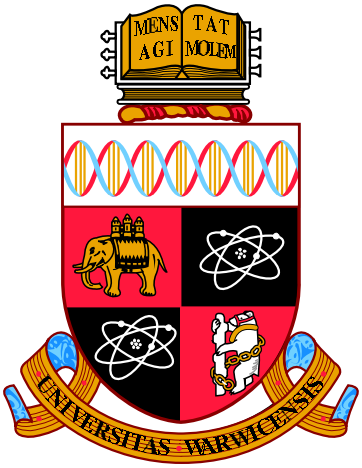


Efficient MCMC and Posterior Consistency for Bayesian Inverse Problems

by

Sebastian Josef Vollmer

Thesis

Submitted to the University of Warwick

for the degree of

Doctor of Philosophy

Mathematical Institute

September 2013

THE UNIVERSITY OF WARWICK

To my parents, Rosi and Alfred, and my fiancée Michaela

Contents

List of Figures

Acknowledgements

This work has been developed during my four years as postgraduate student at the University of Warwick. When I arrived in Warwick in 2009, I actually planned to stay in the UK only for one year before finishing my MSc degree in Germany. However, during this time, I got to know Professor Andrew Stuart who introduced me to the research area of Bayesian inverse problems which has intrigued me ever since. Therefore I did not have to think about it for long when he offered me to stay for four years to participate in the MSc programme and to work as a PhD student under his supervision. The research environment at the University of Warwick and the research topic are great and I am deeply grateful to Andrew for providing this opportunity for me.

In 2010, Professor Martin Hairer joined Professor Andrew Stuart in supervising my dissertation. I would like to thank both of them for the great collaboration and all the fruitful discussions during the last three years. Moreover, I acknowledge all the possibilities the two of them have offered me to participate in conferences and workshops around the world. I have learned so much during this time which would not have been possible without the encouragement of Andrew and Martin.

I am also most grateful for the opportunity to visit Professor Andreas Eberle in Bonn in 2011. Thank you, Andreas for our great discussions and your hospitality in Bonn.

Warm thanks go also to my colleagues and friends that I have met in Warwick. In particular, I would like to mention Hendrik Weber, Krys Łatuszyński and Andrew Duncan who were always receptive to any question. Moreover, I am grateful to my PhD colleagues Andrew Duncan, Dayal Strub, David Kelly, Sergios Agapiou and Alex Thiéry for making my stay at Warwick worthwhile. I thank Dayal especially for our climbing breaks which I really enjoyed.

I would also like to thank my new office mate, Rémi Bardenet, for giving me such a warm welcome to Oxford. Rémi has not only introduced me to all of my new colleagues

in Oxford but he, as well as Andrew Duncan, also helped me with some proof reading of this thesis. Thanks for all of your comments!

Last but not least I would like to thank my parents Rosi and Alfred for always supporting me in what I am doing. Words fail me to express my gratitude to my fiancée for her love, her support and the endless proof reading of this thesis. Without her, this thesis would not have been possible.

Sebastian J. Vollmer

Declaration

This thesis has been submitted to the University of Warwick in support of my application for the degree of Doctor of Philosophy. I hereby declare that it is entirely my own work except for the collaborative research contained in the research Articles **I** and **IV** which are presented in Part B of this thesis. My individual contributions to these four research articles are indicated in detail in the next paragraph on page ix.This thesis has not been submitted in application for a degree at another university.

Sebastian J. Vollmer

List of the Original Research Articles and the Author's Contributions

In this thesis, we present the results of four original research articles written by the author and his collaborators. In the following, we would like to give some fundamental details about these articles and a detailed description of the author's contributions.

I **Martin Hairer, Andrew M. Stuart and Sebastian J. Vollmer**, 2011. Spectral Gaps for a Metropolis-Hastings Algorithm in Infinite Dimensions. *Accepted with minor revisions by the Annals of Applied Probability, 39 pages.*

II **Sebastian J. Vollmer**, 2013. Dimension-Independent MCMC Sampling for Inverse Problems with Non-Gaussian Priors. *Submitted to the SIAM/ASA Journal on Uncertainty Quantification, 22 pages.*

III **Sebastian J. Vollmer**, 2013. Posterior Consistency for Bayesian Inverse Problems through Stability and Regression Results. *Submitted to Inverse Problems, 38 pages.*

IV **Andrew M. Stuart and Sebastian J. Vollmer**, 2013. Notes on a Bayesian Multiscale Elliptic Inverse Problem, *in Preparation, 28 pages.*

The research presented in Articles **II** and **III** is entirely the author's own work. Moreover, the author played an important role for the development of the central ideas for Article **I**. He is responsible for almost all technical details and the writing of this research article. With regard to Article **IV**, the author participated in the development of the key ideas, carried out most of the technical details. He is responsible for all computations and the writing of the article. All four original research articles are contained in Part B of this thesis.

Abstract

Many mathematical models used in science and technology often contain parameters that are not known a priori. In order to match a model to a physical phenomenon, the parameters have to be adapted on the basis of the available data. One of the most important statistical concepts applied to inverse problems is the Bayesian approach which models the a priori and a posteriori uncertainty through probability distributions, called the prior and posterior, respectively. However, computational methods such as Markov Chain Monte Carlo (MCMC) have to be used because these probability measures are only given implicitly. This thesis deals with two major tasks in the area of Bayesian inverse problems: the improvement of the computational methods, in particular, different kinds of MCMC algorithms, and the properties of the Bayesian approach to inverse problems such as posterior consistency.

In inverse problems, the unknown parameters are often functions and therefore elements of infinite dimensional spaces. For this reason, we have to discretise the underlying problem in order to apply MCMC methods to it. Finer discretisations lead to a higher dimensional state space and usually to a slower convergence rate of the Markov chain. We study these convergence rates rigorously and show how they deteriorate for standard methods. Moreover, we prove that slightly modified methods exhibit a dimension independent performance constituting one of the first dimension independent convergence results for locally moving MCMC algorithms.

The second part of the thesis concerns numerical and analytical investigations of the posterior based on artificially generated data corresponding to a true set of parameters. In particular, we study the behaviour of the posterior as the amount of data increases or the noise in the data decreases. Posterior consistency describes the phenomenon that a sequence of posteriors concentrates around the truth. In this thesis, we present one of the first posterior consistency results for non-linear infinite dimensional inverse problems. We also study a multiscale elliptic inverse problem in detail. In particular, we show that it is not posterior consistent but the posterior concentrates around a manifold.

List of Abbreviations

Abbreviation	Full name	Page
(BRP)	Bayesian Regression Problem	101
(EIP)	Elliptic Inverse Problem	14
gPC	general Polynomial Chaos algorithm	16,31
ESJD	Expected Squared Jump Distance	46,64
(IP)	Inverse Problem	101
IS	Independence Sampler	39,43
MALA	Metropolis-Adjusted-Langevin algorithm	37,40
MAP	Maximum Posterior Estimator	11,26
MCMC	Markov Chain Monte Carlo algorithm	33
MCQMC	Markov Chain Quasi Monte Carlo algorithm	30
MSE	Mean Square Error	47
pCN	Preconditioned Crank-Nicolson algorithm	44
pCNL/PIA	Preconditioned Crank-Nicolson-Langevin algorithm	45
QMC	Quasi Monte Carlo algorithm	27,30
RRWM	Reflection Random Walk Metropolis algorithm	36,77
RSRWM	Reflection Standard Random Walk Metropolis algorithm	19,83
RURWM	Revlection Uniform Random Walk Metropolis algorithm	83
RWM	Random Walk Metropolis algorithm	37,40
SMC	Sequential Monte Carlo algorithm	27,30

Chapter 1

Introduction

In many areas of science and technology, it is often the case that important parameters cannot directly be observed in an experiment. A prime example is x-ray computed tomography. For diagnostic purposes, the patient lies down in an x-ray tube, x-rays are sent through the body and their intensity is measured at the rim of the tube. In fact, using the laws of physics, we can construct a mathematical model, known as forward model, that allows us to predict the intensity of the x-rays at the rim of the tube on the basis of certain input data. However, in order to produce tomographic images, the interest lies in the properties of the body and therefore in the reconstruction of the input data from these measurements. Mathematically, this is known as the corresponding inverse problem. This thesis is devoted to the mathematical theory for a particular approach to inverse problems, called the Bayesian approach. The key idea is that not all parameter choices are equally likely. Instead, the a priori uncertainty is modelled as a probability measure called the prior. Given the forward model and the distribution of the observational noise, the parameters and the data can be treated as jointly varying random variables. Under mild assumptions on the quantities involved, the conditional distribution of the parameters, given the data, exists and can be represented as an unnormalised density with respect to the prior. This distribution is called the posterior and is an update of the prior using the data modelling the a posteriori uncertainty. Compared to the classical regularisation approach, that estimates the parameters as the minimiser of an appropriate functional, this approach has three major advantages in the

infinite dimensional case:

1. The assumptions are clearly modelled in the prior.

2. The posterior is continuous in the data.

3. A precise quantification of uncertainty is given.

On the basis of this concept, many interesting mathematical questions arise. For which models does the posterior exist and how can it be represented? Can the difference between the infinite dimensional posterior and the posterior based on a finite dimensional model be bounded in some way? How efficient are computational tools in approximating posterior expectations, the posterior mean or the MAP estimator for a finite dimensional model? If the data is generated for a fixed parameter, does the posterior concentrate around this parameter as the amount of data tends to infinity?

In this work, we discuss the results presented in the research Articles **I**, **II**, **III** and **IV** which are contained in Part B of this thesis. These results address the last two questions. All four articles are concerned with either the theory of Bayesian inverse problems or the computational methods that can be applied to it. The contributions of this work fall naturally into three categories:

1. The ability to approximate posterior expectations is crucial in order to make inference about the parameters with respect to which the problem is formulated. This is not straightforward as the posterior is only given as unnormalised density with respect the prior. Evaluating the density is computationally expensive as each time of the forward model has to be simulated. Markov Chain Monte Carlo (MCMC) algorithms are one of the most important computational methods that can be used to approximate posterior expectations. The approximation is obtained as the sample average of the steps of the corresponding Markov Chain.We study how the number of necessary samples depends on the dimension of the discretised state space.

2. We assume that the data is generated by the model for a fixed parameter called the truth. It is then of interest to show that the posterior concentrates around the

truth. We study properties of the prior and the fixed parameter that guarantee this property of the model, known as posterior consistency, to hold.

3. For an inverse problem with a multiscale structure, we consider the problem of characterising different multiscale structures giving rise to the same effective problem and how it can be used to approximate the posterior.

In the remainder of this chapter, we summarise the work presented in the research Articles **I**, **II**, **III** and **IV** contained in Part B of this thesis before we give an outline of the structure of this work and a more detailed exposition to the framework of Bayesian inverse problems.

1.1 Summary of the Original Research Articles

Bayesian inverse problems often involve continuum forward models. For numerical simulations, a discretised version on a high dimensional state space is used. Articles **I** and **II** address the question of how the performance of MCMC methods depends on an increase of the dimension. The research Article **III** studies the posterior consistency properties of nonlinear inverse problems describing the asymptotic behaviour as the noise goes to zero or the amount of data goes to infinity. In Article **IV**, we investigate how the posterior concentrates around a manifold for an under-determined elliptic multiscale inverse problem.

Article I: Spectral Gaps for a Metropolis-Hastings Algorithm in Infinite Dimensions

We consider MCMC algorithms applied to finite dimensional approximations of infinite dimensional measures given by a density with respect to a Gaussian reference measure. We compare the convergence of the standard Random Walk Metropolis (RWM) algorithm and a slight modification that is known as the preconditioned Crank-Nicolson algorithm (pCN). Heuristics in [131] and [151] suggest that the convergence of the RWM deteriorates as the dimension increases and that there is a dimension independent lower bound on the convergence rate of the pCN. We make these heuristics rigorous by bound-

ing the L^2-spectral gap giving rise to bounds on the asymptotic variance of the CLT and non-asymptotic bounds on the mean square error. Our results are the first dimension independent convergence results for a locally moving MCMC algorithm.

Article II: Dimension-Independent MCMC Sampling for Inverse Problems with Non-Gaussian Priors

We show that a Metropolis-Hastings algorithm has an L^2-spectral gap if the target measure has a density that is bounded from above and below with respect to a reference measure and if the proposal kernel has an L^2-spectral gap with respect to the reference measure. We use this result in order to obtain an efficient Metropolis-Hastings algorithm for an elliptic inverse problem by constructing a proposal accordingly. As the proposal of the pCN algorithm, considered in Article **I**, has a spectral gap with respect to the Gaussian reference measure, this can be seen as generalisation. However, the results in Article **I** also apply to unbounded densities.

Article III: Posterior Consistency for Bayesian Inverse Problems through Stability and Regression Results

We develop a method that proves posterior consistency for non-linear inverse problems. In particular, we consider a sequence of posteriors arising from an increasing amount of artificial data generated for a fixed parameter called the truth. As the prior is a subjective choice, it is desirable to characterise priors leading to posteriors that concentrate around this truth. Whereas there are simple conditions in finite dimensions, the choice of the prior has more impact in infinite dimensions because of almost sure properties of the prior. This work is one of the first to address this question for nonlinear inverse problems in infinite dimensions. In order to illustrate the result, we apply our method to an elliptic inverse problem which is well-known for its application in subsurface geophysics. However, we would like to mention that the method is generally applicable.

Article IV: Notes on a Bayesian Elliptic Multiscale Inverse Problem

We study the inverse problem of reconstructing a multiscale diffusion coefficient. The set of additive multiscale diffusion coefficients, giving rise to the same homogenised diffusion

coefficient, is investigated analytically. We show that this set forms a manifold given by a graph structure. The inverse problem is considered from the Bayesian perspective and MCMC simulations for artificial data are performed. These simulations show that the posterior concentrates around the level set containing parameters used to generate the data. Moreover, we show that the posterior based on the homogenised model is close to that of the multiscale problem.

1.2 Outline of this Thesis

This thesis is divided into two parts. Part A contains an overview of Bayesian inverse problems and a description of the results obtained in Articles **I**, **II**, **III** and **IV**. Part B incorporates the original research articles.

Chapter 2 - Bayesian Inverse Problems

We review the Bayesian framework by first giving an overall exposition in Section 2.1 before applying it to an elliptic inverse problem which forms the guiding example for later chapters. We go into more detail about the existence and representation of the posterior in Section 2.3. Continuity and approximation results for the posterior are reviewed in Section 2.5. We introduce the concept of posterior consistency in Section 2.4 and survey computational methods probing the posterior in Section 2.6.

Chapter 3 - Markov Chain Monte Carlo Algorithms for Bayesian Inverse Problems

In this chapter, we introduce the Metropolis-Hastings algorithm on general state spaces. We provide an in-depth literature review considering both heuristic and rigorous convergence results for the resulting Markov chains. This review sets the stage for the presentation of our results from the research Articles **I** and **II**. Moreover, we would like to point the reader to Section 3.3.1.1 which contains a detailed description of how the conductance can be used in order to bound the spectral gap from above for MCMC algorithms. This presentation is more general and in much more detail than the description contained in Article **I**.

Chapter 4 - Posterior Consistency for Bayesian Inverse Problems

We introduce posterior consistency for Bayesian inverse problems and relate it to the concept of posterior consistency in non-parametric statistics. This is followed by a detailed literature review which is used as a background for our introduction to the results from Article **III**.

Chapter 5 - An Under-Determined Multiscale Elliptic Inverse Problem

We consider a particular elliptic inverse problem for which the diffusion coefficient has an additive structure consisting of a fast and slow scale. After reviewing the basics of homogenisation of elliptic PDEs, we present the results from Article **IV**.

Part A

Context and Presentation

Chapter 2

Bayesian Inverse Problems

The intention of this chapter is to summarise the framework of Bayesian inverse problems which is the common topic of the results presented in this thesis. In particular, Articles **III** and **IV** address this framework directly whereas Articles **I** and **II** study Markov Chain Monte Carlo (MCMC) algorithms, a computational method that can be used for approximations of the posterior, the central probability distribution in Bayesian inverse problems.

This chapter is organised as follows. We start by giving a short exposition of the main idea of the Bayesian approach to inverse problems in Section 2.1. In a nutshell, that is to treat the input of a mathematical model as a random variable based on subjective a priori knowledge. The conditional probability distribution of the input given the data is called posterior distribution and is the main quantity of interest in this area. In Section 2.2, we illustrate the Bayesian approach using the example of an elliptic inverse problem. This example also guides us through most of the results given in Articles **II** and **III** even though they are more generally applicable. In contrast, Article **IV** focuses on a multi-scale version of this example. In Section 2.3, we review the literature on the well-definedness of the posterior. Moreover, the posterior is continuous in the data which is one of the crucial benefits of the Bayesian framework. In Section 2.5, we give details about this fact and a finite dimensional approximation result justifying numerical simulations. This framework can be evaluated by investigating posteriors arising from artificial data which is generated using a fixed input. Closely related is the concept of

posterior consistency which is only briefly introduced in Section 2.4 because Chapter 4 is devoted to this subject. We close this chapter by reviewing different computational techniques for making inference based on the posterior. MCMC, one of the most important computational methods for approximating expectations with respect to the posterior, is the subject of Chapter 3.

This thesis concentrates on the Bayesian approach to inverse problems. For a survey of the main other approach to inverse problems, the so-called frequentist approach, we refer the reader to [31]. The research article [31] introduces estimators based on regularised least-squares problems, their minimax error rates and estimators that adapt to the unknown smoothness of the underlying truth. frequentist methods and Bayesian methods can be combined in the so called empirical Bayes' method. We refer the interested reader to [111].

2.1 Exposition to Bayesian Inverse Problems

We follow [177] and [176] for an introduction to some basic notations and definitions in order to set stage for our results and the forthcoming literature reviews in Chapters 3, 4 and 5. This allows us to put our contributions in the context of the literature and its development.

