.2 above to these existing results, so that we can easily obtain rate of contraction results for many concrete priors.

Theorem 3.3 (Contraction rates for stochastic process priors)
Let the prior Π_n be the law of $\Psi(W_n)$, for $\Psi : \mathbb{R} \to (0, \infty)$ an increasing, differentiable function such that both Ψ and the derivative of $\log \Psi$ are bounded, and $W_n = (W_n(t) : t \in [0, \tau])$ is a stochastic process with square integrable sample paths. Suppose that for sequences $\tilde{\varepsilon}_n, \bar{\varepsilon}_n \to 0$ such that $n(\tilde{\varepsilon}_n \wedge \bar{\varepsilon}_n)^2 \to \infty$ and constants $c_1, c_2 > 0$ it holds that for every $c_3 > 1$, there exist sets $B_n \subset L^2[0, \tau]$ and a constant $c_4 > 0$ such that, for $w_0 \in L^2[0, \tau]$ such that $\lambda_0 = \Psi(w_0)$, conditions (3.8) – (3.10) hold. Then for $\varepsilon_n = \tilde{\varepsilon}_n \vee \bar{\varepsilon}_n$ and all sufficiently large $M > 0$, the conclusions of Theorems 3.1 and 3.2 remain valid.

The theorem assumes implicitly that Ψ is bounded and that $\lambda_0 = \Psi(w_0)$ for some (necessarily unique) function w_0. Clearly, we must then have $\|\lambda_0\|_\infty \leq \|\Psi\|_\infty$. Since λ_0 is unknown, the only way to ensure that this holds in practice is to assume a known uniform bound M on the unknown intensity function and then choose a link function Ψ such that $\|\Psi\|_\infty \geq M$.

Since such an assumption may be undesirable in certain cases, it is of interest to devise ways to avoid it. One possibility is to use a rescaling factor for a fixed link function, and endowing this factor with an additional prior. Theorem 3.4 shows that as long as the tails of the prior on the scaling factor are appropriately thin, there is no loss in terms of rate.

Theorem 3.4 (Contraction rates for rescaled stochastic process priors)
Let the prior Π_n be the law of $M + A\Psi(W_n)$, for $\Psi : \mathbb{R} \to (0, 1)$ an increasing, differentiable function such that both Ψ and the derivative of $\log \Psi$ are bounded, a constant $M > 0$ such that $\lambda_0 > M$, $W_n = (W_n(t) : t \in [0, \tau])$ a

stochastic process with square integrable sample paths, and A an independent $(0,\infty)$*-valued random variable with a positive, continuous Lebesgue density.*

Suppose that for sequences $\tilde{\varepsilon}_n, \bar{\varepsilon}_n \to 0$ *such that* $n(\tilde{\varepsilon}_n \wedge \bar{\varepsilon}_n)^2 \gtrsim \log n$ *and constants* $c_1, c_2 > 0$ *it holds that for every* $c_3 > 1$ *there exist sets* $B_n \subset L^2[0,\tau]$ *and a constant* $c_4 > 0$ *such that, with* $w_0 \in L^2[0,\tau]$ *and* $c \geq 1$ *such that* $\lambda_0 = M + c\|\lambda_0 - M\|_\infty \Psi(w_0)$*, conditions (3.15) – (3.17) are satisfied and in addition there exists a polynomial sequence* $a_n(c_3)$ *such that*

$$\mathbb{P}(A > a_n(c_3)) \leq e^{-c_3 n \tilde{\varepsilon}_n^2}.$$

Then there exists a constant $C > 1$ *such that for* $\varepsilon_n = \tilde{\varepsilon}_n \vee \bar{\varepsilon}_n$ *and for all sufficiently large* $M > 0$*,*

$$\mathbb{E}_{\lambda_0} \Pi(\lambda : \|\lambda - \lambda_0\|_2 \geq M\varepsilon_n \mid N^n) \to 0$$

and

$$\mathbb{E}_{\lambda_0} \Pi(\lambda : \rho(\lambda, \lambda_0) \geq M\varepsilon_n \mid C^n) \to 0$$

as $n \to \infty$*.*

The proof of the theorem is given in Section 3.4.4.

We have introduced constants $c_i > 1$ and M in the assumptions of the theorem. This is to have some more flexibility when checking the conditions of the theorem for specific priors and underlying true intensity function λ_0. For Ψ^{-1} the inverse of the link function, we have

$$w_0 = \Psi^{-1}\Big(\frac{\lambda_0 - M}{c\|\lambda_0 - M\|_\infty}\Big).$$

Hence, if we assume for instance that the truth has Hölder regularity of order, α, say, then this caries over to w_0 as soon as the restriction of Ψ^{-1} to $[(\inf \lambda_0 - M)/(c\|\lambda_0 - M\|_\infty), 1/c]$ has sufficiently many bounded derivatives. Since the derivative of Ψ^{-1} typically blows up at 0 and 1, we have to assume that λ_0 is bounded away from M in such a case and take c strictly larger than 1. As mentioned before, assuming a lower bound on the intensity function is not restrictive. It also allows us to claim that, by construction, the functions in the support of the prior are uniformly bounded from zero.

In the next section we apply the theoretical results derived above to our spline prior.

3.3.2 Contraction rates for our spline prior

Having the general rate of contraction results given by Theorems 3.1 and 3.2 at our disposal, we can use them to study the performance of the spline-based priors outlined in Chapter 4 and in particular, the one defined in Section 3.2.2, in the context of estimating Poisson intensities. We fix the order $q \geq 2$ of the splines that are used. As before, let N^n be the a full path up till time $n\tau$ of an inhomogenous Poisson process N with τ-period intensity λ_0 and let C^n be the discrete-time counts $C^n = (C_{ij} : i = 1, \ldots, n, j = 1, \ldots, m)$, with C_{ij} as in (3.1).

The contraction rate of the posterior will depend on the regularity of the true intensity function, measured in Hölder sense. For $\alpha > 0$, let $\mathcal{H}_\alpha([0,\tau])$ be the space of functions on $[0,\tau]$ with Hölder smoothness α. (For $\lfloor \alpha \rfloor$

the greatest integer strictly smaller than α, having $f \in \mathcal{H}_\alpha([0,\tau])$ means that f has $\lfloor \alpha \rfloor$ derivatives and that the highest derivative $f^{(\lfloor \alpha \rfloor)}$ is Hölder-continuous of order $\alpha - \lfloor \alpha \rfloor$.)

Theorem 3.5 (Contraction rate for the spline prior)
Assume the true intensity function λ_0 belongs to $\mathcal{H}_\alpha([0,\tau])$ for some $\alpha \in (0,q]$, and $M_1 \le \lambda_0 \le M_2$. Consider the prior Π constructed in Section 3.2.2. For all $p > 1$ and all sufficiently large $M > 0$,

$$\mathbb{E}_{\lambda_0} \Pi_n\Big(\lambda \in \Lambda : \|\lambda - \lambda_0\|_2 \ge M\Big(\frac{n}{\log^p n}\Big)^{-\frac{\alpha}{1+2\alpha}} \mid N^n\Big) \to 0$$

and

$$\mathbb{E}_{\lambda_0} \Pi_n\Big(\lambda \in \Lambda : \rho(\lambda,\lambda_0) \ge M\Big(\frac{n}{\log^p n}\Big)^{-\frac{\alpha}{1+2\alpha}} \mid C^n\Big) \to 0$$

as $n \to \infty$.

Note that up to a logarithmic factor, the rate of contraction in the theorem is the optimal rate $n^{-\alpha/(1+2\alpha)}$ for estimating an α-regular function. Moreover, the prior does not depend on α. Hence the procedure automatically adapts to the smoothness of the intensity function, up to the order of the splines that are used. This theorem deals with the case that we have known bounds M_1 and M_2 for the intensity. If no upper bound is know then we can take Ψ as an identity and use Theorem 3.4. The existence lower bound $M_1 > 0$ is not restrictive since it can be enforced by adding a homogeneous Poisson process with known intensity to the data.

In Section 3.3.3 we present one last result. Instead of imposing smoothness conditions on the true intensity λ_0 we can also consider shape restrictions. This is a common practice in reliability analysis. Spline based priors can be used for this purpose since an increasing vector of B-spline coefficients results in an increasing spline function (cf. [81]) but we consider, in the next section, priors based directly on the Dirichlet process.

3.3.3 Contraction rates for monotonous intensities

We assume in this subsection that λ_0 is a non-decreasing, continuous function. Decreasing intensities, which arise for instance in certain applications in Poisson regression and reliability, can be handled analogously.

We employ a prior based on the Dirichlet process [33]. In other statistical settings it is known that when properly constructed, Dirichlet-based priors can yield posterior convergence at the rate $n^{-1/3}$ (up to a logarithmic factor) when the truth is monotone. See for instance Section 6 of [38] for an example in the context of a current status model or [94] for the estimation of a monotone drift function in a diffusion model. We are not aware of a lower bound on the rate for our model in the literature. We expect though that $n^{-1/3}$ is (eventually up to a log factor) the optimal rate.

The proof of the following result can be found in Section 3.4.6.

Theorem 3.6 (Contraction rate for Dirichlet prior)
Let the prior Π be the law of the process $A + BD$ where D is a Dirichlet process on $[0,1]$ with base measure which has a positive, continuous Lebesgue density on $[0,\tau]$ and A and B are independent, positive random variables with

positive and continuous densities, which are independent of D. Moreover, we assume that for $x > 0$ large enough,

$$\mathbb{P}(A > x^2) \asymp \mathbb{P}(B > x^2) \lesssim e^{-cxe^{x/3}} \tag{3.11}$$

for some constant $c > 0$. Let λ_0 be a non-decreasing, continuous function such that $\lambda_0(0) > 0$ and $\lambda_0(\tau) < \infty$. Then, for all sufficiently large $M > 0$,

$$\mathbb{E}_{\lambda_0}\Pi_n\Big(\lambda \in \Lambda : h(\lambda, \lambda_0) \geq M\Big(\frac{n}{\log n}\Big)^{-1/3} \mid X^n\Big) \to 0$$

as $n \to \infty$, for X^n either N^n or C^n.

We remark that if the tails of the variables A and B in the definition of the prior are heavier than the ones in the condition of the theorem then we might still get a rate, but it will typically be sub-optimal. Indeed, inspection of the proof of this result we can simply *define* the radius L_n of the range of the functions in the sieve Λ_n as the smallest number such that

$$\mathbb{P}(A > L_n) \lesssim e^{-n^{1/3}\log n}, \qquad \mathbb{P}(B > L_n) \lesssim e^{-n^{1/3}\log n}. \tag{3.12}$$

The prior and remaining mass conditions are still fulfilled with $\tilde{\varepsilon}_n \sim (n/\log n)^{-1/3}$ and the entropy condition with $\bar{\varepsilon}_n \sim (n/\sqrt{L_n})^{-1/3}$. So as long as $L_n \ll n^2$ we still obtain a rate, but it will be slower than $(n/\log n)^{-1/3}$ if (3.11) does not hold. If the tails of A and B are only sub-exponential, for instance, then (3.12) holds for $L_n \sim n^{1/3}\log n$ and we only get the rate $n^{-5/18}$, up to a logarithmic factor. For power-law tails we do not obtain a rate at all.

If in addition to the monotonicity, an a-priori upper bound L on λ_0 is assumed, then A and B can simply be taken uniform on $[0, L]$, for instance. We then obtain a rate $(n/\log n)^{-1/3}$ relative to the L^2-norm on the intensities for the posteriors $\Pi(\cdot \mid N^n)$ and relative to the semi-metric ρ for $\Pi(\cdot \mid C^n)$. Known lower bounds M on λ_0 can also be incorporated into the prior by taking Π as the law of the process $M + A + BD$ with $A + BD$ as in Theorem 3.4.

3.4 Proofs

3.4.1 Proof of Theorem 3.1

A useful observation is that we can view the statistical problem to which the theorem applies as a density estimation problem for functional data. Indeed, in the full observations case we observe a sample $N^{(1)}, \ldots, N^{(n)}$, which are independent, identically distributed random elements in the Skorohod space $D[0, \tau]$ of càdlàg (right-continuous functions with left-hand limits) on $[0, \tau]$ (see [50], Chapter VI). Under the intensity function λ, the density p_λ of $N^{(1)}$ relative to the law of a standard Poisson process indexed by $[0, \tau]$ is given by

$$p_\lambda(N) = e^{-\int_0^\tau (\lambda(t)-1)\,dt + \int_0^\tau \log(\lambda(t))\,dN_t}$$

(e.g. [50], Chapter III). Hence, the density estimation results of [38], [37] apply in our case.

We want to apply Theorem 2.1 of [37]. This gives conditions for posterior contraction rates in terms of the Hellinger distance on densities and other, related, distance measures. The Hellinger distance $h(p_\lambda, p_{\lambda'})$ is in our case given by $h^2(p_\lambda, p_{\lambda'}) = 2(1 - \mathbb{E}_{\lambda'}\sqrt{p_\lambda(N)/p_{\lambda'}(N)})$, where $\mathbb{E}_\lambda$ is the expectation corresponding to the probability measure $\mathbb{P}_\lambda$ under which the process N is a Poisson process with intensity function λ. The other relevant distance measures are the Kullback-Leibler divergence $K(p_\lambda, p_{\lambda'}) = -\mathbb{E}_{\lambda'}\log(p_\lambda(N)/p_{\lambda'}(N))$ between p_λ and $p_{\lambda'}$ and the related variance measure $V(p_\lambda, p_{\lambda'}) = \mathbb{V}_{\lambda'}\log(p_\lambda(N)/p_{\lambda'}(N))$. For a Poisson process N with intensity λ and a bounded, measurable function f,

$$\begin{aligned}
\mathbb{E}\int_0^\tau f(t)\,dN_t &= \int_0^\tau f(t)\lambda(t)\,dt,\\
\mathbb{V}\int_0^\tau f(t)\,dN_t &= \int_0^\tau f^2(t)\lambda(t)\,dt,\\
\mathbb{E}e^{\int_0^\tau f(t)\,dN_t} &= e^{-\int_0^\tau (1-\exp(f(t)))\lambda(t)\,dt}.
\end{aligned}$$

Using these relations it is straightforward to verify that we have

$$\begin{aligned}
h^2(p_\lambda, p_{\lambda'}) &= 2(1 - e^{-\frac{1}{2}\int_0^\tau \left(\sqrt{\lambda(t)}-\sqrt{\lambda'(t)}\right)^2 dt}),\\
K(p_\lambda, p_{\lambda'}) &= \int_0^\tau (\lambda(t) - \lambda'(t))\,dt + \int_0^\tau \lambda'(t)\log\frac{\lambda'(t)}{\lambda(t)}\,dt,\\
V(p_\lambda, p_{\lambda'}) &= \int_0^\tau \lambda'(t)\log^2\frac{\lambda'(t)}{\lambda(t)}\,dt,
\end{aligned}$$

respectively.

The following lemma relates these statistical distances between densities to certain distances between intensity functions. We denote the minimum and maximum of two numbers a and b by $a \wedge b$ and $a \vee b$, respectively.

Lemma 3.1

We have the inequalities

$$\begin{aligned}
\frac{1}{\sqrt{2}}\Big(\|\sqrt{\lambda} - \sqrt{\lambda'}\|_2 \wedge 1\Big) \le h(p_\lambda, p_{\lambda'}) &\le \sqrt{2}\Big(\|\sqrt{\lambda} - \sqrt{\lambda'}\|_2 \wedge 1\Big),\\
K(p_\lambda, p_{\lambda'}) &\le 3\|\sqrt{\lambda} - \sqrt{\lambda'}\|_2^2 + V(p_\lambda, p_{\lambda'}),\\
\|\sqrt{\lambda} - \sqrt{\lambda'}\|_2^2 &\le \frac{1}{4}\int_0^\tau (\lambda(t) \vee \lambda'(t))\log^2\frac{\lambda(t)}{\lambda'(t)}\,dt.
\end{aligned}$$

Proof: The inequalities for h follow from the fact that $(1/4)(x \wedge 1) \le 1 - \exp(-x/2) \le x \wedge 1$ for $x \ge 0$.

For the Kullback-Leibler divergence we have

$$K(p_\lambda, p_{\lambda'}) = \int_0^\tau \lambda'(t) f(\lambda(t)/\lambda'(t))\,dt,$$

for $f(x) = x - 1 - \log x$. By Taylor's formula, $|f(x)|$ is bounded by a constant times $(\sqrt{x}-1)^2$ in a neighborhood of 1. Since $|f(x)|$ is bounded by $|x|$ for $x \ge 1$ and $|x|/(\sqrt{x}-1)^2 \to 1$ as $x \to \infty$, we have in fact $|f(x)| \le 3(\sqrt{x}-1)^2$

for all $x \in (1/e, \infty)$, say. For $(0, 1/e)$ we have $|f(x)| \leq |\log x|$. It follows that

$$K(p_\lambda, p_{\lambda'}) \leq 3 \int_{\lambda/\lambda' \geq 1/e} (\sqrt{\lambda(t)} - \sqrt{\lambda'(t)})^2 \, dt + \int_{\lambda/\lambda' \leq 1/e} \lambda'(t) \Big| \log \frac{\lambda(t)}{\lambda'(t)} \Big| \, dt.$$

The first term on the right is bounded by $3\|\sqrt{\lambda} - \sqrt{\lambda'}\|_2^2$. For the second term we note that for $\lambda/\lambda' \leq 1/e$, we have $\log|\lambda/\lambda'| \geq 1$ and hence $\log|\lambda/\lambda'| \leq \log^2|\lambda/\lambda'|$. The statement of the lemma follows.

To prove the last inequality, write $\|\sqrt{\lambda} - \sqrt{\lambda'}\|_2^2$ as the sum of an integral over the set $\{\lambda' \leq \lambda\}$ and an integral over the set $\{\lambda' > \lambda\}$ and use the fact that $1 - x \leq |\log x|$ for $x \in (0, 1]$. □

To connect assumptions (3.4)–(3.6) to the corresponding assumptions of Theorem 2.1 of [37] we first note that since λ_0 is bounded away from 0 and infinity by assumption, the same holds for any $\lambda \in \Lambda$ that is uniformly close enough to λ_0. Lemma 3.1 and the definition of V therefore imply that for λ uniformly close enough to λ_0, both $K(p_\lambda, p_{\lambda_0})$ and $V(p_\lambda, p_{\lambda_0})$ are bounded by a constant times the uniform norm $\|\lambda - \lambda_0\|_\infty$. It follows that for n large enough, the Kullback-Leibler-type ball

$$B(\varepsilon_n) = \{\lambda \in \Lambda : K(p_\lambda, p_{\lambda_0}) \leq \tilde{\varepsilon}_n^2, V(p_\lambda, p_{\lambda_0}) \leq \tilde{\varepsilon}_n^2\} \tag{3.13}$$

is larger than a multiple of the uniform ball $\{\lambda \in \Lambda : \|\lambda - \lambda_0\|_\infty \leq \tilde{\varepsilon}_n\}$. Lemma 3.1 also implies that the covering number $N(\bar{\varepsilon}_n, \{p_\lambda : \lambda \in \Lambda_n\}, h)$ is bounded by $N(\bar{\varepsilon}_n/\sqrt{2}, \sqrt{\Lambda_n}, \|\cdot\|_2)$. Hence, assumptions (3.4)–(3.6) imply that the conditions of Theorem 2.1 of [37] are fulfilled. This theorem states that for M large enough, $\mathbb{E}_{\lambda_0} \Pi_n(\lambda : h(p_\lambda, p_{\lambda_0}) \geq M\varepsilon_n) \to 0$. To complete the proof, note that by the fact that $M\varepsilon_n \leq 1$ for n large enough and the first inequality of the lemma, it holds, for n large enough, that $\|\sqrt{\lambda} - \sqrt{\lambda_0}\|_2 \geq \sqrt{2} M\varepsilon_n$ implies that $h(p_\lambda, p_{\lambda_0}) \geq M\varepsilon_n$.

3.4.2 Proof of Theorem 3.2

The proof is similar to the proof of Theorem 3.1, but this time we start from the observation that in the discrete-observations case, the data constitute a sample of n independent, identically distributed random vectors $C^{(1)}, \ldots, C^{(n)}$ in $\mathbb{R}^m$, where

$$C^{(i)} = (C_{i1}, \ldots, C_{im})$$

and C_{ij} is given by (3.1). The coordinates C_{ij} of $C^{(i)}$ are independent Poisson variables with mean λ_j given by (3.2).

Again we apply Theorem 2.1 of [37]. In this case the Hellinger distance h_m, Kullback-Leibler divergence K_m and variance measure V_m are easily seen to be given by

$$\begin{aligned}
h_m^2(\lambda, \lambda') &= 2(1 - e^{-\frac{1}{2}\sum\left(\sqrt{\lambda_j} - \sqrt{\lambda_j'}\right)^2}), \\
K_m(\lambda, \lambda') &= \sum(\lambda_j - \lambda_j') + \sum \lambda_j' \log \frac{\lambda_j'}{\lambda_j}, \\
V_m(\lambda, \lambda') &= \sum \lambda_j' \log^2 \frac{\lambda_j'}{\lambda_j},
\end{aligned}$$

respectively. These quantities satisfy the same bounds as in Lemma 3.1, but with the integrals replaced by the

corresponding sums. Moreover, by expanding the square and using Cauchy-Schwarz we see that

$$\sum \left(\sqrt{\lambda_j} - \sqrt{\lambda'_j}\right)^2 \leq \|\sqrt{\lambda} - \sqrt{\lambda'}\|_2^2,$$

and hence also

$$V_m(\lambda, \lambda') \leq 4 \sum \frac{\lambda'_j}{\lambda_j \wedge \lambda'_j} \left(\sqrt{\lambda_j} - \sqrt{\lambda'_j}\right)^2 \leq 4 \frac{\|\lambda'\|_\infty}{\inf_t |\lambda(t)| \wedge \inf_t |\lambda'(t)|} \|\sqrt{\lambda} - \sqrt{\lambda'}\|_2^2.$$

Using these relations the proof can be completed exactly as in Section 3.4.1.

3.4.3 Proof of Theorem 3.3

Let $\lambda_0 = \Psi(w_0)$ for some function $w_0 \in L^2[0, \tau]$ and $\lambda = \Psi(w)$ for w in the support of the process W. Note that for all $a, b \in \mathbb{R}$,

$$|\log \Psi(a) - \log \Psi(b)| = \left| \int_a^b \frac{\psi(t)}{\Psi(t)} \, dt \right| \leq \left\| \frac{\psi}{\Psi} \right\|_\infty |a - b|.$$

Combining this with the bounds provided by Lemma 3.1, we conclude

$$\|\sqrt{\lambda} - \sqrt{\lambda_0}\|_2^2 \leq \frac{1}{4} \|\Psi\|_\infty \left\| \frac{\psi}{\Psi} \right\|_\infty^2 \|w - w_0\|_2^2.$$

Using this we can connect (3.8)–(3.10) to the corresponding assumptions of Theorem 2.1 of [37] in the same way as we did for the previous two theorems and the result follows.

3.4.4 Proof of Theorem 3.4

Take $\lambda = M + A\Psi(w)$ for w in the support of the process W_n and $\lambda_0 = M + c\|\lambda_0 - M\|_\infty \Psi(w_0)$ for some constant c and function w_0 with M such that $\lambda_0 > M$. We consider first the case of continuous observations. First note that

$$\|\sqrt{\lambda} - \sqrt{\lambda_0}\|_2 \leq \|\sqrt{\lambda} - \sqrt{M + c\|\lambda_0 - M\|_\infty \Psi(w)}\|_2 + \|\sqrt{M + c\|\lambda_0 - M\|_\infty \Psi(w)} - \sqrt{\lambda_0}\|_2.$$

We have for all $a, b, u \geq 0$,

$$|\sqrt{M + au} - \sqrt{M + bu}| = \int_a^b \frac{u}{2\sqrt{M + tu}} \, dt \leq \frac{u}{2\sqrt{M}} |a - b|,$$

from which it follows that

$$\|\sqrt{\lambda} - \sqrt{M + c\|\lambda_0 - M\|_\infty \Psi(w)}\|_2 \leq \frac{\tau}{2\sqrt{M}} \|\Psi\|_\infty \left| A - c\|\lambda_0 - M\|_\infty \right|$$

Also, for $u \geq 0$ and $a, b \in \mathbb{R}$,

$$|\sqrt{M + u\Psi(a)} - \sqrt{M + u\Psi(b)}| = \int_a^b \frac{u\psi(t)}{2\sqrt{M + u\Psi(t)}} \, dt \leq \frac{\sqrt{u}}{2} \left\| \frac{\psi}{\sqrt{\Psi}} \right\|_\infty |a - b|,$$

implying

$$\|\sqrt{M + c\|\lambda_0 - M\|_\infty \Psi(w)} - \sqrt{\lambda_0}\|_2 \leq \frac{\sqrt{c\|\lambda_0 - M\|_\infty}}{2} \left\|\frac{\psi}{\Psi}\right\|_\infty \|w - w_0\|_2.$$

We also have

$$V(p_\lambda, p_{\lambda_0}) \leq (M + c\|\lambda_0 - M\|_\infty)\|\Psi\|_\infty \int_0^\tau \log^2 \frac{\lambda_0(t)}{\lambda(t)}\, dt.$$

Proceeding in the same way as before, for $a, b, u \geq 0$,

$$|\log(M + au) - \log(M + bu)| = \int_a^b \frac{u}{M + tu}\, dt \leq \frac{u}{M}|a - b|,$$

and for $u \geq 0$, $a, b \in \mathbb{R}$,

$$|\log(M + u\Psi(a)) - \log(M + u\Psi(b))| = \int_a^b \frac{u\psi(t)}{M + u\Psi(t)}\, dt \leq \left\|\frac{\psi}{\Psi}\right\|_\infty |a - b|,$$

which give, respectively,

$$\int_0^\tau \log^2 \frac{\lambda(t)}{M + c\|\lambda_0 - M\|_\infty \Psi\big(w(t)\big)}\, dt \leq \frac{\tau}{M}^2 \|\Psi\|_\infty^2 \big|A - c\|\lambda_0 - M\|_\infty\big|^2,$$

and

$$\int_0^\tau \log^2 \frac{\lambda_0(t)}{M + c\|\lambda_0 - M\|_\infty \Psi\big(w(t)\big)}\, dt \leq \left\|\frac{\psi}{\Psi}\right\|_\infty^2 \|w - w_0\|_2^2.$$

We conclude that for d either h^2, K or V,

$$d(\lambda, \lambda_0) \leq C^2\Big(\big|A - c\|\lambda_0 - M\|_\infty\big|^2 + \|w - w_0\|_2^2\Big),$$

for a constant C depending on M, τ, the range of λ_0 and on the link function Ψ. The same computations can be carried out in the case of discrete observations and the statement of the previous display will hold for another constant C depending on the same quantities.

We verify now the conditions of Theorem 2.1 of [37]. Let $\varepsilon > 0$ and recall that $B(\varepsilon)$ is defined as (3.13). By construction it holds that $\Pi_n(B(\varepsilon))$ is bounded from below by

$$\mathbb{P}\big(K(\lambda, \lambda_0) \leq \varepsilon^2, V(\lambda, \lambda_0) \leq \varepsilon^2, A > a\big)$$

for $\lambda = M + A\Psi(W_n)$, $\lambda_0 = M + c\|\lambda_0 - M\|_\infty \Psi(w_0)$ and every $a > 0$. The bound derived above implies that there exists a constant $C > 0$ such that for $a = c\|\lambda_0 - M\|_\infty$, this is further bounded from below by

$$\mathbb{P}\big(|A - c\|\lambda_0 - M\|_\infty|^2 + \|W_n - w_0\|_2^2 \leq C^2\varepsilon^2, A \geq c\|\lambda_0 - M\|_\infty\big).$$

By the triangle inequality and independence this is lower bounded by

$$\mathbb{P}(|A - c\|\lambda_0 - M\|_\infty| \leq C\varepsilon, A \geq c\|\lambda_0 - M\|_\infty)\mathbb{P}(\|W_n - w_0\|_2 \leq C\varepsilon).$$

The first factor in the display is bounded from below by a constant times ε and the last one is lower bounded

by assumption. Hence, by the prior mass assumption of the theorem, we have that the prior mass condition is fulfilled for certain constants $c_1, c_2 > 0$ as in the statement of the theorem.

To verify condition remaining mass condition is suffices to show that for some given constant $c_3 > 1$, there exist sets Λ_n such that $\Pi_n(\Lambda_n^c) \leq 2\exp(-c_3 n\tilde{\varepsilon}_n^2)$. By the assumptions, there exists a set B_n such that $\mathbb{P}(W_n \notin B_n) \leq \exp(-c_3 n\tilde{\varepsilon}_n^2)$ and a sequence $a_n = a_n(c_3) \to \infty$ such that $\mathbb{P}(A > a_n) \leq \exp(-c_3 n\tilde{\varepsilon}_n^2)$ as well. Then for the sieves $\Lambda_n = \{a\Psi(w) : a \leq a_n, w \in B_n\}$ we have

$$\Pi_n(\Lambda_n^c) \leq \mathbb{P}(A > a_n) + \mathbb{P}(W_n \notin B_n) \leq 2e^{-c_3 n\tilde{\varepsilon}_n^2},$$

as required.

By the bound on the Hellinger distance derived above and the entropy assumption of the theorem we have, for a constant $L > 0$,

$$\begin{aligned}\log N(L\bar{\varepsilon}_n, \Lambda_n, h) &\leq \log N(\bar{\varepsilon}_n, [0, a_n], |\cdot|) + \log N(\bar{\varepsilon}_n, B_n, \|\cdot\|_2) \\ &\lesssim \log \frac{a_n}{\bar{\varepsilon}_n} + C_2 n\bar{\varepsilon}_n^2.\end{aligned}$$

The right-hand side is bounded by a multiple of $n\bar{\varepsilon}_n^2$ hence verifying the entropy condition.

3.4.5 Proof of Theorem 3.5

Under the prior Π, the number of knots J has, by construction, a shifted Poisson distribution. By Stirling's approximation, this implies that for large j,

$$\mathbb{P}(J > j) \lesssim e^{-c_1 j\log j}, \qquad \mathbb{P}(J = j) \gtrsim e^{-c_2 j\log j}$$

for some $c_1, c_2 > 0$. For the sequence of inner knots $\mathbf{k}$ constructed in the definition of the prior the mesh size $M(\mathbf{k}) = \max\{|k_j - k_{j-1}|\}$ and the sparsity $m(\mathbf{k}) = \min\{|k_j - k_{j-1}|\}$ satisfy

$$\mathbb{P}(m(\mathbf{k}) < j^{-2} \mid J = j) = 0, \qquad \mathbb{P}(M(\mathbf{k}) \leq 2/j \mid J = j) \gtrsim e^{-j\log j}.$$

The first of these facts follows trivially from the construction, the second one by bounding the probability of interest from below by the probability that every of the consecutive intervals of length τ/j contains at least one knot. For the B-spline coefficients, by independence,

$$\mathbb{P}(\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_\infty \leq \varepsilon \mid J = j) \gtrsim \varepsilon^j$$

for all $\boldsymbol{\theta}_0 \in [M_1, M_2]^j$. Theorem 1 of [7] deals exactly with this situation. In the present setting the theorem asserts that if $\lambda_0 \in \mathcal{H}_\alpha([0,\tau])$ and $M_1 \leq \lambda_0 \leq M_2$, then for $J_n, \bar{J}_n > q$ and positive $\varepsilon_n \geq \bar{\varepsilon}_n$ such that $\varepsilon_n \to 0$, $n\bar{\varepsilon}_n^2 \to \infty$ and

$$2\Big(\frac{\bar{\varepsilon}_n}{\|\lambda_0\|_{C^\alpha}}\Big)^{-1/\alpha} \leq \bar{J}_n, \quad \log \bar{J}_n \lesssim \log \frac{1}{\bar{\varepsilon}_n}, \quad J_n \log \frac{J_n^3}{\varepsilon_n} \lesssim n\varepsilon_n^2, \quad n\bar{\varepsilon}_n^2 \leq J_n \log J_n, \tag{3.14}$$

then there exist function spaces (of splines) Λ_n and a constant $c > 0$ such that

$$\Pi(\lambda : \|\lambda - \lambda_0\|_\infty \leq 2\bar{\varepsilon}_n) \gtrsim e^{-c\bar{J}_n \log \frac{1}{\bar{\varepsilon}_n}}, \tag{3.15}$$

$$\Pi(\lambda \notin \Lambda_n) \lesssim e^{-c_1 n\bar{\varepsilon}_n^2}, \tag{3.16}$$

$$\log N(\varepsilon_n, \Lambda_n, \|\cdot\|_2) \lesssim n\varepsilon_n^2. \tag{3.17}$$

Now observe that the first two inequalities in (3.14) hold for

$$\bar{\varepsilon}_n = n^{-\frac{\alpha}{1+2\alpha}} \log^p n, \quad \bar{J}_n = Kn^{\frac{1}{1+2\alpha}} \log^q n,$$

provided K is large enough and $q \geq -p/\alpha$. The third and fourth inequalities then hold for

$$J_n = Ln^{\frac{1}{1+2\alpha}} \log^r n, \quad \varepsilon_n = n^{-\frac{\alpha}{1+2\alpha}} \log^s n$$

if L is large enough and $2p \leq r + 1 \leq 2s$. To complete the proof we have to link (3.15)–(3.17) to the conditions (3.4)–(3.6) of Theorems 3.1 and 3.2. Note that since (3.5) should hold for all $c_3 > 0$, we need to have

$$\bar{J}_n \log \frac{1}{\bar{\varepsilon}_n} \ll n\bar{\varepsilon}_n^2.$$

For our choices of $\bar{J}_n$ and $\bar{\varepsilon}_n$ this holds if $2p > q + 1$. This amounts to choosing $p > \alpha/(1 + 2\alpha)$. Then if we define

$$\tilde{\varepsilon}_n = \sqrt{\frac{\bar{J}_n}{n} \log \frac{1}{\bar{\varepsilon}_n}}$$

the right-hand side of (3.15) equals $\exp(-2n\tilde{\varepsilon}_n^2)$. Moreover, it holds that $\tilde{\varepsilon}_n \sim n^{-\alpha/(1+2\alpha)} (\log n)^{(q+1)/2}$, so if we make sure that $p > (q + 1)/2$, the desired inequality (3.4) holds. The considerations above imply that (3.5) then holds as well, for any $c_3 \geq 1$. Recall that we found that the entropy condition holds for $\varepsilon_n \sim n^{-\alpha/(1+2\alpha)} (\log n)^s$, provided $s > p$. This means that we should choose p, q, r and s above such that

$$p > \frac{\alpha}{1 + 2\alpha}, \quad r = 2p - 1, \quad s > p, \quad q = -\frac{1}{1 + 2\alpha}.$$

Since the intensities in Λ_n are uniformly bounded by a common constant (see the proof of Theorem 1 of [7]), (3.17) implies that (3.6) is fulfilled.

3.4.6 Proof of Theorem 3.6

We again verify the conditions of Theorem 2.1 of [37].

Note that by the triangle inequality and the fact that λ_0 is increasing,

$$\|A + BD - \lambda_0\|_\infty \leq |A - \lambda_0(0)| + B\left\|D - \frac{\lambda_0 - \lambda_0(0)}{\lambda_0(\tau) - \lambda_0(0)}\right\|_\infty + |B - (\lambda_0(\tau) - \lambda_0(0))|.$$

Hence, by independence, there exists a constant $C_0 > 0$ only depending on λ_0 such that for $\varepsilon \in (0,1)$, with

$\nu_0 = (\lambda_0 - \lambda_0(0))/(\lambda_0(\tau) - \lambda_0(0))$,

$$\begin{aligned}&\Pi(\lambda : \|\lambda - \lambda_0\|_\infty \le C_0\varepsilon)\\ &\quad \ge \mathbb{P}(|A - \lambda_0(0)| \le \varepsilon)\mathbb{P}(|B - (\lambda_0(\tau) - \lambda_0(0))| \le \varepsilon)\mathbb{P}(\|D - \nu_0\|_\infty \le 3\varepsilon).\end{aligned}$$

The first two factors on the right are bounded from below by a constant times ε. To deal with the third one we argue as in Example 6.1 of [38]. Let $\varepsilon > 0$ be fixed for now and let $0 = t_0 < t_1 < \cdots < t_N = \tau$ be points such that for $I_i = (t_{i-1}, t_i]$, $\nu_0(I_i) \le \varepsilon$. Since ν_0 has total mass one, these points can be chosen such that $N \lesssim 1/\varepsilon$. Now it can be verified that $\|D - \nu_0\|_\infty \le \sum_{j\le N} |D(I_j) - \nu_0(I_j)| + \varepsilon$ and hence

$$\mathbb{P}(\|D - \nu_0\|_\infty \le 3\varepsilon) \ge \mathbb{P}\Big(\sum_{j\le N} |D(I_j) - \nu_0(I_j)| \le 2\varepsilon\Big).$$

Lemma 6.1 of [38] implies that for ε small enough, the probability on the right is bounded from below by a constant times $\exp(-c(1/\varepsilon)\log(1/\varepsilon))$ for some $c > 0$. Putting things together we find that for $\varepsilon > 0$ small enough and constants $c_0, C > 0$,

$$\Pi(\lambda : \|\lambda - \lambda_0\|_\infty \le C_0\varepsilon) \gtrsim \varepsilon^2 e^{-c\frac{1}{\varepsilon}\log\frac{1}{\varepsilon}} \gtrsim e^{-c'\frac{1}{\varepsilon}\log\frac{1}{\varepsilon}}.$$

By using the inequalities derived in Lemma 3.1 for continuous data and on Section 3.4.2 for the discrete setup, for all sufficiently small $\varepsilon > 0$,

$$\{\lambda : K(\lambda, \lambda_0) \le C\varepsilon^2, V(\lambda, \lambda_0) \le C\varepsilon^2\} \supset \{\lambda : \|\lambda - \lambda_0\|_\infty \le C_0\varepsilon\},$$

for some $C > 0$. Combining this with the previous statement it follows that the prior mass condition is fulfilled for $\tilde{\varepsilon}_n$ a multiple of $(n/\log n)^{-1/3}$.

Next we define sieves $\Lambda_n = \{f : [0,1] \to [0, 2L_n] \mid f \text{ increasing}\}$, for L_n a sequence of positive numbers further determined below. Since the random function $A + BD$ is an increasing function on $[0,1]$ which takes values in $[0, A + B]$,

$$\Pi(\Lambda_n^c) \le \mathbb{P}(A > L_n) + \mathbb{P}(B > L_n).$$

By the assumption on the tails of A and B, this is bounded by a constant times $\exp(-c\, n^{1/3}\log n)$ for $L_n \sim \log^2 n$, which in turn is bounded by to be bounded by $e^{-C_1 n\tilde{\varepsilon}_n^2}$ for any constant $C_1 > 0$. Hence, the remaining mass condition is satisfied.

We use now the bounds on the Hellinger metric obtained in Lemma 3.1 for continuous data and in Section 3.4.2 for discrete data. For the entropy condition we note that the functions in $\sqrt{\Lambda_n}$ are increasing and take values in $[0, \sqrt{2L_n}]$. Hence, by Theorem 2.7.5 of [98], we have the entropy bound

$$\log N(\varepsilon, \Lambda_n, h) \lesssim \log N(\varepsilon, \sqrt{\Lambda_n}, \|\cdot\|_2) \le \log N_{[\,]}(\varepsilon, \sqrt{\Lambda_n}, \|\cdot\|_2) \lesssim \frac{\sqrt{L_n}}{\varepsilon}.$$

This shows that the entropy condition is satisfied for $\tilde{\varepsilon}_n = (n/\sqrt{L_n})^{-1/3} \sim (n/\log n)^{-1/3}$. We conclude that we have the posterior rate $(n/\log n)^{-1/3}$ relative to the Hellinger distance.

4

Adaptive Priors based on Splines with Random Knots

Splines are useful building blocks when constructing priors on nonparametric models indexed by functions. Recently, it has been established in the literature that hierarchical priors based on splines with a random number of equally spaced knots and random coefficients in the corresponding B-spline basis deliver, under certain conditions, adaptive posterior contraction rates, over certain smoothness functional classes. In this chapter we extend these results for when the location of the knots is also endowed with a prior. This has already been a common practice in MCMC applications, but a theoretical basis in terms of adaptive contraction rates was missing. Under some mild assumptions, we establish a result that provides sufficient conditions for adaptive contraction rates in a range of models, over certain functional classes of smoothness, up to the order of the splines that are used. We also present some numerical results illustrating how such a prior adapts to inhomogeneous variability (smoothness) of the function in the context of nonparametric regression.

4.1 Introduction

The Bayesian approach in statistics has become quite popular in recent years as an alternative to classical *frequentist* methods. The main appeal of the Bayesian methodology is its conceptual simplicity: given a model for the observed data $X \sim P_f$, $f \in \mathcal{F}$, some space of functions, put a prior on the unknown parameter f and draw inferences based on the resulting posterior $\Pi(f|X)$. Knowledge about the model under study can also be incorporated into the inference procedure via the prior. However, some seemingly "correct" priors can lead to unreasonable posteriors, especially in nonparametric models. It is therefore desirable to place ourselves in a setting where it is possible to assess the quality of the resulting posterior from some objective point of view. This gave rise to the development of the notion of contraction rate (cf. [38]), a Bayesian analog of a convergence rate: data is assumed to come from a fixed probability measure $P_0 = P_{f_0}$ for a "true" $f_0 \in \mathcal{F}$; the contraction rate is then the smallest radius such that the posterior mass in balls (with respect to an appropriate distance) of probability measures around P_0 converges to 1 in P_0-probability as some information index such as a sample size goes to infinity.

Some general results about posterior contraction rates establish sufficient conditions on prior distributions such that the resulting posteriors attain a certain contraction rate. In this spirit, when studying specific priors, some authors now choose to present their results in the form of say *meta-theorems* which claim that sufficient conditions (such as the ones in [38]) required to attain a certain range of contraction rates hold for their choice of prior; cf. [25, 86, 95] and further references therein. We adopt this practice here as well.

In the case where f_0 is a function from some functional space of smoothness α, the posterior contraction rate is typically compared to the convergence rate of the minimax risk (called optimal rate) over that space in the estimation problem. For example, if we observe a sample of size n and want to estimate a univariate α-smooth function (e.g., density or regression function), the typical optimal rate is of order $n^{-\alpha/(2\alpha+1)}$, possibly up to a logarithmic factor depending on the risk function. If the smoothness parameter α is unknown, and one wants to build estimators which attain the optimal rate corresponding to α but do not depend explicitly on α, then one speaks of an adaptation problem. In a Bayesian context, the *adaptation problem* consists in finding a prior which leads to the optimal posterior contraction rate (usually up to a logarithmic factor) for any α-smooth function of interest and does not depend on the smoothness parameter α. Such priors are called *rate adaptive*. There is a growing number of papers, where this problem has been studied in different settings; cf. [6, 25, 86, 95, 97] among others.

Splines, in particular, can be used when constructing adaptive priors. A spline (cf. [23]) is a piecewise polynomial function designed to have a certain level of smoothness which is referred to as its order. Splines are easy to store, differentiate, integrate and evaluate on a computer, and are extensively used in practice for constructing good, parsimonious approximations of smooth functions. The points at which the different polynomial pieces of a spline connect are called knots. If an order (read: maximal polynomial degree) and a set of knots is fixed, then the space of all splines with that order and those knots forms a linear space which admits a basis of so-called B-splines. Any spline of a fixed order is consequently characterized by a set of knots and its coordinates in the B-splines basis corresponding to those knots. Randomly generating a number of knots and, given those, generating random coordinates in the corresponding B-spline basis with equally spaced knots results in a random spline whose law can be used as a prior. If, given the number of knots, the coordinates in the corresponding B-spline basis are chosen to be independent and normally distributed, then the resulting spline has a conditionally Gaussian law and was studied by [25] by using Reproducing Kernel Hilbert Space techniques. In [86] a more general, random

series prior was proposed: the coefficients in the series are not necessarily independent or Gaussian and a basis other than the B-spline basis can also be used.

The case where the locations of the knots are also random is not covered by the results of either [25] or [86]. However when practitioners put a prior on the number of knots they almost invariably also put a prior on the locations of the knots (e.g., [27, 30, 85]) – a Poisson process is a popular choice. Their motivation for allowing arbitrarily located knots seems to be twofold. Firstly, this is attractive from the implementation point of view: designing reversible jump MCMC samplers is much simpler if any collection of knots is allowed since new knots can be inserted at arbitrary positions causing only localized changes in the spline. Secondly, the resulting posterior based on the prior with random locations of the knots is expected to be more adaptive with respect to inhomogeneous smoothness of the function of interest: the function may not have a fixed level of smoothness throughout its support, it may consist of rough and smooth pieces. To sustain an adequate level of accuracy over the whole support, more knots are needed in rough pieces and less in smooth ones. Therefore, to make it at least possible for the resulting posterior to pick up eventual spatial features of the function, the prior has to be flexible enough to model random locations of the knots.

In this chapter, we extend the results of [25], and those of [86] with respect to the prior with random knots: we add one more level to the hierarchical spline prior by putting a prior on the location of the knots of the spline as well, making, in fact, the basis functions also random. Under some mild assumptions on the proposed hierarchical spline prior, we establish our main result for the proposed prior, providing sufficient conditions for adaptive, optimal contraction rates of the resulting posterior in a range of models (among others: density estimation, nonparametric regression, binary regression, Poisson regression, and classification). In doing so, we provide a theoretical basis for the common practice of using randomly located knots in spline based priors. Another interesting feature of a prior with random knots locations is that it leads to the posterior of the knots vector which provides (some sort of empirical Bayes) inference on the variability (smoothness inhomogeneity) of the underlying function. We present some numerical results illustrating how such a prior adapts to inhomogeneous variability (smoothness) of the function in the context of nonparametric regression.

4.2 Notation and preliminaries on splines

First we introduce some notation. For $d \in \mathbb{N}$ and $1 \leq p < \infty$ denote by $\|\mathbf{x}\|_p = \left(\sum_{i=1}^d |x_i|^p\right)^{1/p}$ the l_p-norm of $\mathbf{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d$ and by $\|\mathbf{x}\|_\infty = \max_{i=1,\ldots,d} |x_i|$. For $1 \leq p < \infty$ let the L_p-norm of a function f on [0,1] be $\|f\|_p = \left(\int_0^1 |f(x)|^p \, dx\right)^{1/p}$ and $\|f\|_\infty = \sup_{x\in[0,1]} |f(x)|$.

We use $\lesssim$ (respectively $\gtrsim$) to denote smaller (respectively greater) or equal up to a constant, the symbols $a \vee b$ and $a \wedge b$ stand for $\max\{a, b\}$ and $\min\{a, b\}$ respectively. The covering number $N(\varepsilon, S, d)$ of a subset S of a metric space with balls of size ε is the smallest number of balls (with respect to distance d) of radius ε needed to cover S.

