Adaptive posterior contraction rates for diffusions
van Waaij, J.

Citation for published version (APA):
van Waaij, J. (2018). Adaptive posterior contraction rates for diffusions
ADAPTIVE POSTERIOR CONTRACTION RATES FOR DIFFUSIONS

Diffusions have many applications in science and can be described with a stochastic differential equation (SDE). We consider the following SDE, which was for example used in molecular dynamics (see e.g. Papaspiliopoulos et al. (2012)),

\[ dX_t = \theta(X_t)dt + dW_t \]

where \( \theta \) is measurable, one-periodic and \( \int_0^1 \theta(x)^2 dx < 1 \). We are interested in estimating \( \theta \) from observations \( (X_t : t \in [0, T]) \) of eq. (\(^\star\)). We study the posterior rates of contraction for several nonparametric Bayesian methods for diffusions. For Gaussian process priors we derive optimal posterior contraction rates, when the smoothness of the Gaussian process coincides with the smoothness of the target drift function. Adaptivity to the unknown smoothness is achieved by random scaling of the Gaussian process prior, or by equipping the baseline smoothness hyperparameter with a hyperprior.

We derive good adaptive posterior contraction results for priors defined as randomly truncated series priors. We consider expansions in orthonormal bases and in the Faber-Schauder basis, both with inverse gamma scaling. We also study the empirical Bayes approach to selecting the scaling parameter of the Gaussian process prior. Here the parameter is estimated from the data and plugged into the prior. Adaptive optimal contraction rates for the associated posterior are derived.

ISBN: 978-94-028-0883-4
http://hdl.handle.net/11245.1/6b890997-2565-4718-9b0f-71975568d9
Adaptive posterior contraction rates for diffusions

Jan van Waaij
Adaptive posterior contraction rates for diffusions
Promotiecommissie

Promotor:
prof. dr. J.H. van Zanten
*Universiteit van Amsterdam*

Overige leden:
prof. dr. I. Castillo
*Université Pierre et Marie Curie, Paris*

prof. dr. C.A.J. Klaassen
*Universiteit van Amsterdam*

prof. dr. P.J.C. Spreij
*Universiteit van Amsterdam*

prof. dr. A.W. van der Vaart
*Universiteit Leiden*

dr. S.G. Cox
*Universiteit van Amsterdam*

dr. F.H. van der Meulen
*Technische Universiteit Delft*

Faculteit der Natuurwetenschappen, Wiskunde en Informatica

Het hier beschreven onderzoek werd mede mogelijk gemaakt door steun van de Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO).
Voor mijn familie.
Contents

List of notation xi

Preface xiii

1 Introduction 1
  1.1 Diffusions on the line .................. 1
  1.2 Stochastic differential equations ........ 2
  1.3 Periodic drift .......................... 4
  1.4 Statistical inference for diffusions ....... 6
  1.5 Bayesian nonparametric inference for periodic diffusion models .......................... 7

2 Gaussian process methods for one-dimensional diffusions 13
  2.1 Definition of the prior .................. 16
  2.2 Main results ............................. 17
    2.2.1 Fixed hyperparameters ............... 18
    2.2.2 Prior on the scale .................. 18
    2.2.3 Prior on the Gaussian process regularity .... 20
  2.3 Auxiliary results ........................ 21
    2.3.1 Small ball probabilities .......... 21
  2.4 Proof of theorem 2.1 .................... 24
  2.5 Proof of theorem 2.2 .................... 24
    2.5.1 Prior mass condition ............... 24
    2.5.2 Sieves ............................. 25
  2.6 Proof of theorem 2.3 .................... 26
    2.6.1 Prior mass condition ............... 26
    2.6.2 Remaining mass and entropy .......... 27
## CONTENTS

### 3 Randomly truncated orthonormal series prior
- **3.1 Introduction** ........................................ 29
- **3.2 Prior model** .......................................... 32
- **3.3 Main results** .......................................... 33
- **3.4 Specific statistical settings** ........................... 35
  - 3.4.1 Detecting a signal in Gaussian white noise .... 35
  - 3.4.2 Estimating the drift of periodic diffusion .... 36
- **3.5 Proof of theorem 3.1** .................................. 36
  - 3.5.1 Prior mass ............................................ 36
  - 3.5.2 Sieves, remaining mass and entropy .......... 40
- **3.6 Proof of theorem 3.2** .................................. 40

### 4 Drift estimation using Faber-Schauder expansions
- **4.1 Introduction** ........................................ 43
- **4.2 Prior construction** .................................... 46
- **4.3 Theorems on posterior contraction rates** ........... 47
- **4.4 Applications to nonparametric regression and density estimation** .......................... 49
  - 4.4.1 Nonparametric regression model ................. 50
  - 4.4.2 Nonparametric density estimation ............... 50
- **4.5 Proofs** ................................................ 51
  - 4.5.1 Proof of theorem 4.4 ............................... 51
  - 4.5.2 Lemmas ............................................... 56

### 5 Empirical Bayes selection of the scaling parameter
- **5.1 Main result** ........................................ 59
- **5.2 The estimator is with high probability in a favourable set** ................................ 65
- **5.3 Posterior convergence on the favourable event** ........................................ 68
- **5.4 Auxiliary results** .................................... 69

---

**Dankwoord** ........................................ 75

**Discussion** ............................................ 77

**Summary** ............................................... 79

**Samenvatting** ......................................... 81

**Overzicht van de bijdragen van de co-auteurs** ........ 83
<table>
<thead>
<tr>
<th>CONTENTS</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Bibliography</td>
<td>85</td>
</tr>
<tr>
<td>Index</td>
<td>91</td>
</tr>
</tbody>
</table>
List of notation

\[ \mathbb{E} X \]  The expectation of \( X \).
\[ \text{Var} X \]  The variance of \( X \).
\( \mathbb{N} \)  The natural numbers \( 1, 2, \ldots \)
\( \mathbb{R} \)  The real numbers.
\( \mathbb{R}^+ \)  The nonnegative real numbers.
\( \mathbb{R}_{>0} \)  The positive real numbers.
\( \mathbb{Z} \)  The integers \( \ldots, -2, -1, 0, 1, 2, \ldots \)
\( \mathbb{Z}_{\geq 0} \)  The nonnegative integers \( 0, 1, 2, \ldots \)
\[ \| \theta \|_2 \]  Square integrable norm of \( \theta \).
\[ \| \theta \|_\infty \]  Supremum norm of \( \theta \).
\[ a_n \lesssim b_n \]  For some \( N \in \mathbb{N} \) and a constant \( C > 0 \), \( a_n \leq C b_n \) for all \( n \geq N \).
\[ a_n \gtrsim b_n \]  Equivalent to \( b_n \lesssim a_n \).
\[ a_n \precsim b_n \]  Equivalent to \( a_n \lesssim b_n \) and \( b_n \lesssim a_n \).
\( \mathbb{I}_A \)  Indicator function of a set \( A \).
\( \text{IG}(A, B) \)  Inverse gamma distribution with shape parameter \( A > 0 \) and scale parameter \( B > 0 \).
\( N(\mu, \sigma^2) \)  Normal distribution with mean \( \mu \) and variance \( \sigma^2 \).
a.s.  Abbreviation of almost surely.
Preface

Diffusions have many applications in science and mathematical finance and are well studied in the literature. However, the study of nonparametric Bayesian methods for diffusions has only recently begun. In this thesis we add a contribution to this field. We study the asymptotic behaviour of nonparametric Bayesian methods for diffusions, in terms of posterior convergence. Chapter 1 is an introduction to the subject and its most important results.

A first numerical implementation for a nonparametric Gaussian process prior for diffusions on an interval is given in Papaspiliopoulos et al. (2012). It provides a method to sample from the posterior distribution, using a finite element method for differential equations. Consistency for this prior is shown in Pokern, Stuart, and van Zanten (2013). In chapter 2 we investigate whether we can improve on this result in terms of optimal posterior contraction rates via a different mathematical route, by checking the sufficient conditions for posterior convergence of van der Meulen, van der Vaart, and van Zanten (2006). We also investigate the method of putting a hyperprior on the hyperparameters of the prior to obtain adaptivity to the unknown (Sobolev-)smoothness of the drift.

In van der Meulen, Schauer, and van Zanten (2014) a reversible jump MCMC sample scheme is provided to sample from the posterior of a nonparametric prior with observations of a diffusion. The prior is defined as a random function, expanded in some basis, where coefficients are normally distributed, the expansion is randomly truncated and scaled with an inverse gamma distribution. They not only allow for Fourier basis functions, but for general orthonormal bases and the Faber-Schauder basis. In chapter 3 we investigate the asymptotic behaviour of the posterior of this prior with an orthonormal basis, and in chapter 4 we investigate the asymptotic behaviour of this prior with
the Faber-Schauder basis.

The empirical Bayes methodology is another interesting approach that recently garnered a great deal of attention in the literature (see e.g. Donnet et al. (2018), Petrone, Rousseau, and Scricciolo (2014), Rousseau and Szabó (2016), and Rousseau and Szabó (2017) and reference therein). In this methodology the optimal hyperparameters of the prior are estimated in a frequentist manner, and then plugged into the prior, whose posterior is then used for the inference. General conditions for posterior convergence in the nonparametric empirical Bayes setting are given in Donnet et al. (2018) and Rousseau and Szabó (2017) for several models, but these do not apply directly to diffusion models. In chapter 5 we investigate empirical Bayes selection of the scaling parameter of a Gaussian process prior and we study the contraction rates of the corresponding posterior.

This thesis is based on the following publications.


Chapter 5 is joint work with J.H. van Zanten, and a paper based on this chapter is in preparation. A discussion with suggestions for future work and a summary are given at the end of this thesis.
1 Introduction

1.1 Diffusions on the line

In this section we recall some general results about diffusions that we need in this thesis. For a more background information we refer to Kallenberg (2002), Karatzas and Shreve (1999), Revuz and Yor (1999), and Rogers and Williams (1987).

Diffusion processes on the line are defined as continuous strong Markov processes $X = (X_t)_{t \in \mathbb{R}^+}$ taking values in the one-dimensional Euclidean space $\mathbb{R}$. For simplicity we assume that $X_t$ is defined for all $t \in \mathbb{R}^+$. For a more general definition of a diffusion where termination is allowed or for diffusions only defined on a proper subset of $\mathbb{R}$ (with some boundary behaviour) we refer to Kallenberg (2002, chapter 23), as we do not consider them in this thesis. The law of the process under which $X$ starts at $x \in \mathbb{R}$ is denoted by $\mathbb{P}^x$. We call $t$ the time and $X_t$ the state of the process at time $t$. A basic example of a diffusion process is a Brownian motion.

Define $\tau_y$ as the first hitting time of $y$. A diffusion is regular when for all $x, y \in \mathbb{R}$ the process $X$ starting in $x$ hits $y$ in finite time with positive probability. Let $\tau_{a,b}$ be the first time that $X$ hits either $a$ or $b$. For $a < x < b$ this is the first time that $X$ starting at $x$ leaves the inter-
val \((a, b)\). The probability \(\mathbb{P}^x(\tau_b < \tau_a) = \mathbb{P}^x(\tau_{a,b} = \tau_b)\) is the probability of the process hitting \(b\) before \(a\), or to say it differently, the probability of leaving the interval \((a, b)\) via \(b\). The probabilities \(\mathbb{P}^x(\tau_b < \tau_a)\) and \(\mathbb{P}^x(\tau_a < \tau_b)\) are referred to as the exit probabilities. For a Brownian motion it is well known that \(\mathbb{P}^x(\tau_b < \tau_a) = \frac{x - a}{b - a}\) for \(a < x < b\).

A diffusion which has the same exit probabilities as a Brownian motion is said to be in natural scale. For every regular diffusion there is a transformation so that the transformed process is in natural scale. More precisely, there exists a continuous increasing function \(s\), which is unique up to affine transformations, so that for all \(x\), under \(\mathbb{P}^x\), \(s(X_t)\) is a diffusion (starting in \(s(x)\)) and a local martingale. For the process \(X_t\) we have \(\mathbb{P}^x(\tau_b < \tau_a) = \frac{s(x) - s(a)}{s(b) - s(a)}\). In particular the process \(t \mapsto s(X_t)\) is in natural scale and a process \(X\) is in natural scale if and only if \(s\) is an affine transformation. We say that \(s\) is a scale function of the process \(X\).

For an interval \(I = (a, b)\) and \(x \in I\) define \(m_I(x) = \mathbb{E}^x \tau_{a,b}\). There is a unique measure \(M\) on \(\mathbb{R}\) so that \(m_I(x) = \int_I G_I(x, y) dM(y)\), where \(G_I\) is Green’s function

\[
G_I(x, y) = \frac{(s(x \wedge y) - s(a))(s(b) - s(x \vee y))}{s(b) - s(a)}.
\]

As \(s(X_t)\) is a continuous local martingale, it can be written as a time changed Brownian motion \(W\), \(s(X_t) = W_{\sigma_t}\). Let \(\ell : \mathbb{R}^+ \times \mathbb{R} \rightarrow \mathbb{R}^+\) be the local time of \(W\). That is, \(\ell\) is jointly continuous in its variables and satisfies a.s.

\[
\ell(t, x) = \lim_{\varepsilon \downarrow 0} \frac{1}{2\varepsilon} \text{meas} \{s \in [0, t] : X_s \in [x - \varepsilon, x + \varepsilon]\}
\]

, where \text{meas} is the Lebesgue measure. The time change \(\sigma\) can be described as the right-continuous inverse of the process \(A\) defined by \(A_t = \int \ell(t, s(y)) M(dy)\). Moreover \(l(t, x) = \ell(\sigma_t, s(x))\) is the local time of \(X\) with respect to \(M\). For every bounded measurable function \(f : \mathbb{R} \rightarrow \mathbb{R}\) one has the occupation times formula \(\int_0^t f(X_t) dt = \int f(x) l(t, x) dM(x)\).

### 1.2 Stochastic differential equations

Let \(\sigma : \mathbb{R} \rightarrow (0, \infty)\) and \(\theta : \mathbb{R} \rightarrow \mathbb{R}\) be measurable functions and \(W\) a Brownian motion and \(x \in \mathbb{R}\). Under mild regularity assumptions
Kallenberg (2002, Theorems 19.24 & 21.11) a diffusion can be described as a weak solution to the stochastic differential equation (SDE)

\[
\begin{aligned}
\begin{cases}
    dX_t &= \theta(X_t)dt + \sigma(X_t)dW_t, \\
    X_0 &= x
\end{cases}
\end{aligned}
\] (1.1)

in the sense that \(X_t\) is a sum of the initial value \(x\), a Lebesgue integral and an Itô integral,

\[X_t = x + \int_0^t \theta(X_s)ds + \int_0^t \sigma(X_s)dW_s. \]

We call \(\theta\) and \(\sigma\) the drift and diffusion of \(X\) respectively. This gives us two tools to study diffusions, by viewing them as continuous strong Markov processes or as semi-martingales. An SDE (1.1) is said to have a weak solution up to an explosion time \(S\) when there is a filtered probability space \((\Omega, \mathcal{F}, (\mathcal{F}_t)_{t \geq 0}, \mathbb{P})\), with a Brownian motion \(W = (W_t)_{t \geq 0}\) under \((\mathcal{F}_t)_{t \geq 0}\) and a continuous real-valued process \(X = (X_t)_{t \geq 0}\) adapted to the filtration \((\mathcal{F}_t)_{t \geq 0}\), \(X_0 = x\) a.s. and if we define \(S_n = \inf \{t \geq 0 : |X_t| \geq n\}\), then \(X_t\) satisfies for all \(t \geq 0\) and \(n \in \mathbb{N}\),

\[
\int_0^t \left( |\theta(X_s)| + \sigma^2(X_s) \right) I_{\{s \leq S_n \}} ds < \infty \text{ a.s.} \tag{1.2}
\]

and

\[
X_{t \wedge S_n} = x + \int_0^t \theta(X_s) I_{\{s \leq S_n \}} ds + \int_0^t \sigma(X_s) I_{\{s \leq S_n \}} dW_s. \tag{1.3}
\]

We refer to the stopping time \(S = \lim_{n \to \infty} S_n\) as the explosion time of the process \(X\). When \(S < \infty\) on an event of positive measure, then \(|X_t|\) converges with positive probability to \(\infty\) in finite time. When \(S = \infty\) a.s., then \(X_t\) is a.s. defined for every \(t \geq 0\) and we call \(X\) simply a weak solution of eq. (1.1).

A weak solution \(X\) to eq. (1.1) is weakly unique or unique in law when for any other weak solution \(\tilde{X}\) to eq. (1.1) possibly defined on a different probability space or a different filtration with a different Brownian motion, we have that \(X\) and \(\tilde{X}\) have the same finite dimensional distributions. For the notions of strong existence, strong uniqueness and pathwise uniqueness we refer to the literature. Under mild conditions on \(\theta\) and \(\sigma\), there exists a unique weak solution to eq. (1.1),
necessary and sufficient conditions are given in Karatzas and Shreve (1999, §5.5).

It follows easily from the Itô formula that a diffusion \( X_s \) satisfying eq. (1.1) has a scale function

\[
s(x) = \int_0^x \exp \left\{ -2 \int_0^y \frac{\theta(z)}{\sigma^2(z)} \, dz \right\} \, dy
\]

and speed measure

\[
dM(x) = \frac{dx}{s'(x)s^2(x)}.
\]

In particular \( M \) has a density with respect to the Lebesgue measure and we denote its density by \( m \), and we refer to \( m \) as the speed density. When \( \int_0^1 \theta(x) \, dx = 0 \), the speed density is periodic.

### 1.3 Periodic drift

Let us now consider a special case of eq. (1.1) with \( \sigma \equiv 1 \) and the drift function \( \theta : \mathbb{R} \to \mathbb{R} \) is one-periodic and \( \int_0^1 \theta(x)^2 \, dx < \infty \). Hence eq. (1.1) is given by

\[
\begin{cases}
  dX_t = \theta(X_t) \, dt + dW_t, \\
  X_0 = x.
\end{cases}
\]

The set of drift functions satisfying the conditions above is denoted by \( L^2(\mathbb{T}) \). Alternatively, one could see them as the square integrable functions on the circle \( \mathbb{T} \). The \( L^2 \)- or square integrable norm of \( \theta \in L^2(\mathbb{T}) \) is defined by

\[
||\theta||_2 = \sqrt{\int_0^1 \theta(x)^2 \, dx}.
\]

The closed subspace of \( \theta \in L^2(\mathbb{T}) \) where \( \int_0^1 \theta(x) \, dx = 0 \) is denoted by \( \dot{L}^2(\mathbb{T}) \).

**Lemma 1.1.** For every \( \theta \in L^2(\mathbb{T}) \) and initial value \( x \in \mathbb{R} \) the SDE (1.5) has a unique weak solution.
1.3. Periodic drift

Proof. It follows from Karatzas and Shreve (1999, theorem 5.15) that the SDE (1.5) has a unique weak solution up to an explosion time. Define $\tau_0 = 0$ and for $i \geq 1$ the random times $\tau_i = \inf\{t \geq \tau_{i-1}: |X_t - X_{\tau_{i-1}}| = 1\}$. By periodicity of drift and the strong Markov property the random variables $U_i = \tau_i - \tau_{i-1}$ are independent and identically distributed. Note that

$$\inf\{t: X_t - X_0 = \pm n\} \geq \sum_{i=1}^n U_i$$

and hence non-explosion follows from $\lim_{n \to \infty} \sum_{i=1}^n U_i = \infty$ almost surely. The latter holds true since $U_1 > 0$ with probability one, which is clear from the continuity of the diffusion paths.

Let $l$ be the local time of $X$ with respect to the speed measure. Obviously the local time is not periodic in $x$, but for fixed $t$, $l(t, \cdot)$ has a.s. its support on the range of $s \mapsto X_s, s \leq t$, which is a.s. compact. Thus the sums $l(t, x) := \sum_{k=-\infty}^\infty l(t, x + k)$ and $\hat{L}(t, x) := \sum_{k=-\infty}^\infty l(t, x + k)m(x + k)$ have a.s. only finitely many nonzero summands, and are therefore a.s. well defined. It follows that for a bounded measurable one-periodic function $f : \mathbb{R} \to \mathbb{R}$, the following occupation times formula is valid a.s.

$$\int_0^t f(X_t)dt = \int_{-\infty}^\infty f(x)l(t, x)m(x)dx = \int_0^1 f(x)\hat{L}(t, x)dx. \quad (1.6)$$

The process $X_T := (X_t : t \in [0, T])$ induces a distribution $\mathbb{P}_\theta$ on $C[0, T]$, the space of real-valued continuous functions on the interval $[0, T]$. As the process starts in $x$ it gives full mass to the measurable set of functions starting in $x$. We suppress the dependence of $\mathbb{P}_\theta$ on $x$ and $T$ in the notation. When $\theta \equiv 0$, then $\mathbb{P}_0$ denotes the Wiener measure on $C[0, T]$.

Lemma 1.2. For every fixed $x \in \mathbb{R}$ and fixed $T > 0$ and for all $\theta_1, \theta_2 \in L^2(T)$, $\mathbb{P}_{\theta_1}$ and $\mathbb{P}_{\theta_2}$ are equivalent. Moreover, $\mathbb{P}_\theta$ has density

$$p_\theta(X_T) = \exp \left\{ \int_0^T \theta(X_t) dX_t - \frac{1}{2} \int_0^T \theta(X_t)^2 dt \right\} \quad (1.7)$$
1. Introduction

relative to the Wiener measure on $C[0,T]$ and $\mathbb{P}_{\theta_1}$ has the following
density with respect to $\mathbb{P}_{\theta_2}$,

$$
\frac{d\mathbb{P}_{\theta_1}}{d\mathbb{P}_{\theta_2}}(X^T) = \exp\left\{ \int_0^T (\theta_1(X_t) - \theta_2(X_t))dW_t
- \frac{1}{2} \int_0^T (\theta_1(X_t) - \theta_2(X_t))^2 dt \right\}, \tag{1.8}
$$

with $W$ a $\mathbb{P}_{\theta_2}$-Brownian motion.

Proof. Let $t > 0$ and $\theta \in L^2(\mathbb{T})$. As $\dot{L}_t$ is continuous and $[0,1]$ is
compact, $x \mapsto \dot{L}_t(x)$ is bounded a.s. with respect to $\mathbb{P}_0$ and with respect
to $\mathbb{P}_\theta$. It follows that

$$
\int_0^t \theta^2(X_s)ds = \int_0^1 \theta^2(x)\dot{L}(t,x)dx < \infty
$$
a.s. with respect to $\mathbb{P}_0$ and $\mathbb{P}_\theta$. It follows from Jacod and Shiryaev
(2002, Theorem III.5.38c, page 201) that all measures $\mathbb{P}_\theta, \theta \in L^2(\mathbb{T})$
are equivalent and have densities eqs. (1.7) and (1.8).

In the case that $\int_0^1 \theta(x)dx = 0$, $m$ is one-periodic and we simply
have $\dot{L}(t,x) = \dot{l}(t,x)m(x)$. We have the following law of large numbers
result for the periodic local time, see Pokern, Stuart, and van Zanten
(2013, Theorem 4.1):

**Lemma 1.3.** Let $\rho$ be the speed density normalised to be a probability
density on $[0,1]$. For $\theta_0 \in \dot{L}^2(\mathbb{T})$ it holds $\mathbb{P}_{\theta_0}$-almost surely that

$$
\sup_{x \in [0,1]} \left| \frac{1}{T} \dot{L}(T,x) - \rho(x) \right| \to 0,
$$
as $T \to \infty$.

Recent unpublised work by M.R. Schauer suggests that when for
$\theta \in L^2(\mathbb{T})$, $\int_0^1 \theta(x)dx \neq 0$ lemma 1.3 holds as well.

1.4 Statistical inference for diffusions

Diffusions are a natural model for many phenomena in nature and fi-
nance. Examples range from molecular dynamics (e.g. Papaspiliopou-
llos et al. (2012)), climate research (e.g. Ditlevsen, Ditlevsen, and An-
dersen (2002)) and neurobiology (e.g. Hindriks (2011)) to finance (e.g.
Karatzas and Shreve (1998)). In this type of models, the data is assumed to be a (discretely sampled) process which satisfies a certain SDE, for instance (1.1). The model eq. (1.5) with periodic drift we study in this thesis is motivated by measuring angles between atoms of a molecule, see Papaspiliopoulos et al. (2012).