The area of inverse problems is concerned with estimating some unknown parameter $u \in X$ on the basis of a given data set. The data $y \in Y$ is usually modelled as

$$y = \mathcal{G}(u) + \eta, \tag{2.1}$$

with $\mathcal{G}$ denoting the forward operator representing the mathematical model, η being the additive noise with distribution $\mathbb{Q}_0$ and X as well as Y being Banach spaces.

The key idea of the Bayesian approach to inverse problems is to treat the input u of a mathematical model, for example the initial condition of a PDE, as a random variable. Its distribution $\mu_0(du)$ is called the prior and is a modelling choice incorporating a priori knowledge. The Bayesian approach allows us to treat u and y as jointly varying

random variables. The regular conditional probability distribution μ^y of u given the data y, which exists under weak conditions (c.f. Section 2.3), represents the a posteriori knowledge and is henceforth called the posterior. It is in the focus of the Bayesian approach to inverse problems.

As the posterior models the a posteriori uncertainty, a concentrated posterior corresponds to a high amount of certainty. It is possible for the posterior to have multiple modes, an example in Lagrangian data assimilation has been provided in [6]. Another important aspect of the posterior is that it can be used to estimate the unknown u through the posterior mean or the maximum a posteriori (MAP) estimator. The latter is given by the location of an infinitesimal ball with maximum probability and can be linked to the Tikhonov regularisation as demonstrated in [45]. Moreover, the posterior can be used to quantify the uncertainty of such an estimate in terms of, for example, the posterior variance or the posterior probability of a neighbourhood of an estimate.

An appropriate representation of the posterior is needed in order to approximate the quantities above. In particular, the posterior can be represented through Bayes' rule

$$\text{posterior} \;\propto \text{prior} \;\times\; \text{likelihood} \tag{2.2}$$

which is valid if all involved quantities have probability densities. Under appropriate assumptions, Bayes' formula can be generalised in the following way (see Section 2.3). For observational noise η that has a density ρ with respect to the Lebesgue measure λ, the posterior takes the form

$$\frac{d\mu^y}{d\mu_0}(u) \propto \rho\left(\mathcal{G}(u) - y\right). \tag{2.3}$$

In case of Gaussian noise $\eta \sim \mathcal{N}(0, \Gamma)$, that is $\rho(\eta) \propto \exp\left(-\frac{1}{2}\left\|\eta\right\|_\Gamma^2\right)$, where

$$\begin{aligned} \langle x, y\rangle_\Gamma &= \left\langle \Gamma^{-1}x, y\right\rangle \\ \left\|x\right\|_\Gamma^2 &= \langle x, y\rangle_\Gamma . \end{aligned}$$

In this case, we obtain that

$$\frac{d\mu^y}{d\mu_0}(u) \propto \exp\left(-\frac{1}{2}\left\|\mathcal{G}(u)-y\right\|_\Gamma^2\right) \propto \exp\left(-\frac{1}{2}\left\|\mathcal{G}(u)\right\|_\Gamma^2 + \langle y, \mathcal{G}(u)\rangle_\Gamma - \left\|y\right\|_\Gamma^2\right) \tag{2.4}$$
$$\propto \exp\left(-\frac{1}{2}\left\|\mathcal{G}(u)\right\|_\Gamma^2 + \langle y, \mathcal{G}(u)\rangle_\Gamma\right).$$

The last line also holds when η is an infinite dimensional Gaussian random variable as for example has been shown in [177]. In this case, $\|\cdot\|_\Gamma$ denotes the norm of the Cameron-Martin space $(H_{\mathbb{Q}_0}, \langle\cdot,\cdot\rangle_\Gamma)$ of $\mathbb{Q}_0$. If $\rho > 0$, both cases can be written as

$$\frac{d\mu^y}{d\mu_0}(u) \propto \exp\left(-\Phi(u;y)\right). \tag{2.5}$$

In order to approximate the MAP estimator or posterior expectations, such as the mean or the variance, the posterior has to be discretised as it is supported on an infinite dimensional function space. The resulting error can be quantified by a difference in the total variation or the Hellinger distance. This quantity can be bounded in terms of the forward difference, for details consider Section 2.5. The last obstacle for computational methods is that the normalising constants in the Equations (2.3) and (2.4) are unknown. However, there are appropriate computational methods such as Markov Chain Monte Carlo algorithms reviewed in Section 2.6. These algorithms create approximate samples that can be used to approximate posterior expectations. Asymptotic confidence intervals for these approximations can be derived by bounding the convergence rate of the underlying stochastic process. This is one of the main aspects addressed in this thesis. Details are contained in Chapter 3 reviewing the results of Articles **I** and **II**. Deriving more efficient computational methods is a very active area of research as all methods are based on evaluating the posterior density each of which requires a run of the usually expensive forward model.

In general, the literature that is not addressing computational topics falls naturally into the following three parts:

1. the existence and representation of the posterior presented in Section 2.3;

2. continuity and approximation of the posterior with respect to the forward model,

the noise distribution and the data which is addressed in Section 2.5. Even though this is not in the focus of this thesis, these results justify the approximations arising from discretised models on computers;

3. posterior consistency and its relation to the frequentist approach to inverse problems which is dealt with in Chapter 4. It is of interest to study posteriors arising from artificially generated data corresponding to a fixed true parameter. A sequence of posteriors arising from diminishing noise is called posterior consistent if the posteriors concentrate around this truth.

We close this section by commenting on some recent developments before sketching briefly the history of Bayesian inverse problems. The frequentist approach to inverse problems studies estimators that adapt to the smoothness of the underlying truth. In this setting, the empirical Bayes' methods introduce hyper parameters that are then estimated from the data. A proper Bayesian approach to this problem places another prior on the hyper parameters resulting in a hierarchical prior. Both the empirical Bayes' methods as well as hierarchical methods for inverse problems have been studied in [111].

The idea to approach an infinite dimensional linear inverse problem by modelling the unknown as random variable in a way to represent 'a-priori conviction concerning the size and the smoothness'[1] goes back to Franklin [70]. A proper Bayesian approach followed shortly afterwards in a series of papers by Backus [9, 8, 10]. Tarantola developed these ideas further by putting them into a general framework applying Bayesian statistics to finite dimensional discretised versions of the underlying model, an account of this has been, for example, presented in [179]. His approach included the use of MCMC methods to sample from the posterior. A real cornerstone is the book by Kaipio and Somersalo [99]. The authors create an algorithmic framework and demonstrate that the resulting methods are competitive with state-of-the-art regularisation methods, for example, for limited angle tomography, also known as dental x-ray imaging.

[1] p. 690 in J. N. Franklin. Well-posed stochastic extensions of ill-posed linear problems. J. Math. Anal. Appl., 31:682–716, 1970

2.2 The Guiding Example - An Elliptic Inverse Problem

Most results presented in this work are illustrated by an application of the Bayesian approach to a nonlinear elliptic inverse problem. This particular inverse problem is concerned with the reconstruction of the diffusion coefficient a from measurements of the pressure P. The relation between a and P is modelled by the following elliptic partial differential equation

$$\begin{cases} -\nabla \cdot (a \nabla P) &= g \quad \text{in } D \\ P &= 0 \qquad \text{on } \partial D \end{cases} \tag{2.6}$$

where g denotes the forcing and D is a Lipschitz domain in $\mathbb{R}^d$. Notice that the map

$$a \mapsto P(a)$$

is a nonlinear map. For ease of presentation, we abbreviate this inverse problem by (EIP) in our subsequent discussions. The (EIP) has many important applications, for example in subsurface geophysics. The reason for its importance is that Equation (2.6) is a good model for groundwater flow. The derivation is based on Darcy's law which can be derived by homogenising the porous medium equation, for which we refer the reader to [14]. This book also illustrates the use of this model of groundwater flow in nuclear waste management. A review of the (EIP) is contained in [134]. For the model of groundwater flow, the Bayesian approach can be described as conditioning knowledge of the permeability on measurements of the hydraulic head.

Moreover, we would like to point the reader to article [128] containing a comparison of MCMC algorithms and other computational methods, such as the ensemble Kalman filter, by applying them to generalisations of the (EIP).

Because of its importance for many applications and the simplicity of the equation, we use the (EIP) as a guiding example in the following ways:

- In Article **III**, we develop a method to show posterior consistency for nonlinear

inverse problems and apply it to the (EIP). Details can be found in Chapter 4.

- In Article **II**, we prove convergence results for Metropolis-Hastings algorithms in infinite dimensions and construct proposals for the posterior arising from the (EIP). We summarise these results in Section 3.3.2.

- In Article **IV**, we study the strongly under-determined problem if a has an additive multi-scale structure. The corresponding results are reviewed in Chapter 5.

This section is organised as follows. First we set up the corresponding inverse problem following [177] and [92] imposing a prior based on a series expansion with i.i.d. uniformly on $[-1, 1]$ distributed coefficients. We conclude this section by presenting simulations with artificial data and demonstrate the behaviour of the posterior using MCMC samples in Section 2.2.3.

2.2.1 A Literature Review on the Bayesian Approach to the EIP

For the noiseless inverse problem, we refer the reader to [154] and references therein. It has been shown in the article that a can be reconstructed from P under appropriate assumptions on g, D and the class of a under consideration in Equation (2.6). These results are used in Article **III**.

Whereas we concentrate on the Bayesian approach below, we refer the interested reader to [117] for a survey on regularisation techniques for this particular inverse problem. Convergence results for the Tikhonov regularisation of this problem have been obtained in [188].

In the uncertainty quantification literature it has been studied how uncertainty propagates from a to P. The uncertainty is again represented as a probability measure and the resulting uncertainty on P then corresponds to the push-forward measure on P.

The Bayesian method have been applied to the discretisations of the (EIP), see for example [134]. The well-posedness of the posterior in the infinite dimensional problem has been established recently first for log-Gaussian priors in [46]. Besov priors followed shortly afterwards in [44]. Well-definedness of the posterior for a prior based on a

series expansion with i.i.d. uniformly distributed coefficients has been shown in [92]. In the same setting, article [170] provides a sparse general Polynomial Chaos (gPC) representation of the posterior density. Data-adaptive Smolyak integration algorithms can then be used in order to approximate posterior expectations [168].

In [92], the IS algorithm has been compared with two speeded up versions thereof for the posterior arising from the (EIP). One is the multi-level approach which expresses the expectation as difference of posterior expectations corresponding to finer and finer triangulations of the finite-element forward problem. Under a fixed computational budget, more and more samples can be used for coarser discretisations reducing the overall error, see also [103]. The second approach considered in [92] is based on sparse representations of the forward model using gPC method. This representation of the forward model can be evaluated at a reduced computational cost but the representation has to be precomputed resulting in additional fixed computational costs. Some of the above references are considered in more detail in Section 2.6 and Chapters 3 and 4.

2.2.2 The Bayesian Approach with Uniform Series Prior

We apply the Bayesian approach, as presented in Section 2.1, to the (EIP). For sufficiently regular priors, we present formulae for the posterior that are used in simulations in Section 2.2.3. We introduce the uniform series prior which has been used for Bayesian inverse problems in [170] and [92]. It is given through the following parametrisation of the diffusion coefficient a

$$a_u(x) = \bar{a}(x) + \sum_{j\geq 1} u_j \psi_j(x), \quad \text{with } x \in D \tag{2.7}$$

where $u \in U = [-1,1]^\infty$, $\psi_j \in L^\infty(D)$ and $\bar{a} \in L^\infty$ are subject to the subsequent assumption.

Assumption 2.1. *There is a positive constant κ such that*

$$\sum_{j\geq 1} \|\psi_j\|_{L^\infty(D)} \leq \frac{\kappa}{1+\kappa} \bar{K}_{min}$$

where $\bar{a}_{min} = ess\ inf_{x\in D}\bar{a}(x)$.

In particular, this assumption implies that there are $a_{\max} > a_{\min} > 0$ such that

$$0 < a_{\min} \leq a(x) \leq a_{\max}\ \forall x \in D$$

and that the bi-linear form associated with the weak formulation of Equation (2.6) on $V = H_0^1(D)$ is uniformly coercive implying the existence of a solution to Equation (2.6) using the Lax-Milgram lemma [68, 77]. Moreover, it can be shown that the solution operator satisfies the following property.

Proposition 2.2. *Under Assumption 2.1, the map* $K \mapsto P(K)$ *is Lipschitz as a mapping from the appropriate subset of* $\{K_u | u \in U\}$ *to* V.

Proof. See [92]. □

We use the parametrisation in Equation (2.7) to construct a prior on a. In fact, we place a prior on

$$u = (u_1, \dots) \in U = [-1, 1]^{\mathbb{N}}$$

resulting in a simpler presentation which is, by Theorem Appendix B.1 in Article **III**, equivalent to placing a prior on a. Following [170, 92, 177], we choose

$$\mu_0 = \bigotimes_{j=1}^{\infty} \mathcal{U}(-1, 1) \tag{2.8}$$

such that $u_i \overset{\text{i.i.d}}{\sim} \mathcal{U}(-1, 1)$. We denote by $G : U \to V$ the solution operator

$$G(u) = P(a(u))$$

and consider forward operators of the form

$$\mathcal{G}(u) = \mathcal{O}\left(G(u)\right) \tag{2.9}$$

where $\mathcal{O} : V \to W$ is the observation operator. Subsequently, we consider either

1. $\mathcal{O} = \mathrm{Id}$ or

2. $\mathcal{O} = (l_1(P), \dots, l_k(P))$ where $l_i \in V^\star$.

The data is supposed to be modelled based on the forward operator with additive noise

$$y = \mathcal{G}(u) + \eta$$

where $\eta \sim \mathcal{N}(0, \Gamma)$. According to Equation (2.4), the posterior takes the form

$$\frac{d\mu^y}{d\mu_0}(u) \propto \exp\left(-\frac{1}{2}\|\mathcal{G}(u)\|_\Gamma^2 + \langle y, \mathcal{G}(u)\rangle_\Gamma\right).$$

We conclude this section by summarising the formulae for the (EIP)

Model	$-\nabla \cdot (a\nabla P) = g$ in D, $\quad P = 0$ on ∂D (2.6)
Prior	$\mu_0 = \otimes_{j=1}^{\infty} \mathcal{U}(-1, 1)$ on u
Data	$y = \mathcal{G}(u) + \eta$
Posterior	$\frac{d\mu^y}{d\mu_0}(u) \propto \exp\left(-\frac{1}{2}\|\mathcal{G}(u)\|_\Gamma^2 + \langle y, \mathcal{G}(u)\rangle_\Gamma\right).$

(2.10)

For more details about this derivation, we refer the reader to [177].

2.2.3 Markov Chain Monte Carlo Simulations

We perform MCMC simulations for the posterior associated with the inverse problem described by Equation (2.10) on the one dimensional domain $D = (0, 1)$. The main purpose of these simulations is to illustrate the property of posterior consistency and to present an application of an MCMC algorithm to a particular Bayesian inverse problem.

In order to implement our simulations, we have to specify the parametrisation of the problem and the prior which is, in the general case, given by the Equations (2.7) and (2.8). In the following, we recall the parametrisation of the diffusion coefficient K of Equation (2.8)

$$K_u(x) = \bar{K}(x) + \sum_{j=0}^{J} u_j \psi_j(x) \text{ where } u_j \overset{\text{i.i.d.}}{\sim} \mathcal{U}(-1, 1).$$

For the subsequent simulations, we set

$$\begin{aligned} \bar{K}(x) &= 3.5, \\ \psi_{2j-1}(x) &= \frac{1}{j^3}\cos(2\pi jx), \text{ , } J_0 \geq j \geq 1, \\ \psi_{2j}(x) &= \frac{1}{j^3}\sin(2\pi jx), \ \gamma_{2j-1}, \quad J_0 \geq j \geq 1 \psi_0(x) = 1, \\ g(x) &= 10\pi\cos(2\pi x) + 6\cos(0.6\pi x) + 2. \end{aligned}$$

We notice that this choice implies $a(x) \geq 0.1$ independent of $J = 2J_0 + 1$. The data corresponds to evaluations of the pressure uniformly spaced at a distance Δy apart from each other, that is

$$y = \mathcal{G}(K^\dagger) + \eta = (P(K^\dagger)(i\Delta y) + \eta_i)_{i=0}^{\lfloor 1/\Delta y \rfloor}$$

where $\eta \sim \mathcal{N}(0, \sigma^2 I)$. The diffusion coefficient $K^\dagger$, which we call the truth, is generated according to the prior and is fixed for all subsequent simulations. The posterior given in Equation (2.10) involves the solution operator to the PDE which has to be approximated for simulations. In this one dimensional case, the elliptic PDE becomes an ODE which can be explicitly integrated. The resulting ODE has been implemented using the trapezoidal rule.

We visualise the distribution of the pressure, the diffusion coefficient and its constant mode with respect to the prior in Figure 2.1 and with the posterior for a different number of observations (Δy) and magnitude of the observational noise (σ) in the Figures 2.2-2.4. The prior is approximated through Monte Carlo samples of u whereas the posterior is given through MCMC samples. The particular MCMC algorithm used for these simulations is the Reflection Standard Random Walk Metropolis (RSRWM) algorithm which we have constructed in Article **II**. More details about this algorithm can be found in Section 3.3.2.3.

We notice that the posterior variation and in particular, the density become more peaked as the amount of observation (Δy) increases and the magnitude of the observational noise (σ) decreases. If the priors of a sequence of inverse problems converge to a

point mass, this sequence is called posterior consistent. We have studied the posterior consistency of the (EIP) and related problems in Article **III**. The results are summarised in Chapter 4.

2.3 Existence and Representation of the Posterior

In the following, we sketch how the formulae for the posterior (c.f. Equations (2.2), (2.3) and (2.4)) arise from the notion of regular conditional probability distributions. We recall the definition of a conditional probability distribution for a standard probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and random variables $X : (\Omega, \mathcal{F}) \to (S, \mathcal{A})$ and $Y : (\Omega, \mathcal{F}) \to (T, \mathcal{B})$.