Now we provide some preliminaries on splines, which can be found, for example, in [81]. A function is called a spline is of order $q \in \mathbb{N}$, with respect to a certain partition of its support, if it is $q - 2$ times continuously differentiable and when restricted to each interval in this partition, coincides with a polynomial of degree at most $q - 1$. Consider $q \in \mathbb{N}$, $q \geq 2$, which will be fixed throughout the remainder of this text. For any $j \in \mathbb{N}$, such that $j \geq q$ let $\mathcal{K}_j = \{(k_1, \ldots, k_{j-q}) \in (0,1)^{j-q} : 0 < k_1 < \cdots < k_{j-q} < 1\}$. We will refer to a vector $\mathbf{k} = \mathbf{k}_j \in \mathcal{K}_j$ as a set of inner knots; the index j in $\mathbf{k}_j$ will sometimes be used to emphasize the dependence on j. A vector $\mathbf{k} \in \mathcal{K}_j$ will be

said to induce the partition $\{[k_0,k_1),[k_1,k_2),\dots,[k_{j-q},k_{j-q+1}]\}$, with $k_0=0$ and $k_{j-q+1}=1$. For any $\mathbf{k}\in\mathcal{K}_j$ we call $M(\mathbf{k})=\max_{i=1}^{j-q+1}|k_i-k_{i-1}|$ the mesh size of the partition induced by $\mathbf{k}$ and $m(\mathbf{k})=\min_{i=1}^{j-q+1}|k_i-k_{i-1}|$ the sparseness of the partition induced by $\mathbf{k}$. For a $\mathbf{k}\in\mathcal{K}_j$, denote by $\mathcal{S}_\mathbf{k}=\mathcal{S}_\mathbf{k}^q$ the linear space of splines of order q on $[0,1]$ with simple knots $\mathbf{k}$ (see the definition of knot multiplicity in [81]). This space has dimension j and admits a basis of so-called B-splines $\{B_{\mathbf{k},1},\dots,B_{\mathbf{k},j}\}=\{B^q_{\mathbf{k},1},\dots,B^q_{\mathbf{k},j}\}$. The construction of $\{B_{\mathbf{k},1},\dots,B_{\mathbf{k},j}\}$ involves the knots $k_{-q+1},\dots,k_{-1},k_0,k_1,\dots,k_{j-q},k_{j-q+1},k_{j-q+2},\dots,k_j$, with arbitrary extra knots $k_{-q+1}\le\cdots\le k_{-1}\le k_0=0$ and $1=k_{j-q+1}\le k_{j-q+2}\le\cdots\le k_j$. Usually one takes $k_{-q+1}=\cdots=k_{-1}=k_0=0$ and $1=k_{j-q+1}=\cdots=k_j$, and we adopt this choice here as well. These basis functions are nonnegative: $B_{\mathbf{k},i}(x)\ge 0$, for all $x\in[0,1]$. Besides, they have local support and form a partition of unity:

$$B_{\mathbf{k},i}(x)=0 \ \text{ for } x\notin[k_{-q+i},k_i], \qquad \sum_{i=1}^{j}B_{\mathbf{k},i}(x)=1 \ \text{ for all } x\in[0,1]. \tag{4.1}$$

To refer explicitly to the coordinates $\mathbf{a}=(a_1,\dots,a_j)\in\mathbb{R}^j$ of a spline in a specific B-spline basis with inner knots $\mathbf{k}$, we write $s_{\mathbf{k},\mathbf{a}}(x)=\sum_{i=1}^j a_iB_{\mathbf{k},i}(x)$, $x\in[0,1]$. Since $\sum_{i=1}^j B_{\mathbf{k},i}(x)=1$, it is easy to see that for any $s_{\mathbf{k},\mathbf{a}},s_{\mathbf{k},\mathbf{b}}\in\mathcal{S}_\mathbf{k}^q$

$$\|s_{\mathbf{k},\mathbf{a}}-s_{\mathbf{k},\mathbf{b}}\|_2\le\|s_{\mathbf{k},\mathbf{a}}-s_{\mathbf{k},\mathbf{b}}\|_\infty\le\|\mathbf{a}-\mathbf{b}\|_\infty\le\|\mathbf{a}-\mathbf{b}\|_2. \tag{4.2}$$

Splines have good approximation properties for sufficiently smooth functions provided they are defined on a partition with appropriately small mesh size. We say that a function f on $[0,1]$ belongs to a generic smoothness class $\mathcal{F}_\alpha$, $\alpha>0$, if f is Lipschitz, i.e., $f\in\mathcal{L}_{\kappa_\alpha}(L_\alpha,[0,1])=\{f:|f(x_1)-f(x_2)|\le L_\alpha|x_1-x_2|^{\kappa_\alpha},x_1,x_2\in[0,1]\}$ for some $\kappa_\alpha,L_\alpha>0$, and for any set of inner knots $\mathbf{k}$ there exists a spline $s_{\mathbf{k},\mathbf{a}}\in\mathcal{S}_\mathbf{k}^q$ such that for some bounded C_f

$$\|f-s_{\mathbf{k},\mathbf{a}}\|_\infty\le C_fM^\alpha(\mathbf{k}). \tag{4.3}$$

A leading example of a smoothness class $\mathcal{F}_\alpha$ is the Hölder space $\mathcal{H}_\alpha=\mathcal{H}_\alpha(L,[0,1])$, $0<\alpha\le q$, which is the collection of all functions f that have bounded derivatives up to order $\alpha_0=\lfloor\alpha\rfloor=\max\{z\in\mathbb{Z}:z<\alpha\}$ and such that the α_0-th derivative satisfies the Hölder condition $|f^{(\alpha_0)}(x)-f^{(\alpha_0)}(y)|\le L|x-y|^{\alpha-\alpha_0}$, for $L>0$ and $x,y\in[0,1]$. In this case, a well-known spline approximation result (cf. [23]) states that (4.3) holds with $C_f=C_q\|f^{(\alpha)}\|_\infty$ for some constant C_q depending only on q. Other examples of smoothness classes for which the approximation property (4.3) hold, include α-times continuously differentiable functions, Sobolev and Besov spaces; cf. Theorems 6.21, 6.25 and 6.31 in [81].

4.3 Main Result

We begin by describing a hierarchical prior on $\mathcal{S}=\mathcal{S}^q=\cup_{j=q}^\infty\cup_{\mathbf{k}\in\mathcal{K}_j}\mathcal{S}_\mathbf{k}^q$: first draw a number $J\in\mathbb{N}$, $J\ge q$; then, given J, generate independently $(J-q)$ inner knots $\mathbf{K}_J\in\mathcal{K}_j$ and also independently, J B-spline coefficients $\boldsymbol{\theta}\in\mathbb{R}^J$. Our prior on $\mathcal{S}$ will be the law of the random spline $s_{\mathbf{K}_J,\boldsymbol{\theta}}$. We impose the following conditions on this prior. For $c_1,c_2>0$, $0\le t_1,t_2\le 1$ and all sufficiently large j,

$$\mathbb{P}(J>j)\lesssim\exp\big(-c_1j\log^{t_1}j\big), \tag{4.4}$$

$$\mathbb{P}(J=j)\gtrsim\exp\big(-c_2j\log^{t_2}j\big). \tag{4.5}$$

For some $\tau \geq 1$, $c_3 > 0$, $0 \leq t_3 \leq 1$, and all $j \geq q$,

$$\mathbb{P}\big(m(\mathbf{K}_j) < \delta(j) | J = j\big) = 0, \tag{4.6}$$

$$\mathbb{P}\big(M(\mathbf{K}_j) \leq \tau / j | J = j\big) \gtrsim \exp\big(-c_3 j \log^{t_3} j\big), \tag{4.7}$$

where $\delta(i)$ is a positive, strictly decreasing function on $\mathbb{N}$. Without loss of generality assume that $\delta(i) \leq 1$, $i \in \mathbb{N}$. For each $j \geq q$, the conditional distribution of $\boldsymbol{\theta} \in \mathbb{R}^j$ satisfies the following condition: for any $M > 0$ there exists $c_0 = c_0(M)$ such that

$$\mathbb{P}\big(\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_\infty \leq \varepsilon | J = j\big) \gtrsim \exp\big(-c_0 j \log(1/\varepsilon)\big) \tag{4.8}$$

for all $\varepsilon > 0$ and all $\boldsymbol{\theta}_0 \in \mathbb{R}^j$ such that $\|\boldsymbol{\theta}_0\|_\infty \leq M$.

For examples of particular choices on the components of our hierarchical prior which verify these conditions we refer the reader to Section 4.5.

Denote $\mathcal{C}^j(M) = [-M, M]^j$. The following theorem is our main result.

Theorem 4.1 (Prior bounds)
Let $\|f_0\|_\infty < M$ and $f_0 \in \mathcal{F}_\alpha$ so that (4.3) holds with C_{f_0}. Let $\varepsilon_n, \bar{\varepsilon}_n$ be two positive sequences such that $\varepsilon_n \geq \bar{\varepsilon}_n$, $\varepsilon_n \to 0$ as $n \to \infty$ and $n\bar{\varepsilon}_n^2 > 1$. Assume that there exist sequences $J_n, \bar{J}_n > q$, $M_n \geq 1$ and a constant $c_M \geq c_1$ satisfying:

$$J_n \log\Big[\frac{J_n M_n}{\varepsilon_n \delta(J_n)}\Big] \lesssim n\varepsilon_n^2, \tag{4.9}$$

$$n\bar{\varepsilon}_n^2 \leq J_n \log^{t_1} J_n, \quad \mathbb{P}\big(\boldsymbol{\theta} \notin \mathcal{C}^j(M_n) | J = j\big) \lesssim \exp(-c_M n\bar{\varepsilon}_n^2), \quad q \leq j \leq J_n, \tag{4.10}$$

$$\Big[\frac{\bar{\varepsilon}_n}{\tau^\alpha C_{f_0}}\Big]^{-1/\alpha} \leq \bar{J}_n, \quad \log^{t_2 \vee t_3} \bar{J}_n \lesssim \log \frac{1}{\bar{\varepsilon}_n}. \tag{4.11}$$

Let $\mathcal{S}_n = \bigcup_{j=q}^{J_n} \bigcup_{\mathbf{k} \in \mathcal{K}_j^{\delta(j)}} \big\{s_{\mathbf{k},\boldsymbol{\theta}} \in \mathcal{S}_{\mathbf{k}}^q : \|\boldsymbol{\theta}\|_\infty \leq M_n\big\}$, where $\mathcal{K}_j^\delta = \{\mathbf{k} \in \mathcal{K}_j : m(\mathbf{k}) \geq \delta\}$. Then it holds that

$$\log N(\varepsilon_n, \mathcal{S}_n, \|\cdot\|_2) \lesssim n\varepsilon_n^2, \tag{4.12}$$

$$\mathbb{P}\big(s_{\mathbf{K}_J,\boldsymbol{\theta}} \notin \mathcal{S}_n\big) \lesssim \exp\big\{-c_1 n\bar{\varepsilon}_n^2\big\}, \tag{4.13}$$

$$\mathbb{P}\big(\|s_{\mathbf{K}_J,\boldsymbol{\theta}} - f_0\|_\infty \leq 2\bar{\varepsilon}_n\big) \gtrsim \exp\big\{-(c_0(M) + c_2 + c_3)\bar{J}_n \log(1/\bar{\varepsilon}_n)\big\}. \tag{4.14}$$

Proof: First we establish (4.12). Let $L_n(j) = 4M_n j(q+1)(\delta(j))^{-(q+1)}$ and $j > q$. Let $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{m_1}\}$ be an $\varepsilon_n/2$-net of the set $\{\boldsymbol{\theta} \in \mathbb{R}^j : \|\boldsymbol{\theta}\|_\infty \leq M_n\}$ and let $\{\mathbf{x}_1, \ldots, \mathbf{x}_{m_2}\}$ be an $\varepsilon_n/(2L_n(j))$-net of $\{\mathbf{x} \in \mathbb{R}^{j-q} : \mathbf{x} \in (0,1)^{j-q}\}$, both with respect to the $\|\cdot\|_\infty$-norm. Then, by using (4.2) and Lemma 4.2 (Lemma 4.2 is applicable since $\varepsilon_n/(2L_n(j)) \leq 2/(q-1)$ for sufficiently large n), $\{s_{\mathbf{x}_l,\boldsymbol{\theta}_k}, k = 1, \ldots, m_1,\ l = 1, \ldots m_2\}$ forms an ε_n-net of $\bigcup_{\mathbf{k} \in \mathcal{K}_j^{\delta(j)}} \big\{s_{\mathbf{k},\boldsymbol{\theta}} \in \mathcal{S}_{\mathbf{k}}^q :$

$\|\boldsymbol{\theta}\|_\infty \le M_n\}$ with respect to the $\|\cdot\|_\infty$-norm. By using this fact, we obtain for sufficiently large n that

$$N\big(\varepsilon_n, \mathcal{S}_n, \|\cdot\|_2\big) \le N\big(\varepsilon_n, \mathcal{S}_n, \|\cdot\|_\infty\big) \le \sum_{j=q}^{J_n} N\Big(\varepsilon_n, \cup_{\mathbf{k}\in\mathcal{K}_j^{\delta(j)}}\big\{s_{\mathbf{k},\boldsymbol{\theta}} \in \mathcal{S}_{\mathbf{k}}^q : \|\boldsymbol{\theta}\|_\infty \le M_n\big\}, \|\cdot\|_\infty\Big)$$
$$\le \sum_{j=q}^{J_n}\left[N\Big(\frac{\varepsilon_n}{2}, \big\{\boldsymbol{\theta}\in\mathbb{R}^j : \|\boldsymbol{\theta}\|_\infty \le M_n\big\}, \|\cdot\|_\infty\Big)\, N\Big(\frac{\varepsilon_n}{2L_n(j)}, (0,1)^{j-q}, \|\cdot\|_\infty\Big)\right]$$
$$\le J_n\left(\frac{2M_n}{\varepsilon_n}\right)^{J_n}\left(\frac{2L_n(J_n)}{\varepsilon_n}\right)^{J_n-q} \le J_n\left(\frac{16(q+1)M_n^2 J_n}{\varepsilon_n^2(\delta(J_n))^{q+1}}\right)^{J_n}.$$

The last relation and (4.9) imply (4.12):

$$\log N\big(\varepsilon_n, \mathcal{S}_n, \|\cdot\|_2\big) \lesssim J_n \log\left(\frac{J_n M_n}{\varepsilon_n \delta(J_n)}\right) \lesssim n\varepsilon_n^2.$$

Now we check (4.13). From the definition of $\mathcal{S}_n$, the relations (4.4), (4.6) and (4.10), it follows that

$$\mathbb{P}\big(s_{\mathbf{K}_J,\boldsymbol{\theta}} \notin \mathcal{S}_n\big) \le \mathbb{P}\Big(\{J > J_n\} \cup \big[\{q \le J \le J_n\} \cap \big(\{m(\mathbf{K}_j) < \delta(j)\} \cup \{\boldsymbol{\theta} \notin \mathcal{C}^j(M_n)\}\big)\big]\Big)$$
$$\le \mathbb{P}\big(J > J_n\big) + \sum_{j=q}^{J_n}\mathbb{P}\big(J=j\big)\Big(\mathbb{P}\big(m(\mathbf{K}_j) < \delta(j) | J = j\big) + \mathbb{P}\big(\boldsymbol{\theta} \notin \mathcal{C}^j(M_n) | J = j\big)\Big)$$
$$\lesssim \exp\big\{-c_1 J_n \log^{t_1} J_n\big\} + 0 + \exp\big\{-c_M n\bar{\varepsilon}_n^2\big\} \lesssim \exp\big\{-c_1 n\bar{\varepsilon}_n^2\big\}.$$

It remains to prove (4.14). First note that, by using (4.3) and (4.11), for all $j \ge \bar{J}_n$ and for all sets of knots $\mathbf{k}_j \in \mathcal{K}_j$ such that $M(\mathbf{k}_j) \le \tau/j$, there exists a spline $s_{\mathbf{k}_j,\boldsymbol{\theta}_0} \in \mathcal{S}_q^{\mathbf{k}_j}$ (of course, $\boldsymbol{\theta}_0 = \boldsymbol{\theta}_0(\mathbf{k}_j) = \boldsymbol{\theta}_0(\mathbf{k}_j, f_0)$) such that

$$\|f_0 - s_{\mathbf{k}_j,\boldsymbol{\theta}_0}\|_\infty \le C_{f_0} M^\alpha(\mathbf{k}_j) \le C_{f_0}\tau^\alpha \bar{J}_n^{-\alpha} \le \bar{\varepsilon}_n. \tag{4.15}$$

Since $\|f_0\|_\infty < M$, there exists an $\varepsilon > 0$ such that the spline $s_{\mathbf{k}_j,\boldsymbol{\theta}_0}$ from (4.15) satisfies $\|s_{\mathbf{k}_j,\boldsymbol{\theta}_0}\|_\infty \le M - \varepsilon$ for sufficiently large n. Besides, $\bar{J}_n$ must grow with n in view of (4.11). Then, according to Lemma 4.3, there exists a $\delta = \delta(\mathcal{F}_\alpha, \varepsilon)$ such that, for sufficiently large n, $\|\boldsymbol{\theta}_0(\mathbf{k}_j)\|_\infty \le M$ for all sets of knots $\mathbf{k}_j \in \mathcal{K}_j$ such that $M(\mathbf{k}_j) \le \tau/\bar{J}_n \le \delta$ and $j \ge \bar{J}_n$.

Introduce the events: $E_1^j = \{M(\mathbf{K}_j) \le \tau/j\}$, $E_2^j = \{\|f_0 - s_{\mathbf{K}_j,\boldsymbol{\theta}_0(\mathbf{K}_j)}\|_\infty \le \bar{\varepsilon}_n\}$, $E_3^j = \{\|\boldsymbol{\theta}_0(\mathbf{K}_j) - \boldsymbol{\theta}\|_\infty \le \bar{\varepsilon}_n\}$, $E_4^j = \{\|f_0 - s_{\mathbf{K}_j,\boldsymbol{\theta}}\|_\infty \le 2\bar{\varepsilon}_n\}$ and $E_5^j = \{\|\boldsymbol{\theta}_0(\mathbf{K}_j)\|_\infty \le M\}$. Using the argument from the previous paragraph, the triangle inequality, (4.2) and (4.15), we obtain that

$$E_1^{\bar{J}_n} \subseteq E_2^{\bar{J}_n}, \quad E_1^{\bar{J}_n} \subseteq E_5^{\bar{J}_n}, \quad E_2^j \cap E_3^j \subseteq E_4^j, \quad j \ge q. \tag{4.16}$$

Combining (4.5), (4.7), (4.8), (4.11) and (4.16), we prove (4.14):

$$\begin{aligned}
\mathbb{P}\big(\|s_{\mathbf{K}_J,\boldsymbol{\theta}} - f_0\|_\infty \le 2\bar{\varepsilon}_n\big) &= \mathbb{P}(E_4^J) \ge \mathbb{P}(J=\bar{J}_n)\mathbb{P}\big(E_4^{\bar{J}_n}|J=\bar{J}_n\big)\\
&\ge \mathbb{P}(J=\bar{J}_n)\mathbb{P}\big(E_2^{\bar{J}_n}\cap E_3^{\bar{J}_n}|J=\bar{J}_n\big)\\
&\ge \mathbb{P}(J=\bar{J}_n)\mathbb{P}\big(E_1^{\bar{J}_n}\cap E_3^{\bar{J}_n}\cap E_5^{\bar{J}_n}|J=\bar{J}_n\big)\\
&= \mathbb{P}(J=\bar{J}_n)\mathbb{E}\big[P\big(E_1^{\bar{J}_n}\cap E_3^{\bar{J}_n}\cap E_5^{\bar{J}_n}|J=\bar{J}_n,\mathbf{K}_{\bar{J}_n}\big)\big]\\
&= \mathbb{P}(J=\bar{J}_n)\mathbb{E}\big[\mathbb{I}\{\mathbf{K}_{\bar{J}_n}\in E_1^{\bar{J}_n}\cap E_5^{\bar{J}_n}\}\mathbb{P}\big(E_3^{\bar{J}_n}|J=\bar{J}_n,\mathbf{K}_{\bar{J}_n}\big)\big]\\
&\ge \mathbb{P}(J=\bar{J}_n)\mathbb{P}\big(E_1^{\bar{J}_n}|J=\bar{J}_n\big)\inf_{\|\boldsymbol{\theta}_0\|_\infty\le M}\mathbb{P}\big(\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_\infty\le\bar{\varepsilon}_n|J=\bar{J}_n\big)\\
&\gtrsim \exp\big(-(c_2+c_3)\bar{J}_n\log^{t_2\vee t_3}\bar{J}_n\big)\exp\big(-c_0(M)\bar{J}_n\log(1/\bar{\varepsilon}_n)\big)\\
&\gtrsim \exp\big(-(c_0(M)+c_2+c_3)\bar{J}_n\log(1/\bar{\varepsilon}_n)\big).
\end{aligned}$$

□

Remark 4.1 *Condition (4.6) is used in the proof of Theorem 4.1 exclusively to enforce* $\sum_{j=q}^{J}\mathbb{P}\big(J=j\big)\mathbb{P}\big(m(\mathbf{K}_j)<\delta(j)|J=j\big)$ *to be zero, when proving (4.13). Inspection of the proof shows however that, instead of condition (4.6), it would suffice to require this sum to be upper-bounded by a multiple of* $\exp\big(-c_1 n\bar{\varepsilon}_n^2\big)$. *Although this would be a weaker requirement, typically the sequence* $\bar{\varepsilon}_n$ *will depend on the unknown smoothness* α. *To avoid the dependence on* $\bar{\varepsilon}_n$, *a slightly stronger condition (based on the fact that* $n\bar{\varepsilon}_n^2$ *is of a smaller order than* n *as* $n\to\infty$*) can be proposed. Namely, if condition (4.6) is replaced by*

$$\sum_{j=q}^{J_n}\mathbb{P}\big(J=j\big)\mathbb{P}\big(m(\mathbf{K}_j)<\delta(j)|J=j\big)\le c_5\exp(-c_4 n), \tag{6'}$$

for some $c_4, c_5 > 0$ *and a function* $\delta(\cdot)$ *as in (4.6), then the conclusions of Theorem 4.1 remain valid as long as* J_n *is a sequence satisfying (4.9) and (4.10); cf. Section 4.5 for a comparison of (4.6) and (6').*

Remark 4.2 *If the range of the underlying curve* f_0 *is contained in some known interval* $[a,b]\subset\mathbb{R}$, *then, according to Lemma 4.3 and the proof of property (4.14), the prior on* $\boldsymbol{\theta}\in\mathbb{R}^j$ *can be chosen to be supported on, say,* $[a-1,b+1]^j$ *so that (4.8) has to hold only for* $\boldsymbol{\theta}_0\in[a-1,b+1]^j$. *Condition (4.10) will be trivially satisfied for* $M_n > (1-a)\vee(b+1)$.

Remark 4.3 *If (4.26) is assumed instead of (4.7), the proof of (4.14) can then be simplified a lot, as in this case one can condition on the event* $\{\mathbf{K}_{\bar{J}_n}=\bar{\mathbf{k}}_{\bar{J}_n}\}$ *so that* $\boldsymbol{\theta}_0=\boldsymbol{\theta}_0(\bar{\mathbf{k}}_{\bar{J}_n})$ *becomes fixed and* $\mathbb{P}(E_1^J|J=\bar{J}_n,\mathbf{K}_{\bar{J}_n}=\bar{\mathbf{k}}_{\bar{J}_n})=1$.

4.4 Implications of the main result

We clarify now the relevance of our result. Consider a family of models $\mathcal{P}=\{\mathbb{P}_f : f\in\mathcal{F}_{\mathcal{A}}\}$, $\mathcal{F}_{\mathcal{A}}=\cup_{\alpha\in\mathcal{A}}\mathcal{F}_\alpha$, with densities p_f with respect to some common dominating measure. Assume that we observe a sample $\mathbf{X}^{(n)}=\{X_1,\dots,X_n\}\sim p_{f_0}^{(n)}$, $X_i\overset{ind}{\sim}p_{f_0}$, $f_0\in\mathcal{F}_\alpha$ for some unknown smoothness $\alpha\in\mathcal{A}$. The Bayesian approach consists of

putting a prior measure Π on $\mathcal{F} \subseteq \mathcal{F}_{\mathcal{A}}$ which, together with the likelihood $p_f^{(n)}$, leads to the posterior distribution $\Pi(\cdot|\mathbf{X}^{(n)})$ via Bayes' formula:

$$\Pi\big(A|\mathbf{X}^{(n)}\big) = \frac{\int_A p_f^{(n)}(\mathbf{X}^{(n)})\, d\Pi(f)}{\int_{\mathcal{F}} p_f^{(n)}(\mathbf{X}^{(n)})\, d\Pi(f)}$$

for a measurable $A \subseteq \mathcal{F}$. The asymptotic behavior of the posterior distribution can be studied from the point of view of the probability measure $\mathbb{P}_0 = \mathbb{P}_{f_0}$; see [38].

For two densities p_f and p_g with $f, g \in \mathcal{F}_{\mathcal{A}}$, define the (squared) Hellinger metric $h^2(p_f, p_g) = 2\big(1 - \mathbb{E}_g\sqrt{p_f(X)/p_g(X)}\big)$, Kullback-Leibler divergence $K(p_f, p_g) = -\mathbb{E}_g \log\big(p_f(X)/p_g(X)\big)$ and the Csiszár f-divergence $V(p_f, p_g) = \mathbb{E}_g \log^2\big(p_f(X)/p_g(X)\big)$. Define also the ball $B(\varepsilon_n, f_0) = \big\{f \in \mathcal{F} : K(f, f_0) \le \varepsilon^2, V(f, f_0) \le \varepsilon^2\big\}$.

The following theorem is a version of Theorem 2.1 from [37] which makes a statement about the asymptotic behavior of a posterior measure.

Theorem 4.2 (Theorem 2.1 of [38])
Let Π_n be a sequence of priors on $\mathcal{F}$. Suppose that for two positive sequences $\kappa_n \ge \tilde{\kappa}_n$ such that $n\tilde{\kappa}_n^2 \to \infty$ and $\kappa_n \to 0$ as $n \to \infty$, sets $\mathcal{F}_n \subseteq \mathcal{F}$ and constants $b_1, b_2, b_3, b_4 > 0$, the following conditions hold:

$$\log N\big(\kappa_n, \mathcal{F}_n, h\big) \le b_1 n\kappa_n^2, \tag{4.17}$$

$$\Pi_n(\mathcal{F}\backslash\mathcal{F}_n) \le b_2 e^{-(b_3+4)n\tilde{\kappa}_n^2}, \tag{4.18}$$

$$\Pi_n(B(\tilde{\kappa}_n, f_0)) \ge b_4 e^{-b_3 n\tilde{\kappa}_n^2}. \tag{4.19}$$

Then, for large enough $M > 0$, $\Pi_n\big(f \in \mathcal{F} : h(p_f, p_{f_0}) \ge M\kappa_n|\mathbf{X}^{(n)}\big) \to 0$ as $n \to \infty$ in $\mathbb{P}_{f_0}$-probability.

The conditions of this theorem require the existence of a *sieve* $\mathcal{F}_n$ with small entropy (4.17) which contains most of the prior mass (4.18) and with enough prior mass around the parameter f_0 which indexes the "true" underlying measure of the data. Assume now that the models in $\mathcal{P}$ are such that for d^2 being h^2, K or V, $d^2(p_f, p_{f_0}) \lesssim \|f - f_0\|_2^2$. If in addition one can prove that in the considered model $h(p_f, p_{f_0}) \gtrsim \|f - f_0\|_2$, then Theorem 4.2 delivers a contraction rate κ_n with respect to the L_2-distance as well. Some examples of models for which the above relations between norms can be established are, white noise, density estimation, non-parametric regression, binary regression, Poisson regression and classification; cf. [25, 38, 86]. We should note here that it requires a fair piece of effort to implement this idea for many concrete models, only for the white noise model the above relations between norms are straightforward. Once these relation between norms are established, one can apply our meta-theorem (Theorem 4.1) to obtain an adaptive contraction rate that essentially verifies (4.17)–(4.19) for our spline-based prior. We summarize this in the following theorem.

Theorem 4.3 (Contraction)
Let Π be the spline prior described in Section 4.3. Consider a family of models $\mathcal{P} = \big\{\mathbb{P}_f : f \in \mathcal{F}_{\mathcal{A}}\big\}$, $\mathcal{F}_{\mathcal{A}} = \cup_{\alpha\in\mathcal{A}}\mathcal{F}_\alpha$, with densities p_f with respect to some common dominating measure. Assume also that the models in $\mathcal{P}$ are such that $d^2(p_f, p_{f_0}) \lesssim \|f - f_0\|_2^2$ for d^2 being h^2, K or V. Take an i.i.d. sample $\mathbf{X}^{(n)} = \{X_1, \ldots, X_n\}$, $X_i \sim p_{f_0}$, $f_0 \in \mathcal{F}_\alpha$, $\|f_0\|_\infty < M$, for some unknown smoothness $\alpha \in \mathcal{A}$, $\alpha \le q$. Consider a prior Π that verifies (4.4) through (4.8) for

certain constants c_1, c_2, c_3, t_1, t_2 and t_3. Assume that at least one of the two conditions, $\alpha > 1$ or $t_2 \wedge t_3 < 1$, is fulfilled.

Then, for large enough $C > 0$, $\Pi\big(f \in \mathcal{F} : h(p_f, p_{f_0}) \geq C r_n | \mathbf{X}^{(n)}\big) \to 0$ as $n \to \infty$ in $\mathbb{P}_0$-probability for $r_n = n^{-\alpha/(2\alpha+1)} (\log n)^{\alpha/(2\alpha+1)+(1-t_1)/2}$. If $h(p_f, p_{f_0}) \gtrsim \|f - f_0\|_2$ then in the previous statement the Hellinger distance may be replaced by the L_2-distance and the statement remains valid.

Proof: We have that for some constant $\rho > 0$ and $\mathcal{F} = \mathcal{S}$, $\mathcal{F}_n = \mathcal{S}_n$,

$$N\big(\kappa_n, \mathcal{F}_n, h\big) \leq N(\kappa_n/\rho, \mathcal{F}_n, \|\cdot\|_2), \tag{4.20}$$

$$\Pi(\mathcal{F} \backslash \mathcal{F}_n) = \mathbb{P}\big(s_{\mathbf{K}_J, \boldsymbol{\theta}} \notin \mathcal{F}_n\big), \tag{4.21}$$

$$\Pi(B(\bar{\kappa}_n, f_0)) \geq \mathbb{P}\big(\|s_{\mathbf{K}_J, \boldsymbol{\theta}} - f_0\|_\infty \leq \bar{\kappa}_n/\rho\big). \tag{4.22}$$

The first inequality follows from the fact that by assumption $h(p_f, p_g) \leq \rho \|f - g\|_2$ and so a κ/ρ-cover of $\mathcal{F}_n$ according to $\|\cdot\|_2$ induces a κ-cover of $\mathcal{F}_n$ according to h. Then, since $d^2(p_f, p_{f_0}) \leq \rho \|f - f_0\|_2^2$ for d^2 being K or V, we have $B(\bar{\kappa}_n, f_0) \supset \big\{f \in \mathcal{F} : \|f - f_0\|_2 \leq \bar{\kappa}/\rho\big\}$ and the last inequality follows.

By assumption $f_0 \in \mathcal{F}_\alpha$ satisfies the conditions of Theorem 4.1; assume (4.3) holds for some C_{f_0}. Consider then a prior that satisfies (4.4)–(4.8). Let us present a choice of quantities M_n, $\delta(j)$, J_n, $\bar{J}_n$, ε_n and $\bar{\varepsilon}_n$ that meet conditions (4.9)–(4.11). First of all, the sequence M_n can be taken a polynomial in n (for instance, for normal or exponential conditional priors for $\boldsymbol{\theta} \in \mathbb{R}^j$ in (4.10)) and $1/\delta(j)$ a polynomial in j. Next, note that there is no $\bar{J}_n$ that satisfies (4.11) if both $\alpha \leq 1$ and $t_2 \wedge t_3 = 1$ hold. If either $\alpha > 1$ or $t_2 \wedge t_3 < 1$, then the best possible choices are $\bar{J}_n = \bar{J}_n(C_1) = \tau C_{f_0}^{1/\alpha} (\bar{\varepsilon}_n(C_1))^{-1/\alpha}$ so that the first inequality of (4.11) is satisfied, $\bar{\varepsilon}_n = \bar{\varepsilon}_n(C_1) = C_1 (\log n/n)^{\alpha/(2\alpha+1)}$ for sufficiently large $C_1 \geq 1$ so that the second inequality of (4.11) is satisfied, $J_n = C_2 n^{1/(2\alpha+1)} (\log n)^{2\alpha/(2\alpha+1)-t_1}$ for sufficiently large C_2 (any $C_2 \geq C_1^2$ will do) so that the first inequality of (4.10) is satisfied, and finally,

$$\varepsilon_n = C_3 n^{-\alpha/(2\alpha+1)} (\log n)^{\alpha/(2\alpha+1)+(1-t_1)/2}$$

for sufficiently large C_3 so that (4.9) is satisfied. Since these quantities satisfy (4.9)–(4.11), Theorem 4.1 implies conditions (4.12)–(4.14) for the quantities defined above. Besides, we take constants C_1, C_2, C_3 so big that (4.13) and (4.14) also hold for $\bar{\varepsilon}_n(\sqrt{C_1})$ and $\bar{J}_n(\sqrt{C_1})$.

Now, take $\kappa_n = 2\rho\varepsilon_n$ and $\bar{\kappa}_n = 2\rho\bar{\varepsilon}_n(\sqrt{C_1})$. Then it follows from (4.12) and (4.20) that

$$N\big(\kappa_n, \mathcal{F}_n, h\big) \leq N(\kappa_n/\rho, \mathcal{F}_n, \|\cdot\|_2) = N(\varepsilon_n, \mathcal{F}_n, \|\cdot\|_2) \lesssim n\varepsilon_n^2 \lesssim n\kappa_n^2. \tag{4.23}$$

Next, using (4.21) and (4.13) for $\bar{\varepsilon}_n(C_1)$ and $\bar{J}_n(C_1)$, we obtain that

$$\begin{aligned} \Pi(\mathcal{F} \backslash \mathcal{F}_n) = \mathbb{P}\big(s_{\mathbf{K}_J, \boldsymbol{\theta}} \notin \mathcal{F}_n\big) &\lesssim \exp\big\{-c_1 n \bar{\varepsilon}_n^2(C_1)\big\} \\ &= \exp\big\{-c_1 (2\rho)^{-2} C_1 n \bar{\kappa}_n^2\big\} \leq \exp\big\{-5 n \bar{\kappa}_n^2\big\} \end{aligned} \tag{4.24}$$

for sufficiently large C_1. Denote $K = (c_0(M) + c_2 + c_3) \tau C_{f_0}^{1/\alpha} (2\rho)^{-2} \alpha/(2\alpha+1)$, then

$$(c_0(M) + c_2 + c_3) \bar{J}_n(\sqrt{C_1}) \log(1/\bar{\varepsilon}_n(\sqrt{C_1})) = K C_1^{-(1+1/\alpha)} n \bar{\kappa}_n^2 (1 + o(1))$$

as $n \to \infty$. The last relation, (4.22) and (4.14) for $\bar{\varepsilon}_n(\sqrt{C_1})$ and $\bar{J}_n(\sqrt{C_1})$ imply that

$$\begin{aligned}\Pi(B(\bar{\kappa}_n, f_0)) &\geq \mathbb{P}\big(\|s_{\mathbf{K}_J,\boldsymbol{\theta}} - f_0\|_\infty \leq 2\bar{\varepsilon}_n(\sqrt{C_1})\big)\\ &\gtrsim \exp\big\{-(c_0(M) + c_2 + c_3)\bar{J}_n(\sqrt{C_1})\log(1/\bar{\varepsilon}_n(\sqrt{C_1}))\big\}\\ &\gtrsim \exp\big\{-KC_1^{-(1+1/\alpha)} n\bar{\kappa}_n^2\big\} \geq \exp\big\{-n\bar{\kappa}_n^2\big\} \qquad (4.25)\end{aligned}$$

for sufficiently large C_1. Thus, for sufficiently large C_1 (and C_2, C_3), the relations (4.17)–(4.19) follow from (4.23), (4.24) and (4.25) respectively.

Finally, applying Theorem 4.2 (since (4.17)–(4.19) are fulfilled), we conclude that the contraction rate of the resulting posterior is at most ε_n, which appears to be optimal (up to a logarithmic factor) in a minimax sense over the Hölder class $\mathcal{H}_\alpha$ (also over α-smooth Sobolev class). □

Remark 4.4 *A priori, it may be unknown whether $\alpha > 1$ or not, or it may be simply known that $\alpha \leq 1$. We can however always ensure the condition $t_2 \wedge t_3 < 1$ by an appropriate choice of prior. For example, we take a geometric prior on J so that $t_2 = 0$ and a prior on $\mathbf{K}_j$ such that (4.26) (which implies (4.7)) holds with, say, $t_3 = 0$.*

4.5 Examples of Priors

We give now examples of particular choices for the several components of our hierarchical prior that verify conditions (4.4)–(4.8), (6') and the second relation in (4.10).

As for the prior on the number of basis functions, assumptions (4.4) and (4.5) hold for the geometric, Poisson and negative binomial distributions; cf. [86] (assumption (5) is slightly different from the corresponding assumption (B1) in [86]). Assumption (4.8), in turn, will trivially hold if we assume, for example, the coordinates of $\boldsymbol{\theta} \in \mathbb{R}^j$ to be (conditionally on $J = j$) independent and identically distributed according to a density ϕ that is uniformly bounded away from zero on the interval $[-M, M]$. On the other hand, the prior distribution on $\boldsymbol{\theta} \in \mathbb{R}^j$ (conditionally on $J = j$) should have sufficiently light tails so that the second requirement in (4.10) holds for a sequence M_n that converges to infinity as $n \to \infty$ not faster than polynomially in n. It can easily be checked for normal and Gamma densities ϕ. Let us consider standard normal ϕ. As $q \leq j \leq J_n$ and taking $M_n = n$, we immediately derive the required relation:

$$\mathbb{P}\big(\boldsymbol{\theta} \notin C^j(M_n) | J = j\big) \leq j\mathbb{P}(|\theta_1| \geq M_n | J = j) \leq \frac{J_n 2\exp(-M_n^2/2)}{\sqrt{2\pi}M_n} \leq \exp(-c_M n\bar{\varepsilon}_n^2).$$

There is an ample choice of priors on $\mathbf{K}_J$, given $J = j$, that satisfy condition (4.6). First note that this condition enforces the prior on the location of the knots, for each $J = j$, to be such that, with probability 1, adjacent knots are at least $\delta(j)$ apart. The function $1/\delta(j)$ can be taken a polynomial in j of high degree which makes the requirement less restrictive. If a certain sequence ε_n verifies the conditions of Theorem 4.1, then an increase in the exponent of $1/\delta(j)$ can be accommodated by making ε_n larger by a multiplicative factor (cf. condition (4.9)).

A simple choice for the prior on $\mathbf{K}_J$, given $J = j$, is to pick $(j - q)$ knots uniformly at random, without replacement, on a uniform $\delta(j)$-sparse grid. This construction is possible if δ is chosen in such a way that $\lfloor 1/\delta(j) \rfloor > j - q$ for all j. Another way is to generate the $(j - q)$ inner knots in $\mathbf{K}_j$ sequentially in the following

way: add a knot K_1 uniformly at random on the interval $[\delta(j), 1-\delta(j)]$, then a knot K_2 uniformly at random on the interval $[\delta(j), 1-\delta(j)]\backslash(K_1-\delta(j), K_1+\delta(j))$ and so on. Finally, take the ordered $\mathbf{K}_j = (K_{(1)}, \ldots, K_{(j-q)})$. This construction is always possible if $1/\delta(j)$ grows faster than $2(j-q)$. If J is Poisson distributed, these points are simply distributed like a homogeneous Poisson process, conditioned to have all points at least $\delta(J)$ apart. Clearly, condition (4.6) is satisfied for these two constructions since all prior mass is concentrated on partitions with sparseness larger than $\delta(j)$.

It is also easy to see that condition (4.7) is verified for the knot vectors obtained from one of these two constructions. In fact, condition (4.7) is trivially fulfilled if, for some $0 \le t_3 < 1$,

$$\mathbb{P}(\mathbf{K}_j = \bar{\mathbf{k}}_j) \gtrsim \exp\big(-c_3 j \log^{t_3} j\big), \tag{4.26}$$

where $\bar{\mathbf{k}}_j \in \mathcal{K}_j$ is the set of $(j-q)$ equally spaced inner knots. This suggests a mechanism to assure that any prior which verifies (4.6) can be slightly modified to also verify (4.7): given $J = j$, generate a Bernoulli random variable X with success probability, say, $\exp(-c_3 j \log^{t_3} j)$; if $X = 1$, then take $\mathbf{K}_j = \bar{\mathbf{k}}_j$, otherwise pick the knots in $\mathbf{K}_j$ according to any procedure which verifies (4.6), for instance, one of the two procedures described above. The resulting prior will trivially satisfy both (4.6) and (4.7).

Condition (4.6) necessarily excludes some knot vectors from the support of the prior (and then also from the support of the posterior.) It is therefore of interest to design a weaker alternative for condition (4.6). Condition (6′) plays this role in that it allows priors on $\mathbf{K}$ which can have any set of knots of $[0,1]$ in its support. Assuming condition (6') instead of (4.6) consequently allows us to put positive mass on any vector of simple knots in a straightforward way: generate a Bernoulli random variable with success probability $1-c_5\exp(-c_4 n)$; if $X = 1$ take $\mathbf{K}_j = \bar{\mathbf{k}}_j$; if $X = 0$, then take an arbitrary $\mathbf{K}_j$ (for example, independent, uniformly distributed points on $[0,1]$). If we take $1/\delta(j) = j$ and $\tau \ge q$, then conditions (6') and (4.7) are verified. This procedure, although simple, does place little prior mass on knot vectors with inhomogeneous distributions.

An alternative, less degenerate prior, which verifies (6') and (4.7) can be obtained in the following way. Given $J = j$, first generate a Bernoulli random variable X_1 with success probability $c_5\exp(-c_4 n)$; if $X_1 = 1$ distribute the $(j-q)$ knots arbitrarily; if $X_1 = 0$, then generate another Bernoulli random variable X_2 with success probability $\exp(-j)$; if $X_2 = 1$, then take $(j-q)$ equally spaced knots $\bar{\mathbf{k}}_j$; if $X_2 = 0$, then place the knots in such a way that (4.6) is verified. This procedure allows good control of the prior on the knots while not excluding arbitrary knot vectors.

Note that the priors described above which verify (4.4)–(4.8) do not depend on the sample size n, as prescribed by the Bayesian paradigm. Condition (6') is a weaker requirement than condition (4.6), but it will introduce a dependence on the sample size n in the prior.

Remark 4.5 *The common practice, in applications, of endowing the location of the knots with a Poisson process prior results in a prior that does not verify assumption (4.6). Assumption (6'), however, will be satisfied if the prior is modified in such a way that a large enough prior mass is assigned to an equally spaced knot vector.*

4.6 Numerical Example

We present here some numerical results. By applying the reversible jump MCMC (RJMCMC) method [40], we compare two hierarchical priors in the nonparametric regression model. Both priors are based on splines, as described in Section 4.3, and they satisfy conditions (4.4)–(4.8). The first has a.s. equally spaced knots and in the second the locations of the knots are random; we therefore refer to these priors as the fixed knots prior and the free knots prior. We also look into the possibility of using data-driven priors on the knots based on a two stage empirical Bayes procedure. We say that vector $\mathbf{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d$ is ordered if $x_1 \leq \ldots \leq x_d$.

Consider $n = 1000$ observations $\mathbf{X}^{(n)} = \{(t_i, Y_i), i = 1, \ldots, n\}$ from the non-parametric regression model with regular design points $\mathbf{t}^{(n)} = (t_1, \ldots, t_n)$, $t_i = i/n$:

$$Y_i = f(t_i) + \xi_i, \quad i = 1, \ldots, n, \tag{M}$$

where the ξ_i's are independent, standard Gaussian random variables. It is well known that the relations between appropriate norms required to apply Theorem 4.3 hold for the model (M). The regression function $f(\cdot)$ is taken to be the so-called Doppler function

$$f(t) = 10\sqrt{t(1-t)}\,\sin\left(\frac{2\pi \cdot 1.05}{t + 0.05}\right), \quad t \in [0,1], \tag{4.27}$$

which we plot in Figure 4.6.1 together with the observations from the model (M).

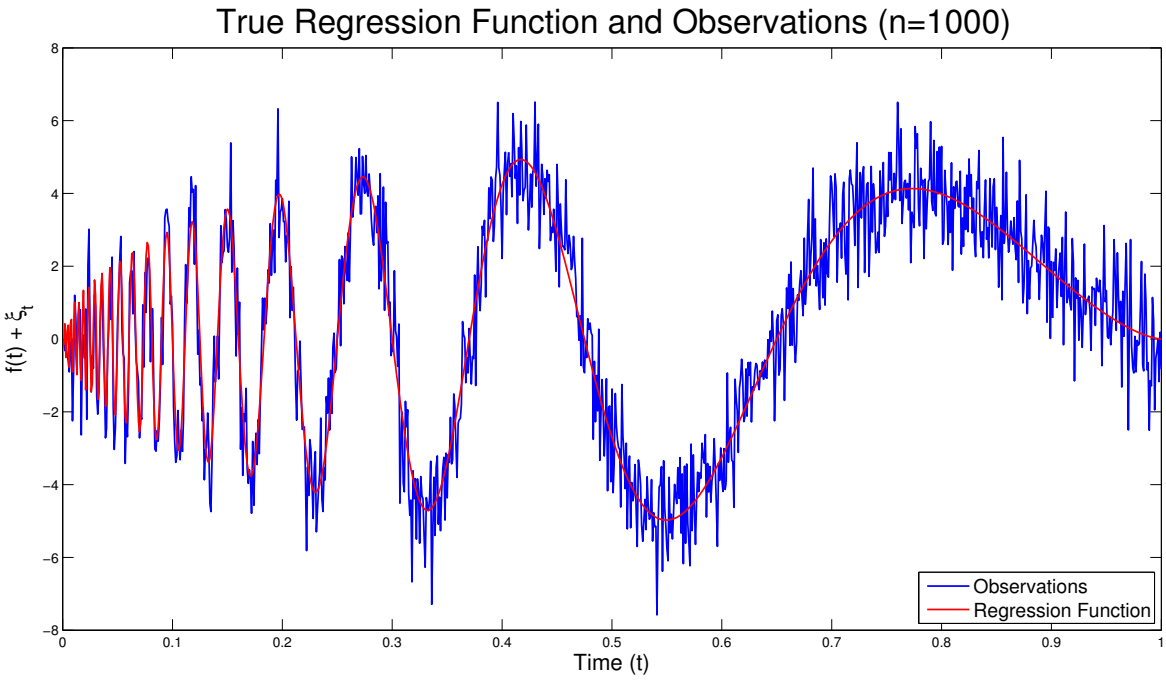

Figure 4.6.1: Simulated data from model (M) used in this section (in blue) and the true regression function (in red).

Now we describe the two priors, the fixed knots prior and the free knots prior. In both hierarchical priors we endow J with a (shifted, with support starting with $q \geq 2$) Poisson prior with mean ν and on each spline coefficient we put a uniform prior on $[-M, M]$. In the fixed knots prior, given $J = j$, the $j - q$ inner knots are taken to be equally spaced: $k_i = i/(j-q+1)$, $i = 1, \ldots, j-q$. In the free knots prior, given $J = j$, first generate $U_1, \ldots, U_{j-q}$, uniformly on $\big[0, 1-(j-q+1)\delta(j)\big)$ with $\delta(j) = 1/j^2$, and let $U_{(1)} \leq \ldots \leq U_{(j-q)}$. Next, take the knot vector $\mathbf{K}_j$ with entries $K_{i,j} = U_{(i)} + i\delta(j)$, $i = 1, \ldots, j-q$. We represent the fixed knots posterior density as $\bar{\pi}\big(j, \mathbf{k}_j, \boldsymbol{\theta}_j \mid \mathbf{X}^{(n)}\big)$ and the free knots posterior as $\tilde{\pi}\big(j, \mathbf{k}_j, \boldsymbol{\theta}_j \mid \mathbf{X}^{(n)}\big)$. We have

$$\bar{\pi}\big(j, \mathbf{k}_j, \boldsymbol{\theta}_j \mid \mathbf{X}^{(n)}\big) \propto \varphi_n\big(\mathbf{X}^{(n)} - s_{\mathbf{k}_j, \boldsymbol{\theta}_j}(\mathbf{t}^{(n)})\big)\, \nu^{j-q}\, (2M)^{-j},$$
$$\tilde{\pi}\big(j, \mathbf{k}_j, \boldsymbol{\theta}_j \mid \mathbf{X}^{(n)}\big) \propto \varphi_n\big(\mathbf{X}^{(n)} - s_{\mathbf{k}_j, \boldsymbol{\theta}_j}(\mathbf{t}^{(n)})\big)\, \nu^{j-q}\, (2M)^{-j}\, \big(1-(j-q+1)\delta(j)\big)^{j-q},$$

where $s_{\mathbf{k}_j, \boldsymbol{\theta}_j}(\mathbf{t}^{(n)}) = (s_{\mathbf{k}_j, \boldsymbol{\theta}_j}(t_1), \ldots, s_{\mathbf{k}_j, \boldsymbol{\theta}_j}(t_n))$ represents the spline $s_{\mathbf{k}_j, \boldsymbol{\theta}_j}$ evaluated at the design points $\mathbf{t}^{(n)}$ and φ_n stands for the density of n independent standard Gaussian random variables.