We speak about continuous observations when the whole sample path $X^T = (X_t : t \in [0, T])$ of eq. (1.1) up to a time $T > 0$ is observed. In this case $\sigma$ can be estimated from the data without error with probability 1 on the range of $X^T$, so that we usually assume that $\sigma$ is known. After rescaling one can as well assume that $\sigma \equiv 1$. Estimating the drift from $X^T$ is however a nontrivial task. As the drift is a function, it is most natural to assume a nonparametric model for the drift. The problem of nonparametric estimation of the drift with continuous observations in the frequentist setting is studied by multiple authors. For an overview of the parametric and the nonparametric frequentist estimation of the drift parameter in the continuous observation setting, one consults Kutoyants (2004).

Bayesian (nonparametric) methods are often applied for their conceptual simplicity, ease of implementation and numerical advantages. In Bayesian statistics one equips the parameter space with a probability distribution and inference is done via the posterior, the distribution of the parameter given the data. In the nonparametric setting, numerical schemes to sample from the posterior are developed in the papers Papaspiliopoulos et al. (2012) and van der Meulen, Schauer, and van Zanten (2014). Asymptotic properties of Bayesian methods for nonparametric diffusion models are studied in van der Meulen, van der Vaart, and van Zanten (2006) in a general setting and for ergodic diffusion in Panzar and van Zanten (2009). In this thesis we study posterior contraction rates for several nonparametric priors.

1.5 Bayesian nonparametric inference for periodic diffusion models

We assume continuous observations $X^T = (X_t : t \in [0, T]), T > 0$ satisfying eq. (1.5) with $\theta = \theta_0$ the unknown parameter. We equip the parameter space $L^2(\mathbb{T})$ (or $\dot{L}^2(\mathbb{T})$) with the Borel-$\sigma$-algebra induced by the $L^2$-metric. We endow this measurable space with a probability
measure (prior) $\Pi$. The prior might depend on $T$, but we will suppress this in the notation. As the $p_\theta$ are equivalent and have densities $p_\theta$ relative to a common measure, the posterior of a measurable set $A \subseteq L^2(\mathbb{T})$ is given by

$$
\Pi(A \mid X^T) = \frac{\int_A p_\theta(X^T) d\Pi(\theta)}{\int p_\theta(X^T) d\Pi(\theta)},
$$

provided the denominator is a.s. positive, and $\theta \mapsto p_\theta(X^T)$ is a.s. Borel measurable on $L^2(\mathbb{T})$, which is the content of the following lemma.

**Lemma 1.4.** Suppose that $\Pi$ is a Borel probability measure (prior) on $L^2(\mathbb{T})$. Then for every $\theta_0 \in L^2(\mathbb{T})$ it $\mathbb{P}_{\theta_0}$-a.s. holds that

(i) the random map $\theta \mapsto p_\theta(X^T)$ admits a version that is Borel measurable on $L^2(\mathbb{T})$,

(ii) for the denominator we have $0 < \int p_\theta(X^T) \Pi(d\theta) < \infty$.

**Proof.** (i). We deal with the Lebesgue integral and the stochastic integral in (1.7) separately. First note that by the occupation times formula eq. (1.6),

$$
\int_0^T \theta(X_t)^2 dt = \int_0^1 \theta(x)^2 \hat{L}(t, x) dx.
$$

Since $\mathbb{P}_{\theta_0}$-a.s. we have $\|\hat{L}(T, \cdot)\|_\infty < \infty$, this implies that $\theta \mapsto \int_0^T \theta(X_t)^2 dt$ is a continuous and hence measurable functional on $L^2(\mathbb{T})$.

Using the SDE for $X$, the stochastic integral in (1.7) can be written as the sum of a Lebesgue integral and a Brownian integral. The Lebesgue integral can be handled as in the preceding paragraph. To show that the Brownian integral $\theta \mapsto \int_0^T \theta(X_t) dW_t$ is measurable on $L^2(\mathbb{T})$ we write

$$
L^2(\mathbb{T}) = \bigcup_{K \in \mathbb{N}} B_K,
$$

where $B_K = \{\theta \in L^2(\mathbb{T}) : \|\theta\|_2 \leq K\}$. On every ball $B_K$ the measurability follows from the first statement of the Stochastic Fubini theorem as given in Theorem 2.2 of Veraar (2012). Indeed, condition (2.1) of Veraar (2012) translates into the requirement that, $\mathbb{P}_{\theta_0}$-a.s.,

$$
\int_{B_K} \left( \int_0^T \theta(X_t)^2 dt \right)^{1/2} \Pi(d\theta) < \infty.
$$
This is clearly fulfilled since, by the occupation times formula again, the left-hand side is bounded by $K\|\hat{L}(T, \cdot)\|_\infty^{1/2}$.

(ii). For the upper bound we note that the $\mathbb{P}_0$-expectation of the denominator equals 1, hence it is $\mathbb{P}_0$-a.s. finite. But then also $\mathbb{P}_\theta_0$-a.s., since the measures are equivalent by Lemma 1.2.

For the lower bound we first observe that since $\Pi$ is a probability measure on $L^2(\mathbb{T})$, there exists a $K > 0$ such that $\Pi(B_K) > 0$. Let $\tilde{\Pi}$ be the restriction of $\Pi$ to $B_K$, renormalised so that it is a probability measure again. Then it follows from Jensen’s inequality that

$$\int p_\theta(X^T)d\Pi(\theta) \geq \Pi(B_K) \int p_\theta(X^T)d\tilde{\Pi}(\theta) \geq \Pi(B_K) \exp\left(\int \log p_\theta(X^T)d\tilde{\Pi}(\theta)\right).$$

Hence, it suffices to show that $\mathbb{P}_\theta_0$-a.s.,

$$\left|\int \log p_\theta(X^T)\tilde{\Pi}(\theta)\right| < \infty.$$

As before the log-likelihood can be written as a sum of Lebesgue and stochastic integrals. Dealing with the Lebesgue integrals is straightforward, in view of the occupation times formula again and the a.s. finiteness of $\|\hat{L}(T, \cdot)\|_\infty$. It remains to show that $\mathbb{P}_\theta_0$-a.s.,

$$\left|\int \left(\int_0^T \theta(X_t)dW_t\right)\tilde{\Pi}(\theta)\right| < \infty.$$

But this follows from the stochastic Fubini theorem of Veraar (2012) again, since as shown above the necessary condition for the theorem to hold is fulfilled.

In this thesis we propose priors whose posteriors asymptotically (as $T \to \infty$) concentrate all mass in smaller and smaller balls around the true parameter.

**Definition 1.5.** We say that the posterior contracts at $\varepsilon_T$ converges with rate $\varepsilon_T \downarrow 0$ when for some constant $M > 0$

$$\mathbb{P}_\theta_0 \Pi(\theta \in L^2(\mathbb{T}) : \|\theta - \theta_0\|_2 \leq M\varepsilon_T \mid X^T) \to 1, \text{ as } T \to \infty.$$
Thus we have posterior contraction at rate $\varepsilon_T$ when asymptotically all posterior mass concentrates in $M\varepsilon_T$-$L^2$-balls around the true parameter with probability converging to one. As we only study non-parametric models in this thesis it is not necessary to have a (slowly) diverging sequence $M_T \to \infty$ instead of $M$.

The main result in van der Meulen, van der Vaart, and van Zanten (2006) gives sufficient conditions for deriving posterior contraction rates in Brownian semi-martingale models. The following theorem is an adaptation and refinement of Theorem 2.1 and Lemma 2.2 of van der Meulen, van der Vaart, and van Zanten (2006) for diffusions defined on the circle. The $\varepsilon$-covering number of a set $A$ for a semi-metric $d$, denoted by $N(\varepsilon, A, d)$, is defined as the minimal number of $d$-balls of radius $\varepsilon$ needed to cover the set $A$. The logarithm of the covering number is referred to as the entropy.

The following theorem characterises the rate of posterior contraction for diffusions on the circle in terms of properties of the prior and parameter space.

**Theorem 1.6.** Let $\Pi$ be a prior on $L^2(\mathbb{T})$ (which might depend on $T$), $\theta_0 \in L^2(\mathbb{T})$ and $(\varepsilon_T)_{T \geq 0}$ a sequence of positive numbers such that $T\varepsilon_T^2 \to \infty$. Assume that there is a constant $\xi > 0$ such that

$$\Pi(\theta \in L^2(\mathbb{T}) : \|\theta_0 - \theta\|_2 < \varepsilon_T) \geq e^{-\xi T\varepsilon_T^2} \quad (1.10)$$

and for every $K > 0$ there is a measurable set $\Theta_T \subseteq L^2(\mathbb{T})$ so that for every $a > 0$ there is a constant $C > 0$ so that for $T$ large enough,

$$\log N(a\varepsilon_T, \{\theta \in \Theta_T : \|\theta_0 - \theta\|_2 < \varepsilon_T\}, \|\cdot\|_2) \leq CT\varepsilon_T^2, \quad (1.11)$$

and

$$\Pi(L^2(\mathbb{T}) \setminus \Theta_T) \leq e^{-K T\varepsilon_T^2}. \quad (1.12)$$

Then for some $M > 0$

$$\mathbb{P}_{\theta_0} \Pi(\theta \in L^2(\mathbb{T}) : \|\theta - \theta_0\|_2 \leq M\varepsilon_T \mid X_T) \to 1.$$

Equations (1.10), (1.11) and (1.12) are referred to as the prior mass (or small ball), entropy and remaining mass condition of theorem 1.6 respectively.
1.5. Bayesian nonparametric inference for periodic diffusion models

Proof. A general result for deriving contraction rates for Brownian semi-martingale models is proved in van der Meulen, van der Vaart, and van Zanten (2006). Theorem 1.6 follows upon verifying the assumptions of this result for the diffusion on the circle. These assumptions are easily seen to boil down to:

1. For every $T > 0$ and $\theta_1, \theta_2 \in L^2(\mathbb{T})$ the measures $\mathbb{P}_{\theta_1}$ and $\mathbb{P}_{\theta_2}$ are equivalent.

2. The posterior as defined in equation eq. (1.9) is well defined.

3. Define the random metric $h$ (depending on $T$) on $L^2(\mathbb{T})$ by

$$h(\theta_1, \theta_2) := \sqrt{\int_0^T \left( \theta_1(X_t) - \theta_2(X_t) \right)^2 dt}, \quad \theta_1, \theta_2 \in L^2(\mathbb{T}).$$

There are constants $0 < c < C$ for which

$$\lim_{T \to \infty} \mathbb{P}_{\theta_0} \left( c\sqrt{T} \| \theta_1 - \theta_2 \| \leq h(\theta_1, \theta_2) \leq C \sqrt{T} \| \theta_1 - \theta_2 \|, \forall \theta_1, \theta_2 \in L^2(\mathbb{T}) \right) = 1.$$

Conditions 1 and 2 are the content of lemma 1.2 and lemma 1.4, respectively. We continue checking by the third condition. We already derived that

$$\int_0^T f(X_t)dt = \int_0^1 f(x)\hat{L}_T(x)dx$$

for every measurable function $f$ for which the above integrals are defined. As $\theta_0 \in \hat{L}^2(\mathbb{T})$, it follows from lemma 1.3 that $\hat{L}_T/T$ converges uniformly to a positive deterministic function depending only on $\theta_0$ and which is bounded away from zero and infinity. Since the random metric $h$ can be written as

$$h(\theta_1, \theta_2) = \sqrt{T} \sqrt{\int_0^1 (\theta_1(x) - \theta_2(x))^2 \frac{\hat{L}_T(x)}{T} dx}$$

it follows that the third assumption is satisfied. The fact that the statement holds for some constant $M > 0$ instead of a (slowly) diverging
sequence $M_T \to \infty$ (as in van der Meulen, van der Vaart, and van Zanten (2006)) in the case that $T \varepsilon_T^2 \to \infty$ is easily seen from the second display on page 883 in the proof of Theorem 2.1 of van der Meulen, van der Vaart, and van Zanten (2006), where one notes that $\mu_n$ in van der Meulen, van der Vaart, and van Zanten (2006) is $T \varepsilon_T^2$ in our case.

As already remarked below lemma 1.3, recent unpublished work by M.R. Schauer suggest that the law of large numbers for the local time (lemma 1.3) also holds in the case that $\int_0^1 \theta_0(x)dx \neq 0$, so that by a slight adaption to the proof of theorem 1.6 this theorem holds for every $\theta_0 \in L^2(\mathbb{T})$. 

Gaussian process methods for one-dimensional diffusions: Optimal rates and adaptation

Various papers have recently considered nonparametric Bayesian procedures for one-dimensional stochastic differential equations (SDEs) with periodic drift. This is motivated among others by problems in which SDEs are used for the dynamic modelling of angles in different contexts. See for instance Hindriks (2011) for applications in the modelling of neuronal rhythms and Pokern (2007) for the use of SDEs in the modelling of angles in molecular dynamics.

The first paper to propose a concrete nonparametric Bayesian method in this context and to study its implementation was Papaspiliopoulos et al. (2012). In Pokern, Stuart, and van Zanten (2013) the first theoretical results were obtained for this procedure. These papers consider observations \((X_t : t \in [0, T])\) from the basic SDE model

\[
dX_t = \theta(X_t) \, dt + dB_t, \quad X_0 = 0,
\]

where \(B\) is a Brownian motion, and the drift function \(\theta\) belongs to the space \(\dot{L}^2(\mathbb{T})\) of square integrable, periodic functions on \([0, 1]\) with
zero mean, i.e. \( \int_0^1 \theta(x) \, dx = 0 \). For the function \( \theta \) of interest a Gaussian process prior is proposed with mean zero and precision (inverse covariance) operator

\[
\eta((-\Delta)^{\alpha+1/2} + \kappa I),
\]

where \( \Delta \) is the one-dimensional Laplacian, \( I \) is the identity operator and \( \eta, \kappa > 0 \) and \( \alpha + 1/2 \in \{2, 3, \ldots\} \) (\( p = \alpha + 1/2 \) in Pokern, Stuart, and van Zanten (2013)) are fixed hyperparameters. It can be proved that this defines a valid prior on \( \dot{L}^2(\mathbb{T}) \), cf. Pokern, Stuart, and van Zanten (2013), Section 2.2.

The main convergence result proved in Pokern, Stuart, and van Zanten (2013) asserts that if in this setup the true drift \( \theta_0 \) generating the data has (Sobolev) regularity \( \alpha + 1/2 \), then the corresponding posterior distribution of \( \theta \) contracts around \( \theta_0 \) at the rate \( T^{-\alpha/(1+2\alpha)} \) as \( T \to \infty \), with respect to the \( L^2 \)-norm. In the concluding section of Pokern, Stuart, and van Zanten (2013) it is already conjectured that this result is not completely sharp. More specifically, it is anticipated that the rate \( T^{-\alpha/(1+2\alpha)} \) should already be attainable under the less restrictive assumption that the drift \( \theta_0 \) has regularity of order \( \alpha \). The first main result of this chapter confirms that this is indeed the case. Since the degree of regularity of the Gaussian process with precision eq. (2.1) is (essentially) \( \alpha \) (see e.g. Pokern, Stuart, and van Zanten (2013), Lemma 2.2.), this reconciles the result for this SDE model with the general message from the Gaussian process prior literature, which says that to obtain optimal rates with Gaussian process priors, one should match the regularities of the prior and the truth (see van der Vaart and van Zanten (2008a)). Although lower bounds for the minimax rate appear to be unknown for the exact model we consider in this chapter, results for closely related models suggest it is of the order \( T^{-\alpha/(1+2\alpha)} \) for an \( \alpha \)-Sobolev smooth drift function (e.g. Kutoyants (2004, section 4.5)).

We are able to obtain the improved result by following a different mathematical route than in Pokern, Stuart, and van Zanten (2013). The latter paper uses more or less explicit representations of the posterior mean and covariance in terms of weak solutions of certain differential equations to study the asymptotic behaviour of the posterior using techniques from PDE theory. In the present chapter we follow instead the approach of van der Meulen, van der Vaart, and van Zanten (2006), which is essentially an adaptation to the SDE case of the
general “testing approach” which has by now become well known in Bayesian nonparametrics. These ideas, combined with results about the asymptotic behaviour of the so-called periodic diffusion local time from Pokern, Stuart, and van Zanten (2013), allow us to obtain the new, sharp result for the Gaussian process prior with precision eq. (2.1).

The scope of this result is still somewhat limited, since it is a non-adaptive statement. Indeed, it is not realistic to assume that we know the regularity of the truth exactly and hence it is unlikely that we guess the correct smoothness of the prior leading to the optimal contraction rate. We therefore also consider several ways of obtaining adaptation to smoothness for this problem. A first option we explore is putting a prior on the multiplicative constant $\eta$ in eq. (2.1), instead of taking it fixed as in Papaspiliopoulos et al. (2012) and Pokern, Stuart, and van Zanten (2013). This leads to a hierarchical, conditionally Gaussian process prior on the drift $\theta$. Our second main result shows that if the hyperprior on $\eta$ is appropriately chosen, then adaptation is obtained for the whole range of regularities between 0 and $\alpha + 1/2$. More precisely, if the degree of regularity $\beta$ of the true drift belongs to $(0, \alpha + 1/2]$, then we attain the posterior contraction rate $T^{-\beta/(1+2\beta)}$.

It is obviously desirable to have a large range of regularities to which we can adapt. At first sight, the result just discussed might suggest to let $\alpha$ tend to infinity with $T$. However, it turns out that the parameter $\alpha$ appears in the constant multiplying the rate of contraction. A straightforward adaptation of the proof of the previous result (which we will not carry out in this chapter, since it contains no new ideas) shows that although taking a hyperparameter $\alpha_T \to \infty$ would indeed lead to adaptation over $(0, \infty)$, the rate would deteriorate by a factor $(\alpha_T)^c$ for some constant $c > 0$.

The preceding observations indicate that in order to obtain adaptation to the full range of possible regularities for the drift, using a prior on the multiplicative scale parameter $\eta$ is perhaps not the best option. Therefore we also consider another possibility, namely putting a prior on the hyperparameter $\alpha$ that controls the regularity of the prior directly. We prove that this is, from the theoretical perspective at least, indeed preferable. We can obtain the optimal contraction rate for any regularity of the truth, without suffering a penalty in the rate.

In this chapter we focus on deriving theoretical results. We do not consider the related numerical issues, since this requires a completely
different analysis, but these are clearly of interest as well. For instance, it is quite conceivable that the last option we consider, putting a prior on $\alpha$, is numerically quite demanding, more so than putting a prior on $\eta$. Therefore in practice it might actually be worthwhile to accept non-optimal statistical rates or only a limited range of adaptation, in order to gain speed on the numerical side.

The chapter is organised as follows. In the next section we define the several priors on the diffusion parameter $\theta$ of eq. (1.5). In section 2.2 we present and discuss the main results described briefly in the introduction. Some auxiliary results that we use in the proofs are prepared in section 2.3. The proofs themselves are given in sections 2.4 to 2.6.

2.1 Definition of the prior

To make Bayesian inference about the drift function we consider a Gaussian process prior on the space of drift functions $\dot{L}^2(\mathbb{T})$. We are interested in the Gaussian process with mean zero and precision operator eq. (2.1). As shown in section 2.2 of Pokern, Stuart, and van Zanten (2013), the Gaussian process $V$ with this mean and covariance can be written as

$$V = \frac{1}{\sqrt{\eta}} \sum_{k=1}^{\infty} \sqrt{\lambda_k} \phi_k Z_k,$$

where the $Z_k$ are independent standard normal variables, the $\phi_k$ are the orthonormal eigenfunctions of the Laplacian, given by

$$\phi_{2k-1}(x) = \sqrt{2} \sin(2\pi k x),$$
$$\phi_{2k}(x) = \sqrt{2} \cos(2\pi k x),$$

for $k \in \mathbb{N}$, and

$$\lambda_k = \left( \left( \frac{4\pi^2}{2} \right) \frac{k^2}{2} \right)^{\alpha+1/2} + \kappa \right)^{-1}. \tag{2.2}$$

The results we derive in this chapter actually do not depend crucially on the exact form of the eigenfunctions and eigenvalues $\phi_k$ and $\lambda_k$. The $\phi_k$ can in fact be any orthonormal basis of $\dot{L}^2(\mathbb{T})$ (provided the smoothness spaces defined ahead are changed accordingly). Moreover, the specific value of the hyperparameter $\kappa$ in eq. (2.2) is irrelevant for
our results. For the $\lambda_k$ we only need that there exist constants $c, C > 0$ and $\alpha > 0$ such that

$$ck^{-1/2-\alpha} \leq \sqrt{\lambda_k} \leq Ck^{-1/2-\alpha}. \quad (2.3)$$

Note that the $\lambda_k$'s in eq. (2.2) satisfy these bounds. For notational convenience we will work with $\sqrt{\lambda_k} = k^{-1/2-\alpha}$ throughout the chapter, but all results hold if this exact choice is replaced by $\lambda_k$'s satisfying eq. (2.3).

Introducing the notation $L = 1/\sqrt{\eta}$ for the scaling constant, the priors for the drift function $\theta$ that we consider take the general form

$$\theta \sim L \sum_{k=1}^{\infty} k^{-1/2-\alpha} \phi_k Z_k, \quad (2.4)$$

where the $Z_k$ are independent standard Gaussian variables and $(\phi_k)$ is an arbitrary, fixed orthonormal basis of $\dot{L}^2(\mathbb{T})$. We will consider various setups in which the scale $L$ is either a constant or a random factor, and also the regularity parameter $\alpha$ will be either a fixed constant or random.

The regularity of the true drift function $\theta_0$ that generates the data will be measured in Sobolev sense relative to the basis $(\phi_k)$. For $\beta > 0$ we define

$$\dot{H}^\beta(\mathbb{T}) = \left\{ f = \sum_{k=1}^{\infty} f_k \phi_k \in \dot{L}^2(\mathbb{T}) : \sum_{k=1}^{\infty} f_k^2 k^{2\beta} < \infty \right\}.$$

Note that in the case that the $\phi_k$ are the eigenfunctions of the Laplacian given above, this is the usual $L^2$-Sobolev regularity.

2.2 Main results

In this section we present the main rate of contraction results for the posteriors corresponding to the various priors of the form eq. (2.4), with different choices for the hyperparameters $L$ and $\alpha$. The proofs of the results are given in Sections 2.4 to 2.6.

For simplicity the prior on $\theta$ will always be denoted by $\Pi$, but it will be clearly described in each case.
2. Gaussian process methods for one-dimensional diffusions

2.2.1 Fixed hyperparameters

Our first main result deals with the case that the scaling parameter and the regularity parameters of the Gaussian process are fixed, positive constants. Specifically, we fix $L > 0$ and $\beta > 0$ and define the prior $\Pi$ on the drift function structurally as

$$\theta \sim L \sum_{k=1}^{\infty} k^{-1/2-\beta} \phi_k Z_k,$$

where the $Z_k$ are independent standard Gaussian variables and $(\phi_k)$ is the chosen orthonormal basis of $L^2(\mathbb{T})$. Note that the expected squared $L^2$-norm of $\theta$ under this prior is $L^2 \sum_{k} k^{-1-2\beta} < \infty$, hence by lemma 1.4 the posterior is well defined. Recall the definition of posterior contraction, definition 1.5. We have the following result for the prior just described.

**Theorem 2.1.** Let the prior be given by eq. (2.5), with $\beta, L > 0$ fixed. If $\theta_0 \in \dot{H}^{\beta}(\mathbb{T})$ for $\beta > 0$, then the posterior contracts around $\theta_0$ at the rate $\varepsilon_T = T^{-\beta/(1+2\beta)}$.

As noted in the introduction, this theorem improves Theorem 5.2 of Pokern, Stuart, and van Zanten (2013). The latter corresponds to the case that the $\phi_k$ are the eigenfunctions of the Laplacian and $\beta + 1/2 \in \{2, 3, \ldots\}$. In Pokern, Stuart, and van Zanten (2013) the obtained rate for this prior is also (essentially) $T^{-\beta/(1+2\beta)}$, but this is obtained under the stronger condition that $\theta_0$ belongs to $\dot{H}^{\beta+1/2}(\mathbb{T})$. Additionally, the new result is valid for all $\beta > 0$.

2.2.2 Prior on the scale

The fact that we get the optimal rate $T^{-\beta/(1+2\beta)}$ in theorem 2.1 strongly depends on the fact that the degree of smoothness $\beta$ of the true drift $\theta_0$ matches the choice of the regularity parameter of the prior. Although strictly speaking it has not been established for the SDE setting of this chapter, results from the Gaussian process prior literature for analogous settings indicate that if these regularities are not matched exactly, then sub-optimal rates will be obtained (see for instance van der Vaart and
van Zanten (2008a) and Castillo (2008)). We would obviously prefer a method that does not depend on knowledge of the true regularity $\beta$ of the truth and that *adapts* to this degree of smoothness automatically.