Definition 2.3. *(From [63]) A collection of probability measures $\mathbb{P}_{X|Y=y}$ on S is a conditional distribution of X given $Y = y$ under $\mathbb{P}$ if*

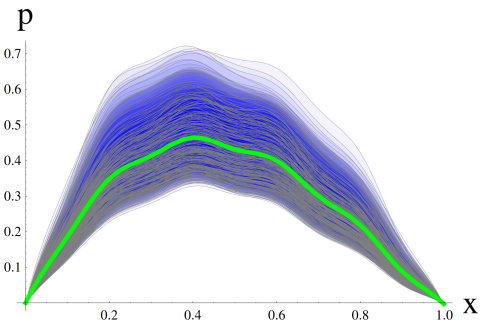
1. *for $A \in \mathcal{A}$ the map $y \mapsto \mathbb{P}_{X|Y=y}(A)$ is measurable from $(T, \mathcal{A})$ into $(\mathbb{R}, \mathcal{B}(\mathbb{R})$ and*
2. *for $A \in \mathcal{A}$ and $B \in \mathcal{B}$ $\mathbb{P}(A \times B) = \int_B \mathbb{P}_{X|Y=y}(Y_\star \mathbb{P})(dy)$.*

Regular conditional probability distributions exist under quite weak conditions, see for example [63, 100, 27]. However, in order to use computational methods, an appropriate presentation of the posterior is needed, for example as density with respect to the prior. For this reason, we focus on the representation of the conditional probability distribution and illustrate how the representations in the Equations (2.2), (2.3) and (2.4) can be obtained from the following general conditioning lemma.

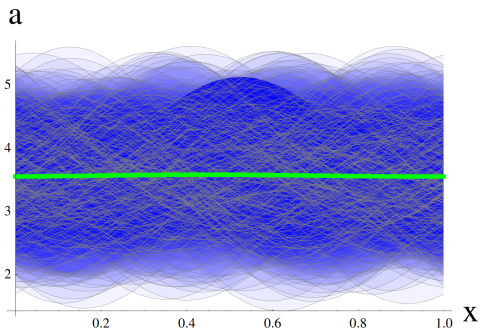
Lemma 2.4. *(Lemma 5.3 in [85]) Let ν and π be two probability measures on $(\Omega, \mathcal{F})$. Assume that π has a density φ with respect to ν and that the conditional distribution $\nu_{X|Y=y}$ exists. Then $\pi_{X|Y=y}$ exists and is given by*

$$\frac{d\pi_{X|Y=y}}{d\nu_{X|Y=y}}(x) = \begin{cases} \frac{1}{c(y)}\varphi(x,y) & \text{if } c(y) > 0 \\ 1 & \text{otherwise} \end{cases}$$

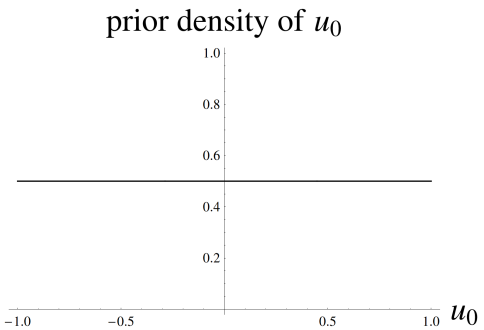
with $c(y) = \int_S \varphi(x,y)\nu_{X|Y=y}(dx)$ for all $y \in T$.



(a) The prior variation of the pressure P

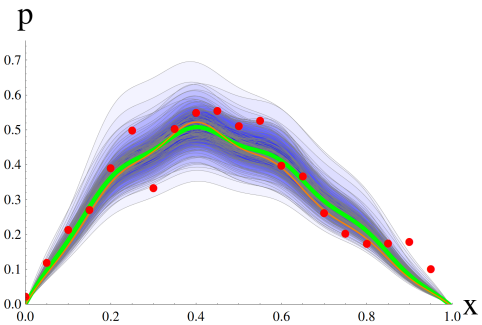


(b) The prior variation of diffusion coefficient K

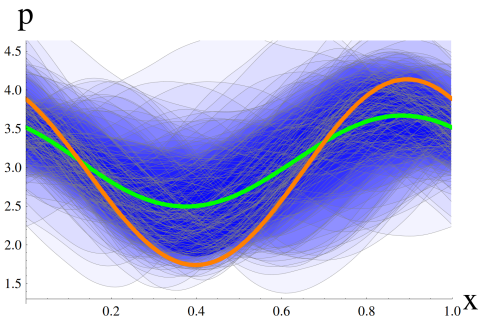


(c) The prior density of constant Fourier mode u_0 of diffusion coefficient K

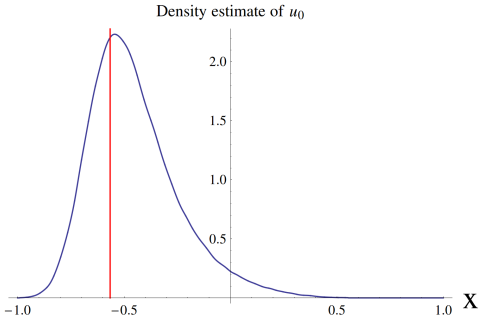
Figure 2.1: Visualisation of the prior



(a) The pressure P corresponding to the truth (orange), the observations (red), the posterior mean (green, MCMC approximated) and the posterior variation (blue)

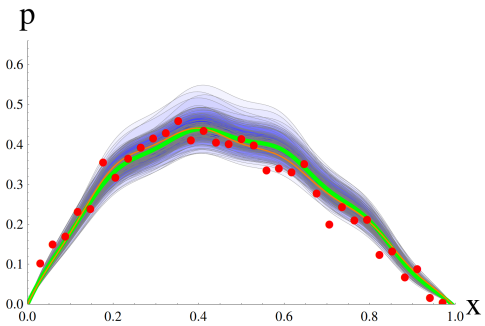


(b) The diffusion coefficient K corresponding to the truth (orange), the posterior mean (green, MCMC approximated) and the posterior variation (blue)

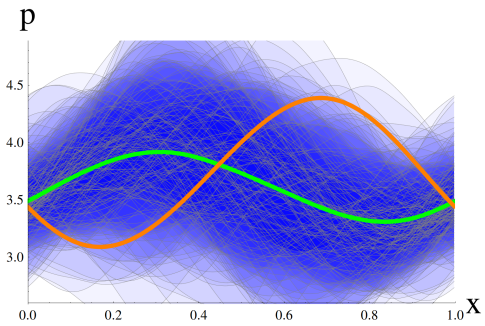


(c) The constant Fourier mode u_0 of the diffusion coefficient K corresponding to the truth (red), posterior density (blue given by smoothed histogram of MCMC samples)

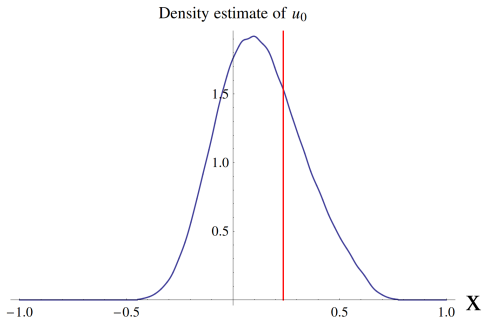
Figure 2.2: MCMC Simulations for $\sigma = 0.05$ and $\Delta y = 0.05$



(a) Pressure P corresponding to the truth (orange), the observations (red), the posterior mean (green, MCMC approximated) and the posterior variation (blue)

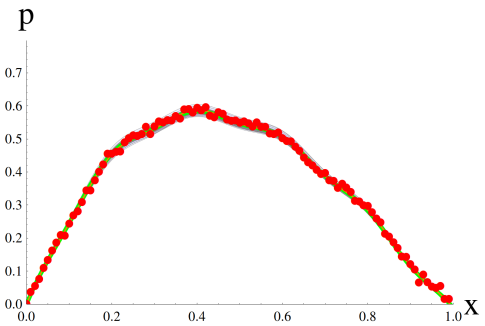


(b) The diffusion coefficient K corresponding to the truth (orange), the posterior mean (green, MCMC approximated) and the posterior variation (blue)

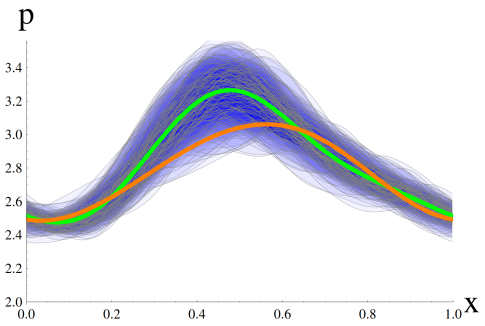


(c) The constant Fourier mode u_0 of the diffusion coefficient K corresponding to the truth (red) and the posterior density (blue given by smoothed histogram of MCMC samples)

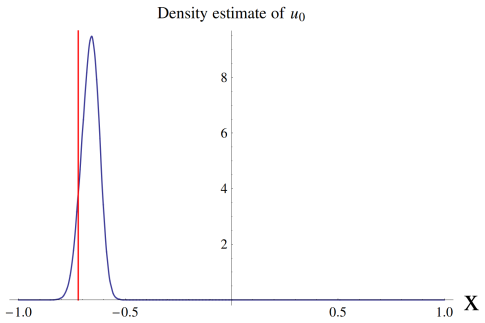
Figure 2.3: MCMC Simulations for $\sigma = 0.03$ and $\Delta y = 0.03$



(a) The pressure P corresponding to the truth (orange), the observations (red), the posterior mean (green, MCMC approximated) and the posterior variation (blue)



(b) The diffusion coefficient K corresponding to the truth (orange), the posterior mean (green, MCMC approximated) and the posterior variation (blue)



(c) The constant Fourier mode u_0 of the diffusion coefficient K corresponding to the truth (red) and the posterior density (blue given by smoothed histogram of MCMC samples)

Figure 2.4: MCMC Simulations for $\sigma = 0.01$ and $\Delta y = 0.01$

We use the above lemma to derive a Bayes' rule for inverse problems as introduced in Section 2.1. We recall that the data is modelled as

$$y = \mathcal{G}(u) + \eta \tag{2.1}$$

for u and y being elements of the Banach spaces X and Y, respectively. By placing a prior μ_0 on u and assuming that $\eta \sim \mathbb{Q}_0$ is independent of u, the joint distribution π of (u, y) takes the form

$$(u, y) \sim \mu_0(du) \left(\mathsf{T}_{\mathcal{G}(u)\star}\mathbb{Q}_0\right)(dy) =: \nu$$

where

$$\mathsf{T}_y(x) := x + y.$$

is the translation operator. In the following, we postulate that $\mathbb{Q}_0$ is quasi-translational invariant with respect to translations by $\mathcal{G}(u)$ for a.e. u with respect to μ_0.

Assumption 2.5. *(Quasi-translational invariance of the noise measure) We assume there is* $\Phi(u;\, y) : X \times Y \to \mathbb{R}$ *such that*

$$\frac{d\left(\mathsf{T}_{\mathcal{G}(u)\star}\mathbb{Q}_0\right)}{d\mathbb{Q}_0}(y) = \exp\left(-\Phi\left(u; y\right)\right). \tag{2.11}$$

This assumption allows us to derive Bayes' rule on the basis of Theorem 2.4.

Theorem 2.6. *(Bayes' rule from [177]) Assume* $\Phi\left(u; y\right)$ *is* ν*-measurable and that for* $\mathbb{Q}_0$*-a.e.* y

$$Z := \int_x \exp\left(-\Phi(u; y)\right) \mu_0(du) > 0.$$

Then the conditional distribution μ^y *of* u *given* y *exists and for* ν*-a.e.* y *it takes the form*

$$\frac{d\mu^y}{d\mu_0}(u) = \frac{1}{Z} \exp\left(-\Phi\left(u; y\right)\right).$$

A very important special case of this result corresponds to $\mathbb{Q}_0 = \mathcal{N}\left(0, \Gamma\right)$ which is

also mentioned in Section 2.1. In this case Assumption 2.5 holds with

$$\Phi(u;y) = -\|\mathcal{G}(u)\|_\Gamma^2 + \langle \mathcal{G}(u), \eta\rangle_\Gamma$$

because of the Cameron-Martin theorem which can be found in [40] and [82]. Moreover, it is worth mentioning that Lemma 2.4 can be used to derive the standard finite dimensional Bayes' rule. The joint law can be written as a density with respect to a product probability measure by introducing an artificial probability measure with everywhere positive density for normalisation.

A more general discussion of representing the posterior in infinite dimensions is available in [118]. The author also discusses the problem of choosing a version of the conditional distribution based on the continuity of $y \mapsto \mathbb{P}_{X|Y=y}$.

2.4 Posterior Consistency

The Bayesian method can be assessed by fixing a true input $u^\dagger$ and by considering the posterior μ^y resulting from artificial data $y = \mathcal{G}(u^\dagger) + \eta$. As the aim of inverse problems is to reconstruct u, it is desirable that μ^y concentrates around $u^\dagger$ for a large class of $u^\dagger$. This property is called posterior consistency. In Article **III**, we present one of the first posterior consistency results for nonlinear infinite-dimensional inverse problems. We give a thorough overview about the existing literature on posterior consistency and our own contribution in more detail in Chapter 4.

2.5 Approximation and Continuity of the Posterior

In Section 2.3 we justify Bayes' rule for the posterior density in different infinite dimensional settings. However, infinite dimensional forward models have to be approximated by discretised versions for simulations. The simulations are only reasonable if the bias introduced by the discretisation can be quantified. The appropriate approximation results along with continuity of the posterior in data are the subject of this section.

The Bayesian framework is particularly appealing as under appropriate conditions,

see for example [36] or [177], the posterior μ^y depends in a Lipschitz continuous way on the data, that is

$$d_{\text{Hell}}(\mu^{y_1}, \mu^{y_2}) \leq C \|y_1 - y_2\|$$

where d_{Hell} is the Hellinger distance. For two probability measures μ_1 and μ_2 with common reference measure, the Hellinger distance is given by

$$d_{\text{Hell}}(\mu_1, \mu_2) := \sqrt{\frac{1}{2} \int \left(\sqrt{\frac{d\mu_1}{d\nu}} - \sqrt{\frac{d\mu_2}{d\nu}} \right)^2 d\nu}.$$

For a function $f \in L^2_{\mu_1} \cap L^2_{\mu_1}$ a bound on the Hellinger distance gives rise to

$$|\mathbb{E}^{\mu_1} f - \mathbb{E}^{\mu_2} f| \leq 2 \left(\mathbb{E}^{\mu_1} f^2 + \mathbb{E}^{\mu_2} f^2 \right)^{\frac{1}{2}} d_{\text{Hell}}(\mu_1, \mu_2) \tag{2.12}$$

which can be found in Lemma 6.37 in [176].

Similar results are available for perturbing or approximating the forward model $\mathcal{G}$ by a discretised version $\mathcal{G}^N$, for example, based on finite element methods for the (EIP) in [92] or spectral methods for the Navier-Stokes equation in [37]. More generally, $\Phi(u; y)$ is approximated by $\Phi^N(u; y)$ in Equation (2.5). A very important special case is when $\Phi^N(u; y)$ represents a finite dimensional approximation to a continuum model. If the forward model can be diagonalised by a spectral representation in a Hilbert space, a natural option is $\Phi^N(u; y) := \Phi(\Pi_N u; y)$ where $\Pi_N u$ is the projection onto the space spanned by $\phi_1, \ldots, \phi_N \in X$. The approximated posterior then corresponds to

$$\frac{d\mu^y_N}{d\mu_0}(u) \propto \exp\left(-\Phi^N(u; y)\right).$$

If there is an appropriate bound of the form

$$\left|\Phi(u; y) - \Phi^N(u; y)\right| \leq \psi(N) M(u),$$

such that $M(u) \in L^1_{\mu_0}$, then

$$d_{\text{Hell}}(\mu^y, \mu^y_N) \leq C\psi(N).$$

This crucial result justifies the application of computational methods to the discretised posterior in order to approximate posterior expectations. For more details on these approximation results, we refer the reader to [177]. The first results of this type have been obtained in [37]. We also refer the reader to [119] which studies finite dimensional approximations to posterior distributions in quite generality. However, this article quantifies the convergence in distribution and the total variation distance which do not guarantee that the conditional variance converges as the dimension increases.

The references presented in this section contain important approximation results justifying computational methods presented in the next section for discretised forward models.

2.6 Computational Methods for Bayesian Inference

Making inference based on the posterior is not straightforward since it is only represented as an unnormalised density with respect to the prior and it involves evaluations of the usually expensive forward model, for details consult the Equations (2.3) and (2.4). In this section, we review appropriate computational methods for probing the posterior. These methods aim to approximate

1. the Maximum A Posteriori (MAP) estimator,
2. posterior expectations or
3. minimisers of posterior expectations.

We do not go into detail about the computational methods approximating minimisers of posterior expectations but just remark that they can be obtained by combining computational methods for approximating posterior expectations and stochastic optimisation algorithms as the Robbins-Monro algorithms. This algorithm has been introduced in [155] and we recommend [28] and [5] for further reading.

In finite dimensions, the MAP estimator corresponds to finding the maximum of the Lebesgue density of the posterior. This is equivalent to determining the location of an infinitesimal ball with maximal a posteriori probability. This formulation generalises to

the infinite dimensional setting. Recently, it has been shown rigorously that the MAP estimator for posteriors arising from Gaussian priors and noise corresponds to minimisers of an appropriate Tikhonov L^2-regularisation functional [45]. In this way, the choice of norms in the functional can be related to the prior and the noise covariance. For the same reason, all optimisation methods that can be used for Tikhonov functionals can also be used for approximating the MAP estimator. We refer the reader to [90] for an introduction and overview of PDE-based optimisation which is a suitable method for the resulting functional.