We implemented RJMCMC procedures for these two priors to sample from the corresponding posteriors. A generic state of the sampler is a vector $\big(j, \mathbf{k}_j, \boldsymbol{\theta}_j\big) \in \mathbb{N} \times \mathbb{R}^{j-q} \times \mathbb{R}^j$. To sample from the posterior corresponding to the fixed knots prior, we consider three types of moves: a) changing the coefficients of a spline, b) adding a knot and c) removing a knot. In addition to these moves, the sampler for the posterior corresponding to the free knots prior has an extra move: d) changing the location of the knots. These moves are attempted with probabilities p_a, p_b, p_c, p_d, $(p_a + p_b + p_c + p_d = 1)$ respectively, which are parameters of the sampler.

A move of type a) corresponds to jumping from the state $\big(j, \mathbf{k}_j, \boldsymbol{\theta}_j\big)$ to a proposal $\big(j, \mathbf{k}_j, \boldsymbol{\vartheta}_j\big)$ where $\boldsymbol{\vartheta}_j = \boldsymbol{\theta}_j + \eta_a \mathbf{u}$ and $\mathbf{u}$ is a vector of j independent standard normal random variables. This move is attempted with probability p_a. Both η_a and p_a are parameters of the sampler. Moves of type d) correspond to jumping from the state $\big(j, \mathbf{k}_j, \boldsymbol{\theta}_j\big)$ to a proposal $\big(j, \boldsymbol{\kappa}_j, \boldsymbol{\theta}_j\big)$ where $\boldsymbol{\kappa}_j$ is obtained from $\mathbf{k}_j$ by perturbing its i-th entry, with the index i chosen uniformly at random, and then ordering the resulting vector. The perturbation is $\kappa_{i,j} = k_{i,j} + \eta_d v$, with v a standard normal random variable. This move is attempted with probability p_d and again, both η_d and p_d are parameters of the sampler. The acceptance probabilities for moves of type a) and moves of type d) are given by, respectively,

$$\min\left(1, \frac{\pi\big(j, \mathbf{k}_j, \boldsymbol{\vartheta}_j \mid \mathbf{X}^{(n)}\big)}{\pi\big(j, \mathbf{k}_j, \boldsymbol{\theta}_j \mid \mathbf{X}^{(n)}\big)}\right) \qquad \text{and} \qquad \min\left(1, \frac{\tilde{\pi}\big(j, \boldsymbol{\kappa}_j, \boldsymbol{\theta}_j \mid \mathbf{X}^{(n)}\big)}{\tilde{\pi}\big(j, \mathbf{k}_j, \boldsymbol{\theta}_j \mid \mathbf{X}^{(n)}\big)}\right),$$

where π is either $\bar{\pi}$ or $\tilde{\pi}$.

Now we specify how proposals for moves of type b), where we add an extra knot to the current state of the chain $\big(j, \mathbf{k}_j, \boldsymbol{\theta}_j\big)$, are designed. The proposal will, for both priors, be a state $\big(j+1, \boldsymbol{\kappa}_{j+1}, \boldsymbol{\vartheta}_{j+1}\big)$. For the fixed knots prior we propose $\kappa_{i,j+1} = i/(j-q+2)$, for $i = 1, \ldots, j-q+1$. For the free knots prior, generate a new knot k uniformly on $(0,1)$ so that $k \in [k_{i-1,j}, k_{i,j}]$ (with $k_{0,j} = 0$ and $k_{j-q+1,j} = 1$) for some $i \in \{1, \ldots, j-q+1\}$, and propose $\boldsymbol{\kappa}_{j+1} = (k_{1,j}, \ldots, k_{i-1,j}, k, k_{i,j}, \ldots, k_{j-q,j})$.

For moves of type b), it remains to describe how the coefficient vector $\boldsymbol{\vartheta}_{j+1}$ is generated in the proposal. Whatever the vector $\boldsymbol{\kappa}_{j+1}$ is, for the sake of comparing the priors, the procedure for proposing $\boldsymbol{\vartheta}_{j+1}$ is the same for both priors. To ease the notation, we abbreviate the current state and the proposed state as $\big(j, \mathbf{k}, \boldsymbol{\theta}\big)$ and $\big(j+1, \boldsymbol{\kappa}, \boldsymbol{\vartheta}\big)$, where both $\boldsymbol{\kappa}$ and $\boldsymbol{\vartheta}$ have one more element than $\mathbf{k}$ and $\boldsymbol{\theta}$, respectively. The coefficients $\boldsymbol{\vartheta}$ will be obtained via (perturbed) interpolation at $j+1$ points $\mathbf{t} = \mathbf{t}_{j+1} = (t_1, \ldots, t_{j+1})$. Of these $j+1$ points, $j-q+2$ points are taken

to be the midpoints of the intervals comprised between the adjacent points of the vector $(0, \boldsymbol{\kappa}, 1) \in [0,1]^{j-q+3}$; the remaining $q-1$ points are the first $q-1$ elements from the list $0, 1, \kappa_1, \kappa_{j-q+1}, \kappa_2, \kappa_{j-q}, \kappa_3, \ldots$. The vector $\mathbf{t}$ is assumed to be ordered.

Consider now the $(j+1) \times (j+1)$ matrices $C = C_j(\boldsymbol{\kappa}, \mathbf{t})$ with (i,l)-entry $B_l^{\boldsymbol{\kappa}}(t_i)$ and the $(j+1) \times j$ matrices $D = D_j(\mathbf{k}, \mathbf{t})$ with (i,l)-entry $B_l^{\mathbf{k}}(t_i)$. One can show that for our choice of interpolation points C and D are of full column rank. For a matrix M denote by $\mathscr{L}(M)$ the linear space spanned by the columns of matrix M. Then $\mathscr{L}(C) = \mathbb{R}^{j+1}$, $\mathscr{L}(D) \subseteq \mathbb{R}^{j+1}$ with $\dim(\mathscr{L}(D)) = j$. Let $\mathbf{w} \in \mathscr{L}^{\perp}(D)$ (the orthogonal complement of $\mathscr{L}(D)$ so that $D^T\mathbf{w} = \mathbf{0}$) which is unique up to scaling. Clearly, the interpolation problem $s_{\boldsymbol{\kappa},\boldsymbol{\vartheta}}(\mathbf{t}) = s_{\mathbf{k},\boldsymbol{\theta}}(\mathbf{t})$ corresponds to the system of linear equations $C\boldsymbol{\vartheta} = D\boldsymbol{\theta}$. Because of the mismatch in the dimensions of $\boldsymbol{\theta}$ and of $\boldsymbol{\vartheta}$, this relation between $\boldsymbol{\theta}$ and $\boldsymbol{\vartheta}$ is not a bijection. Indeed, $\boldsymbol{\theta} = (D^TD)^{-1}D^TC\boldsymbol{\vartheta}_\rho$ for all $\boldsymbol{\vartheta}_\rho = \boldsymbol{\vartheta}_\rho(\boldsymbol{\theta}) = C^{-1}(D\boldsymbol{\theta} + \rho\mathbf{w})$, $\rho \in \mathbb{R}$.

Assume that by default all vectors are column vectors. Our proposals for $\boldsymbol{\vartheta}$ is the following linear function that matches the dimensions of $\boldsymbol{\theta}$ and $\boldsymbol{\vartheta}$:

$$\boldsymbol{\vartheta} = g(\boldsymbol{\theta}, \rho) = C^{-1}\begin{bmatrix} D & \mathbf{w} \end{bmatrix}\begin{bmatrix} \boldsymbol{\theta} \\ \eta\rho + \mathring{\rho} \end{bmatrix},$$

where ρ is a standard Gaussian random variable, $\eta > 0$ and $\mathring{\rho} = \mathring{\rho}(\boldsymbol{\theta})$ is taken to be

$$\mathring{\rho}(\boldsymbol{\theta}) = \arg\min_{\rho \in \mathbb{R}} \| s_{\boldsymbol{\kappa},\boldsymbol{\vartheta}_\rho(\boldsymbol{\theta})} - s_{\mathbf{k},\boldsymbol{\theta}} \|_2^2 = \frac{\langle s_{\boldsymbol{\kappa},\boldsymbol{\omega}_1},\, s_{\mathbf{k},\boldsymbol{\theta}} - s_{\boldsymbol{\kappa},\boldsymbol{\omega}_2} \rangle}{\langle s_{\boldsymbol{\kappa},\boldsymbol{\omega}_1},\, s_{\boldsymbol{\kappa},\boldsymbol{\omega}_1} \rangle},$$

$\boldsymbol{\omega}_1 = C^{-1}\mathbf{w}$, $\boldsymbol{\omega}_2 = C^{-1}D\boldsymbol{\theta}$ and $\langle s_1, s_2 \rangle$ represents the inner product $\int_0^1 s_1(t)s_2(t)dt$. The interpretation of $\mathring{\rho}$ is that our proposal for $s_{\boldsymbol{\kappa},\boldsymbol{\vartheta}}$ is "centered" (cf. [16]) around a good approximation $s_{\boldsymbol{\kappa},\boldsymbol{\vartheta}_{\mathring{\rho}}}$ of the previous state $s_{\mathbf{k},\boldsymbol{\theta}}$. This central state $s_{\boldsymbol{\kappa},\boldsymbol{\vartheta}_{\mathring{\rho}}}$ can be seen as an ideal interpolator.

It is straightforward to check that the Jacobian matrix of the mapping g is

$$J_g = J_g(\eta) = C^{-1}\left(\begin{bmatrix} D & \eta\mathbf{w} \end{bmatrix} + \begin{bmatrix} \mathbf{w} & \ldots & \mathbf{w} \end{bmatrix} \operatorname{diag}\left((\nabla_{\boldsymbol{\theta}}\mathring{\rho}(\boldsymbol{\theta}), \eta) \right) \right),$$

where $\operatorname{diag}(\mathbf{v})$ denotes a square matrix with the entries of $\mathbf{v}$ in its main diagonal and $\nabla_{\boldsymbol{\theta}}\mathring{\rho}(\boldsymbol{\theta})$ is the gradient of $\mathring{\rho}(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$. Note that the determinant of this Jacobian does not depend on the gradient of $\mathring{\rho}$ and is given by

$$\det\left(J_g\right) = \eta \frac{\det\left(\begin{bmatrix} D & \mathbf{w} \end{bmatrix}\right)}{\det\left(C\right)}.$$

We then take $\eta = \eta_b \det\left(C\right)/\det\left(\begin{bmatrix} D & \mathbf{w} \end{bmatrix}\right)$, where η_b becomes a parameter of the sampler.

We propose moves of type b) and c) with probabilities $p_{b,j}$ and $p_{c,j}$, respectively, which depend on j, p_a and p_d ($0 < p_a + p_d < 1$): $p_{b,j} = (1 - p_a - p_d)/2$, $p_{c,j} = (1 - p_a - p_d)/2$, $j \geq q$ and $p_{b,q-1} = (1 - p_a - p_d)$, $p_{c,q-1} = 0$; for the fixed knots prior take $p_d = 0$. These choices make sure if there are no inner knots in the current state, no

knot is removed. For moves of type b), the acceptance probability of the proposed state $\left(j+1, \boldsymbol{\kappa}, \boldsymbol{\vartheta}\right)$ is as follows:

$$\min\left(1, \frac{\tilde{\pi}\left(j+1, \boldsymbol{\kappa}, \boldsymbol{\vartheta} \mid \mathbf{X}^{(n)}\right) p_{c,j+1}}{\tilde{\pi}\left(j, \mathbf{k}, \boldsymbol{\theta} \mid \mathbf{X}^{(n)}\right) p_{b,j} \varphi_1(\rho)} \eta_b\right) \quad \text{and}$$
$$\min\left(1, \frac{\tilde{\pi}\left(j+1, \boldsymbol{\kappa}, \boldsymbol{\vartheta} \mid \mathbf{X}^{(n)}\right) p_{c,j+1} (j-q+1)^{-1}}{\tilde{\pi}\left(j, \mathbf{k}, \boldsymbol{\theta} \mid \mathbf{X}^{(n)}\right) p_{b,j} \varphi_1(\rho)} \eta_b\right),$$

for the fixed knots prior and the free knots prior, respectively. Moves of type c) are simply the reverse move to a move of type b), so we omit the details. For this type of move, we remove one knot from the current state of the chain, uniformly at random, and recompute the spline coefficients via the inverse of the mapping g.

We let both reversible jump MCMC samplers run for the same number of iterations, both starting from the state $\left(40, (1/37, 2/37, \ldots, 36/37), \mathbf{0}\right)$ which corresponds to a constant function equal to zero with 38 equally spaced inner knots. We then collect 10.000 states from the each chain. The results of the MCMC procedures are given in Figures 4.6.2 and 4.6.3. In both priors we use cubic splines ($q = 4$) and $n = 1000$. We take for the fixed knots prior $v = 40$, $M = 15$, $p_a = 0.5$, $p_d = 0$, $\eta_a = 1.12 \times 10^{-1}$, $\eta_b = 3 \times 10^{-2}$ and $\eta_d = 0$. For the free knots prior, we choose $v = 40$, $M = 15$, $p_a = 0.66$, $p_d = 0.33$, $\eta_a = 1.18 \times 10^{-1}$, $\eta_b = 6 \times 10^{-3}$ and $\eta_d = 4 \times 10^{-2}$.

For both priors, we compute the proposed spline coefficients in the same way (described above), for the sake of comparing their performance. This is, however, not strictly necessary for the free knots prior. In this case, the insertion of a new knot only has a local effect on the spline: if all coefficients are kept the same, it is simple to propose a reasonable procedure for the new coefficient associated with the newly added B-spline. In case of the free knots prior, adding and removing knots from the current state of the chain can be made in a straightforward and computationally efficient way which does not involve recomputing all of the coefficients of the spline in the proposal.

As the simulations results in Figures 4.6.2 and 4.6.3 show, the free knots prior seems to outperform the fixed knots prior: the free knots posterior detects better the high and low variability regions of the regression function and facilitates the placement of more knots in the high variability region. In its turn, the fixed knots prior uses roughly 25% more knots to actually achieve a worse fit: 29 knots for the fixed knots posterior against around 23 knots for the free knots posterior. The fixed knots posterior fails to assign a number of knots that is compatible with the (inhomogeneous) structure of the true regression function f over the whole interval $[0, 1]$. As a consequence, the posterior seems to compromise on a number of knots which is clearly not sufficient for the high variability region close to zero (resulting in oversmoothing) and excessive for the low variability region close to 1 (undersmoothing the data).

Bayesian analysis based on the free knots prior has the advantage of providing relevant information about how the posterior chooses to place the knots. The bottom display of Figure 4.6.3 clearly shows a concentration of knots close to 0. This concentration, accompanied with the wider credible bands in the top display, suggests that the regression function is more variable ("volatile") in this region. This can be used to make an inference on the variability (smoothness inhomogeneity, volatility) of the underlying function and to try and improve estimation procedures.

In fact, this leads the following data-driven, empirical procedure for selecting a more appropriate prior on

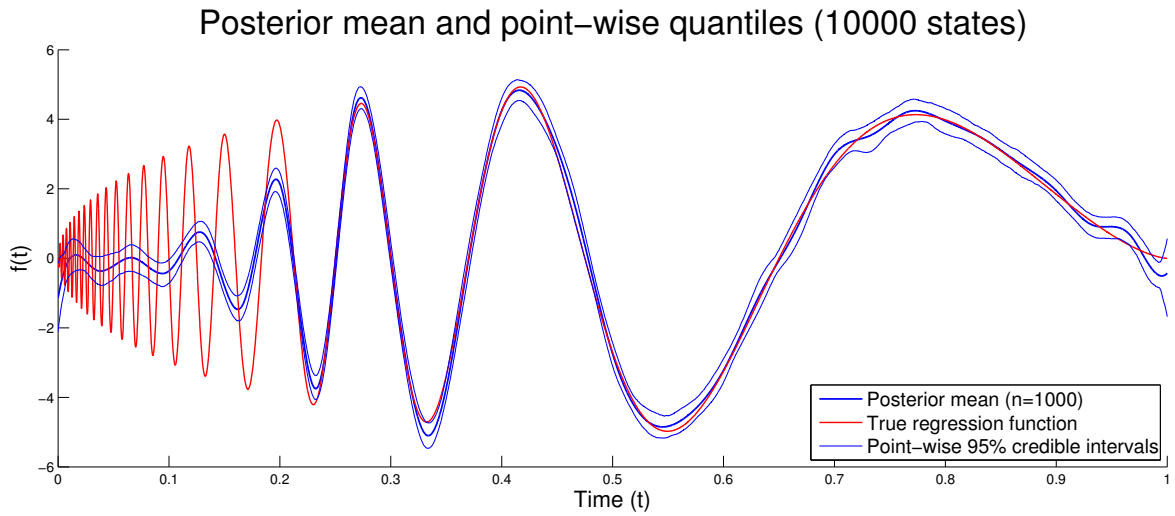

Figure 4.6.2: Results of the MCMC sampler for the fixed knots prior: posterior mean and respective 95% point-wise credible bands (in blue) and the true regression function (in red).

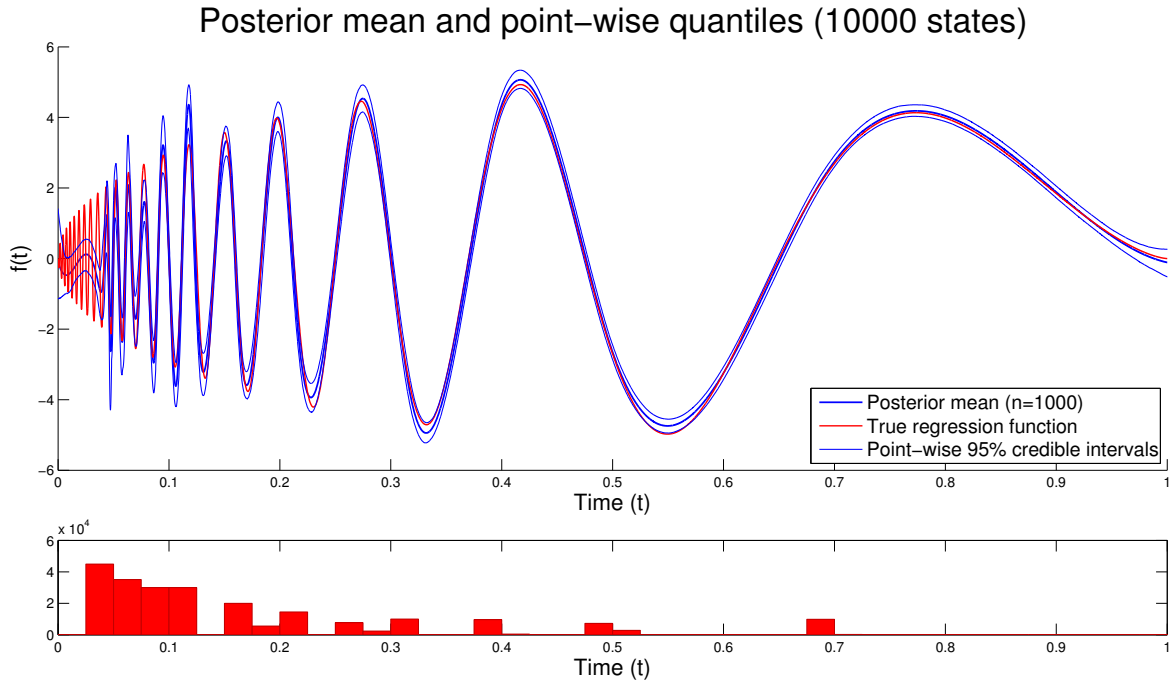

Figure 4.6.3: Results of the MCMC sampler for the free knots hierarchical prior. Above: posterior mean and respective 95% point-wise credible bands (in blue) and the true regression function (in red). Below: histogram of all the knots in all the sampled states.

the location of the knots: sample $j - q$ knots, i.i.d., from the empirical knot distribution presented in the bottom display of Figure 4.6.3 instead of our original prior on the knots. Actually, since some regions of the support had

no knots, we mixed the distribution presented in the histogram in Figure 4.6.3 with a uniform distribution: where before the numbers U_i used to define our prior on the knots were uniform on $\big[0, 1-(j-q+1)\delta(j)\big)$ they are now this mixture. This was done to give positive mass to the knot locations over the entire support of the regression function thus facilitating the mobility of the knots. We re-ran the MCMC procedure using such a prior on the knots. The results are given in Figure 4.6.4. This data-driven prior, at least in our numerical study, does not seem

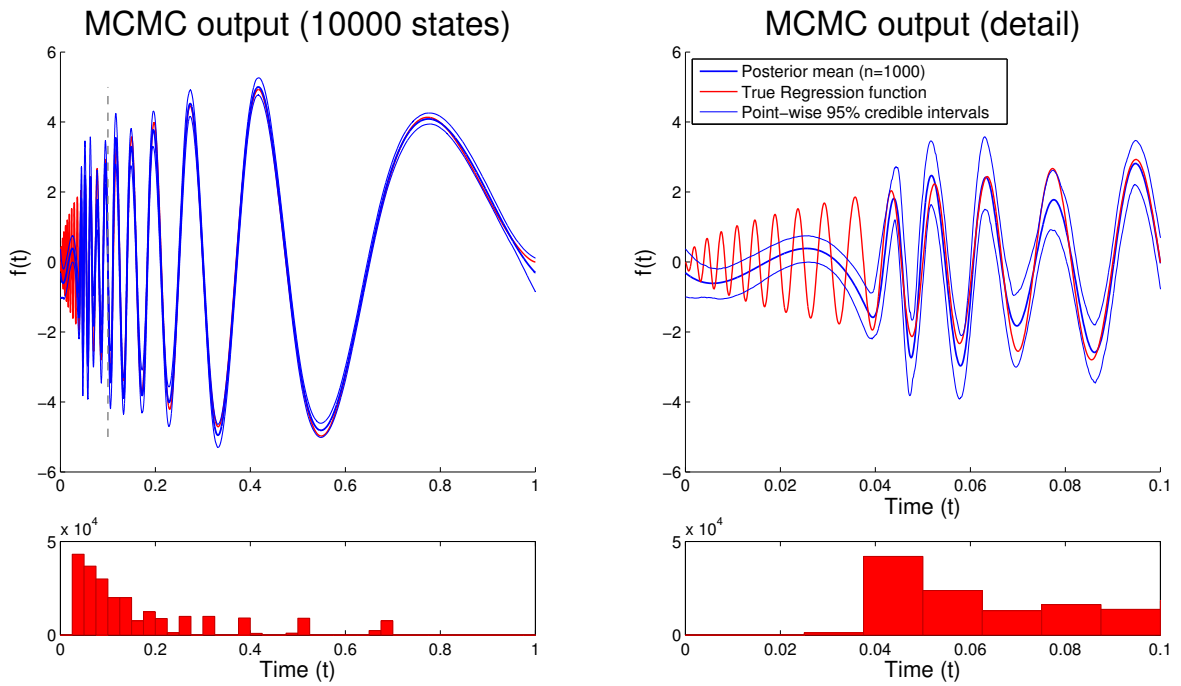

Figure 4.6.4: Results of the MCMC sampler for the free knots prior with a data-driven prior on the knots locations. On the left side we display the same figures as on the right side, but on the interval $[0, 0.1]$. Above: posterior mean and respective 95% point-wise credible bands (in blue) and the true regression function for comparison (in red). Below: histogram of all the knots in all the sampled states.

to improve significantly upon the free knots prior. This might simply mean that the free knots prior is already managing the location of the spline knots adequately, and reinforcing this via a data-driven prior does not give an extra advantage in the inference procedure. Note that Theorem 4.1 may still be applied to such a data-driven prior so that the resulting posterior retains (at least) the same theoretical properties as the free knots posterior.

Remark 4.6 *To summarize the above discussion, one can obtain the two stage sampler via the following procedure: a) split the dataset into two (independent) collections of observations; b) run the MCMC procedure on half of the data to obtain the posterior for the knots and use this to construct an empirical distribution for the knots; c) construct a new prior, using the empirical distribution of the knots obtained from the first sampler (mixed with any prior distribution on knots for which (4.6) and (4.7) hold, with a small, yet large enough weight in the mixture); d) run the MCMC sampler on the remaining data with the prior described in c).*

4.7 Technical results

In this section we collect some technical results. Lemmas 4.1 and 4.2 are needed to bound the entropy number of the sieves $\mathcal{S}_n$ in Theorem 4.1. Lemma 4.3 claims in essence that if some bounds on the range of the function f_0 are known, then this knowledge can be incorporated into the prior on the coefficients $\boldsymbol{\theta}$.

Theorem 4.26 of [81] claims that if all the inner knots of a B-spline are simple, then the B-spline is continuous, uniformly over its support, with respect to its knots. In Lemma 4.2 we establish a slightly stronger result (a Lipschitz-type property): if we take two splines with the same coefficients in their respective B-spline basis, then the L_∞ distance between the splines can be bounded by a multiple of the l_∞ distance between the two sets of knots, as long as the sets of knots are sufficiently sparse. First, we present a preliminary lemma. Denote the $(r+1)$-th order divided difference of a function h over the points $t_1 \le \cdots \le t_{r+1}$ as $[t_1,\dots,t_{r+1}]h = ([t_2,\dots,t_{r+1}]h - [t_1,\dots,t_r]h)/(t_{r+1}-t_1)$, with $[t_i]h = h(t_i)$. If $t_1 = \cdots = t_{r+1}$, then define $[t_1,\dots,t_{r+1}]h = h^{(r)}(t_1)/r!$ for a function h with enough derivatives at t_1.

Lemma 4.1 *Let $i \in \{1,\dots,r\}$, $r \ge 2$, $(k_1,\dots,k_{r+1}) \in (0,1)^{r+1}$. Assume $k_{\nu+1} - k_\nu \ge \delta > 0$ for $\nu = 0,\dots,i-1,i+1,\dots,r$ and $k_{i+1} - k_i = 0$. For fixed $x \in [0,1]$ take the function $h(y) = (x-y)_+^{q-1}$ with $y \in [0,1]$ and $q \ge 2$. Then the divided difference $\left|[k_1,\dots,k_{r+1}]h\right| \le 4/\delta^r$ for $x \ne k_i$ and any $\delta \le 2/(q-1)$.*

Proof: Notice that $|h'(y)| = (q-1)(x-y)_+^{q-2} \le (q-1) \le 2/\delta$ for $x \ne y$, as $q \ge 2$ and $\delta \le \frac{2}{q-1}$. Next, if $\nu = i-1$, $\left|[k_{\nu+1},k_{\nu+2}]h\right| = |h'(k_{\nu+1})| \le 1/\delta$; if $\nu \ne i-1$, $\left|[k_{\nu+1},k_{\nu+2}]h\right| = |h(k_{\nu+2}) - h(k_{\nu+1})|/|k_{\nu+2} - k_{\nu+1}| \le 2/\delta$. We conclude $\left|[k_{\nu+1},k_{\nu+2}]h\right| \le 2/\delta$ as long as $x \ne k_i$.

For $j = 2,\dots,r$, define $\gamma_j = \min_{\nu=1,\dots,r+1-j} |k_{\nu+j} - k_\nu| \ge (j-1)\delta$. Now we make use of Theorem 2.56 from [81] and the previous bound:

$$\left|[k_1,\dots,k_{r+1}]h\right| \le \sum_{\nu=0}^{r-1} \binom{r-1}{\nu} \frac{\left|[k_{\nu+1},k_{\nu+2}]h\right|}{\gamma_2 \dots \gamma_r} \le \frac{2^r}{(r-1)!\delta^r} \le \frac{4}{\delta^r}$$

holds for all $x \ne k_i$. This completes the proof of the Lemma. □

Recall that $s_{\mathbf{k},\boldsymbol{\theta}}(x)$, $x \in [0,1]$, is a spline of order $q \ge 2$ with the coordinates $\boldsymbol{\theta}$ in the B-spline basis and the inner knots vector $\mathbf{k}$.

Lemma 4.2 *Let $\boldsymbol{\theta} \in \mathbb{R}^j$ satisfies $\|\boldsymbol{\theta}\|_\infty \le M$ and let $\mathbf{k},\boldsymbol{\kappa} \in \mathcal{K}_j^\delta = \{\mathbf{k} \in \mathcal{K}_j : m(\mathbf{k}) \ge \delta\}$. Then $\|s_{\mathbf{k},\boldsymbol{\theta}} - s_{\boldsymbol{\kappa},\boldsymbol{\theta}}\|_\infty \le L\|\mathbf{k} - \boldsymbol{\kappa}\|_\infty$, for $L = 4j(q+1)M\delta^{-(q+1)}$ and any $\delta \le 2/(q-1)$.*

Proof: Define $\mathbf{k}^l = (k_1^l,\dots,k_{j-q}^l) = (\kappa_1,\dots,\kappa_l,k_{l+1},\dots,k_{j-q})$ for $l = 0,\dots,j-q$, such that $\mathbf{k}^0 = \mathbf{k}$ and $\mathbf{k}^{j-q} = \boldsymbol{\kappa}$.

By (4.1) and the triangle inequality, we get

$$\begin{aligned}\left\|s_{\mathbf{k},\boldsymbol{\theta}} - s_{\boldsymbol{\kappa},\boldsymbol{\theta}}\right\|_\infty &= \left\|\sum_{i=1}^{j}\theta_i B_i^{\mathbf{k}^0} - \sum_{i=1}^{j}\theta_i B_i^{\mathbf{k}^{j-q}}\right\|_\infty \le M\left\|\sum_{i=1}^{j}(B_i^{\mathbf{k}^0} - B_i^{\mathbf{k}^{j-q}})\right\|_\infty \\ &\le jM\max_{1\le i\le j}\left\|B_i^{\mathbf{k}^0} - B_i^{\mathbf{k}^{j-q}}\right\|_\infty \le jM\max_{1\le i\le j}\sum_{l=0}^{j-q-1}\left\|B_i^{\mathbf{k}^l} - B_i^{\mathbf{k}^{l+1}}\right\|_\infty \\ &\le (q+1)jM\max_{1\le i\le j}\max_{0\le l\le j-q-1}\left\|B_i^{\mathbf{k}^l} - B_i^{\mathbf{k}^{l+1}}\right\|_\infty.\end{aligned}$$

The last inequality in the above display follows from (4.1). Indeed, the inner knots of $B_i^{\mathbf{k}^l}$ and $B_i^{\mathbf{k}^{l+1}}$ differ only at the $(l+1)$-th entry. Therefore, according to (4.1), for each i there are at most $(q+1)$ nonzero terms $\left\|B_i^{\mathbf{k}^l} - B_i^{\mathbf{k}^{l+1}}\right\|_\infty$ in the last sum.

Theorem 4.27 of [81] gives explicit expressions for the derivative of a B-spline with respect to one of its knots. These expressions are in terms of the divided differences which satisfy the conditions of Lemma 4.1, so that combining this with Lemma 4.1 for $r = q+1$ (the maximal number of knots in the support of a B-spline) yields that this derivative is bounded in absolute value by $4\delta^{-(q+1)}$, except at $x = k_{l+1}^l$, where it is not defined. Then, as $\|\mathbf{k}^l - \mathbf{k}^{l+1}\|_\infty \le \|\mathbf{k} - \boldsymbol{\kappa}\|_\infty$, we obtain that, for $x \ne k_{l+1}^l$, $l = 0, \dots, j-q-1$,

$$\left|B_i^{\mathbf{k}^l}(x) - B_i^{\mathbf{k}^{l+1}}(x)\right| \le |k_{l+1}^{l+1} - k_{l+1}^l| \sup_{k_{l+1}^l \in (0,1)}\left|\frac{\partial B_i^{\mathbf{k}^l}(x)}{\partial k_{l+1}^l}\right| \le \frac{4\|\mathbf{k} - \boldsymbol{\kappa}\|_\infty}{\delta^{q+1}}.$$

Since splines are continuous for all $q > 1$, so is $s_{\mathbf{k},\boldsymbol{\theta}} - s_{\boldsymbol{\kappa},\boldsymbol{\theta}}$ and we conclude that the same bound must also hold for $x = k_{l+1}^l$. Combining the above two relations concludes the proof. □

The properties of B-splines allow to relate the range of the coefficients of the approximating spline to the range of the approximated function. The following lemma generalizes Lemma 1 of [86] for non-equally spaced knots.

Lemma 4.3 *Let $f \in \mathcal{F}_\alpha$ (so that (4.3) holds), $a < b$, $\varepsilon > 0$. Assume that $f(x) \in [a+\varepsilon, b-\varepsilon]$ for all $x \in [0,1]$. Then there exists a positive constant $\delta = \delta(\mathcal{F}_\alpha, \varepsilon)$ such that for any $\mathbf{k} \in \mathcal{K}_j$, $j \ge q$, such that $M(\mathbf{k}) \le \delta$, the coefficients $\mathbf{a}$ of the approximating spline $s_{\mathbf{k},\mathbf{a}}$ in (4.3) can be taken to be contained in (a,b).*

Proof: Fix q, j and inner knots $\mathbf{k}$, assume $I = [a,b]$, $a < b$ and $a + \varepsilon < f < b - \varepsilon$, for some $\varepsilon > 0$.

We use results from Section 4.6 of [81] on dual basis of B-splines. If $\{B_{\mathbf{k},1}, \dots, B_{\mathbf{k},j}\}$ is the B-spline basis associated with the inner knots $\mathbf{k}$, then there exists a dual basis $\lambda_1, \dots, \lambda_j$ of linear functionals such that, for each $i, r = 1, \dots, j$, $\lambda_r B_{\mathbf{k},i} = 1$ if $i = r$ and is 0 otherwise. As a consequence, we obtain that $\lambda_i s_{\mathbf{k},\mathbf{a}} = a_i$, and since $\sum_{i=1}^j B_{\mathbf{k},i}(x) = 1$, it follows that $\lambda_i c = c$ for any constant c and all $i = 1, \dots, j$. This dual basis is not necessarily unique and, according to Theorem 4.41 from [81], can be taken such that $|\lambda_i f| \le C_1 \sup_{x \in I_i} |f(x)|$ where I_i represents the support of $B_{\mathbf{k},i}$ and constant C_1 depends only on q. Each I_i consists of at most q adjacent intervals in the partition induced by $\mathbf{k}$ and thus the length of I_i is bounded by $qM(\mathbf{k})$.

Let $s_{\mathbf{k},\mathbf{a}}$ be such that (4.3) is fulfilled for f. Then, for any constant c,

$$\begin{aligned} |a_i - c| &= \left|\lambda_i s_{\mathbf{k},\mathbf{a}} - \lambda_i f + \lambda_i f - c\right| \le \left|\lambda_i (s_{\mathbf{k},\mathbf{a}} - f)\right| + \left|\lambda_i (f - c)\right| \\ &\le C_1 C_f M^{\alpha}(\mathbf{k}) + C_1 \sup_{x \in I_i} |f(x) - c|. \end{aligned}$$

Take $c = \inf_{x\in I_i} f(x)$ and recall that $f \in \mathcal{F}_\alpha \subseteq \mathcal{L}(\kappa_\alpha, L_\alpha)$. Using the Lipschitz property, we derive that $\sup_{x\in I_i} |f(x) - c| = \sup_{x\in I_i} f(x) - \inf_{x\in I_i} f(x) \le L_\alpha (q\, M(\mathbf{k}))^{\kappa_\alpha}$ and therefore

$$|a_i - \inf_{x\in I_i} f(x)| \le C_1 C_f M^{\alpha}(\mathbf{k}) + C_1 L_\alpha (q\, M(\mathbf{k}))^{\kappa_\alpha} \le C_2 M^{\alpha\wedge\kappa_\alpha}(\mathbf{k}).$$

In the same way, if we take $c = \sup_{x\in I_i} f(x)$, we derive that $\sup_{x\in I_i} |f(x) - c| \le L_\alpha (q\, M(\mathbf{k}))^{\kappa_\alpha}$ and thus $\left|a_i - \sup_{x\in I_i} f(x)\right| \le C_2 M^{\alpha\wedge\kappa_\alpha}(\mathbf{k})$.

Now for $\delta = (\varepsilon/(2C_2))^{1/(\alpha\wedge\kappa_\alpha)}$ we conclude that if $M(\mathbf{k}) \le \delta$, then $a_i \ge \inf_{x\in I_i} f(x) - C_2 M^{\alpha\wedge\kappa_\alpha}(\mathbf{k}) \ge \inf_{x\in I_i} f(x) - \varepsilon/2 > a$. For the same choice of δ we have $a_i \le \sup_{x\in I_i} f(x) + C_2 M^{\alpha\wedge\kappa_\alpha}(\mathbf{k}) \le \sup_{x\in I_i} f(x) + \varepsilon/2 < b$. □

Part II

Tracking

5

Tracking of Conditional Quantiles

We consider the problem of constructing an on-line, recursive algorithm for tracking a conditional quantile in a general setting. The observations are assumed to be a time series, whose terms need not be independent. We propose a recursive algorithm to track a conditional quantile of a level of our choice; both the level of the quantile and the quantile itself do not need to be fixed. We establish an upper bound for the error of the algorithm, which we then specify for different conditions on the variability of the quantile function. From these results we derive the convergence rates for the considered examples and present some numerical results.

5.1 Introduction

Often in applications one wishes to recover a functional dependence between different parameters of the underlying distribution based on observations from that distribution. Non-parametric model regression is one common approach to this problem. Strictly speaking, in regression analysis we are interested in estimating the conditional expectation of one random variable given another one. Thus, certain moment conditions in regression models are unavoidable. Moreover, in additive regression, stronger structural conditions on the noise are usually imposed, for example, normality of errors. However, sometimes it is desirable to minimize the conditions on the moments (and the form) of the distribution of the noise. If, for example, we only assume that the error at each moment has a quantile of certain fixed level (for identifiability purposes), then we obtain the so-called quantile regression model, first introduced into the literature in [2]; see [55] for a nice account on this topic.

The quantile regression model is quite important in fields such as econometrics, social sciences and ecology. There, one often studies response variables whose relation with its measured predictors is complex. In these cases the conditional expectation of the response variable might simply be insensitive to these relations and will provide a poor description of the underlying phenomenon. Error bounds on certain regression estimates can be viewed as crude quantile regressions [92] but these quantiles can be estimated directly. By estimating conditional quantiles of different levels, rather than the conditional mean we get a much more comprehensive and robust description of the data. This seems to be of particular relevance in applications, for example in ecology, where data often displays heterogeneous variances [20].

In this chapter we treat the problem of recovering a quantile regression function in an on-line fashion. Precisely, suppose that at each time point $k \in \mathbb{N}$ we observe a random variable X_k and the problem is to recover its quantile of level $\alpha_k \in (0,1)$ by using observations available at this time moment. An important complicating factor of our framework is that we do not assume the traditional independence of the observations. In fact, the observations can be arbitrarily dependent so that by the time moment k we have observed $\mathbf{X}_k = (X_1, \dots, X_k)$ and we would like to recover the conditional α_k-quantile of X_{k+1} given $\mathbf{X}_k$.

Moreover, it is desirable to design an algorithm such that the estimate of the quantile at the current time moment is based on the estimate of the quantile at the previous time moment and a small correction based on the current observation. This allows us to bypass the need to use optimization algorithms to recuperate the quantile regression function which is standard in quantile regression and brings us to the area of stochastic approximation algorithms. The idea of stochastic approximation algorithm was first proposed by [76] and since then a huge amount of literature has appeared on this topic (cf. Chapter 6). These algorithms run in parallel with the collection of data and are driven by a *gain function*. The gain, when properly rescaled by a so-called *step size*, can be used to improve any approximation for a quantity of interest (a conditional quantile in our case). Sequential application of this procedure results in a recursive, tracking algorithm.

5.2 Preliminaries

Suppose that at each time $k \in \mathbb{N}$ we observe a random variable X_k so that by a time moment n we have n observations $X_1, \dots, X_n$. Denote $\mathbf{X}_k = (X_1, \dots, X_k)$, $k \in \mathbb{N}$, with the convention that $\mathbf{X}_0$ is an empty vector and let $\mathcal{X}$ represent the (common) support of each observation. Hereafter, vectors are represented by bold symbols and can be upper or lowercase letters.

For some known fixed values $\alpha_k \in (0,1)$, $k \in \mathbb{N}$, let $\theta_k = \theta_k(\mathbf{x}_{k-1}) = \theta_k(\mathbf{x}_{k-1}, \alpha_k)$, the α_k-quantile of the conditional distribution of X_k given $\mathbf{X}_{k-1} = \mathbf{x}_{k-1} = (x_1, \ldots, x_{k-1})$:

$$\theta_k(\mathbf{x}_{k-1}) = \inf\left\{\theta : \mathbb{P}(X_k \le \theta | \mathbf{X}_{k-1} = \mathbf{x}_{k-1}) \ge \alpha_k\right\}.$$

Let $F_k(x_k|\mathbf{x}_{k-1})$ denote the (unknown) conditional distribution function of $X_k - \theta_k(\mathbf{x}_{k-1})$ given $\mathbf{X}_{k-1} = \mathbf{x}_{k-1}$. Thus, $F_k(0|\mathbf{x}_{k-1}) = \alpha_k$, $k \in \mathbb{N}$.

Our goal is, loosely speaking, to track down the conditional quantile θ_k based on the information available to us at time k. More precisely, at each time moment $k \in \mathbb{N}$ we want to estimate $\theta_k(\mathbf{X}_{k-1})$ by using the observations $\mathbf{X}_k = (X_1, \ldots, X_k)$ available at that moment. Ideally, we would like our procedure to approach the evolving conditional quantile $\theta_k(\mathbf{X}_{k-1})$ as time progresses. If, however, this is impossible, then the procedure should at least stay in proximity of $\theta_k(\mathbf{X}_{k-1})$, the evolving conditional quantile, as close as possible. Until now we have not imposed any assumption on the observations $X_1, X_2, \ldots$, these are arbitrarily distributed and have an arbitrary dependence structure. Clearly, the stated problem in its full generality has no feasible solution. Thus, in order to obtain some non-void results, we need to impose some assumptions on the conditional distributions of X_k (given $\mathbf{X}_{k-1} = \mathbf{x}_{k-1}$), $k \in \mathbb{N}$, while at the same time trying to keep these conditions as weak as possible.

Now we are ready to introduce the conditions on the conditional distributions $F_k(x_k|\mathbf{x}_{k-1})$ which we are going to use in the derivation of the main result.

(C1) For some positive b, B, δ, the following inequality holds for any $\varepsilon \in [-\delta, \delta]$:

$$b|\varepsilon| \le \inf_{\mathbf{x}_{k-1} \in \mathcal{X}^{k-1}} |F_k(\varepsilon|\mathbf{x}_{k-1}) - \alpha_k| \le \sup_{\mathbf{x}_{k-1} \in \mathcal{X}^{k-1}} |F_k(\varepsilon|\mathbf{x}_{k-1}) - \alpha_k| \le B|\varepsilon|, \quad k \in \mathbb{N}.$$

(C2) The conditional quantiles $\theta_k(\mathbf{X}_{k-1})$ take values in some compact set Θ so that

$$\sup_{k \in \mathbb{N}} \sup_{\mathbf{x}_{k-1} \in \mathcal{X}^{k-1}} \left|\theta_k(\mathbf{x}_{k-1})\right| \le \sup_{\theta \in \Theta} |\theta| = C_\Theta,$$

for some (known) constant C_Θ.

Condition (C1) is fulfilled if, for example, the conditional distributions $F_k(x_k|\mathbf{x}_{k-1})$, are absolutely continuous with the conditional densities $f_k(x_k|\mathbf{x}_{k-1})$ such that for some positive b, B, δ,

$$0 < b \le \inf_{\mathbf{x}_{k-1} \in \mathcal{X}^{k-1}} f_k(x_k|\mathbf{x}_{k-1}) \le \sup_{\mathbf{x}_{k-1} \in \mathcal{X}^{k-1}} f_k(x_k|\mathbf{x}_{k-1}) \le B, \quad k \in \mathbb{N}.$$

for almost all (with respect to the Lebesgue measure) $x_k \in [-\delta, \delta] \cap \mathcal{X}$.

If $\mathcal{X}$ is a discrete set then condition (C1) is inadequate. We require instead

(D1) For positive constants $c < 1 < \delta$, the following inequality holds for any $x_k \in [-\delta, \delta] \cap \mathcal{X}$ with $\mathcal{X} = \mathbb{Z}$ (w.l.g.):

$$0 < c \le \inf_{\mathbf{x}_{k-1} \in \mathcal{X}^{k-1}} p_k(x_k|\mathbf{x}_{k-1}) \le \sup_{\mathbf{x}_{k-1} \in \mathcal{X}^{k-1}} p_k(x_k|\mathbf{x}_{k-1}), \quad k \in \mathbb{N}.$$

Remark 5.1 *Notice that, even under the above conditions, we deal with a rather general framework: the obser-*

vations can be dependent and not identically (marginally) distributed. Besides, our problem is stated in the robust setting, in the sense that we do not assume anything about the moments of the observations X_k – they simply may not exist.

Conditions (C1), (D1) and (C2) are rather natural for the most important particular case of independent observations X_k, $k \in \mathbb{N}$. In this case the conditional α_k-quantiles θ_k become unconditional (θ_k does not depend on $\mathbf{X}_k$) and bounded uniformly in k according to condition (C2). The observations can then be expressed in the form $X_k = \theta_k + \xi_k$, $k \in \mathbb{N}$, with independent noises ξ_k. Condition (C1) means that the noises ξ_k have zero α_k-quantiles respectively and their probability distributions behave regularly in the neighborhood of zero in the sense that they degenerate neither into zero nor into delta-function. Alternatively, when $\mathcal{X}$ is discrete, (D1) requires the distribution function of the noises to jump by at least c close to zero.

Conditions (C1), (D1) and (C2) do not seem too restrictive for another important case of Markov model observations: in this case the conditional density f_k depends only on two arguments x_k, x_{k-1}.

Introduce the indicator function $\mathbb{1}\{A\}$ of a set A, the function $\operatorname{sign}(x) = x/|x|$ for $x \neq 0$ and $\operatorname{sign}(0) = 1$, and the function

$$\bar{S}_k(u, v) = \bar{S}_k(u, v, \alpha_k) = \alpha_k - \mathbb{1}\{u - v < 0\} + c/2, \quad k \in \mathbb{N}. \tag{5.1}$$

where c is the constant from condition (D1) if $\mathcal{X}$ is discrete and $c = 0$ otherwise.

Let $\{\gamma_k,\ k \in \mathbb{N}\}$ be a nonnegative sequence bounded by some constant Γ:

$$0 \le \gamma_k \le \Gamma, \quad k \in \mathbb{N}. \tag{5.2}$$

Next, for some positive, fixed κ introduce the constant

$$\bar{C} = C_\Theta + \Gamma(1 + c/2) + \kappa, \tag{5.3}$$

where constants C_Θ and Γ are from conditions (C2) and (5.2), respectively, and as before, c is the constant from condition (D1) if $\mathcal{X}$ is discrete and $c = 0$, otherwise.

We are now ready to define our algorithm for tracking a conditional quantile:

$$\hat{\theta}_k = \Pi_{\bar{C}}\big(\hat{\theta}_{k-1} + \gamma_k \bar{S}_k(X_k, \hat{\theta}_{k-1})\big), \quad \hat{\theta}_0 \in \Theta,\ k \in \mathbb{N}, \tag{5.4}$$

where $\hat{\theta}_0 \in \Theta$ some initial value[1], the sequence of *step sizes* $\{\gamma_k,\ k \in \mathbb{N}\}$ satisfies the restriction (5.2), the constant $\bar{C}$ is defined by (5.3) and $\Pi_d x = [x]_{-d}^{d} = x\,\mathbb{1}\{|x| \le d\} + d\,\operatorname{sign}(x)\,\mathbb{1}\{|x| > d\}$ is the projection operator on the interval $[-d, d]$.

If $\mathcal{X}$ is a discrete set, then we define $\Pi_{\mathcal{X}} x$, the projection operator on $\mathcal{X}$, as $\Pi_{\mathcal{X}} x = \arg\min_{y \in \mathcal{X}} |x - y|$. We then use a modified version of (5.4),

$$\hat{\theta}_k = \Pi_{\bar{C}}\big(\hat{\theta}_{k-1} + \gamma_k \bar{S}_k(X_k, \Pi_{\mathcal{X}}\hat{\theta}_{k-1})\big), \quad \hat{\theta}_0 \in \Theta,\ k \in \mathbb{N}. \tag{5.5}$$

[1] We can take for example $\hat{\theta}_0 = 0$.