In this section we consider a first method to achieve this. This involves putting a prior distribution on the scaling parameter $L$ instead of taking it fixed. We employ a hierarchical prior $\Pi$ on $\theta$ that can be described as follows:

$$L \sim \frac{E^{1/2+\alpha}}{\sqrt{T}}, \quad (2.6)$$

$$\theta \mid L \sim L \sum_{k=1}^{\infty} k^{-1/2-\alpha} \phi_k Z_k. \quad (2.7)$$

Here $\alpha > 0$ is a fixed hyperparameter, which should be thought of as describing the “baseline smoothness” of the prior. The $Z_k$ and $\phi_k$ are as before and $E$ is a standard exponential, independent of the $Z_k$. Note that we could equivalently describe the prior on $L$ as a Weibull distribution with scale parameter $1/\sqrt{T}$ and shape parameter $2/(1+2\alpha)$. Lemma 1.4 ensures again that the posterior is well defined, since by conditioning we see that the expected squared $L^2$-norm of $\theta$ is now given by $c_L \sum k^{-1-2\alpha}$, where $c_L$ is the second moment of $L$ under the prior, which is finite.

The specific choice of the prior for $L$ is convenient, but the proof of the following theorem shows that it can actually be slightly generalised. It is for instance enough that the random variable $E$ in eq. (2.6) has a density that satisfies exponential lower and upper bounds in the tail. Our proof breaks down however if we deviate too much from the choice above. For instance, without the dependence on $T$ we would only be able to derive sub-optimal rates. We stress that this does not mean that other priors cannot lead to optimal rates, only that such results cannot be obtained using our technical approach. An alternative route, for instance via empirical Bayes as in Knapik et al. (2015), might lead to less restrictive assumptions on the hyperprior for $L$. This will require a completely different analysis however.

**Theorem 2.2.** Let the prior be given by eqs. (2.6) and (2.7), with $\alpha > 0$ fixed. If $\theta_0 \in H^\beta(\mathbb{T})$ for $\beta \in (0, \alpha + 1/2]$, then the posterior contracts around $\theta_0$ at the rate $\varepsilon_T = T^{-\beta/(1+2\beta)}$.  

19
So indeed with a prior on the multiplicative scale we can achieve adaptation for a range of smoothness levels $\beta$. Note however that the range is limited by the baseline smoothness $\alpha$ of the prior. Putting a prior on the scale $L$ does allow to adapt to truths that are arbitrarily rougher than the prior, but if the degree of smoothness of the truth is larger than $\alpha + 1/2$, the procedure does not achieve optimal rates. This phenomenon has been observed in the literature in different statistical settings as well. See for instance Szabó, van der Vaart, and van Zanten (2013) for similar results in the white noise model.

2.2.3 Prior on the Gaussian process regularity

To circumvent the potential problems described in the preceding section, we consider an alternative method for achieving adaptation to all smoothness levels. Instead of taking a fixed baseline prior smoothness and putting a prior on the scale, we put a prior on the Gaussian process smoothness itself. Specifically, we use a prior on $\alpha$ that is truncated to the growing interval $(0, \alpha_T]$ and that has a density proportional to $x \mapsto \exp(-T^{1/(1+2x)})$ on that interval. For convenience we take $\alpha_T = \log T$, but other choices are possible as well. We define the probability density $\lambda_T$, with support $[0, \log T]$, by

$$
\lambda_T(x) = C_T^{-1} e^{-T^{1/(1+2x)}}, \quad x \in [0, \log T],
$$

where $C_T$ is the normalising constant. The full prior $\Pi$ on $\theta$ that we employ is now described as follows:

$$
\alpha \sim \lambda_T, \quad \tag{2.8}
$$

$$
\theta \mid \alpha \sim \sum_{k=1}^{\infty} k^{-1/2-\alpha} \phi_k Z_k, \quad \tag{2.9}
$$

where the $\phi_k$ and $Z_k$ are again as before. Note that for this prior we have that for every $\alpha > 0$, the conditional prior probability that $\|\theta\|_2 < \infty$ given $\alpha$ equals 1, hence the unconditional prior probability that the norm is finite is 1 as well. Lemma 1.4 thus implies the posterior is well defined again and we can formulate the following result.

**Theorem 2.3.** Let the prior be given by eqs. (2.8) and (2.9). If $\theta_0 \in \dot{H}^\beta(\mathbb{T})$ for $\beta > 0$, then the posterior contracts around $\theta_0$ at the rate $\varepsilon_T = T^{-\beta/(1+2\beta)}$. 

20
So by placing a prior on $\alpha$ we obtain adaptation to all smoothness levels, without paying for it in the rate. A similar result has recently been obtained in the setting of the white noise model in Knapik et al. (2015). We note however that the results in the latter paper rely on rather explicit computations specific for that model. The results we present here for the SDE model are derived in a completely different way, by using the testing approach proposed in van der Meulen, van der Vaart, and van Zanten (2006). We note that the rates we obtain are slightly better than those in Knapik et al. (2015), in the sense that we don’t obtain additional slowly varying factors. We expect that similar results can be obtained for white noise model and other related models by adapting our proofs. A downside of our approach is that we can only prove the desired result for somewhat contrived hyperpriors on $\alpha$ such as $\lambda_T$, which may appear unnatural at first sight. The result is however in accordance with similar findings for other statistical models obtained for instance in Lember and van der Vaart (2007) and Ghosal, Lember, and van der Vaart (2008). Our prior on $\alpha$ has a density proportional (on $(0, \alpha_T]$) to $\exp(-T\varepsilon_{\alpha,T}^2)$, where $\varepsilon_{\alpha,T}$ is the rate we would get when using the unconditional Gaussian process prior on the right of eq. (2.9). Hence our theorem is in accordance with the results in the cited papers, which state that in some generality, such a choice of hyperprior leads to rate-adaptive procedures. Other priors on $\alpha$ may lead to adaptation as well, including potentially priors that do not depend on the sample length $T$. But to prove such results, different mathematical techniques seem to be required.

The main point we want to make here however, and that is supported by the theorems we present, is that if the goal is to achieve adaptation to an unrestricted range of smoothness levels, then, from the theoretical point of view at least, putting a prior on a smoothness hyperparameter is preferable to fixing the baseline smoothness of the prior and putting a prior on a multiplicative scaling parameter.

2.3 Auxiliary results

2.3.1 Small ball probabilities

In this section we prepare a result that allows us to verify the prior mass condition eq. (1.10) of theorem 1.6 for the various priors in section 2.2.
For $\alpha, L > 0$ we define the Gaussian process

$$W^{\alpha,L} = L \sum_{k=1}^{\infty} k^{-1/2-\alpha} \phi_k Z_k,$$

where the $Z_k$ are independent standard Gaussian variables and $(\phi_k)$ is an arbitrary orthonormal basis of $\dot{L}^2(\mathbb{T})$.

**Lemma 2.4.** There exists a positive, continuous function $f$ on $(0, \infty)$ and constants $c_0, c_1 > 0$ such that $c_0 \alpha \leq f(\alpha) \leq c_1 \alpha$ for $\alpha$ large enough and

$$-\log \mathbb{P}(\|W^{\alpha,L}\|_2 < \varepsilon) \leq f(\alpha) \left(\frac{L}{\varepsilon}\right)^{1/\alpha},$$

for all $\alpha > 0$ and for $\varepsilon/L > 0$ small enough.

**Proof.** Note that $\mathbb{P}(\|W^{\alpha,L}\|_2 < \varepsilon) = \mathbb{P}(\|W^{\alpha,1}\|_2 < \varepsilon/L)$, so the case $L = 1$ implies the general case. Since $(\phi_k)$ is an orthonormal basis, $\mathbb{P}(\|W^{\alpha,1}\|_2 < \varepsilon) = \mathbb{P}(\sum_{k=1}^{\infty} k^{-2\alpha-1} Z_k^2 < \varepsilon^2)$. The result then follows from Corollary 4.3 of Dunker, Lifshits, and Linde (1998) and straightforward algebra.

Next we consider the reproducing kernel Hilbert space (RKHS) $\mathbb{H}^{\alpha,L}$ associated to the Gaussian process $W^{\alpha,L}$. It follows from the series representation eq. (2.10) that $\mathbb{H}^{\alpha,L} = \dot{H}^{1/2+\alpha}(\mathbb{T})$, and that the associated RKHS norm of an element $h \in \mathbb{H}^{\alpha,L}$ satisfies $L \|h\|_{\mathbb{H}^{\alpha,L}} = \|h\|_{2,1/2+\alpha}$, where for $\beta > 0$, the Sobolev norm $\|h\|_{2,\beta}$ of a function $h = \sum h_k \phi_k$ is defined by

$$\|h\|_{2,\beta}^2 = \sum_{k=1}^{\infty} h_k^2 k^{2\beta}.$$

For these facts and more general background on reproducing kernel Hilbert spaces of Gaussian processes with a view towards Bayesian nonparametrics, see van der Vaart and van Zanten (2008b).

**Lemma 2.5.** Suppose that $\theta_0 \in \dot{H}^{\beta}(\mathbb{T})$ for $\beta \leq \alpha + 1/2$. Then for $\varepsilon > 0$ small enough,

$$\inf_{h \in \mathbb{H}^{\alpha,L}, \|h - \theta_0\|_2 \leq \varepsilon} \|h\|_{\mathbb{H}^{\alpha,L}}^2 \leq \|\theta_0\|_{2,\beta}^2 \frac{1}{L^2 \varepsilon} \frac{2\beta - 2\alpha - 1}{\beta}.$$
Proof. Consider the expansion \( \theta_0 = \sum_{k=1}^{\infty} \theta_k \phi_k \) and define \( h = \sum_{k \leq I} \theta_k \phi_k \), where \( I \) will be determined below. We have that \( h \in \mathbb{H}^{\alpha,L} \), and from the smoothness condition on \( \theta_0 \) it follows that

\[
\|h - \theta_0\|_2^2 = \sum_{k > I} \theta_k^2 \leq I^{-2\beta} \sum_{k > I} \theta_k^2 k^{2\beta}.
\]

Since \( \theta_0 \in \dot{H}^\beta(\mathbb{T}) \) the sum on the right vanishes for \( I \to \infty \), hence \( \|h - \theta_0\|_2^2 \leq I^{-2\beta} \) for \( I \) large enough. Setting \( I = \lfloor \varepsilon^{-1/\beta} \rfloor \) we obtain that, for \( \varepsilon \) small enough, the infimum in the statement of the lemma is bounded by

\[
\frac{1}{L^2} \sum_{k \leq I} \theta_k^2 k^{1+2\alpha} \leq \frac{1}{L^2} \sum_{k \leq I} \theta_k^2 k^{1+2\alpha-2\beta} \leq \frac{1}{L^2} \|\theta_0\|_{2,\beta}^2 I^{1+2\alpha-2\beta},
\]

since \( \beta \leq \alpha + 1/2 \). The proof is completed by recalling the choice of \( I \).

Lemmas 2.4 and 2.5 together give a non-centered small ball probability bound for the Gaussian process \( W^{\alpha,L} \). This will be used to verify the prior mass condition eq. (1.10) of theorem 1.6 for the various priors.

**Lemma 2.6.** Suppose that \( \alpha > 0 \) and \( \theta_0 \in \dot{H}^\beta(\mathbb{T}) \) for \( \beta \leq \alpha + 1/2 \). There exist a constant \( C > 0 \), depending only on \( \theta_0 \), such that

\[
P(\|W^{\alpha,L} - \theta_0\|_2 < \varepsilon) \geq \exp \left(-C \left(f(\alpha) \left(\frac{L}{\varepsilon} \right)^{1/\alpha} + \frac{1}{L^2} \frac{\varepsilon^{2\beta-2\alpha-1}}{\beta} \right)\right).
\]

for \( \varepsilon/L > 0 \) small enough.

Proof. This follows directly from lemmas 2.4 and 2.5 using, for instance, Lemma 5.3 of van der Vaart and van Zanten (2008b). \( \Box \)
2.4 Proof of theorem 2.1

In this case the prior $\Pi$ is the law of the Gaussian process $W^{\beta,L}$. Applying lemma 2.6 with $\alpha = \beta$ we obtain, for $\theta_0 \in \dot{H}^\beta(\mathbb{T})$, the bound

$$
\Pi(\theta : \|\theta - \theta_0\|_2 \leq \varepsilon) \geq e^{-C\varepsilon^{-1/\beta}},
$$

for a constant $C > 0$ and $\varepsilon > 0$ small enough. It follows that the prior mass condition eq. (1.10) of theorem 1.6 is satisfied for $\varepsilon_T$ a constant times $T^{-\beta/(1+2\beta)}$. By the general result for Gaussian process priors given by Theorem 2.1 of van der Vaart and van Zanten (2008a), the other assumptions of theorem 1.6 are then automatically satisfied as well. Hence, the desired result follows from an application of that theorem.

2.5 Proof of theorem 2.2

We will again verify the conditions of theorem 1.6. We note that in this case, the conditional distribution of $\theta$ under the prior, given the value of $L$, is the law of $W^{\alpha,L}$.

2.5.1 Prior mass condition

Denoting the prior density of $L$ by $g$, and assuming again that $\theta_0 \in \dot{H}^\beta(\mathbb{T})$, we have, by lemma 2.6, that there exists a constant $C > 0$ such that for $\varepsilon$ small enough,

$$
\Pi(\theta : \|\theta - \theta_0\|_2 \leq \varepsilon) = \int \mathbb{P}(\|W^{\alpha,L} - \theta_0\|_2 \leq \varepsilon) g(L) dL \\
\geq \int_{\varepsilon^{(\beta-\alpha)/\beta}}^{2\varepsilon^{(\beta-\alpha)/\beta}} e^{-C((L/\varepsilon)^{1/\alpha} + \varepsilon(2\beta-2\alpha-1)/\beta)/L^2)} g(L) dL.
$$

On the range of integration the exponential in the integrand is bounded from below by $e^{-C'\varepsilon^{-1/\beta}}$ for some $C' > 0$. Moreover, the assumptions on the prior on $L$ imply that for $\varepsilon$ a multiple of $T^{-\beta/(1+2\beta)}$,

$$
\int_{\varepsilon^{(\beta-\alpha)/\beta}}^{2\varepsilon^{(\beta-\alpha)/\beta}} g(L) dL = \mathbb{P}(cT^{1/(1+2\beta)} < E < 2cT^{1/(1+2\beta)}) \geq e^{-3cT^{1/(1+2\beta)}}
$$
for a constant $c > 0$ and $T$ large enough. It follows that there exist constants $c_1, c_2 > 0$ such that for $\varepsilon_T = c_1 T^{-\beta/(1+2\beta)}$,

$$\Pi(\theta : \|\theta - \theta_0\|_2 \leq \varepsilon_T) \geq e^{-c_2 T \varepsilon_T^2},$$

which covers condition eq. (1.10) of theorem 1.6.

### 2.5.2 Sieves

Recall from section 2.3.1 that the RKHS unit ball $\mathbb{H}^{\alpha,L}_1$ of $W^{\alpha,L}$ is the ball $\dot{H}^{\alpha+1/2}_L(\mathbb{T})$ of radius $L$ in the Sobolev space $\dot{H}^{\alpha+1/2}(\mathbb{T})$ of regularity $\alpha + 1/2$. This motivates the definition of sieves $B_T$ of the form

$$B_T = R\dot{H}^{\alpha+1/2}_1(\mathbb{T}) + \varepsilon_T \dot{L}^2_1(\mathbb{T}),$$

where $R$ will be determined below and $\dot{L}^2_1(\mathbb{T})$ is the unit ball in $\dot{L}^2(\mathbb{T})$.

### Remaining mass condition

By conditioning we have, for any $L_0 > 0$,

$$\Pi(\theta \notin B_T) = \int \mathbb{P}(W^{\alpha,L} \notin B_T) g(L) dL \leq \int_0^{L_0} \mathbb{P}(W^{\alpha,L} \notin B_T) g(L) dL + \int_{L_0}^{\infty} g(L) dL. \quad (2.11)$$

The second term on the right is bounded by $\exp(-(L_0^2 T)^{1/(1+2\alpha)})$, by the assumptions on the prior on $L$. For $L_0$ a large enough multiple of $T^{(\alpha-\beta)/(1+2\beta)}$ this is bounded by $e^{-DT^{1/(1+2\beta)}}$, for a given constant $D > 0$.

As for the first term, note that the probability in the integrand is increasing in $L$. Since $B_T = (R/L_0) \mathbb{H}^{\alpha,L_0}_1 + \varepsilon_T \dot{L}^2_1(\mathbb{T})$, the Borell-Sudakov inequality (see van der Vaart and van Zanten (2008b), Theorem 5.1) implies that

$$\mathbb{P}(W^{\alpha,L_0} \notin B_T) \leq 1 - \Phi(\Phi^{-1}(\mathbb{P}(\|W^{\alpha,L_0}\|_2 \leq \varepsilon_T)) + R/L_0).$$

By lemma 2.4, the probability on the right is bounded from below by $\exp(-C(L_0/\varepsilon_T)^{1/\alpha})$ for some $C > 0$. Furthermore, since for $y \in (0, 1/2)$,

$$\Phi^{-1}(y) \geq -\sqrt{\frac{2}{\pi}} \log(1/y)$$

25
2. Gaussian process methods for one-dimensional diffusions

and for \( x \geq 1, 1 - \Phi(x) \leq \exp(-x^2/2) \), we have

\[
P(W^{\alpha,L_0} \not\in \mathbb{B}_T) \leq \exp\left(-\frac{1}{2} \left( \frac{R}{L_0} - \sqrt{C'(\frac{L_0}{\varepsilon_T})^{1/\alpha}} \right)^2 \right),
\]

for some \( C' > 0 \). The choices of \( L_0 \) and \( \varepsilon_T \) imply that if \( R \) is chosen to be a large multiple of \( T^{(1/2 + \alpha - \beta)/(1 + 2\beta)} \), then the first term on the right of eq. (2.11) is bounded by \( e^{-DT^{1/(1 + 2\beta)}} \) as well.

**Entropy**

It remains to verify that \( \mathbb{B}_T \) satisfies the entropy condition eq. (1.11) of theorem 1.6. By the known entropy bound for Sobolev balls (see for instance Ghosal and van der Vaart (2017, Proposition C.7)) we have

\[
\log N(\varepsilon, RH^{\alpha+1/2}_1, \| \cdot \|_2) \leq C\left( \frac{R}{\varepsilon} \right)^{2/(1+2\alpha)}
\]

for some \( C > 0 \). Recalling the definitions of \( \mathbb{B}_T, \varepsilon_T \) and \( R \), it follows that

\[
\log N(2\varepsilon_T, \mathbb{B}_T, \| \cdot \|_2) \leq C\left( \frac{R}{\varepsilon_T} \right)^{2/(1+2\alpha)} \leq C'T^{1/(1 + 2\beta)}
\]

for some \( C' > 0 \). This concludes the proof of the theorem.

### 2.6 Proof of theorem 2.3

Note that in this case the conditional prior law of \( \theta \), given \( \alpha \), is the law of the Gaussian process \( W^{\alpha,1} \).

#### 2.6.1 Prior mass condition

By lemma 2.6, there exist a constant \( C > 0 \) such that for \( \varepsilon \) small enough, \( \delta > 0 \) and \( \theta_0 \in \dot{H}^\beta(T) \),

\[
\Pi(\theta : \| \theta - \theta_0 \|_2 \leq \varepsilon) \geq \int_\beta^{\beta+\delta} \mathbb{P}(\| W^{\alpha,1} - \theta_0 \|_2 \leq \varepsilon) \lambda_T(\alpha) \, d\alpha
\]

\[
\geq \int_\beta^{\beta+\delta} e^{-C((1/\varepsilon)^{1/\alpha} + \varepsilon^{(2\beta - 2\alpha - 1)/\beta})} \lambda_T(\alpha) \, d\alpha.
\]
2.6. Proof of theorem 2.3

On the range of integration the exponential in the integrand is bounded from below by \(\exp\left(-C'\varepsilon^{-1/(1+2\delta)/\beta}\right)\) for some \(C' > 0\). Since \(\lambda_T\) is increasing, we get

\[
\Pi(\theta : \|\theta - \theta_0\|_2 \leq \varepsilon) \geq \delta C_T^{-1} e^{-T^{1/(1+2\beta)}} e^{-C'\varepsilon^{-1/(1+2\delta)/\beta}}.
\]

Since \(C_T \leq \log T\) and by choosing \(\delta\) to be a multiple of \(1/\log T\), it follows that, for \(\varepsilon_T\) a multiple of \(T^{-\beta/(1+2\beta)}\), condition eq. (1.10) of theorem 1.6 is fulfilled.

2.6.2 Remaining mass and entropy

In this case we take sieves of the form \(B_T = R\dot{H}_{1}^{\gamma+1/2}(\mathbb{T}) + \varepsilon_T\dot{L}_{1}^{2}(\mathbb{T})\), where \(\gamma\) and \(R\) will be determined below.

For the remaining mass we have

\[
\Pi(\theta \notin B_T) \leq \int_0^\gamma \lambda_T(\alpha) \, d\alpha + \int_\gamma^\infty \mathbb{P}(W^{\alpha,1} \notin B_T) \lambda_T(\alpha) \, d\alpha.
\]

For \(\alpha \geq \gamma\) we have \(B_T \supset R\dot{H}_{1}^{\gamma+1/2}(\mathbb{T}) + \varepsilon_T\dot{L}_{1}^{2}(\mathbb{T})\). Hence, by the Borell-Sudakov inequality,

\[
\mathbb{P}(W^{\alpha,1} \notin B_T) \leq 1 - \Phi\left(\Phi^{-1}\left(\mathbb{P}(\|W^{\alpha,1}\|_2 \leq \varepsilon_T)\right) + R\right).
\]

Note that \(\|W^{\alpha,1}\|_2 \leq \|W^{\gamma,1}\|_2\), so \(\mathbb{P}(\|W^{\alpha,1}\|_2 \leq \varepsilon_T) \geq \mathbb{P}(\|W^{\gamma,1}\|_2 \leq \varepsilon_T).\) By lemma 2.4, the latter is bounded from below by \(\exp(-C_\gamma\varepsilon_T^{-1/\gamma})\) for a \(C_\gamma > 0\). We note that \(C_\gamma\) depends continuously on \(\gamma\), through the continuous function \(f\) in lemma 2.4. Below we will chose \(\gamma\) to be in a shrinking neighbourhood of \(\beta\), which is fixed. Hence, for this choice of \(\gamma\), we have that \(\mathbb{P}(\|W^{\gamma,1}\|_2 \leq \varepsilon_T) \geq \exp(-C_\gamma\varepsilon_T^{-1/\gamma})\) for a constant \(C > 0\) that is independent of \(\gamma\). We conclude that for \(\gamma \leq \alpha\),

\[
\mathbb{P}(W^{\alpha,1} \notin B_T) \leq \exp\left(-\left(R - \sqrt{C'}\left(\frac{1}{\varepsilon_T}\right)^{1/\gamma}\right)^2\right)
\]

for some \(C' > 0\). Taking \(R\) a large multiple of \(\varepsilon_T^{-1/(2\gamma)}\) this is bounded by \(\exp(-D\varepsilon_T^{-1/\gamma})\) for a given constant \(D > 0\). For the other term, observe that by definition of \(\lambda\),

\[
\int_0^\gamma \lambda(\alpha) \, d\alpha \leq \gamma C_T^{-1} e^{-T^{1/(1+2\gamma)}} \leq \gamma e^{-T^{1/(1+2\gamma)}},
\]
since $C_T \geq \frac{\log T}{2 \exp(e)}$. Putting things together, we have

$$
\Pi(\theta \notin \mathbb{B}_T) \leq e^{-De_T^{-1/\gamma}} + \gamma e^{-T^{1/(1+2\gamma)}}.
$$

If we choose $\gamma = \beta/(1 + C'/\log T)$ for a large enough constant $C > 0$, then the right-hand side is smaller than $\exp(-DT\varepsilon_T^2)$, as desired.

For the entropy we have, as before,

$$
\log N(2\varepsilon_T, \mathbb{B}_T, \| \cdot \|_2) \leq C(\frac{R}{\varepsilon_T})^{2/(1+2\gamma)}.
$$

For the choice of $R$ that we made the right side is a constant times $\varepsilon_T^{-1/\gamma}$, which by the choice of $\gamma$ is bounded by a constant times $T\varepsilon_T^2$. 
Randomly truncated orthonormal series prior with Gaussian coefficients and inverse gamma scaling

3.1 Introduction

In Bayesian function estimation, a common approach to putting a prior distribution on a function $f$ of interest, for instance a regression function in nonparametric regression models or a drift function in diffusion models, is to expand the function in a particular basis and to endow the coefficients in the expansion with prior weights. For computational or other reasons the series is often truncated after finitely many terms, and the truncation level is endowed with a prior as well. The coefficients in the expansion are often chosen to be independent under the prior and distributed according to some given probability density.