Posteriors with the same MAP estimator can look quite differently. They might be unimodal, multimodal, concentrated or not concentrated around the MAP estimator at all. For this reason, it is important to consider computational methods that approximate posterior expectations which characterise the posterior. Posterior expectations also contain meaningful information such as how likely it is that the unknown lies in a certain set or characterises the a posteriori uncertainty, for example through the posterior variance. Approximating posterior expectations by evaluating the posterior on a grid is only possible in low dimensions as the computational complexity increases exponentially in the dimension. Sampling algorithms constitute a large class of algorithms that under appropriate assumptions suffer less from the dimension and often do not need a separate estimation of the normalising constant. The common idea is to generate samples and to approximate the expectation using appropriate averages. Subsequently, we present a brief overview of different sampling algorithms such as the Monte Carlo, the Markov Chain Monte Carlo (MCMC), the Sequential Monte Carlo (SMC), the Quasi Monte Carlo (QMC) algorithms. We also briefly mention general Polynomial Chaos (gPC) algorithms which are not based on sampling. The purpose of the remainder of this chapter is to review all major computational methods for approximating posterior expectations before concentrating on MCMC algorithms in the following chapter.

Monte Carlo Algorithms

A prime example are Monte Carlo methods which use i.i.d. samples from the target measure in order to approximate its expectation using the sample average [7]. The cent-

ral limit theorem guarantees that the difference between the expectation and the average based on n samples is of order $\mathcal{O}(\frac{1}{\sqrt{n}})$. Importance sampling consists in sampling from an importance distribution different from the target distribution, and then replacing the sample average by a weighted average to cancel the influence of the importance distribution. Both methods rely on the ability to produce i.i.d. samples with appropriate distributions. A common strategy is to apply a transformation random variables which can be sampled easily in order to obtain samples from the desired distribution. Whereas it is difficult to construct functions exactly, a recent approach to Bayesian inverse problems is to construct a function that approximately maps the prior to the posterior [66]. Another reason why it is difficult to create independent samples is that for many methods, precise bounds on the density are needed for them to be feasible. For example, this is the case for rejection sampling as described in [156].

Even if it is not possible to sample from the target measure directly it is often possible to sample from an arbitrary close approximation. A typical example is the solution of an SDE. It is possible to sample from the Euler-Maruyama with arbitrary small step size. However, each step size introduces a bias. For a fixed computational budget there is therefore a trade off between both errors. In this case it is also possible to express the expectation of interest as a telescopic sum of expectation with respect ti a sequence of increasingly finer step-size. This benefits in two ways:

1. more samples can be used for summands with large step size leading to small Monte Carlo error

2. summands corresponding to small step sizes the integrand is quite small leading to small Monte Carlo error even for a small amount of samples.

This is the so called multilevel approach which has been introduced in article [78].

Markov Chain Monte Carlo Algorithms

The limited applicability of standard Monte Carlo algorithms can be overcome by allowing the samples to be dependent on each other. One of the simplest ways for samples to be dependent is that of a Markov chain as the next in the sequence is conditionally

independent of all the previous given the current. Under appropriate conditions on the Markov chain, the resulting samples can be used to approximate expectations through averages. Algorithms providing the appropriate Markov chains are called Markov Chain Monte Carlo (MCMC) algorithms. They are much more generally applicable than standard Monte Carlo algorithms which can be seen as special cases. For a prescribed target measure, there are many suitable algorithms. Choosing a 'fast' one is difficult as it is hard to obtain an error bound, that is a confidence set for the expectation of interest. For a prescribed confidence set, the complexity of an MCMC algorithm can be quantified as follows

$$\text{number of necessary steps} \times \text{cost of one step.} \tag{2.13}$$

Whereas the cost of one step is usually problem dependent and straightforward to quantify, the number of necessary steps for a prescribed confidence bound is not. The difference between expectation and sample average is partly due to the finite sample size, the Monte Carlo error and partly due to the bias which arises because a discretised model is used for approximating the posterior density. The latter can be bounded using the approximation results from Section 2.5 and the former decreases asymptotically as $\mathcal{O}(\frac{1}{\sqrt{n}})$ if a Markov chain CLT holds, see [29, 122]. It is of interest to show how the leading constant in $\mathcal{O}(\frac{1}{\sqrt{n}})$ depends on relevant problem parameters. For Bayesian inverse problems, it is of interest to characterise the dependence on the noise level and the dimension of the approximations. In particular, the dependence on the dimension is studied in Articles **I** and **II** and is the subject of Chapter 3.

One way to speed up an MCMC algorithm is to reduce the cost of evaluating the target density. For Bayesian inverse problems this can be achieved by an application of a representation of the model which can be precomputed. If an appropriate representation is available, each evaluation is cheaper at the expense of an initial cost. This approach has been taken in [92] using a sparse-tensor representation of the forward model based on general Polynomial Chaos (gPC) for the (EIP).

Another possibility to speed up an MCMC algorithm for Bayesian inverse problems is to split the expectation of interest into a telescopic sum of expectations corresponding

to increasingly precise approximations of the forward model. In this way, more samples can be used in order to decrease the Monte Carlo error for the coarser models and less samples are then needed as the difference of the expectations is comparatively small. This is the multi-level approach which has originally been introduced for SDEs and has recently been applied to the (EIP) in [92] and [103].

Sequential Monte Carlo Algorithms

Whereas we just discussed extensions of the standard Monte Carlo method, it is also possible to extend importance sampling by considering a sequence of probability distributions using each as the importance distribution for the next. These methods are called Sequential Monte Carlo (SMC) methods and are reviewed in [62]. Recently, these methods have been applied to the inverse problem of reconstructing the initial condition of the Navier-Stokes equations in [101].

Quasi-Monte Carlo Algorithms

All sampling methods discussed previously are based on constructing a sequence of random quantities such that an appropriate average converges to the expectation of interest. The idea of Quasi-Monte Carlo (QMC) methods is to consider a deterministic series instead. For this theory, the model problem is to integrate a function on the unit cube against the Lebesgue measure. Depending on the smoothness of the integrand using a space filling sequence, these methods are able to beat the algebraic order $\mathcal{O}(\frac{1}{\sqrt{n}})$ of the Monte Carlo error. For an introduction, we refer the reader to an overview in [58]. An interesting recent approach is to combine QMC and MCMC methods into MCQMC. So far, theoretical results only guarantee the existence of a deterministic driver sequence with a rate that agrees with the standard Monte Carlo rate of $\mathcal{O}(\frac{1}{\sqrt{n}})$, for which we refer the reader to [57]. Another drawback is that both QMC and MCQMC are not as generally applicable as SMC or MCMC algorithms. So far, neither QMC or MCQMC has been tested with posteriors arising in Bayesian inverse problems.

General Polynomial Chaos Algorithms

The method of general Polynomial Chaos (gPC) is not limited to speeding up MCMC by representing the forward model as described above more efficiently, it can also be used to represent the posterior density. Deterministic integration schemes, like adaptive Smolyak quadrature, can be applied to gPC representation of leading convergence rate that beats the usual $\mathcal{O}(n^{-\frac{1}{2}})$ of Monte Carlo algorithms. For a recent account of the theory, we refer the interested reader to [168].

Whereas the purpose of this section is to review all major computational methods probing the posterior, we concentrate on MCMC algorithms in the following chapter. Even though MCMC algorithms are well-studied and their qualitative convergence rate is often of order $\mathcal{O}(n^{-\frac{1}{2}})$, statistical error bounds that are relevant for practice are still rare. We present results contained in Articles **I** and **II** quantifying the dependence of the statistical error bound on the dimension of a discretisation of a Bayesian inverse problem in a rigorous way.

Chapter 3

Markov Chain Monte Carlo Algorithms for Bayesian Inverse Problems

Markov Chain Monte Carlo (MCMC) algorithms are one of the most widely used computational methods for approximating expectations with respect to a given target probability measure μ on the space E. In particular, MCMC algorithms play an important role in dealing with Bayesian inverse problems. For this reason, we have briefly introduced them in Section 2.6. In addition, this chapter is devoted to summarising the deeper analysis of MCMC algorithms provided in Articles **I** and **II** and to relate them to the existing literature.

The common idea of MCMC algorithms is to approximate $\mathbb{E}_\mu f$ for some function $f : E \to \mathbb{R}$ by the sample average

$$S_{n,n_0}(f) = \frac{1}{n} \sum_{i=n_0}^{n+n_0} f(X_i) \tag{3.1}$$

where $\{X_i\}_{i\in\mathbb{N}}$ denotes the evolution of an appropriate Markov chain, n is the sample size and n_0 is called the burn-in which aims at reducing the bias. This approximation is based on the ergodicity of the underlying process guaranteeing that the time average converges to the space average as the time tends to infinity. In this chapter as well as in Articles **I** and **II**, we focus on a particular subclass of MCMC algorithms known as Metropolis-Hastings algorithms. The basic idea of Metropolis-Hasting algorithms is to

obtain an appropriate Markov chain by accepting or rejecting a move from the proposal kernel in a way such that the resulting Markov chain is invariant with respect to the target measure. Under mild conditions (we refer the reader to Theorem 3.5), the sample average converges to the expectation of interest that is

$$\lim_{n\to\infty} S_{n,n_0}(f) = \mu(f).$$

However, as the stochastic error term can converge to zero arbitrarily slowly, it is of interest to derive (asymptotic) confidence intervals for the error

$$\mathcal{E}^X_{n,n_0}(f) = \mathbb{E}_\mu f - S_{n,n_0}(f).$$

For standard Monte Carlo methods corresponding to i.i.d. $X_i \sim \mu$, the central limit theorem (CLT) guarantees that the size of the asymptotic confidence interval for a prescribed level decays like $\mathcal{O}(\frac{1}{\sqrt{n}})$ for $f \in L^2_\mu$. These considerations can be turned into rigorous statistical error bounds for standard Monte Carlo algorithms. It is worth pointing out that the confidence intervals for MCMC algorithms also often exhibit an $\mathcal{O}(\frac{1}{\sqrt{n}})$ behaviour.

In this chapter, as well as in Articles **I** and **II**, we consider target measures of the form

$$\mu(dx) = M \exp(-\Phi(x))\gamma(dx), \tag{3.2}$$

where γ is a reference probability measure on an infinite dimensional Banach or Hilbert space, and approximations to the measure μ. These d-dimensional approximations are given by

$$\mu_d(dx) = M \exp(-\Phi_d(x))\gamma_d(dx).$$

Target measures of this kind arise as posteriors in Bayesian inverse problems described in Section 2.3 and 2.5.

We concentrate on the subclass of Metropolis-Hastings algorithms as many MCMC algorithms used in practice and most of the algorithms known to be well-defined on

infinite dimensional state spaces exhibit this particular form. One exception has been presented in [147]. For most MCMC algorithms, it is natural to expect that the cost per step of the algorithm increases when applied to μ_d for an increasing dimension d. However, in order to satisfy a fixed stochastic error bound on $\mathcal{E}^X_{n,n_0}(f)$, in general, the number of necessary steps increases as well. This leads to tremendously higher computational costs constituting a major problem for application for Bayesian inverse problems.

In Article **I**, we provide the first dimension independent convergence result for an appropriately modified random-walk type Metropolis-Hastings algorithm, reviewed in Section 3.3.1.2. While there are some heuristic arguments quantifying the increase in the number of steps, for example, for the RWM algorithm, we also quantify this in a rigorous way using the method of conductance. For many algorithms, the dimension dependence has been quantified in terms of the acceptance probability and diffusion limits [22]. Using the method presented in Section 3.3.1.1, many of these results can be turned into rigorous upper bounds on the rate of convergence.

The remainder of this chapter is organised as follows. In Section 3.1, we give a brief overview of Metropolis-Hastings algorithms on general state spaces allowing us to consider abstract MCMC algorithms on an infinite dimensional Hilbert space H. We provide a thorough literature review on both heuristic and rigorous convergence results for the resulting Markov chain in Section 3.2. Using this background, we summarise the results obtained in Articles **I** and **II** which are contained in Part B of this thesis.

Research Article I

We study the standard Random Walk Metropolis (RWM) algorithm, a Metropolis-Hastings algorithm with a particular Random Walk proposal, and a slight modification of it, known as the preconditioned Crank-Nicolson (pCN) algorithm. We apply both algorithms to d-dimensional approximations μ_d of the target measure μ given in Equation (3.2) with γ being Gaussian. We pose conditions on the target measure such that the $L^2_{\mu^d}$-spectral gap of the pCN does not depend on the dimension d. Standard results then imply confidence bounds for the sample average of any $L^2_{\mu^d}$-function f independent of

the dimension d. In contrast, we show that the spectral gap of the RWM deteriorates as the dimension increases.

Research Article II

We show that the lazy version of the Metropolis-Hastings algorithm has an L^2-spectral gap with respect to a target measure if it has a density with respect to a reference probability measure for which the proposal kernel has an L^2-spectral gap. Prior to this result, the independence sampler has been the only sound MCMC algorithm applicable, for example, to the Bayesian inverse problem considered in Section 2.2. For this reason, only the independence sampler has been considered in [92]. Moreover, we construct a class of algorithm which we call the Reflection Random Walk Metropolis (RRWM) algorithms which have explicitly been tailored for the (EIP) with uniform series priors as introduced in Section 2.2.

In order to present this work in more detail in Section 3.3, we review Metropolis-Hastings algorithms and provide a detailed literature review in the following sections.

3.1 Exposition of Metropolis-Hastings Algorithms

The most relevant subclass of MCMC algorithms are the Metropolis-Hastings algorithms [136, 89]. For an extensive overview of standard MCMC algorithms, we refer the reader to [29] and for a concise survey we recommend [156]. Metropolis-Hastings algorithms are given by a Markov chain which is constructed on the basis of a given homogeneous proposal Markov chain by accepting or rejecting its moves. The aim is that the resulting Metropolis-Hastings Markov chain can be used to sample from a given target measure. Under appropriate assumptions on the target measure and the proposal Markov chain, the acceptance probability can be chosen in a way that the target measure is an invariant measure of the Metropolis-Hastings Markov chain. Following [181], we introduce the Metropolis-Hastings algorithm in full generality in this section. This allows us to consider abstract algorithms for the full infinite dimensional posterior arising in Bayesian inverse problems. Numerical simulations suggest that these algorithms perform well

when applied to high dimensional discretisations of the target measure.

First, we introduce Metropolis-Hastings algorithms for arbitrary proposal Markov kernels and present the common proposal choices corresponding to

- the independence sampler (IS),
- the Random Walk Metropolis (RWM) algorithm and
- the Metropolis-Adjusted-Langevin (MALA) algorithm

for target measures that have a density with respect to the Lebesgue measure. In a second step, we discuss modifications of these algorithms that perform better as the dimension of the approximations increases. This type of target measure is considered in Article **I** and arises from Bayesian inverse problems if the prior is chosen to have a density with respect to a Gaussian measure. More details about the target measures and algorithms considered in Article **II** can be found in Section 3.3.2.

All MCMC algorithms give rise to a particular Markov chain. Therefore we present the relevant notions first before concentrating on Metropolis-Hastings algorithms.

3.1.1 Time Homogeneous Markov Chains

In this thesis, we only consider time homogeneous Markov chains on a Polish space E. Homogeneous Markov chains are uniquely determined through their initial distribution $\mathcal{L}(X_0)$ and their transition probability kernel P, which we recall in the following.

Definition 3.1. *(From [137]) $P : E \times \mathcal{B}(E) \to \mathbb{R}$ is a probability transition kernel if*

- *for each $A \in \mathcal{B}(E)$, $P(\cdot, A)$ is a non-negative measurable function on E;*
- *for each $x \in E$, $P(x, \cdot)$ is a probability measure on $\mathcal{B}(E)$.*

We use the standard notation $P^n(x, y)$ for the n-step kernel. Moreover, P acts on a function $f : E \to \mathbb{R}$ and a measure ν on $(E, \mathcal{B}(E))$ as follows

$$(\mu P)(dz) := \int_E P(x, dy)\mu(dx)$$
$$Pf(x) := \int_E f(y)P(x, dy).$$

3.1.2 The Abstract Metropolis-Hastings Algorithm

The Metropolis-Hastings algorithm accepts a move from the proposal kernel $Q(x, dy)$ with probability $\alpha(x, y)$, specified below, giving rise to the following abstract algorithm.

Algorithm 3.2 (Metropolis-Hastings algorithm)**.** *Initialise* X_0*. For* i*=0,...,*n *do: Generate* $Y \sim Q(X_i, \cdot)$ *and* $U \sim \mathcal{U}(0, 1)$ *independently, then set*

$$X_{i+1} = \begin{cases} Y & \alpha(X_i, Y) \geq U \\ X_i & \textit{otherwise} \end{cases}.$$

It is straightforward to show that the resulting sequence is again a Markov chain with transition kernel

$$P(x, dz) = \alpha(x, z)Q(x, dz) + \delta_x(dz)\left(1 - \int_E Q(x, dv)\alpha(x, v)\right)$$

where δ_x denotes the Dirac measure centred at x. The main idea is to choose the acceptance probability in a way that makes the target measure invariant for P, that is

$$\mu P = \mu.$$

Note that a strong law of large numbers holds because of Birkhoff's ergodic theorem if μ is the unique invariant measure of P. The corresponding result can be found in [100]. If the Radon-Nikodym derivative $\frac{d(\mu(dy)Q(y,dx))}{d(\mu(dx)Q(x,dy))}$ exists, this measure μ is invariant for P if

$$\alpha(x, y) : \quad = \quad \min\left(1, \frac{d\left(\mu(dy)Q(y, dx)\right)}{d\left(\mu(dx)Q(x, dy)\right)}\right). \tag{3.3}$$

In fact, this choice implies that P is reversible with respect to μ, that is

$$\mu(dx)P(x, dy) = \mu(dy)P(y, dx). \tag{3.4}$$

For an explicit calculation we refer the reader to [181]. So far, we have introduced the

Metropolis-Hastings algorithm in an abstract setting. In the next section, we present particular algorithms based on choices for the proposal distribution and derive their acceptance probability α.