By definition, the sequence $\{\hat{\theta}_k,\ k \in \mathbb{N}\}$ is trivially bounded :

$$|\hat{\theta}_k| \le \bar{C}, \quad k \in \mathbb{N}. \tag{5.6}$$

It is easy to see that the algorithm (5.4) can be rewritten as

$$\hat{\theta}_k = \hat{\theta}_{k-1} + \gamma_k S_k(X_k, \hat{\theta}_{k-1}), \quad \hat{\theta}_0 \in \Theta,\ k \in \mathbb{N}, \tag{5.7}$$

where the function $S_k(u, v)$ (which we call *shift function*) is defined as follows:

$$S_k(u, v) = \bar{S}_k(u, v) + \tilde{S}_k(u, v), \quad k \in \mathbb{N}. \tag{5.8}$$

Here $\bar{S}_k(u, v)$ is given by (5.1) and, for all $k \in \mathbb{N}$,

$$\tilde{S}_k(u, v) = \gamma_k^{-1}\big[\bar{C}\ \mathrm{sign}\left(v + \gamma_k \bar{S}_k(u, v)\right) - (v + \gamma_k \bar{S}_k(u, v))\big]\ \mathbb{1}\{|v + \gamma_k \bar{S}_k(u, v)| > \bar{C}\} \tag{5.9}$$

if $\gamma_k \neq 0$, and without loss of generality we put $\tilde{S}_k(u, v) = 0$ if $\gamma_k = 0$. In (5.9), the constant $\bar{C}$ is defined by (5.3) and the sequence $(\gamma_k,\ k \in \mathbb{N})$ satisfies the restriction (5.2).

5.3 Main results

In this section we formulate the main result of the chapter. We start with two technical lemmas which we shall need in the proof of the main theorem. For the sake of brevity, denote $\theta_k = \theta_k(\mathbf{X}_{k-1})$.

Lemma 5.1 *Let the functions $\tilde{S}_k(u, v)$, $k \in \mathbb{N}$, be defined by (5.9) and constant $\bar{C} = C_\Theta + \Gamma(1 + c/2) + \kappa$ by (5.3), where c is as in (D1) if $\mathcal{X}$ is discrete and $c = 0$ otherwise. Then $|\tilde{S}_k(u, v)| \le 1 + c/2$, uniformly over $u \in \mathbb{R}$ and $|v| \le \bar{C}$. Also, the relations*

$$\tilde{S}_k(u, v) = -\tilde{G}_k(\theta_k, u, v)(v - \theta_k), \quad k \in \mathbb{N}, \tag{5.10}$$

hold for some functions $\tilde{G}_k(\theta_k, u, v)$ such that $0 \le \tilde{G}_k(\theta_k, u, v) \le \kappa^{-1}$, uniformly over $u \in \mathbb{R}$, $\theta_k \in \Theta$ and $|v| \le \bar{C}$.

Lemma 5.2 *Let the functions $S_k(u, v)$, $k \in \mathbb{N}$, be defined by (5.8). Let c be as in condition (D2) for discrete $\mathcal{X}$ and $c = 0$ for continuous $\mathcal{X}$. Then $|S_k(u, v)| \le 2 + c$, $k \in \mathbb{N}$, uniformly over $u \in \mathbb{R}$ and $|v| \le \bar{C}$. Moreover, if conditions (C1), (C2) are fulfilled, then*

$$\mathbb{E}\big[S_k(X_k, v)\big|\mathbf{X}_{k-1}\big] = -G_k(\theta_k, v)(v - \theta_k), \quad k \in \mathbb{N}, \tag{5.11}$$

for some functions $G_k(\theta_k, v) = G_k(\theta_k, v, \mathbf{X}_{k-1})$ such that $0 < h \le G_k(\theta_k, v) \le H$ with probability 1, uniformly over $\theta_k \in \Theta$ and $|v| \le \bar{C}$ (the constants h, H depend on δ, C_Θ, $\bar{C}$ and κ; they also depend on either b and B for continuous $\mathcal{X}$ or on c for discrete $\mathcal{X}$.)

Remark 5.2 *An informal interpretation of Lemma 5.2 is as follows. Firstly, the shift function $S_k(X_k, v)$ gives the*

"right average direction" from v towards the conditional quantile value θ_k. Secondly, the "average length" of the shift $S_k(X_k, v)$ is a controlled multiple of the distance between v and the conditional quantile value θ_k.

The next theorem is the main result of this chapter.

Theorem 5.1 (Error bound)
Let conditions (C1), (C2) be satisfied, let the estimator $\hat{\theta}_k$ be defined by (5.4) with step sizes $\{\gamma_k,\ k \in \mathbb{N}\}$ satisfying (5.2) and let the constants h and H be from Lemma 5.2. Define $\delta_k = \hat{\theta}_k - \theta_k$ and $\Delta\theta_k = \theta_k - \theta_{k-1}$, $k \in \mathbb{N}$ (define $\theta_0 = 0$). Then for any $p \geq 1$, any $n, n_0 \in \mathbb{N}$ such that $n_0 \leq n$ and $\gamma_k H \leq 1$ for all $n_0 \leq k \leq n$, the following relation holds

$$\mathbb{E}|\delta_n|^p \leq C_1 \exp\left(-ph \sum_{k=n_0+1}^{n} \gamma_k\right) + C_2 \left(\sum_{k=n_0+1}^{n} \gamma_k^2\right)^{p/2} + C_3\, \mathbb{E}\left(\max_{n_0+1\leq k\leq n} \left|\theta_k - \theta_{n_0}\right|^p\right) + C_4\, \mathbb{E}\left(\sum_{k=n_0+1}^{n} \gamma_k |\Delta\theta_k|\right)^p \quad (5.12)$$

for some positive C_1, C_2, C_3 and C_4, constants which depend only on p, C_Θ, $\bar{C}$ and H.

The proofs of the theorem and the lemma are deferred to the last section.

Remark 5.3 *The upper bound (5.31) depends on the levels α_k, $k \in \mathbb{N}$ via the constants C_Θ, h and H. The closer α_k is to one (resp. zero), the bigger (resp. smaller) the quantile's value θ_k, and therefore the constant C_Θ becomes bigger. There is too little probability mass in a neighborhood of an extreme quantile, so condition (C1) is more difficult to fulfill as α_k gets closer to 1 (or zero). This makes constants δ and b smaller, which in turn makes constant h smaller and constant H bigger. These changes in C_Θ, h and H in turn lead to an increase of the final constants C_1, C_2, C_3 and C_4 in inequality (5.31). As for the dependence on p, as it appears from the proof of the theorem, the bigger p, the bigger the constants C_1, C_2, C_3 and C_4.*

Remark 5.4 *In the relation (5.38) below one can derive an alternative bound*

$$\max_{n_0+1\leq k\leq n} |R_k| \leq \max_{n_0+1\leq k\leq n} \left|\sum_{i=n_0+1}^{k} \gamma_i M_i(X_i, \theta_i, \hat{\theta}_{i-1})\right| + \sum_{i=n_0+1}^{n} e^{-h\gamma_i} |\Delta\theta_i|,$$

which would lead to an alternative final statement for the theorem:

$$\mathbb{E}|\delta_n|^p \leq C_1 \exp\left(-ph \sum_{k=n_0+1}^{n} \gamma_k\right) + C_2 \left(\sum_{k=n_0+1}^{n} \gamma_k^2\right)^{p/2} + C_5\, \mathbb{E}\left(\sum_{k=n_0+1}^{n} e^{-h\gamma_k} |\Delta\theta_k|\right)^p.$$

Remark 5.5 *By analyzing the proof of the theorem, one can see that the particular form of the shift function $S_k(u, v)$ is not important, it is the property (5.26) for the quantity $\mathbb{E}\big[S_k(X_k, v)|\mathbf{X}_{k-1}\big]$ and the fact that the $\hat{\theta}_k$, $k \in \mathbb{N}$, are bounded that are really needed in the proof. Therefore, any shift function $S_k(X_k, v)$ for which Lemma 5.2 holds and $\hat{\theta}_k$ are bounded will do the job. For example, Lemma 5.2 (and therefore Theorem 5.1) holds for the*

following shift function $S_k(u,v)$*:*

$$S_k(u,v) = \bar{S}_k(u,v)\,\mathbb{1}\{|v| \le C_\Theta + \delta\} - v\,\mathbb{1}\{|v| > C_\Theta + \delta\}, \quad k \in \mathbb{N}. \tag{5.13}$$

Thus, one can use the shift function (5.13) *in the algorithm* (5.7)*. Actually, we can establish Lemma* 5.2 *for the shift function* (5.13) *uniformly over all* $v \in \mathbb{R}$ *(instead of just uniformly over* $|v| \le \bar{C}$*).*

The result of the theorem – inequality (5.31) – is given in a non-asymptotic form as an explicit upper bound for the error of the algorithm and is in essence determined by the choice of a time moment n_0 and the step sizes γ_k, $k = n_0, \dots, n$.

If we analyze the right hand side of (5.31), then we see that the second term is small if the sum $\sum_{k=n_0+1}^{n} \gamma_k^2$ is small. This will hold if, for example, the series $\sum_{k\in\mathbb{N}} \gamma_k^2$ is convergent and n_0 is sufficiently large. On the other hand, in order to make the first term small the sum $\sum_{k=n_0+1}^{n} \gamma_k$ should be sufficiently large, which will hold if, for example, the series $\sum_{k\in\mathbb{N}} \gamma_k$ diverges and the difference between n_0 and n is large enough. These are the classical conditions for the step sizes of the Robbins-Monro type algorithms and well known in the literature. Intuitively, if the sum $\sum_{i=n_0+1}^{n} \gamma_i^2$ is small, then the algorithm can "approach" θ_n arbitrarily closely, and if the sum $\sum_{k=n_0+1}^{n} \gamma_k$ is big, then algorithm can "reach" any point $\theta \in \Theta$. The value n_0 and the difference between the time moments n_0 and n represent a "burn-in" time for the algorithm. Recall that the algorithm starts from an arbitrary point $\hat{\theta}_0$ and therefore some time is needed for the algorithm to get adjusted and to start to really track the drifting quantile parameter θ_n.

The third and the forth terms in the right hand side of (5.31) can be arbitrarily large in general if we do not impose conditions that regulate the evolution of the parameter θ_k, $k \in \mathbb{N}$. We discuss this problem in more detail in the next section, where we also consider examples of such conditions. The basic idea is as follows: the less the conditional quantile is allowed to vary, the better the tracking algorithm performs.

5.4 Applications of the main result

In this section we consider some examples of situations when we can apply Theorem 5.1. From now on, by c and C we denote universal constants which can be different in different expressions.

First of all note that we can write $X_k = \theta_k(\mathbf{X}_{k-1}) + \eta_k(\mathbf{X}_k)$, $k \in \mathbb{N}$, for η_k such that $\mathbb{P}(\eta_k(\mathbf{X}_k) \le 0 | \mathbf{X}_{k-1}) = \alpha_k$. Conditions (C1) and (C2) are satisfied if, for example, it is possible to write $\eta_k(\mathbf{X}_k) = \sigma_k(\mathbf{X}_k)\xi_k$, i.e.,

$$X_k = \theta_k(\mathbf{X}_{k-1}) + \sigma_k(\mathbf{X}_k)\xi_k, \quad k \in \mathbb{N},$$

and if the following requirements hold: for some constants v_0, v_1

$$0 < v_0 \le \inf_{\mathbf{x}_k \in \mathbb{R}^k} \sigma_k(\mathbf{x}_k) \le \sup_{\mathbf{x}_k \in \mathbb{R}^k} \sigma_k(\mathbf{x}_k) \le v_1 < \infty, \quad k \in \mathbb{N}; \tag{5.14}$$

the noise terms ξ_k are independent with densities $f_k(x)$, respectively, such that, for some constants δ, a and A,

$$0 < a \le \inf_{|x|\le\delta} f_k(x) \le \sup_{|x|\le\delta} f_k(x) \le A, \quad k \in \mathbb{N};$$

and

$$\sup_{\mathbf{x}_{k-1}\in\mathbb{R}^{k-1}} |\theta_k(\mathbf{x}_{k-1})| \le C_\Theta, \quad k \in \mathbb{N}.$$

These conditions become somewhat unnatural as k increases – the functions σ_k and θ_k have arguments of growing dimensions. However, the above conditions are reasonable if the observations $\{X_k, k \in \mathbb{N}\}$ form a Markov chain of order, say, p:

$$X_k = \theta_k(X_{k-1}, \dots, X_{k-p}) + \sigma_k(X_{k-1}, \dots, X_{k-p})\xi_k, \quad k \in \mathbb{N},$$

with some initial $X_0, X_{-1}, \dots, X_{1-p}$.

Of course, it is impossible to provide any bound for the quality of the recursive algorithm by using Theorem 5.1 if nothing is known about the behavior of the increments $\Delta\theta_k = \theta_k - \theta_{k-1}$, $k \in \mathbb{N}$.

5.4.1 Constant quantile

Apart from conditions (C1) and (C2), assume now the strong model condition $\theta_k(\mathbf{X}_{k-1}) = \theta_0$, almost surely, $k \in \mathbb{N}$, for some unknown constant quantile θ_0, (i.e., $\Delta\theta_k = 0$, $k \in \mathbb{N}$). In essence, we have a parametric setup. Clearly, in this case the third and the forth terms in the right hand side of (5.31) vanish. Take $\gamma_k = (C_\gamma \log k)/k$ and $n_0 = \lfloor qn \rfloor$ for some $q \in (0,1)$, where $\lfloor a \rfloor$ denotes a whole part of the number $a \in \mathbb{R}$. Let $n_0 \ge 2$, which is satisfied if $n \ge 2/q = N_q$. Then since for sufficiently large C_γ and $n \ge N_q$

$$\sum_{k=n_0+1}^{n} \gamma_k \ge C_\gamma \log n_0 \sum_{k=n_0+1}^{n} \frac{1}{k} \ge \frac{\log n}{2h},$$

the first term in the right hand side of (5.31) is bounded as follows:

$$C_1 \exp\left(- ph \sum_{k=n_0+1}^{n} \gamma_k \right) \le C_1 n^{-p/2}. \tag{5.15}$$

Using $\sum_{k=n_0+1}^{n} \gamma_k^2 \le c(\log n)^2 n^{-1}$, we bound the second term:

$$C_2 \left(\sum_{k=n_0+1}^{n} \gamma_k^2 \right)^{p/2} \le C\big(n^{-1/2} \log n\big)^p, \tag{5.16}$$

which leads to the parametric (up to a logarithmic term) convergence rate

$$\max_{n \ge N_q} \mathbb{E}\left(\frac{\sqrt{n}}{\log n} |\delta_n| \right)^p \le c. \tag{5.17}$$

Interestingly, we cannot get rid of the log factor in the above statement. In some sense, one can see this as "payment" for recursiveness and robustness. Indeed, the bound (5.17) holds for any moment $p \ge 1$, whereas no moment conditions were assumed for the original observations $\{X_k, k \in \mathbb{N}\}$.

Besides, from inequality (5.17) one can derive the convergence $\hat{\theta}_n \to \theta_0$ with probability 1 with the rate

$n^{-(1/2-\varepsilon)}$ for any $\varepsilon > 0$. Indeed, by using (5.17) for $p > \varepsilon^{-1}$ and the Markov inequality, we obtain that for any $c > 0$

$$\sum_{n=1}^{\infty} \mathbb{P}\big(n^{1/2-\varepsilon}|\hat{\theta}_n - \theta_0| > c\big) \le \sum_{n=1}^{\infty} \frac{n^{p/2-p\varepsilon}\mathbb{E}|\delta_n|^p}{c^p} \le C \sum_{n=1}^{\infty} \frac{(\log n)^p}{n^{p\varepsilon}} < \infty \tag{5.18}$$

and the convergence $\hat{\theta}_n \to \theta_0$ as $n \to \infty$ with probability 1 with the rate $n^{-(1/2-\varepsilon)}$ follows by the Borel-Cantelli Lemma.

5.4.2 Polynomially decreasing quantile increments

Suppose now that the increments of the conditional quantile $\Delta\theta_k = \theta_k - \theta_{k-1}$, $k \in \mathbb{N}$, satisfy the following restrictions

$$\mathbb{E}|\Delta\theta_k|^p \le \rho_k^p, \quad k \in \mathbb{N}, \tag{5.19}$$

for some $p \ge 1$ and some positive sequence ρ_k decreasing to zero. One can interpret this condition as a requirement for the p-th moment of the oscillations $\Delta\theta_k$ to "slow down" over time. Throughout this section, assume the polynomial restriction $\rho_k = c_\rho k^{-\beta}$ for some $c_\rho > 0$, $\beta \ge 0$.

Consider first the case $\beta \ge 3/2$. It turns out that in this case the oscillations of the sequence θ_k slow down so quickly that an appropriately chosen algorithm step of leads to the same quality as if function $\theta(t)$ were constant. Indeed, take γ_k and n_0 to be the same as in the case of constant quantile (see Section 5.4.1). Then the first and the second terms of (5.31) can be bounded in the same way as in (5.15) and (5.16), whereas the third and the forth terms are bounded by a multiple of $\mathbb{E}\big(\sum_{k=n_0+1}^{n}|\Delta\theta_k|\big)^p$. By the Hölder inequality, we evaluate

$$\begin{aligned}\mathbb{E}\Big(\sum_{k=n_0+1}^{n}|\Delta\theta_k|\Big)^p &\le (n-n_o)^{p-1}\sum_{k=n_0+1}^{n}\mathbb{E}|\Delta\theta_k|^p \le C(n-n_o)^p\rho_{n_0}^p \\ &\le c\big((n-n_0)n_0^{-\beta}\big)^p \le Cn^{-(\beta-1)p} \le Cn^{-p/2},\end{aligned} \tag{5.20}$$

which leads to the same bound as (5.17) with another constant c.

Now consider the case $0 < \beta < 3/2$. Let $\gamma_k = C_\gamma(\log k)^{1/3}k^{-2\beta/3}$, $n_0 = n - n^{2\beta/3}(\log n)^{2/3}$. By using the elementary inequality $(1+x)^\alpha \le 1 + \alpha x$ for $0 < \alpha < 1$ and $x \ge -1$, we obtain that for sufficiently large n (i.e., $n \ge N_1 = N_1(\beta)$) and sufficiently large constant C_γ

$$\begin{aligned}\sum_{k=n_0+1}^{n}\gamma_k &\ge C_\gamma(\log n_0)^{1/3}\sum_{k=n_0+1}^{n}\frac{1}{k^{2\beta/3}} \ge C_\gamma(\log n_0)^{1/3}\int_{n_0}^{n}\frac{dx}{x^{2\beta/3}} \\ &= \frac{C_\gamma(\log n_0)^{1/3}}{1-2\beta/3}\Big(n^{1-2\beta/3} - n^{1-2\beta/3}\big(1-n^{2\beta/3-1}(\log n)^{2/3}\big)^{1-2\beta/3}\Big) \\ &\ge \frac{C_\gamma(\log n_0)^{1/3}}{1-2\beta/3}\Big(n^{1-2\beta/3} - n^{1-2\beta/3}\big(1-n^{2\beta/3-1}(\log n)^{2/3}(1-2\beta/3)\big)\Big) \\ &= C_\gamma(\log n_0)^{1/3}(\log n)^{2/3} \ge \frac{\log n}{2h}.\end{aligned}$$

This yields the same bound for the first term of the right-hand side of the inequality (5.31), similar to (5.15): for

$n \geq N_1$ and sufficiently large constant C_γ

$$C_1 \exp\left(- ph \sum_{k=n_0+1}^{n} \gamma_k \right) \leq C_1 n^{-p/2}.$$

Let us bound the second term on the right-hand side of inequality (5.31): for $n \geq N_2 = N_2(\beta)$

$$C_2 \Big(\sum_{k=n_0+1}^{n} \gamma_k^2 \Big)^{p/2} \leq C\big((\log n)^{2/3} n_0^{-4\beta/3} (n - n_0) \big)^{p/2} \leq c\big((\log n)^{2/3} n^{-\beta/3} \big)^p.$$

For sufficiently large n (i.e., $n \geq N_3 = N_3(\beta)$) the third and the fourth terms on the right-hand side of inequality (5.31) are bounded similarly to (5.20) by the expression

$$\mathbb{E}\left(C \sum_{k=n_0+1}^{n} |\Delta\theta_k| \right)^p \leq c\big((n - n_0) n_0^{-\beta} \big)^p \leq C\big((\log n)^{2/3} n^{-\beta/3} \big)^p.$$

Finally we obtain that for $0 < \beta < 3/2$ and sufficiently large constant C_γ in the algorithm step $\gamma_k = C_\gamma (\log k)^{1/3} k^{-2\beta/3}$

$$\max_{n \geq N_\beta} \mathbb{E}\left(\frac{n^{\beta/3}}{(\log n)^{2/3}} |\delta_n| \right)^p \leq c,$$

where $N_\beta = \max(N_1, N_2, N_3)$ is the burn-in period of the algorithm.

Remark 5.6 *If we choose $\gamma_k = C_\gamma (\log k)^{\alpha_1} k^{-\alpha}$ and $n_0 = n - n^\alpha (\log n)^{\alpha_2}$, for some $0 < \alpha < 1$, $\alpha_1, \alpha_2 \geq 0$ and $\alpha_1 + \alpha_2 \geq 1$ in case $0 < \beta < 3/2$, then we get the following bound of the convergence rate: for sufficiently large n and sufficiently large constant C_γ*

$$\mathbb{E}|\delta_n|^p \leq C\Big(n^{-\min(\beta - \alpha,\, \alpha/2)} (\log n)^{\max(\alpha_2,\, \alpha_1 + \alpha_2/2)} \Big)^p.$$

Thus , the choice $\alpha = 2\beta/3$, $\alpha_1 = 1/3$, $\alpha_2 = 2/3$ is optimal in the sense of the minimum of the right-hand side of the above inequality.

Remark 5.7 *Much in the same way as for (5.18), we can establish that for any $\varepsilon > 0$, $\lim_{n\to\infty} n^{\beta/3-\varepsilon} |\delta_n| = 0$ with probability 1.*

Finally, consider the case $\beta = 0$, i.e., we assume the following weak requirement: $\mathbb{E}|\Delta\theta_k|^p \leq c$, $k \in \mathbb{N}$, for some uniform constant c. Take $n - n_0 = N$, $\gamma_k = \gamma$ for some $N \in \mathbb{N}$, $\gamma > 0$. Theorem 5.1 then implies that

$$\max_{n \geq N} \mathbb{E}|\delta_n|^p \leq C_1 e^{-phN\gamma} + C_2 N^{p/2} \gamma^p + C_3 N^p c + C_4 N^p \gamma^p c = E.$$

We thus have that the algorithm will track down the conditional quantile in the proximity of size E, which we can try to minimize by choosing appropriate constants N and γ.

5.4.3 Lipschitz quantile: asymptotics in frequency of observations

Consider the model

$$X_k = \theta(k/n) + \sigma_k(\mathbf{X}_k)\xi_k, \quad k = 1, 2, \dots, n,$$

where the σ_k's satisfy (5.14) and $\theta(t)$, $t \in [0,1]$ is an unknown Lipschitz function, i.e., $\theta(\cdot) \in \mathcal{L}(L,\beta) = \{g(\cdot) : |g(t_1) - g(t_2)| \le L|t_1 - t_2|^\beta,\ t_1, t_2 \in [0,1]\}$, for some $0 < \beta \le 1$ and $L > 0$. The parameter n has a meaning of frequency of the observations, the number of observations per time unit. This setting is typical for the problem of nonparametric regression estimation. The nonparametric median estimation problem has been studied in [13] and [14]) for such an asymptotic regime. Although this asymptotic regime is not really practical for recursive procedures – the step size is constant and depends on the observation frequency n, and so if n changes, the whole model changes – we derive the asymptotic results for this setting as consequences of our non-asymptotic general Theorem 5.1.

Let $\gamma_k \equiv C_\gamma (\log n)^{(2\beta-1)/(2\beta+1)} n^{-2\beta/(2\beta+1)}$ for $k = 1, \dots, n$, and define

$$k_0 = k_0(n) = k - (\log n)^{2/(2\beta+1)} n^{2\beta/(2\beta+1)},$$

for all $k \ge K_n = (\log n)^{2/(2\beta+1)} n^{2\beta/(2\beta+1)}$. For sufficiently large C_γ

$$\sum_{i=k_0+1}^{k} \gamma_i = C_\gamma (\log n)^{(2\beta-1)/(2\beta+1)} n^{-2\beta/(2\beta+1)} (k - k_0) \ge C_\gamma \log n \ge \frac{\log n}{2h},$$

which leads to

$$\exp\left(- ph \sum_{i=k_0+1}^{k} \gamma_i \right) \le c n^{-p/2}.$$

Now we have

$$\left(\sum_{i=k_0+1}^{k} \gamma_i^2 \right)^{p/2} \le C\Big((\log n)^{\frac{2\beta-1}{2\beta+1}} n^{-\frac{2\beta}{2\beta+1}} (k - k_0)^{1/2} \Big)^p = C\Big((\log n)^{\frac{2\beta}{2\beta+1}} n^{-\frac{\beta}{2\beta+1}} \Big)^p.$$

By the Lipschitz property of the function $\theta(t)$, with $\theta_i = \theta(i/n)$,

$$\max_{k_0+1 \le i \le k} |\theta_i - \theta_{k_0}|^p \le c \left| \frac{k - k_0}{n} \right|^{\beta p} \le C\Big((\log n)^{\frac{2\beta}{2\beta+1}} n^{-\frac{\beta}{2\beta+1}} \Big)^p.$$

Since $|\Delta\theta_i| \le L\left|\frac{i}{n} - \frac{i-1}{n}\right|^\beta \le c n^{-\beta}$,

$$\left(\sum_{i=k_0+1}^{k} \gamma_i |\Delta\theta_i| \right)^p \le c\Big((k - k_0) \gamma_{k_0} n^{-\beta} \Big)^p \le C\big(n^{-\beta} \log n \big)^p.$$

Combining the last four inequalities with the bound (5.31), we derive that for sufficiently large C_γ

$$\sup_{\theta(\cdot) \in \mathcal{L}(L,\beta)} \max_{k \ge K_n} \mathbb{E}|\delta_k|^p \le C\Big((\log n)^{\frac{2\beta}{2\beta+1}} n^{-\frac{\beta}{2\beta+1}} \Big)^p. \tag{5.21}$$

In the derivation of this inequality, we used the uniformity of all the bounds over $k \geq K_n = (\log n)^{2/(2\beta+1)} n^{2\beta/(2\beta+1)}$ and over the Lipschitz functional class $\theta(\cdot) \in \mathcal{L}(L, \beta)$. The sequence K_n has a meaning of "burn-in" period.

Notice that the resulting convergence rate coincides (up to a logarithmic factor) with the minimax rate over the Lipschitz class $\mathcal{L}(L, \beta)$ in the problem of minimax regression function estimation.

5.5 Numerical example

We now treat a small numerical example to illustrate our results. This example can be found in [92]. For $\mathbf{t}^{(n)} = \big(1/n - 1, 3/n - 1, \ldots, 1 - 3/n, 1 - 1/n\big)$ we make observations from

$$X_i = f(t_i) + \sigma(t_i)\xi_i, \quad i = 1, \ldots, n,$$

where the ξ_i's are independent standard normal random variables. The function f and standard deviation of the noise σ are taken, for $t \in [-1, 1]$, as

$$f(t) = \frac{\sin(t)}{t}, \quad \sigma(t) = 0.1 \exp(1 - t).$$

As explained in the previous section, we technically have a different model for each value of n and the bound given in Theorem 5.1 becomes a statement about asymptotics in the sampling frequency n. At time k, for $\alpha \in (0, 1)$, we are interested in estimating the α-th quantile of X_k, call it θ_k, based on the data $\mathbf{X}_k = (X_1, \ldots, X_k)$. It is straightforward to see that we can write $\theta_k = \theta_{\alpha,k} = \vartheta_\alpha(t_k)$, where

$$\vartheta_\alpha(t) = f(t) + \sigma(t)\Phi^{-1}(\alpha), \quad t \in [-1, 1],$$

where Φ is the cumulative distribution function of a standard normal random variable from where we assume that the quantile function ϑ_α is in $\mathcal{L}_1\big([-1, 1]\big)$.

In our numerical study we took $\gamma_k \equiv C_\gamma(\log(n)/n^2)^{1/3}$ for $C_\gamma = 2.5$, $n \in \{100, 250, 500, 1000, 10000\}$ and $\alpha \in \{0.1, 0.25, 0.5, 0.75, 0.9\}$. Our tracking sequence is then defined as

$$\hat{\theta}_k = \Pi_{\bar{C}}\Big(\hat{\theta}_{k-1} + \gamma_k\big(\alpha - \mathbb{1}\{X_k < \hat{\theta}_{k-1}\}\big)\Big), \quad k = 1, \ldots, n, \tag{5.22}$$

as defined in (5.4) where we took $\hat{\theta}_0 = 2$ and $\bar{C} = 5$. In Figure 5.5.1 we show the functions $\theta_\alpha(t)$ which are obtained by linearly interpolating the sequence (5.22), for each n and each α. (Note that to get the tracking sequence only requires knowledge of the value of the indicators $\mathbb{1}\{X_k < \hat{\theta}_{k-1}\}$ and not of the actual observations X_k.)

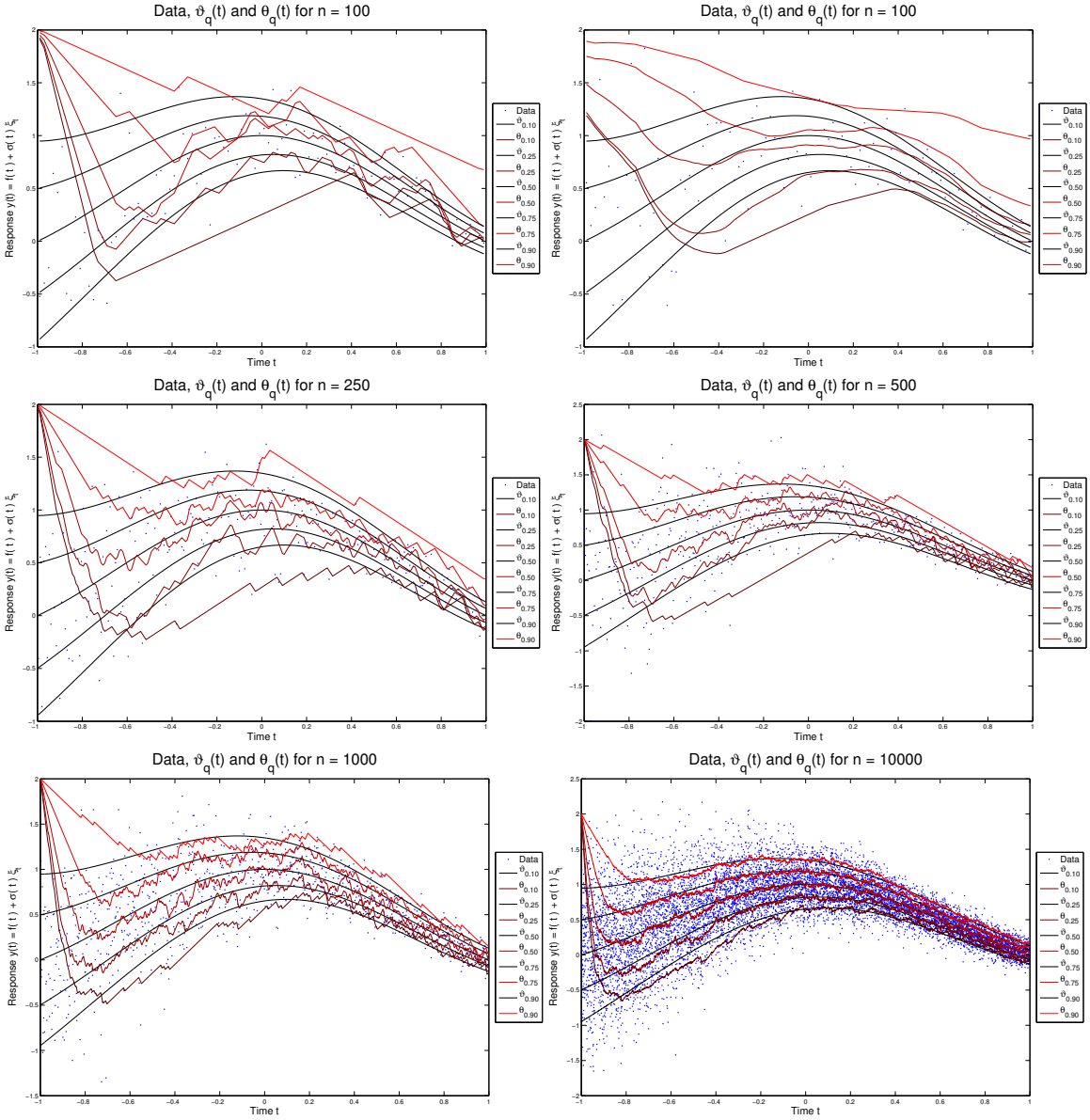

Figure 5.5.1: Results of the tracking algorithm. All pictures contain the data (blue dots), the true quantile function for the chosen values of α (black lines), and the respective tracking sequences (tones of red). To each picture corresponds a specific sample size. On the first row we compare, for $n = 100$, the *raw* tracking sequence (left) with a smoothed version of it (right).

The burn-in period of the algorithm is quite noticeable in the pictures in Figure 5.5.1 as is the improvement in the quality of the approximations as n increases. The constant C_γ has to be picked large enough so as not to hinder the capability of the tracking sequence to "catch up" with the signal. Picking C_γ too large will however cause the tracking sequence to make large jumps (since γ_k is larger for small n). This can be partially compensated for, by smoothing out the resulting tracking sequence via, say, a moving window average, i.e., replacing each $\hat\theta_k$ by the mean of $\hat\theta_{1\vee(k-w)},\dots,\hat\theta_k,\dots,\hat\theta_{n\wedge(k+w)}$ for some $w\in\mathbb{N}$. (For the top right picture in Figure 5.5.1 we took $w=\lfloor 2.5n^{1/3}\rfloor$.) This somewhat improves the results although its clear that the approximation of the more extreme quantiles (0.1 and 0.9) is quite crude for low values of n. We will revisit this example in the next chapter, in Section 6.7.1.

5.6 Proofs

In this section we omit mentioning $k\in\mathbb{N}$ as we agree that all the relations where the index k is involved hold for all $k\in\mathbb{N}$. We will also agree that the constant c is the constant from condition (D1) if $\mathcal{X}$ is discrete and $c=0$ otherwise.

Lemma 5.1 *Let the functions $\tilde S_k(u,v)$, $k\in\mathbb{N}$, be defined by (5.9) and constant $\bar C=C_\Theta+\Gamma(1+c/2)+\kappa$ by (5.3), where c is as in (D1) if $\mathcal{X}$ is discrete and $c=0$ otherwise. Then $|\tilde S_k(u,v)|\le 1+c/2$, uniformly over $u\in\mathbb{R}$ and $|v|\le\bar C$. Also, the relations*

$$\tilde S_k(u,v)=-\tilde G_k(\theta_k,u,v)(v-\theta_k),\quad k\in\mathbb{N},\tag{5.23}$$

hold for some functions $\tilde G_k(\theta_k,u,v)$ such that $0\le\tilde G_k(\theta_k,u,v)\le\kappa^{-1}$, uniformly over $u\in\mathbb{R}$, $\theta_k\in\Theta$ and $|v|\le\bar C$.

Proof: For $u\in\mathbb{R}$, $\theta_k\in\Theta$, $|v|\le\bar C$ put

$$\tilde G_k(\theta_k,u,v)=-\frac{\tilde S_k(u,v)}{v-\theta_k}\quad\text{if}\quad v\ne\theta_k.\tag{5.24}$$

and $\tilde G_k(\theta_k,u,\theta_k)=0$. Let us show that functions $\tilde G_k(\theta_k,u,v)$ satisfy the assertions of the lemma. First of all, note that the representation (5.23) holds true. In view of (5.24), it is trivial if $v\ne\theta_k$. The representation (5.23) holds true also for $v=\theta_k$ since $\tilde G_k(\theta_k,u,\theta_k)=0$ by the definition and $\tilde S_k(u,\theta_k)=0$ for all $u\in\mathbb{R}$ and $\theta_k\in\Theta$, as $|\theta_k+\gamma_k\tilde S_k(u,\theta_k)|\le C_\Theta+\Gamma(1+c/2)\le\bar C$.

Now we need to prove that $0\le\tilde G_k(\theta_k,u,v)\le\kappa^{-1}$ uniformly over $u\in\mathbb{R}$, $\theta_k\in\Theta$ and $|v|\le\bar C$. Since $\tilde G_k(\theta_k,u,\theta_k)=0$, we consider only the case $v\ne\theta_k$. If $\gamma_k=0$ or $|v+\gamma_k\tilde S_k(u,v)|\le\bar C$, then $\tilde S_k(u,v)=0$ so that, according to (5.24), $\tilde G_k(\theta_k,u,v)=0$ and the lemma is proved also for these cases.

It remains to consider the case when $v\ne\theta_k$, $\gamma_k>0$ and $|v+\gamma_k\tilde S_k(u,v)|>\bar C$. If we have additionally that $|v|\le\bar C$, then $\bar C<|v+\gamma_k\tilde S_k(u,v)|\le|v|+|\gamma_k\tilde S_k(u,v)|\le\bar C+\gamma_k(1+c/2)$. Then $0\le\gamma_k^{-1}(|v+\gamma_k\tilde S_k(u,v)|-\bar C)\le 1+c/2$, or equivalently

$$0\le\gamma_k^{-1}|v+\gamma_k\tilde S_k(u,v)-\bar C\,\mathrm{sign}(v+\gamma_k\tilde S_k(u,v))|\le 1+c/2,\quad u\in\mathbb{R},\ |v|\le\bar C,$$

which implies in passing the first assertion of the lemma: $|\tilde S_k(u,v)|\le 1+c/2$, $u\in\mathbb{R}$, $|v|\le\bar C$. Obviously, $v+$

$\gamma_k \bar{S}_k(u,v)$ is of the same sign as v since $|\bar{S}_k(u,v)| \le 1+c/2$ and $\gamma_k \le \Gamma$, whereas $\bar{C} = C_\Theta + \Gamma(1+c/2) + \kappa > \Gamma(1+c/2)$. This implies that $v + \gamma_k \bar{S}_k(u,v) - \bar{C}\,\mathrm{sign}(v + \gamma_k \bar{S}_k(u,v))$ is also of the same sign as v because $|v + \gamma_k \bar{S}_k(u,v)| > \bar{C}$. Now, from $|v| \ge |v + \gamma_k \bar{S}_k(u,v)| - |\gamma_k \bar{S}_k(u,v)| \ge \bar{C} - \Gamma(1 + c/2) = C_\Theta + \kappa$, it follows that $v - \theta_k$ is of the same sign as v for all $\theta_k \in \Theta$. Moreover, $|v - \theta_k| \ge |v| - |\theta_k| \ge C_\Theta + \kappa - C_\Theta = \kappa$. Thus we showed that if $u \in \mathbb{R}$, $\theta_k \in \Theta$, $|v| \le \bar{C}$ and $v \ne \theta_k$, then

$$0 \le \frac{\gamma_k^{-1}\big(v + \gamma_k \bar{S}_k(u,v) - \bar{C}\,\mathrm{sign}(v + \gamma_k \bar{S}_k(u,v))\big) I\{|v + \gamma_k \bar{S}_k(u,v)| > \bar{C}\}}{v - \theta_k} \le \frac{1}{\kappa}. \tag{5.25}$$

In view of (5.9) and (5.24), the expression in the middle of the relation (5.25) is nothing else but $\tilde{G}_k(\theta_k, u, v)$ for $v \ne \theta_k$. By using $\tilde{G}_k(\theta_k, u, \theta_k) = 0$ and the relation (5.25), we obtain the second assertion of the lemma:

$$0 \le \tilde{G}_k(\theta_k, u, v) \le \frac{1}{\kappa}, \quad u \in \mathbb{R},\ \theta_k \in \Theta,\ |v| \le \bar{C}.$$

□

Lemma 5.2 *Let the functions $S_k(u,v)$, $k \in \mathbb{N}$, be defined by (5.8). Let c be as in condition (D2) for discrete $\mathcal{X}$ and $c = 0$ for continuous $\mathcal{X}$. Then $|S_k(u,v)| \le 2 + c$, $k \in \mathbb{N}$, uniformly over $u \in \mathbb{R}$ and $|v| \le \bar{C}$. Moreover, if conditions (C1), (C2) are fulfilled, then*

$$\mathbb{E}\big[S_k(X_k, v)\big|\mathbf{X}_{k-1}\big] = -G_k(\theta_k, v)(v - \theta_k), \quad k \in \mathbb{N}, \tag{5.26}$$

for some functions $G_k(\theta_k, v) = G_k(\theta_k, v, \mathbf{X}_{k-1})$ such that $0 < h \le G_k(\theta_k, v) \le H$ with probability 1, uniformly over $\theta_k \in \Theta$ and $|v| \le \bar{C}$ (the constants h, H depend on δ, C_Θ, $\bar{C}$ and κ; they also depend on either b and B for continuous $\mathcal{X}$ or on c for discrete $\mathcal{X}$.)

Proof: The definition (5.1) implies the obvious bound $|\bar{S}_k(u,v)| \le 1 + c/2$. By Lemma 5.1, $|\tilde{S}_k(u,v)| \le 1 + c/2$ uniformly in $u \in \mathbb{R}$ and $|v| \le \bar{C}$. Therefore $|S_k(u,v)| \le |\bar{S}_k(u,v)| + |\tilde{S}_k(u,v)| \le 2 + c$ uniformly in $u \in \mathbb{R}$ and $|v| \le \bar{C}$, which proves the first assertion of the lemma.

For $\theta_k \in \Theta$ and $|v| \le \bar{C}$ define $G_k(\theta_k, \theta_k) = \frac{b+B}{2}$, and for $v \ne \theta_k$ define

$$G_k(\theta_k, v) = -\frac{\mathbb{E}\big[S_k(X_k, v)|\mathbf{X}_{k-1}\big]}{v - \theta_k} = -\frac{\mathbb{E}\big[\bar{S}_k(X_k, v)|\mathbf{X}_{k-1}\big]}{v - \theta_k} - \frac{\mathbb{E}\big[\tilde{S}_k(X_k, v)|\mathbf{X}_{k-1}\big]}{v - \theta_k}. \tag{5.27}$$

Representation (5.26) holds immediately for $v \ne \theta_k$. In case $v = \theta_k$, $\tilde{S}_k(u, \theta_k) = 0$ for all $u \in \mathbb{R}$ and $\theta_k \in \Theta$ since $|\theta_k + \gamma_k \bar{S}_k(u, \theta_k)| \le \bar{C}$, and $\mathbb{E}\big[\tilde{S}_k(X_k, \theta_k)|\mathbf{X}_{k-1}\big] = \alpha_k - \mathbb{P}\big(X_k - \theta_k(\mathbf{X}_{k-1}) < 0|\mathbf{X}_{k-1}\big) = \alpha_k - F_k(0|\mathbf{X}_{k-1}) = 0$ by the definition of the conditional distribution function F_k. Therefore, in case $v = \theta_k \in \Theta$, we obtain that $\mathbb{E}\big[S_k(X_k, \theta_k)|\mathbf{X}_{k-1}\big] = \mathbb{E}\big[\bar{S}_k(X_k, \theta_k)|\mathbf{X}_{k-1}\big] + \mathbb{E}\big[\tilde{S}_k(X_k, \theta_k)|\mathbf{X}_{k-1}\big] = 0$ and the relation (5.26) holds again. It remains to prove that $G_k(\theta_k, v)$ satisfies the inequality $0 < h \le G_k(\theta_k, v) \le H$ uniformly over $\theta_k \in \Theta$ and $|v| \le \bar{C}$.

For $\theta_k \in \Theta$, $|v| \le \bar{C}$ and $v \ne \theta_k$, introduce the function

$$\bar{G}_k(\theta_k, v) = -\frac{\mathbb{E}\big[\bar{S}_k(X_k, v)|\mathbf{X}_{k-1}\big]}{v - \theta_k} = \frac{F_k(v - \theta_k|\mathbf{X}_{k-1}) - \alpha_k - c/2}{v - \theta_k}. \tag{5.28}$$

Since the conditional distribution function F_k has zero quantile of level $\alpha_k \in (0,1)$, function $\bar{G}_k(\theta_k, v)$ is always non-negative. In view of (C1) and (D1), we have that for $0 < |v - \theta_k| \leq \delta$, almost surely,

$$b \leq \bar{G}_k(\theta_k, v) \leq B \quad \text{and} \quad \frac{c}{2} \leq \bar{G}_k(\theta_k, v) \leq 2,$$

for respectively $\mathcal{X}$ a continuous set and for $\mathcal{X}$ a discrete set.

Suppose now that $|v - \theta_k| > \delta$. Then obviously $\bar{G}_k(\theta_k, v) \leq (1 + c/2)\delta^{-1}$. On the other hand, as $|\theta_k| \leq C_\Theta$ and $|v| \leq \bar{C}$, $|v - \theta_k| \leq |v| + |\theta_k| \leq \bar{C} + C_\Theta$, which in turn implies by (C1) that

$$\bar{G}_k(\theta_k, v) \geq \frac{\min\left(F_k(\delta|\mathbf{X}_{k-1}) - \alpha_k - c/2,\ c/2 + \alpha_k - F_k(-\delta|\mathbf{X}_{k-1})\right)}{|v - \theta_k|}.$$

The last display is lower-bounded, almost surely, by $b\delta/(\bar{C} + C_\Theta)$ if $\mathcal{X}$ is a continuous set and lower-bounded, almost surely, by $c(\delta - 1/2)/(\bar{C} + C_\Theta)$ if $\mathcal{X}$ is discrete. Thus, for any $\theta_k \in \Theta$, $|v| \leq \bar{C}$, $v \neq \theta_k$, we have established that

$$\min\left(b,\ b\delta(\bar{C} + C_\Theta)^{-1}\right) \leq \bar{G}_k(\theta_k, v) \leq \max\left(B,\ \delta^{-1}\right) \tag{5.29}$$

almost surely, for data supported on a continuous set $\mathcal{X}$ and

$$\min\left(c/2,\ c(\delta - 1/2)(\bar{C} + C_\Theta)^{-1}\right) \leq \bar{G}_k(\theta_k, v) \leq \max\left(2, (1 + c/2)\delta^{-1}\right) \tag{5.30}$$

almost surely, for discrete $\mathcal{X}$, with in both cases the functions $\bar{G}_k(\theta_k, v)$ defined by (5.28).

Recall that we defined $G_k(\theta_k, \theta_k) = \frac{b+B}{2}$, $\theta_k \in \Theta$, so that $b \leq G_k(\theta_k, \theta_k) \leq B$. Using this fact, relations (5.27), (5.28), (5.29) and Lemma 5.1, we obtain for $\mathcal{X}$ continuous that,

$$h = \min\left(b,\ \frac{b\delta}{\bar{C} + C_\Theta}\right) \leq G_k(\theta_k, v) \leq \max\left(B + \frac{1}{\kappa},\ \frac{1}{\delta} + \frac{1}{\kappa}\right) = H$$

and for discrete $\mathcal{X}$, using (5.27), (5.28), (5.30) and Lemma 5.1,

$$h = \min\left(\frac{c}{2},\ \frac{c(\delta - 1/2)}{\bar{C} + C_\Theta}\right) \leq G_k(\theta_k, v) \leq \max\left(2 + \frac{1}{\kappa},\ \frac{1 + c/2}{\delta} + \frac{1}{\kappa}\right) = H$$

both almost surely, uniformly in $\theta_k \in \Theta$, $|v| \leq \bar{C}$. This establishes (5.26) and the lemma is proved. □

Theorem 5.1

Let conditions (C1), (C2) be satisfied, let the estimator $\hat{\theta}_k$ be defined by (5.4) with step sizes $\{\gamma_k,\ k \in \mathbb{N}\}$ satisfying (5.2) and let the constants h and H be from Lemma 5.2. Define $\delta_k = \hat{\theta}_k - \theta_k$ and $\Delta\theta_k = \theta_k - \theta_{k-1}$, $k \in \mathbb{N}$ (define $\theta_0 = 0$). Then for any $p \geq 1$, any $n, n_0 \in \mathbb{N}$ such that $n_0 \leq n$ and $\gamma_k H \leq 1$ for all $n_0 \leq k \leq n$, the following relation holds

$$\mathbb{E}|\delta_n|^p \leq C_1 \exp\left(-ph \sum_{k=n_0+1}^{n} \gamma_k\right) + C_2\left(\sum_{k=n_0+1}^{n} \gamma_k^2\right)^{p/2} + C_3\,\mathbb{E}\left(\max_{n_0+1 \leq k \leq n} \left|\theta_k - \theta_{n_0}\right|^p\right) + C_4\,\mathbb{E}\left(\sum_{k=n_0+1}^{n} \gamma_k |\Delta\theta_k|\right)^p \tag{5.31}$$

for some positive C_1, C_2, C_3 and C_4, constants which depend only on p, C_Θ, $\bar{C}$ and H.