It is of interest to understand whether, in addition to their attractive conceptual and computational aspects, nonparametric priors of this type enjoy favourable theoretical properties as well. Examples
of papers in which this was studied for various families of series priors include Zhao (2000), Shen and Wasserman (2001), de Jonge and van Zanten (2012), Rivoirard and Rousseau (2012), Arbel, Gayraud, and Rousseau (2013), Shen and Ghosal (2015). The results in these papers show that when appropriately constructed, random series priors can yield posteriors that contract at optimal rates and that adapt automatically to the smoothness of the function that is being estimated.

To ensure that the nonparametric Bayes procedure not only adapts to smoothness, but is also flexible with respect to the multiplicative scale of the function of interest, a multiplicative hyperparameter with an independent prior distribution is often employed as well. Theoretically this is usually not needed for an optimal contraction rate of the posterior, but it can greatly improve performance in practice.

In van der Meulen, Schauer, and van Zanten (2014) it is explained why it is computationally attractive in certain settings to use Gaussian priors on the series coefficients in combination with a multiplicative (squared) scaling parameter with an inverse gamma prior. For a given truncation level, the prior is conjugate and allows for posterior computations using standard Gibbs sampling. The existing theoretical results do not cover this important case however. This is mainly due to the fact that essentially, the available rate of contraction theorems for series priors require that hyper priors have (sub-)exponential tails, which excludes the inverse gamma distribution. (For example the second part of condition (A2) of Shen and Ghosal (2015) is not satisfied in our setting.) The theoretical properties of random series priors with inverse gamma scaling have therefore remained unexplored. Here we fill this gap.

Concretely, we consider statistical models in which the unknown object of interest is a square integrable function \( f \) on \([0, 1]\). We endow this function with a prior that is hierarchically specified as follows:

\[
J \sim \text{Poisson or geometric},
\]

\[
s^2 \sim \text{inverse gamma},
\]

\[
f \mid s, J = \sum_{j \leq J} f_j \phi_j, \quad \text{with } (f_1, \ldots, f_J) \sim N(0, \text{diag}(s^2 j^{-1-2\alpha}))_{j \leq J},
\]

where \((\phi_j)\) is a fixed orthonormal basis of \(L^2[0, 1]\) and \(\alpha > 0\) is a hyperparameter. (In fact, we will consider a somewhat broader class
of hyper priors on $J$ and $s^2$, see section 3.2.)

In recent years, general rate of contraction theorems have been derived for a variety of nonparametric statistical problems. Roughly speaking, such theorems give sufficient conditions for having a certain rate of contraction in terms of (i) the amount of mass that the prior gives to neighbourhoods of the true function and (ii) the existence of growing subsets of the support of the prior, so-called sieves, that contain all but an exponentially small amount of the prior mass and whose metric entropy is sufficiently small. See for instance theorem 1.6. The statements of our main theorem in this chapter match the conditions of some of these existing general results. This means that we automatically obtain results for different statistical settings, including for instance signal estimation in white noise and drift estimation for SDEs.

A simple but important observation that we make in this and the next chapter is that in order to obtain sharp rates for the priors that we consider, it is necessary to use versions of the general contraction rate theorems that give entropy conditions on the intersection of the sieves with balls around the true function, as can be found for instance in Ghosal, Ghosh, and van der Vaart (2000), van der Meulen, van der Vaart, and van Zanten (2006) and Ghosal and van der Vaart (2007). As remarked in these papers, it is in many nonparametric problems sufficient to consider only the entropy of the sieves themselves, without intersecting them with a ball around the truth. For the priors we consider in this paper however, which in some sense are finite-dimensional in nature in certain regimes, this is not the case. It turns out that since the inverse gamma distribution has polynomial tails, we need to make the sieves relatively large in order to ensure that they receive sufficient prior mass. Without intersecting them with a small ball around the truth, this would make their entropy too large, or even infinite.

The proofs of our main results indicate that the good adaptation properties of series priors like eq. (3.1) are really due to the fact that \textit{both} the truncation level $J$ and the scaling constant $s$ are random. If the true function that is being estimated is relatively smooth, the prior can approximate it well by letting $J$ be small. If it is relatively rough however, the prior can adapt to it by letting $J$ be essentially infinite, or very large, to pick up all the fluctuations. The correct bias-variance trade-off is in that case achieved automatically by adapting the
multiplicative scale. In some sense, priors like eq. (3.1) can switch with sufficient probability between being essentially finite-dimensional, and being essentially infinite-dimensional. In combination with a random multiplicative scale, this gives them the ability to adapt to all levels of smoothness.

The remainder of the chapter is organised as follows. In the next section we describe in detail the class of priors we consider. In section 3.3 we present the main results of the paper, which give bounds on the amount of mass that the priors give to $L^2$-neighbourhoods of functions with a given degree of (Sobolev-type) smoothness, and the existence of appropriate sieves within the support of the prior. In section 3.4 we link these general theorems to existing rate of contraction results for two different SDE models, to obtain concrete contraction results for signal estimation in white noise and drift estimation of a one-dimensional SDE with priors of the form eq. (3.1). The proofs of the main results are given in sections 3.5 and 3.6.

### 3.2 Prior model

We consider problems in which the unknown function of interest (e.g. a drift function of an SDE, a signal observed in noise, . . .) is a square integrable function on $[0, 1]$, i.e. an element of $L^2[0, 1] = \{ f : [0, 1] \rightarrow \mathbb{R} : \| f \|_2 < \infty \}$, where the $L^2$-norm is as usual defined by $\| f \|_2^2 = \int_0^1 f^2(x) \, dx$. We fix an arbitrary orthonormal basis $(\phi_j)$ of $L^2[0, 1]$ (for instance the standard Fourier basis). Every element of $f \in L^2[0, 1]$ can be represented as a series $f = \sum_j \langle f, \phi_j \rangle \phi_j$ where the convergence is in the $L^2$-norm and by the Plancherel formula $\| f \|_2^2 = \sum_j |\langle f, \phi_j \rangle|^2$. Finite series $\sum_{j \leq J} \langle f, \phi_j \rangle \phi_j$ approximate $f$ and the quality of this approximation depends on the decay of the coefficients $\langle f, \phi_j \rangle$, which also determines the “smoothness” of the function. The class of $\beta$-Sobolev smooth functions $H^\beta[0, 1]$ is given by all $f \in L^2[0, 1]$ for which the $\beta$-Sobolev norm

$$\| f \|_\beta := \sqrt{\sum_j k^{2\beta} |\langle f, \phi_j \rangle|^2}$$

is finite. If $\phi_j$ is the classical Fourier series basis, these are the classical $\beta$-Sobolev spaces.
3.3. Main results

We define a series prior on a function $f \in L^2[0,1]$ through a hierarchical scheme which involves a prior on the point $J$ at which the series is truncated, a prior on the multiplicative scaling constant $s$ and conditionally on $s$ and $J$, a series prior with Gaussian coefficients on $f$.

Specifically, the prior on $J$ is defined through a probability mass function $p$ that is assumed to satisfy, for constants $C, C' > 0$,

$$p(j) \gtrsim e^{-Cj \log j}, \quad \sum_{i > j} p(i) \lesssim e^{-C'j}$$

for all $j \in \mathbb{N}$. This includes for instance the cases of a Poisson or a geometric prior on $J$. For the scaling parameter we assume that the density $g$ of $s^2$ is positive and continuous and satisfies, for some $q < -1$ and $C'' > 0$,

$$g(x) \gtrsim e^{-C''/x} \quad \text{near } 0, \quad g(x) \gtrsim x^q \quad \text{near } \infty.$$  

Hence in particular, the popular and computationally convenient choice of an inverse gamma prior on $s^2$ is included in our setup. The full prior $\Pi$ is then specified as follows:

$$J \sim p$$

$$s^2 \sim g$$

$$f \mid s, J \sim s \sum_{j=1}^{J} j^{-1/2-\alpha} Z_j \phi_j,$$

where $\alpha$ is a positive constant which determines the baseline smoothness of the prior, $p$ satisfies eq. (3.2), $g$ satisfies eq. (3.3) and the $Z_j$ are independent standard Gaussians.

3.3 Main results

Our main abstract result gives properties of the truncated series prior that link directly to the conditions of existing general theorems for posterior contraction in a variety of statistical settings. Combined with such existing results, we obtain concrete results for, for instance, signal estimation in white noise, drift estimation in diffusion models, et cetera. We give concrete examples in the next section.
As usual, if $\mathcal{F}$ is a subset of a normed vector space with norm $\| \cdot \|$, then we denote by $N(\varepsilon, \mathcal{F}, \| \cdot \|)$ the minimal number of balls of $\| \cdot \|\text{-radius } \varepsilon$ needed to cover the set $\mathcal{F}$.

**Theorem 3.1.** Let the prior $\Pi$ on $f$ be as defined in eqs. (3.4) to (3.6), with $\alpha > 0$ and $p$ and $g$ satisfying eqs. (3.2) and (3.3). Let $f_0 \in H^\beta[0,1]$ for $\beta > 0$. Then there exists a constant $c > 0$ such that for every $K > 1$, there exist $\mathcal{F}_n \subset L^2[0,1]$ such that with

$$
\varepsilon_n = c\left(\frac{n}{\log n}\right)^{-\beta/(1+2\beta)},
$$

we have

$$
\Pi(f : \|f - f_0\|_2 \leq \varepsilon_n) \geq e^{-n\varepsilon_n^2}, \quad (3.7)
$$

$$
\Pi(f \notin \mathcal{F}_n) \leq e^{-Kn\varepsilon_n^2}, \quad (3.8)
$$

$$
\log N(a\varepsilon_n, \{f \in \mathcal{F}_n : \|f - f_0\|_2 \leq \varepsilon_n\}, \| \cdot \|_2) \lesssim n\varepsilon_n^2, \quad (3.9)
$$

for all $a \in (0,1)$.

The proof of the theorem is given in section 3.5. The result matches with the sufficient conditions of existing posterior contraction theorems, provided that the relevant statistical distance-type quantities (e.g. Hellinger, Kullback-Leibler, . . .) in the model can be appropriately linked to the $L^2$-norm on the parameter $f$. In the next section we give two concrete SDE-related examples, which motivated the present study.

Theorem 3.1 shows that with truncated series priors of the type eqs. (3.4) to (3.6) we can have adaption to arbitrary degrees of smoothness in certain function estimation problems, and achieve posterior contraction rates that are optimal up to a logarithmic factor. Inspection of the proof of theorem 3.1 shows that in the range $\beta \leq \alpha + 1/2$, i.e. if the “baseline smoothness” $\alpha$ of the prior happens to have been chosen large enough relative to the smoothness $\beta$ of the true function, then we actually get the optimal rate $n^{-\beta/(1+2\beta)}$ without additional logarithmic factors. This is true under a slightly stronger condition on the prior on the cut-off point $J$. Instead of eq. (3.2), we need to assume that for constants $C, C' > 0$ it holds that

$$
p(j) \gtrsim e^{-Cj}, \quad \sum_{i>j} p(i) \lesssim e^{-C'j}. \quad (3.10)
$$
for all $j \in \mathbb{N}$. This means that the prior on $J$ can still be geometric, but that the Poisson prior on $J$ is excluded.

**Theorem 3.2.** Let the prior $\Pi$ on $f$ be as defined in eqs. (3.4) to (3.6), with $\alpha > 0$ and $p$ and $g$ satisfying eqs. (3.3) and (3.10). Let $f_0 \in H^\beta[0, 1]$ for $0 < \beta \leq \alpha + 1/2$. Then there exists a constant $c > 0$ such that for every $K > 1$, there exist $F_n \subset L^2[0, 1]$ such that with

$$\varepsilon_n = cn^{-\beta/(1+2\beta)},$$

we have

$$\Pi(f : \|f - f_0\|_2 \leq \varepsilon_n) \geq e^{-n\varepsilon_n^2}, \quad (3.11)$$

$$\Pi(f \notin F_n) \leq e^{-Kn\varepsilon_n^2}, \quad (3.12)$$

$$\log \mathcal{N}(a\varepsilon_n, \{f \in F_n : \|f - f_0\|_2 \leq \varepsilon_n\}, \|\cdot\|_2) \lesssim n\varepsilon_n^2, \quad (3.13)$$

for all $a \in (0, 1)$.

The proof of this theorem is given in section 3.6.

### 3.4 Specific statistical settings

#### 3.4.1 Detecting a signal in Gaussian white noise

Suppose we observe a sample path $X^{(n)} = (X^{(n)}_t : t \in [0, 1])$ of stochastic process satisfying the SDE

$$dX^{(n)}_t = f_0(t) \, dt + \frac{1}{\sqrt{n}} \, dW_t,$$

where $W$ is a standard Brownian motion and $f_0 \in L^2[0, 1]$ is an unknown signal. To make inference about the signal we endow it with the truncated series prior $\Pi$ described in section 3.2 and we compute the corresponding posterior $\Pi(\cdot | X^{(n)})$. Theorem 3.1 of van der Meulen, van der Vaart, and van Zanten (2006) or Theorem 6 of Ghosal and van der Vaart (2007), combined by our main result theorem 3.1, imply that if $f_0 \in H^\beta[0, 1]$ for $\beta > 0$, then we have the posterior contraction

$$\Pi(f : \|f - f_0\|_2 > M(n/\log n)^{-\beta/(1+2\beta)} | X^{(n)}) \overset{P}{\to} 0$$

for some positive constant $M$, where the convergence is in probability under the true model corresponding to the signal $f_0$.  

3. Randomly truncated orthonormal series prior

3.4.2 Estimating the drift of periodic diffusion

The results in this chapter are readily adapted to the case that $\Pi$ is a prior on $\dot{L}^2(\mathbb{T})$ and $(\phi_k)$ is an orthonormal basis of $\dot{L}^2(\mathbb{T})$. Let, as in chapters 1 and 2, $X^T$ an observation of the SDE (1.5), which is given by

\[
\begin{align*}
    dX_t &= \theta_0(X_t)dt + dW_t, \\
    X_0 &= x_0,
\end{align*}
\]

where $\theta_0 \in \dot{L}^2(\mathbb{T})$ is the unknown one-periodic drift parameter. Let the prior on the parameter space $\dot{L}^2(\mathbb{T})$ be as described in section 3.2, only with $(\phi_k)$ an orthonormal basis of $\dot{L}^2(\mathbb{T})$. Then we obtain posterior contraction rate $(T/\log T)^{-\beta/(1+2\beta)}$.

3.5 Proof of theorem 3.1

3.5.1 Prior mass

The following theorem implies that eq. (3.7) holds with $\varepsilon_n$ as specified.

**Theorem 3.3.** Let the prior $\Pi$ on $f$ be defined according to eqs. (3.4) to (3.6), with $\alpha > 0$ and $p$ and $g$ satisfying eqs. (3.2) and (3.3), and let $f_0 \in H^\beta[0,1]$ for $\beta > 0$. Then, for a constant $C > 0$, it holds that

\[-\log \Pi(f : \|f - f_0\|_2 \leq 2\varepsilon) \leq C\varepsilon^{-1/\beta} \log 1/\varepsilon,\]

for all $\varepsilon > 0$ small enough.

**Proof.** Recall that $s^2$ has density $g$ under the prior. Hence, by conditioning we see that the probability of interest is bounded from below by

\[
p\left(\left[\varepsilon/\|f_0\|_\beta\right]^{-1/\beta}\right) \times \int_0^\infty \Pi\left(\sum_{j=1}^{\left[\varepsilon/\|f_0\|_\beta\right]^{-1/\beta}} j^{-1/2 - \alpha} Z_j \phi_j - f_0 \right)_2^2 \leq 2\varepsilon \right) g(\eta) d\eta,
\]
Now suppose first that $1 + 2\alpha - 2\beta \leq 0$. Then by lemma 3.5 below, the preceding is further lower bounded by

$$\exp \left(-C_1 \varepsilon^{-1/\beta} \log 1/\varepsilon\right) p\left(\left[ (\varepsilon/\|f_0\|_{\beta})^{-1/\beta} \right] \right) \int_{\varepsilon^{1/\beta}}^{2\varepsilon^{1/\beta}} g(\eta) \, d\eta \quad (3.14)$$

for a constant $C_1 > 0$. By the assumption on $p$ (eq. (3.2))

$$p\left(\left[ (\varepsilon/\|f_0\|_{\beta})^{-1/\beta} \right] \right) \gtrsim \exp \left(-C_2 \varepsilon^{-1/\beta} \log 1/\varepsilon\right),$$

for some constant $C_2 > 0$, and by the behaviour of $g$ near zero (eq. (3.3)),

$$\int_{\varepsilon^{1/\beta}}^{2\varepsilon^{1/\beta}} g(\eta) \, d\eta \gtrsim e^{-C_3 \varepsilon^{-1/\beta}},$$

for some $C_3 > 0$. Hence eq. (3.14) is bounded from below by a constant times $\exp(-C_4 \varepsilon^{-1/\beta} \log 1/\varepsilon)$ for $\varepsilon$ small enough, for some constant $C_4 > 0$.

In the other case $1 + 2\alpha - 2\beta > 0$ we restrict the integral over $\eta$ to a different region to obtain instead the lower bound

$$\exp \left(-C_1 \varepsilon^{-1/\beta} \log 1/\varepsilon\right) p\left(\left[ (\varepsilon/\|f_0\|_{\beta})^{-1/\beta} \right] \right) \int_{\varepsilon^{(2\beta - 2\alpha)/\beta}}^{2\varepsilon^{(2\beta - 2\alpha)/\beta}} g(\eta) \, d\eta \gtrsim \exp \left(-C_2 \varepsilon^{-1/\beta} \log 1/\varepsilon\right) \int_{\varepsilon^{(2\beta - 2\alpha)/\beta}}^{2\varepsilon^{(2\beta - 2\alpha)/\beta}} g(\eta) \, d\eta \quad (3.15)$$

for some positive constants $C_1$ and $C_2$. For $\alpha < \beta < \alpha + 1/2$ making use of the behaviour of $g$ near zero

$$\int_{\varepsilon^{(2\beta - 2\alpha)/\beta}}^{2\varepsilon^{(2\beta - 2\alpha)/\beta}} g(\eta) \, d\eta \gtrsim e^{-C_3 \varepsilon^{-(2\beta - 2\alpha)/\beta}} \geq e^{-C_3 \varepsilon^{-1/\beta}},$$

for some constant $C_3 > 0$ and using that $1 > 2\beta - 2\alpha$. Hence eq. (3.15) is up to a constant bounded from below by $e^{-C_4 \varepsilon^{-1/\beta} \log 1/\varepsilon}$, for some constant $C_4 > 0$. When $\alpha = \beta$, using that $g$ is positive, we have that

$$\int_{\varepsilon^{(2\beta - 2\alpha)/\beta}}^{2\varepsilon^{(2\beta - 2\alpha)/\beta}} g(\eta) \, d\eta = \int_{1}^{2} g(\eta) \, d\eta$$
is a positive constant, and again eq. (3.15) is up to a constant bounded from below by $e^{-C_4 \varepsilon^{-1/\beta} \log 1/\varepsilon}$, for some constant $C_4 > 0$.

For the range $\beta < \alpha$ using the behaviour of $g$ near infinity, we obtain

$$\int_{\varepsilon(2\beta - 2\alpha)/\beta}^{2\varepsilon(2\beta - 2\alpha)/\beta} g(\eta) \, d\eta \gtrsim \varepsilon^{-|C_3(2\beta - 2\alpha)/\beta|} \geq e^{-\varepsilon^{-1/\beta}}.$$ 

for some constant $C_3 > 0$. Hence eq. (3.15) is up to a constant lower bounded by $e^{-C_4 \varepsilon^{-1/\beta} \log 1/\varepsilon}$, for some positive constant $C_4$.

Lemma 3.4. Let $Z_1, Z_2, \ldots$ be independent and standard normal. There exists a universal constant $K > 1$ such that for every $s > 0$, $\varepsilon > 0$, $J \in \mathbb{N}$ and $a \in \ell^2$,

$$- \log \mathbb{P} \left( \left\| s \sum_{j=1}^{J} a_j Z_j \phi_j \right\|_2 \leq \varepsilon \right) \leq 2J \log \left( K \lor \frac{s \|a\|_2}{\varepsilon} \right).$$

Proof. Since the $\phi_j$ form an orthonormal basis, the probability we have to lower bound equals

$$\mathbb{P} \left( s^2 \sum_{j=1}^{J} a_j^2 Z_j^2 \leq \varepsilon^2 \right) \geq \mathbb{P} \left( \max_{j \leq J} |Z_j| \leq \frac{\varepsilon}{s \|a\|_2} \right)$$

$$= \left( \mathbb{P} \left( |Z_1| \leq \frac{\varepsilon}{s \|a\|_2} \right) \right)^J.$$ 

If $\varepsilon/(s \|a\|_2) \geq \xi_3/4$, with $\xi_p$ the $p$-quantile of the standard normal distribution, $\mathbb{P}(|Z_1| \leq \varepsilon/(s \|a\|_2)) \geq 1/2$. In the other case, it is at least $\phi(\xi_3/4) \times 2\varepsilon/(s \|a\|_2)$, with $\phi$ the standard normal density. So in either case, it is at least a constant $C \in (0, 1)$ times $1 \land \varepsilon/(s \|a\|_2)$. It follows that

$$- \log \mathbb{P} \left( \left\| s \sum_{j=1}^{J} a_j Z_j \phi_j \right\|_2 \leq \varepsilon \right) \geq J \log C + J \log \left( 1 \land \frac{\varepsilon}{s \|a\|_2} \right) \geq 2J \log \left( C \land \frac{\varepsilon}{s \|a\|_2} \right).$$

This implies the statement of the lemma. \qed
Lemma 3.5. Let \( Z_1, Z_2, \ldots \) be independent and standard normal. Let \( \beta > 0 \) and \( f_0 \in H^\beta[0, 1] \) be given. There exists a constant \( K > 1 \) such that for all \( \varepsilon, s, \alpha > 0 \) and \( J \geq (\varepsilon/\|f_0\|_\beta)^{-1/\beta} \),

\[
- \log \mathbb{P} \left( \left\| \sum_{j=1}^{J} j^{-1/2-\alpha} Z_j \phi_j - f_0 \right\|_2 \leq 2\varepsilon \right) \\
\leq 2J \log \left( K \vee \frac{s}{\varepsilon} \right) + \frac{\|f_0\|_\beta^2}{s^2} J^{(1+2\alpha-2\beta)\vee 0}.
\]

Proof. For fixed \( J, s \), the sum \( s \sum_{j=1}^{J} j^{-1/2-\alpha} Z_j \phi_j \) is a centered Gaussian random element in \( L^2[0, 1] \) and has associated reproducing kernel Hilbert space (RKHS), which is the space \( H^{s,J} \) of all functions \( h = \sum_{j \leq J} h_j \phi_j \), with RKHS-norm

\[
\left\| \sum_{j \leq J} h_j \phi_j \right\|_{H^{s,J}}^2 = \frac{1}{s^2} \sum_{j \leq J} j^{1+2\alpha} h_j^2.
\]

The function \( f_0 \) admits a series expansion \( f_0 = \sum f_j \phi_j \). For \( J_0 \leq J \), consider the function \( h_0 = \sum_{j \leq J_0} f_j \phi_j \) in the RKHS. It holds that

\[
\|f_0 - h_0\|_2^2 = \sum_{j > J_0} f_j^2 \leq J_0^{-2\beta} \|f_0\|_\beta^2.
\]

Hence for \( J_0 = \lfloor (\varepsilon/\|f_0\|_\beta)^{-1/\beta} \rfloor \), we have that \( \|f_0 - h_0\|_2 \leq \varepsilon \). The condition on \( J \) ensures that \( h_0 \) is an element of the RKHS, and

\[
\|h_0\|_{H^{s,J}}^2 = \frac{1}{s^2} \sum_{j \leq J_0} j^{1+2\alpha-2\beta} f_j^2 \leq \frac{\|f_0\|_\beta^2}{s^2} J_0^{(1+2\alpha-2\beta)\vee 0}.
\]

It follows that

\[
\inf_{h \in H^{s,J}, \|h - f_0\| \leq \varepsilon} \|h\|_{H^{s,J}}^2 \leq \frac{\|f_0\|_\beta^2}{s^2} J^{(1+2\alpha-2\beta)\vee 0}.
\]

Combining this with the preceding lemma and Lemma 5.3 of van der Vaart and van Zanten (2008b) completes the proof. \( \square \)
3. Randomly truncated orthonormal series prior

3.5.2 Sieves, remaining mass and entropy

Let the sequence $\varepsilon_n \to 0$ and $\beta > 0$ be given. We consider sieves of growing dimension of the form

$$\mathcal{F}_n = \left\{ h = \sum_{j \leq J_n} h_j \phi_j \right\}, \quad (3.17)$$

where

$$J_n = K_1 \varepsilon_n^{-1/\beta} \log 1/\varepsilon_n \quad (3.18)$$

for a constant $K_1 > 0$ specified below.