3.1.3 The IS, RWM and MALA Algorithm for Target Measures with Lebesgue Density

We consider basic Metropolis-Hastings algorithms for target measures μ given as a density π with respect to the Lebesgue measure λ. In this case, common choices for the proposal distribution are those of the Independence Sampler (IS), the Random Walk Metropolis (RWM) and the Metropolis-Adjusted-Langevin (MALA) algorithm. Whereas these algorithms are well-known, see for example [156, 29], we concentrate on modified versions which are more suitable for target measures arising in Bayesian inverse problems. We introduce the standard algorithms in this section and their modified versions in Section 3.1.4. However, before introducing their modifications, we devote this section to the standard IS, RWM and MALA algorithm in the following assuming that the target measure takes the form $\mu \propto \pi d\lambda$. For the standard IS, RWM and MALA algorithm, we state the proposal and corresponding acceptance probabilities according to Equation (3.3). They are given by a Metropolis-Hastings algorithm (see Algorithm 3.2) with the appropriate choices for Q and α.

Independence Sampler (IS)

The IS gets its name because the proposals do not depend on the current state of the chain and are given by $\rho(y)dy$ where ρ is an appropriate density. In this setting, the IS sampler takes the form

$$\begin{aligned} Q_{\mathrm{IS}}(x, dy) &= \rho(y)dy \\ \alpha_{\mathrm{IS}}(x, y) &= \frac{\pi(y)\rho(x)}{\pi(x)\rho(y)} \wedge 1. \end{aligned} \tag{3.5}$$

Subsequent steps of the resulting Markov chain only depend on the current state through the acceptance probability. This leads to a fast decorrelation of the samples if the acceptance probability is large enough.

Random Walk Metropolis (RWM)

In contrast to the IS algorithm, the RWM is based on the random walk proposal which consists of local moves based on a normal distribution centred at the current sample. For a covariance matrix C and a step-size parameter δ the proposal and acceptance probabilities are given by

$$\begin{aligned} Q_{\mathrm{RWM}}(x, dy) &= \mathcal{N}(x, 2\delta C)(dy) \\ \alpha_{\mathrm{RWM}}(x, y) &= \frac{\pi(y)}{\pi(x)} \wedge 1. \end{aligned} \tag{3.6}$$

It is worth mentioning that the proposal Q_{RWM}, as any symmetric random walk, is reversible with respect to the Lebesgue measure.

Metropolis-Adjusted-Langevin (MALA)

Another common Metropolis-Hastings algorithm, known as MALA algorithm, is based on the continuous time dynamics of the over-damped Langevin stochastic differential equation (SDE)

$$dX_t = C\nabla \log \pi dt + \sqrt{2}C^{\frac{1}{2}} dB_t.$$

The resulting continuous time dynamics preserve the target measure which can be verified using the forward Kolmogorov equation, for details consider [149]. However, just using the steps of a numerical approximation as samples introduces a non-vanishing bias which is demonstrated and quantified in [162, 132, 133]. This can be overcome by using the steps of a numerical integrator such as the Euler-Maruyama algorithm as the proposal for a Metropolis-Hastings algorithm. Following [109], the Euler-Maruyama time step takes the form

$$X_{t+\Delta t} = X_t + C\nabla \log \pi(X_t)\Delta t + \sqrt{2}C^{\frac{1}{2}}(B_{t+\Delta t} - B_{t+\Delta t}).$$

According to this, the resulting Metropolis-Hastings algorithm takes the form of Algorithm 3.2 with the following proposal kernel and acceptance probability

$$\begin{aligned} Q_{\mathrm{MALA}}(x, dy) &= \mathcal{N}(x + \delta C \nabla \log \pi, 2\delta C)(dy) \\ \alpha_{\mathrm{MALA}}(x, y) &= \frac{\pi(y) \exp\left(-\frac{1}{2} \|x - y - \delta C \nabla \log \pi(y)\|_C^2\right)}{\pi(x) \exp\left(-\frac{1}{2} \|y - x - \delta C \nabla \log \pi(x)\|_C^2\right)} \wedge 1. \end{aligned}$$

It is possible to apply the RWM and the MALA algorithm to discretisations of Bayesian inverse problems of the measures. However, it is well-known from simulations, for example in [173], that the performance of the RWM and MALA deteriorate if the dimension increases. We present heuristic arguments based on [159], [131] and [152] in Section 3.2.2 suggesting that the number of steps for a prescribed confidence level increases like $\mathcal{O}(d)$ and $\mathcal{O}(d^{\frac{1}{3}})$ for the RWM and the MALA algorithm, respectively. We make this rigorous in terms of the L^2_μ-spectral gap which is one of our major contributions of Article **I** and is reviewed in Section 3.3.1.1. Because of the undesirable dimension dependence of the RWM and the MALA algorithm, we consider appropriate modifications in the following section.

3.1.4 Metropolis-Hastings Algorithms for Target Measures based on Gaussian Probability Measures

Subsequently, we present appropriate modifications of the RWM and the MALA algorithm that are not only well-defined for Gaussian based infinite dimensional target measures arising from Bayesian inverse problems, but also outperform the standard algorithms for high-dimensional discretisations. For this presentation, we briefly recall the structure of the posteriors arising in Bayesian inverse problems, see Chapter 2, with Gaussian-based priors and finite dimensional approximations based on a Karhunen-Loeve expansion. These measures take the form

$$\mu(dx) = M \exp(-\Phi(x))\gamma(dx) \tag{3.2}$$

where γ is a Gaussian reference probability measure. These measures do not only arise in Bayesian inverse problems, but also in sampling conditioned diffusions [87, 86, 85]. In the case of Bayesian inverse problems, it might be a helpful illustration to assume that γ coincides with the prior and μ with the posterior. For more details, we refer the reader to Section 2.1. For ease of presentation, we consider a central Gaussian reference measure $\gamma = \mathcal{N}(0, C)$ on a Hilbert space H. At this point, we would like to remark that even though this exposition only considers Gaussian reference measures, the results from Article **II** apply also to non-Gaussian reference measures, consider also Section 3.3.

In the Gaussian case, γ can be represented through the Karhunen-Loeve expansion

$$\gamma(dx) = \mathcal{L}\left(\sum_{i=1}^{\infty} \gamma^i \varphi_i \xi_i\right)(dx), \tag{3.7}$$

where $\left((\gamma^i)^2, \varphi_i\right)$ is an orthogonal eigensystem of the trace class covariance operator. This expansion motivates the following finite dimensional approximation

$$\begin{aligned} \mu_d(dx) &= M_d \exp(-\Phi_d(x))\gamma_d(dx) \text{ with} \\ \gamma_d(dx) &= \mathcal{L}\left(\sum_{i=1}^{d} \gamma^i \varphi_i \xi_i\right)(dx) \end{aligned} \tag{3.8}$$

on the space

$$H_d = \text{span}\left\{\phi_1, \ldots, \phi_d\right\},$$

where Φ_d denotes a d-dimensional approximation of Φ. A natural choice for Φ_d is, for example, $\Phi(\Pi_d \cdot)$ where Π_d denotes the orthogonal projection onto H_d. Note that the covariance operator C_d of γ_d on H_d can be represented as a diagonal matrix with entries $\left((\gamma^1)^2, \ldots, (\gamma^d)^2\right)$ with respect to the basis $\{\phi_1, \ldots, \phi_d\}$. Under appropriate assumptions on Φ, the finite dimensional approximations of μ_m converge weakly to the target measure μ. Results of this type are contained in [38], [177] and [180]. We abuse the notation by writing $\mu_\infty = \mu$ and $\gamma_\infty = \gamma$. In this way, we obtain the same notation for the MCMC algorithm on H_d and $H_\infty := H$.

The algorithms presented in Section 3.1.3 can be applied to the target specified by

Equation (3.8). For example, the acceptance probability for the RWM algorithm takes the form

$$\alpha(x,y) = \exp\left(\Phi(x) + \frac{1}{2}\left\langle C_d^{-1}x, x\right\rangle - \Phi(y) - \frac{1}{2}\left\langle C_d^{-1}y, y\right\rangle\right) \wedge 1.$$

Whereas for each finite d this acceptance probability results in a well-defined algorithm, an application of the RWM to μ is not well-defined because $\left\langle C^{-1}y, y\right\rangle$ is γ-a.s. infinite. The reason for this is that the Cameron-Martin norm associated to a Gaussian measure of a draw of the same measure is almost surely infinite, consider also [82, 26]. This fact underlies the development in[22] and is explained further in the recent review [38]. This suggests that the RWM algorithm does not perform well for large d. In Article **I**, we have made this fact rigorous which is also reviewed in Section 3.3.1. Subsequently, we discuss appropriate modifications for the IS, the RWM and the MALA algorithm.

The Independence Sampler (IS)

It is straightforward to modify the IS algorithm by choosing a proposal distribution ν that has a density ρ with respect to γ_d such that the proposal and the acceptance probability are given by

$$\begin{aligned} Q_{\mathrm{IS}}(x, dy) &= \rho(y) d\gamma_d(dy) \\ \alpha_{\mathrm{IS}}(x,y) &= \exp[\Phi(x) - \Phi(y)]\frac{\rho(x)}{\rho(y)} \wedge 1. \end{aligned} \tag{3.9}$$

Note that this construction of the independence sampler does not depend on the fact that γ_d is Gaussian. One of the first applications of this algorithm to an infinite dimensional state space has been presented in [161] by considering for the problem of sampling diffusion bridges.

The independence sampler proposes independent moves based on $\rho\gamma_d$ such that the next step of the Markov chain only depends on the previous step through the acceptance probability. In this way, the independence sampler usually performs global moves.

The Preconditioned Crank-Nicolson (pCN) Algorithm

Locally moving Metropolis-Hastings algorithms that are well-defined on infinite dimensional spaces are considered much later. The following modification of the random walk is specific to Gaussian reference measures and consists of centring the proposal around $(1-2\delta)^{\frac{1}{2}}x$ instead of x. This construction gives rise to the following proposal kernel and acceptance probability

$$\begin{aligned} Q_{\mathrm{pCN}}(x,dy) &= \mathcal{N}\left((1-2\delta)^{\frac{1}{2}}x, 2\delta C_d\right)(dy) \\ \alpha_{\mathrm{pCN}}(x,y) &= \exp\left(\Phi(y)-\Phi(x)\right)\wedge 1. \end{aligned} \tag{3.10}$$

It is interesting to note that the proposal corresponds to the steps of an Ornstein-Uhlenbeck process. Moreover, the proposal can also be derived as the preconditioned Crank-Nicolson discretisation of an appropriate SPDE, for more details we refer the reader to [38]. For target measures as in Equation 3.7, this algorithm has been first considered in [22] (and has previously been called P-RWM and PIA algorithm with $(\alpha,\theta)=(0,\frac{1}{2})$). Articles [22],[18] and [38] review different algorithms that are valid on function spaces and present numerical simulations. Moreover, the proposal in Equation (3.10) has already been used in the finite dimensional context, for example in [140].

Note that the proposal kernel Q_{pCN} is reversible with respect to γ_d (this is also the case for $d=\infty$). This follows from the fact that

$$\mathcal{N}(0,C)(dy)\mathcal{N}\left((1-2\delta)^{\frac{1}{2}}y, 2\delta C\right)(dx) = \mathcal{N}(0,C)(dx)\mathcal{N}\left((1-2\delta)^{\frac{1}{2}}x, 2\delta C\right)(dy)$$

which can be checked by calculating the mean and the covariance operator on both sides. In this way, the pCN can be seen as natural generalisation of the standard random walk which is reversible with respect to the Lebesgue measure.

In general, if the proposal kernel Q is reversible with respect to the reference measure γ, the acceptance probability takes a particularly simple form

$$\alpha(x,y) = 1\wedge\frac{d\mu(dy)Q(y,dx)}{d\mu(dx)Q(x,dy)} = 1\wedge\frac{\exp\left(-\Phi(y)\right)}{\exp\left(-\Phi(x)\right)}\frac{d\gamma(dy)Q(y,dx)}{d\gamma(dx)Q(x,dy)} = 1\wedge\exp\left(\Phi(x)-\Phi(y)\right).$$

The Preconditioned Crank-Nicolson Langevin (pCNL) Algorithm

The standard MALA algorithm is based on the Langevin SDE, likewise its generalisation is based on the preconditioned Crank-Nicolson discretisation of the following appropriate SPDE

$$\frac{dx}{dt} = -x - CD\Phi + \sqrt{2C}\frac{dw}{dt}. \tag{3.11}$$

The Δt-time step of the Crank-Nicolson algorithm is given by

$$x_{t+\Delta t} - x_t = (-\frac{1}{2}x_t - \frac{1}{2}x_{t+\Delta t} - CD\Phi)\Delta t + \sqrt{2\Delta t}\xi$$

with $\xi \sim \mathcal{N}(0, C)$. Rearranging gives rise to

$$x_{t+\Delta t} = \frac{1 - \frac{1}{2}\Delta t}{1 + \frac{1}{2}\Delta t}x_t - CD\Phi\frac{\Delta t}{1 + \frac{1}{2}\Delta t} + \frac{\sqrt{2\Delta t}}{1 + \frac{1}{2}\Delta t}\xi.$$

This presentation goes back to [22] and [38]. The Metropolis-Hastings algorithm based on this proposal is called preconditioned Crank-Nicolson-Langevin proposal (pCNL) and can be summarised as

$$\begin{aligned} Q_{\text{pCNL}}(x, dy) &= \mathcal{N}(x + \delta C\nabla \log \pi, 2\delta A)(dy) \\ \alpha_{\text{pCNL}}(x, y) &= 1 \wedge \frac{\rho(y, x)}{\rho(x, y)} \text{ with} \\ \log p(x, y) &= c - \Phi(x) + \frac{1}{2}\langle v - u, D\Phi\rangle + \frac{\Delta t}{4}\langle u + v, D\Phi\rangle + \frac{\Delta t}{4}\left\|C^{\frac{1}{2}}D\Phi\right\|^2. \end{aligned}$$

Note that the algorithm has also been called PIA algorithm with $(\alpha, \theta) = (1, \frac{1}{2})$ in [22] and a semi-implicit MALA scheme in [64].

The pCN algorithm can also be derived from Equation (3.11) for $\Phi = 0$. This can also be seen by substituting $\Delta t = \frac{2\left(-\delta - \sqrt{1-2\delta}+1\right)}{d}$ for the pCNL giving rise to

$$x_{t+\Delta t} = (1 - 2\delta)^{\frac{1}{2}}x_t - CD\Phi(1 - (1 - 2\delta)^{\frac{1}{2}}) + \sqrt{2\delta}\xi$$

which agrees with the pCN for $\Phi = 0$. We would like to mention that there are also non-preconditioned versions of the pCN and the MALA algorithm [22, 38].

3.2 Heuristics and Convergence Rates - A Literature Review

One major focus of the research on Metropolis-Hastings algorithms is the quantification and improvement of their performance. We describe different notions of Markov chain convergence, their implications for the sample average and describe conditions on the target measure under which a given MCMC algorithm converges with respect to a specific notion of convergence. In particular, we focus on the concepts of geometric ergodicity, L^2_μ-spectral gaps and Wasserstein convergence. We close this section by presenting heuristic convergence results in terms of diffusion limits and the expected squared jump distance (ESJD) as they motivate some of the results given in Article **I**.

There exist many different MCMC algorithms that can be used to approximate expectations for a given target measure. However, the performance of each algorithm can vary dramatically depending on parameters of the algorithm and the target measures. There are different ways to quantify the performance of an algorithm, for example, convergence rates to equilibrium, spectral gaps, expected squared jump distance considered subsequently. As the main aim of MCMC algorithms is to approximate expectations computationally, on way to quantify their performance is to compare their computational costs for a given level of accuracy. More precisely, statistical bounds on the error $\mathcal{E}^X_{n,n_0}(f)$ for a function $f : E \to \mathbb{R}$ are expressed in terms of confidence sets. The aim is to obtain a small confidence set for a high confidence level for a small number of steps. In this way, the computational costs can be measured as

$$\text{number of necessary steps} \times \text{cost of one step.} \tag{2.13}$$

In general, the number of necessary steps for a prescribed confidence level depends both on the Monte Carlo error and the bias due to approximations of the posterior density. This relation is described in more detail in Section 2.6 along with ways to speed up MCMC using multi-level methods or appropriate parametric representations of the forward model. However, in the following, we concentrate on the Monte Carlo error of

a single Markov chain given by an MCMC algorithm. Whereas this is straightforward for standard Monte Carlo methods, due to the central limit theorem, much of theoretical research on Metropolis-Hastings algorithms concentrates on obtaining qualitative or quantitative bounds on the error. Under appropriate assumptions on the Markov chain and a function $f : E \to \mathbb{R}$, $f(X_i)$ satisfies a CLT

$$\sqrt{n}\mathcal{E}^X_{n,n_0}(f) \xrightarrow{w} \mathcal{N}(0, \sigma^2_{f,X}). \tag{3.12}$$

In this setting, $\sigma_{f,X}$ is called the asymptotic variance and depends on the Markov chain and f. We refer the reader to [81] for explicit expressions for the asymptotic variance in terms of the integrated autocorrelation or the spectral measure and remark that these expressions can usually not be evaluated.

Asymptotic confidence intervals can be obtained using CLTs. However, the asymptotic variance $\sigma^2_{f,P}$ is usually unknown and has to be estimated. Different estimators are reviewed in Chapter 7 of [29]. Non-asymptotic confidence intervals recently became the focus of the research on MCMC algorithms. Good references are [121, 120] and [167] for results based on geometric and polynomial ergodicity and results based on L^2_μ-spectral gaps, respectively. It is also possible to determine confidence intervals on the basis of bounds on the Mean Square Error (MSE) given by

$$\mathbb{E}\mathcal{E}^X_{n,n_0}(f)^2. \tag{3.13}$$

Whereas a CLT only provides asymptotic confidence intervals, the MSE can be used to construct non-asymptotic confidence intervals using Chebyshev's inequality. Moreover, we note that the size of the confidence intervals can be improved by the median trick which estimates $\mathbb{E}f$ through the median of multiple shorter runs leading to exponential tight bounds. This trick has been developed for MCMC algorithms in [142]. Another good reference is given by [121] (consider Remark 4.6 for the variance of the same estimator).