Proof: In this proof, we impose the convention that the summation and the product over an empty set of indices is zero and one respectively. In particular, $\sum_{i=m+1}^{m} b_i = 0$ and $\prod_{i=m+1}^{m} b_i = 1$. Recall the notations $\delta_k = \hat{\theta}_k - \theta_k$, $\Delta\theta_k = \theta_k - \theta_{k-1}$, and introduce further

$$g_k(\theta_k, \nu) = \mathbb{E}\big[S_k(X_k, \nu)|\mathbf{X}_{k-1}\big], \quad M_k(X_k, \theta_k, \nu) = S_k(X_k, \nu) - g_k(\theta_k, \nu).$$

By Lemma 5.2, $g_k(\theta_k, \hat{\theta}_{k-1}) = -G_k(\theta_k, \hat{\theta}_{k-1})(\hat{\theta}_{k-1} - \theta_k)$ for some functions $G_k(\theta_k, \nu)$ such that almost surely $0 < h \le G_k(\theta_k, \nu) \le H$, uniformly over $\theta_k \in \Theta$ and $|\nu| \le \bar{C}$. Then

$$g_k(\theta_k, \hat{\theta}_{k-1}) = -G_k(\theta_k, \hat{\theta}_{k-1})(\hat{\theta}_{k-1} - \theta_k) = -G_k(\theta_k, \hat{\theta}_{k-1})\delta_{k-1} + G_k(\theta_k, \hat{\theta}_{k-1})\Delta\theta_k$$

and the algorithm (5.7) can thus be written in the following form:

$$\begin{aligned}\delta_k &= \delta_{k-1} + \gamma_k\big(M_k(X_k, \theta_k, \hat{\theta}_{k-1}) + g_k(\theta_k, \hat{\theta}_{k-1})\big) - \Delta\theta_k \\ &= \delta_{k-1}\big(1 - \gamma_k G_k(\theta_k, \hat{\theta}_{k-1})\big) + \gamma_k M_k(X_k, \theta_k, \hat{\theta}_{k-1}) - \big(1 - \gamma_k G_k(\theta_k, \hat{\theta}_{k-1})\big)\Delta\theta_k \\ &= \delta_{k-1}q_k + r_k, \quad k \in \mathbb{N}. \end{aligned} \tag{5.32}$$

Here and from now on we use the following notations:

$$q_k = 1 - \gamma_k G_k(\theta_k, \hat{\theta}_{k-1}), \quad r_k = \gamma_k M_k(X_k, \theta_k, \hat{\theta}_{k-1}) - q_k\Delta\theta_k, \quad R_k = \sum_{i=n_0+1}^{k} r_i. \tag{5.33}$$

Now we bound the random variables q_k, $n_0 \le k \le n$ defined above by (5.33). According to the conditions of the theorem, $0 \le 1 - \gamma_k H$ if $n_0 \le k \le n$ and, by Lemma 5.2 and (C2), $0 < h \le G_k(\theta_k, \hat{\theta}_{k-1}) \le H$ almost surely, uniformly over $\theta_k \in \Theta$. Then

$$0 \le 1 - \gamma_k H \le q_k = 1 - \gamma_k G_k(\theta_k, \hat{\theta}_{k-1}) \le 1 - \gamma_k h \le e^{-\gamma_k h} \le 1, \quad n_0 \le k \le n, \tag{5.34}$$

almost surely. Here we also used the elementary inequality $1 - x \le e^{-x}$.

By iterating the relation (5.32) from n up to n_0, $0 \le n_0 \le n$, and by applying the Abel transformation for series,

$$\begin{aligned}\delta_n &= \delta_{n_0}\prod_{k=n_0+1}^{n} q_k + \sum_{k=n_0+1}^{n} r_k \prod_{j=k+1}^{n} q_j \\ &= \delta_{n_0}\prod_{k=n_0+1}^{n} q_k + R_n q_n + \sum_{k=n_0+1}^{n-1} R_k(q_{k+1} - 1)\prod_{j=k+1}^{n} q_j. \end{aligned} \tag{5.35}$$

Moreover, note that

$$\sum_{k=n_0+1}^{n-1}(1 - q_{k+1})\prod_{j=k+2}^{n} q_j = \sum_{k=n_0+1}^{n-1}\Big(\prod_{j=k+2}^{n} q_j - \prod_{j=k+1}^{n} q_j\Big) = 1 - \prod_{j=n_0+2}^{n} q_j. \tag{5.36}$$

From relations (5.35), (5.36) and (5.34) it follows that

$$|\delta_n| \le |\delta_{n_0}| \prod_{k=n_0+1}^{n} (1-\gamma_k h) + 2 \max_{n_0+1\le k\le n} |R_k|. \tag{5.37}$$

By using (5.33) and (5.34), we bound the term $\max_{n_0+1\le k\le n} |R_k|$:

$$\begin{aligned} \max_{n_0+1\le k\le n} |R_k| \le & \max_{n_0+1\le k\le n} \Big| \sum_{i=n_0+1}^{k} \gamma_i M_i(X_i, \theta_i, \hat\theta_{i-1}) \Big| \\ & + \max_{n_0+1\le k\le n} \Big| \sum_{i=n_0+1}^{k} \Delta\theta_i \Big| + H \sum_{i=n_0+1}^{n} \gamma_i |\Delta\theta_i|, \end{aligned} \tag{5.38}$$

Condition (C2) and (5.6) ensure that the estimation accuracy δ_k is always bounded:

$$|\delta_k| \le |\hat\theta_k| + |\theta_k| \le \bar C + C_\Theta, \quad k \in \mathbb{N}.$$

Taking into account this fact, the inequalities (5.37), (5.38) and again the elementary inequality $1 + x \le e^x$, $x \in \mathbb{R}$, we conclude that

$$\begin{aligned} |\delta_n| \le (\bar C + C_\Theta) \exp\Big(- h \sum_{k=n_0+1}^{n} \gamma_k \Big) & + 2 \max_{n_0+1\le k\le n} \Big| \sum_{i=n_0+1}^{k} \gamma_i M_i(X_i, \theta_i, \hat\theta_{i-1}) \Big| \\ & + 2 \max_{n_0+1\le k\le n} \Big| \sum_{i=n_0+1}^{k} \Delta\theta_i \Big| + 2H \sum_{i=n_0+1}^{n} \gamma_i |\Delta\theta_i|. \end{aligned} \tag{5.39}$$

Now note that the sequence $\{\gamma_k M_k(X_k, \theta_k, \hat\theta_{k-1}),\ k \in \mathbb{N}\}$ is nothing else but a martingale difference with respect to the natural filtration $\{\mathfrak{F}_k,\ k \in \mathbb{N}\}$, i.e., $\mathfrak{F}_k = \sigma(X_1, \dots, X_k)$ is the σ-algebra generated by the random variables $X_1, \dots, X_k$. Indeed, by Lemma 5.2 and (5.2) this sequence is bounded almost surely

$$|\gamma_k M_k(X_k, \theta_k, \hat\theta_{k-1})| \le \gamma_k \big(|S_k(X_k, \hat\theta_{k-1})| + |g_k(\theta_k, \hat\theta_{k-1}|\big) \le (4 + 2c)\Gamma, \quad k \in \mathbb{N},$$

and

$$\mathbb{E}\big[M(X_{k+1}, \theta_{k+1}, \hat\theta_k,) \big| \mathfrak{F}_k \big] = g_k(\theta_k, \hat\theta_k) - g_k(\theta_k, \hat\theta_k) = 0, \quad k \in \mathbb{N}.$$

Therefore the sequence $\big\{ \sum_{i=n_0+1}^{k} \gamma_i M_i(X_i, \theta_i, \hat\theta_{i-1}),\ k \ge n_0 + 1 \big\}$ is a martingale with respect to the same filtration. Since $|M_i(X_i, \theta_i, \hat\theta_{i-1})| \le 4 + 2c$ due to Lemma 5.2, we can apply the maximal Burkholder inequality in case $p > 1$ and the Davis inequality for $p = 1$ (see, for example, [87]) to this martingale: for any $p \ge 1$ there exists a constant B_p such that

$$\begin{aligned} \mathbb{E}\Big[\max_{n_0+1\le k\le n} \Big| \sum_{i=n_0+1}^{k} \gamma_i M_i(X_i, \theta_i, \hat\theta_{i-1}) \Big|^p \Big] & \le B_p\, \mathbb{E}\Big[\sum_{i=n_0+1}^{n} \gamma_i^2 M_i^2(X_i, \theta_i, \hat\theta_{i-1}) \Big]^{p/2} \\ & = (4 + 2c)^p B_p \Big(\sum_{i=n_0+1}^{n} \gamma_i^2 \Big)^{p/2}. \end{aligned} \tag{5.40}$$

For $p > 1$, one can take $B_p = \Big[(18 p^{5/2}/(p-1)^{3/2} \Big]^p$; cf. [87].

Now take the pth power ($p \geq 1$) of both sides of the inequality and apply the Hölder inequality $\left(\sum_{i=1}^{m} a_i\right)^p \leq m^{p-1} \sum_{i=1}^{m} |a_i|^p$ for $m = 4$. Next, take the expectations of the both sides of the resulting inequality and use (5.40) to derive the statement of the theorem:

$$\mathbb{E}|\delta_n|^p \leq C_1 \exp\left(-ph \sum_{k=n_0+1}^{n} \gamma_k\right) + C_2\Big(\sum_{i=n_0+1}^{n} \gamma_i^2\Big)^{p/2} + C_3\mathbb{E}\Big[\max_{n_0+1\leq k\leq n} \Big|\sum_{i=n_0+1}^{k} \Delta\theta_i\Big|^p\Big] + C_4\mathbb{E}\Big[\sum_{i=n_0+1}^{n} \gamma_i|\Delta\theta_i|\Big]^p.$$

□

6

Tracking of Drifting Parameters of a Time Series

In this chapter we present an online algorithm for tracking a multivariate time-varying parameter of a time series. The algorithm is driven by a gain function. Under assumptions on the gain function, we derive uniform error bounds on the tracking algorithm in terms of the chosen step size for the algorithm and on the variation of the parameter of interest. We give examples of a number of different variational setups for the parameter where our result can be applied, and we also outline how appropriate gain functions can be constructed. We treat in some detail the tracking of time varying parameters of an AR(d) model as a particular application of our method and present two small numerical studies.

6.1 Introduction

When one analyzes data that arrive sequentially over time, it is important to detect changes in the underlying model which can then be adjusted accordingly. Estimation or tracking of time-varying parameters in stochastic systems is therefore of fundamental interest in sequential analysis. Furthermore, it arises in many engineering, econometric and biomedical applications and has an extensive literature widely scattered in these fields. Motivated by many applications in signal processing, speech recognition, communication systems, neural physiology, environmental and economic modeling, we consider in this chapter the problem of recursive (online) estimation of the multivariate time-varying parameter of a time series.

Consider then an $\mathcal{X}$-valued time series $(X_k,\ k \in \mathbb{N}_0)$, $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$, $\mathcal{X} \subseteq \mathbb{R}^l$, such that at time moment $t = 0$ the first observation $X_0 \sim P_{\theta_0}$ and subsequently at each time moment $k \in \mathbb{N}$ a new datum X_k arrives according to the model $X_k|\mathbf{X}_{k-1} \sim P_{\theta_k}(\cdot|\mathbf{X}_{k-1})$ with transition law depending on some multivariate parameter $\theta_k \in \Theta \subseteq \mathbb{R}^d$ and where $\mathbf{X}_{k-1} = (X_0, X_1, \ldots, X_{k-1})$. Thus, the growing statistical model is, at time $t = n$, $\mathcal{P}^{(n)} = \mathcal{P}^{(n)}(\Theta^{n+1}) = \{\prod_{k=0}^{n} P_{\theta_k}(x_k|\mathbf{x}_{k-1}) : (\theta_0, \ldots, \theta_n) \in \Theta^{n+1}, \mathbf{x}_n \in \mathcal{X}^{n+1}\}$ with the convention that $P_{\theta_0}(y_0|\mathbf{x}_{-1}) = P_{\theta_0}(y_0)$. This time series formulation represents the most general sequential setting, sequences of independent observations and Markov chains of arbitrary order are typical examples of models that fit into this framework.

The multivariate parameter $\theta_k \in \Theta \subseteq \mathbb{R}^d$, $k \in \mathbb{N}$, is time-varying and the goal is to estimate (or to track) its value based on the data $\mathbf{X}_k$ (and prior information) available by that time moment. Since the data arrives in a successive manner, conventional methods based on samples of a fixed size are not easy to use. A more appropriate approach is based on sequential methods, stochastic recursive algorithms, which allow fast updating of parameter or state estimates at each instant as new data arrive and therefore can be used to perform "online" inference, that is, during the operation of the system. Stochastic recursive algorithms, also known as stochastic approximation, take many forms and have numerous applications in the biomedical, socio-economic and engineering sciences, which highlights the interdisciplinary nature of the subject.

There is a vast literature on stochastic approximation beginning with the seminal papers [76] and [53]. There is a large variety of techniques in the area of stochastic approximation which have been developed and inspired by the applications from other fields. We mention here the books [15, 59, 61, 66, 72, 93, 101].

A classical topic in adaptive control concerns the problem of tracking drifting parameters of a linear regression model, or somewhat equivalently, tracking the best linear fit when the parameters change slowly. This problem also occurs in communication theory for adaptive equalizers and noise cancellation, etc., where the signal, noise, and channel properties change with time. Successful stochastic approximation schemes for tracking in the time-varying case were given in [17, 26, 60, 61] (see further references therein).

In [60] (see also [15, 17]) the authors discuss the important problem of the choice of the step sizes in the tracking algorithm which we also address in this chapter. In general, the step size of the tracking algorithm is not necessarily decreasing to zero because of considerations concerning robustness of the actual physical model in practical online applications and to allow some tracking of the desired parameter as the system changes over time. In signal processing applications, it is usual to keep the step size bounded away from zero.

Coming back to our model $\mathcal{P}^{(n)}$ with time-varying parameter $(\theta_k \in \Theta,\ k \in \mathbb{N}_0)$, the problem of tracking a signal θ_k is clearly unfeasible, especially in such general formulation, without some conditions on the model $\mathcal{P}^{(n)}$. In general, some knowledge about the structure of underlying time seres and some control over the variability of the parameter θ_k over time are needed. Interestingly, in this seemingly very general time series framework,

we actually do not require the knowledge of the model $\mathcal{P}^{(n)}$. Instead, all we need is to be able to compute a so-called *gain vector* at each time moment $k \in \mathbb{N}$, which is a certain (vector) function of the previous estimate of the parameter θ_k, new observation X_k and history $\mathbf{X}_{k-1}$. The essential property of such a gain vector is that it, roughly speaking, "pushes" in the right direction of the current value of true parameter to track. Although the assumption about the existence of that gain vector seems to be rather strong, we demonstrate on a number of interesting examples when such an assumption indeed holds. Basically, in case of observations from a Markov chain, if the form of transition density is known as a function of the underlying parameter and it satisfies certain regularity assumptions, then the gain vector can always be constructed, for example, as a score function corresponding to the conditional maximum likelihood method. Under appropriate regularity conditions (the existence of the conditional Fisher information and L_2-differentiability of the conditional log likelihood), such a score function always has the property of being a gain vector, at least locally.

A gain function, together with a step sequence and new observations from the model, can be used to adjust the current approximation of the drifting parameter, resulting in a tracking algorithm. To ease the verification of our assumptions on the gain function, we formulate them in two equivalent forms. Under some assumptions on the gain vectors, we establish a uniform non-asymptotic bound on the L_1 error of the resulting tracking algorithm, in terms of the variation of the drifting parameter. Under the extra assumption that the gain function is bounded, we can strengthen this result to a uniform bound on the L_p error (and then an almost sure bound). These error bounds constitute our main result and they also guide us in the choice of the step size for the algorithm. Some extensions are also presented where we allow for approximation terms and approximate gains.

Based on our main result, we specify the appropriate choice for the step sequence in three different variational setups for the drifting parameter. We treat first the simple case of a constant parameter. Although we are mainly concerned with tracking time-varying parameters, our algorithm is still of interest in the constant parameter case since it should result in an algorithm which is both recursive and robust. We also consider a setup where the parameter is stabilizing. This covers both the case where the parameter is converging and where we sample the signal with increasing frequency. The third variational setup covers the important case of tracking smooth signals. This setup is somewhat different in that we make observations with a certain frequency from an underlying continuous-time process which is indexed by a parameter changing like a Lipschitz function. Our result can then either be interpreted as a uniform, non-asymptotic result for each fixed sampling frequency or as an asymptotic statement in the observation frequency.

Examples are also given for different possible gain functions. These fall into two categories: general, score based gain functions for tracking multidimensional parameters in regular models and specialized gains for tracking more specific quantities. The latter include gains to track level sets or maxima of drifting functions (extending the classical Robbins-Monro and Kiefer-Wolfowitz algorithms) and gains to track drifting conditional quantiles. We also propose modifications for a given gain function (rescaling, truncation, projection) which can be used to design gains tailored specifically to verify our assumptions.

We illustrate our method by treating some concrete applications of the proposed algorithm but we focus mostly on the problem of tracking drifting parameters in autoregressive models. Results on tracking algorithms for these models already exist in the literature (cf. [5, 71]) and we can derive similar results by choosing an appropriate gain function. Using our approach, obtaining error bounds on the resulting tracking algorithm reduces to verifying our assumptions for the chosen gain function which considerably simplifies the derivation of results.

This chapter is structured as follows. In Section 6.2 we summarize the notation that will be used througout the remainder of the chapter, as well as our model and two equivalent formulations for our assumptions. Section 6.3 contains our main result and respective proof as well as some straightforward extensions of the main result. The construction and modification of gain functions for different models and different parameters of interest is explained in Section 6.4. Section 6.5 contains three examples of variational setups for the time-varying parameter for which we specify the tracking error implied by our main result. We collect in Section 6.6 some examples of applications and in Section 6.7 some numerical examples. Section 6.8 contains the proofs for our lemmas.

6.2 Preliminaries

First we introduce some notation that we are using throughout the chapter. All vectors are always column vectors unless explicitly transposed. We use bold uppercase letters to represent sets of vectors. For vectors $x, y \in \mathbb{R}^d$, denote by $\|x\|_2$ and $\langle x, y\rangle = x^T y$ the usual Euclidean norm and the inner product in $\mathbb{R}^d$, respectively, and by $\|x\|_p$ the l_p norm on vectors in $\mathbb{R}^d$. We will represent the indicator of the event A as $\mathbb{1}_A$. For a symmetric $d \times d$ matrix M, let $\lambda_{(1)}(M)$ and $\lambda_{(d)}(M)$ be the smallest and the largest eigenvalues of M respectively. Denote $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$. Let also O denote the zero matrix, I the identity matrix and J the exchange matrix whose dimensions will be determined by the context. We will use the convention that $\sum_{i\in\varnothing} A_i = O$ and $\prod_{i\in\varnothing} B_i = I$ for matrices A_i and B_i with such dimensions that these matrix operations (summation and product) are well defined. When applied to matrices, the symbol $|||\cdot|||_p$ will represent the operator norm induced by the l_p vector norm, which is a matrix norm defined as

$$|||A|||_p = \max_{x\neq 0} \frac{\|Ax\|_p}{\|x\|_p} = \max_{x=1} \|Ax\|_p = \max_{x\leq 1} \|Ax\|_p.$$

Assume that by time $n \in \mathbb{N}$, we have observed $\mathbf{X}_n = (X_0, X_1, \ldots, X_n)$ according to the following model:

$$X_0 \sim P_{\theta_0}, \qquad X_k|\mathbf{X}_{k-1} \sim P_{\theta_k}(\cdot|\mathbf{X}_{k-1}), \quad k \in \mathbb{N}. \tag{6.1}$$

Here the time series $(X_k,\ k \in \mathbb{N}_0)$ takes value in some set $\mathcal{X} \subseteq \mathbb{R}^l$, i.e., $\mathbb{P}(X_k \in \mathcal{X}) = 1$, $k \in \mathbb{N}_0$. Let $\mathcal{F}_k = \sigma(\mathbf{X}_k)$ denote the σ-algebra generated by $\mathbf{X}_k = (X_0, X_1, \ldots, X_k)$. The time-varying parameter $\theta_k = \theta_k(\mathbf{X}_{k-1})$, $k \in \mathbb{N}_0$, is allowed to depend on the past of the time series, i.e., it is assumed to be predictable with respect to the filtration $(\mathcal{F}_k)_{k\in\mathbb{N}}$. Further, θ_k is assumed to take values in some convex compact subset Θ of $\mathbb{R}^d$, to be precise, $P(\theta_k(\mathbf{X}_{k-1}) \in \Theta) = 1$ for all $k \in \mathbb{N}_0$. We are interested in tracking the drifting parameter $\theta_k(\mathbf{X}_{k-1})$ which we will often abbreviate as θ_k. Denote from now on

$$C_\Theta = \sup_{\theta\in\Theta} \|\theta\|_2. \tag{6.2}$$

At time $n \in \mathbb{N}$, the underlying (growing) statistical model is $\mathcal{P}^{(n)} = \mathcal{P}^{(n)}(\Theta^{n+1})$, which we can write as

$$\mathcal{P}^{(n)}(\Theta^{n+1}) = \left\{ \prod_{k=0}^{n} P_{\theta_k}(x_k|\mathbf{x}_{k-1}) : (\theta_0, \ldots, \theta_n) \in \Theta^{n+1},\ \mathbf{x}_n \in \mathcal{X}^{n+1} \right\},$$

where $P_{\theta_0}(y_0|\mathbf{x}_{-1})$ should be understood as $P_{\theta_0}(y_0)$. For $k = 0, \dots, n$, each conditional measure belongs to

$$\mathcal{P}_k = \mathcal{P}_k(\Theta) = \left\{P_\theta(\cdot|\mathbf{x}_{k-1}) : \theta \in \Theta, \mathbf{x}_{k-1} \in \mathcal{X}^k\right\}.$$

At time k, given $\mathbf{X}_k$, the model $\mathcal{P}_{k+1}$ contains all the relevant information about the next observation but we do not consider it to be (completely) known. Instead, we assume that our prior knowledge about the model is formalized as follows: for each $k \in \mathbb{N}$ we have certain $\mathbb{R}^d$-valued functions $G_k(x, \theta|\mathbf{x}_{k-1})$ at our disposal (which we will call *gain vectors* or *gain function*), $x \in \mathcal{X}$, $\mathbf{x}_{k-1} \in \mathcal{X}^k \subset \mathbb{R}^{lk}$, $\theta \in \mathbb{R}^d$, i.e., $G_k : \mathcal{X}^{k+1} \times \mathbb{R}^d \to \mathbb{R}^d$, and these gain vectors satisfy conditions (A1) and (A2) below.

(A1) For all $k \in \mathbb{N}$ and all $\theta, \vartheta \in \Theta$ the following statements hold almost surely:

$$g_k(\theta, \vartheta|\mathbf{X}_{k-1}) = \int G_k(x, \theta|\mathbf{X}_{k-1})\, dP_\vartheta(x|\mathbf{X}_{k-1}) \tag{6.3}$$

is well defined, there exists a symmetric positive-definite matrix $M_k = M_k(\mathbf{X}_{k-1})$ with (random) eigenvalues $0 < \Lambda_{(1)}(M_k) \le \cdots \le \Lambda_{(d)}(M_k)$ and constants $0 < \lambda_1 \le \lambda_2 < \infty$ such that

$$g_k(\theta, \vartheta|\mathbf{X}_{k-1}) = -M_k(\mathbf{X}_{k-1})(\theta - \vartheta), \tag{6.4}$$

with $0 < \lambda_1 \le \mathbb{E}[\Lambda_{(1)}(M_k)|\mathbf{X}_{k-2}] \le \Lambda_{(d)}(M_k) \le \lambda_2 < \infty$.

(A2) There exists a constant $C > 0$ such that for all $k \in \mathbb{N}$ and all $\theta, \vartheta \in \Theta$,

$$\mathbb{E}\|G_k(X_k, \theta|\mathbf{X}_{k-1}) - g_k(\theta, \vartheta|\mathbf{X}_{k-1})\|_2^2 \le C. \tag{6.5}$$

Note that assumption (A2) is redundant if, for example, the gain vectors $G_k(x, \theta|\mathbf{X}_{k-1})$ are almost surely bounded. Condition (A1) means, in a way, that, on average, the gain vector $G_k(X_k, \hat{\theta}_k|\mathbf{X}_{k-1})$ shifts $\hat{\theta}_k$ towards the "true" value $\theta_k = \theta_k(\mathbf{X}_{k-1})$:

$$\mathbb{E}\big[G_k(X_k, \hat{\theta}_k|\mathbf{X}_{k-1})|\mathcal{F}_{k-1}\big] = g_k(\hat{\theta}_k, \theta_k|\mathbf{X}_{k-1}) = -M_k(\mathbf{X}_{k-1})\big(\hat{\theta}_k - \theta_k\big),$$

for some symmetric, almost surely positive-definite matrix $M_k(\mathbf{X}_{k-1})$ such that $0 < \lambda_1 \le \mathbb{E}[\lambda_{(1)}(M_k)|\mathcal{F}_{k-2}] \le \lambda_{(d)}(M_k) \le \lambda_2 < \infty$.

Condition (A1) can be reformulated as (Ã1), which gives some intuition as to the role of the function g_k and which may, in certain situations, be simpler to verify.

(Ã1) The quantity $g_k(\theta, \vartheta|\mathbf{X}_{k-1})$ defined by (6.3) satisfies, almost surely, the following conditions: there exist random variables $\Lambda_1(\mathbf{X}_{k-1})$ and $\Lambda_2(\mathbf{X}_{k-1})$ and constants $0 < \lambda_1 \le \lambda_2 < \infty$, $0 < L < \infty$ such that for all $\theta, \vartheta \in \Theta$,

$$\begin{aligned} \Lambda_1(\mathbf{X}_{k-1})\|\theta - \vartheta\|_2^2 \le -(\theta - \vartheta)^T g_k(\theta, \vartheta|\mathbf{X}_{k-1}) \le \Lambda_2(\mathbf{X}_{k-1})\|\theta - \vartheta\|_2^2 \\ \|g_k(\theta, \vartheta|\mathbf{X}_{k-1})\|_2 \le L\|\theta - \vartheta\|_2 \end{aligned} \tag{6.6}$$

with $0 < \lambda_1 \le \mathbb{E}[\Lambda_1(\mathbf{X}_{k-1})|\mathbf{X}_{k-2}] \le \Lambda_2\big(\mathbf{X}_{k-1}\big) \le \lambda_2 < \infty$.

In view of the lemma below, if (A1) holds, then (Ã1) will also hold (and vice versa); the values of the constants λ_1 and λ_2 appearing in the assumptions are different, though. The proof of this lemma is deferred to Section 6.8.

Lemma 6.1 *Let $x, y \in \mathbb{R}^d$. If there exists a symmetric positive-definite matrix M such that $y = Mx$ and $0 < \lambda_1 \le \lambda_{(1)}(M) \le \lambda_{(d)}(M) \le \lambda_2 < \infty$ for some $\lambda_1, \lambda_2 \in \mathbb{R}$, then $0 < \lambda_1'\|x\|^2 \le \langle x, y\rangle \le \lambda_2'\|x\|^2 < \infty$ and $\|y\| \le C\|x\|$ for some $\lambda_1', \lambda_2', C \in \mathbb{R}$ (depending only on λ_1, λ_2) such that $0 < \lambda_1' \le \lambda_2' < \infty$ and $C > 0$.*

Conversely, if $0 < \lambda_1'\|x\|^2 \le \langle x, y\rangle \le \lambda_2'\|x\|^2 < \infty$ and $\|y\| \le C\|x\|$ for some $\lambda_1', \lambda_2', C \in \mathbb{R}$ such that $0 < \lambda_1' \le \lambda_2' < \infty$ and $C > 0$, then there exists a symmetric positive-definite matrix M such that $y = Mx$ and $0 < \lambda_1 \le \lambda_{(1)}(M) \le \lambda_{(d)}(M) \le \lambda_2 < \infty$ for some constants $\lambda_1, \lambda_2 \in \mathbb{R}$ depending only on λ_1', λ_2' and C.

At each time $k \in \mathbb{N}$, the observer should be able to calculate the gain vector at $(X_k, \mathbf{X}_{k-1})$ and an estimator $\hat\theta_k$, $G_k(X_k, \hat\theta_k|\mathbf{X}_{k-1})$, in order use it to update the estimate $\hat\theta_k$. In Section 6.4 we will show how gain functions can be constructed, but before that,in the next section, we present our tracking algorithm based on the gain function and our main result describing the quality of the algorithm.

6.3 Main result

Consider the recursive algorithm for tracking the sequence $\theta_k = \theta_k(\mathbf{X}_{k-1}) \in \Theta \subset \mathbb{R}^d$ from the observations (6.1):

$$\hat\theta_{k+1} = \hat\theta_k + \gamma_k G_k(X_k, \hat\theta_k|\mathbf{X}_{k-1}), \quad k \in \mathbb{N}, \tag{6.7}$$

for some positive sequence of step sizes $\gamma_k \le \Gamma$ and some (arbitrary) initial value $\hat\theta_0 \in \Theta \subset \mathbb{R}^d$.

Heuristically, since the gain vector $G_k(X_k, \hat\theta_k|\mathbf{X}_{k-1})$ moves, on average, $\hat\theta_k$ towards θ_k and the sequence $\theta_k \in \Theta$ (since Θ is compact) is bounded, the resulting estimating sequence $\hat\theta_k$ should also be well-behaved. The following lemma states that the second moment of $\hat\theta_k$ is uniformly bounded in $k \in \mathbb{N}$.

Lemma 6.2 *For sufficiently small γ_k there exists a constant $\bar C_\Theta$ such that*

$$\mathbb{E}\|\hat\theta_k\|_2^2 \le \bar C_\Theta^2, \qquad k \in \mathbb{N}.$$

The proof of this lemma is given in the Section 6.8. In fact, it is enough to assume that γ_k is sufficiently small for all $k \ge N$ for some fixed $N \in \mathbb{N}$. This lemma will be used in the proof of the main theorem below.

Theorem 6.1 (Error bound)
Let Assumptions (A1) and (A2) hold and $p \ge 1$. Let the tracking sequence $\hat\theta_k$ be defined by (6.7) with the sequence γ_k satisfying the conditions of Lemma 6.2, $\delta_k = \delta_k(\mathbf{X}_{k-1}) = \hat\theta_k - \theta_k$ and $\Delta_k = \Delta_k(\mathbf{X}_k) = \theta_k - \theta_{k+1}$, $k \in \mathbb{N}$. Then for any $k_0, k \in \mathbb{N}$ such that $k_0 \le k$ and $\gamma_i\lambda_2 < 1$ for all $k_0 \le i \le k$, the following relation holds:

$$\mathbb{E}\|\delta_{k+1}\|_p \le C_1 \exp\left(-\frac{\lambda_1}{2}\sum_{i=k_0}^{k}\gamma_i\right) + C_2\left(\sum_{i=k_0}^{k-1}\gamma_i^2\right)^{1/2} + C_3 \max_{i=k_0,\dots,k}\mathbb{E}\|\theta_{i+1} - \theta_{k_0}\|_2, \tag{6.8}$$

where $C_1 = (2d)^{1/2}\big(\bar C_\Theta + C_\Theta\big)$, $C_2 = d^{1/2}C^{1/2}\big(1 + \lambda_2/\lambda_1\big)$, $C_3 = d^{1/2}\big(1 + \lambda_2/\lambda_1\big)$ and C is from Assumption (A2).

If, in addition, $\Lambda_{(1)}(M_k) \ge \lambda_1$ (in Assumption (A1)) and $|G_k(X_k, \hat\theta_k|\mathbf{X}_{k-1})| \le C$ almost surely, then for any

$k_0, k \in \mathbb{N}$ such that $k_0 \le k$ and $\gamma_i \lambda_2 < 1$ for all $k_0 \le i \le k$,

$$\mathbb{E}\|\delta_{k+1}\|_p^p \le C_1 \exp\left(- p\lambda_1 \sum_{i=k_0}^{k} \gamma_i \right) + C_2 \left(\sum_{i=k_0}^{k-1} \gamma_i^2 \right)^{p/2} + C_3 \max_{i=k_0,\dots,k} \mathbb{E}\|\theta_{i+1} - \theta_{k_0}\|_p^p, \tag{6.9}$$

where $C_1 = 2^{p-1} K_p^p \mathbb{E}\|\delta_{k_0}\|_p^p$, $C_2 = d\, 2^{2p-1} B_p C^p \left(1 + K_p^2 \lambda_2/\lambda_1\right)^p$ and $C_3 = 2^{p-1}\left(1 + K_p^2 \lambda_2/\lambda_1\right)^p$.

Proof: For the sake of brevity, denote $\theta_k = \theta_k(\mathbf{X}_{k-1})$, $G_k = G(X_k, \hat{\theta}_k | \mathbf{X}_{k-1})$ and $g_k = g(\hat{\theta}_k, \theta_k | \mathbf{X}_{k-1})$, $k \in \mathbb{N}$. Recall that $\mathcal{F}_k = \sigma(\mathbf{X}_k)$ is the σ-field generated by $\mathbf{X}_k = (X_0, X_2, \dots, X_k)$.

We have

$$\mathbb{E}[G_k | \mathcal{F}_{k-1}] = g_k(\hat{\theta}_k, \theta_k | \mathbf{X}_{k-1}) = g_k, \quad k \in \mathbb{N}.$$

It follows that $D_k = G_k - g_k$, $k \in \mathbb{N}$, is a (vector) martingale difference sequence with respect to the filtration $\{\mathcal{F}_k,\ k \in \mathbb{N}_0\}$.

Rewrite the algorithm equation (6.7) as

$$\delta_{k+1} = \delta_k + \Delta\theta_k + \gamma_k D_k + \gamma_k g_k, \quad k \in \mathbb{N}.$$

In view of Assumption (A1), we have the decomposition $g_k = -M_k \delta_k$, with a symmetric positive-definite matrix $M_k = M(\hat{\theta}_k, \theta_k | \mathbf{X}_{k-1})$ so that

$$\delta_{k+1} = \Delta\theta_k + \gamma_k D_k + (I - \gamma_k M_k)\delta_k, \quad k \in \mathbb{N}. \tag{6.10}$$

By iterating the above relation, we obtain that for any $k_0 = 0, \dots, k$

$$\begin{aligned} \delta_{k+1} &= (1 - \gamma_k M_k)(I - \gamma_{k-1} M_{k-1})\delta_{k-1} + \Delta\theta_k + \gamma_k D_k \\ &\quad + (1 - \gamma_k M_k)(\Delta\theta_{k-1} + \gamma_{k-1} D_{k-1}) \\ &= \Big[\prod_{i=k_0}^{k} (I - \gamma_i M_i) \Big] \delta_{k_0} + \sum_{i=k_0}^{k} \Big[\prod_{j=i+1}^{k} (I - \gamma_j M_j) \Big] (\Delta\theta_i + \gamma_i D_i). \end{aligned} \tag{6.11}$$

Denote $A_i = \sum_{j=k_0}^{i} \gamma_j D_j$, $B_i = \sum_{j=k_0}^{i} \Delta\theta_j$ and $C_i = A_i + B_i$. Applying the vector version of the Abel transformation (Lemma 6.4) to the second term of the right hand side of (6.11) yields

$$\sum_{i=k_0}^{k} \Big[\prod_{j=i+1}^{k} (I - \gamma_j M_j) \Big] (\Delta\theta_i + \gamma_i D_i) = C_k - \sum_{i=k_0}^{k-1} \gamma_{i+1} M_{i+1} \Big[\prod_{j=i+2}^{k} (I - \gamma_j M_j) \Big] C_i. \tag{6.12}$$

Note in particular that, if we take $M_j = \lambda_1$ for $j = k_0, \dots, k$, $\Delta\theta_j = 0$, for $j = k_0, \dots, k$, $D_{k_0} = 1$ and $D_j = 0$ for $j = k_0 + 1, \dots, k$, we derive that (if $0 \le \gamma_j \lambda_1 \le 1$ for $j = k_0, \dots, k$)

$$\sum_{i=k_0}^{k-1} \lambda_1 \gamma_{i+1} \prod_{j=i+2}^{k} (1 - \gamma_j \lambda_1) = 1 - \prod_{j=k_0+1}^{k} (1 - \gamma_j \lambda_1) \le 1, \tag{6.13}$$

which we will use later.

Using (6.12), we can rewrite our expansion of δ_{k+1} in (6.11) as

$$\delta_{k+1} = \Big[\prod_{i=k_0}^{k}(I-\gamma_i M_i)\Big]\delta_{k_0} + C_k - \sum_{i=k_0}^{k-1}\gamma_{i+1}M_{i+1}\Big[\prod_{j=i+2}^{k}(I-\gamma_j M_j)\Big]C_i.$$

Take $p \in \mathbb{N}$. The previous display, the triangle inequality and the sub-multiplicative property of the operator norm ($\|MN\|_p \le \|M\|_p\|N\|_p$) imply that

$$\begin{aligned}\|\delta_{k+1}\|_p \le \|\delta_{k_0}\|_p \prod_{i=k_0}^{k}\|I-\gamma_i M_i\|_p + \|C_k\|_p \\ + \sum_{i=k_0}^{k-1}\gamma_{i+1}\|M_{i+1}\|_p\|C_i\|_p\prod_{j=i+2}^{k}\|I-\gamma_j M_j\|_p.\end{aligned} \tag{6.14}$$

Due to Assumption (A1), the matrix M_i has smallest and largest eigenvalues $\Lambda_{(1),i}$ and $\Lambda_{(d),i}$, respectively, such that almost surely $0 \le \gamma_i\Lambda_{(1),i} \le \gamma_i\Lambda_{(d),i} \le \gamma_i\lambda_2 < 1$, $k_0 \le i \le k$, and $\mathbb{E}[\Lambda_{(1),i}|\mathcal{F}_{i-2}] \ge \lambda_1 > 0$. Then,

$$\mathbb{E}\Big[(1-\gamma_k\Lambda_{(1),k})^2\Big|\mathcal{F}_{k-2}\Big] \le \mathbb{E}\Big[1-\gamma_k\Lambda_{(1),k}\Big|\mathcal{F}_{k-2}\Big] \le 1-\gamma_k\lambda_1,$$

almost surely. Similarly, $\mathbb{E}\big[1-\gamma_k\Lambda_{(1),k}\big|\mathcal{F}_{k-2}\big] \le 1-\gamma_k\lambda_1$ almost surely. It then follows by Lemma 6.3 that

$$\begin{aligned}\mathbb{E}\prod_{i=k_0}^{k}\|1-\gamma_i M_i\|_2^2 &= \mathbb{E}\prod_{i=k_0}^{k}(1-\gamma_i\Lambda_{(1),i})^2 = \mathbb{E}\left[\mathbb{E}\Big[\prod_{i=k_0}^{k}(1-\gamma_i\Lambda_{(1),i})^2\Big|\mathcal{F}_{k-2}\Big]\right] \\ &= \mathbb{E}\left[\mathbb{E}\Big[(1-\gamma_k\Lambda_{(1),k})^2\Big|\mathcal{F}_{k-2}\Big]\prod_{i=k_0}^{k-2}(I-\gamma_i\Lambda_{(1),i})^2\right] \\ &\le (1-\gamma_k\lambda_1)\mathbb{E}\prod_{i=k_0}^{k-1}(I-\gamma_i\Lambda_{(1),i})^2 \le \prod_{i=k_0}^{k}(1-\gamma_i\lambda_1),\end{aligned} \tag{6.15}$$

by iterating the recursion.

Let D_{kl} denote the l-th coordinate of the vector D_k. Clearly, for each $l = 1,\dots,d$, $\{D_{kl},\ k \in \mathbb{N}\}$ is a martingale difference with respect to the filtration $\{\mathcal{F}_k,\ k \in \mathbb{N}_0\}$. Using the fact that martingale increments are uncorrelated, we derive that for all $i = k_0,\dots,k$

$$\mathbb{E}\|A_i\|_2^2 = \mathbb{E}\sum_{l=1}^{d}\Big(\sum_{j=k_0}^{i}\gamma_j D_{jl}\Big)^2 = \sum_{l=1}^{d}\sum_{j=k_0}^{i}\gamma_j^2\mathbb{E}D_{jl}^2 = \sum_{j=k_0}^{i}\gamma_j^2\mathbb{E}\|D_j\|_2^2 \le C\sum_{j=k_0}^{k}\gamma_j^2.$$

Since B_i is a telescopic sum, we also have, for all $p \in \mathbb{N}$ and $i = k_0,\dots,k$,

$$\mathbb{E}\|B_i\|_p = \mathbb{E}\Big\|\sum_{j=k_0}^{i}\Delta\theta_j\Big\|_p = \mathbb{E}\|\theta_{i+1}-\theta_{k_0}\|_p \le \max_{i=k_0,\dots,k}\mathbb{E}\|\theta_{i+1}-\theta_{k_0}\|_p.$$

Since $\|C_i\|_2$ is $\mathcal{F}_j$-measurable for all $j \geq i$, it follows that

$$\begin{aligned}\mathbb{E}\Big[\|C_i\|_2 \prod_{j=i+2}^{k} \|I - \gamma_j M_j\|_2\Big] &= \mathbb{E}\mathbb{E}\Big[\|C_i\|_2 \prod_{j=i+2}^{k} \|I - \gamma_j M_j\|_2 \Big| \mathcal{F}_{k-2}\Big] \\ &= \mathbb{E}\Big[\mathbb{E}\big[1 - \gamma_k \Lambda_{(1),j} \big| \mathcal{F}_{k-2}\big] \|C_i\|_2 \prod_{j=i+2}^{k-1} (1 - \gamma_j \Lambda_{(1),j})\Big] \\ &\leq (1 - \gamma_k \lambda_1) \mathbb{E}\Big[\|C_i\|_2 \prod_{j=i+2}^{k-1} (1 - \gamma_j \Lambda_{(1),j})\Big] \leq \mathbb{E}\|C_i\|_2 \prod_{j=i+2}^{k} (1 - \gamma_j \lambda_1).\end{aligned}$$

Combining the last three displays, relations (6.13), (6.14) and (6.15), Lemma 6.3, the Hölder and triangle inequalities and the elementary inequality $1 - x \leq e^{-x}$, we finally get that

$$\begin{aligned}&\mathbb{E}\|\delta_{k+1}\|_2 \\ &\leq \Big(\mathbb{E}\|\delta_{k_0}\|_2^2 \mathbb{E} \prod_{i=k_0}^{k} \|I - \gamma_i M_i\|_2^2\Big)^{1/2} + \mathbb{E}\|C_k\|_2 + \mathbb{E}\Big[\sum_{i=k_0}^{k-1} \gamma_{i+1} \|M_{i+1}\|_2 \|C_i\|_2 \prod_{j=i+2}^{k} \|I - \gamma_j M_j\|_2\Big] \\ &\leq \Big(\mathbb{E}\|\delta_{k_0}\|_2^2 \prod_{i=k_0}^{k} (1 - \gamma_i \lambda_1)\Big)^{1/2} + \mathbb{E}\|C_k\|_2 + \sum_{i=k_0}^{k-1} \gamma_{i+1} \lambda_2 \mathbb{E}\Big[\|C_i\|_2 \prod_{j=i+2}^{k} \|I - \gamma_j M_j\|_2^2\Big] \\ &\leq \big(\mathbb{E}\|\delta_{k_0}\|_2^2\big)^{1/2} \exp\Big(-\frac{\lambda_1}{2} \sum_{i=k_0}^{k} \gamma_j\Big) + \max_{i=k_0,\ldots,k} \mathbb{E}\|C_i\|_2 \Big(1 + \sum_{i=k_0}^{k-1} \gamma_{i+1} \lambda_2 \prod_{j=i+2}^{k} (1 - \gamma_j \lambda_1)\Big) \\ &\leq \sqrt{2}\Big(\bar{C}_\Theta + C_\Theta\Big) \exp\Big(-\frac{\lambda_1}{2} \sum_{i=k_0}^{k} \gamma_i\Big) + \Big(1 + \frac{\lambda_2}{\lambda_1}\Big)\Big(\Big(C \sum_{i=k_0}^{k} \gamma_i^2\Big)^{1/2} + \max_{i=k_0,\ldots,k} \mathbb{E}\|\theta_{i+1} - \theta_{k_0}\|_2\Big),\end{aligned}$$

since $\mathbb{E}\|\delta_{k_0}\|_2^2 \leq 2\mathbb{E}\|\hat{\theta}_{k_0}\|_2^2 + 2\mathbb{E}\|\theta_{k_0}\|_2^2 \leq 2(\bar{C}_\Theta^2 + C_\Theta^2)$, by (6.2) and Lemma 6.2. Note that $\|\delta_{k_0}\|_2 \geq \|\delta_{k_0}\|_p$ for $p \geq 2$, $d^{1/2}\|\delta_{k_0}\|_2 \geq \|\delta_{k_0}\|_p$ for $1 \leq p < 2$. We have established the first statement of the theorem.