By assumption eq. (3.2) we have

$$\Pi(f \not\in \mathcal{F}_n) = \Pi(J > J_n) \lesssim e^{-C' K_1 \varepsilon_n^{-1/\beta} \log 1/\varepsilon_n}.$$ 

This implies that statement eq. (3.8) of theorem 3.1 holds if $K_1$ is chosen large enough.

As for the entropy condition eq. (3.9), we note that if the function $f_0$ admits the series expansion $f_0 = \sum_j f_{0,j} \phi_j$, then a function $f \in \mathcal{F}_n$ which satisfies $\|f - f_0\|_2 \leq \varepsilon$ is of the form $f = \sum_{j \leq J_n} f_j \phi_j$, and $\sum_{j \leq J_n} (f_j - f_{0,j})^2 \leq \varepsilon^2$. Hence, the covering number in eq. (3.9) is bounded by the $a\varepsilon$-covering number of a ball of radius $\varepsilon$ in $\mathbb{R}^{J_n}$, which is bounded by $(3/a)^{J_n}$ (see, for instance, Pollard (1990)). In view of the choice eq. (3.18) of $J_n$ it follows that eq. (3.9) holds.

3.6 Proof of theorem 3.2

Under the conditions of theorem 3.2 we can replace the result of theorem 3.3 by the following, which implies that eq. (3.11) holds.

**Theorem 3.6.** Let the prior $\Pi$ on $f$ be defined according to eqs. (3.4) to (3.6), with $\alpha > 0$ and $p$ and $g$ satisfying eqs. (3.3) and (3.10), and let $f_0 \in H^\beta[0,1]$ with $0 < \beta \leq \alpha + 1/2$. Then, for a constant $C > 0$, it holds that

$$- \log \Pi(f : \|f - f_0\|_2 \leq \varepsilon) \leq C \varepsilon^{-1/\beta},$$

for all $\varepsilon > 0$ small enough.
Proof. Instead of using lemma 3.4 we simply note that for $s > 0$ and $J \in \mathbb{N}$, and $Z_1, Z_2, \ldots$ independent and standard normal,

$$- \log \mathbb{P} \left( \left\| s \sum_{j=1}^{J} j^{-1/2-\alpha} Z_j \phi_j \right\|_2 \leq \varepsilon \right)$$

$$\leq - \log \mathbb{P} \left( \left\| \sum_{j=1}^{\infty} j^{-1/2-\alpha} Z_j \phi_j \right\|_2 \leq \varepsilon / s \right).$$

By lemma 2.4 the right-hand side is bounded by a constant times $(\varepsilon / s)^{-1/\alpha}$. Using eq. (3.16) and Lemma 5.3 of van der Vaart and van Zanten (2008b), we see that for $J \geq (\varepsilon / \| f_0 \|_\beta)^{-1/\beta}$,

$$- \log \mathbb{P} \left( \left\| s \sum_{j=1}^{J} j^{-1/2-\alpha} Z_j \phi_j - f_0 \right\|_2 \leq 2\varepsilon \right)$$

$$\lesssim \left( \frac{s}{\varepsilon} \right)^{1/\alpha} + \frac{1}{s^2} \vee \varepsilon^{-(1+2\alpha-2\beta)/\beta}.$$

For $\beta \leq \alpha + 1/2$ the two terms on the right are balanced for $s$ of the order $\varepsilon^{(\beta-\alpha)/\beta}$, in which case the right-hand side of bounded by a constant times $\varepsilon^{-1/\beta}$. It follows by conditioning that

$$\Pi(f : \| f - f_0 \|_2 \leq 2\varepsilon)$$

$$\geq \exp \left( - c_1 \varepsilon^{-1/\beta} \right) p \left( \left\lfloor c_2 \varepsilon^{-1/\beta} \right\rfloor \right) \int_{\varepsilon^{(2\beta-2\alpha)/\beta}}^{2\varepsilon^{(2\beta-2\alpha)/\beta}} g(\eta) \, d\eta.$$

Using the assumptions on $p$ (now with eq. (3.10)) and $g$ (eq. (3.3)) one can show in a similar way as in the proof of theorem 3.3 that this is bounded from below by a constant times $\exp(-C\varepsilon^{-1/\beta})$, for some constants $C, c_1, c_2 > 1$.

To complete the proof of theorem 3.2 we note that in this case we can use the same sieves $F_n$ as defined in eq. (3.17), but with a different choice for the dimension $J_n$, namely $J_n = \left\lceil K_1 \varepsilon_n^{-1/\beta} \right\rceil$, for some $K_1 > 0$. The tail condition in eq. (3.10) then ensures that eq. (3.12) holds if $K_1$ is chosen large enough. The entropy bound eq. (3.13) is obtained by the same argument as in section 3.5.2, but now using the new choice of $J_n$. \qed
3. Randomly truncated orthonormal series prior
Adaptive nonparametric drift estimation for diffusion processes using Faber-Schauder expansions

4.1 Introduction

Assume we have continuous-time observations $X^T = (X_t : t \in [0, T])$ from a diffusion process $X$ defined as a weak solution to the stochastic differential equation (SDE)

$$dX_t = \theta(X_t)dt + dW_t, \quad X_0 = x_0.$$  \hfill (4.1)

Here $W$ is a Brownian motion and the drift $\theta$ is assumed to be a real-valued measurable function on the real line that is one-periodic and square integrable on $[0,1]$. This model is studied in chapter 1. In chapter 2 we propose a prior distribution on $\theta$ which is defined as

$$\theta = L \sum_{k=1}^{\infty} k^{-1/2-\alpha} \phi_k Z_k,$$  \hfill (4.2)

where $(\phi_k)$ is an orthonormal basis of $\hat{L}^2(\mathbb{T})$, $(Z_k)$ is a sequence of independent standard normally distributed random variables and $\alpha$ is
4. Drift estimation using Faber-Schauder expansions

a positive constant. It is shown that when $L$ and $\alpha$ are fixed and $\theta_0$ is assumed to be $\alpha$-Sobolev smooth, then the rate $T^{-\alpha/(1+2\alpha)}$ is obtained. Note that this result is nonadaptive, as the regularity of the prior must match the regularity of $\theta_0$. To obtain good posterior contraction rates for the full range of possible regularities of the drift, two options are investigated: endowing either $L$ or $\alpha$ with a hyperprior. Only the second option results in the desired adaptivity over all possible regularities.

While the prior eq. (4.2) (with additional prior on $\alpha$) has good asymptotic properties, from a computational point of view the infinite series expansion is inconvenient. Clearly, in any implementation this expansion needs to be truncated. Random truncation of a series expansion is a well known method for defining priors in Bayesian nonparametrics, see for instance Shen and Ghosal (2015) and references therein. Exactly this idea is exploited in van der Meulen, Schauer, and van Zanten (2014). The prior is defined as the law of a random function. Two options are investigated, the random function defined by

$$ S \sum_{k=1}^{R} Z_k \phi_k, \quad (4.3) $$

where $(\phi_k)$ is an orthonormal basis of $\dot{L}^2(\mathbb{T})$, and the $Z_k$ are normally distributed, or by

$$ SZ_1 \psi_1 + S \sum_{j=0}^{R} \sum_{k=1}^{2^j} Z_{jk} \psi_{jk}, \quad (4.4) $$

where the functions $\psi_{jk}$ constitute the Faber-Schauder basis (see fig. 4.1), the prior coefficients $Z_{jk}$ are equipped with a Gaussian distribution, and $R$ and $S$ are independent random variables.

The first choice is investigated in chapter 3, the latter option is investigated in this chapter. Despite the similarities between the two priors, the proofs of this chapter are more elaborate due to the fact that the Faber-Schauder functions are not orthogonal. A benefit of using this prior, is that generally a function is easier expanded in this basis than in the Fourier basis for instance, as expanding in the Faber-Schauder basis does not involve integrating.

The results obtained in this chapter are stronger than in chapter 3, in the sense that we obtain a prior mass and entropy result with respect
4.1. Introduction

Figure 4.1: Elements $\psi_1$ and $\psi_{j,k}$, $0 \leq j \leq 2$ of the Faber-Schauder basis

to the supremum norm. We can combine this result with Ghosal, Ghosh, and van der Vaart (2000, Theorem 2.1) and van der Vaart and van Zanten (2008a, Lemma 3.1), to obtain a posterior contraction result for the density estimation setting.

Draws from the posterior can be computed using a reversible jump Markov Chain Monte Carlo (MCMC) algorithm (cf. van der Meulen, Schauer, and van Zanten (2014)). Fast computation is facilitated by leveraging inherent sparsity properties stemming from the compact support of the functions $\psi_{jk}$. In the discussion of van der Meulen, Schauer, and van Zanten (2014) it is argued that inclusion of both the scaling and random truncation in the prior is beneficial. However, this claim is only supported by simulations results. In this chapter we support this claim theoretically by proving adaptive contraction rates of the posterior distribution in case the prior eq. (4.4) is used. We verify the sufficient conditions for posterior convergence given in theorem 1.6. A consequence of our results is that if the true drift function is $B^\beta_{\infty,\infty}$-Besov smooth, $\beta \in (0, 2)$, then by appropriate choice of the variances of $Z_{jk}$, as well as the priors on $R$ and $S$, the posterior for the drift $\theta$ contracts at the rate $(T/\log T)^{-\beta/(1+2\beta)}$ around the true drift. When the true function $\theta_0 = z_1 \psi_1 + \sum_{j=0}^{\infty} \sum_{k=1}^{2^j} z_{jk} \psi_{jk}$, to be estimated, can be expanded in this basis, and the coefficients decay with order $|z_{jk}| \lesssim 2^{-\beta j}$, then posterior convergence rate $(T/\log T)^{-\beta/(1+2\beta)}$ is achieved, no matter how large $\beta > 0$ is. However, for $\beta \geq 2$, we were
not able to relate this smoothness property to a well known class of smooth functions studied in the literature.

Comparisons with similar models suggest that the obtained rate is minimax-optimal, possibly up to the log factor (see for instance Kutoyants (2004, section 4.5)). Moreover, it is adaptive: the prior does not depend on $\beta$.

The chapter is organised as follows. In the next section we give a precise definition of the prior. The main results are in section 4.3. Many results of section 4.3 concern general properties of the prior and their application is not confined to drift estimation of diffusion processes. To illustrate this, we show in section 4.4 how these results can easily be adapted to nonparametric regression and nonparametric density estimation. Proofs are gathered in section 4.5.

### 4.2 Prior construction

Define the “hat” function $\Lambda$ by

$$\Lambda(x) = (2x)\mathbb{I}_{[0,\frac{1}{2})}(x) + 2(1-x)\mathbb{I}_{[\frac{1}{2},1]}(x).$$

The Faber-Schauder basis functions are given by

$$\psi_{j,k}(x) = \Lambda(2^j x - k + 1), \quad j \geq 0, \quad k = 1, \ldots, 2^j$$

Let

$$\psi_1 = (\psi_{0,1}(x - \frac{1}{2}) + \psi_{0,1}(x + \frac{1}{2})) \mathbb{I}_{[0,1]}(x).$$

In fig. 4.1 we have plotted $\psi_1$ together with $\psi_{j,k}$ where $j \in \{0, 1, 2\}$.

We define our prior as in eq. (4.4) with Gaussian coefficients $Z_1$ and $Z_{jk}$, where the truncation level $R$ and the scaling factor $S$ are equipped with (hyper)priors. We make the following assumptions on $R$ and $S$.

**Assumption 4.1.** The prior on the truncation level satisfies, for some positive constants $c_1, c_2$,

$$\begin{align*}
\mathbb{P}(R > r) & \leq \exp(-c_1 2^r r), \\
\mathbb{P}(R = r) & \geq \exp(-c_2 2^r r).
\end{align*}$$

For the prior on the scaling we assume existence of constants $0 < p_1 < p_2$, $q > 0$ and $C > 1$ with $p_1 > q|\alpha - \beta|$ such that

$$\mathbb{P}(S \in [x^{p_1}, x^{p_2}]) \geq \exp(-x^q) \quad \text{for all } x \geq C.$$
4.3. Theorems on posterior contraction rates

The prior on $R$ can be defined as $R = \lfloor 2\log(Y + 1) \rfloor$, where $Y$ is Poisson distributed. Equation (4.6) is satisfied for a whole range of distributions, including the popular family of inverse gamma distributions. We extend $\theta$ periodically if we want to consider $\theta$ as function on the real line. We divide the basis functions into levels. The functions $\psi_1$ and $\psi_{0,1}$ are said to belong to level 0, which is convenient for notational purposes. For every $j \geq 1$ the functions $\psi_{j,1}, \ldots, \psi_{j,2^j}$ belong to level $j$. For levels $j \geq 1$ the basis functions are per level orthogonal with essentially disjoint support. For notational convenience, let $\mathcal{I}_r$ denote the set of indices $i = 1$ and $i = (j, k), j = 0, \ldots, r, k = 1, \ldots, 2^j$ and for an index $i$, let $\ell(i)$ denote the level of the basis function $\psi_i$ (so $\ell(1) = 0$ and $\ell((j, k)) = j$).

The prior is defined as follows

$$\theta \mid R, S = S \left( Z_1 \psi_1 + \sum_{j=0}^{R-1} \sum_{k=1}^{2^j} Z_{jk} \psi_{jk} \right),$$

(4.7)

Where $Z_1 \sim N(0,1)$, $Z_{jk} \sim N(0,2^{-\alpha j})$ (for some constant $\alpha > 0$) and $Z_1, Z_{jk}, j \in \mathbb{N}_0, k = 1, \ldots, 2^j$, $R$ and $S$ are all independent random variables, $R$ satisfies eq. (4.5) and $S$ satisfies eq. (4.6). The prior distribution of $\theta$ is denoted by $\Pi$.

If the coefficients $Z_{jk}$ have standard deviation $2^{-j/2}$ (so $\alpha = 1$), the random draws from this prior are scaled piecewise linear interpolations on a dyadic grid of a Brownian bridge on $[0, 1]$ plus the random function $Z_1 \psi_1$, see for instance Bhattacharya and Waymire (2007, paragraph 10.1) for a similar process. The choice of $\psi_1$ is motivated by the fact that in this case $\text{Var} \theta(t)$ conditioned on $S = s$ and $R = \infty$ is constant $s^2$, independent of $t$.

4.3 Theorems on posterior contraction rates

The main result of this section, theorem 4.5, characterises the frequentist rate of contraction of the posterior probability around a fixed parameter $\theta_0$ of unknown smoothness using the truncated series prior from section 4.2.

We make the following assumption on the true drift function.
Assumption 4.2. The true drift $\theta_0$ can be expanded in the Faber-Schauder basis, $\theta_0 = z_1\psi_1 + \sum_{j=0}^{\infty} \sum_{k=1}^{2^j} z_{jk}\psi_{jk}$ and there exists a $\beta \in (0, \infty)$ such that
\[
\|\theta_0\|_\beta := |z_1| \vee \sup_{j \in \mathbb{Z}_{\geq 0}, k \in \{1, \ldots, 2^j\}} 2^{\beta j} |z_{jk}| < \infty. \tag{4.8}
\]
(Note that we use a slightly different symbol for the norm, as we denote the $L^2$-norm by $\| \cdot \|_2$.) A well known class of smooth functions that quantify the local oscillations of functions, is the space $B^{s}_{p,q}[0,1]$ of $B^{s}_{p,q}$-Besov smooth real valued functions on $[0,1]$ (or another domain). For the definition and main results, one consults for instance Giné and Nickl (2016, Section 4.3). In fact the space of functions with finite $\| \cdot \|_\beta$-norm, as defined in eq. (4.8) resembles closely the space of $B^{\beta}_{\infty,\infty}$-Besov smooth functions.

Lemma 4.3. When $\beta \in (0,2)$ and $\theta_0 \in L^2(T)$ is $B^{\beta}_{\infty,\infty}$-Besov smooth, then eq. (4.8) holds.

Proof. Let $\theta_0 = z_1\psi_1 + \sum_{j=0}^{\infty} \sum_{k=1}^{2^j} z_{jk}\psi_{jk}$ be the Faber-Schauder basis expansion of $\theta_0$. It follows from the definition of the basis functions that
\[
z_{jk} = \theta_0 \left( (2k-1)2^{-(j+1)} \right) - \frac{1}{2} \theta_0 \left( 2^{-(j+1)}(2k-2) \right) - \frac{1}{2} \theta_0 \left( 2^{-(j+1)}2k \right).
\]
Hence, it follows from equations (all in section 4.3 Giné and Nickl (2016)) (4.72) (with $r = 2$) and (4.73) (with $p = \infty$) in combination with equation (4.79) (with $q = \infty$) that eq. (4.8) of this chapter holds.

We obtain the following result for our prior.

Theorem 4.4. Assume $\theta_0$ satisfies assumption 4.2. Suppose the prior is defined by eq. (4.7) and satisfies assumption 4.1. Let $(\varepsilon_T)_{T>0}$ be a sequence of positive numbers that converges to zero as $T \to \infty$. There is a constant $C_1 > 0$ so that for $T$ large enough,
\[
\log \Pi (\|\theta - \theta_0\|_\infty < \varepsilon_T) \geq -C_1\varepsilon_T^{-1/\beta} |\log \varepsilon_T|,
\]
and for any $C_2 > 0$ there is a measurable set $\mathcal{B}_T \subseteq L^2(\mathbb{T})$ such that for every $\varepsilon > 0$ there is a positive constant $C_3$ so that for $T$ sufficiently large

$$\log \Pi (\theta \notin \mathcal{B}_T) \leq -C_2 \varepsilon^{-1/\beta} |\log \varepsilon_T|,$$

$$\log N(a\varepsilon_T, \{\theta \in \mathcal{B}_T : \|\theta - \theta_0\|_2 \leq \varepsilon_T\}, \|\cdot\|_\infty) \leq C_3 \varepsilon^{-1/\beta} |\log \varepsilon_T|.$$

The following theorem is obtained by combining these bounds with theorem 1.6 after taking $\varepsilon_T = (T / \log T)^{-\beta/(1+2\beta)}$.

**Theorem 4.5.** Let $X^T$ be observations of eq. (4.1), with $\theta$ replaced by $\theta_0$, where $\theta_0$ satisfies assumption 4.2 and in addition $\int_0^1 \theta_0(x)dx = 0$. Let the prior satisfy assumption 4.1. Then for some constant $M > 0$,

$$\mathbb{P}_{\theta_0} \Pi \left( \theta : \|\theta - \theta_0\|_2 \geq M \left( \frac{1}{\log T} \right)^{-\frac{\beta}{1+2\beta}} \left| X^T \right| \right) \to 0$$

as $T \to \infty$.

When a different function $\Lambda$ is used, defined on a compact interval of $\mathbb{R}$, and the basis elements are defined by $\psi_{jk} = \sum_{m \in \mathbb{Z}} \Lambda(2^j (x - m) + k - 1)$; forcing them to be one-periodic, then theorem 4.5 and derived results for applications still hold provided $\|\psi_{jk}\|_\infty = 1$ and $\psi_{j,k} \cdot \psi_{j,l} \equiv 0$ when $|k - l| \geq d$ for a fixed $d \in \mathbb{N}$ and the smoothness assumptions on $\theta_0$ are changed accordingly. A finite number of basis elements may be added or redefined as long as they are one-periodic.

### 4.4 Applications to nonparametric regression and density estimation

Our general results also apply to other models. The following results are obtained for $\theta_0$ satisfying assumption 4.2 and the prior eq. (4.7) satisfying assumption 4.1. As already remarked, a finite number of basis functions might be added or redefined, to constitute a basis that is more appropriate in the models considered.
4. Drift estimation using Faber-Schauder expansions

4.4.1 Nonparametric regression model

As a direct application of the properties of the prior shown in the previous section, we obtain the following result for a nonparametric regression problem. Assume

\[ X^n_i = \theta_0(i/n) + \eta_i, \quad 0 \leq i \leq n, \]

with independent Gaussian observation errors \( \eta_i \sim N(0, \sigma^2) \). When we apply Ghosal and van der Vaart (2007), example 7.7 to our theorem 4.4 we obtain, for some constant \( M > 0 \),

\[ \Pi \left( \theta : \|\theta - \theta_0\|_2 \geq M \left( \frac{n}{\log n} \right)^{-\frac{\beta}{1+2\beta}} \left| X^n \right| \right) \xrightarrow{\mathbb{P}_{\theta_0}} 0 \]
as \( n \to \infty \).

4.4.2 Nonparametric density estimation

Let us consider \( n \) independent observations \( X^n := (X_1, \ldots, X_n) \) with \( X_i \sim p_0 \) where \( p_0 \) is an unknown density on \([0,1]\) relative to the Lebesgue measure. Let \( \mathcal{P} \) denote the space of densities on \([0,1]\) relative to the Lebesgue measure. The natural distance for densities is the Hellinger distance \( h \) defined by

\[ h(p, q)^2 = \int_0^1 \left( \sqrt{p(x)} - \sqrt{q(x)} \right)^2 dx. \]

Define the prior on \( \mathcal{P} \) by \( p = e^{\theta \|e\|_1} \), where \( \theta \) is endowed with the prior of theorem 4.5 or its non-periodic version. Assume that \( \log p_0 \) is \( \beta \)-smooth in the sense of assumption 4.2. Applying Ghosal, Ghosh, and van der Vaart (2000), theorem 2.1 and van der Vaart and van Zanten (2008a), lemma 3.1 to our theorem 4.4, we obtain for some constant \( M > 0 \)

\[ \Pi \left( p \in \mathcal{P} : h(p, p_0) \geq M \left( \frac{n}{\log n} \right)^{-\frac{\beta}{1+2\beta}} \left| X^n \right| \right) \xrightarrow{\mathbb{P}_{p_0}} 0, \]
as \( n \to \infty \).
4.5 Proofs

4.5.1 Proof of theorem 4.4

By assumption 4.2 the true drift can be represented as

\[ \theta_0 = z_1 \psi_1 + \sum_{j=0}^{\infty} \sum_{k=1}^{2^j} z_{jk} \psi_{jk}. \]

For \( r \geq 0 \), define its truncated version by

\[ \theta_0^r = z_1 \psi_1 + \sum_{j=0}^{r} \sum_{k=1}^{2^j} z_{jk} \psi_{jk}. \]

**Small ball probability**

For \( \varepsilon > 0 \) choose an integer \( r_\varepsilon \) with

\[ C_\beta \varepsilon^{-1/\beta} \leq 2^{r_\varepsilon} \leq 2C_\beta \varepsilon^{-1/\beta} \]

where \( C_\beta = \left\lceil \frac{\left\lVert \theta_0 \right\rVert_{\beta}^{1/\beta}}{(2^\beta - 1)^{1/\beta}} \right\rceil. \) (4.9)

For notational convenience we will write \( r \) instead of \( r_\varepsilon \) in the remainder of the proof. By lemma 4.8 we have \( \| \theta_0^r - \theta_0 \|_\infty \leq \varepsilon \). Therefore

\[ \| \theta^{r,s} - \theta_0 \|_\infty \leq \| \theta^r - \theta_0 \|_\infty + \| \theta_0^r - \theta_0 \|_\infty \leq \| \theta^{r,s} - \theta^r_0 \|_\infty + \varepsilon \]

which implies

\[ \mathbb{P}(\| \theta^{r,s} - \theta_0 \|_\infty < 2\varepsilon) \geq \mathbb{P}(\| \theta^{r,s} - \theta^r_0 \|_\infty < \varepsilon). \]

Let \( f_S \) denotes the probability density of \( S \). For any \( x > 0 \), we have

\[ \mathbb{P}(\| \theta^{R,s} - \theta_0 \|_\infty < 2\varepsilon) \]

\[ = \sum_{r \geq 1} \mathbb{P}(R = r) \int_0^{\infty} \mathbb{P}(\| \theta^{r,s} - \theta_0 \|_\infty < 2\varepsilon) f_S(s) \, ds \]

\[ \geq \mathbb{P}(R = r) \inf_{s \in [L_\varepsilon, U_\varepsilon]} \mathbb{P}(\| \theta^{r,s} - \theta^r_0 \|_\infty < \varepsilon) \int_{L_\varepsilon}^{U_\varepsilon} f_S(s) \, ds, \] (4.10)

where

\[ L_\varepsilon = \varepsilon^{-\frac{p_1}{q_\beta}} \quad \text{and} \quad U_\varepsilon = \varepsilon^{-\frac{p_2}{q_\beta}} \]

and \( p_1, p_2 \) and \( q \) are taken from assumption 4.1. For \( \varepsilon \) sufficiently small, we have by the second part of assumption 4.1

\[ \int_{L_\varepsilon}^{U_\varepsilon} f_S(s) \, ds \geq \exp \left( -\varepsilon^{-\frac{1}{\beta}} \right). \]
4. Drift estimation using Faber-Schauder expansions

By choice of $r$ and the first part of assumption 4.1, there exists a positive constant $C$ such that

$$\mathbb{P}(R = r) \geq \exp\left(-c_2 2^r r\right) \geq \exp\left(-C\varepsilon^{-\frac{1}{2}} |\log \varepsilon|\right),$$

for $\varepsilon$ sufficiently small.