We focus on error estimates in terms of CLTs and MSEs in the following which are usually obtained through:

1. a formulation of appropriate conditions on the Markov Chain implying the desired properties of $\mathcal{E}^X_{n,n_0}(f)$. These include convergence rates of the Markov chain to equilibrium for different notions of convergence;

2. a formulation of appropriate conditions on the target measure such that Markov chains resulting from an MCMC algorithm satisfy the conditions above.

For a given target measure, there are many different Markov chains satisfying a CLT. Even for the subclass of Metropolis-Hastings algorithms, there is great freedom of picking the proposal distribution. It is then of interest to choose the proposal or the parameters in the proposal, such as the variance of the proposal in the random walk, in a way that reduces $\mathcal{E}_{n,n_0}(f)$. It is usually quite difficult to relate the choices in the proposal kernel to the confidence bound $\mathcal{E}_{n,n_0}(f)$. Therefore heuristics have been developed to indicate the performance of MCMC algorithms reviewed in Section 3.2.2.

For Bayesian inverse problems, it is of particular interest to study the dependence of the error on the dimension of the approximation to the forward model. The first theoretical results addressing the performance of MCMC algorithms with increasing dimension have been obtained in terms of diffusion limits in [157]. More and more steps are needed in order to approximate the diffusion up to a time of order $\mathcal{O}(1)$. In this way, it is possible to quantify heuristically how many additional steps are needed in order to obtain fixed confidence bounds for an increasing dimension of the state space.

We review the literature accordingly by first describing different notions of Markov chain convergence and how they apply to MCMC algorithms before describing optimal scaling results.

3.2.1 Convergence to Equilibrium

We introduce different notions of convergence of a Markov chain to its equilibrium. In particular, we deal with geometric ergodicity and related concepts as the theory is used for a large part of the literature on MCMC. We also present L^2-spectral gaps and Wasserstein convergence because these concepts play an important role for the derivation of our results presented in Articles **I** and **II**.

In each subsequent section, we introduce a different notion of convergence, how it applies to general Markov chains and what implications arise for the Monte Carlo error $\mathcal{E}_{n,n_0}(f)$. In the next step, we discuss sufficient conditions for the target measure of an MCMC algorithm in order for the corresponding chain to satisfy the appropriate conditions. We conclude this section by briefly relating these notions of convergence to others available in the literature. In the following, we employ the notation introduced in Section 3.1.1.

3.2.1.1 Convergence in the Total Variation Distance and Geometric Ergodicity

Most of the theory on MCMC algorithm is formulated with respect to the total variation distance. In this section, we introduce the key concept of geometric ergodicity of Markov chains. We discuss both the implications and how this property can be verified.

For the presentation, we follow [160] containing a concise introduction to geometric ergodicity and CLTs for Markov chains and in particular, how they apply to MCMC algorithms. For a full development of the theory and historical remarks, we point the reader to [137]. A Markov chain with transition kernel P is defined to be geometrically ergodic if

$$\|P^n(x,\cdot) - \mu\|_{\mathrm{TV}} \leq M(x)t^n \quad \text{with } 0 \leq t < 1 \tag{3.14}$$

where the total variation distance is given by

$$\|\nu_1 - \nu_2\|_{\mathrm{TV}} = \sup_A |\nu_1(A) - \nu_2(A)|\,.$$

A very useful characterisation of the total variation distance is given by

$$\|\nu_1 - \nu_2\|_{\mathrm{TV}} = \inf_{X_i \sim \nu_i} \mathbb{P}(X_1 \neq X_2). \tag{3.15}$$

In this way, the total variation can be bounded from above by constructing a coupling between ν_1 and ν_2. The concept of geometric ergodicity has been introduced in [102] for a finite state space and has been extended to general state spaces in [145]. We

refer the reader to [137] for many sufficient conditions and equivalent characterisations of geometric ergodicity. Most often these are formulated under the assumptions of irreducibility and aperiodicity.

Definition 3.3. *(Adapted from [160]) A Markov chain is ϕ-irreducible if there exists a non-zero σ-finite measure ϕ on E such that for all measurable sets $A \subseteq E$ with $\phi(A) > 0$ and for all $x \in E$ there exists a positive integer $n = n(x, A)$ such that $P^n(x, A) > 0$.*

Definition 3.4. *(Adapted from [160]) A Markov chain with stationary distribution μ is aperiodic if there do not exist $m \geq 2$ disjoint subsets $E_1, \ldots, E_m \subseteq E$ with $P(x, E_{i+1}) = 1$ for all $x \in E_i$ $(1 \leq i \leq m)$ and $P(x, E_1) = 1$ for all $x \in E_m$ such that $\mu(E_1) > 0$.*

As most Metropolis-Hastings chains used in practice are aperiodic and this simplifies the presentation, we only concentrate only on aperiodic chains subsequently. For an example of periodic chains in the MCMC literature, we refer to [139].

Implications of Geometric Ergodicity for the Sample Average

Establishing both properties, irreducibility and aperiodicity, is already sufficient to guarantee that the Monte Carlo error of the sample average goes to zero with probability 1 for μ-almost every deterministic starting point. For this result, we refer the reader to Chapter 7 of [156] and Chapter 17 of [137]. For Metropolis-Hastings algorithms, we quote the following result.

Theorem 3.5. *(Theorem 7.2 in [156] and Theorem 4 and Fact 5 in [160]) Suppose that the Metropolis-Hastings Markov chain $(X_n)_{n\in\mathbb{N}}$ is ϕ-irreducible.*

1. *If $h \in L^1_\mu$, then for μ-a.e. deterministic starting point $X_0 = x_0$*

$$\lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} h(X_i) = \mathbb{E}_\mu h.$$

2. *If, in addition, $(X_n)_{n\in\mathbb{N}}$ is aperiodic, then*

$$\lim_{n\to\infty} \|P^n(x, \cdot) - \mu\| .$$

Under the assumption of irreducibility, geometric ergodicity implies a CLT for $S_{n,n_0}(f)$ for all $f \in L_\mu^{2+\delta}$ with $\delta > 0$ as shown for example in [137] and [160]. For reversible Markov chains, in particular, those arising as a Metropolis-Hastings chain, this also holds for $\delta = 0$ as geometric ergodicity in this case implies an L_μ^2-spectral gap. A more detailed discussion of this fact is presented in Section 3.2.1.2. In fact, if the Markov chain is reversible, an L_μ^2-spectral gap and geometric ergodicity are equivalent, see also [164]. Even though CLTs allow the derivation of asymptotic confidence intervals, it is a priori not clear how geometric ergodicity, which measures the convergence to equilibrium, directly translates into the sample error $\mathcal{E}_{n,n_0}(f)$. This relation has rigorously been established in [121]. Even if the Markov chain is started at equilibrium, conditionally the first sample is a point mass and fast convergence to equilibrium therefore causes a fast decay of the autocorrelation. This in turn indicates that the Monte Carlo error is small.

Quantitative bounds on the rate of geometric ergodicity have been obtained in [165] on the basis of a small set and drift condition and are reviewed in [166]. Sharper bounds have been obtained using renewal theory in [13]. Recent progress in renewal theory and a detailed comparison of results of this type can be found in [16].

In particular, these quantitative bounds can be used in order to obtain bounds on the MSE in [121]. Confidence intervals are then obtained using Chebyshev's inequality or the median trick as described at the beginning of Chapter 3. However, working directly with small sets and drift conditions, as has been done in [120], seems to yield better bounds on the Monte Carlo error.

Conditions for Geometric Ergodicity

The most well-known condition geometric ergodicity is based on drift conditions and small sets.

Definition 3.6. *A subset $S \subseteq E$ is small if there exists a positive integer n_0, $\epsilon > 0$ and a probability measure ν on E such that the following minorisation condition holds*

$$P^{n_0}(x, \cdot) \geq \epsilon \nu(\cdot) \qquad \forall x \in S. \tag{3.16}$$

The idea of small sets is due to [59] and has been introduced to unbounded state spaces by Harris in [88]. It is instructive to study the case $S = E$ and $n_0 = 1$. Notice that the small set condition implies that P can be written as

$$P(x, \cdot) = \epsilon\nu(\cdot) + (1 - \epsilon)\tilde{P}(x, \cdot).$$

Thus, the transition kernel can be implemented by throwing an ϵ-coin. If it comes up heads, X_{n+1} is drawn from ν and otherwise from $\tilde{P}(x, \cdot)$. As P preserves μ, $P^n(x, \cdot)$ can be coupled to μ by running a chain X_n started at x and a chain Y started at a random draw from μ. We couple the two chains by using the same coin and the same draw of ν. The characterisation of the total variation distance through coupling allows us to conclude that

$$\|P^n(x, \cdot) - \mu\|_{\mathrm{TV}} \leq \mathbb{P}(X_n \neq Y_n) \leq (1 - \epsilon)^n.$$

It follows that in this case $M(x)$, as in Equation (3.14), does not depend on x, a property called uniform ergodicity. If $S \neq E$, the following drift condition ensures that both X_n and Y_n are sufficiently often in S.

Definition 3.7. *A Markov chain satisfies a drift condition for the set S if there is a function $V : E \to [1, \infty)$, $0 < l < 1$, $b \in \mathbb{R}$ such that*

$$PV \leq lV + b\mathbb{1}_S.$$

Theorem 3.8. *(Harris theorem, see also [160] or [84]) If a Markov chain is ϕ-irreducible and aperiodic, then having a small set and satisfying a drift condition is equivalent to the Markov chain being geometrically ergodic.*

Whereas the proof in [160] is based on the coupling argument, this theorem is classically proved using renewal theory by splitting the evolution of the Markov chain into excursions from the small set and analysing their return times, see [137].

Metropolis-Hastings Algorithms

On the basis of the Markov chain convergence theory presented above, it is possible to derive necessary and sufficient conditions on the target measure μ and the proposal Q such that the Metropolis-Hastings algorithm is geometrically ergodic. This in turn gives rise to CLTs and bounds on the MSE as described above. This approach has, for example, been taken in [135], [163], [93] and [95]. For distributions on $\mathbb{R}$ and under appropriate assumptions on a random-walk proposal, it is both sufficient and necessary for the target measure to have exponential tails [135]. In higher dimensions, this condition is necessary but not sufficient [95]. One of the most general sufficient conditions for appropriate random walk proposals has been formulated in [93] and reads as follows

$$\begin{aligned} \limsup_{|x|\to\infty} \frac{x}{|x|} \cdot \frac{\nabla\pi}{|\nabla\pi|} &< 0 \\ \lim_{|x|\to\infty} \frac{x}{|x|} \cdot \nabla \log \pi &= -\infty. \end{aligned} \tag{3.17}$$

An interesting approach is an application of a transformation h to μ such that $h_\star\mu$ satisfies the conditions in Equation (3.17). This approach has been taken in [97].

In this section, we concentrated on geometric ergodicity, that is exponential convergence of $P(x,\cdot)$ to μ in the total variation distance with a constant depending on x. Polynomial ergodicity is a weaker assumption only assuming that the convergence in the total variation happens at a polynomial rate. Similar to Theorem 3.8, this can be verified in terms of small sets and a weakened version of the drift condition. The seminal paper [94] contains results (c.f. Theorem 3.6) that simplified the existent theory by only assuming one drift condition. Moreover, polynomial ergodicity also implies a CLT for the Monte Carlo error $\mathcal{E}_{n,n_0}(f)$ under stronger assumptions on f. An appropriate result can be found in Theorem 4.2 [94].

At the beginning of Section 3.3, we explain why the total variation distance is not well-suited to study convergence to equilibrium on infinite dimensional state spaces. Therefore we study the convergence with respect to the L^2_μ-spectral gap and the Wasserstein distance in the following section.

3.2.1.2 Spectral Gaps

The entire Section 3.2.1 is devoted to summarising the main convergence results for Markov chains, especially for MCMC algorithms, to their equilibrium. An early focus of the subject has been on Markov chains with finite state spaces for which the transition kernel P can be represented as a matrix. Because convergence concerns the limits of P^n, it is natural to study its eigenvalues. One reason is that the powers of P can be more easily expressed through an eigenvalue decomposition. In particular, if P is reversible, then its transition matrix is symmetric in an appropriate basis and can be diagonalised. We would like to point out the reader to the interesting example of card shuffling that has been studied through the eigenvalues of P in [53]. Moreover, research on card shuffling has lead to a fruitful development in mathematics reviewed in [55]. For a standard introduction to convergence of Markov chains on discrete state spaces, we refer the reader to [126].

A central concept introduced in this section is the notion of L^2-spectral gaps. For a transition kernel P the L^2-spectral gap is given by the absolute difference of largest two elements in the spectrum. It is crucial because not the whole spectrum but just the L^2-spectral gap is sufficient to characterise the convergence speed of a Markov chain.

The remainder of this section is structured as follows. First, we introduce the appropriate spectral theory for the transition kernel allowing us to define L^2-spectral gaps. In a second step, we review the implications for the sample average. We close this section by introducing the method of conductance which is a crucial tool for the results obtained in Articles **I**.

If the transition kernel P of a Markov chain X_n is reversible with respect to the invariant measure μ, then Jensen's inequality implies that P can be identified as a self-adjoint linear operator from L^2_μ to itself. The convergence properties of the Markov chain can now be formulated in terms of the spectrum of the operator P. We setup our notation for the basic concepts but assume that the reader is familiar with the spectral theory of bounded linear operators, the relevant material can be found in [107]. In

general, the spectrum $\sigma(P)$ is defined as

$$\left\{\lambda \in \mathbb{C} \middle| P - \lambda I \text{ is not invertible}\right\}.$$

It is well-known that a self-adjoint operator P satisfies $\sigma(P) \subseteq \mathbb{R}$. We consider Markov operators for which Jensen's inequality implies that $\sigma(P) \subseteq [-1, 1]$. Moreover, the property $P1 = 1$ implies that 1 is the largest eigenvalue of P. The spectral gap $1 - \beta$ is the difference between 1 and the modulus of the second largest element of $\sigma(P)$. This can be expressed as difference between 1 and the spectral radius of the operator P restricted to the orthogonal complement of the space of constant functions denoted by $L^2_{\mu,0}$.

Definition 3.9. *(The L^2_μ-spectral gap) A Markov operator $\mathcal{P}$ with invariant measure μ has an L^2_μ-spectral gap $1 - \beta$ if*

$$\beta = \sup_{f \in L^2_\mu} \frac{\|P(f - \mu(f))\|_2}{\|f - \mu(f)\|_2} = \sup_{f \in L^2_{\mu,0}} \frac{\|Pf\|_2}{\|f\|_2} < 1. \tag{3.18}$$

Observe that for discrete state spaces, one can decompose f in an eigenbasis of P. It follows that all coefficients corresponding to basis elements with eigenvalues in $(-1, 1)$ decay exponentially if P is applied repeatedly.

Note that because of the self-adjointness, we can express the smallest and largest eigenvalue of $P|_{L^2_{\mu,0}}$ by

$$\lambda = \inf_{\|f\|_{L^2_{\mu,0}}=1} \langle Pf, f\rangle \text{ and } \Lambda = \sup_{\|f\|_{L^2_{\mu,0}}=1} \langle Pf, f\rangle \tag{3.19}$$

respectively which allows us to express the spectral gap as

$$1 - \beta = \min\{1 - |\lambda|, 1\text{-}\Lambda\}. \tag{3.20}$$

Definition 3.10. *For a Markov kernel P, we refer to the quantities $1 - |\lambda|$ and 1-Λ as the lower and upper L^2_μ-spectral gap, respectively.*

These notions are introduced because in Article **II** we only obtain a lower bound on the upper L^2_μ-spectral gap. In some sense, an upper L^2_μ-spectral gap is sufficient as it

is possible to modify the chain resulting in an L^2_μ-spectral gap of almost the same size. The modification consists of adding an additional rejection step resulting in the so-called lazy-chain reviewed in Section 3.3.2. Before reviewing conductance and other methods to bound the L^2_μ-spectral gap from below, we give an overview of the implications of such a bound.

Implications of L^2-Spectral Gaps for the Sample Average

The importance of lower bounds on the L^2_μ-spectral gap lies in the fact that it implies a CLT as well as a bound on the MSE. As introduced at the beginning of Section 3.2, both can be used to quantify the Monte Carlo error of the sample average.

Proposition 3.11. *(From [106]) Consider an ergodic Markov chain with transition operator P which is reversible with respect to a probability measure π. Let $f \in L^2$ be such that*

$$\sigma^2_{f,P} = \left\langle \frac{1+P}{1-P} f, f \right\rangle \leq \frac{2\mu((f^2 - \mu(f)))}{(1-\Lambda)} \leq \frac{2\mu((f^2 - \mu(f)))}{(1-\beta)} < \infty,$$

then for $X_0 \sim \mu$ the expression $\sqrt{n}(S_n - \mu(f))$ converges weakly to $\mathcal{N}(0, \sigma^2_{f,P})$.

A lower bound on the L^2-spectral gap can also be used to bound the MSE form above which can be seen in the following proposition.