Let now the components of the gain function G_k be almost surely bounded, in absolute value, by a certain constant C. Using Lemma 6.3 and the elementary inequality $1 - x \leq e^{-x}$, we have that, for each $p \in \mathbb{N}$, and then some constant K_p, we can derive the following alternative expression to (6.14).

$$\begin{aligned}\|\delta_{k+1}\|_p &\leq K_p \|\delta_{k_0}\|_p \prod_{i=k_0}^{k} (1 - \gamma_i \lambda_1) + \max_{i=k_0,\ldots,k} \|C_i\|_p \Big(1 + K_p^2 \sum_{i=k_0}^{k-1} \gamma_{i+1} \lambda_2 \prod_{j=i+2}^{k} (1 - \gamma_i \lambda_1)\Big) \\ &\leq K_p \|\delta_{k_0}\|_p \exp\Big(-\lambda_1 \sum_{i=k_0}^{k} \gamma_j\Big) + \Big(1 + K_p^2 \frac{\lambda_2}{\lambda_1}\Big) \max_{i=k_0,\ldots,k} \|C_i\|_p,\end{aligned}$$

where we again use (6.13). Take now the p-th power ($p \geq 1$) of both sides of the inequality and apply the Hölder inequality $(\sum_{i=1}^{m} a_i)^p \leq m^{p-1} \sum_{i=1}^{m} |a_i|^p$ for $m = 2$ to get

$$\|\delta_{k+1}\|_p^p \leq 2^{p-1} K_p^p \|\delta_{k_0}\|_p^p \exp\Big(-p\lambda_1 \sum_{i=k_0}^{k} \gamma_j\Big) + 2^{p-1} \Big(1 + K_p^2 \frac{\lambda_2}{\lambda_1}\Big)^p \max_{i=k_0,\ldots,k-1} \|C_i\|_p^p,$$

Remember that the sequence $\big\{\sum_{j=k_0}^{i} \gamma_j D_j(\mathbf{X}_j, \hat{\theta}_j, \theta_j),\ i \geq k_0\big\}$ is a martingale with respect to the filtration $\{\mathcal{F}_i, i \in \mathbb{N}\}$ and that the entries of D_j verify $|D_{jl}| \leq 2C$, almost surely. Applying the maximal Burkholder for $p > 1$

and the Davis inequality for $p = 1$ (cf. [21, 87]) we conclude that for any $p \geq 1$, with $B_p = ((18p^{5/2})/(p-1)^{3/2})^p$,

$$\mathbb{E} \max_{i=k_0,\ldots,k-1} \|A_i\|_p^p = \mathbb{E} \max_{i=k_0,\ldots,k-1} \sum_{l=1}^{d} \left| \sum_{j=k_0}^{i} \gamma_j D_{jl} \right|^p \leq \sum_{l=1}^{d} \mathbb{E} \max_{i=k_0,\ldots,k-1} \left| \sum_{j=k_0}^{i} \gamma_j D_{jl} \right|^p$$
$$\leq B_p \sum_{l=1}^{d} \mathbb{E} \left| \sum_{j=k_0}^{k-1} \gamma_j^2 D_{jl}^2 \right|^{p/2} \leq d B_p 2^p C^p \left| \sum_{j=k_0}^{k-1} \gamma_j^2 \right|^{p/2},$$

The second inequality of the theorem now follows by taking expectations on both sides of the bound on $\|\delta_{k+1}\|_p^p$ above, by using the last inequality, (6.13) and the fact that $\|C_i\|_p^p \leq 2^{p-1}\|A_i\|_p^p + 2^{p-1}\|B_i\|_p^p$. □

Remark 6.1 *Sometimes we will not be interested in tracking the, say,* natural *parameter θ_k of the model but some other parameter ϑ_k which is, on average, close to θ_k. The difference $\|\theta_k - \vartheta_k\|_p$ can be seen as an approximation term in that the parameter θ_k driving the time series is actually an approximation for our parameter of interest ϑ_k. Denoting $\hat{\theta}_k - \vartheta_k$ as δ_k^*, the following expansion can be derived,*

$$\begin{aligned}
\delta_{k+1}^* &= \delta_k^* + \Delta\vartheta_k + \gamma_k D_k - \gamma_k M_k(\hat{\theta}_k - \theta_k) \\
&= \Delta\vartheta_k + \gamma_k M_k(\theta_k - \vartheta_k) + \gamma_k D_k + (I - \gamma_k M_k)\delta_k^* \\
&= \Big[\prod_{i=k_0}^{k}(I - \gamma_i M_i)\Big]\delta_{k_0}^* + \sum_{i=k_0}^{k}\Big[\prod_{j=i+1}^{k}(I - \gamma_j M_j)\Big](\Delta\vartheta_k + \gamma_k M_k(\theta_k - \vartheta_k) + \gamma_i D_i).
\end{aligned}$$

The same note could be made for situations where $g_k = -M_k(\hat{\theta}_k - \theta_k - \eta_k)$ where η_k is a remainder term which may be random so long as it is measurable with respect to $\sigma(\mathbf{X}_{k-1})$; it would then follow that

$$\delta_{k+1} = \Big[\prod_{i=k_0}^{k}(I - \gamma_i M_i)\Big]\delta_{k_0} + \sum_{i=k_0}^{k}\Big[\prod_{j=i+1}^{k}(I - \gamma_j M_j)\Big](\Delta\theta_i - \eta_i + \gamma_i D_i).$$

Noting that $\|\gamma_k M_k(\theta_k - \vartheta_k)\|_p < \lambda_2 \gamma_k K_p \|\theta_k - \vartheta_k\|_p$ we conclude, for the same constants C_1, C_2, C_3 as before and all $p \in \mathbb{N}$, that the following also hold

$$\begin{aligned}
\mathbb{E}\|\delta_{k+1}\|_p \leq & C_1 \exp\left(-\frac{\lambda_1}{2}\sum_{i=k_0}^{k}\gamma_j\right) + C_2\left(\sum_{i=k_0}^{k-1}\gamma_i^2\right)^{1/2} \\
& + C_3 \mathbb{E} \max_{i=k_0,\ldots,k} \|\vartheta_{i+1} - \vartheta_{k_0}\|_2 + \lambda_2 K_2 \mathbb{E} \sum_{i=k_0}^{k} \gamma_i \|\eta_i\|_2,
\end{aligned} \tag{6.16}$$

$$\begin{aligned}
\mathbb{E}\|\delta_{k+1}\|_p^p \leq & C_1 \exp\left(-p\lambda_1\sum_{i=k_0}^{k}\gamma_j\right) + C_2\left(\sum_{i=k_0}^{k-1}\gamma_i^2\right)^{p/2} \\
& + C_3 \mathbb{E}\left(\max_{i=k_0,\ldots,k} \|\vartheta_{i+1} - \vartheta_{k_0}\|_p + \lambda_2 K_p \sum_{i=k_0}^{k} \gamma_i \|\eta_i\|_p\right)^p,
\end{aligned} \tag{6.17}$$

where either a) $\delta_k = \hat{\theta}_k - \vartheta_k$ and $\eta_k = \theta_k - \vartheta_k$, b) $\delta_k = \hat{\theta}_k - \theta_k$, $\vartheta_k = \theta_k$ and η_k such that $g_k = -M_k(\delta_k - \eta_k)$; (6.16) and (6.17) generalize then the bounds in (6.8) and (6.9) where we had c) $\delta_k = \hat{\theta}_k - \theta_k$, $\vartheta_k = \theta_k$ and $\eta_k = 0$.

Remark 6.2 *If we are interested in tracking $\vartheta_k = \varphi(\theta_k)$, functional of the parameter θ_k with uniformly bounded derivatives, then by using Taylor's Theorem, our Theorem 6.1 above straightforwardly delivers a bound on the expectation of $\|\hat{\vartheta}_k - \vartheta_k\|_p = \|\varphi(\hat{\theta}_k) - \varphi(\theta_k)\|_p$ and its powers.*

6.4 Construction of gain functions

In this section we address the construction, or choice, of appropriate gain functions to be used with the algorithm (6.7). Any gain function for which conditions (A1) and (A2) hold may be used in our algorithm, and whether a particular gain function is suitable or not depends exclusively on the model under study. Namely, this will depend on the way in which the distributions in the model depend on the parameter which we are interested in tracking. For certain types of models, there might be natural choices for the gain function. As before we abbreviate $\theta_k = \theta_k(\mathbf{X}_{k-1})$.

A situation, which essentially extends the original setup in which [76] developed their classical algorithm, is when the data, $\mathbf{X}_k = (X_1, \ldots, X_k)$, is such that

$$X_k = \vartheta_k(\mathbf{X}_{k-1}) + \xi_k(\mathbf{X}_{k-1}),$$

where the $\vartheta_k(\cdot)$ are functions of $\mathbf{X}_{k-1}$, and $\xi_k(\mathbf{X}_{k-1})$ are martingale difference noise terms which may also depend on $\mathbf{X}_{k-1}$. In this case, given $\mathbf{X}_{k-1}$ we may simply take

$$G_k(x, \theta|\mathbf{X}_{k-1}) = x - \theta \tag{6.18}$$

since for each θ,

$$g_k(\theta, \vartheta_k(\mathbf{X}_{k-1})|\mathbf{X}_{k-1}) = \mathbb{E}_\vartheta[G_k(X_k, \theta|X_{k-1})|\mathbf{X}_{k-1}] = -(\theta - \vartheta_k(\mathbf{X}_{k-1})).$$

Non-parametric regression is an example of a model which fits into this situation and for which our results may be used.

It could also be that $\mathbb{E}_\theta[X_k|\mathbf{X}_{k-1}]$, the conditional expectation of the data, given the past, is not θ but instead $\phi(\theta)$ for some some smooth function ϕ. In this case, given $\mathbf{X}_{k-1}$, one should consider instead,

$$G_k(x, \theta|\mathbf{X}_{k-1}) = x - \phi(\theta) \tag{6.19}$$

and then, for each θ,

$$g_k(\theta, \vartheta_k|\mathbf{X}_{k-1}) = \mathbb{E}_\vartheta[G_k(X_k, \theta|X_{k-1})|\mathbf{X}_{k-1}] = -\big(\phi(\theta) - \phi(\vartheta_k)\big).$$

The term on the far right should then be comparable to $-(\theta - \vartheta_k)$. Autoregressive models, for example, fall into this category (cf. Section 6.6.4).

One may also consider more dynamical situations where the observations themselves depend on our tracking sequence. An example of such a setup is the [53] algorithm where we would like to track the sequence of (unique) maxima of a sequence of functions $\vartheta_k : \Theta \subset \mathbb{R}^d \mapsto \mathbb{R}$, $k \in \mathbb{N}$, which we may observe at any point, corrupted with

white noise. One possibility (cf. [61]) is to use gain functions defined using random directions. Let then D_k, $k \in \mathbb{N}$, denote a random sequence of independent unit vectors. We would consider, for a positive sequence e_k, $k \in \mathbb{N}$, the gain function

$$G_k(X_k^-, X_k^+, \hat{\theta}_k | \mathbf{X}_{k-1}^-, \mathbf{X}_{k-1}^+, \mathbf{D}_{k-1}) = D_k \frac{X_k^-(\hat{\theta}_k) - X_k^+(\hat{\theta}_k)}{2e_k}, \tag{6.20}$$

where, with some abuse of notation, the observations $X_{k+1}^{\pm}(\theta_k)$, are given by

$$X_{k+1}^{\pm}(\hat{\theta}_k) = \vartheta_k\big(\hat{\theta}_k \pm e_k D_k\big) + \xi_k^{\pm},$$

for $\hat{\theta}_k$ the tracking sequence defined by the gain (6.20) and $\xi_k^{\pm}$ independent, zero mean noise. Let for each $k \in \mathbb{N}$, θ_k be the unique maximum of $\vartheta_k(\cdot)$. In this case we would have, for the filtration $\mathcal{F}_k = \sigma\big(\mathbf{X}_k^{\pm}, \mathbf{D}_k\big)$,

$$\begin{aligned} g_k(\hat{\theta}_k, \theta_k | \mathcal{F}_{k-1}) &= \mathbb{E}\Big[- D_k D_k^T \nabla \vartheta_k(\hat{\theta}_k) + H_k(\hat{\theta}_k) + D_k \frac{\xi_k^- - \xi_k^+}{2e_k} \Big| \mathcal{F}_{k-1} \Big] \\ &= -\mathbb{E}\big[D_k D_k^T\big] \nabla \vartheta_k(\hat{\theta}_k) + \mathbb{E}\big[H_k(\hat{\theta}_k) \big| \mathcal{F}_{k-1}\big] + \frac{\mathbb{E}\big[D_k(\xi_k^- - \xi_k^+)\big]}{2e_k} \\ &= -\mathbb{E}\big[D_k D_k^T\big] \nabla^2 \vartheta_k(\theta_k^*)(\hat{\theta}_k - \theta_k) + \eta_k, \end{aligned}$$

where $\nabla^2 \vartheta_k(\cdot)$ is the Hessian of $\vartheta_k(\cdot)$, $\theta_k^* \in \Theta$ and, for $\theta \in \Theta$,

$$H_k(\theta) = D_k D_k^T \nabla \vartheta_k(\theta) - D_k \frac{\vartheta_k(\theta + e_k D_k) - \vartheta_k(\theta - e_k D_k)}{2e_k}.$$

Conditions (A1) and (A2) will hold if, for example, we assume that the random directions where chosen such that $\mathbb{E}\big[D_k D_k^T\big]$ are positive-definite matrices, that the Hessian $\nabla^2 \vartheta_k(\cdot)$ is positive-definite over Θ and that for appropriately small e_k the expectation $\mathbb{E}[\|\eta_k\|_p]$ is appropriately small, uniformly over $\theta \in \Theta$. These conditions are comparable to the ones in the original formulation of the Kiefer-Wolfowitz algorithm, and can be significantly relaxed by, for example, considering different types of expansions for g_k depending on how large the norm of $\delta_k = \hat{\theta}_k - \theta_k$ is.

Consider now a different example. Say $\mathcal{X} \subset \mathbb{R}$ and, given the past of the process, $\mathbf{X}_{k-1}$, we would like to track a conditional quantile of a certain distribution, i.e., we would like to track $\vartheta_k = \vartheta_k(\mathbf{X}_{k-1})$ such that $\vartheta_k = \inf\big\{x \in \mathcal{X} : F_k(x|\mathbf{X}_{k-1}) \geq \alpha_k\big\}$, where α_k is a sequence in $(0,1)^{\mathbb{N}}$ of our choice and $F_k(\cdot|\mathbf{X}_{k-1})$ the cumulative distribution function of $X_k|\mathbf{X}_{k-1}$. In this case it makes sense to use

$$G_k(x, \theta | \mathbf{X}_{k-1}) = \alpha_k - I\{x - \theta \leq 0\} \tag{6.21}$$

since we see that

$$g_k(\theta, \vartheta_k | \mathbf{X}_{k-1}) = \mathbb{E}[G_k(X_k, \theta | \mathbf{X}_{k-1}) | \mathbf{X}_{k-1}] = -\big(F_k(\theta - \vartheta_k | \mathbf{X}_{k-1}) - \alpha_k\big),$$

where we assume without loss of generality that the distribution is centered around the quantile ϑ_k. The quantity in the last display clearly has the same sign as $\vartheta_k - \theta$. Note also that the algorithm based on this gain function

only requires knowledge of the values of the indicators $\mathbb{1}\{X_k - \theta \leq 0\}$ which means that we may still track the required quantiles without explicitly observing X_k. This problem is treated in detail for the case of independent observations in [7] and for the more general case where the observations are not independent in Chapter 5.

For certain models it might, however, not be obvious how gain functions can be constructed, especially when tracking multi-dimensional parameters. It is therefore important to have a general procedure that can be used to construct candidate gain functions that can either be used directly or, if needed, modified to verify (A1) and (A2).

Assume that for each $k \in \mathbb{N}$, each distribution from the family of conditional distributions $\mathcal{P}_k = \{P_\theta(x|\mathbf{X}_{k-1}), \theta \in \Theta\}$ has a density with respect to some σ-finite dominating measure μ and denote this conditional density by $p_\theta(x|\mathbf{X}_{k-1})$, $\theta = (\theta_1, \ldots, \theta_d) \in \Theta$. Assume also that there is a common support $\mathcal{X}$ for these densities, and that for any $x \in \mathcal{X}$ and $\theta \in \Theta \subset \mathbb{R}^d$, the partial derivatives $\partial p_\theta(x|\mathbf{X}_{k-1})/\partial\theta_i$, $i = 1, \ldots, d$, exist and are finite, almost surely. Under these assumptions, the *conditional* gradient vector

$$\nabla_\theta \log p_\theta(x|\mathbf{X}_{k-1}) = \Big(\partial \log p_\theta(x|\mathbf{X}_{k-1})/\partial\theta_1, \ldots, \partial \log p_\theta(x|\mathbf{X}_{k-1})/\partial\theta_d\Big) \tag{6.22}$$

and the square, random matrices $I_k(\theta|\mathbf{X}_{k-1})$ with entries

$$I_{k,i,j}(\theta|\mathbf{X}_{k-1}) = \mathbb{E}_\theta\left[\frac{\partial}{\partial\theta_i} p_\theta(x|\mathbf{X}_{k-1}) \cdot \frac{\partial}{\partial\theta_j} p_\theta(x|\mathbf{X}_{k-1})\right] \tag{6.23}$$

for $i, j = 1, \ldots, d$, can be defined, almost surely. A possible gain function is simply the conditional score of the model, i.e. the gradient vector

$$G_k(x, \theta|\mathbf{X}_{k-1}) = \nabla_\theta \log p_\theta(x|\mathbf{X}_{k-1}). \tag{6.24}$$

If (6.23) is almost surely non-singular then one might also consider

$$G_k(x, \theta|\mathbf{X}_{k-1}) = I_k^{-1}(\theta|\mathbf{X}_{k-1})\nabla_\theta \log p_\theta(x|\mathbf{X}_{k-1}). \tag{6.25}$$

We justify now why these choices are reasonable. Take $\vartheta = (\vartheta_1, \ldots, \vartheta_d) \in \Theta$. It is not uncommon for the Kullback-Leibler divergence $K\big(P_\vartheta(x|\mathbf{X}_{k-1}), P_\theta(x|\mathbf{X}_{k-1})\big)$ to be a quadratic form in the distance between the parameters θ and ϑ, i.e., equal to a multiple of $(\theta - \vartheta)^T M(\theta - \vartheta)$ for some (eventually random) positive semi-definite matrix M. If so, under the assumption that we can interchange integration and differentiation and that M does not depend on θ, $g_k(\theta, \vartheta|\mathbf{X}_{k-1})$ will almost surely reduce to

$$\begin{aligned}
&\int \nabla_\theta \log p_\theta(x|\mathbf{X}_{k-1}) dP_\vartheta(x|\mathbf{X}_{k-1}) = \nabla_\theta \int \log p_\theta(x|\mathbf{X}_{k-1}) dP_\vartheta(x|\mathbf{X}_{k-1}) \\
=&\nabla_\theta\Big(\int \log \frac{p_\theta(x|\mathbf{X}_{k-1})}{p_\vartheta(x|\mathbf{X}_{k-1})} dP_\vartheta(x|\mathbf{X}_{k-1}) + \int \log p_\vartheta(x|\mathbf{X}_{k-1}) dP_\vartheta(x|\mathbf{y}_{k-1})\Big) \\
=&\nabla_\theta \int \log \frac{p_\theta(x|\mathbf{X}_{k-1})}{p_\vartheta(x|\mathbf{X}_{k-1})} dP_\vartheta(x|\mathbf{X}_{k-1}) = -\nabla_\theta K\big(P_\vartheta(x|\mathbf{X}_{k-1}), P_\theta(x|\mathbf{X}_{k-1})\big) \\
=&-\nabla_\theta(\theta - \vartheta)^T M(\theta - \vartheta) = -2M(\theta - \vartheta).
\end{aligned}$$

The score will in principle depend on the past of the chain $\mathbf{X}_{k-1}$ and the previous argument might only be

valid for a certain subset of values $\mathbf{X}_{k-1}$ in $\mathcal{X}^{k-1}$. This dependence could prevent (A1) from holding. In these cases, using the form (6.25) might be a good alternative since the matrix $I_k^{-1}(\theta|\mathbf{X}_{k-1})$ will act as an appropriate scaling factor.

The dependence of the gain function on the past of the time series is in fact one of the main issues one has to deal with when checking (A1) and (A2). On one hand, to ensure that the gain function has, on average, the right direction, as required by (6.3), the gain will often need to depend on previous observations. This might, however, affect either the range or the variance of the gain. Gain function, such as (6.24) and (6.25), can be modified, or rescaled, to ensure that the respective conditional expectation $g_k(\theta, \vartheta|\mathbf{X}_{k-1})$ verifies the assumptions of Theorem 6.1. One can for example truncate certain entries or factors in both $G_k(x, \theta|\mathbf{X}_{k-1})$ and $I_k(\theta|\mathbf{X}_{k-1})$ to ensure that the resulting $g_k(\theta, \vartheta|\mathbf{X}_{k-1})$ follows the required assumptions. Another possibility is to rescale, or directly truncate, the length of a given gain vector and consider, for example, one of the following gains

$$\begin{aligned}
\tilde{G}_k(x, \theta|\mathbf{X}_{k-1}) &= \frac{G_k(x, \theta|\mathbf{X}_{k-1})}{1 + \|G_k(x, \theta|\mathbf{X}_{k-1})\|_2}, \\
\mathring{G}_k(x, \theta|\mathbf{X}_{k-1}) &= G_k(x, \theta|\mathbf{X}_{k-1})\Big(1 + \frac{\kappa - \|G_k(x, \theta|\mathbf{X}_{k-1})\|_2}{\|G_k(x, \theta|\mathbf{X}_{k-1})\|_2} \mathbb{1}\{\|G_k(x, \theta|\mathbf{X}_{k-1})\|_2 \geq \kappa\}\Big), \\
\bar{G}_k(x, \theta|\mathbf{X}_{k-1}) &= G_k(x, \theta|\mathbf{X}_{k-1}) \frac{\min\big(s(\mathbf{X}_{k-1}), \kappa\big)}{s(\mathbf{X}_{k-1})},
\end{aligned}$$

for G_k an arbitrary gain function, $\kappa > 0$ and some function $s : \mathcal{X}^k \mapsto \mathbb{R}^+$. Note that $\tilde{G}_k$, $\mathring{G}_k$ and $\bar{G}_k$ all preserve the direction of G_k and have norm bounded by respectively 1, κ, and the norm of G_k, almost surely.

The gain $\bar{G}_k$ is specifically rescaled for situations where we have a conditional gain g_k almost surely of the form $g_k = -s(\mathbf{X}_{k-1})M_k(\theta - \vartheta)$, where M_k has eigenvalues as prescribed by (A1). Consequently we will have that $\bar{g}_k = -\min\big(s(\mathbf{X}_{k-1}), \kappa\big)M_k(\theta - \vartheta)$ from where it follows that the largest eigenvalue of the matrix $\min\big(s(\mathbf{X}_{k-1}), \kappa\big)M_k$ will then be almost surely upper-bounded; in certain situations it will be possible to use the fact that almost surely $E[\Lambda_{(1)}(M_k)|\mathbf{X}_{k-2}] \geq \lambda_1$, to show that $E[\min\big(s(\mathbf{X}_{k-1}), \kappa\big)\Lambda_{(1)}(M_k)|\mathbf{X}_{k-2}] \geq c\lambda_1$ for some $0 < c \leq 1$ and sufficiently large κ. Using the fact that the function $\min(x, \kappa)/x \leq 1$ we have, again abbreviating $\bar{G}_k(X_k, \theta|\mathbf{X}_{k-1})$ and $\bar{g}_k(\theta, \vartheta|\mathbf{X}_{k-1})$

$$\begin{aligned}
&\mathbb{E}\mathbb{E}_\vartheta\big[\|\bar{G}_k - \bar{g}_k\|_2^2|\mathbf{X}_{k-1}\big] = \\
&\qquad \mathbb{E}\Big[\Big(\frac{\min\big(s(\mathbf{X}_{k-1}), \kappa\big)}{s(\mathbf{X}_{k-1})}\Big)^2 \mathbb{E}_\vartheta\big[\|G_k - g_k\|_2^2|\mathbf{X}_{k-1}\big]\Big] \leq \mathbb{E}\|G_k - g_k\|_2^2,
\end{aligned} \tag{6.26}$$

such that if G_k verifies (A2) then so will $\bar{G}_k$.

Another possible modification one might consider, is to truncate the iterates of the our algorithm (6.7). This might be motivated by practical considerations in the case where the parameter being tracked has some sort of physical meaning and is therefore bounded; it stands to reason then that the algorithm itself should be restricted as well. We would then, for a parameter set Θ, consider the sequence

$$\hat{\theta}_{k+1} = \Pi_{\bar{\Theta}}\big(\hat{\theta}_k + \gamma_k G_k(X_k, \hat{\theta}_k|\mathbf{X}_{k-1})\big), \quad k \in \mathbb{N}, \tag{6.27}$$

where $\Pi_{\bar{\Theta}}(\cdot)$ acts as a projection on a convex set $\bar{\Theta} \supset \Theta$ in that $\Pi_{\bar{\Theta}}(\cdot)$ is an identity on $\bar{\Theta}$ and maps points in $\bar{\Theta}^c$ to $\bar{\Theta}$.

We will provide concrete examples of gain functions later in Section 6.6. Before this, we present in Section 6.5 some examples of different types of variation that the parameter of the model may have such that our algorithm is capable of adequately tracking it.

6.5 Variational setups for the drifting parameter

It is clear – and in fact explicit in (6.8) and (6.9) – that the changes in the parameter have a non-negligible contribution to the accuracy of our tracking algorithm. This is reasonable since, if the parameter changes arbitrarily in-between observations, we should not expect it to the "trackable". We must then specify how the parameter is allowed to vary and, based on that assumption, pick an appropriate sequence γ_k which minimizes the general bounds in (6.8) or (6.9). We will specify in this section what these bounds reduce to for concrete examples for the variation of the parameter being tracked. These examples refer only to how the parameter is assumed to change and are unrelated to the actual model in question; examples of specific models can be found in Section 6.6.

6.5.1 Static parameter

We assume in this section that $\theta_j(\mathbf{X}_{j-1}) = \theta_0$, almost surely, $\forall\, j \in \mathbb{N}$ for some unknown $\theta_0 \in \Theta$ such that in fact $\Delta\theta_j = \mathbf{0}$, almost surely, and we are actually in a parametric setup. Note that, in this case, the second terms in both (6.8) and (6.9) obviously vanish.

Take then $\gamma_j = C_\gamma j^{-1} \log j$ and for $q \in (0,1)$, $n_0 = [qn]$, where $[a]$ is the whole part of $a \in \mathbb{R}$. Let $n \geq 2/q = N_q$ such that $n_0 \geq 2$. For large enough C_γ and all $n \geq N_q$ we have,

$$\sum_{j=n_0}^{n} \gamma_j \geq c_\gamma \log n_0 \sum_{j=n_0}^{n} \frac{1}{k} \geq \frac{\log n}{2\lambda_1},$$

from where for all $p \in \mathbb{N}$,

$$\exp\Big(- p\lambda_1 \sum_{j=n_0}^{n} \gamma_j \Big) \leq n^{-p/2}.$$

Note that in the case where we have $\mathbb{E}\|\delta_{n_0}\|_p^p \leq C_0 n_0^p$ we can take the constant C_γ to be larger (say take rC_γ, $r > 2$) in which case

$$C_1 \exp\Big(- p\lambda_1 \sum_{j=n_0}^{n} \gamma_j \Big) \leq c_1 n^p n^{-rp/2} \leq C_1 n^{-p/2}.$$

Using now the fact that $\sum_{j=n_0}^{n} \gamma_j^2 \leq c(\log n)^2 n^{-1}$ for some constant $c > 0$ we have

$$\left(\sum_{j=n_0}^{n} \gamma_j^2 \right)^{p/2} \leq (n^{-1/2} \log n)^p.$$

We conclude that we can rewrite (6.8) and (6.9) respectively as

$$\max_{n \geq N_q} \mathbb{E} \frac{\sqrt{n}}{\log n} \|\delta_n\|_p \leq C \quad \text{and} \quad \max_{n \geq N_q} \mathbb{E}\Big(\frac{\sqrt{n}}{\log n} \|\delta_n\|_p \Big)^p \leq C,$$

for all $p \in \mathbb{N}$. The log term in the rate cannot be avoided and is a consequence of the recursiveness of the algorithm.

Note that by taking $p > \varepsilon^{-1}$ and, by using Markov's inequality and the second bound in the previous display, we conclude that

$$\begin{aligned}\sum_{n=1}^{\infty} P\big(n^{1/2-\varepsilon}\|\hat{\theta}_n - \theta_0\|_1 > c\big) &\leq \sum_{n=1}^{\infty} P\big(d^{\frac{p-1}{p}} n^{1/2-\varepsilon}\|\hat{\theta}_n - \theta_0\|_p > c\big) \\ &\leq \sum_{n=0}^{\infty} \frac{d^{p-1} n^{p/2-p\varepsilon}\mathbb{E}\|\delta_n\|_p^p}{c^p} \leq C\sum_{n=1}^{\infty} \frac{(d\log n)^p}{n^{p\varepsilon}} < \infty.\end{aligned} \tag{6.28}$$

By application of the Borel-Cantelli Lemma, we conclude that $\|\hat{\theta}_n - \theta_0\|_1 \to 0$ as $n \to 0$ takes place with probability 1 at a rate $n^{1/2-\varepsilon}$ for all $\varepsilon > 0$.

The particular setup presented in this section, where the parameter is fixed, might seem out of place since we are mainly concerned with tracking time-changing parameters. We would like to point out, however, that our algorithm is recursive and, as such, always produces estimates in a fast, straightforward fashion. This is an advantage especially over "offline" estimators obtained, say, as solutions to a certain system, which require iterative likelihood or least squares optimization or are obtained via other indirect methods, a situation which is common when dealing with Markov models (cf. Section 6.6.4.)

6.5.2 Stabilizing parameter

Suppose now that the parameter we want to track is stabilizing. This situation might arise if the expectation of the sequence of values that the parameter takes is converging to some limiting value. It could also be the case that the data is being sampled, with increasing frequency, from an underlying, continuous-time process which depends on a parameter varying continuously; in this case, the parameter varies less and less since it is allowed less time to change. Regardless, we assume that $\Delta\theta_i = \theta_i(\mathbf{X}_{i-1}) - \theta_{i+1}(\mathbf{X}_i)$ verifies

$$\mathbb{E}\|\Delta\theta_i\|_p^p \leq \rho_i^p, \, i \in \mathbb{N}$$

for $p \geq 1$ and some decreasing sequence ρ_i. Assume then that we have $\rho_i = c_\rho i^{-\beta}$ for some constant $c_\rho > 0$ and $\beta \geq 0$.

Consider first the case $\beta \geq 3/2$. In this case, the variation of the parameter vanishes so quickly that we are essentially in the setup of the previous section. Indeed, take γ_i and n_0 as in the previous section. The first and third term in both (6.8) and (6.9) can be bounded in the same way as in the previous section. As for the second term, by the Hölder inequality,

$$\begin{aligned}\mathbb{E}\Big(\sum_{i=n_0}^{n}\|\Delta\theta_i\|_p\Big)^p &\leq (n-n_0)^{p-1}\sum_{i=n_0}^{n}\mathbb{E}\|\Delta\theta_i\|_p^p \leq C(n-n_0)^p\rho_{n_0}^p \\ &\leq c\big((n-n_0)n_0^{-\beta}\big)^p \leq Cn^{-(\beta-1)p} \leq Cn^{-p/2},\end{aligned} \tag{6.29}$$

leading to the same bounds as in the previous section

Consider now the case where $0 < \beta < 3/2$. Let $\gamma_i = C_\gamma(\log i)^{1/3} i^{-2\beta/3}$, $n_0 = n - n^{2\beta/3}(\log n)^{2/3}$. By using the elementary inequality $(1+x)^\alpha \leq 1 + \alpha x$ for $0 < \alpha < 1$ and $x \geq -1$, we obtain that for sufficiently large n (i.e.,

$n \ge N_1 = N_1(\beta)$) and sufficiently large constant C_γ

$$\begin{aligned}\sum_{i=n_0}^{n} \gamma_i &\ge C_\gamma (\log n_0)^{1/3} \sum_{i=n_0}^{n} \frac{1}{i^{2\beta/3}} \ge C_\gamma (\log n_0)^{1/3} \int_{n_0}^{n} \frac{dx}{x^{2\beta/3}} \\ &= \frac{C_\gamma (\log n_0)^{1/3}}{1-2\beta/3}\Big[n^{1-2\beta/3} - n^{1-2\beta/3}\big(1 - n^{2\beta/3-1}(\log n)^{2/3}\big)^{1-2\beta/3}\Big] \\ &\ge \frac{C_\gamma (\log n_0)^{1/3}}{1-2\beta/3}\Big[n^{1-2\beta/3} - n^{1-2\beta/3}\big(1 - n^{2\beta/3-1}(\log n)^{2/3}(1-2\beta/3)\big)\Big] \\ &= C_\gamma (\log n_0)^{1/3}(\log n)^{2/3} \ge \frac{\log n}{2h}.\end{aligned}$$

This yields the same bound for the first terms in (6.8) and (6.9): for $n \ge N_1$, sufficiently large constant C_γ and all $p \in \mathbb{N}$,

$$C_1 \exp\Big(-ph \sum_{i=n_0}^{n} \gamma_i\Big) \le C_1 n^{-p/2}.$$

Let us now bound the last terms in (6.8) and (6.9): for $n \ge N_2 = N_2(\beta)$ and all $p \in \mathbb{N}$,

$$\Big(\sum_{i=n_0}^{n} \gamma_i^2\Big)^{p/2} \le C\big((\log n)^{2/3} n_0^{-4\beta/3}(n-n_0)\big)^{p/2} \le c\big((\log n)^{2/3} n^{-\beta/3}\big)^p.$$

For sufficiently large n (i.e., $n \ge N_3 = N_3(\beta)$) the second terms in (6.8) and (6.9) are bounded similarly to (6.29) by

$$\mathbb{E}\Big(\sum_{i=n_0}^{n} \|\Delta\theta_i\|_p\Big)^p \le c\big((n-n_0)n_0^{-\beta}\big)^p \le C\big((\log n)^{2/3} n^{-\beta/3}\big)^p.$$

Finally we obtain that for $0 < \beta < 3/2$ and a sufficiently large constant C_γ in the algorithm step $\gamma_i = C_\gamma (\log i)^{1/3} i^{-2\beta/3}$, (6.8) and (6.9) can be rewritten respectively as

$$\max_{n \ge N_\beta} \mathbb{E} \frac{n^{\beta/3}}{(\log n)^{2/3}} \|\delta_n\|_2 \quad \text{and} \quad \max_{n \ge N_\beta} \mathbb{E}\Big(\frac{n^{\beta/3}}{(\log n)^{2/3}} \|\delta_n\|_p\Big)^p \le c,$$

where $N_\beta = \max(N_1, N_2, N_3)$ is the burn-in period of the algorithm.

Remark 6.3 *If we choose $\gamma_i = C_\gamma (\log i)^{\alpha_1} i^{-\alpha}$ and $n_0 = n - n^\alpha (\log n)^{\alpha_2}$, $0 < \alpha < 1$, $\alpha_1, \alpha_2 \ge 0$, $\alpha_1 + \alpha_2 \ge 1$ in case $0 < \beta < 3/2$, then we get the following bound of the convergence rate: for sufficiently large n and sufficiently large constant C_γ*

$$\mathbb{E}\|\delta_n\|_p^p \le C\Big(n^{-\min\{\beta-\alpha, \alpha/2\}} (\log n)^{\max\{\alpha_2, \alpha_1+\alpha_2/2\}}\Big)^p.$$

Thus , the choice $\alpha = 2\beta/3$, $\alpha_1 = 1/3$, $\alpha_2 = 2/3$ is optimal in the sense of the minimum of the right-hand side of the above inequality.

Remark 6.4 *Much in the same way as for (6.28), we can establish that for any $\varepsilon > 0$, $\lim_{n\to\infty} n^{\beta/3-\varepsilon}\|\delta_n\|_1 = 0$ with probability 1.*

Finally, consider the case $\beta = 0$, i.e., we assume the following weak requirement: $\mathbb{E}\|\Delta\theta_i\|_p^p \le c$, $i \in \mathbb{N}$, for some uniform constant c. Take $n - n_0 = N$, $\gamma_i = \gamma$ for some $N \in \mathbb{N}$, $\gamma > 0$. Then Theorem 6.1 implies that

$$\max_{n \ge N} \mathbb{E}\|\delta_n\|_p^p \le C_1 e^{-phN\gamma} + C_2 N^{p/2}\gamma^p + C_3 N^p c = D.$$

We thus have that the algorithm will track down the parameter in the proximity of size D, which we can try to minimize by choosing appropriate constants N and γ.

6.5.3 Lipschitz signal with asymptotics in the sampling frequency

We consider now a slightly different setup where we assume that the parameter is changing, on average, like a Lipschitz function. In this setup we let the time series (6.1) be sampled from a continuous-time process X_t, $t \in [0,1]$ which we observe with frequency n. This means that for each $n \in \mathbb{N}$ we have a different model, namely,

$$X_0^n \sim P_{\theta_0^n}, \qquad X_k^n | \mathbf{X}_{k-1}^n \sim P_{\theta_k^n}(\cdot | \mathbf{X}_{k-1}^n), \quad k \le n \in \mathbb{N}, \tag{6.30}$$

where the parameter $\theta_k^n = \theta_k^n(\mathbf{X}_{k-1}^n)$ verifies, for some $p \in \mathbb{N}$ and $\kappa_{d,p} < \infty$

$$\mathbb{E}\|\theta_k^n(\mathbf{X}_{k-1}^n) - \theta_{k_0}^n(\mathbf{X}_k^n)\|_p^p \le \kappa_{d,p}^p \Big(\frac{k-k_0}{n}\Big)^{\beta p}.$$

We could have for example that $\theta_k^n(\mathbf{X}_{k-1}^n) = \vartheta(k/n)$, almost surely, where $\vartheta(\cdot) \in \mathcal{L}_\beta(M, [0,1]) = \{g(\cdot) : \|g(t_1) - g(t_2)\|_1 \le M|t_1 - t_2|^\beta, t_1, t_2 \in [0,1]\}$ for some $0 < \beta \le 1$ and $M > 0$, a space of vector-valued Lipschitz functions.

Let $\gamma_k \equiv C_\gamma (\log n)^{(2\beta-1)/(2\beta+1)} n^{-2\beta/(2\beta+1)}$, (i.e. γ_k is constant in k) for $k = 1, \dots, n$, and

$$k_0 = k_0(n) = k - (\log n)^{2/(2\beta+1)} n^{2\beta/(2\beta+1)},$$

for $k \ge K_n = (\log n)^{2/(2\beta+1)} n^{2\beta/(2\beta+1)}$. Note that for $K_n/n \to 0$ as $n \to \infty$ for any $0 < \beta \le 1$.

For sufficiently large C_γ

$$\sum_{i=k_0}^{k} \gamma_i = C_\gamma (\log n)^{(2\beta-1)/(2\beta+1)} n^{2\beta/(2\beta+1)} (k - k_0) \ge C_\gamma \log n \ge \frac{\log n}{3\lambda_1},$$

leading to

$$\exp\Big(- p\lambda_1 \sum_{i=k_0}^{k} \gamma_i \Big) \le c n^{-p/3}.$$

In much the same way,

$$\Big(\sum_{i=k_0}^{k} \gamma_i^2 \Big)^{p/2} \le C \Big((\log n)^{\frac{2\beta-1}{2\beta+1}} n^{-\frac{2\beta}{2\beta+1}} (k-k_0)^{1/2} \Big)^p = C \Big((\log n)^{\frac{2\beta}{2\beta+1}} n^{-\frac{\beta}{2\beta+1}} \Big)^p.$$

From our assumption on the variation of the parameter, we have

$$\max_{i=k_0,\dots,k} \mathbb{E}\|\theta_{i+1}^n - \theta_{k_0}^n\|_p^p \le c \Big(\frac{k-k_0}{n}\Big)^{-p\beta} \le C \Big((\log n)^{\frac{2\beta}{2\beta+1}} n^{-\frac{\beta}{2\beta+1}} \Big)^p.$$

Combining the three bounds, we get that (6.8) and (6.9) imply

$$\sup_{\vartheta\in\mathcal{L}(L,\beta)} \max_{i\geq K_n} \mathbb{E}\|\delta_i\|_2 \leq C(\log n)^{\frac{2\beta}{2\beta+1}} n^{-\frac{\beta}{2\beta+1}}, \tag{6.31}$$

$$\sup_{\vartheta\in\mathcal{L}(L,\beta)} \max_{i\geq K_n} \mathbb{E}\|\delta_i\|_p^p \leq C\left((\log n)^{\frac{2\beta}{2\beta+1}} n^{-\frac{\beta}{2\beta+1}}\right)^p. \tag{6.32}$$

6.6 Some applications of the main result

In this section we present some examples of particular models to which our algorithm may be applied. We start with two toy examples and present thereafter some more involved examples. The toy examples illustrate the type of results that can be obtained from our main result and its extensions, how a gain function can be picked and modified, and how conditions (A1) and (A2) can be checked.

6.6.1 Tracking the intensity function of a Poisson process

Let us say that we are monitoring n independent Poisson processes on $[0,1]$ with unknown intensity function $\lambda(\cdot)$, for fixed $n \in \mathbb{N}$. This is equivalent to observing $N(t) = N(t,n)$, a Poisson process with intensity $n\lambda(t)$, $0 \leq t \leq 1$. We would like to track the intensity function $\lambda(\cdot)$ which we will assume is uniformly upper-bounded by L.

Let us say that we observe the process with frequency n, in that our observations are $X_k^n = N(k/n)$, such that for each $n \in \mathbb{N}$ we have the model

$$X_0^n = 0, \quad X_{k+1}^n | X_k^n \sim P_{\theta_k^n}(\cdot|X_k^n) = P_{\theta_k^n}(\cdot - X_k^n), \quad k = 1,\dots,n,$$

where $P_\theta(\cdot)$ represents a Poisson law with parameter $\theta \in \mathbb{R}^+$. We will work then with $p_\theta(\cdot|y)$ a conditional, shifted Poisson mass function given by

$$p_\theta(x|y) = \exp(-\theta)\frac{\theta^{x-y}}{(x-y)!},$$

for $x \in \mathbb{N}$, $x \geq y$. The moving parameter θ_k^n is given, for $k = 1,\dots,n$, by

$$\theta_k^n = \int_{\frac{k-1}{n}}^{\frac{k}{n}} n\lambda(t)\,dt.$$

Consider now the gain function G_k of the type (6.25) and its conditional expectation g_k, respectively given by

$$\begin{aligned} G_k(x,\theta|X_{k-1}^n) &= x - X_{k-1}^n - \theta, \\ g_k(\theta,\vartheta|X_{k-1}^n) &= \mathbb{E}_\vartheta[X_k^n - X_{k-1}^n - \theta|X_{k-1}^n] = -(\theta-\vartheta). \end{aligned} \tag{6.33}$$

with $\mathbb{E}_\vartheta[\,\cdot\,|X_{k-1}^n]$ the expectation with respect to $p_\vartheta(\cdot|X_{k-1}^n)$. It is also simple to see that

$$\mathbb{E}|G(X_k^n,\theta|X_{k-1}^n) - g(\theta,\vartheta|X_{k-1}^n)|^2 = \mathbb{E}\mathbb{E}_\vartheta[|X_k^n - X_{k-1}^n - \vartheta|^2|X_{k-1}^n] = \vartheta \leq L.$$

We conclude then that the gain function displayed in (6.33) satisfies both (A1) and (A2).

This gain function can now be used for the three setups outlined in Section 6.5 and attains the rates indicated there. For a constant intensity function $\lambda(\cdot) \equiv \vartheta, 0 < \vartheta \leq L$, the parameter of the model θ_k^n reduces to the constant ϑ and we simply track the rate of the process. Note that this happens since we have matched the sampling frequency $1/n$ with the sample size n. If we were to have sampled the process with frequency $2/n$, say, then $\theta_k^n = 2\vartheta$ in which case the algorithm would track 2ϑ and not ϑ. The tracking sequence would then have to be rescaled by a factor $1/2$ to obtain a tracking sequence for ϑ itself.

In the setup where we assume that the parameter is stabilizing, take $n = 1$ and call $\vartheta_k = \theta_n^1 = \int_{k-1}^k \lambda(t)\,dt$ the mean number of events per time unit. Note that

$$\left|\Delta\vartheta_k\right| = \left|\int_{k-1}^k \lambda(t)\,dt - \int_k^{k+1} \lambda(t)\,dt\right| = \left|\theta_k^1 - \theta_{k+1}^1\right|.$$

We then assume that the average number of events is stabilizing in such a way that the previous display is upper bounded, for $\beta \geq 0$ and $c_\beta > 0$, by $c_\beta k^{-\beta}$. The algorithm will then track the mean number of events per time unit.

We can also assume that the intensity function $\lambda(\cdot)$ belongs to $\mathcal{L}_\beta\big(M, [0,1]\big) = \{g(\cdot) : |g(t_1) - g(t_2)| \leq M|t_1 - t_2|^\beta, t_1, t_2 \geq 0\}$ for some $0 < \beta \leq 1$ and $M > 0$. Call $\vartheta_k^n = \lambda(k/n)$, $k, n \in \mathbb{N}$. It follows that

$$\begin{aligned}\left|\Delta\vartheta_k^n\right| &= \left|\lambda\big(k/n\big) - \lambda\big((k+1)/n\big)\right| \leq M\,n^{-\beta},\\ \left|\theta_k^n - \vartheta_k^n\right| &= \left|\int_{(k-1)/n}^{k/n} n\lambda(t)\,dt - \lambda(k/n)\right| \leq n\int_{(k-1)/n}^{k/n} \left|\lambda(t) - \lambda(k/n)\right| dt \leq M\,n^{-\beta}.\end{aligned}$$

The tracking sequence based on the gain (6.33) will then track the sequence $\vartheta_k^n = \lambda(k/n)$, $k, n \in \mathbb{N}$ (as well as θ_k^n) with the asymptotics seen in Section 6.5 (cf. Remark 6.1.)

6.6.2 Tracking the mean function of a conditionally Gaussian process

Assume that we observe, with fixed frequency $n \in \mathbb{N}$, a process X_t, $t \in [0,1]$, taking values on $\mathcal{X} \subset \mathbb{R}^d$, $d \in \mathbb{N}$. In this way, for $k = 1, \ldots, n$, the observations available to us at time k/n will be a random vector $\mathbf{X}_k^{(n)} = \big(X_0, X_{1/n}, \ldots, X_{k/n}\big)$. The increments $X_{k/n} - X_{(k-1)/n}$ will be assumed to be conditionally Gaussian in the sense that given the past of the process, each increment has a multivariate normal distribution, and so,

$$X_0^n \sim N\Big(\theta_0^n,\ \Sigma_0^n\Big), \quad X_{(k+1)/n}|\mathbf{X}_k^n \sim N\Big(\theta_k^n\big(\mathbf{X}_{k-1}^n\big),\ \Sigma_k^n\big(\mathbf{X}_{k-1}^n\big)\Big), \quad k = 1, \ldots, n.$$

The dependence on the past in the model comes from the fact that both the mean and the covariance of the process are allowed to depend on the past of the process. Here, for each $n \in \mathbb{N}$, θ_k^n is an arbitrary sequence in k depending on $\mathbf{X}_{k-1}^n$, and Σ_k^n a sequences in $k \in \mathbb{N}$ of (positive-definite) covariance matrices or order d which, as already mentioned, may also depend on $\mathbf{X}_{k-1}^n$.

In the case where the covariance structure of the process is known, we can use the gain (6.24) for which it is straightforward to check that it verifies, given $\mathbf{X}_{k-1}^n$, for $\mathbf{x}, \theta, \vartheta \in \mathbb{R}^d$, $k = 1, \ldots, n$,

$$\begin{aligned}G_k\big(\mathbf{x}, \theta|\mathbf{X}_{k-1}^n\big) &= \big(\Sigma_k^n(\mathbf{X}_{k-1}^n)\big)^{-1}\big(\mathbf{x} - \theta\big),\\ g_k\big(\theta, \vartheta(\mathbf{X}_{k-1}^n)|\mathbf{X}_{k-1}^n\big) &= -\big(\Sigma_k^n(\mathbf{X}_{k-1}^n)\big)^{-1}\big(\theta - \vartheta(\mathbf{X}_{k-1}^n)\big).\end{aligned} \tag{6.34}$$

If this gain function is used, then we assume that, almost surely, for $k = 1, \ldots, n$, the eigenvalues of the covariance

matrices $\Sigma_k^n(\mathbf{X}_{k-1}^n)$ are $0 < \Lambda_{(1),k}^n(\mathbf{X}_{k-1}^n) \leq \cdots < \Lambda_{(d),k}^n(\mathbf{X}_{k-1}^n) < \infty$, so that for constants λ_1^n, λ_2^n,

$$0 < \lambda_1^n \leq \Lambda_{(1),k}^n(\mathbf{X}_{k-1}^n) \leq \Lambda_{(d),k}^n(\mathbf{X}_{k-1}^n) \leq \lambda_2^n < \infty,$$

almost surely. We then have for all $\theta, \vartheta \in \mathbb{R}^d$,

$$\begin{aligned}
&\mathbb{E}\|G(X_k^n, \theta|\mathbf{X}_{k-1}^n) - g(\theta, \vartheta(\mathbf{X}_{k-1}^n)|\mathbf{X}_{k-1}^n)\|_2^2 = \\
&\quad = \mathbb{E}\big((\Sigma_k^n(\mathbf{X}_{k-1}^n))^{-1}\big(X_k^n - \vartheta(\mathbf{X}_{k-1}^n)\big)\big)^T(\Sigma_k^n(\mathbf{X}_{k-1}^n))^{-1}\big(X_k^n - \vartheta(\mathbf{X}_{k-1}^n)\big) \\
&\quad \leq (\lambda_1^n)^{-2}\,\mathbb{E}\mathbb{E}_\vartheta\big[\|X_k^n - \vartheta(\mathbf{X}_{k-1}^n)\|_2^2\big|\mathbf{X}_k^n\big] = (\lambda_1^n)^{-2}\mathbb{E}\,\mathrm{tr}\,\big(\Sigma_k^n(\mathbf{X}_{k-1}^n)\big) \leq d\,\lambda_2^n\,(\lambda_1^n)^{-2}.
\end{aligned}$$

Assumptions (A1) and (A2) are then met for the gain in (6.34).

Let us now assume that the covariance matrix of the process is unknown, difficult to invert or that assumption on the eigenvalues of the covariance matrix does not hold. In this case we can use the gain (6.25) which gives us, for $\mathbf{x}, \theta, \vartheta \in \mathbb{R}^d$, $k = 1, \ldots, n$,

$$\begin{aligned}
G(\mathbf{x}, \theta|\mathbf{X}_{k-1}^n) &= \mathbf{x} - \theta, \\
g(\theta, \vartheta(\mathbf{X}_{k-1}^n)|\mathbf{X}_{k-1}^n) &= -\big(\theta - \vartheta(\mathbf{X}_{k-1}^n)\big).
\end{aligned} \tag{6.35}$$

If we now assume that, almost surely, for $k = 1, \ldots, n$, the largest eigenvalue of the covariance matrices $\Sigma_k^n(\mathbf{X}_{k-1}^n)$ is upper bounded by some constant $\lambda_2^n < \infty$, then, for all $\theta, \vartheta \in \mathbb{R}^d$,

$$\begin{aligned}
&\mathbb{E}_\vartheta\|G(X_k^n, \theta|\mathbf{X}_{k-1}^n) - g(\theta, \vartheta(\mathbf{X}_{k-1}^n)|\mathbf{X}_{k-1}^n)\|_2^2 = \\
&\quad = \mathbb{E}\mathbb{E}_\vartheta\big[\|X_k^n - \vartheta(\mathbf{X}_{k-1}^n)\|_2^2\big|\mathbf{X}_{k-1}^n\big] = \mathbb{E}\,\mathrm{tr}\,\big(\Sigma_k^n(\mathbf{X}_{k-1}^n)\big) \leq d\,\lambda_2^n,
\end{aligned}$$

and so assumptions (A1) and (A2) are met for the gain in (6.35).