In what follows we bound the middle term in eq. (4.10) from below, uniformly over $[L_\varepsilon, U_\varepsilon]$. Recall that $\mathcal{I}_r$ denotes the set of indices $i = 1$ and $i = (j, k), j = 0, \ldots, r, k = 1, \ldots, 2^j$. Note that

$$\theta^{r,s} - \theta^r_0 = (sZ_1 - z_1)\psi_1 + \sum_{j=0}^{r} \sum_{k=1}^{2^j} (sZ_{jk} - z_{jk})\psi_{jk}$$

which implies

$$\|\theta^{r,s} - \theta^r_0\|_\infty \leq |sZ_1 - z_1| + \sum_{j=0}^{r} \max_{1 \leq k \leq 2^j} |sZ_{jk} - z_{jk}| \leq (r + 2) \max_{i \in \mathcal{I}_r} |sZ_i - z_i|.$$

This gives the lower bound

$$\mathbb{P}(\|\theta^{r,s} - \theta^r_0\|_\infty < \varepsilon) \geq \prod_{i \in \mathcal{I}_r} \mathbb{P}\left(|sZ_i - z_i| < \frac{\varepsilon}{r + 2}\right).$$

By choice of the $Z_i$, we have for all $i \in \mathcal{I}_r$, that $2^{\alpha \ell(i)}Z_i$ is standard normally distributed, where we recall that $\ell(i)$ denotes the level of function $\psi_i$ (so $\ell(1) = 0$ and $\ell((j, k)) = j$), and hence

$$\log \mathbb{P}\left(|sZ_i - z_i| < \frac{\varepsilon}{r + 2}\right)$$

$$= \log \mathbb{P}\left(|2^{\alpha \ell(i)}Z_i - 2^{\alpha \ell(i)}z_i/s| < \frac{2^{\alpha \ell(i)}\varepsilon}{(r + 2)s}\right)$$

$$\geq \log \left(\frac{2^{\alpha \ell(i)}\varepsilon}{(r + 2)s}\right) - \frac{2^{2\alpha \ell(i)}\varepsilon^2 s^2}{(r + 2)^2 s^2} - \frac{2^{2\alpha \ell(i)}z_i^2}{s^2} + \frac{1}{2} \log \left(\frac{2}{\pi}\right),$$

where the inequality follows from lemma 4.9. The third term can be further bounded as we have

$$2^{2\alpha \ell(i)}z_i^2 = 2^{2(\alpha - \beta)\ell(i)}/2^{2\beta \ell(i)}z_i^2 \leq 2^{2(\alpha - \beta)\ell(i)}\|\theta_0\|_\beta^2.$$
Hence
\[ \log \mathbb{P} \left( |sZ_i - z_i| < \frac{\varepsilon}{r+2} \right) \geq \log \left( \frac{2^{2\alpha \ell(i)\varepsilon}}{(r+2)s} \right) - \frac{2^{2\alpha r\varepsilon^2}}{(r+2)^2 s^2} - \frac{2^2(\alpha-\beta)\ell(i)\|\theta_0\|_\beta^2}{s^2} + \frac{1}{2} \log \left( \frac{2}{\pi} \right). \] (4.11)

For \( s \in [L_\varepsilon, U_\varepsilon] \) and \( i \in \mathcal{I} \), we will now derive bounds on the first three terms on the right of eq. (4.11). For \( \varepsilon \) sufficiently small we have \( r \leq r+2 \leq 2r \) and then eq. (4.9) implies
\[ \log C_\beta \leq r + 2 \leq 2 \log(2C_\beta) + \frac{2}{\beta} |\log \varepsilon|. \]

**Bounding the first term on the RHS of eq. (4.11).** For \( \varepsilon \) sufficiently small, we have
\[ \log \left( \frac{(r+2)s}{2^{2\alpha \ell(i)\varepsilon}} \right) \leq \log \left( \frac{(r+2)U_\varepsilon}{\varepsilon} \right) = \log \left( (r+2)\varepsilon^{-(1+\frac{p_2}{q^2})} \right) \leq \log \left\{ 2 \log(2C_\beta) + \frac{2}{\beta} |\log \varepsilon| \right\} + \left( 1 + \frac{p_2}{q^2} \right) |\log \varepsilon| \leq \tilde{C}_{p_2,q,\beta} |\log \varepsilon|, \]

where \( \tilde{C}_{p_2,q,\beta} \) is a positive constant.

**Bounding the second term on the RHS of eq. (4.11).** For \( \varepsilon \) sufficiently small, we have
\[ \frac{2^{2\alpha r\varepsilon^2}}{(r+2)^2 s^2} \leq \frac{2^{2\alpha r\varepsilon^2}}{(\log C_\beta)^2 L_\varepsilon^2} \leq \frac{(2C_\beta)^{2\alpha}}{(\log C_\beta)^2} \varepsilon^{2(\alpha+\beta+p_1/q)} \leq 1. \]
The final inequality is immediate in case \( \alpha = \beta \), else it suffices to verify that the exponent is positive under the assumption \( p_1 > q|\alpha - \beta| \).

**Bounding the third term on the RHS of eq. (4.11).** For \( \varepsilon \) sufficiently small, in case \( \beta \geq \alpha \) we have
\[ \frac{2^{2(\alpha-\beta)\ell(i)\|\theta_0\|_\beta^2}}{s^2} \leq \|\theta_0\|_\beta^2 L_\varepsilon^{-2} \leq 1. \]
In case \( \beta < \alpha \) we have
\[ \frac{2^{2(\alpha-\beta)\ell(i)\|\theta_0\|_\beta^2}}{s^2} \leq \frac{2^{2(\alpha-\beta)r\|\theta_0\|_\beta^2}}{L_\varepsilon^2} \leq (2C_\beta)^{2(\alpha-\beta)} \varepsilon^{\beta(\frac{p_1}{q}-1+\beta)} \leq 1. \]
as the exponent of $\varepsilon$ is positive under the assumption $p_1 > q|\alpha - \beta|$. Hence for $\varepsilon$ small enough, we have

$$\log \mathbb{P}\left( \left| sZ_i - z_i \right| < \frac{\varepsilon}{r + 2} \right) \geq -\tilde{C}_{p_2,q,\beta} |\log \varepsilon| - 3.$$ 

As $-2^{r+1} \geq -4C_\beta \varepsilon^{-1/\beta}$ we get

$$\log \inf_{s \in [x_1,x_2]} P (\|\theta^{r,s} - \theta_0^r\|_\infty < \varepsilon) \geq -4C_\beta \varepsilon^{-1/\beta} \left( \tilde{C}_{p_2,q,\beta} |\log \varepsilon| + 3 \right) \geq -\varepsilon^{-1/\beta} |\log \varepsilon|.$$ 

We conclude that the right hand side of eq. (4.10) is bounded below by $\exp \left( -C_1 \varepsilon^{-1/\beta} |\log \varepsilon| \right)$, for some positive constant $C_1$ and sufficiently small $\varepsilon$.

**Entropy and remaining mass conditions**

For $r \in \{0,1,\ldots\}$ denote by $\mathcal{C}_r$ the linear space spanned by $\psi_1$ and $\psi_{jk}$, $0 \leq j \leq r$, $k \in 1,\ldots,2^j$, and define

$$\mathcal{C}_{r,t} := \{ \theta \in \mathcal{C}_r, [\theta]_\alpha \leq t \}.$$ 

**Proposition 4.6.** For any $\varepsilon > 0$

$$\log N(\varepsilon, \mathcal{C}_{r,t}, \| \cdot \|_\infty) \leq 2^{r+1} \log (3A_\alpha t \varepsilon^{-1}),$$

where $A_\alpha = \sum_{k=0}^\infty 2^{-k\alpha}$.

**Proof.** We follow van der Meulen, van der Vaart, and van Zanten (2006, §3.2.2). Choose $\varepsilon_0, \ldots, \varepsilon_r > 0$ such that $\sum_{j=0}^r \varepsilon_j \leq \varepsilon$. Define

$$U_j = \begin{cases} 
[-2^{-\alpha j} t, 2^{-\alpha j} t]^{2^j} & \text{if } j \in \{1,\ldots, r\} \\
[-t,t]^{2^j} & \text{if } j = 0 
\end{cases}.$$ 

For each $j \in \{1,\ldots, r\}$, let $E_j$ be a minimal $\varepsilon_j$-net with respect to the max-distance on $\mathbb{R}^{2^j}$, that is a set $E_j$ of minimal cardinality that satisfies

$$U_j \subseteq \bigcup_{y \in E_j} \left\{ x \in \mathbb{R}^{2^j} : \|x - y\|_\infty \leq \varepsilon_j \right\},$$

for each $j \in \{1,\ldots, r\}$. Let $E := \bigcup_{j=0}^r E_j$.
and let $E_0$ be a minimal $\varepsilon_0$-net with respect to the max-distance on $\mathbb{R}^2$. Hence, if $x \in U_j$, then there exists a $e_j \in E_j$ such that $\max_k |x_k - e_k| \leq \varepsilon_j$.

Take $\theta \in \mathcal{C}_{r,t}$ arbitrary: $\theta = z_1 \psi_1 + \sum_{j=0}^{r} \sum_{k=1}^{2^j} z_{jk} \psi_{jk}$. Let $\tilde{\theta} = e_1 \psi_1 + \sum_{j=0}^{r} \sum_{k=1}^{2^j} e_{jk} \psi_{jk}$, where $(e_1, e_{0,1}) \in E_0$ and $(e_{j,1}, \ldots, e_{j,2^j}) \in E_j$ (for $j = 1, \ldots, 2^j$). We have

$$\|\theta - \tilde{\theta}\|_\infty \leq |z_1 - e_1| \|\psi_1\|_\infty + \sum_{j=0}^{r} \max_{1 \leq k \leq 2^j} |z_{jk} - e_{jk}| \|\psi_{jk}\|_\infty$$

$$\leq |z_1 - e_1| + \sum_{j=0}^{r} \max_{1 \leq k \leq 2^j} 2^j |2^{-j\alpha} z_{jk} - 2^{-j\alpha} e_{jk}|.$$ 

This can be bounded by $\sum_{j=0}^{r} \varepsilon_j$ by an appropriate choice of the coefficients in $\tilde{\theta}$. In that case we obtain that $\|\theta - \tilde{\theta}\|_\infty \leq \varepsilon$. This implies

$$\log N(\varepsilon, \mathcal{C}_{r,t}, \|\cdot\|_\infty) \leq \sum_{j=0}^{r} \log |E_j| \leq \sum_{j=0}^{r} 2^j \log \left(\frac{3 \cdot 2^{-ajt}}{\varepsilon_j}\right).$$

The asserted bound now follows upon choosing $\varepsilon_j = \varepsilon 2^{-j\alpha}/A_\alpha$.

**Proposition 4.7.** There exists a positive constant $K$ such that

$$\log N (a\varepsilon, \{\theta \in \mathcal{C}_r : \|\theta - \theta_0\|_2 \leq \varepsilon\}, \|\cdot\|_\infty) \leq 2^{r+1} \log \left(6\sqrt{3}A_{\alpha}K2^{(\alpha+1/2)}a^{-1}\varepsilon^{-1}\right).$$

**Proof.** There exists a positive $K$ such that

$$\{\theta \in \mathcal{C}_r : \|\theta - \theta_0\|_2 \leq a\varepsilon\} \subset \{\theta \in \mathcal{C}_r : \|\theta\|_2 \leq K\}.$$ 

By lemma 4.11 below, this set is included in the set

$$\left\{ \theta \in \mathcal{C}_r : \|\theta\|_\infty \leq \sqrt{3} \cdot 2^{(r+1)/2} K \right\}. \quad (4.12)$$

By lemma 4.10 below, for any $\theta = z_1 \psi_1 + \sum_{j=0}^{r} \sum_{k=1}^{2^j} z_{jk} \psi_{jk}$ in this set we have

$$\max \{|z_1|, |z_{jk}|, j = 0, \ldots, r, k = 1 \ldots, 2^j\} \leq 2\|\theta\|_\infty \sqrt{3} \cdot 2^{(r+1)/2} K.$$
Hence, the set eq. (4.12) is included in the set
\[ \{ \theta \in C_r : \| \theta \|_\alpha \leq a(r, \varepsilon) \} = \mathcal{C}_{r,a(r, \varepsilon)}, \]
where \( a(r, \varepsilon) = 2^{1+\alpha r} \sqrt{3} \cdot 2^{(r+1)/2} K. \)

Hence,
\[ N \left( a \varepsilon, \left\{ \theta \in C_r : \| \theta - \theta_0 \|_2 \leq \varepsilon \right\}, \| \cdot \|_\infty \right) \leq N \left( a \varepsilon, \mathcal{C}_{r,a(r, \varepsilon)}, \| \cdot \|_\infty \right). \]
The result follows upon applying proposition 4.6.

We can now finish the proof for the entropy and remaining mass conditions. Choose \( r_T \) to be the smallest integer so that \( 2^{r_T} \geq L \varepsilon_T^{-\frac{1}{\beta}} \), where \( L \) is a constant, and set \( \mathcal{B}_T = \mathcal{C}_{r_T} \). The entropy bound then follows directly from proposition 4.7.

For the remaining mass condition, using assumption 4.1, we obtain
\[ \mathbb{P} ( \theta^{R.S} \notin \mathcal{B}_T ) = \mathbb{P} ( R > r_T ) \leq \exp \left( -c_1 2^{r_T} r_T \right) \leq \exp \left( -C_2 \varepsilon_T^{-\frac{3}{2}} | \log \varepsilon_T | \right), \]
and note that the constant \( C_2 \) can be made arbitrarily large by choosing \( L \) large enough.

### 4.5.2 Lemmas

**Lemma 4.8.** Suppose \( z \) has Faber-Schauder expansion
\[ z = z_1 \psi_1 + \sum_{j=0}^{\infty} \sum_{k=1}^{2^j} z_{jk} \psi_{jk}. \]
If \( \| z \|_\beta < \infty \) (with the norm defined in eq. (4.8)), then for \( r \geq 1 \)
\[ \left\| z - \sum_{i \in I_r} z_i \psi_i \right\|_\infty \leq \| z \|_\beta \frac{2^{-r \beta}}{2^\beta - 1}. \] (4.13)

**Proof.** This follows from
\[ \left\| z - \sum_{i \in I_r} z_i \psi_i \right\|_\infty \leq \sum_{j=r+1}^{\infty} \left\| \sum_{k=1}^{2^j} z_{jk} \psi_{jk} \right\|_\infty \]
\[ = \sum_{j=r+1}^{\infty} 2^{-j \beta} \max_{1 \leq k \leq 2^j} 2^{j \beta} | z_{jk} | \leq \| z \|_\beta \sum_{j=r+1}^{\infty} 2^{-j \beta}. \]
\[ \square \]
Lemma 4.9. Let $X \sim N(0, 1)$, $\theta \in \mathbb{R}$ and $\epsilon > 0$. Then
\[
P(|X - \theta| \leq \epsilon) \geq \frac{e^{-\theta^2}}{\sqrt{2}} \frac{e^{\log \epsilon - \epsilon^2 + \log \sqrt{\pi}}}{\sqrt{2}}.
\]

Proof. Note that
\[
\int_{\theta - \epsilon}^{\theta + \epsilon} e^{-\frac{1}{2}x^2} dx = \int_{-\epsilon}^{\epsilon} e^{-\frac{1}{2}(x+\theta)^2} dx
\]
and
\[
\frac{e^{-\theta^2} e^{-\frac{1}{2}(x+\theta)^2}}{e^{-\frac{1}{2}(\sqrt{2}x)^2}} = e^{\theta^2 - \frac{1}{2}(x+\theta)^2 + x^2} = e^{\frac{1}{2}(x-\theta)^2} \geq 1,
\]
thus $e^{-\frac{1}{2}(x+\theta)^2} \geq e^{-\frac{1}{2}(x+\theta)^2} e^{-\frac{1}{2}(\sqrt{2}x)^2}$, hence
\[
\int_{\theta - \epsilon}^{\theta + \epsilon} e^{-\frac{1}{2}x^2} dx \geq e^{-\theta^2} \int_{-\epsilon}^{\epsilon} e^{-\frac{1}{2}(\sqrt{2}x)^2} dx = \frac{e^{-\theta^2}}{\sqrt{2}} \int_{-\sqrt{2}\epsilon}^{\sqrt{2}\epsilon} e^{-\frac{1}{2}u^2} du.
\]
Now the elementary bound $\int_{-y}^{y} e^{-\frac{1}{2}x^2} dx \geq 2 ye^{-\frac{1}{2}y^2}$ gives
\[
P(|X - \theta| \leq \epsilon) = \frac{1}{\sqrt{2\pi}} \int_{\theta - \epsilon}^{\theta + \epsilon} e^{-\frac{1}{2}x^2} dx
\]
\[
\geq \frac{1}{\sqrt{2\pi}} \frac{e^{-\theta^2}}{\sqrt{2}} \int_{-\sqrt{2}\epsilon}^{\sqrt{2}\epsilon} e^{-\frac{1}{2}u^2} du
\]
\[
= \frac{e^{-\theta^2}}{\sqrt{2}} \frac{\sqrt{2} \epsilon}{\sqrt{2}} \int_{-\epsilon}^{\epsilon} e^{-\frac{1}{2}x^2} dx
\]
\[
\geq \frac{1}{\sqrt{2\pi}} \frac{e^{-\theta^2}}{\sqrt{2}} 2 \sqrt{2} \epsilon e^{-\epsilon^2}
\]
\[
= e^{\log \epsilon - \epsilon^2 + \log \sqrt{\pi}}.
\]
\]

Lemma 4.10. Let
\[
f = z_1 \psi_1 + \sum_{j=1}^{r} \sum_{k=1}^{2^j} z_{j,k} \psi_{j,k}.
\]
Then
\[
\sup_{i : t(i) \leq r} |z_i| \leq 2 \|f\|_{\infty}.
\]
4. Drift estimation using Faber-Schauder expansions

Proof. Note that \(|z_1| = |f(0)| \leq 2\|f\|_{\infty}\), and \(|z_{0,1}| = |f(1/2)| \leq 2\|f\|_{\infty}\) and, for \(j \geq 1\), \(z_{jk} = f((2k - 1)2^{-(j+1)}) - \frac{1}{2}f(2^{-(j+1)}(2k - 2)) - \frac{1}{2}f(2^{-(j+1)}2k)\), from which follows that \(|z_{jk}| \leq 2\|f\|_{\infty}\).

Lemma 4.11. Let \(\mathcal{C}_r\) the linear space spanned by \(\psi_1\) and \(\psi_{jk}\), \(0 \leq j \leq r\), \(k \in \{1, \ldots, 2^j\}\). Then

\[
\sup_{0 \neq f \in \mathcal{C}_r} \frac{\|f\|_{\infty}}{\|f\|_2} \leq \sqrt{3} \cdot 2^{(r+1)/2}.
\]

Proof. Let \(f \in \mathcal{C}_r\) be nonzero. Note that for any constant \(c > 0\),

\[
\frac{\|cf\|_{\infty}}{\|cf\|_2} = \frac{\|f\|_{\infty}}{\|f\|_2}.
\]

Hence, we may and do assume that \(\|f\|_{\infty} = 1\). Furthermore, since the \(L^2\) and \(L^\infty\) norm of \(f\) and \(|f|\) are the same, and \(|f|\) is also in \(\mathcal{C}_r\), we may also assume that \(f\) is nonnegative.

Let \(x_0\) be a global maximiser of \(f\). Clearly \(f(x_0) = 1\). Since \(f\) is a linear interpolation between the points \(\{k2^{-j-1} : k = 0, 1, \ldots, 2^{r+1}\}\), we may choose \(x_0\) of the form \(x_0 = k2^{-j-1}\). We consider two cases (i) \(0 \leq k < 2^{r+1}\) and (ii) \(k = 2^{r+1}\). In case (i) we have that \(f(x) \geq (1 - 2^{r+1}(x-k2^{-r-1})) \mathbb{I}_{[k2^{-r-1},(k+1)2^{-r-1}])(x)\), for all \(x \in [k2^{-r-1},(k+1)2^{-r-1}]\). In case (ii) \(f(x) \geq 2^{r+1}(x-1+2^{-r-1}) \mathbb{I}_{[1-2^{-r-1},1]}(x)\), for all \(x \in [1-2^{-r-1},1]\). Hence, in both cases,

\[
\|f\|_2^2 \geq 2^{2r+2} \int_0^{2^{-r-1}} x^2 dx = \frac{1}{3} 2^{2r+2} 2^{-3r-3} = \frac{1}{3} 2^{-r-1}.
\]

Thus

\[
\frac{\|f\|_{\infty}}{\|f\|_2} \leq \frac{1}{\sqrt{3}} 2^{-(r+1)/2} = \sqrt{3} \cdot 2^{(r+1)/2},
\]

uniformly over all nonzero \(f \in \mathcal{C}_{r,s}\). \(\Box\)
Empirical Bayes selection of the scaling parameter in Gaussian process priors for drift estimation

In this chapter we study an empirical Bayes method for nonparametric estimation of periodic drift. We consider the same model as studied in chapter 1 and the Gaussian process prior of chapter 2, where we this time estimate the scaling parameter from the data and plug it in the prior, instead of equipping it with a distribution. Recall that we assume a continuous observation \( X^T = (X_t : t \in [0, T]) \) which is a weak solution to the stochastic differential equation

\[
dX_t = \theta_0(X_t)dt + dW_t,
\]

where \( W \) is a Brownian motion and the unknown parameter \( \theta_0 \) is a one-periodic real-valued function on \( \mathbb{R} \), \( \int_0^1 \theta_0(x)dx = 0 \) and \( \int_0^1 \theta_0(x)^2dx < \infty \). Equivalently \( \theta_0 \) can be seen as a square integrable function on the unit circle with mean zero. The space of functions with these properties is denoted by \( \dot{L}^2(\mathbb{T}) \).

Diffusions have many applications in science and mathematical finance. For instance, in computational chemistry eq. (5.1) is used to
model the angle between planes spanned by atoms in a molecule, see Papaspiliopoulos et al. (2012). In this paper a Gaussian process prior on the unknown drift is proposed and a numerical scheme to sample from the posterior is given. Their prior is basically defined as follows

$$\theta = s \sum_{k=1}^{\infty} k^{-\alpha - 1/2} Z_k \phi_k,$$

(5.2)

where $s > 0$ is a constant (referred to as the scaling parameter), $(\phi_k)_{k=1}^{\infty}$ is an orthonormal basis of $L^2(T)$, and $(Z_k)_{k=1}^{\infty}$ is a sequence of independent standard normally distributed random variables. An example of such an orthonormal basis is the Fourier basis defined by

$$\phi_{2k-1}(x) = \sqrt{2} \sin(2\pi k x),$$

$$\phi_{2k}(x) = \sqrt{2} \cos(2\pi k x),$$

(5.3)

with $k \in \mathbb{N}$. In Pokern, Stuart, and van Zanten (2013) posterior consistency is shown for this prior. In chapter 2 optimal rates for this prior are derived, when the smoothness of the prior coincides with the smoothness of the true function. Adaptation to the (generally unknown) smoothness of true function can be achieved by randomly scaling the Gaussian process. It is shown that the prior hierarchically defined as

$$\theta \mid s = s \sum_{k=1}^{\infty} k^{-\alpha - 1/2} Z_k \phi_k,$$

$$s \sim G,$$

where $G$ is a certain probability measure on $\mathbb{R}_{>0}$, attains the optimal posterior convergence rate $T^{-\beta/(1+\beta)}$ as long as $0 < \beta \leq \alpha + 1/2$, when the true function is $\beta$-Sobolev smooth (when using the Fourier basis).