Proposition 3.12. *(From [167]) Suppose that we have a Markov chain with Markov operator $\mathcal{P}$ which has an L^2_μ-spectral gap $1-\beta$. For $p \in (2, \infty]$, let $n_0(p)$ be the smallest natural number which is greater or equal to*

$$\frac{1}{\log(\beta^{-1})} \begin{cases} \frac{p}{2(p-2)} \log(\frac{32p}{p-2}) \left\| \frac{d\nu}{d\mu} - 1 \right\|_{\frac{p}{p-2}} & p \in (2,4), \\ \log(64) \left\| \frac{d\nu}{d\mu} - 1 \right\|_{\frac{p}{p-2}} & p \in [4, \infty]. \end{cases} \tag{3.21}$$

Then

$$\sup_{\|f\|_p \leq 1} \mathbb{E}\left[(\mu^y(g(u)) - S_{n,n_0}(g))^2 \right] \leq \frac{2}{n(1-\beta)} + \frac{2}{n^2(1-\beta)^2}.$$

We notice that for this proposition it is not enough to know that $\Lambda < 1$. Moreover, we observe that these MSE bounds only hold if the Markov chain is started from an appropriately distributed random point. In contrast, the results obtained in [121] and [120] formulated in terms of geometric ergodicity, small sets and drift conditions can be used even if an MCMC algorithm is started from a deterministic point.

Both results hold under the assumption of a lower bound on the L^2_μ-spectral gap. We review methods to obtain such a bound in the next section.

Bounding the L^2_μ-Spectral Gap

Except for special cases, the L^2_μ-spectral gap cannot be bounded in a straightforward manner. Our main interest lies in the lower bounds because they imply upper bounds on the Monte Carlo error of the sample average which can be seen form the Propositions 3.11 and 3.12. Instead, there are mainly four commonly used methods for obtaining (lower) bounds on the L^2_μ-spectral gap:

1. The rate of geometric ergodicity, see Section 3.2.1.1, allows us to bound the L^2_μ-spectral gap from below. Details on this result can also be found in [158].

2. Functional inequalities can be used to bound the L^2-spectral gap [187]. In particular, we would like to mention Poincaré inequalities which have a similar structure as Equation (3.2.1.2) for the generator of a continuous time process. For a comparison to the approach based on Lyapunov-functions, we recommend [11].

3. The L^2_μ-spectral gap can also be bounded in terms of exponential convergence in a Wasserstein distance reviewed in Section 3.2.1.3.

4. The method of conductance reviewed on the following two pages.

Depending on the Markov kernel P, different methods are applicable and yield bounds with different qualities.

The Method of Conductance

Following [184], we define the conductance of a Markov chain with transition probability P and invariant distribution μ by

$$\mathsf{C} = \inf_{\mu(A)\leq\frac{1}{2}} \frac{\int_A P(x, A^c)d\mu(x)}{\mu(A)}. \tag{3.22}$$

Note that this quotient can be interpreted as the quotient of how likely it is in equilibrium to go from a set A to its complement A^c compared to the equilibrium probability mass of A. A small conductance means that an algorithm can be trapped in some part of the state space for a long while. For this reason a lower bound on the conductance seems desirable. This fact can be made explicit by an application of Cheeger's inequality which allows to bound the upper L^2_μ-spectral gap in the following way

$$\frac{\mathsf{C}^2}{2} \leq 1 - \Lambda \leq 2\mathsf{C}. \tag{3.23}$$

The name Cheeger's inequality is due to a related result in differential geometry which can be found in [32].

For general state spaces, this inequality has been proved in [123] (note that their notion of conductance k satisfies $\mathsf{C} \leq k \leq 2\mathsf{C}$). This way of bounding the spectral gap has been mainly used for discrete state spaces. We recommend [54] for an application to Markov chains on graphs whereas we refer the reader to [175] for an approximation of counting in computer science. In the discrete setting, it is possible to introduce the notion of local conductance. Bounds on the local conductance can be combined in order to obtain a bound on the conductance. This idea has been introduced in [129] and has been used in [96] to prove convergence of Metropolis-Hastings algorithms applied to measures on $\mathbb{R}$ with monotone and log-concave tails.

Conductance has also been used to study the complexity of integrating a function over a convex domain. Article [184] surveys a sequence of articles on this question concentrating on the ball-walk and the hit and run algorithm. This particular problem has also been addressed in an adaptive fashion in [130].

For reversible transition kernels, a lower bound on the L^2_μ-spectral gap is implied by

exponential convergence in different Wasserstein distances. As this type of convergence is important in its own right, we have devoted Section 3.2.1.3 to a summary and detailed literature review.

3.2.1.3 Exponential Convergence in Wasserstein Distances

The analysis of Markov processes in terms of the Wasserstein distance has become increasingly popular. This can be seen in the areas of optimal transport and Wasserstein gradient flows for which we refer the reader to [185]. For many infinite dimensional stochastic processes, convergence to equilibrium can be quantified naturally in a Wasserstein distance. One interesting example is the stochastic 2d Navier-Stokes equation [83]. In contrast to the total variation convergence or the convergence in L^2, the Wasserstein distance depends on a metric d of the state space E. For two measures ν_1 and ν_2, the 1-Wasserstein distance is given by

$$d_W(\nu_1, \nu_2) = \inf_{\pi \in \Gamma(\nu_1, \nu_2)} \int_{\mathbf{E}\times\mathbf{E}} d(x, y)\pi(dx, dy). \tag{3.24}$$

Note that for $\mathbb{R}$ with the standard metric, this implies $d_W(\delta_0, \delta_\epsilon) = \epsilon$ as opposed to $d_{TV}(\delta_0, \delta_\epsilon) = 1$. In fact, d does not need to be a metric. It is sufficient to require that $d : E \times E \to \mathbb{R}$ is symmetric and lower semi-continuous so that $d(x, y) = 0$ is equivalent to $x = y$. However, if d is a metric, the Monge-Kantorovich duality [185]

$$d_W(\nu_1, \nu_2) = \sup_{\|f\|_{Lip(d)}=1} \int f d\nu_1 - \int f d\nu_2$$

holds. We would like to point out that the Wasserstein distance coincides with the total variation distance for the discrete metric. Details for this result can be found in [160]. Convergence of the Markov chain can then be quantified through a decay of $d_W(P^n(x, \cdot), \mu)$ to zero. Note that geometric ergodicity postulates an exponential decay of the same expression for the total variation as can be seen in Equation (3.14). The following notion of the Wasserstein spectral gap is the natural generalisation for Wasserstein distances.

Definition 3.13. *A Markov chain with transition kernel $P(x, dy)$ is said to have a Wasserstein spectral gap if there are $\lambda > 0$ and $C > 0$ such that for any probability measures ν_1 and ν_2*

$$d(\nu_1 \mathcal{P}^n, \nu_2 \mathcal{P}^n) \leq C \exp(-\lambda n) d(\nu_1, \nu_2) \text{ for all } n \in \mathbb{N}. \tag{3.25}$$

Using results from the literature, we demonstrate that this notion is very useful for MCMC algorithms because

1. it implies appropriate bounds on the Monte Carlo error
2. can be verified using the weak Harris theorem from [84].

Implications of the Wasserstein Spectral Gap for the Sample Average

Because we concentrate on Metropolis-Hastings algorithms in this thesis, all the resulting Markov chains are reversible with respect to the target measure μ. In this case, the Wasserstein spectral gap implies an L^2_μ-spectral gap and therefore a CLT and a bound on the mean square error for the sample average as reviewed in the previous section. This implication has been stated in continuous time in Theorem 2.1 in [186]. We also refer the reader to Article **I** which contains a short proof that has been shown to the author by F. - Y. Wang for the discrete case.

Articles [146] and [98] assume a stronger condition than the one stated in Equation (3.25). More precisely, they consider

$$d\left(P(x_1, \cdot), P(x_2, \cdot)\right) \leq c\, d(x, y)$$

for some $c \in [0, 1)$ which they call a Ricci curvature condition. Under this condition, they derive results for the Monte Carlo error of the sample average for Lipschitz functions such as bounds on the MSE. Furthermore, a CLT for the Lipschitz functional can be derived under appropriate assumptions formulated in [114].

The Weak Harris Theorem - A Condition for Wasserstein Spectral Gaps

Verifying a Wasserstein-spectral gap has become much easier with the emergence of the weak Harris theorem stated in [84]. The following can be seen as generalisation of the Harris theorem 3.8. The key idea is that the notion of small-sets can be weakened to that of a d-small set.

Proposition 3.14. *(Weak Harris theorem [84]) Let P be a Markov kernel in a Polish space E and assume that*

- *there is a Lyapunov function $V : E \to [1, \infty)$ such that*

$$\mathcal{P}^n V(x) \leq l^n V(x) + K \text{ for all } x \text{ and } n;, \tag{3.26}$$

- *P is d-contracting for a distance-like function, that is there is a $c \in (0, 1)$ such that for all x, y with $d(x, y) < 1$*

$$d(\mathcal{P}(x, \cdot), \mathcal{P}(y, \cdot)) \leq c \cdot d(x, y); \tag{3.27}$$

- *the set $S = \{x \in E : V(x) \leq 4K\}$ is d-small, that is there is $\epsilon > 0$ such that*

$$d(\mathcal{P}(x, \cdot), \mathcal{P}(y, \cdot)) \leq 1 - \epsilon \quad \forall x, y \in S. \tag{3.28}$$

Then there exists $\tilde{n}$ such that for any two probability measures ν_1 and ν_2 on E we have

$$\tilde{d}(\nu_1 \mathcal{P}^{\tilde{n}}, \nu_2 \mathcal{P}^{\tilde{n}}) \leq \frac{1}{2} \tilde{d}(\nu_1, \nu_2)$$

where $\tilde{d}(x, y) = \sqrt{d(x, y)(1 + V(x) + V(y))}$ and $\tilde{n}(l, K, c, s)$ is increasing in l, K, c and s. This result also implies that there is at most one invariant measure. Moreover, if there exists a complete metric d_0 on E such that $d_0 \leq \sqrt{d}$ and P is Feller on E, then there exists a unique invariant measure μ for P.

Remark 1. *It is possible to trace the constants through the result and to optimise the overall bound. However, it seems that the resulting bounds do not bear practical relevance as they are only a crude lower bound on the convergence rate. Nevertheless, it is*

important to obtain a quantitative theory for exponential convergence in a Wasserstein metric. These results would be in analogy to the quantitative exponential convergence results that are available for the total variation distance. More details can be found in Section 3.2.1.1.

In this thesis, we only consider exponential Wasserstein convergence as stated in Definition 3.13. Recently, sub-exponential convergence results have been obtained for example in [30]. However, so far, it is not clear what the significance of subgeometric convergence in a Wasserstein distance has for the sample average. This is in contrast to the exponential convergence in a Wasserstein distance for which the implications for the sample average have been discussed in the previous section. We would like to mention again that for reversible chains, a Wasserstein spectral gap in fact implies an L^2_μ-spectral gap. Therefore the consequences for the sample average from Section 3.2.1.2 hold, too.

Using the weak Harris theorem, we have shown in Article **I** that the pCN algorithm has a dimension independent Wasserstein spectral gap. As the Wasserstein spectral gap implies an L^2_μ-spectral gap of the same size, the latter is dimension independent, too.

3.2.1.4 Relations and Other Notions

For completeness, we would like to mention other notions of convergence to equilibrium briefly in the following. Closely related to the L^2_μ-spectral gap and equivalent for reversible chains are L^2_μ-exponential convergence and L^2_μ-geometric ergodicity. The thesis [167] shows that all three notions are equivalent for reversible Markov chains. These notions can be generalised to L^p_μ-exponential convergence and L^p_μ-geometric ergodicity. In particular, L^1_μ and L^∞_μ-exponential convergence imply an L^2_μ spectral gap for reversible Markov chains. The corresponding definitions and results can be found in [167]. We also like to mention the log-Sobolev inequality which constitutes a stronger notion than the L^2_μ-spectral gap. It can be formulated in terms of the action of the transition kernel P on a class of functions. For this and related notions, we refer the reader to [80], [52] and [187].

3.2.2 Heuristics for the Choice of the Proposal Distribution

In using Metropolis-Hastings algorithms, there is a great freedom in choosing the proposal distribution. However, this does also include a burden as the performance varies dramatically. For example, the performance of the standard RWM with $\mathcal{N}(0, \sigma^2 I)$ for target measures with a continuous density on $\mathbb{R}^d$ deteriorates as $\sigma \to \{0, \infty\}$. On the one hand, very large proposal variances σ lead to a rejection rate close to one as the proposal typically lies in the tails of the target measure. On the other hand, very small proposal variances σ lead to an acceptance rate close to one but the samples are very correlated. Thus, the problem is to choose σ in a way that is appropriate for the target measure. Because we focus on the case when $d \to \infty$, we consider a sequence of target measures μ_d and proposal variances σ_d^2. In this case, it is of interest to study the choice of σ_d in an asymptotic way. In order to characterise the choice of σ_d, which is optimal for the performance of the Markov chain, its performance has to be quantified. At this point, the heuristic comes into play. As it is seldom possible to find an explicit expression for the convergence rate for a given σ_d, a proxy is used. Typical proxies are the expected squared-jump-distance (ESJD) or the convergence rate of a limiting diffusion process.

In the following, we review the heuristic scaling methods first for general high dimensional target measures before specialising on target measures arising in Bayesian inverse problems.

Heuristics for General High Dimensional Target Measures

For general target measures, we start from the seminal paper [157] by Roberts, Gelman and Gilks. In this article, the authors study quantitatively the optimal choice of σ for the proposal $\mathcal{N}(0, \sigma^2 I)$ in the limit of the dimension $d \to \infty$. They consider sequence of RWM chains X^d with proposal variance σ_d^2 for target measures of the form $\mu(dx) = \pi(x)dx$ where

$$\pi_d(x) = \prod_{i=1}^{d} f(x_i).$$

It is shown that the correct scaling is $\sigma_d = \frac{l}{d}$ and give guidance for the choice of l in terms of the acceptance probability. We can state the result of [157] more precisely by

introducing $T_d : \mathbb{R}^d \to \mathbb{R}$ denoting the projection onto the first coordinate. Then it follows that

$$Z^d(t) = T_d(X^d_{\lfloor t \cdot d \rfloor}) \tag{3.29}$$

converges to the appropriate Langevin diffusion equation

$$dU_t = (h(l))^{\frac{1}{2}}\, dB_t + h(l)\frac{1}{2}\frac{f'(U_t)}{f(U_t)}dB_t.$$

In this equation, $h(l)$ can be seen as linear rescaling of the time. This suggests a larger $h(l)$ corresponding to a faster exploration of the state space in equilibrium and therefore to a small error of the estimator $S_n(f)$. It has been shown in [157] that $h(l)$ is maximised for the choice of l corresponding to an acceptance probability of 0.234. Moreover, in order to define Z^d on the interval $[0, 1]$, d steps of X^d are needed suggesting that the number of steps needed to explore the state space grows like $\mathcal{O}(d^1)$. For the MALA algorithm the same approach suggests that the steps have to increase like $\mathcal{O}(d^{\frac{1}{3}})$ and the acceptance rate should be tuned to 0.574.

Even though this was only shown in a very special context, this fact has been verified later for more general cases under much weaker assumptions, see [159] and [15] for the case of target measures that do not have a product structure. This scaling method can, for example, also be used to study the behaviour of MCMC algorithms for targets that concentrate closer and closer around a manifold [180].

A more direct approach is to consider the expected squared jump distance (ESJD)

$$\text{ESJD} = \mathbb{E}\,\|X_{n+1} - X_n\|^2$$

as an indicator for the performance of an MCMC algorithm. Considering the problem of choosing σ, the ESJD behaves qualitatively in the right way. On the one hand, for a too large proposal variance, the Metropolis-Hastings algorithm rejects with large probability leading to a small ESJD. On the other hand, a small proposal variance leads to a small ESJD as the ESJD can trivially be bounded in terms of the proposal variance. It has been justified that the 0.234 acceptance rule is asymptotically optimal in greater

generality by showing that in this case the ESJD is optimised. For details we refer the reader to [174] and [172]. Moreover, we recommend Chapter 4 of [29] for a recent overview of optimal scaling results.

Heuristics for Target Measures Arising in Bayesian Inverse Problems

The remainder of this section discusses diffusion limits and the ESJD available in the literature for target measures arising in Bayesian inverse problems (see also Section 2.3 and 3.1.4 and). The ESJD has been studied for the RWM and the MALA algorithm applied to target measures similar to those in Equation (3.8) in article [21]. They considered diagonalised covariance matrices with eigenvalues λ_i^2 decaying algebraically like $i^{-2\kappa}$ and they assumed that the proposal variance of the RWM and the MALA algorithm is of the form $i^{-\rho}$. Imposing these conditions, they have proved that if ρ is larger or smaller than $i^{-2\kappa-1}$ for the RWM and $i^{-2\kappa-\frac{1}{3}}$ for the MALA algorithm, then the expected acceptance probability goes to 0 or 1. If the ρ is correctly chosen, then the ESJD is optimised if the acceptance probability is scaled to 0.234 and 0.574, respectively. Both results have also been obtained in terms of diffusion limits in articles [131] for the RWM algorithms and [152] for the MALA algorithm. These results indicate more clearly that the number of necessary steps for a prescribed level of accuracy increases like d^1 for the RWM algorithm and $d^{\frac{1}{3}}$ for the MALA algorithm. In contrast to early scaling results like [157], these results also apply to non-product measures and the diffusion limit is obtained in the Hilbert space on which the Gaussian reference measure is defined. Moreover, it has been shown in [150] that a diffusion limit for the pCN exists with a scaling that is independent of the dimension.

However, the heuristic results in terms of scaling limits and ESJD do not replace rigorous results that quantify the convergence rates. In the next section, we fill this gap by providing rigorous results for the RWM and pCN algorithm through a quantification of their L^2_μ-spectral gaps.

3.3 Contributions of Articles I and II

Reviewing both the literature on rigorous convergence results in Section 3.2.2 and the literature on heuristic scaling results in Section 3.2.1, it is apparent that the heuristic results are much more generally applicable and that they quantify the performance in a much more explicit way. In particular, for infinite dimensional target measures arising from Bayesian inverse problems and approximations thereof, many scaling results are available. For a deeper discussion, we refer the reader again to [22], [131] and [151]. One of the major contributions of Articles **I** and **II** is to provide rigorous convergence results in this setting.