The results of Section 6.5 can be applied to the algorithm based on the gain functions presented above. If, for each $n \in \mathbb{N}$, the mean of the process is constant, $\theta_k^n(\mathbf{X}_{k-1}^n) \equiv \vartheta^n$ then the algorithm will track the (fixed) mean of the process. Alternatively, we may assume that the parameter is not constant but is stabilizing. We take then $n = 1$, and assume that the changes in the mean vector of the process are such that, for $k \in \mathbb{N}$,

$$\mathbb{E}\|\Delta\theta_k^n\|_2^2 = \mathbb{E}\|\theta_k^n(\mathbf{X}_{k-1}^n) - \theta_{k+1}^n(\mathbf{X}_k^n)\|_2^2 \leq c_\beta k^{-\beta},$$

for some $\beta \geq 0$, and a constant $c_\beta > 0$. The other possibility is to assume that for $n \in \mathbb{N}$, the mean of the process is obtained from a function $\theta(\cdot, \mathbf{X}_{k-1}^n)$ which is, on average, Lipschitz in the sense that it belongs to $\mathcal{L}_\beta\big(M, [0,1], \mathbf{X}_{k-1}^n\big) = \{g : \mathbb{E}\|g(t_1, \mathbf{X}_{k-1}^n) - g(t_2, \mathbf{X}_{k-1}^n)\|_1 \leq M|t_1 - t_2|^\beta, t_1, t_2 \geq 0\}$ for some $0 < \beta \leq 1$ and $M > 0$. Call $\vartheta_k^n = \theta(k/n, \mathbf{X}_{k-1}^n)$, $k, n \in \mathbb{N}$. It follows that

$$\mathbb{E}\big\|\Delta\vartheta_k^n\big\|_1 = \mathbb{E}\big\|\theta\big(k/n, \mathbf{X}_{k-1}^n\big) - \theta\big((k+1)/n, \mathbf{X}_k^n\big)\big\|_1 \leq Mn^{-\beta}.$$

In this case the algorithm tracks the mean function $\theta(k/n, \mathbf{X}_{k-1}^n)$ at times k/n, with $k \in \mathbb{N}$.

6.6.3 Tracking an ARCH(1) parameter

Consider the following ARCH(1) model with drifting parameter

$$X_k = \left(1 + \theta_k X_{k-1}^2\right)^{1/2} \xi_k, \quad X_0 = 0 \text{ (a.s.)}, \tag{6.36}$$

where ξ_k, $k \in \mathbb{N}$, form a martingale difference sequence with variance $\sigma^2 = 1$. The drifting parameter θ_k belongs to some interval $[0, \rho]$ for some ρ such that $\rho^2 \mathbb{E}\xi_k^4 \leq 1$ for all $k \in \mathbb{N}$.

Consider the gain function

$$G(X_k, \theta | X_{k-1}) = \frac{\min(X_{k-1}^2, c\sigma^2)}{X_{k-1}^2}(X_k^2 - 1 - \theta X_{k-1}^2) \tag{6.37}$$

such that, since $\mathbb{E}_\vartheta X_k = 0$ and $\mathbb{E}_\vartheta[X_k^2 | X_{k-1}] = \sigma^2(1 + \vartheta X_{k-1}^2) = 1 + \vartheta X_{k-1}^2$,

$$g(\theta, \vartheta | X_{k-1}) = \mathbb{E}_\vartheta\Big[\frac{\min(X_{k-1}^2, c\sigma^2)}{X_{k-1}^2}(X_k^2 - 1 - \theta X_{k-1}^2)\Big| X_{k-1}\Big] = -\min(X_{k-1}^2, c\sigma^2)(\theta - \vartheta),$$

for some constant $c > 0$. We then have that $\Lambda_{(1)} \leq c\sigma^2$, almost surely. Note that

$$\mathbb{E}\Big[\min(X_{k-1}^2, c\sigma^2)\Big| X_{k-2}\Big] = \mathbb{E}\Big[\min\left((1 + \theta_{k-1} X_{k-2}^2)\xi_{k-1}^2, c\sigma^2\right)\Big| X_{k-2}\Big] \geq \mathbb{E}\Big[\min\left(\xi_{k-1}^2, c\sigma^2\right)\Big].$$

By using the fact that $\min(a, b) = (a + b)/2 - |a - b|$ and the Hölder inequality, it is straightforward to check that

$$2\,\mathbb{E}\Big[\min\left(\xi_{k-1}^2, c\sigma^2\right)\Big] = (c + 1)\sigma^2 - \mathbb{E}|\xi_{k-1}^2 - c\sigma^2| \geq (c + 1)\sigma^2 - \left(\mathbb{E}\Big[\left(\xi_{k-1}^2 - c\sigma^2\right)^2\Big]\right)^{1/2} \geq \sigma^2,$$

as long as for every $k \in \mathbb{N}$, $2c\,\sigma^2 \geq \mathbb{E}\xi_k^4$. We conclude that (A1) holds for the gain (6.37).

To check (A2) note first that

$$\mathbb{E}[X_k^2 | X_{k-1}^2] = \sigma^2(1 + \theta_k X_{k-1}^2)$$

and then

$$\mathbb{E}X_k^2 \leq \sigma^2(1 + \rho \mathbb{E}X_{k-1}^2).$$

Since $\rho^2 \mathbb{E}\xi_k^4 \leq 1$, it follows that $\rho\sigma^2 \leq 1$ by Jensen's inequality. Using this recursion we get that

$$\mathbb{E}X_k^2 \leq \sigma^2 + \sigma^2 \rho \mathbb{E}X_{k-1}^2 \leq \sigma^2 + \sigma^4 \rho + \sigma^4 \rho^2 \mathbb{E}X_{k-2}^2 \leq \sigma^2 \sum_{i=1}^{k} (\rho\sigma^2)^{i-1} \leq \frac{\sigma^2}{1 - \sigma^2 \rho}.$$

In the same way,

$$\mathbb{E}[X_k^4 | X_{k-1}] = (1 + 2\theta_k X_{k-1}^2 + \theta_k^2 X_{k-1}^4)\mathbb{E}\xi_k^4,$$

and then, since $\rho^2 \mathbb{E}\xi_k^4 \leq 1$,

$$\mathbb{E}X_k^4 \leq (1 + 2\frac{\sigma^2 \rho}{1 - \sigma^2 \rho} + \rho^2 \mathbb{E}X_{k-1}^4)\mathbb{E}\xi_k^4 \leq \mathbb{E}\xi_k^4 (1 + 2\frac{\sigma^2 \rho}{1 - \sigma^2 \rho}) \sum_{i=1}^{k} (\rho^2 \mathbb{E}X_{k-1}^4)^{i-1} < \infty.$$

Using the same argument as for (6.26) we see that (A2) holds since by the Hölder inequality

$$\mathbb{E}G^2(X_k, \theta | X_{k-1}) \leq 3(\mathbb{E}X_k^2 + \rho^2 \mathbb{E}X_{k-1}^2 + 1),$$

which is bounded, uniformly over $k \in \mathbb{N}$.

6.6.4 Tracking an AR(d) parameter

We consider now an autoregressive model with d time-varying auto-regressive parameters:

$$X_k = \sum_{i=1}^{d} \theta_{k,i} X_{k-i} + \xi_k, \quad k \in \mathbb{N},\ k \geq d, \tag{6.38}$$

where $X_0, X_1, \ldots, X_{d-1}$ have p bounded moments (cf. the end of this section). We would like to track the vector $\theta_k = (\theta_{k,1}, \theta_{k,2}, \ldots, \theta_{k,d})$, which may be random but must be measurable with respect to the σ-algebra generated by $\mathbf{X}_{k-2d-1}$. In this section we will use the notation $\mathbf{X}_{k,d} = \big(X_k, X_{k-1}, \ldots, X_{k-(d-1)}\big)$ for the vector of the d observations leading up to X_k.

In analogy with the non-drifting AR(d) model, we can associate with the model its (drifting) autoregressive polynomial $z \mapsto 1 - \sum_{i=1}^{d} \theta_{k,i} z^i$; write then

$$t(z, \theta) = 1 - \sum_{i=1}^{d} \theta_i z^i, \quad z \in \mathbb{C}. \tag{6.39}$$

It is well knows that an AR(p) model with autoregressive parameters θ has a stationary distribution if, and only if, the (complex) zeros of the polynomial $t(z, \theta)$ are outside the unit circle. This motivates the definition of the parameter sets $\Theta(\rho)$, (cf. [71]) which we define as the closure of

$$\big\{\theta \in \mathbb{R}^d : \text{for all } |z| < \rho^{-1}, t(z, \theta) \neq 0\big\}, \tag{6.40}$$

for any $0 < \rho < 1$. One can show that if $\mathcal{B}(r)$ is a uniform ball in $\mathbb{R}^d$ with radius $r > 0$ around the origin, then the following embeddings hold:

$$\mathcal{B}\big((\rho^{-2} + \cdots + \rho^{-2d})^{-1/2}\big) \subseteq \Theta(\rho) \subseteq \mathcal{B}\big((1+\rho)^d - 1\big),$$

which gives us some feeling as to the size of the parameter set (cf. [71]). This implies in particular that for all $\rho \in (0,1)$, the set $\Theta(\rho)$ is non-empty and bounded.

The AR(d) model (6.38) can also be described by the following inhomogeneous difference equation

$$\mathbf{X}_{k,d} = C(\theta_k)\mathbf{X}_{k-1,d} + I\mathbf{e}_1 \xi_k, \tag{6.41}$$

where $\mathbf{e}_1 = (1, 0, \ldots, 0) \in \mathbb{R}^d$ and, for any $\theta \in \mathbb{R}^d$, $C(\theta)$ is the square matrix of order d

$$C(\theta) = \begin{bmatrix} \theta_1 & \theta_2 & \cdots & \theta_{d-1} & \theta_d \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix}. \tag{6.42}$$

This matrix is usually called the *companion matrix* to the autoregressive polynomial $t(z, \theta)$; it is also sometimes called the *state transition matrix*. One can show that the eigenvalues of $C(\theta)$ are exactly the reciprocals of the zeros of $t(z, \theta)$. This means that all the eigenvalues of $C(\theta)$ for $\theta \in \Theta(\rho)$ are at most $\rho < 1$. This in turn implies that for any sequence of vectors $\theta_d, \theta_{d+1}, \cdots \in \Theta(\rho)$, the pair of sequences

$$\Big(\big(C(\theta_d), C(\theta_{d+1}), \cdots\big), \big(I_d, I_d, \cdots\big)\Big)$$

forms a so-called *exponentially stable pair*. Among other things, this gives us that so long as the p-th moments of both the initial $\mathbf{X}_{d-1,d}$ and the noise terms ξ_k are bounded, then the p-th moments of all X_k, $k \geq d$ will be bounded as well (cf. Proposition 10 of [71]).

A particular gain function that can be used to track the parameters of an autoregressive model can be found in [71]. The gain function considered there is an appropriately rescaled version of the gain (6.19), namely,

$$G(X_k, \theta | \mathbf{X}_{k-1,d}) = (X_k - \theta^T \mathbf{X}_{k-1,d}) \frac{\mathbf{X}_{k-1,d}}{1 + \mu \mathbf{X}_{k-1,d}^T \mathbf{X}_{k-1,d}},$$

for an appropriately chosen $\mu > 0$. It is straightforward to check that the conditions in (Ã1) on the corresponding conditional gain g hold in this case; the lower bound in (6.6) is established in Lemma 17 of [71] and the upper bounds are straightforward to check; assumption (A2) can be reduced to moment conditions on the observations of the autoregressive process which are verified if the signal θ lives in $\Theta(\rho)$ as mentioned above. In a sense, conditions (A1) and (A2) capture the essential properties that a gain function must have such that resulting tracking algorithm behaves properly and, in fact, these conditions will hold even if the noise terms are not Gaussian; we discuss this issue again at the end of this section. In the following we propose an alternative gain function. We will first treat the one-dimensional case where we can obtain a stronger result.

Consider $d = 1$ and assume that the sequence $\theta_k(\mathbf{X}_{k-1}) \in \Theta(\rho)$ is almost surely bounded, in absolute value, by $\rho < 1$. Assume also that $\mathbb{E}X_0^2$ and $\mathbb{E}X_0^4$ are bounded and that for all $k \in \mathbb{N}$, $\mathbb{E}\xi_k = \mathbb{E}\xi_k^3 = 0$, $\mathbb{E}\xi_k^2 = \sigma^2 > 0$ and $\mathbb{E}\xi_k = c\,\sigma^4$, for some constant $0 \leq c < 5$. (We can take $c = 3$ if the noise is Gaussian, for example.) Let us say we would like to use the following gradient type gain function

$$\begin{aligned} G_k\big(X_k, \theta | X_{k-1}\big) &= X_{k-1}^2 (X_k / X_{k-1} - \theta), \\ g_k\big(\theta, \vartheta | X_{k-1}\big) &= -X_{k-1}^2 (\theta - \vartheta), \end{aligned} \tag{6.43}$$

almost surely. The random eigenvalues in (A1) reduce, in this case, to $\Lambda_{(1)}(X_{k-1}) = X_{k-1}^2$. Note that if for all $k \in \mathbb{N}$

X_{k-2} are integrable, we have

$$\mathbb{E}\big[X_{k-1}^2\big|X_{k-2}\big] = \mathbb{E}\big[(X_{k-2}\theta_{k-2} + \xi_{k-1})^2\big|X_{k-2}\big] = X_{k-2}^2\theta_{k-2}^2 + \mathbb{E}\xi_{k-1}^2 \geq \sigma^2, \tag{6.44}$$

but still X_{k-1}^2 would not be almost surely upper-bounded by a constant. To remedy this we will truncate X_{k-1}^2 and consider

$$\begin{aligned} G_k\big(X_k, \theta|X_{k-1}\big) &= \min\Big(X_{k-1}^2, \frac{9-c}{4}\sigma^2\Big)(X_k/X_{k-1} - \theta), \\ g_k\big(\theta, \vartheta|X_{k-1}\big) &= -\min\Big(X_{k-1}^2, \frac{9-c}{4}\sigma^2\Big)(\theta - \vartheta). \end{aligned} \tag{6.45}$$

(Note that this is a rescaled gain function of the same type as $\bar{G}$ at the end of Section 6.4.) We now have an almost sure upper-bound for $\Lambda_{(1)}(X_{k-1}) = \min\big(X_{k-1}^2, (9-c)\sigma^2/4\big)$; we truncate X_{k-1}^2 at this specific value since one can prove (cf. Lemma 6.5) that

$$\mathbb{E}\Big[\min\Big(X_{k-1}^2, \frac{9-c}{4}\sigma^2\Big)\Big|X_{k-2}\Big] \geq \frac{5-c}{4}\sigma^2 > 0,$$

so that (A1) holds. Assumption (A2) also holds since

$$\begin{aligned} \mathbb{E}\big|G_k\big(X_k, \theta|X_{k-1}\big) - g_k\big(\theta, \vartheta|X_{k-1}\big)\big|^2 &= \mathbb{E}\frac{\min\big(X_{k-1}^2, \frac{5-c}{4}\sigma^2\big)^2}{X_{k-1}^2}\mathbb{E}_\vartheta\big[|X_k - \vartheta X_{k-1}|^2\big|X_{k-1}\big] \\ &\leq \Big(\frac{5-c}{4}\Big)^2\sigma^4\mathbb{E}\xi_k^2 = \Big(\frac{5-c}{4}\Big)^2\sigma^6. \end{aligned}$$

The previous truncation argument is still valid if we truncate X_{k-1}^2 at a higher value. In that case, we also still have that (A1) and (A2) hold, with a larger constant λ_2 in (A1) and larger C in (A2). This means that in order to use the previous gain function we don't need to know the exact value of σ^2 but only an upper bound for it. Also, in practice, for a truncation at a high enough value, the effect of the truncation will be innocuous and trajectories of (6.43) and (6.45) will coincide, with high probability; the truncation is simply an artifact to enforce the fulfillment of (A1) and should be of little practical importance. Up to the requirement that the distribution of the noise be symmetrical about 0, the previous result generalizes that of [5] where the noise terms are assumed to be almost surely bounded.

Now we turn our attention to the general AR(d) model. As we will see in what follows, assumptions (A1) and (A2) can be easily checked. In the d dimensional case we assume that the noise terms ξ_k in (6.38) form a Gaussian white noise sequence with mean zero and variance $\sigma^2 > 0$. Assume first that the autoregressive parameters do not depend on k, i.e. $\theta_k \equiv \theta = (\theta_1, \ldots, \theta_d) \in \Theta(\rho) \subset \mathbb{R}^d$. Given the vector of past observations $\mathbf{X}_{k-d,d} = \big(X_{k-d}, X_{k-d-1}, \ldots, X_{k-2d+1}\big)$, we can see $\mathbf{X}_{k,d}$ as a system of d equations in $X_k, X_{k-1}, \ldots, X_{k-(d-1)}$, depending on $X_{k-d}, X_{k-d-1}, \ldots, X_{k-2d+1}$ and θ, which, for $\boldsymbol{\xi}_{k,d} = \big(\xi_k, \xi_{k-1}, \ldots, \xi_{k-(d-1)}\big)$, can be written as

$$A(\theta)\mathbf{X}_{k,d} = B(\theta)\mathbf{X}_{k-d,d} + \boldsymbol{\xi}_{k,d}; \tag{6.46}$$

the matrices $A(\theta)$ and $B(\theta)$ are Toeplitz matrices created from the vectors $\mathbf{a}(\theta) = (0, \ldots, 0, 1, -\theta_1, \ldots, -\theta_{d-1})$ and $\mathbf{b}(\theta) = (\theta_1, \ldots, \theta_{d-1}, \theta_d, 0, \ldots, 0)$ respectively. (For $\mathbf{m} = (m_{-(d-1)}, m_{-(d-2)}, \ldots, m_0, \ldots, m_{d-2}, m_{d-1})$, the

Toeplitz matrix of order d associated with that vector is the square matrix M of order d with entries $m_{i,j} = m_{i-j}$, such that the entries of the matrix are constant over descending diagonals.) The matrix $A(\theta)$ is upper triangular with a diagonal consisting of ones whence invertible. We conclude, then, that given the full past of the process, $\mathbf{X}_{k-d}$,

$$\mathbf{X}_{k,d}|\mathbf{X}_{k-d} \sim N\big(A^{-1}(\theta)B(\theta)\mathbf{X}_{k-d,d},\ \sigma^2 A^{-1}(\theta)A^{-T}(\theta)\big). \tag{6.47}$$

Alternatively, we could have derived the (6.46) by applying the recursion in (6.41) d times.

In this case we will consider a gain of the type (6.24) such that

$$G_{dk}\big(\mathbf{x}, \theta|\mathbf{X}_{d(k-1),d}\big) = \nabla_\theta \log p_\theta\big(\mathbf{x}|\mathbf{X}_{d(k-1)}\big), \tag{6.48}$$

where $p_\theta(\cdot|\mathbf{X}_{d(k-1)})$ is the conditional density of (6.47). At the end of this section we explicitly compute (6.48); see (6.50).

For us, each data point will be a vector $\mathbf{X}_{dk,d}$, $k \in \mathbb{N}$ such that the tracking sequence is updated with batches of d observations from the autoregressive process. (Below, to ease the notation, we will mostly write $\mathbf{x}$ and $\mathbf{y}$ instead of $\mathbf{X}^n_{dk,d}$ and $\mathbf{X}^n_{d(k-1),d}$, respectively.) This is necessary to make sure that the representation (6.47) is valid even if the parameter θ is allowed to change among different batches of observations; otherwise the system (6.46) would be under-determined. We must now establish that this gain function verifies (A1).

As explained in Section 6.4, the expectation g_{dk} can be seen as minus the gradient of the Kullback-Leibler divergence between the transition kernel with two different parameters. This observation is particularly useful if we are able to write this Kullback-Leibler divergence as an appropriate quadratic form. One can show that the Kullback-Leibler divergence between two d-dimensional multivariate normal distributions $P_0 = N(\boldsymbol{\mu}_0, \Sigma_0)$ and $P_1 = N(\boldsymbol{\mu}_1, \Sigma_1)$ is given by

$$K(P_0, P_1) = \frac{1}{2}\left(\log\frac{\det\Sigma_1}{\det\Sigma_0} + \operatorname{tr}(\Sigma_1^{-1}\Sigma_0) - d + (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^T\Sigma_1^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)\right). \tag{6.49}$$

Write, for $\mathbf{y} \in \mathbb{R}^d$, $\mu(\theta, \mathbf{y}) = A^{-1}(\theta)B(\theta)\mathbf{y}$ and $\Sigma(\theta) = \sigma^2 A^{-1}(\theta)A^{-T}(\theta)$. Let also $S = S_d$ be the Toeplitz matrix associated with the vector $\mathbf{s} = (0, \dots, 0, 1, 0, \dots, 0) \in \mathbb{R}^{2d-1}$ where the 1 occupies the $(d+1)$-th position; these are sometimes called *shift matrices*. For $i = 2, \dots, d-1$, the powers S^i are the Toeplitz matrices associated with the vectors $(0, \dots, 0, 1, 0, \dots, 0) \in \mathbb{R}^{2d-1}$ where the 1 occupies the $(d+i)$-th position; S^d is $O = O_d$, the null matrix of order d, and S^0 should be read as $I = I_d$, the identity matrix of order d. It follows from the definition of the matrix $A(\cdot)$ that for $\theta, \vartheta \in \Theta_d$,

$$A(\theta) - A(\vartheta) = S(\vartheta_1 - \theta_1) + S^2(\vartheta_2 - \theta_2) + \cdots + S^d(\vartheta_d - \theta_d),$$

from where we conclude

$$A(\theta)A^{-1}(\vartheta) = I + SA^{-1}(\vartheta)(\vartheta_1 - \theta_1) + S^2A^{-1}(\vartheta)(\vartheta_2 - \theta_2) + \cdots + S^dA^{-1}(\vartheta)(\vartheta_d - \theta_d).$$

We will compute now $K\big(N(\mu(\vartheta, \mathbf{y}), \Sigma(\vartheta)),\ N(\mu(\theta, \mathbf{y}), \Sigma(\theta))\big)$. For all θ, the matrices $A(\theta)$ have all eigenvalues equal to one (so then also their inverses) whence $\det\Sigma(\theta) \equiv \sigma^{2d}$; we conclude that the logarithm in (6.49)

is null. Also, using basic properties of the trace and the representation for $A(\theta)A^{-1}(\vartheta)$ derived above,

$$\begin{aligned}
&\operatorname{tr}\left(\Sigma^{-1}(\theta)\Sigma(\vartheta)\right) - d = \operatorname{tr}\left(\left(A^{-1}(\theta)A^{-T}(\theta)\right)^{-1}\left(A^{-1}(\vartheta)A^{-T}(\vartheta)\right)\right) - d \\
&= \operatorname{tr}\left(A^{T}(\theta)A(\theta)A^{-1}(\vartheta)A^{-T}(\vartheta)\right) - d = \operatorname{tr}\left(\left(A(\theta)A^{-1}(\vartheta)\right)^{T}A(\theta)A^{-1}(\vartheta)\right) - d \\
&= 2\sum_{i=1}^{d}\operatorname{tr}\left(S^{i}A^{-1}(\vartheta)\right)(\vartheta_i - \theta_i) + \sum_{i=1}^{d}\sum_{j=1}^{d}\operatorname{tr}\left(A^{-T}(\vartheta)(S^{i})^{T}S^{j}A^{-1}(\vartheta)\right)(\vartheta_i - \theta_i)(\vartheta_j - \theta_j).
\end{aligned}$$

The inverse of an upper-triangular matrix is upper-triangular and so, for all $i = 1, \ldots, d$ and all ϑ, the matrices $S^{i}A^{-1}(\vartheta)$ have null trace. Denote now for any n by m matrix M, $\operatorname{vect}(M)$ as the column vector containing the nm entries of M in any (fixed) order. Write then for $i = 1, \ldots, d$, $v_i(\vartheta) = \operatorname{vect}\left(S^{i}A^{-1}(\vartheta)\right)$; $v_d(\vartheta)$ is always a null vector. Note that the i, j-the element of the double sum in the previous display is given by $v_i^{T}(\vartheta)v_j(\vartheta)$, for $i, j = 1, \ldots, d$. We conclude that the previous display can be written as

$$(\vartheta - \theta)^{T}\left[v_1(\vartheta)v_2(\vartheta)\ldots v_d(\vartheta)\right]^{T}\left[v_1(\vartheta)v_2(\vartheta)\ldots v_d(\vartheta)\right](\vartheta - \theta),$$

where the matrices are written by columns.

The quadratic form in the Kullback-Leibler divergence (6.49) can be written, for any $\theta, \vartheta \in \Theta_d$ and $\mathbf{y} \in \mathbb{R}^{d}$, as

$$\begin{aligned}
&\left(\mu(\theta, \mathbf{y}) - \mu(\vartheta, \mathbf{y})\right)^{T}\Sigma^{-1}(\theta)\left(\mu(\theta, \mathbf{y}) - \mu(\vartheta, \mathbf{y})\right) = \\
&\quad = \sigma^{-2}\mathbf{y}^{T}\left(B(\theta) - A(\theta)A^{-1}(\vartheta)B(\vartheta)\right)^{T}\left(B(\theta) - A(\theta)A^{-1}(\vartheta)B(\vartheta)\right)\mathbf{y}.
\end{aligned}$$

Note that the matrix $B(\cdot)$ is linear in its argument and so se can write the expansion

$$B(\theta) - B(\vartheta) = B(\theta - \vartheta) = \left(S^{d-1}\right)^{T}(\vartheta_1 - \theta_1) + \cdots + S^{T}(\vartheta_{d-1} - \theta_{d-1}) + I(\vartheta_d - \theta_d),$$

from where, using the representation for $A(\theta)A^{-1}(\vartheta)$ derived above, we have

$$B(\theta) - A(\theta)A^{-1}(\vartheta)B(\vartheta) = C_1(\vartheta)(\vartheta_1 - \theta_1) + C_2(\vartheta)(\vartheta_2 - \theta_2) + \cdots + C_d(\vartheta)(\vartheta_d - \theta_d),$$

where $C_i(\vartheta) = \left(S^{d-i}\right)^{T} - S^{i}A^{-1}(\vartheta)B(\vartheta)$ for $i = 1, \ldots, d$. We can then write, for $\theta, \vartheta \in \Theta(\rho)$ and $\mathbf{y} \in \mathbb{R}^{d}$,

$$\left(B(\theta) - A(\theta)A^{-1}(\vartheta)B(\vartheta)\right)\mathbf{y} = \left[C_1(\vartheta)\mathbf{y}\cdots C_d(\vartheta)\mathbf{y}\right]\left(\theta - \vartheta\right).$$

We conclude that the following representation holds

$$\begin{aligned}
g_{dk}\left(\theta, \vartheta|\mathbf{X}_{d(k-1),d}\right) = -\sigma^{-2}\Big(&\sigma^{2}\left[v_1(\vartheta)v_2(\vartheta)\ldots v_d(\vartheta)\right]^{T}\left[v_1(\vartheta)v_2(\vartheta)\ldots v_d(\vartheta)\right] + \\
&+ \left[C_1(\vartheta)\mathbf{X}_{d(k-1),d}\cdots C_d(\vartheta)\mathbf{X}_{d(k-1),d}\right]^{T}\left[C_1(\vartheta)\mathbf{X}_{d(k-1),d}\cdots C_d(\vartheta)\mathbf{X}_{d(k-1),d}\right]\Big)\left(\theta - \vartheta\right).
\end{aligned}$$

Note that the matrix that precedes the vector $(\theta - \vartheta)$ does not depend on θ and is clearly positive semi-definite. We bound now the eigenvalues of this sum of matrices.

The first matrix in the sum above is positive semi-definite but has at least one null eigenvalue. It is also clear that the entries of this matrix are polynomials in $\vartheta_1, \ldots, \vartheta_{d-1}$, such that, since $\Theta(\rho)$ is a bounded set, we have that

the largest eigenvalue of this matrix is upper bounded, uniformly over Θ_d, by some constant, say, K_1, depending only on d and the diameter of $\Theta(\rho)$; we remind that this diameter is at most $(1+\rho)^d - 1 < 2^d - 1$.

We have that

$$\mathrm{tr}\Big(\Big[C_1(\vartheta)\mathbf{y} \quad \cdots \quad C_d(\vartheta)\mathbf{y}\Big]^T \Big[C_1(\vartheta)\mathbf{y} \quad \cdots \quad C_d(\vartheta)\mathbf{y}\Big]\Big) = \mathbf{y}^T C_1^T(\vartheta)C_1(\vartheta)\mathbf{y} + \cdots + \mathbf{y}^T C_d^T(\vartheta)C_d(\vartheta)\mathbf{y}.$$

For each $i = 1, \ldots, d-1$, the entries of the matrices $C_i^T(\vartheta)C_i(\vartheta)$, are polynomials in $\vartheta_1, \ldots, \vartheta_d$; the previous display is then also upper-bounded uniformly over $\Theta(\rho)$ by, say, $K_2\mathbf{y}^T\mathbf{y}$, where K_2 is a constant which like K_1 above, depends only on d and the diameter of $\Theta(\rho)$.

To derive a lower bound on the smallest eigenvalue of the matrix in the representation for g_{dk}, note that this matrix can be rewritten in the form

$$\left[\begin{array}{ccc|c} v_{1,1}(\vartheta) & \cdots & v_{1,d-1}(\vartheta) & 0 \\ \vdots & \ddots & \vdots & \vdots \\ v_{d-1,1}(\vartheta) & \cdots & v_{d-1,d-1}(\vartheta) & 0 \\ \hline c_{d,1}(\vartheta) & \cdots & c_{d,d-1}(\vartheta) & \mathbf{y}^T\mathbf{y} \end{array}\right] + \left[\begin{array}{ccc|c} c_{1,1}(\vartheta) & \cdots & c_{1,d-1}(\vartheta) & c_{1,d}(\vartheta) \\ \vdots & \ddots & \vdots & \vdots \\ c_{d-1,1}(\vartheta) & \cdots & c_{d-1,d-1}(\vartheta) & c_{d-1,d}(\vartheta) \\ \hline 0 & \cdots & 0 & 0 \end{array}\right]$$

for $v_{i,j}(\vartheta) = \sigma^2 v_i^T(\vartheta)v_j(\vartheta)$ and $c_{i,j}(\vartheta) = \mathbf{y}^T C_{d-i+1}^T(\vartheta)C_{d-j+1}(\vartheta)\mathbf{y}$, where we swapped the last rows of the matrices. (Note that $C_d(\vartheta) \equiv I$ so that $c_{d,d}(\vartheta) = \mathbf{y}^T\mathbf{y}$ and also $v_d(\vartheta) = \mathbf{0}^T$.)

Note that the top left matrices in the block matrices above are Gram matrices and therefore positive semi-definite; the full block matrices are triangular by blocks. The matrix $[v_{i,j}(\vartheta)]_{i,j=1,\ldots,d-1}$ is the Gram matrix associated with the vectors $v_1(\vartheta), \ldots, v_{d-1}(\vartheta)$. It is simple to see that these vectors are linearly independent (this follows from the fact that $A^{-1}(\vartheta)$ is a triangular matrix with 1's in its main diagonal) whence the associated Gramian is actually positive-definite for each ϑ. Note also that the determinant of this Gramian is a polynomial in the entries of the matrix which in turn are a polynomial in $\vartheta_1, \ldots, \vartheta_d$. Since $\vartheta \in \Theta(\rho)$, which is a compact set, we conclude that the infimum of the determinant of this matrix over $\vartheta \in \Theta(\rho)$ is lower bounded by some positive constant say, K_3. Using the same reasoning we can see that its determinant is upper bounded by some constant K_4. A lower bound on the smallest eigenvalue can then be obtained by noting that for any positive-definite matrix M of order d,

$$\lambda_{(1)}(M) \geq \frac{\det(M)}{\lambda_{(d)}^{d-1}(M)} \geq \frac{K_3}{K_4^{d-1}} \geq \nu > 0,$$

for some constant ν depending only on d and say, the diameter of the parameter set $\Theta(\rho)$.

We conclude that the smallest eigenvalue of the block matrix on the left is at least $\min(\nu, \mathbf{y}^T\mathbf{y})$. The block matrix on the right is clearly positive semi-definite. We conclude that the smallest eigenvalue of the matrix in the representation for g_{dk} above is lower bounded by $\min(\nu, \mathbf{y}^T\mathbf{y})$ by using Weyl's Monotonicity Theorem; cf. [3].

Condition (A2) is simpler to check. Let $D(\mathbf{X}_{dk,d}, \mathbf{X}_{d(k-1),d}) = D(\mathbf{X}_{dk,2d})$ be the Toeplitz matrix associated with the vector $\mathbf{d}(\mathbf{X}_{dk,2d}) = (X_{d(k-2)+1}, \ldots, X_{dk-2}, X_{dk-1})$ which is simply the vector $\mathbf{X}_{dk-1,2d-1}$ written backwards. Based on this, the gain (6.48) can be written, up to a constant depending only on σ^2 and with $\mathbf{x} = \mathbf{X}_{dk,d}$,

$\mathbf{y} = \mathbf{X}_{d(k-1),d}$, in the following form

$$\begin{aligned} G_{dk}\big(\mathbf{x}, \theta|\mathbf{y}\big) &= -\nabla_\theta \big(A(\theta)\mathbf{x} - B(\theta)\mathbf{y}\big)^T \big(A(\theta)\mathbf{x} - B(\theta)\mathbf{y}\big) \\ &= -2\big(A(\theta)\mathbf{x} - B(\theta)\mathbf{y}\big)^T \frac{\boldsymbol{\partial}\big(A(\theta)\mathbf{x} - B(\theta)\mathbf{y}\big)}{\boldsymbol{\partial}\theta} \\ &= 2\big(A(\theta)\mathbf{x} - B(\theta)\mathbf{y}\big)^T D(\mathbf{x}, \mathbf{y})J, \end{aligned} \tag{6.50}$$

where $\boldsymbol{\partial}/\boldsymbol{\partial}\theta$ represents the Jacobian operator. To verify (A2) it suffices to check that the expectation of the norm of G_{dk} is bounded. We omit the details but it is clear from the expression derived above that the norm of the gain function squared is a polynomial of degree 4 in the elements of $\mathbf{X}_{dk-1,2d-1}$. We have already mentioned that so long as the initial values for the autoregressive process and the noise terms have uniformly bounded p-th moments, then this transfers to the each observation X_k, as long as the sequence of parameters of the model, θ_k, lives in the parameter set $\Theta(\rho)$, for some $\rho < 1$.

As we saw above, the eigenvalues of the matrix appearing in the conditional gain vector g_{dk} are upper and lower bounded by multiples of $\|\mathbf{X}_{d(k-1),d}\|_2^2$. We can easily get rid of this dependence by using the scaled gain $\bar{G}_{dk}$ defined at the end of Section 6.4, for $s(x) = \|x\|_2^2$ and large enough κ. The derivation in (6.26) shows that (A2) still holds for this rescaled gain. The largest eigenvalue of the matrix in $\bar{g}_{dk}$ corresponding to $\bar{G}_{dk}$ is going to be almost surely bounded by construction. We need then to verify that the smallest eigenvalue of the matrix in $\bar{g}_{dk}$ has conditional expectation bounded away from zero such that (A1) holds. Note that

$$\begin{aligned} &\mathbb{E}\big[\|\mathbf{X}_{d(k-1),d}\|_2^2 \big| \mathbf{X}_{d(k-2),d}\big] \geq \mathbb{E}\big[X_{d(k-1)}^2 \big| \mathbf{X}_{d(k-2),d}\big] = \\ &\mathbb{E}\big[\big(\theta_{dk,1}X_{d(k-1)-1} + \cdots + \theta_{dk,d}X_{d(k-2)} + \xi_{d(k-1)}\big)^2 \big| \mathbf{X}_{d(k-2),d}\big]. \end{aligned}$$

There are three different types of terms in the sum above: a) error terms which are independent of the filtration, b) observations which are measurable with respect to the filtration, and c) observations which can be written as an error term which is independent of the filtration and a linear combination of previous observations of the process. The sum can therefore be written as the sum of two terms, namely: a) a linear combination of terms which are measurable with respect to the filtration, and b) a linear combination of error terms which are independent of the filtration. This can then be bounded in the same way as (6.43). We conclude that the previous display is lower bounded by σ^2.

One can then proceed as in Lemma 6.5 to show that for an appropriately large κ, $\mathbb{E}\big[\min(X_{d(k-1)}^2, \kappa)\big|\mathbf{X}_{d(k-2),d}\big]$ is positive; we omit this derivation.

For the most part, the requirement that the errors be Gaussian is not used extensively so we expect that the same results hold simply under appropriate moment assumptions: one could still use the gain (6.50) and bound its conditional expectation directly instead of using the Kullback-Leibler representation in (6.49) and assure the validity of (A1) and (A2) based on moment assumptions on the error terms and on the initial conditions for the model as we did in the one-dimensional case.

6.7 Numerical examples

6.7.1 Quantile tracking

In this section we return to the numerical example from Section 5.5 where we considered, for $n \in \mathbb{N}$, the model

$$X_i = f(t_i) + \sigma(t_i)\xi_i, \quad i = 1, \ldots, n,$$

where the ξ_i are independent standard normal random variables, $\mathbf{t}^{(n)} = \left(1/n - 1, 3/n - 1, \ldots, 1 - 3/n, 1 - 1/n\right)$ and, for $t \in [-1, 1]$,

$$f(t) = \frac{\sin(t)}{t}, \quad \sigma(t) = 0.1 \exp(1 - t).$$

Our quantity of interest was, for $\alpha \in (0, 1)$, the sequence $\theta_k = \theta_{\alpha,k} = \vartheta_\alpha(t_k)$, where

$$\vartheta_\alpha(t) = f(t) + \sigma(t)\Phi^{-1}(\alpha), \quad t \in [-1, 1],$$

a quantile function of level α.

In Section 5.5 we proposed a tracking sequence for θ_k based on a specific choice of gain function. Our main result from this chapter provides us with an alternative approach for tracking the drifting quantile θ_k. We can express the quantiles θ_k as a functional of $\mu_k = f(t_k)$ and $v_k = \sigma(t_k)$, i.e., $\theta_k = \phi_\alpha(\mu_k, v_k)$ for

$$\phi_\alpha(s, t) = s + t\,\Phi^{-1}(\alpha),$$

where Φ is the cumulative distribution function of a standard normal random variable. If we have sequences $\hat{\mu}_k$ and $\hat{v}_k$ which respectively track μ_k and v_k then

$$\hat{\theta}_k = \phi_\alpha(\hat{\mu}_k, \hat{v}_k)$$

is a tracking sequence for θ_k.

Consider the gains

$$G_1(x, \mu) = x - \mu, \quad G_2(x, v|\mu) = (x - \mu)^2 - v, \tag{6.51}$$

based on (6.18). Assuming that by time k we have observed $\mathbf{X}_k = (X_1, \ldots, X_k)$ we can use this data and the gains (6.51) to define the tracking sequences

$$\begin{aligned} \hat{\mu}_k &= \Pi_{\bar{C}}\Big(\hat{\mu}_{k-1} + \gamma_{1,k} G_1(X_k, \hat{\mu}_{k-1})\Big), \\ \hat{v}_k &= \Pi_{\bar{C}}\Big(\hat{v}_{k-1} + \gamma_{2,k} G_2(X_k, \hat{v}_{k-1} \mid \hat{\mu}_{k-1})\Big), \\ \hat{\theta}_k &= \phi_\alpha(\hat{\mu}_k, \hat{v}_k), \end{aligned} \tag{6.52}$$

where $\hat{\mu}_0 = \hat{v}_0 = 2$, $\bar{C} = 5$ and $\gamma_{1,k} \equiv C_{1,\gamma}(\log(n)/n^2)^{1/3}$ and $\gamma_{2,k} \equiv C_{2,\gamma}(\log(n)/n^2)^{1/3}$ are the step sequences.

We repeated the numerical study from Section 5.5 for this new tracking sequence for θ_k. We took $C_{1,\gamma} = C_{2,\gamma} = 2.5$, $n \in \{100, 250, 500, 1000, 10000\}$ and $\alpha \in \{0.1,\ 0.25,\ 0.5,\ 0.75,\ 0.9\}$. We present the results in Figure 6.7.1.

There seems to be some improvement over the results from Section 5.5, especially for low sample size. The

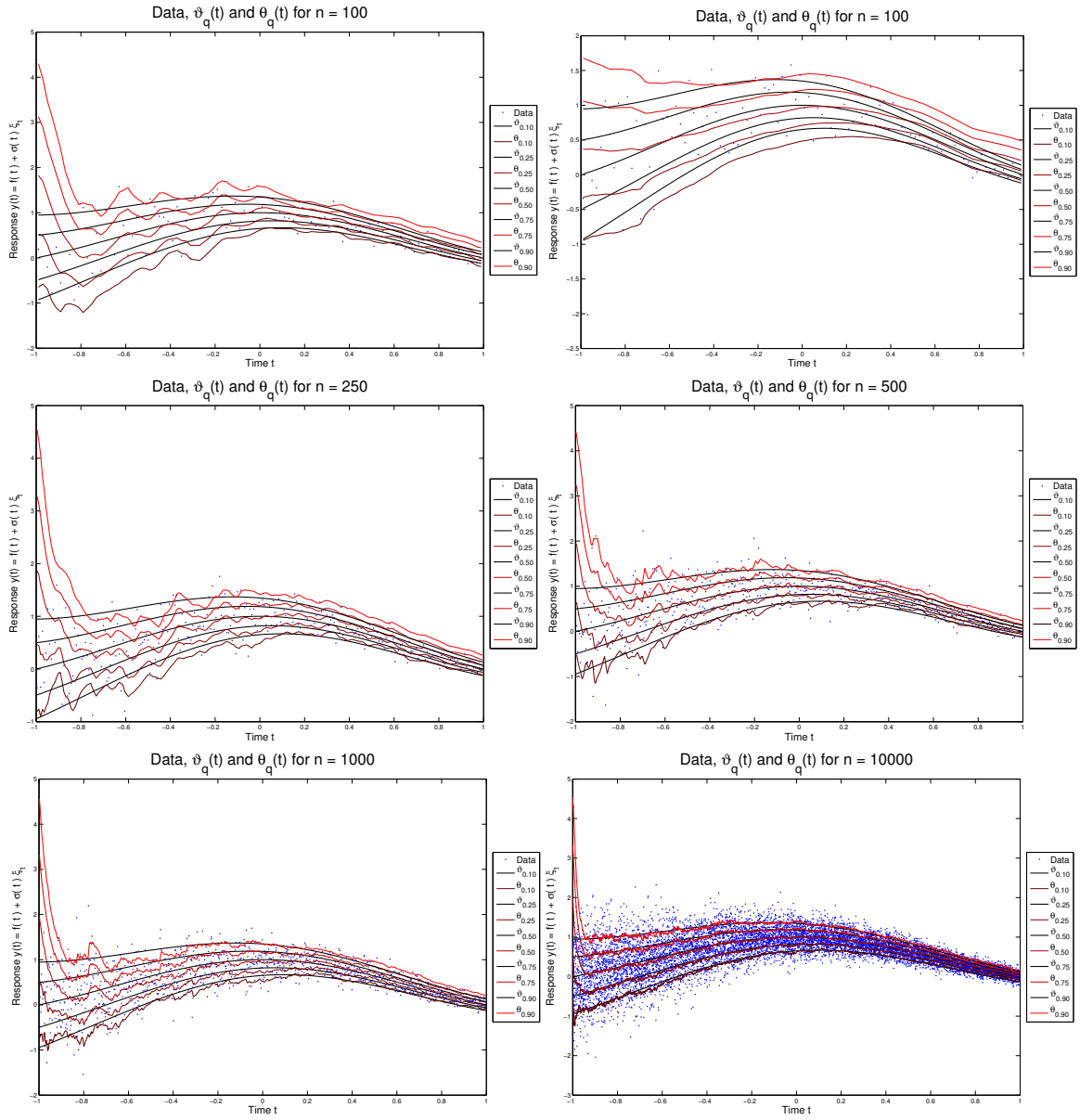

Figure 6.7.1: Results of the tracking algorithm. All pictures contain the data (blue dots), the true quantile function for the chosen values of α (black lines), and the respective tracking sequences (tones of red). To each picture corresponds a specific sample size. On the first row we compare, for $n = 100$, the *raw* tracking sequence (left) with a smoothed version of it (right).

tracking sequences also seem to be less noisy than the ones obtained in Section 5.5. These improvements are not surprising since the gain considered there only depends on the value of the indicators $\mathbb{1}\{X_k < \hat{\theta}_{k-1}\}$ rather than on the actual observation X_k as is the case with the gain (6.52).

6.7.2 Poisson rain

In this section we present some more numerical results. Our model is a 3 dimensional Poisson process with two space dimensions and one time dimension. More concretely, we will take a Poisson process on the unit square, evolving in time. This kind of model is sometimes called *Poisson rain* – we can imagine the Poisson events as raindrops falling on the unit square over time. Our Poisson process $\big(N_t : t \in [0,1]\big)$ will have an intensity function $\lambda(x,y,t) : [0,1]^3 \mapsto \mathbb{R}^+$ such that an event is a point in $[0,1]^2$.

To fit this model into our framework we will discretize the unit square into a grid of 400 equally sized, 1/20 by 1/20 squares, and make observations of the process every 1/1000 time units. We will observe the process for 1 time unit such that by time $t = 1$ our data will be, say, a 20 by 20 by 1000 (three-dimensional) matrix M. The (i,j,k)-entry in the matrix, call it $m_{i,j,k}$, $i,j = 1,\ldots,20$, $k = 1,\ldots,1000$ is an observation of a Poisson random variable with intensity

$$\lambda_{i,j,k} = \int_{(k-1)/1000}^{k/1000} \int_{(i-1)/20}^{i/20} \int_{(j-1)/20}^{j/20} \lambda(x,y,t)\, dx\, dy\, dt,$$

for $i,j = 1,\ldots,20$, $k = 1,\ldots,1000$; these 400.000 Poisson random variables are mutually independent.

The specific intensity function used in our simulation is obtained, appropriately enough, from a picture of a cloud. The objective is to see the resulting Poisson process as a crude model for rainfall. On a computer, images are arrays of pixels, which are small squares characterized by a potentially different color. More precisely, image files can be seen as a collection of 3 matrices of the same size, each corresponding to a color channel – red, green and blue. Each entry of each matrix contains a number in $\{0,1,\ldots,255\}$, corresponding to the level of the respective primary color; each triplet of colors corresponding to the a fixed position in the matrices characterizes the corresponding pixel. When combined, each triplet encodes one of roughly 16 million colors ($256^3 = 2^{24}$) which is called the 24-bit color palette. For example, the triplet $(0,0,0)$ corresponds to black, $(255,0,0)$ to red, $(255,255,255)$ to white. We took then a 1000×1000 pixel image and discarded all but the red channel. The resulting 1000×1000 matrix contains entries in $\{0,1,\ldots,255\}$ and encodes the image seen in Figure 6.7.2.

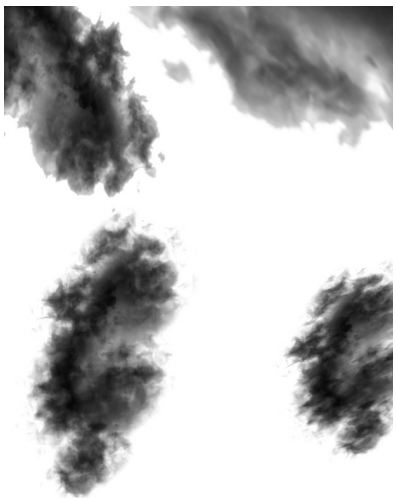

Figure 6.7.2: Picture used to define the intensity function of the Poisson process. Image obtained from [31].