In this chapter we choose the empirical Bayes approach to select the scaling parameter. We use the marginal maximum likelihood estimator to estimate the ‘optimal’ scaling parameter $s$ of the Gaussian process prior $\Pi_s$ defined by eq. (5.2) from the data and use the posterior with the estimated hyperparameter $s$ plugged in for the inference. We derive the same rate of convergence as in chapter 2. The marginal likelihood is defined as

$$s \mapsto \int p_\theta(X^T) d\Pi_s(\theta).$$
5.1 Main result

The marginal maximum likelihood estimator (MMLE) is the argument that maximises the marginal likelihood over a set Λ of positive numbers. Clearly Λ should contain an s for which the posterior \( \Pi_s(\cdot \mid X^T) \) attains the optimal convergence rate. To simplify computations, we choose Λ to be finite, which simplifies the proofs quite substantially, compared to Donnet et al. (2018) and Rousseau and Szabó (2017). This is because we don’t need to control the supremum of \( p_\theta(X^T)d\Pi_s(\theta) \) over small intervals of s. But choosing Λ finite might also give insight in how to choose this set in applications.

Sufficient conditions for posterior convergence for data-driven priors are recently studied in a general setting in the papers Donnet et al. (2018) and Rousseau and Szabó (2017). We follow their general approach in this chapter, and we adapt it to the setting of diffusions, using ideas of van der Meulen, van der Vaart, and van Zanten (2006).

The ideas of Donnet et al. (2018) and Rousseau and Szabó (2017) are the following. Find a subset \( \Lambda_0 \) of Λ where the rate of contraction of \( \Pi_s(\cdot \mid X^T) \) is good. Then show that the estimator belongs to this set with probability converging to one, and finally show that on the event where the estimator is in \( \Lambda_0 \), the desired rate of contraction is attained.

However, the general approach of Donnet et al. (2018) and Rousseau and Szabó (2017) cannot be applied directly, as the test functions in the diffusion setting are in the random metric

\[
h(\theta, \eta) = \sqrt{\int_0^T (\theta(X_t) - \eta(X_t))^2 dt}, \quad \theta, \eta \in \dot{L}^2(\mathbb{T}).
\]

(Note that \( h \) depends on \( T \).) Hence more effort needs to be done, to relate this metric to the deterministic \( L^2 \)-metric.

In the next section the prior and the marginal maximum likelihood estimator are formulated, and the rate of contraction result is stated, together with its proof. The proof is divided in several lemmas whose proofs are in sections 5.2 and 5.3. Section 5.4 contains some auxiliary results.

5.1 Main result

We consider the Gaussian prior \( \Pi_s \) defined by eq. (5.2). In order for a prior to have good asymptotic properties, it should put enough prior...
mass around the true parameter. Failing to do so may lead to suboptimal rates, see Castillo (2008). Therefore it is natural to choose $\Pi_s$ so that it puts most mass around the true parameter $\theta_0$, out of all priors $\Pi_s, s \in \Lambda$. But as $\theta_0$ is unknown, as well as its smoothness, we will define an estimator for this $s$. The hyperparameter $s$ is estimated from the data in the following way. The marginal likelihood is

$$\int p_\theta(X^T) d\Pi_s(\theta),$$

where

$$p_\theta(X^T) = \exp \left\{ \int_0^T \theta(X_t) dX_t - \frac{1}{2} \int_0^T \theta(X_t)^2 dt \right\}$$

is the (Girsanov) density relative to the Wiener measure (see lemma 1.2). It follows from lemma 1.4 that this integral is well defined. The estimator for $s$ will be the value over a properly chosen set which maximises the marginal likelihood.

By the Plancherel formula, every function $\theta \in \dot{L}^2(\mathbb{T})$ can be expanded in the chosen orthonormal basis, $\theta = \sum_{k=1}^{\infty} \theta_k \phi_k, \theta_k = \langle \theta, \phi_k \rangle$. We say that $\theta$ is $\beta$-smooth, $\beta > 0$, when $\sum_{k=1}^{\infty} k^{2\beta} \theta_k^2 < \infty$, and the space of these functions is denoted by $\dot{H}_\beta(\mathbb{T})$. When $(\phi_k)$ is the Fourier basis eq. (5.3), this is the usual $\beta$-Sobolev space. We denote the set over which we maximise the the marginal likelihood by $\Lambda$. We will show that $\Pi_s$ puts most prior mass around the true parameter when $s \cong T^{\alpha-\beta}/4t+\alpha}$, where $\beta$ is the smoothness of the true parameter, as long as $0 < \beta \leq \alpha + 1/2$. This motivates the following choice of $\Lambda$,

$$\Lambda = \left\{ kT^{-\frac{1}{4t+\alpha}} : k \in \mathbb{N}, kT^{-\frac{1}{4t+\alpha}} \leq T^\alpha \right\}.$$ 

(Note that $\Lambda$ depends on $T$.) We show in section 5.4 that this set contains an element $s$ satisfying $\frac{1}{2} T^{\frac{\alpha-\beta}{1+2\beta}} \leq s \leq T^{\frac{\alpha-\beta}{1+2\beta}}$, for every $0 < \beta \leq \alpha + 1/2$. With slight changes to the proofs, one could choose any other grid $\Lambda$ as long at the number of elements grows at most polynomially with $T$ and for every $0 < \beta \leq \alpha + 1/2$ it contains a parameter proportional to $T^{\frac{\alpha-\beta}{1+2\beta}}$.

The MMLE $\hat{s}$ is defined as

$$\hat{s} = \arg \max_{s \in \Lambda} \int p_\theta(X^T) d\Pi_s(\theta).$$

(5.4)
For fixed $s$ the posterior of a measurable set $A \subseteq L^2(\mathbb{T})$ is given by

$$\Pi_s(A \mid X^T) = \frac{\int_A p_\theta(X^T)d\Pi_s(\theta)}{\int p_\theta(X^T)d\Pi_s(\theta)}. \quad (5.5)$$

It follows from lemma 1.4 that the posterior eq. (5.5) is well defined.

We study posterior contraction rates of the plug-in posterior

$$\Pi_{\hat{s}}(\cdot \mid X^T) = \Pi_s(\cdot \mid X^T)\bigg|_{s=\hat{s}},$$

under the law $\mathbb{P}_{\theta_0}$ of the process with the true drift function. We obtain the following result.

**Theorem 5.1.** Let $\alpha > 1/2$. When $\theta_0$ is $\beta$-smooth and $\beta \in (0, \alpha+1/2]$, then, for some constant $M > 0$, we have

$$\Pi_{\hat{s}}\left(\theta \in L^2(\mathbb{T}) : \|\theta - \theta_0\| \leq MT^{-\frac{\beta}{1+2\beta}} \mid X^T\right) \xrightarrow{\mathbb{P}_{\theta_0}} 1, \text{ as } T \to \infty.$$

**Proof.** Through the equivalence of the measures $\mathbb{P}_\theta$, for fixed $T$, (lemma 1.2) the posterior eq. (5.5) can be written as

$$\Pi_s(A \mid X^T) = \frac{\int_A p_\theta(X^T)/p_{\theta_0}(X^T)d\Pi_s(\theta)}{\int p_\theta(X^T)/p_{\theta_0}(X^T)d\Pi_s(\theta)},$$

and

$$\bar{p}_\theta(X^T) := \frac{p_\theta(X^T)}{p_{\theta_0}(X^T)}$$

$$= \exp\left\{\int_0^T (\theta(X_t) - \theta_0(X_t))dW_t - \frac{1}{2} \int_0^T (\theta(X_t) - \theta_0(X_t))^2 dt\right\} \quad (5.6)$$

is the density of $\mathbb{P}_\theta$ relative to $\mathbb{P}_{\theta_0}$ (see lemma 1.2). The $s$ that maximises $\int p_\theta(X^T)d\Pi_s(\theta)$ over $\Lambda$, maximises

$$\int \bar{p}_\theta(X^T)d\Pi_s(\theta) \quad (5.7)$$

over $\Lambda$ as well, as $\bar{p}_\theta(X^T) = p_\theta(X^T)/p_{\theta_0}(X^T)$ differs from $p_\theta(X^T)$ only by a multiplicative constant not depending on $\theta$ or $s$. For technical reasons we introduce a positive constant $K$. It follows from Dunker, Lifshits, and Linde (1998, Theorem 3.1) and elementary algebra that
for every \( \theta_0 \in L^2(\mathbb{T}) \) and positive \( s \) and \( T \) there is a unique \( \varepsilon_s \) (depending on \( T \)) so that

\[
\Pi_s(\|\theta - \theta_0\|_2 < K\varepsilon_s) = e^{-T\varepsilon_s^2}. \tag{5.8}
\]

Let

\[
\varepsilon_0 = \min_{s \in \Lambda} \varepsilon_s. \tag{5.9}
\]

(Which depends on \( T \).) One can think of \( e^{-T\varepsilon_0^2} \) as the most amount of mass one of the priors \( \Pi_s, s \in \Lambda \) can put in an small ball around the true parameter, and thus favours the true parameter the most of all priors \( \Pi_s, s \in \Lambda \). For Gaussian process priors, the amount of prior mass that a prior puts around the true parameter is closely related to the rate of convergence (see for instance van der Vaart and van Zanten (2008a) and Castillo (2008)). In eq. (5.15) below we show that \( \varepsilon_0 \lesssim T^{-\frac{\beta}{1+2\beta}} \), which is (up to a constant) the optimal rate. We define for a constant \( L > 1 \)

\[
\Lambda_0 = \{ s \in \Lambda : \varepsilon_s \leq L\varepsilon_0 \}. \tag{5.10}
\]

Thus \( \Lambda_0 \) are all \( s \) of \( \Lambda \) where \( \Pi_s \) puts enough prior mass around \( \theta_0 \). This set is nonrandom, but depends on the true parameter. In section 5.2 we show that for \( L \) large enough, there exist an event \( F \) on which \( \hat{s} \in \Lambda_0 \), and whose \( \mathbb{P}_{\theta_0} \)-probability converges to one. In section 5.3 we show that for some \( L' > 0 \),

\[
\mathbb{E}_{\theta_0} \left[ \Pi_{\hat{s}}(\|\theta - \theta_0\|_2 \geq LL'\varepsilon_0 \mid X^T) \mathbb{I}_F \right] \to 0
\]

as \( T \to \infty \). The result follows from

\[
\mathbb{E}_{\theta_0} \left[ \Pi_{\hat{s}}(\|\theta - \theta_0\|_2 \geq LL'\varepsilon_0 \mid X^T) \right] \\
\leq \mathbb{E}_{\theta_0} \left[ \Pi_{\hat{s}}(\|\theta - \theta_0\|_2 \geq LL'\varepsilon_0 \mid X^T) \mathbb{I}_F \right] + \mathbb{P}_{\theta_0}(F^c) \to 0 \text{ (as } T \to \infty),
\]

and \( \varepsilon_0 \lesssim T^{-\frac{\beta}{1+2\beta}}. \)
5.2 The estimator is with high probability in a favourable set

It follows from Pokern, Stuart, and van Zanten (2013, Theorem 4.1(ii)) that with probability converging to one, the random metric

\[ h(\theta, \theta') = \sqrt{\int_0^T (\theta(X_s) - \theta'(X_s))^2 ds} \]

and the square integrable \((L^2)\) norm are equivalent. For any constants \(0 < c_\rho < 1 < C_\rho\) satisfying

\[ c_\rho^2 \leq \inf_{x \in [0,1]} \rho(x) \leq \sup_{x \in [0,1]} \rho(x) < C_\rho^2 \]

with

\[ \rho(x) = \frac{\exp \left\{ 2 \int_0^x \theta_0(y) dy \right\}}{\int_0^1 \exp \left\{ 2 \int_0^y \theta_0(z) dz \right\} dy} \]

the \(P_{\theta_0}\)-probability of the event

\[ E = \left\{ c_\rho \sqrt{T} \|\theta - \theta'\|_2 \leq h(\theta, \theta') \leq C_\rho \sqrt{T} \|\theta - \theta'\|_2, \forall \theta, \theta' \in L^2(\mathbb{T}) \right\} \]

converges to 1.

In the next lemma we show that we can restrict \(E\) even more, so that asymptotically this set still has mass one, and \(\hat{s} \in \Lambda_0\) on this event.

**Lemma 5.2.** For every \(T > 0\), there is an \(s_0 \in \Lambda_0\) so that \(\varepsilon_{s_0} = \varepsilon_0\). If \(\alpha > 1/2\), for \(K\) and \(A > 1\) large enough and \(B > A\), and \(L > \sqrt{B}\) in eq. (5.10), the \(P_{\theta_0}\)-probability of the event

\[ F := E \cap \left\{ \max_{s \in \Lambda \setminus \Lambda_0} \int \tilde{p}_\theta(X^T) d\Pi_s(\theta) \leq e^{BT\varepsilon_0^2} \right\} \]

\[ \cap \left\{ \int \tilde{p}_\theta(X^T) d\Pi_{s_0}(\theta) \geq e^{-AT\varepsilon_0^2} \right\}, \]

converges to one, as \(T\) grows to infinity. On this event \(\hat{s} \in \Lambda_0\).
5. Empirical Bayes selection of the scaling parameter

Proof. The measurability of $F$ follows from lemma 1.4 and the fact that $\Lambda_0$ and $\Lambda \setminus \Lambda_0$ are finite.

As $\Lambda$ is finite, the minimum $\varepsilon_0$ in eq. (5.9) is actually attained for some $s_0 \in \Lambda$. From eq. (5.10) follows that $s_0 \in \Lambda_0$. As $\hat{s}$ maximises the quantity $\int \bar{p}_\theta(X^T)d\Pi_s$ over all $s \in \Lambda$, we have in particular

$$\int \bar{p}_\theta(X^T)d\Pi_s(\theta) = \int \bar{p}_\theta(X^T)d\Pi_{s_0}(\theta).$$

On the event $F$,

$$\int \bar{p}_\theta(X^T)d\Pi_{s_0}(\theta) \geq e^{-AT\varepsilon_0^2}$$

$$> e^{-BT\varepsilon_0^2} \geq \sup_{s \in \Lambda \setminus \Lambda_0} \int \bar{p}_\theta(X^T)d\Pi_s(\theta).$$

Hence $\hat{s}$ belongs to $\Lambda_0$ on the event $F$.

Using van der Meulen, van der Vaart, and van Zanten (2006, Lemma 4.2) we derive

$$\mathbb{P}_{\theta_0} \left( \left\{ \int \bar{p}_\theta(X^T)d\Pi_{s_0}(\theta) > e^{-(C^2\rho^2K^2+1)T\varepsilon_0^2} \right\} \cap E \right)$$

$$\geq \mathbb{P}_{\theta_0}(E) - \exp \left\{ -\frac{1}{8} C^2\rho^2 T\varepsilon_0^2 \right\}.$$

It follows from eq. (5.16) below that the last term of the last expression converges to zero, and we already saw that $\mathbb{P}_{\theta_0}(E) \to 1$ (see eq. (5.11)). Hence the left hand side converges to one.

Let $B > (C^2\rho^2K^2 + 1)$ be a constant. Let us now concentrate on the event

$$E \cap \left\{ \max_{s \in \Lambda \setminus \Lambda_0} \int \bar{p}_\theta(X^T)d\Pi_s(\theta) > e^{-BT\varepsilon_0^2} \right\}. \quad (5.13)$$

We have to show that the $\mathbb{P}_{\theta_0}$-probability of eq. (5.13) converges to zero. Note that the number of elements in $\Lambda \setminus \Lambda_0$ is bounded by $\frac{T^{\alpha+\frac{1}{4+4\alpha}}}{T^{-\frac{1}{3+4\alpha}}}$.

66
5.2. The estimator is with high probability in a favourable set

We have

\[
\mathbb{P}_{\theta_0}\left(\left\{ \max_{s \in \Lambda \setminus \Lambda_0} \int \bar{p}_\theta(X^T) d\Pi_s(\theta) > e^{-BT\varepsilon_0^2} \right\} \cap E \right) \\
\leq T^{\alpha + \frac{1}{2 + 4\alpha}} \max_{s \in \Lambda \setminus \Lambda_0} \mathbb{P}_{\theta_0}\left(\left\{ \int \bar{p}_\theta(X^T) d\Pi_s(\theta) > e^{-BT\varepsilon_0^2} \right\} \cap E \right).
\]

It is only left to show that

\[
\max_{s \in \Lambda \setminus \Lambda_0} \mathbb{P}_{\theta_0}\left(\left\{ \int \bar{p}_\theta(X^T) d\Pi_s(\theta) > e^{-BT\varepsilon_0^2} \right\} \cap E \right) = o\left( T^{\alpha - \frac{1}{2 + 4\alpha}} \right),
\]
as \(T \to \infty\). Let \(s \in \Lambda \setminus \Lambda_0\). Let \(\varphi_s\) be the test function of lemma 5.4 below, with \(\varepsilon = \varepsilon_s\) and \(U = K\) and \(\Theta_s\) the corresponding sieves \(\Theta_s\). We have

\[
\mathbb{P}_{\theta_0}\left(\left\{ \int \bar{p}_\theta(X^T) d\Pi_s(\theta) > e^{-BT\varepsilon_0^2} \right\} \cap E \right) \\
\leq \mathbb{E}_{\theta_0} \varphi_s(X^T) \\
+ \mathbb{E}_{\theta_0} \left[ \int \bar{p}_\theta(X^T) d\Pi_s(\theta) > e^{-BT\varepsilon_0^2} \right] (1 - \varphi_s(X^T)) \mathbb{I}_E.
\]

For the first term we have \(\mathbb{P}_{\theta_0} \varphi_s(X^T) \leq e^{-C_1 K^2 \varepsilon_0^2} \leq e^{-C_1 K^2 L^2 \varepsilon_0^2}\), using that for \(s \in \Lambda \setminus \Lambda_0\), \(\varepsilon_s \geq L\varepsilon_0\). After applying first the Markov inequality and then Fubini’s theorem, we see that the second term is bounded by

\[
e^{BT\varepsilon_0^2} \int \mathbb{E}_\theta \left[ (1 - \varphi_s(X^T)) \mathbb{I}_E \right] d\Pi_s(\theta).
\]

Using the defining property of \(\varepsilon_s\), eq. (5.8), and lemma 5.4, we obtain

\[
\int \mathbb{E}_\theta \left[ (1 - \varphi_s(X^T)) \mathbb{I}_E \right] d\Pi_s(\theta) \\
\leq \mathbb{I}_s(\theta \in \hat{L}^2(\mathbb{T}) : \|\theta - \theta_0\|_2 < K\varepsilon_s) \\
+ \int_{\theta \in \Theta_s : \|\theta - \theta_0\|_2 \geq K\varepsilon_s} \mathbb{E}_\theta \left[ (1 - \varphi_s(X^T)) \mathbb{I}_E \right] d\Pi_s(\theta) \\
+ \mathbb{I}_s(\hat{L}^2(\mathbb{T}) \setminus \Theta_s) \\
\leq e^{-T\varepsilon_0^2} + e^{-C_2 K T \varepsilon_0^2} + e^{-K^2 T \varepsilon_0^2} \\
\leq 3e^{-L^2 T \varepsilon_0^2},
\]

67
for $K \geq C_2^{-1} \lor 1$ (note that $C_2$ does not depend on $K$) and using that $\varepsilon_s \geq L\varepsilon_0$. Since $T^{\alpha+\frac{1}{4+4\alpha}}/e^{cT\varepsilon_0^2} \rightarrow 0$ as $T \rightarrow \infty$, for every $c > 0$, we have that for $L > \sqrt{B}$ the $\mathbb{P}_{\theta_0}$-probability of eq. (5.13) converges to zero, which completes the proof of the theorem.

5.3 Posterior convergence on the favourable event

In the previous section it is shown that with $\mathbb{P}_{\theta_0}$-probability converging to one, the estimator $\hat{s}$ is in the set $\Lambda_0$ of $s$ values where the prior $\Pi_s$ puts most prior mass around the true parameter out of all priors $\Pi_s, s \in \Lambda$. In this section we we show that on this event, the posterior $\Pi_{\hat{s}}(\cdot | X^T)$ attains the rate of convergence $\varepsilon_0$, up to a constant.

**Theorem 5.3.** Let $F$ be the event eq. (5.12). For some constant $L' > 0$,

$$
\mathbb{E}_{\theta_0} \left[ \Pi_{\hat{s}}(\|\theta - \theta_0\|_2 \geq L'L\varepsilon_0 | X^T) \mathbb{I}_F \right]
$$

converges to zero, as $T \rightarrow \infty$.

**Proof.** By the definitions of $\hat{s}$ (eq. (5.4)) and $F$ (eq. (5.12)) it follows that on the event $F$,

$$
\int \tilde{p}_\theta(X^T)d\Pi_{\hat{s}}(\theta) \geq \int \tilde{p}_\theta(X^T)d\Pi_{s_0}(\theta) \geq e^{-AT\varepsilon_0^2}.
$$

Hence, on $F$,

$$
\Pi_{\hat{s}}(\|\theta - \theta_0\|_2 \geq LL'\varepsilon_0 | X^T) = \frac{\int_{\|\theta - \theta_0\|_2 \geq LL'\varepsilon_0} \tilde{p}_\theta(X^T)d\Pi_{\hat{s}}(\theta)}{\int \tilde{p}_\theta(X^T)d\Pi_{\hat{s}}(\theta)} \leq e^{AT\varepsilon_0^2} \int_{\|\theta - \theta_0\|_2 \geq LL'\varepsilon_0} p_\theta(X^T)d\Pi_{\hat{s}}(\theta).
$$

Similar as in section 5.2 one can show that $\#\Lambda_0 \leq T^{\alpha+\frac{1}{4+4\alpha}}$. Using that $L\varepsilon_0 \geq \varepsilon_s$, for every $s \in \Lambda_0$ and using that $\hat{s} \in \Lambda_0$ on $F$, we can
bound for a constant $A' > A$,

$$\mathbb{P}_{\theta_0} \left( \left\{ \int_{\|\theta - \theta_0\|_2 \geq LL'\varepsilon_0} p_\theta(X^T) d\Pi_\hat{s}(\theta) > e^{-A'T\varepsilon_0^2} \right\} \cap F \right)$$

$$\leq \mathbb{P}_{\theta_0} \left( \left\{ \max_{s \in \Lambda_0} \int_{\|\theta - \theta_0\|_2 \geq LL'\varepsilon_0} p_\theta(X^T) d\Pi_\hat{s}(\theta) > e^{-A'T\varepsilon_0^2} \right\} \cap F \right)$$

$$\leq T^{\alpha + \frac{1}{4+4\alpha}} \cdot \max_{s \in \Lambda_0} \mathbb{P}_{\theta_0} \left( \left\{ \int_{\|\theta - \theta_0\| \geq L'\varepsilon_s} p_\theta(X^T) d\Pi_\hat{s}(\theta) > e^{-A'T\varepsilon_0^2} \right\} \cap F \right).$$

Let $s \in \Lambda_0$. Let $\varphi_s$ be the test functions of lemma 5.4 below with $s = s, \varepsilon = \varepsilon_s$ and $U = L' \geq K$, which exist for $K$ large enough. With a similar calculation as in section 5.2, one can show that

$$\mathbb{P}_{\theta_0} \left( \left\{ \int_{\|\theta - \theta_0\| \geq L'\varepsilon_s} p_\theta(X^T) d\Pi_\hat{s}(\theta) > e^{-A'T\varepsilon_0^2} \right\} \cap F \right)$$

$$\leq \mathbb{E}_{\theta_0} \varphi_s(X^T) + e^{A'T\varepsilon_0^2} \int_{\|\theta - \theta_0\| \geq L'\varepsilon_s} \mathbb{E}_{\theta} \left[ (1 - \varphi_s(X^T)) I_F \right] d\Pi_\hat{s}(\theta)$$

$$\leq e^{-C_1L'^2T\varepsilon_s^2} + e^{A'T\varepsilon_0^2} \left( e^{-C_2L'^2T\varepsilon_0} + e^{-L'^2T\varepsilon_s^2} \right).$$

Using that $\varepsilon_s \geq \varepsilon_0$ we bound the latter by

$$e^{-C_1L'^2T\varepsilon_0} + e^{A'T\varepsilon_0^2} \left( e^{-C_2L'^2T\varepsilon_0} + e^{-L'^2T\varepsilon_s^2} \right).$$

Using that $T^{\alpha + \frac{1}{4+4\alpha}}/c^{T\varepsilon_0^2} \to 0$ as $T \to \infty$, for every $c > 0$, we obtain for $A' > A$ and $L' > K \lor \sqrt{\frac{A'}{C^2\Lambda_1}}$

$$\mathbb{E}_{\theta_0} \left[ \Pi_\hat{s}(\|\theta - \theta_0\| \geq LL'\varepsilon_0 \mid X^T) I_F \right] \to 0,$n

as $T \to \infty$.

\[ 5.4 \quad \text{Auxiliary results} \]

In this section we show a few auxiliary results that are used in the proofs of this section. First, we use the small ball results for $\Pi_\hat{s}$ from
chapter 2 to derive upper and lower bounds for $\varepsilon_s$. As announced in section 5.1 we will prove the existence of an $s \in \Lambda$ for which $\frac{1}{2} T^{\frac{\alpha-\beta}{1+2\beta}} \leq s \leq T^{\frac{\alpha-\beta}{1+2\beta}}$ and $\varepsilon_0 \lesssim T^{-\frac{\alpha}{1+2\beta}}$. We will also show that $T\varepsilon_0^2 \rightarrow \infty$ as $T \rightarrow \infty$. In the second part of this section we will use these bounds, together with results from chapter 2 to prove the existence of the right sieves and test functions which we use in sections 5.2 and 5.3.