The intensity function at time t, $(x, y) \mapsto \lambda(x, y, t)$ is obtained by taking a 500×500 sub-matrix of the cloud image and seeing each entry as a height for a two-dimensional histogram, which is then linearly interpolated and whose support we rescale to the unit square. The resulting function obtained from this 500×500 window, maps the unit square to the interval $[0, 255]$ and is 255 minus our intensity function at time t – we invert the color such that the white in the picture represents low intensity (0) and the black represents high intensity (255). At time $t = 0$ we take this window to be in the bottom left side of Figure 6.7.2 and then we slowly move the window along a clock-wise spiral towards the center of the image. This represents the evolution of the intensity function along the time axis.

We simulated data from this model. For each $i, j = 1, \dots, 20$ the sequence $(m_{i,j,k} : k = 1, \dots, 1000)$ is a sequence of observations from the model described in Section 6.6.1 with $n = 1000$ and $\theta_k^n = \lambda_{i,j,k}$, $k = 1, \dots, 1000$. For the purpose of this implementation we assumed that for each (i, j), the intensity function is a Lipschitz function with smoothness parameter $\alpha = 1$. By this we mean that for $i, j = 1, \dots, 20$, $\lambda_{i,j}(t) \in \mathcal{L}_1([0, 1])$, where

$$\lambda_{i,j}(t) = \int_{(i-1)/20}^{i/20} \int_{(j-1)/20}^{j/20} \lambda(x, y, t)\, dx\, dy.$$

For this setup, our algorithm tracks the matrix valued parameter

$$\vartheta_k = \left[\lambda_{i,j,k}\right]_{i,j} = \left[\lambda_{i,j}(k/n)\right]_{i,j} \quad i, j = 1, \dots, 20,$$

by running 20×20 tracking sequences in parallel where each component $\hat{\theta}_{i,j,k}$ evolves according to

$$\hat{\theta}_{i,j,k+1} = \hat{\theta}_{i,j,k} + \gamma_k G(m_{i,j,k}, \hat{\theta}_{i,j,k}), \quad i, j = 1, \dots, 20,$$

with $\gamma_k = C_\gamma n^{-2/3} \log^{1/3} n$ and for the gain $G(x, \theta) = x - \theta$ as in (6.21); for the results seen in Figure 6.7.3 we took $C_\gamma = 2.25$. In [83] and [84] we also made available videos depicting the procedure of simulating the data and the algorithm running, respectively. For the results seen in the video, we have smoothed the tracking sequence spatially by replacing each $\hat{\theta}_{i,j,k}$ with the mean of its neighboring cells.

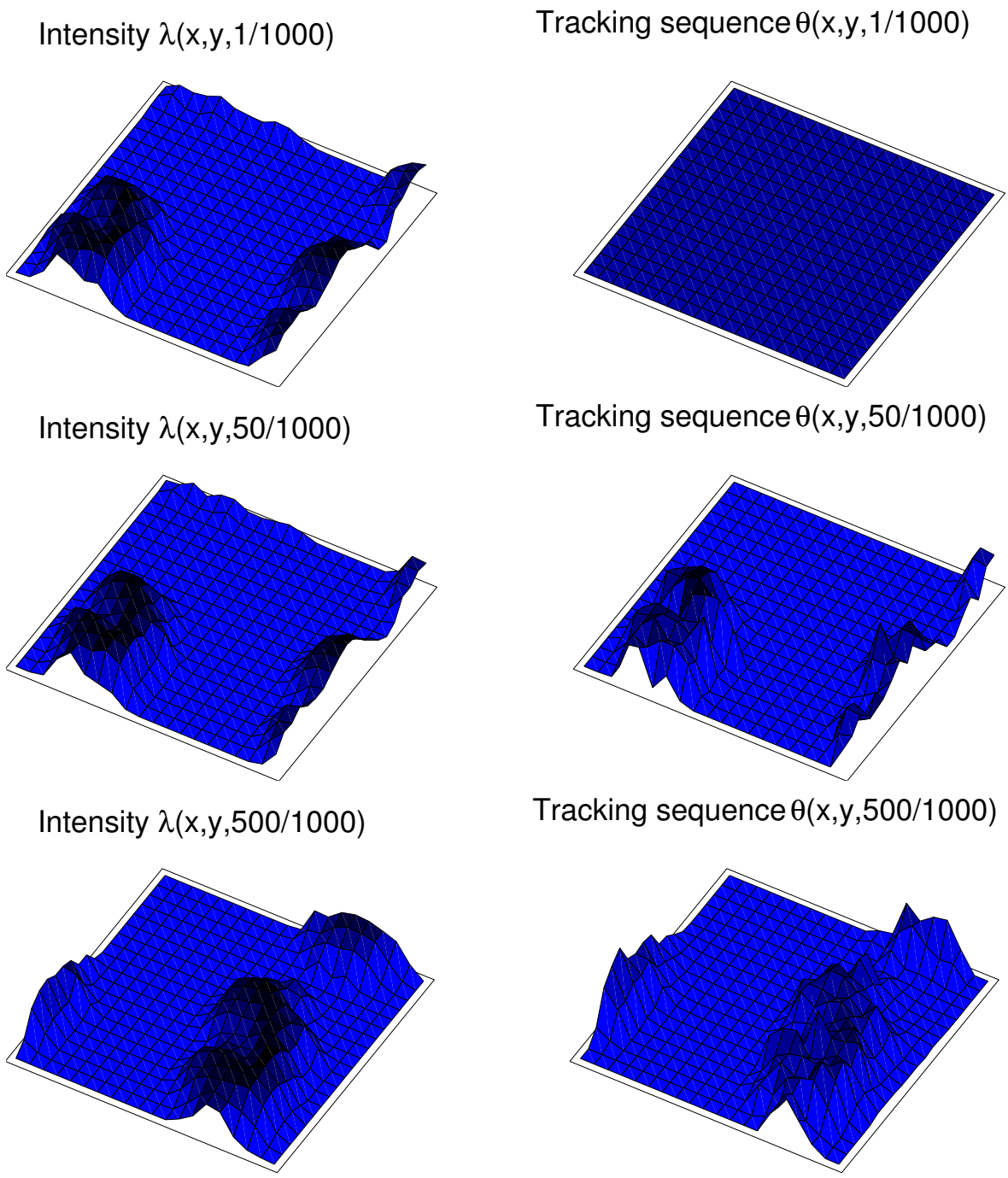

Figure 6.7.3: True intensity (left column) and the obtained tracking sequence (right column) at t=1/1000,50/1000,500/1000 (top to bottom). The displayed tracking sequence corresponds to a linear interpolation of the actual tracking sequence.

The algorithm was initiated from a zero matrix, i.e., we took $\theta_{i,j,0} = 0$, for $i, j = 1, \ldots, 20$. It moves quite quickly to a reasonable approximation of the true intensity function and then accompanies the evolution of the intensity. The parameter of the algorithm C_γ needs to be picked large enough as seen in Section 6.5.3. If it is taken too large, though, it will affect the variance of the tracking sequence. To make sure that C_γ can be taken appropriately large and that the resulting tracking sequence does not have too large variability, we can smooth the tracking sequence along the space dimensions. In this sense, we can, for fixed t, replace each approximation by the average of its neighbors and smooth out the tracking sequence spatially. Alternatively, one can also treat the discretization parameter for the grid (which we took here as 20) as a parameter for the algorithm and pick it appropriately, depending on n.

The method seems to work reasonably well at capturing the shape of the underlying intensity function and its evolution in time. Without any prior knowledge on the intensity function, the choice of an adequate step size for the algorithm, including the value of the constant in the sequence γ_k requires some experimentation. These choices can, however, be motivated by tuning the algorithm using training data.

6.8 Proofs of the lemmas

Proof:[Proof of Lemma 6.1] First suppose that $y = Mx$ for some symmetric positive-definite matrix M such that $0 < \lambda_1 \le \lambda_{(1)}(M) \le \lambda_{(d)}(M) \le \lambda_2 < \infty$. Then $\langle x, y\rangle = x^T Mx$ and therefore

$$0 < \lambda_1 \|x\|_2^2 \le \lambda_{(1)}(M)\|x\|_2^2 \le \langle x, y\rangle \le \lambda_{(d)}(M)\|x\|_2^2 \le \lambda_2 \|x\|_2^2$$

and

$$\|y\|_2^2 = \langle y, y\rangle = x^T M^T Mx = x^T M^2 x \le \lambda_2^2 \|x\|_2^2.$$

Now we prove the converse assertion. Suppose $x, y \in \mathbb{R}^d$ and $0 < \lambda_1' \|x\|_2^2 \le \langle x, y\rangle \le \lambda_2' \|x\|_2^2 < \infty$ for some $\lambda_1', \lambda_2' \in \mathbb{R}$ such that $0 < \lambda_1' \le \lambda_2' < \infty$ and that $\|y\|_2 \le C\|x\|_2$. Let $V = \{v = ax + by : a, b \in \mathbb{R}\}$ be the linear space spanned by x and y. First consider the case $\dim(V) = 1$, i.e., $y = \alpha x$ for some $\alpha \in \mathbb{R}$. Then $\langle y, x\rangle = \alpha\|x\|_2^2$ so that $0 < \lambda_1' \le \alpha \le \lambda_2' < \infty$. Thus $y = \alpha x = Mx$ with symmetric and positive $M = \alpha I$ so that $0 < \lambda_1' \le \alpha = \lambda_{(1)}(M) = \lambda_{(d)}(M) \le \lambda_2' < \infty$.

Now consider the case $\dim(V) = 2$. Let $e_1 = x/\|x\|_2$ and $\{e_1, e_2\}$ be an orthonormal basis of V. Then

$$\begin{aligned} x &= \|x\|_2 e_1 \\ y &= \alpha e_1 + \beta e_2. \end{aligned}$$

The conditions $\lambda_1'\|x\|_2^2 \le \langle x, y\rangle = \alpha\|x\|_2 \le \lambda_2'\|x\|_2^2$ and $\|y\|_2 = \sqrt{\alpha^2 + \beta^2} \le C\|x\|_2$ imply that

$$\lambda_1'\|x\|_2 \le \alpha \le \min\{\lambda_2', C\}\|x\|_2, \quad |\beta| \le C\|x\|_2.$$

Let e_2 be chosen in such a way that $\beta > 0$ (which is always possible.) Now, we change the basis of V as follows:

$$\begin{aligned} e_1' &= \cos(\theta)e_1 - \sin(\theta)e_2, \\ e_2' &= \sin(\theta)e_1 + \cos(\theta)e_2. \end{aligned}$$

We thus rotate the basis $\{e_1, e_2\}$ by the angle θ. In this new basis,

$$x = \|x\|_2 \cos(\theta) e_1' + \|x\|_2 \sin(\theta) e_2' = \alpha_x e_1' + \beta_x e_2',$$
$$y = (\alpha \cos(\theta) - \beta \sin(\theta)) e_1' + (\alpha \sin(\theta) + \beta \cos(\theta)) e_2' = \alpha_y e_1' + \beta_y e_2'.$$

Recall that $\alpha, \beta > 0$. Take $\theta \in (0, \pi/2)$ such that $\alpha \cos(\theta) - \beta \sin(\theta) = \frac{1}{2}\alpha \cos(\theta)$ (i.e., $\tan(\theta) = \frac{\alpha}{2\beta}$). Then,

$$\frac{\lambda_1'}{2} \le \frac{\alpha}{2\|x\|_2} = \frac{\alpha_y}{\alpha_x} \le \frac{\min\{\lambda_2', C\}}{2}, \quad \lambda_1' \le \frac{\alpha}{\|x\|_2} \le \frac{\beta_y}{\beta_x} \le \frac{\alpha}{\|x\|_2} + \frac{2\beta^2}{\alpha\|x\|_2} \le \min\{\lambda_2', C\} + \frac{2C^2}{\lambda_1'}.$$

Take then $\lambda_1 = \lambda_1'/2$ and $\lambda_2 = \min\{\lambda_2', C\} + 2C^2/\lambda_1'$.

Let $\{e_3, \ldots, e_d\}$ be the orthonormal basis of $V^\perp$, so that $b = \{e_1', e_2', e_3, \ldots, e_d\}$ is an orthonormal basis of $\mathbb{R}^d$. Take

$$M' = \left[\begin{array}{c|c} D & 0 \\ \hline 0 & I_{d-2} \end{array}\right] \quad \text{with} \quad D = \begin{bmatrix} \alpha_y/\alpha_x & 0 \\ 0 & \beta_y/\beta_x \end{bmatrix}$$

where the 0's indicate null matrices of the appropriate orders. We then have that $y = M'x$ in the basis b and $\lambda_1 \le \lambda_{(1)}(M') \le \lambda_{(d)}(M') \le \lambda_2$. We can finally obtain M by using the (orthogonal) change of basis matrix E from basis b to the canonical basis of $\mathbb{R}^d$ as $M = E^{-1}M'E = E^T M'E$. Note that M has the same eigenvalues as M' (which are all positive and finite) and is symmetric. □

Proof:[Proof of Lemma 6.2] For the sake of brevity, we use the notations $\theta_k = \theta_k(\mathbf{X}_{k-1})$, $G_k = G(X_k, \hat{\theta}_k|\mathbf{X}_{k-1})$ and $g_k = g(\hat{\theta}_k, \theta_k|\mathbf{X}_{k-1})$, $k \in \mathbb{N}$, $\mathcal{F}_k = \sigma(\mathbf{X}_k)$ is the σ-field generated by $\mathbf{X}_k = (X_0, X_2, \ldots, X_k)$.

Recall that Θ is compact so that $\sup_{\theta\in\Theta} \|\theta\|_2 \le C_\Theta$. First assume $\mathbb{E}\|\hat{\theta}_k\|_2^2 \le KC_\Theta^2$. By (6.5) and (6.6), we obtain

$$\mathbb{E}\|G_k\|_2^2 = \mathbb{E}\|G_k - g_k + g_k)\|_2^2 \le 2C + 4L(\mathbb{E}\|\theta_k\|_2^2 + \mathbb{E}\|\hat{\theta}_k\|_2^2) \le 2C + 4L(K+1)C_\Theta^2 = C_1,$$

which implies, in view of (6.7) and $\gamma_k \le \Gamma$,

$$\mathbb{E}\|\hat{\theta}_{k+1}\|_2^2 \le 2\mathbb{E}\|\hat{\theta}_k\|_2^2 + 2\gamma_k^2\mathbb{E}\|G_k\|_2^2 \le 2KC_\Theta^2 + 2\Gamma^2 C_1 = C_2.$$

Next, consider the case $\mathbb{E}\|\hat{\theta}_k\|_2^2 > KC_\Theta^2$ which of course implies $\mathbb{E}\|\hat{\theta}_k\|_2^2 > K\mathbb{E}\|\theta_k\|_2^2$. As M_k is a symmetric positive-definite matrix such that $0 < A \le \lambda_{(1)}(M_k) \le \lambda_{(d)}(M_k) \le B < \infty$, by the Cauchy-Schwarz inequality,

$$\hat{\theta}_k^T M_k \theta_k \le |\hat{\theta}_k^T M_k \theta_k| \le \left(\hat{\theta}_k^T M_k \hat{\theta}_k\right)^{1/2} \left(\theta_k^T M_k \theta_k\right)^{1/2} \le B\|\hat{\theta}_k\|_2 \|\theta_k\|_2.$$

By using the last relation, (6.2), (6.5), (6.6) and (6.7), we evaluate $\mathbb{E}\|\hat{\theta}_{k+1}\|_2^2$:

$$\begin{aligned}
\mathbb{E}\|\hat{\theta}_{k+1}\|_2^2 \le& \mathbb{E}\|\hat{\theta}_k\|_2^2 + 2\gamma_k \mathbb{E}\left[\hat{\theta}_k^T \mathbb{E}(G_k|\mathcal{F}_{k-1})\right] + \gamma_k^2 \mathbb{E}\|G_k\|_2^2 \\
\le& \mathbb{E}\|\hat{\theta}_k\|_2^2 - 2\gamma_k \mathbb{E}\left(\hat{\theta}_k^T M_k(\hat{\theta}_k - \theta_k)\right) + \gamma_k^2\left[2C + 4L(\mathbb{E}\|\theta_k\|_2^2 + \mathbb{E}\|\hat{\theta}_k\|_2^2)\right] \\
\le& \mathbb{E}\|\hat{\theta}_k\|_2^2 - 2\gamma_k\left[A\mathbb{E}\|\hat{\theta}_k\|_2^2 - \mathbb{E}\left(\hat{\theta}_k^T M_k \theta_k\right)\right] + \gamma_k^2\left[2C + 4LC_\Theta^2 + 4L\mathbb{E}\|\hat{\theta}_k\|_2^2)\right] \\
\le& \mathbb{E}\|\hat{\theta}_k\|_2^2 - 2\gamma_k\left[A\mathbb{E}\|\hat{\theta}_k\|_2^2 - B\mathbb{E}\left(\|\hat{\theta}_k\|_2\|\theta_k\|_2\right)\right] + \gamma_k^2\left[2C + 4LC_\Theta^2 + 4L\mathbb{E}\|\hat{\theta}_k\|_2^2)\right].
\end{aligned}$$

From $\mathbb{E}\|\hat{\theta}_k\|_2^2 > K\mathbb{E}\|\theta_k\|_2^2$ and the Cauchy-Schwarz inequality, it follows that $\mathbb{E}\|\hat{\theta}_k\|_2\|\theta_k\|_2 \le \left(\mathbb{E}\|\hat{\theta}_k\|_2^2\mathbb{E}\|\theta_k\|_2^2\right)^{1/2} \le$

$\mathbb{E}\|\hat{\theta}_k\|_2^2/\sqrt{K}$. Using this, we proceed by bounding the previous display as follows:

$$\begin{aligned}
&\leq \mathbb{E}\|\hat{\theta}_k\|_2^2 - 2\gamma_k\big(A\mathbb{E}\|\hat{\theta}_k\|_2^2 - B(\mathbb{E}\|\theta_k\|_2^2\mathbb{E}\|\hat{\theta}_k\|_2^2)^{1/2}\big) + \gamma_k^2\big[2C + 4LC_\Theta^2 + 4L\mathbb{E}\|\hat{\theta}_k\|_2^2)\big] \\
&\leq \mathbb{E}\|\hat{\theta}_k\|_2^2 - \gamma_k\mathbb{E}\|\hat{\theta}_k\|_2^2\Big(2A - \frac{2B}{\sqrt{K}} - \gamma_k 4L\Big) + \gamma_k^2(2C + 4LC_\Theta^2) \\
&\leq \mathbb{E}\|\hat{\theta}_k\|_2^2 - \gamma_k\mathbb{E}\|\hat{\theta}_k\|_2^2\Big(2A - \frac{2B}{\sqrt{K}} - \gamma_k 4L\Big) + \gamma_k^2(2C + 4LC_\Theta^2)\frac{\mathbb{E}\|\hat{\theta}_k\|_2^2}{KC_\Theta^2} \\
&= \mathbb{E}\|\hat{\theta}_k\|_2^2 - \gamma_k\mathbb{E}\|\hat{\theta}_k\|_2^2\Big(2A - \frac{2B}{\sqrt{K}} - \gamma_k\frac{4LC_\Theta^2(K+1) + 2C}{KC_\Theta^2}\Big) \leq \mathbb{E}\|\hat{\theta}_k\|_2^2,
\end{aligned}$$

for sufficiently large K and sufficiently small γ_k. Thus, for sufficiently large K and sufficiently small γ_k, $\mathbb{E}\|\hat{\theta}_{k+1}\|_2^2 \leq C_2$. □

Lemma 6.3 *Let M be a symmetrical positive-definite matrix of order d with (increasing) eigenvalues $\lambda_{(i)}(M)$, the smallest and largest of which we denote as $\lambda_{(1)}(M)$ and $\lambda_{(d)}(M)$ respectively. Then, for $\gamma > 0$ such that $\gamma\lambda_{(d)}(M) < 1$, and constants $K_p > 0$, $p \in \mathbb{N}$,*

$$\begin{aligned}
|||M|||_p &\leq K_p|||M|||_2 = K_p\lambda_{(d)}(M), \\
0 < \lambda_{(1)}(I - \gamma M) &\leq \lambda_{(d)}(I - \gamma M) = 1 - \gamma\lambda_{(1)}(M) < 1,
\end{aligned}$$

where for $p \in \mathbb{N}$, $|||M|||_p$ is the operator norm induced by l_p.

Proof: Note that for $x \in \mathbb{R}^d$, if

$$R_2^p = \max_{x\neq 0}\frac{\|x\|_p}{\|x\|_2}, \quad R_p^2 = \max_{x\neq 0}\frac{\|x\|_2}{\|x\|_p},$$

then it follows (cf. Horn and Johnson [47, Theorem 5.6.18])

$$\max_{M\neq 0}\frac{|||M|||_p}{|||M|||_2} = R_2^p R_p^2 = K_p.$$

We then have (c.f. Horn and Johnson [47, Section 5.6.6]) that for M a real, symmetrical, positive-definite matrix, where $\lambda_{(i)}(M)$ is the i-th largest eigenvalue of a matrix M,

$$|||M|||_2 = \max_i\sqrt{\lambda_i(M^TM)} = \max_i\sqrt{\lambda_{(i)}(M^2)} = \lambda_{(d)}(M).$$

The first statement then follows. Note that by application of the Hölder inequality, we have $\|x\|_p \leq d^{(q-p)/(qp)}\|x\|_q$ for $p \leq q$ and so we can take $K_p = d^{(p-1)/(2p)}$ if $p \geq 2$ and $K_p = d^{1/2}$ if $p = 1$.

It is straightforward to check that the matrix $I - \gamma M$ has eigenvalues $1 - \gamma\lambda_i$. Now, if $\gamma\lambda_{(d)}(M) < 1$ then for all $i = 1, \ldots, d$, $0 < \gamma\lambda_{(1)}(M) \leq \gamma\lambda_i \leq \gamma\lambda_{(d)}(M) < 1$ implying $1 > 1 - \gamma\lambda_{(1)}(M) \geq 1 - \gamma\lambda_i \geq 1 - \gamma\lambda_{(d)}(M) > 0$ and so $\max_{i=1,\ldots,d}|1 - \gamma\lambda_i| = 1 - \gamma\lambda_{(1)}(M) < 1$. □

Lemma 6.4 (Abel Tranformation) *For $k_0, k \in \mathbb{N}$ such that $k_0 \leq k$, let $a_i \in \mathbb{R}^d$, $i = k_0, \ldots, k$, B_i, $i = k_0, \ldots, k$, be square $d \times d$ matrices and $A_i = \sum_{j=k_0}^{i} a_j$, $i = k_0, \ldots, k$. Then*

$$\sum_{i=k_0}^{k} B_i a_i = \sum_{i=k_0}^{k-1} (B_i - B_{i+1}) A_i + B_k A_k.$$

Proof: We prove this by induction on k. For $k = k_0$ we simply have $B_{k_0} a_{k_0} = B_{k_0} A_{k_0} = B_{k_0} a_{k_0}$ and the assertion holds. Let us assume then that the equality holds for $k = n$ and let us prove the result for $k = n + 1$. We have

$$\begin{aligned}
\sum_{i=k_0}^{n+1} B_i a_i &= \sum_{i=k_0}^{n} B_i a_i + B_{n+1} a_{n+1} = \sum_{i=k_0}^{n-1} (B_i - B_{i+1}) A_i + B_n A_n + B_{n+1} a_{n+1} \\
&= \sum_{i=k_0}^{n} (B_i - B_{i+1}) A_i - (B_n - B_{n+1}) A_n + B_n A_n + B_{n+1} a_{n+1} \\
&= \sum_{i=k_0}^{n} (B_i - B_{i+1}) A_i + B_{n+1} A_{n+1}.
\end{aligned}$$

□

Lemma 6.5 *Consider an AR(1) model with a random, drifting parameter θ_k,*

$$X_k = X_{k-1}\theta_k + \xi_k, \quad k \in \mathbb{N},$$

where the random variables ξ_k are independent of $\sigma(X_0, \ldots, X_{k-1})$, the σ-algebra generated by $\mathbf{X}_{k-1}$ and for all $k \in \mathbb{N}$, $\mathbb{E}\xi_k = \mathbb{E}\xi_k^3 = 0$, $\mathbb{E}\xi_k^2 = \sigma^2 > 0$ and, for some constant $0 \leq c < 5$, $\mathbb{E}\xi_k^4 = c\,\sigma^4$. Let also X_0 be such that $\mathbb{E}X_0^2$ and $\mathbb{E}X_0^4$ are bounded. We assume that the drifting parameter θ_k is measurable with respect to $\sigma(\mathbf{X}_{k-1})$, and verifies $|\theta_k| \leq q < 1$, almost surely, for every $k \in \mathbb{N}$. Then, for any s such that $4s \geq (9 - c)\sigma^2$,

$$\mathbb{E}\big[\min\left(X_t^2, s\right)\big|X_{t-1}\big] \geq \frac{5 - c}{4}\sigma^2.$$

Proof: Note first that since $\sigma^2 > 0$, if $\mathbb{E}X_0^2$ and $\mathbb{E}X_0^4$ are bounded then we can write $\mathbb{E}X_0^2 \leq c_1\sigma^2$ and $\mathbb{E}X_0^4 \leq c_2\sigma^4$ for some $c_1, c_2 \geq 0$. Using the independence of the noise and the bound on the norm of the autoregressive parameters we have that

$$\mathbb{E}X_k^2 = \mathbb{E}(X_{k-1}\theta_k + \xi_k)^2 = \mathbb{E}[X_{k-1}^2\theta_k^2] + 2\mathbb{E}[X_{k-1}\theta_k]\mathbb{E}\xi_k + \mathbb{E}\xi_k^2 \leq q^2\mathbb{E}X_{k-1}^2 + \sigma^2,$$

and by using this recursion we conclude that

$$\mathbb{E}X_k^2 \leq q^{2k}\mathbb{E}X_0^2 + \sigma^2 \sum_{i=1}^{k-1} q^{2i} \leq \sigma^2\left(c_1 + \frac{1}{1 - q^2}\right) < \infty.$$

Using the previous display and proceeding in the same way,

$$\mathbb{E}X_k^4 = \mathbb{E}(X_{k-1}\theta_k + \xi_k)^4 \leq q^4\mathbb{E}X_{k-1}^4 + 6q^2\sigma^2\mathbb{E}X_{k-1}^2 + c\,\sigma^4 \leq q^4\mathbb{E}X_{k-1}^4 + \sigma^4\kappa,$$

with $\kappa = c + 6q^2c_1 + 6q^2/(1-q^2)$. Using this recursion we have that

$$\mathbb{E}X_k^4 \leq q^{4k}\mathbb{E}X_0^4 + \sigma^4\kappa\sum_{i=1}^{k-1} q^{4i} \leq \sigma^4\left(c_2 + \frac{\kappa}{1-q^4}\right) < \infty.$$

We can now use basic properties of the conditional expectation to see that,

$$\begin{aligned}
\mathbb{E}\big[X_k^2\big|X_{t-1}\big] &= X_{k-1}^2\theta_k^2 + 2X_{k-1}\theta_k\mathbb{E}\xi_k + \mathbb{E}\xi_k^2 = X_{k-1}^2\theta_k^2 + \sigma^2,\\
\mathbb{E}\big[X_k^4\big|X_{t-1}\big] &= X_{k-1}^4\theta_k^4 - 4X_{k-1}^3\theta_k^3\mathbb{E}\xi_k + 6X_{k-1}^2\theta_k^2\mathbb{E}\xi_k^2 - 4X_{k-1}\theta_k\mathbb{E}\xi_k^3 + \mathbb{E}\xi_k^4 =\\
&= X_{k-1}^4\theta_k^4 + 6X_{k-1}^2\theta_k^2\sigma^2 + c\,\sigma^4.
\end{aligned}$$

For $a, b \in \mathbb{R}$ we have $\min(a,b) = (a+b)/2 - |a-b|/2$ and so, by the Cauchy-Schwarz inequality and the last display,

$$\begin{aligned}
\mathbb{E}\Big[\min\left(X_t^2, \rho\,\sigma^2\right)\Big|X_{t-1}\Big] &= \mathbb{E}\Big[\frac{1}{2}X_k^2 + \frac{\rho}{2}\sigma^2 - \frac{1}{2}\Big|X_k^2 - \rho\sigma^2\Big|\Big|X_{t-1}\Big]\\
&\geq \frac{1}{2}X_{k-1}^2\theta_k^2 + \frac{\rho+1}{2}\sigma^2 - \frac{1}{2}\Big(\mathbb{E}\Big[\left(X_k^2 - \rho\sigma^2\right)^2\Big|X_{t-1}\Big]\Big)^{1/2},
\end{aligned}$$

for $\rho > 0$. We now have, by plugging in the expressions derived above and simplifying,

$$\begin{aligned}
\mathbb{E}\Big[\left(X_k^2 - \rho\sigma^2\right)^2\Big|X_{t-1}\Big] &= \mathbb{E}\big[X_k^4\big|X_{t-1}\big] - 2\rho\sigma^2\mathbb{E}\big[X_k^2\big|X_{t-1}\big] + \rho^2\sigma^4 =\\
&= X_{k-1}^4\theta_k^4 + 2(3-\rho)X_{k-1}^2\theta_k^2\sigma^2 + (c - 2\rho + \rho^2)\sigma^4 = \left(X_{k-1}^2\theta_k^2 + \frac{c+3}{4}\sigma^2\right)^2,
\end{aligned}$$

if we pick $\rho = (9-c)/4 > 1$. Combining the previous two displays we conclude that

$$\mathbb{E}\Big[\min\left(X_t^2, \frac{9-c}{4}\sigma^2\right)\Big|X_{t-1}\Big] \geq \frac{5-c}{4}\sigma^2,$$

and the statement of the lemma follows a fortiori. □

References

[1] R. P. Adams, I. Murray, and D. J. C. MacKay. Tractable nonparametric bayesian inference in poisson processes with gaussian process intensities. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 9–16, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-516-1. doi: 10.1145/1553374.1553376.

[2] G. Bassett and R. Koenker. Asymptotic theory of least absolute error regression. *Journal of the American Statistical Association*, 73(363):618–622, 1978.

[3] R. Bathia. *Matrix Analysis*, volume 169 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1997.

[4] M. Bebbington and R. Zitikis. A robust heuristic estimator for the period of a Poisson intensity function. *Methodology And Computing In Applied Probability*, 6(4):441–462, 2004.

[5] E. Belitser. Recursive estimation of a drifting autoregressive parameter. *Ann. Statist.*, 28(3):860–870, 2000.

[6] E. Belitser and S. Ghosal. Adaptive Bayesian inference on the mean of an infinite-dimensional normal distribution. *Ann. Statist.*, 31(2):536–559, 2003.

[7] E. Belitser and P. Serra. On properties of the algorithm for pursuing a drifting quantile. *Automation and Remote Control*, 74(4):613–627, April 2013.

[8] E. Belitser and P. Serra. Adaptive priors based on splines with random knots. *arXiv:1303.3365 [math.ST]*, 2013.

[9] E. Belitser and P. Serra. On-line tracking of a conditional quantile. *(preprint)*, 2013.

[10] E. Belitser and P. Serra. Online tracking of a drifting parameter of a time series. *arXiv:1306.0325 [math.ST]*, 2013.

[11] E. Belitser, P. Serra, and H. van Zanten. Estimating the period of a cyclic non-homogeneous Poisson process. *Scand. J. Stat.*, 40(2):204–218, June 2013.

[12] E. Belitser, P. Serra, and H. van Zanten. Rate optimal Bayesian intensity smoothing for inhomogeneous Poisson processes. *arXiv:1304.6017 [math.ST]*, 2013.

[13] E. N. Belitser and A. P. Korostelev. Pseudovalues and minimax filtering algorithms pseudovalues and minimax filtering algorithms for the nonparametric median. *Adv. in Sov. Math.*, 12:115–124, 1992.

[14] E. N. Belitser and S. van de Geer. *On robust recursive nonparametric curve estimation*, volume 47 of *Progress in Probability*, pages 391–404. Birkhäuser Boston, 2000.

[15] A. Benveniste, M. Metivier, and P. Priouret. *Adaptive Algorithms and Stochastic Approximation*. New York, Berlin: Springer-Verlag., 1990.

[16] S. P. Brooks, P. Giudici, and G. O. Roberts. Efficient construction of reversible jump markov chain monte carlo proposal distributions. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 65 (1):3–55, 2003.

[17] J.-M. Brossier. *Egalization Adaptive et Estimation de Phase: Application aux Communications Sous-Marines.* PhD thesis, Institut National Polytechnique de Grenoble, 1992.

[18] L. Brown, N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, and L. Statistical analysis of a telephone call center: a queueing-science perspective. *J. Amer. Statist. Assoc.*, 100(469):36–50, 2005. ISSN 0162-1459. doi: 10.1198/016214504000001808.

[19] H. D. Brunk. Maximum likelihood estimates of monotone parameters. *Ann. Math. Statist.*, 26(4):607–616., 1955.

[20] B. Cade and B. Noon. A gentle introduction to quantile regression for ecologists. *Frontiers in Ecology and the Environment*, 1(8):412–420, 2003. ISSN 1540-9295. doi: 10.1890/1540-9295(2003)001[0412:agitqr]2.0.co;2.

[21] Y. S. Chow and H. Teicher. *Probability Theory. Independence, Interchangeability, Martingales*. Springer texts in Statistics. Springer Verlag, New York, second edition, 1988.

[22] I. P. Cornfeld, S. V. Fomin, and Y. G. Sinai. *Ergodic Theory*, volume 245 of *A Series of Comprehensive Studies in Mathematics*. Springer-Verlag, New York, 1982.

[23] C. de Boor. *A practical guide to splines*. Springer-Verlag, New York, 1978.

[24] C. de Boor. *A practical guide to splines*, volume 27 of *Applied Mathematical Sciences*. Springer-Verlag, New York, revised edition, 2001. ISBN 0-387-95366-3.

[25] R. de Jonge and J. H. van Zanten. Adaptive estimation of multivariate functions using conditionally Gaussian tensor-product spline priors. *Electron. J. Stat.*, 6:1984–2001, 2012.

[26] B. Delyon and A. Juditsky. Asymptotical study of parameter tracking algorithms. *SIAM Journal on Control and Optimization*, 33(1):323–345, January 1995.

[27] D. Denison, B. Mallick, and A. Smith. Bayesian MARS. *Statistics and Computing*, 8:337–346, 1998.

[28] D. G. T. Denison, B. K. Mallick, and A. F. M. Smith. Automatic Bayesian curve fitting. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 60(2):333–350, 1998. ISSN 1369-7412. doi: 10.1111/1467-9868.00128.

[29] P. Diaconis and D. Freedman. On the consistency of Bayes estimates. *Ann. Statist.*, 14(1):1–67, 1986. ISSN 0090-5364. doi: 10.1214/aos/1176349830. With a discussion and a rejoinder by the authors.

[30] I. DiMatteo, C. R. Genovese, and R. E. Kass. Bayesian curve-fitting with free-knot splines. *Biometrika*, 88 (4):1055–1071, 2001. ISSN 0006-3444. doi: 10.1093/biomet/88.4.1055.

[31] D. S. Ebert, 2013. URL https://engineering.purdue.edu/~ebertd/cloud/cloud_big.jpg.

[32] S. Efromovich and M. S. Pinsker. An adaptive algorithm of nonparametric filtering. *Automat. Remote Control*, 11:1434–1440, 1984.

[33] T. S. Ferguson. A Bayesian analysis of some nonparametric problems. *Ann. Statist.*, 1:209–230, 1973.

[34] A. Gelman, W. R. Gilks, and G. O. Roberts. Weak convergence and optimal scaling of random walk Metropolis algorithms. *Ann. Appl. Probab.*, 7(1):110–120, 1997.

[35] S. Ghosal. The Dirichlet process, related priors and posterior asymptotics. In *Bayesian nonparametrics*, Camb. Ser. Stat. Probab. Math., pages 35–79. Cambridge Univ. Press, Cambridge, 2010.

[36] S. Ghosal and A. van der Vaart. Convergence rates of posterior distributions for non-i.i.d. observations. *Ann. Statist.*, 35(1):192–223, 2007. ISSN 0090-5364. doi: 10.1214/009053606000001172.

[37] S. Ghosal and A. W. van der Vaart. Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. *Ann. Statist.*, 29(5):1233–1263, 2001. ISSN 0090-5364. doi: 10.1214/aos/1013203453.

[38] S. Ghosal, J. K. Ghosh, and A. W. van der Vaart. Convergence rates of posterior distributions. *Ann. Statist.*, 28(2):500–531, 2000. ISSN 0090-5364. doi: 10.1214/aos/1016218228.

[39] S. Ghosal, J. Lember, and A. van der Vaart. Nonparametric Bayesian model selection and averaging. *Electron. J. Stat.*, 2:63–89, 2008. ISSN 1935-7524. doi: 10.1214/07-EJS090.

[40] P. J. Green. Reversible jump Markov Chain Monte Carlo computation and bayesian model determination. *Biometrika*, 82(4):711–32, 1995.

[41] P. J. Green and D. I. Hastie. Model choice using reversible jump Markov Chain Monte Carlo. *Statistica Neerlandica*, 66:309–338, 2012.

[42] W. K. Hastings. Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, 21(6):1087–1092, 1970.

[43] R. Helmers and I. W. Mangku. On estimating the period of a cyclic Poisson process. In *Mathematical statistics and applications: Festschrift for Constance van Eeden*, pages 345–356, 2003.

[44] R. Helmers and I. W. Mangku. Estimating the intensity of a cyclic Poisson process in the presence of linear trend. *Ann. Inst. Statist. Math.*, 61(3):599–628, 2009.

[45] R. Helmers, I. W. Mangku, and R. Zitikis. Statistical properties of a kernel-type estimator of the intensity function of a cyclic Poisson process. *J. Multivariate Anal.*, 92(1):1–23, 2005.

[46] R. Helmers, Q. Wang, and R. Zitikis. Confidence regions for the intensity function of a cyclic Poisson process. *Stat. Inference Stoch. Process.*, 12(1):21–36, 2009.

[47] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, 1988.

[48] P. J. Huber. *Robust Statistics*. Wiley series in probability and mathematical statistics. Wiley, 1981.

[49] I. A. Ibragimov and R. Z. Has'minskii. *Statistical estimation : asymptotic theory*. Number 16 in Applications of mathematics. Springer-Verlag, New York, 1981.

[50] J. Jacod and A. N. Shiryaev. *Limit theorems for stochastic processes*, volume 288 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, second edition, 2003. ISBN 3-540-43932-3.

[51] M. E. Johnson. *Multivariate Statistical Simulation*. Wiley, New York, 1987.

[52] I. Karatzas and S. Shreve. *Brownian Motion and Stochastic Calculus*, volume 113 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1991.

[53] J. Kiefer and J. Wolfowitz. Stochastic estimation of the maximum of a regression function. *The Annals of Mathematical Statistics*, 23(3):462–466, 1952.

[54] J. F. C. Kingman. *Poisson Processes*, volume 3 of *Oxford Studies in Probability*. Oxford University Press, 1992.

[55] R. Koenker. *Quantile Regression*. Cambridge University Press, 2005.

[56] A. Kottas and B. Sansó. Bayesian mixture modeling for spatial Poisson process intensities, with applications to extreme value analysis. *J. Statist. Plann. Inference*, 137(10):3151–3163, 2007. ISSN 0378-3758. doi: 10.1016/j.jspi.2006.05.022.

[57] U. Krengel. *Ergodic Theorem*. Number 6 in Studies in Math. de Gruyter, 1985.

[58] H. J. Kushner. Stochastic approximation: a survey. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2:87–96, 2010.

[59] H. J. Kushner and D. S. Clark. *Stochastic Approximation for Constrained and Unconstrained Systems.* Springer Verlag, New York, 1978.

[60] H. J. Kushner and J. Yang. Analysis of adaptive step-size sa algorithms for parameter tracking. *IEEE Trans. Autom. Control*, 40:1403–1410, 1995.

[61] H. J Kushner and G. Yin. *Stochastic Approximation and Recursive Algorithms and Applications.* Berlin and New York: Springer-Verlag, 2003.

[62] Y. A. Kutoyants. On nonparametric estimation of intensity function of inhomogeneous Poisson process. *Problems of Control and Information Theory*, 13(4):253–258, 1984.

[63] Y. A. Kutoyants. *Statistical Inference for Spatial Poisson Processes, Lecture Notes in Statistics* **134**. Springer-Verlag, New York, 1998.

[64] T. L. Lai. Stochastic approximation: invited paper. *Ann. Statist.*, 31(2):391–406, 2003.

[65] L. M. le Cam and G. L. Yang. *Asymptotics in Statistics: Some Basic Concepts*. Springer Series in Statistics. Springer-Verlag, New York, 1990.

[66] L. Ljung and T. Söderström. *Theory and Practice of Recursive Identification.* MIT Press, Cambridge, MA., 1983.

[67] I. W. Mangku. *Estimating the intensity function of a cyclic Poisson process*. PhD thesis, University of Amsterdam, Amsterdam, 2001.

[68] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1953.

[69] J. Møller and R. P. Waagepetersen. *Statistical Inference and Simulation for Spatial Point Processes*. Chapman & Hall / CRC, Boca Raton, 2004.

[70] J. Møller, A. R. Syversveen, and R. P. Waagepetersen. Log Gaussian Cox processes. *Scand. J. Statist.*, 25(3): 451–482, 1998. ISSN 0303-6898. doi: 10.1111/1467-9469.00115.

[71] E. Moulines, P. Priouret, and F. Roueff. On recursive estimation for time varying autoregressive processes. *Ann. Statist.*, 33(6):2610–2654, 2005.

[72] M.B. Nevelson and R. Z. Khasminskii. *Stochastic Approximation and Recursive Estimation.*, volume 47 of *Translation of Mathematical Monographs*. American American Mathematical Society, 1976.

[73] J. A. Palacios and V. N. Minin. Gaussian process-based Bayesian nonparametric inference of population size trajectories from gene genealogies. *Biometrics*, 2013. ISSN 1541-0420. doi: 10.1111/biom.12003. to appear.

[74] L. Panzar and H. van Zanten. Nonparametric Bayesian inference for ergodic diffusions. *J. Statist. Plann. Inference*, 139(12):4193–4199, 2009. ISSN 0378-3758. doi: 10.1016/j.jspi.2009.06.003.

[75] M. S. Pinsker. Optimal filtering of square integrable signals in gaussian white noise. *Probl. Peredachi Inf.*, 16(2):52–68, 1980.

[76] H. Robbins and S. Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22 (3):400–407, 1951.

[77] C. P. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer Texts in Statistics. Springer-Verlag, New York, 2000.

[78] C. P. Robert and G. Casella. *Introducing Monte Carlo Introducing Monte Carlo Methods with R*. Springer-Verlag, New York, 2009.

[79] C. B. Rorabaugh. *DSP Primer*. McGraw-, 1998.

[80] S. M. Ross. *Simulation*. Academic Press, fourth edition, 2006.

[81] L. L. Schumaker. *Spline functions: basic theory*. Cambridge Mathematical Library. Cambridge University Press, Cambridge, third edition, 2007. ISBN 978-0-521-70512-7. doi: 10.1017/CBO9780511618994.

[82] L. Schwartz. On Bayes procedures. *Z. Wahrsch. Verw. Gebiete*, 4(1):10–26, 1965.

[83] P. Serra, 2013. URL http://www.youtube.com/watch?v=4uu3LSUbUzA.

[84] P. Serra, 2013. URL http://www.youtube.com/watch?v=3m0zHKiPLns.

[85] E. Sharef, R. Strawderman, D. Ruppert, M. Cowen, and L. Halasyamani. Bayesian adaptive B-spline estimation in proportional hazards frailty models. *Electron. J. Stat.*, 4:606–642, 2010. ISSN 1935-7524. doi: 10.1214/10-EJS566.

[86] W. Shen and S. Ghosal. MCMC-free adaptive Bayesian procedures using random series prior. *arXiv:1204.4238 [math.ST]*, 2012.

[87] A. N. Shiryaev. *Probability*. Springer-Verlag, New York, second edition, 1996.

[88] M. Smith and R. Kohn. Nonparametric regression using Bayesian variable selection. *Journal of Econometrics*, 75(2):317–343, 1996.

[89] D. L. Snyder. *Random Point Processes*. Wiley, 1975.

[90] C. J. Stone. Optimal rates of convergence for nonparametric estimators. *Ann. Statist.*, 8(6):1348–1360, 1980.

[91] R. L. Streit. *Poisson Point Processes: Imaging, Tracking, and Sensing*. Springer-Verlag, New York, 2010.

[92] I. Takeuchi, Q. V. Le, T. D. Sears, and A. J. Smola. Nonparametric quantile estimation. *J. Mach. Learn. Res.*, 7:1231–1264, 2006. ISSN 1532-4435.

[93] YaZ. Tsypkin. *Adaptation and Learning in Automatic Systems*. Academic Press, New York, 1971.

[94] F. H. van der Meulen, A. W. van der Vaart, and J. H. van Zanten. Convergence rates of posterior distributions for Brownian semimartingale models. *Bernoulli*, 12(5):863–888, 2006. ISSN 1350-7265. doi: 10.3150/bj/1161614950.

[95] A. van der Vaart and H. van Zanten. Rates of contraction of posterior distributions based on Gaussian process priors. *Ann. Statist.*, 36(3):1435–1463, 2008.

[96] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge series in statistical and probabilistic Mathematics. Cambridge University Press, 1998.

[97] A. W. van der Vaart and J. H. van Zanten. Adaptive Bayesian estimation using a Gaussian random field with inverse gamma bandwidth. *Ann. Statist.*, 37(5B):2655–2675, 2009.

[98] A. W. van der Vaart and J. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer-Verlag, New York, 1996.

[99] D. Vere-Jones. On estimation of frequency in point-process data. *J. Appl. Prob.*, 19(A):383–394, 1982.

[100] P. Walters. *An Introduction to Ergodic Theory*. Number 79 in Graduate Texts in Mathematics. Springer-Verlag, New York, 1982.

[101] M. T. Wasan. *Stochastic Approximation*. Cambridge University Press, 1969.

[102] F. T. Wright. The asymptotic behavior of monotone regression estimates. *Ann. Statist.*, 9(2):443–448, 1981.

Subject Index

Curriculum Vitæ

Paulo Jorge de Andrade Serra was born on the 22nd of April, 1981 in Lisbon, Portugal. After finishing the Natural Sciences program in 1999 at the Secondary School of Mem-Martins in Mem-Martins, Sintra, he went on to study Mathematics at the Faculty of Science and Technology of the New University of Lisbon. In 2005 he completed a five year degree in Applied Mathematics. Between March 2005 and July 2006 he was employed by the CA3 group of the UNINOVA research institute at Costa da Caparica, Setúbal, and worked on a project commissioned and financed by the European Space Agency. This gave him an opportunity to engage in research for the first time and present his work at a few international conferences. In August 2006 he moved to Utrecht, the Netherlands to start his Master studies in Mathematical Sciences at Utrecht University which he completed *cum laude* in 2008. In 2009 he started his PhD studies at the Utrecht University, Utrecht and the research center EURANDOM, Eindhoven, under the supervision of prof. dr. Harry van Zanten and dr. Eduard Belitser. The following year he moved full-time to the Department of Mathematics and Computer Science of the Technical University of Eindhoven, in Eindhoven, the Netherlands. In September 2013 he completed his PhD in Mathematical Statistics at this university. The results obtained during these four years are presented in this dissertation.

Colophon

This thesis was typeset using LaTeX, originally developed by Leslie Lamport and based on Donald Knuth's TeX. The body text is set in 11 point Minion Pro, designed by Robert Slimbach in the late Renaissance-era type style, and issued by Adobe in 1990. This is a modified version of a template which can be found online at github.com/suchow/ or from its author at suchow@post.harvard.edu and which was released under the permissive MIT (X11) license. The front cover features a collage made by the author of this thesis using illustrations from *Abbildung Der Hundert Deutschen Wilden Holz-Arten Nach Dem Nummern-Verzeichnis Im Forst-Handbuch Von F. A. L. Von Burgsdorf, 1790–1794* by Johann Daniel Reitter and Gottlieb Friedrich Abel. The egg illustration on the back cover was obtained from *OEUFS*, illustration by Adolphe Millot from *Nouveau Larousse Illustré, 1897–1904*. All the images are in the public domain and were obtained from the Wikimedia Commons commons.wikimedia.org/. The code for the barcode on the back cover was obtained from tex.stackexchange.com/. This thesis was printed by Wöhrmann Print Services, Zutphen, the Netherlands.

ISBN 978-90-386-3439-5