For $\theta_0 \in \dot{H}^\beta(\mathbb{T})$, $0 < \beta \leq \alpha + 1/2$, there are positive constants $D_1, D_2$ and $D_3$, so that for $\varepsilon/s > 0$ small enough,

$$
\Pi_s(\|\theta - \theta_0\|_2 < \varepsilon) \in \left[ \exp \left\{ -D_1 \left( D_2 \left( \frac{s}{\varepsilon} \right)^{1/\alpha} + \frac{1}{s^2} \varepsilon^{\frac{2\beta-2\alpha-1}{\beta}} \right) \right\} , \exp \left\{ -D_3 \left( \frac{s}{\varepsilon} \right)^{1/\alpha} \right\} \right],
$$

where the lower bound is the content of lemma 2.6 and the upper bound follows from lemma 2.4 and van der Vaart and van Zanten (2008b, lemma 5.3). Hence

$$
\exp \left\{ -D_3 \left( \frac{s}{\varepsilon} \right)^{1/\alpha} \right\} \geq \Pi_s(\|\theta - \theta_0\|_2 < K\varepsilon_s) = e^{-T\varepsilon_s^2}.
$$

It follows that

$$
T\varepsilon_s^2 \geq D_3 \left( \frac{s}{(K\varepsilon_s)} \right)^{1/\alpha} \Leftrightarrow \varepsilon_s \geq D_3^{\frac{\alpha}{1+2\alpha}} K^{-\frac{1}{1+2\alpha}} T^{-\frac{\alpha}{1+2\alpha}} s^{\frac{1}{1+2\alpha}}.
$$

We also have

$$
\exp \left\{ -D_1 \left( D_2 \left( \frac{s}{(K\varepsilon_s)} \right)^{1/\alpha} + \frac{1}{s^2} (K\varepsilon_s)^{\frac{2\beta-2\alpha-1}{\beta}} \right) \right\} \leq \Pi_s(\|\theta - \theta_0\|_2 < K\varepsilon_s) = e^{-T\varepsilon_s^2}.
$$

Hence

$$
D_1 \left( D_2 \left( \frac{s}{(K\varepsilon_s)} \right)^{1/\alpha} + \frac{1}{s^2} (K\varepsilon_s)^{\frac{2\beta-2\alpha-1}{\beta}} \right) \geq T\varepsilon_s^2.
$$

So either

$$
D_1 D_2 \left( \frac{s}{(K\varepsilon_s)} \right)^{1/\alpha} \geq T\varepsilon_s^2/2 \quad \text{or} \quad D_1 \frac{1}{s^2} (K\varepsilon_s)^{\frac{2\beta-2\alpha-1}{\beta}} \geq T\varepsilon_s^2/2.
$$

The first case is equivalent to

$$
\varepsilon_s \leq (D_1 D_2)^{1+2\alpha} K^{-\frac{1}{1+2\alpha}} T^{-\frac{\alpha}{1+2\alpha}} s^{\frac{1}{1+2\alpha}},
$$
the second case to
\[ \varepsilon_s \leq (2D_1)^{\frac{\beta}{1+2\alpha}} K^{\frac{2\beta-2\alpha-1}{1+2\alpha}} T^{-\frac{\beta}{1+2\alpha}} s^{-\frac{2\beta}{1+2\alpha}}. \]

Combing the results gives
\[ D_3^{\frac{\alpha}{1+2\alpha}} K^{-\frac{1}{1+2\alpha}} T^{-\frac{\alpha}{1+2\alpha}} s^{\frac{1}{1+2\alpha}} \leq \varepsilon_s \]
\[ \leq (D_1D_2)^{\frac{\alpha}{1+2\alpha}} K^{-\frac{1}{1+2\alpha}} T^{-\frac{\alpha}{1+2\alpha}} s^{\frac{1}{1+2\alpha}} \sqrt{(2D_1)^{\frac{\beta}{1+2\alpha}} K^{\frac{2\beta-2\alpha-1}{1+2\alpha}} T^{-\frac{\beta}{1+2\alpha}} s^{-\frac{2\beta}{1+2\alpha}}} \]
(5.14)

Using eq. (5.14) we see that the best possible upper bound for \( \varepsilon_0 \) (up to a constant) is attained when
\[ T^{-\frac{\alpha-\beta}{1+2\beta}} \lesssim s^{-\frac{2\beta}{1+2\alpha}} T^{-\frac{\beta}{1+2\alpha}}, \]
that is, when \( s \approx T^{1+2\beta} \). Note that on the interval \((0, \alpha + 1/2]\) the quantity
\[ T^{-\frac{1}{4+4\alpha}} \leq T^{\frac{\alpha-\beta}{1+2\beta}} \leq T^\alpha. \]

Write
\[ T^{\frac{\alpha-\beta}{1+2\beta}} = T^{\frac{\alpha-\beta}{1+2\beta}} + \frac{1}{4+4\alpha} T^{-\frac{1}{4+4\alpha}}. \]
Since \( \frac{\alpha-\beta}{1+2\beta} + \frac{1}{4+4\alpha} \geq 0 \), we have for \( T \geq 1 \), \( \left[ T^{\frac{\alpha-\beta}{1+2\beta}} + \frac{1}{4+4\alpha} \right] \in \mathbb{N} \) and,
\[ \frac{1}{2} T^{\frac{\alpha-\beta}{1+2\beta}} + \frac{1}{4+4\alpha} \leq \left[ T^{\frac{\alpha-\beta}{1+2\beta}} + \frac{1}{4+4\alpha} \right] \leq T^{\frac{\alpha-\beta}{1+2\beta}} + \frac{1}{4+4\alpha} \]

hence there is a \( k \in \mathbb{N} \) so that \( kT^{-\frac{1}{4+4\alpha}} \leq T^\alpha \) and
\[ \frac{1}{2} T^{\frac{\alpha-\beta}{1+2\beta}} \leq kT^{-\frac{1}{4+4\alpha}} \leq T^{\frac{\alpha-\beta}{1+2\beta}}. \]

Using the upper bound of eq. (5.14) we have in particular
\[ \varepsilon_0 \lesssim T^{-\frac{\beta}{1+2\beta}}. \] (5.15)

Using the lower bound of eq. (5.14), we see that for all \( s \in \Lambda \),
\[ T \varepsilon_s^2 \geq T \varepsilon_0^2 \lesssim T^{\frac{1}{\alpha+2\beta}} \to \infty, \text{ as } T \to \infty. \] (5.16)

We now show that good test functions exist, with the associated sieves which satisfy the right remaining mass condition. As \( \varepsilon_s \) is implicitly defined, we will make use of the lower bound for \( \varepsilon_s \) derived in the first part of this section.
5. Empirical Bayes selection of the scaling parameter

**Lemma 5.4.** Let $\alpha > 1/2$. There are positive constants $C_1, C_2$ and $K$, only depending on $c_\rho$ and $C_\rho$, so that for every $s \in \Lambda$ and $U \geq K$ there are measurable sets (sieves) $\Theta_s \subseteq \hat{L}^2(\mathbb{T})$ satisfying

$$\Pi_s(\hat{L}^2(\mathbb{T}) \setminus \Theta_s) \leq e^{-U^2 T \varepsilon_s^2},$$

and measurable maps $\varphi_s : C[0, T] \to \{0, 1\}$ which satisfy,

$$\mathbb{E}_{\theta_0} \varphi_s(X^T) \leq e^{-C_1 U^2 T \varepsilon_s^2},$$

and for all $\theta \in \Theta_s$, $\|\theta - \theta_0\|_2 \geq U \varepsilon_s$,

$$\mathbb{E}_{\theta}[(1 - \varphi_s(X^T)) I_E] \leq e^{-C_2 U^2 T \varepsilon_s^2}.$$  

**Proof.** It follows from van der Meulen, van der Vaart, and van Zanten (2006, Lemma 4.1) and the equivalence of the $h$- and the $L^2$-metric on $E$ (eq. (5.11)) that there are constants $C_1$, $C_2$ and $c > 0$, only depending on $c_\rho$ and $C_\rho$, so that when $\Theta_s$ is a measurable set satisfying

$$\log N \left( \frac{c_\rho \varepsilon_s}{8 C_\rho}, \{\theta \in \Theta_s : \|\theta - \theta_0\|_2 \leq \varepsilon_s\}, \| \cdot \|_2 \right) \leq c^2 U^2 T \varepsilon_s^2,$$

for $U \geq K$, then there are measurable maps $\varphi_s$ (depending on $s$, $T$ and $U$) taking values in $\{0, 1\}$ which satisfy eqs. (5.18) and (5.19).

We continue by showing that such $\Theta_s$ actually exist and satisfy eq. (5.17) as well. For this we follow section 2.5.2. Every $\theta \in \hat{L}^2(\mathbb{T})$ has an expansion $\theta = \sum_{k=1}^{\infty} \theta_k \phi_k$ in the chosen orthonormal basis of $\hat{L}^2(\mathbb{T})$. Recall that $\hat{H}_1^{\alpha+1/2}$ is the set of $\theta \in \hat{L}^2(\mathbb{T})$ for which $\sum_{k=1}^{\infty} k^{2\alpha+1} \theta_k^2 \leq 1$ and $\hat{L}^2_1(\mathbb{T})$ is the closed unit ball in $\hat{L}^2(\mathbb{T})$. We take the same sieves as in section 2.5.2,

$$\Theta_s = R \hat{H}_1^{\alpha+1/2}(\mathbb{T}) + \frac{c_\rho \varepsilon_s}{16 C_\rho} \varepsilon_s \hat{L}^2_1(\mathbb{T}).$$

Then for some constant $C > 0$ (only depending on $\alpha$, $c_\rho$ and $C_\rho$),

$$\Pi_s(\hat{L}^2(\mathbb{T}) \setminus \Theta_s) \leq \exp \left\{ -\frac{1}{2} \left( \frac{R}{s} - \sqrt{C(s/\varepsilon_s)^{1/\alpha}} \right)^2 \right\}.$$
Let
\[ R = s \left( U \sqrt{2T \varepsilon_s} + \sqrt{C(s/\varepsilon_s)^{1/\alpha}} \right), \]
then \( \Pi_s(\hat{L}^2(\mathbb{T}) \setminus \Theta_s) \leq e^{-U^2T\varepsilon_s^2} \), hence eq. (5.17) is satisfied. For some constant \( \tilde{C} > 0 \) (depending on \( \alpha, c_\rho \) and \( C_\rho \)) the entropy is bounded as follows

\[ \log N \left( \frac{c_\rho s \varepsilon_s}{8C_\rho}, \{ \theta \in \Theta_s : \| \theta - \theta_0 \|_2 \leq \varepsilon_s \}, \| \cdot \|_2 \right) \leq \tilde{C} \left( \frac{R}{\varepsilon_s} \right)^{2/(1+2\alpha)}. \]

This is bounded, for some constant \( \tilde{C} > 0 \), by
\[
\tilde{C} \left( \frac{2}{s^{2+2\alpha} U^{2+2\alpha} T^{-\frac{1}{1+2\alpha}} \sqrt{s^{1/\alpha} \varepsilon_s^{-1/\alpha}}} \right)
= \tilde{C} T \varepsilon_s^2 \left( \frac{2}{s^{2+2\alpha} U^{2+2\alpha} T^{-\frac{2\alpha}{1+2\alpha} \varepsilon_s^{-2}} \sqrt{T^{-1} s^{1/\alpha} \varepsilon_s^{-1+2\alpha}}} \right)
\]
Inserting the lower bound of eq. (5.14) gives this is bounded, for some constant \( \tilde{C} > 0 \), by
\[
\tilde{C} T \varepsilon_s^2 \left( (U K)^{\frac{2}{1+2\alpha}} \sqrt{K^{1/\alpha}} \right).
\]
The latter is, for \( \alpha > 1/2 \), for any constant \( c > 0 \), for \( K \) large enough and for any \( U \geq K \), bounded by \( cU^2 T \varepsilon_s^2 \), which establish eq. (5.20).
5. Empirical Bayes selection of the scaling parameter
Dankwoord

Dit proefschrift dat hier voor u ligt is het resultaat van vier jaar hard werken. Ik beschouw deze vier jaar als de mooiste van mijn leven tot nu toe. Ik heb veel geleerd en ik heb mooie resultaten bereikt. Ik voel mij zeer gezegend dat ik deze kans heb gehad en ik ben God dankbaar voor deze tijd aan de UvA.

Ik wil mijn promotor, Harry van Zanten, heel hartelijk bedanken voor de plezierige samenwerking. Harry ik vond het prettig om met jou te werken en ik heb veel geleerd van jouw wijze van onderzoek doen en na elke afspraak ging ik weg met veel nieuwe ideeën voor mijn onderzoek.

Frank en Moritz bedankt voor de goede samenwerking dat resulteerde in een artikel in *Statistical Inference for Stochastic Processes* (chapter 4 in gewijzigde vorm van dit proefschrift).

Ik wil ook al mijn collega’s van het KDVI bedanken voor de fijne werksfeer. In het bijzonder mijn kamergenoten Alice, Fabian, Guusje, Jarno, Julia en Nanang. Ik wil Alice en Jarno bedanken voor de onvergetelijke dertigste verjaardag die jullie mij bezorgd hebben in Parijs bij het *Aquaboulevard*. Ik wil mijn lieve ouders bedanken dat ze altijd voor mij klaar staan en ook mijn lieve zussen, zwagers, oma’s, familie en vrienden.
Dankwoord
In chapter 2 we show that Gaussian process priors obtain the optimal rate of convergence when the smoothness of the prior coincides with that of the parameter that generates the data. This confirms the results found in van der Vaart and van Zanten (2008a) and Castillo (2008). We also show that adaptivity, with optimal rates, can be obtained by equipping the hyperparameters with a probability distribution. We even show adaptivity to every Sobolev smoothness by equipping the baseline smoothness parameter $\alpha$ with a certain probability distribution. However, the choice for the hyperparameter distributions is limited, and does for instance not include the popular inverse gamma distribution for the scaling parameter $S$. In Rousseau and Szabó (2017) it is shown (for different models) that adaptive contraction rates for hierarchically defined Gaussian process priors can be obtained as byproduct of the empirical Bayes approach, even allowing for inverse gamma scaling. This approach might be worthwhile to investigate in our context as well.

The results from chapters 3 and 4 are surprising and confirm the strong results for randomly truncated series priors in Shen and Ghosal (2015). Unlike Shen and Ghosal (2015) we also allow for inverse gamma scaling, due to the fact that we check the entropy only for a small ball in a sieve around the true parameter, instead of using the entropy of the whole sieve as is usually done (for instance in chapter 2). The small ball and entropy results in chapter 4 are derived in the stronger supremum norm, which make it possible to apply the results to nonparametric density estimation as well.

As far as we know, chapter 5 develops the first theoretical results for a nonparametric empirical Bayes method for a diffusion model. We show adaptive optimal rates for the empirical Bayes posterior. Our proofs are easier than those in Donnet et al. (2018) and Rousseau and
Szabó (2017), due to the fact that we choose a discrete parameter set (the set $\Lambda$) for the scaling factor, which has the additional benefit to advise on how to choose this set in applications, as one would maximise the marginal maximum likelihood over a discrete set. Further simulation studies need to be done to investigate the (computational) advantages and disadvantages of empirical Bayes methods compared to hierarchical Bayes methods.
Diffusions have many applications in science and can be described with a stochastic differential equation (SDE). We consider the following SDE, which was for example used in molecular dynamics (see e.g. Papaspiliopoulos et al. (2012)),

\[ dX_t = \theta(X_t)dt + dW_t, \]  

where \( \theta \) is measurable, one-periodic and \( \int_0^1 \theta(x)^2 dx < \infty \). We are interested in estimating \( \theta \) from observations \( (X_t : t \in [0, T]) \) of eq. (\( \star \)).

We study the posterior rates of contraction for several nonparametric Bayesian methods for diffusions. For Gaussian process priors we derive optimal posterior contraction rates, when the smoothness of the Gaussian process coincides with the smoothness of the target drift function. Adaptivity to the unknown smoothness is achieved by random scaling of the Gaussian process prior, or by equipping the baseline smoothness hyperparameter with a hyperprior.

We derive good adaptive posterior contraction results for priors defined as randomly truncated series priors. We consider expansions in orthonormal bases and in the Faber-Schauder basis, both with inverse gamma scaling. We also study the empirical Bayes approach to selecting the scaling parameter of the Gaussian process prior. Here the parameter is estimated from the data and plugged into the prior. Adaptive optimal contraction rates for the associated posterior are derived.
Samenvatting

“Adaptive posterior contraction rates for diffusions”

Diffusies hebben vele toepassingen in de wetenschap en kunnen beschreven worden door stochastische differentiaalvergelijkingen. Vaak is men geïnteresseerd in het schatten van de onbekende driftfunctie. In dit proefschrift bestuderen wij niet-parametrische Bayesiaanse methoden voor een stochastische differentiaalvergelijking, welke bijvoorbeeld gebruikt werd in moleculaire dynamica (zie bijvoorbeeld Papaspiliopoulos e.a. (2012)) en gegeven wordt door

$$dX_t = \theta(X_t)dt + dW_t,$$  

waar \(\theta\) meetbaar is met periode één en \(\int_0^1 \theta(x)^2 dx < \infty\) en \(W\) een Brownse beweging is. We zijn geïnteresseerd in het schatten van \(\theta\) met behulp van de waarneming \((X_t : t \in [0, T])\) van eq. (\(\star\)). Wij bestuderen de a-posteriori convergentiesnelheid van verschillende niet-parametrische Bayesiaanse methoden voor diffusies. Voor Gaussische processen priors bewijzen wij optimale a-posteriori convergentiesnelheden wanneer de gladheid van het Gaussische proces overeenkomt met de gladheid van de te schatten driftfunctie. Het aanpassingsvermogen van de a-posteriori verdeling aan de onbekende gladheid is bereikt door stochastische schaling van de Gaussische proces prior, of door de hyperparameter die de gladheid van de Gaussische proces prior bepaald te voorzien van een extra specifieke prior.

We bereiken goede adaptieve a-posteriori convergentiesnelheden voor priors die gedefinieerd zijn als stochastisch afgekapte reeks priors.
Samenvatting

We beschouwen reeks priors met orthonormale basissen en met de Faber-Schauder basis, beide met inverse gamma schaling. We bestuderen ook de empirisch Bayes methode om de optimale schaling van de Gaussische proces prior te schatten. Hierbij wordt de optimale schalingsparameter geschat uit de data en vervolgens ingeplugd in de prior. Optimale adaptieve a-posteriori convergentiesnelheden worden bereikt voor de geassocieerde a-posteriori-verdeling.
Overzicht van de bijdragen van de co-auteurs

**Chapter 2** J. van Waaij (kandidaat) onderzoek en eerste opzet voor artikel, J.H. van Zanten discussie en ontwikkeling van het concept, begeleiding en uiteindelijke versie van het artikel.

**Chapter 3** J. van Waaij (kandidaat) onderzoek en eerste opzet artikel, J.H. van Zanten begeleiding en discussie van het project en uiteindelijke versie van het artikel.

**Chapter 4** M.R. Schauer en J. van Waaij (kandidaat) onderzoek, F.H. van der Meulen discussie en begeleiding van het concept met enkele bijdragen aan het onderzoek. Alle auteurs droegen in gelijke wijze bij tot het totstandkomen van het artikel.
Bibliography


Index

\[B_\infty^\beta, 48\]
\[C_p, 65\]
\[E, 65\]
\[F, 65\]
\[H_\beta[0, 1], 32\]
\[L, 17\]
\[L^2(\mathbb{T}), 4\]
\[T^{\alpha+\beta}, 62, 71\]
\[W^{\alpha,L}, 22\]
\[Z_1, 47\]
\[Z_j, 33\]
\[Z_k, 17, 60\]
\[Z_{jk}, 47\]
\[\Lambda(x), 46\]
\[\Lambda, 62\]
\[\Lambda_0, 64\]
\[\mathbb{P}^x, 1\]
\[\mathbb{P}_0, 5\]
\[\mathbb{P}_\theta, 5\]
\[\alpha, 17, 33\]
\[\beta\text{-Sobolev norm}, 32\]
\[\beta\text{-Sobolev space}, 32, 62\]
\[\dot{H}_\beta(\mathbb{T}), 17\]
\[\dot{H}_\beta(\mathbb{T}), 62\]
\[\dot{L}^2(\mathbb{T}), 4\]
\[\ell(i), 47\]
\[\varepsilon_0, 64\]
\[\varepsilon_s, 64\]
\[\text{bounds for}, 71\]
\[\hat{s}, 62\]
\[\lambda_T(x), 20\]
\[\llbracket \cdot \rrbracket_\beta, 48\]
\[\text{relationship with} \quad B_\infty^\beta\text{-Besov norm}, 48\]
\[\tilde{p}_\theta(X^T), 63\]
\[\mathcal{I}_r, 47\]
\[\phi_k, 17, 60\]
\[\psi_1, 46\]
\[\text{motivation for}, 47\]
\[\psi_{j,k}, 46\]
\[\rho, 65\]
\[\tau_y, 1\]
\[\tau_{a,b}, 1\]
\[c_p, 65\]
\[g, 33\]
\[h, 11, 65\]
\[p, 33\]
\[s, 60\]

Brownian bridge, 47

continuous observations, 7

contraction rate, 9

general theorem, 10

convergence rate, see

contraction rate
covering number, 10

density estimation, 50
diffusion, 1, 3
  applications, 6
  in natural scale, 2
  periodic, 7–12, 36
  regular, 1
  representation by an SDE, 3
drift, 3

drifting, 3

empirical Bayes
  existence of sieves and tests, 72
  remaining mass, 72
  sieves, 73
entropy, 10
  empirical Bayes, 72–73
  Gaussian process prior with hyperprior on $\alpha$, 27–28
  Gaussian process prior with hyperprior on $L$, 26
  of sieve intersected with ball around the true function, 31
randomly truncated and scaled prior with Faber-Schauder basis, 54–56
randomly truncated and scaled prior with orthonormal basis, 40
Sobolev balls, 26
exit probability, 2

Faber-Schauder basis, 46
motivation for, 44

Gaussian white noise, 35

Green’s function, 2

inverse covariance operator, 14

level, 47
local time, 2
  law of large numbers, 6
  periodic, 5

marginal likelihood, 62
marginal maximum likelihood estimator, 62
MMLE, 62

nonparametric density estimation, 50
nonparametric regression, 50
normal distribution
  small ball probability, 57

occupation times formula, 2
plug-in posterior, 63
posterior, 8
  well defined, 8
precision operator, 14
prior, 8
  Gaussian process prior, 18
  on the Gaussian process regularity, 20–21
  on the scale, 18–20, 33, 46
  on truncation point, 33, 46
randomly truncated and scaled prior with Faber-Schauder basis, 47
randomly truncated and scaled prior with orthonormal basis, 33
prior mass, 10
empirical Bayes, 64, 69–71
Gaussian process prior, 22, 23
Gaussian process prior with hyperprior on $\alpha$, 26–27
Gaussian process prior with hyperprior on $L$, 24
randomly truncated and scaled prior with Faber-Schauder basis, 51
randomly truncated and scaled prior with orthonormal basis, 36
random metric, 65
regression, 50
regularity parameter, 17
remaining mass, 10
Gaussian process prior with hyperprior on $\alpha$, 27–28
Gaussian process prior with hyperprior on $L$, 25–26
randomly truncated and scaled prior with Faber-Schauder basis, 56
randomly truncated and scaled prior with orthonormal basis, 40
reproducing kernel Hilbert space, 22
RKHS, 22
scaling constant, 17
scaling parameter, 60
SDE, 3
  applications, 6
density, 6
explosion time, 3
initial value, 3
periodic drift, 4–6
unique in law, 3
weak solution, 3
weak uniqueness, 3
sieve
  empirical Bayes, 72–73
Gaussian process prior with hyperprior on $\alpha$, 27
Gaussian process prior with hyperprior on $L$, 25
randomly truncated and scaled prior with Faber-Schauder basis, 56
randomly truncated and scaled prior with orthonormal basis, 40
small ball, see prior mass
Sobolev norm, 32
Sobolev space, 32
speed density, 4
speed measure
  SDE, 4
stochastic differential equation, see SDE
time change, 2
white noise, 35