A large part of the available literature on MCMC algorithms is formulated with respect to the total variation distance which relies on small sets and more fundamentally on the irreducibility of the underlying chain. We have reviewed this approach also in Section 3.2.1.1. It is interesting to observe that the most recent handbook [29] does not even mention other approaches based, for example, on L^2_μ-spectral gaps or the Wasserstein distance. However, locally moving Markov chains on infinite dimensional spaces often do not exhibit a small set and are not irreducible. This can be illustrated for the pCN algorithm which has been introduced in Section 3.1.4 as follows. Consider the algorithm started at x_1 and x_2. In this case, the transition kernel of the underlying Markov chain takes the form

$$P(x_i, dy) \quad = \quad \alpha(x_i, y)\mathcal{N}\left((1-2\delta)^{\frac{1}{2}}x_i, 2\delta C_d\right)(dy) + r_i\delta_{x_i}(dy), \quad i = 1, 2.$$

For $x_1 - x_2$ not being an element of the Cameron-Martin space $H_{\mathcal{N}(0,2\delta C_d)}$ of $\mathcal{N}(0, 2\delta C_d)$, the $P(x_i, dy)$ are mutually singular ruling out both irreducibility and existence of the small set. The singularity of the transition kernels is due to the Feldman-Hajek theorem [40, 82]. Moreover, we note that $P^n(x_i, \cdot)$ takes the form of Gaussian mixtures. Then the same argument can be used to see that measures in the mixture for $P(x_1, \cdot)$ and $P(x_2, \cdot)$ are mutually singular if $x_1 - x_2$ is not an element of the Cameron-Martin space. This does not rule out the small set approach that can be used for d-dimensional approximations.

However, it is difficult to obtain lower bounds on the convergence using small sets that do not or only decay slowly as d increases except for the independence sampler. For this reason, we use both the method of L^2 and Wasserstein spectral gaps to tackle convergence issues of the Metropolis-Hastings algorithm in infinite dimensions in Articles **I** and **II**. Subsequently, we introduce these results under consideration of the literature reviewed in Section 3.2.

3.3.1 Article I

In Article **I**, we consider the standard RWM and the pCN algorithm, which is a slight modification of the former, for target measures that arise as approximations to measures that have a density with respect to infinite dimensional Gaussian reference measures. We show that

1. the $L^2_{\mu_d}$-spectral gap of the RWM deteriorates with the dimension of the approximation and

2. that the pCN algorithm has a dimension independent $L^2_{\mu_d}$-spectral gap.

In the following, we give an outline of these contributions in detail. First, we recall the form of the target measures arising from Gaussian-based priors from Section 3.1.4 before describing the deterioration of the RWM in more detail using the conductance. In contrast, we show that the pCN has a dimension independent $L^2_{\mu_d}$-spectral gap. This is established by first proving a Wasserstein spectral gap which implies an $L^2_{\mu_d}$-spectral gap of the same size.

A more direct and explicit approach is taken in [64]. The author considers the pCN algorithm (which is called Ornstein Uhlenbeck proposal in the article) and a function space version of the MALA algorithm. Instead of considering an explicit bound version of a given normed space, a general weaker norm is considered. The bound on the Wasserstein distance is obtained by bounding the rejection probability and its dependence on the current state from above. However, their main results only show the Wasserstein contraction property for starting points in a bounded set and for log-concave measures.

Target Measures

In Article **I**, we consider Metropolis-Hastings algorithms applied to measures with a density with respect to an infinite dimensional Gaussian reference measure and approximations thereof. These measures arise in Bayesian inverse problems as described in Chapter 2 and in the area of conditioned diffusion [85, 22, 18]. As described in Section 3.1.4, these measures are of the form

$$\mu(dx) = M \exp(-\Phi(x))\gamma(dx) \tag{3.2}$$

and natural finite dimensional approximations are given by the truncated Karhunen-Loeve expansion

$$\begin{aligned} \mu_d(dx) &= M_d \exp(-\Phi(x))\gamma_d(dx) \text{ with} \\ \gamma_d(dx) &= \mathcal{L}\left(\sum_{i=1}^{d} \gamma^i \varphi_i \xi_i\right)(dx). \end{aligned} \tag{3.30}$$

Having introduced the relevant target measures and their approximations, we study the performance of the RWM and the pCN applied to μ_d in terms of L^2_μ-spectral gaps subsequently.

3.3.1.1 Bounding the Spectral Gap from Above

The scaling analysis in [21] suggests that the RWM algorithm deteriorates quickly as the dimension increases if the proposal variance is not rescaled appropriately. The deterioration is quantified through the expected acceptance rate going to zero. However, their analysis can be pursued further in order to obtain rigorous upper bounds on the L^2-spectral gap. The following is a slight generalisation of Section 2.4 in Article **I**. We recall the definition of the conductance

$$\mathsf{C} = \inf_{\mu(A)\leq \frac{1}{2}} \frac{\int_A P(x, A^c) d\mu(x)}{\mu(A)} \tag{3.22}$$

and Cheeger's inequality

$$\frac{\mathsf{C}^2}{2} \leq 1 - \Lambda \leq 2\mathsf{C} \tag{3.23}$$

from Section 3.2.1.2.

Our main observation is that there is a simple upper bound for the conductance of a Metropolis-Hastings algorithm because it can only move from a set A if

- the proposed move lies in A^c and
- the proposed move is accepted.

Just considering either event gives rise to simple upper bounds that can be used to make many results from the scaling analysis rigorous. We denote the expected acceptance probability for a proposal from x as

$$\alpha(x) = \int_E \alpha(x, y) dQ(x, dy).$$

Considering only the acceptance of the proposal gives rise to

$$C \leq \inf_{\mu(A) \leq \frac{1}{2}} \frac{\int_A \alpha(x)\mu(dx)}{\mu(A)}.$$

In particular, for any set B such that $\mu(B) \leq \frac{1}{2}$, it follows that

$$C \leq \sup_{x \in B} \alpha(x)$$

and also that

$$C \leq 2\mathbb{E}_\mu \alpha(x).$$

The last result allows us to make scaling results like those in [21] rigorous. Similarly, just supposing that the Metropolis-Hastings algorithm accepts all proposals gives rise to the following bound

$$C \leq \inf_{\mu(A) \leq \frac{1}{2}} \frac{\int_A Q(x, A^c) d\mu(x)}{\mu(A)}.$$

We summarise these results in the subsequent proposition which is an extension to the result presented in Article **I**.

Proposition 3.15. *Let $\mathcal{P}$ be a Metropolis-Hastings transition kernel for a target measure μ with proposal kernel $Q(x, dy)$ and acceptance probability $\alpha(x, y)$. The L^2_μ-spectral gap can be bounded by*

$$1 - \beta \leq 1 - \Lambda \leq 2C \leq 2 \begin{cases} \sup_{x \in B} \alpha(x) & \text{for any } \mu(B) \leq \frac{1}{2} \\ 2\mathbb{E}_\mu \alpha(x) \end{cases} \tag{3.31}$$

and

$$1 - \beta \leq 1 - \Lambda \leq 2C \leq 2 \inf_{\mu(A) \leq \frac{1}{2}} \frac{\int_A Q(x, A^c) d\mu(x)}{\mu(A)}. \tag{3.32}$$

We study the behaviour of the L^2_μ-spectral gap $1 - \beta_d$ for the target measure μ_d in Equation (3.30) as $d \to \infty$ on δ_d. If δ_d decays too slowly, the algorithms propose too large moves leading to small acceptance probabilities allowing us to bound the spectral gap using Equation (3.31). For δ_d decaying too quickly, the behaviour of the proposal can be used as described in Equation (3.32). Asymptotic analysis can then be used to obtain the appropriate bounds. We have executed this research programme for the RWM giving rise to the following result.

Theorem 3.16. *Let $\mathcal{P}_m$ be the Markov kernel and α be the acceptance probability associated with the RWM algorithm applied to γ_m as in (3.30).*

1. *For $\delta_m \sim m^{-a}$, $a \in [0, 1)$ and any p there exists a $K(p, a)$ such that the spectral gap of $\mathcal{P}_m$ satisfies*
$$1 - \beta_m \leq K(p, a) m^{-p}.$$

2. *For $\delta_m \sim m^{-a}$, $a \in [1, \infty)$ there exists a $K(a)$ such that the spectral gap of $\mathcal{P}_m$ satisfies*
$$1 - \beta_m \leq K(a) m^{-\frac{a}{2}}.$$

We observe that the resulting scaling differs from the expected scaling expected from [131]. There is no contradiction as this is just an upper bound on the L^2-spectral gap. It might be possible to obtain a smaller lower bound by choosing a different set for bounding the conductance. However, the bound on the conductance might even be

sharp because the lower bound on the upper spectral gap involves a square, compared to Equation 3.23.

We conclude this section by noting that this simple observation gives a rigorous foundation for many results obtained through scaling.

3.3.1.2 Wasserstein and L^2-Spectral Gaps for the pCN

The second but not less important result concerns the performance of the pCN algorithm introduced in Section 3.1.4. Conditions on the target measures μ and μ_d in the Equations (3.2) and (3.30) are formulated leading to a uniform lower bound on the Wasserstein spectral gap (c.f. Definition 3.13). The bound on the Wasserstein spectral gap is obtained by an application of the weak Harris theorem given in Theorem 3.14. We impose the assumption that Φ is locally Lipschitz and satisfies a growth assumption at infinity giving rise to an appropriate lower bound on the acceptance probability. In the following presentation, we restrict ourselves to the case of Φ being globally Lipschitz which is much more instructive. For the case of Φ being locally Lipschitz, we refer the reader to Article **I**. We impose the following assumptions on Φ:

Assumption 3.17. *There is $R > 0$, $\alpha_l > -\infty$ and a function $r : \mathbb{R}^+ \mapsto \mathbb{R}^+$ satisfying $r(s) \leq \frac{\rho}{2} s$ for all $|s| \geq R$ such that for all $x \in B_R(0)^c$*

$$\inf_{z \in B_{r(\|x\|)}((1-\rho)x)} \alpha(x, z) = \inf_{z \in B_{r(\|x\|)}((1-\rho)x)} \exp\left(-\Phi(z) + \Phi(x)\right) > \exp(\alpha_l). \tag{3.33}$$

Assumption 3.18. *Let Φ in (3.2) be globally Lipschitz with constant L and assume that $\exp(-\Phi)$ is γ-integrable.*

Theorem 3.19 (from Article **I**). *Let the Assumptions 3.17 and 3.18 be satisfied with either*

- *$r(\|x\|) = r\|x\|^a$ where $r \in \mathbb{R}^+$ for any $a \in (\frac{1}{2}, 1)$, then we consider $V = \|x\|^i$ with $i \in \mathbb{N}$ or $V = \exp(v\|x\|)$ or*
- *$r(\|x\|) = r \in R$ for $r \in \mathbb{R}^+$, then we take $V = \|x\|^i$ with $i \in \mathbb{N}$.*

Imposing these assumptions, μ_m (μ) is the unique invariant measure for the Markov chain associated with the pCN algorithm applied to μ_m (μ). Moreover, we define

$$\begin{aligned} \tilde{d}(x,y) &= \sqrt{d(x,y)(1+V(x)+V(y))} \text{ with} \\ d(x,y) &= 1 \wedge \frac{\|x-y\|}{\epsilon}. \end{aligned}$$

Then for ϵ small enough there exists an $\tilde{n}$ such that for all probability measures ν_1 and ν_2 on $\mathcal{H}$ and $P_m\mathcal{H}$, respectively, the following inequalities hold

$$\begin{aligned} \tilde{d}(\nu_1 P^{\tilde{n}}, \nu_2 P^{\tilde{n}}) &\leq \frac{1}{2}\tilde{d}(\nu_1,\nu_2), \\ \tilde{d}(\nu_1 P_m^{\tilde{n}}, \nu_2 P_m^{\tilde{n}}) &\leq \frac{1}{2}\tilde{d}(\nu_1,\nu_2) \end{aligned}$$

for all $m \in \mathbb{N}$.

This result implies an L^2_μ-spectral gap for the transition kernel $P^{\tilde{n}}$ as shown in Proposition 2.8 in Article **I** based on a private communication with F.-Y. Wang. In turn, this yields an L^2_μ-spectral gap for P which can be seen using the spectral theorem. As discussed in Section 3.2.1.2, a lower bound on the L^2_μ-spectral gap gives rise to favourable properties of the sample average such as a CLT which is due to [106] and a bound on the MSE for which we refer to [167].

Idea of the Proof

Theorem 3.19 is proved by verifying the three conditions of the weak Harris theorem (see Proposition 3.14) for the Markov kernel of the pCN algorithm for the distance $d(x,y) = \frac{|x-y|}{\epsilon} \wedge 1$ with ϵ being chosen in due course. For the convenience of the reader, we recall these three conditions

1. the existence of a Lyapunov function for the transition kernel of the pCN algorithm, see Equation (3.26);

2. the transition kernel P is d-contracting, consult Equation (3.27);

3. and the existence of a d-small set for the transition kernel P, see Equation (3.28).

Our presentation follows this structure. Moreover, we use $q_x(\xi)=(1-2\delta)^{\frac{1}{2}}x+\sqrt{2}\xi$ to denote the proposal from the position x with noise ξ.

Existence of a Lyapunov Function for the Transition Kernel of the pCN Algorithm

In the following, we verify that Assumption 3.17 implies that $V = \exp(v\|x\|)$ and $V = \|x\|^i$ are Lyapunov functions for P, that is

$$\mathcal{P}^n V(x) \leq l^n V(x) + K \text{ for all } x \text{ and } n.$$

For x with $\|x\| \geq R$ for a large R, Assumption 3.17 yields a lower bound on the probability that the pCN algorithm moves to a point z such that

$$\|z\| \leq (1-2\delta)^{\frac{1}{2}}\|x\| + r(\|z\|) \leq (1-\delta)^{\frac{1}{2}}\|x\|$$

. In particular, $V(z) \leq \theta V(x)$. Let A be the event that

$$A = \left\{\sqrt{2\delta}\xi \leq r(x) \text{ and } q_x(\xi) \text{ is accepted}\right\},$$

then the above implies that

$$\mathbb{E}_x\left(V(X_1); A\right) \leq \theta V(x)\mathbb{P}_x(A).$$

The Lyapunov property can then be deduced from obtaining a bound of the form

$$\mathbb{E}_x\left(V(X_1); A^c\right) \leq V(x)\mathbb{P}_x(A^c) + K.$$

For details, we refer the reader to Lemma 3.2 in Article **I**.

Coupling of the Transition Kernels $P(x,\cdot)$ and $P(y,\cdot)$

Both the d-contraction property and the characterisation of d-small sets require a bound on the Wasserstein distance of the transition kernel because the Wasserstein distance is

defined through the infimum over all couplings. Hence an upper bound can be obtained by picking a particular coupling.

We choose the basic coupling which drives the pCN algorithm started at x and y with the same noise and which uses the same random variable for the accept and reject step. More precisely, this can be represented as follows

$$q_x(\xi) = (1-2\delta)^{\frac{1}{2}}x + \sqrt{2\delta}\xi \qquad q_y(\xi) = (1-2\delta)^{\frac{1}{2}}y + \sqrt{2\delta}\xi$$
$$\tilde{x} = q_x(\xi)\chi_{[0,\alpha(x,q_x)]}(U) + x \cdot \chi_{(\alpha(x,q_x),1]} \qquad \tilde{y} = q_y(\xi)\chi_{[0,\alpha(y,q_y)]}(U) + y \cdot \chi_{(\alpha(y,q_y),1]}$$

with $U \sim \mathcal{U}(0,1)$. Note that in this case, $\mathcal{P}(x,\cdot) = \mathcal{L}(\tilde{x})$ and $\mathcal{P}(y,\cdot) = \mathcal{L}(\tilde{y})$. This implies that the following expression

$$\pi_{\text{Basic}} = \mathcal{L}\left((\tilde{x},\tilde{y})\right)$$

defines a coupling which we call the basic coupling.

The d-Contraction Property of the Transition Kernel of the pCN Algorithm

For x and y such that $d(x,y) < 1$, we use the basic coupling to obtain an upper bound on $d\left(P(x,\cdot), P(y,\cdot)\right)$. This can be achieved by considering the following three cases:

1. Both the algorithm started at x and the algorithm started at y accept the proposals $\tilde{x}$ and $\tilde{y}$.
2. Both the algorithm started at x and the algorithm started at y reject the proposals $\tilde{x}$ and $\tilde{y}$.
3. One algorithm accepts the proposal and the other rejects the proposal.

In the first case, we obtain $d(\tilde{x},\tilde{y}) \leq (1-2\delta)^{\frac{1}{2}}d(x,y)$ and in the second we get $d(\tilde{x},\tilde{y}) = d(x,y)$. If one algorithm rejects, then $d(x,\tilde{y})$ and $d(\tilde{x},y)$ can be bounded from above by one due to the choice of d. Overall, this allows us to bound

$$\begin{aligned} d\left(P(x,\cdot), P(y,\cdot)\right) \quad \leq \quad & (1-2\delta)^{\frac{1}{2}}d(x,y)\mathbb{P}(\text{both accept}) + d(x,y)\mathbb{P}(\text{both reject}) \\ & +1 \cdot \mathbb{P}(\text{only one accepts}). \end{aligned}$$

The probability of the case when only one algorithm accepts can be bounded as follows

$$\begin{aligned}\mathbb{P}(\text{only one accepts}) &\leq \int_X |\alpha(x, q_x)(\xi) - \alpha(y, q_y)(\xi)| \, d\gamma(\xi) \\ &\leq \int |\Phi(q_x) - \Phi(q_y)| + |\Phi(x) - \Phi(y)| \, d\gamma(\xi) \\ &\leq 2L \, |x - y| \leq 2L\epsilon d(x, y).\end{aligned}$$

By choosing ϵ small enough, we obtain a d-contracting property as $\mathbb{P}$(both accept) can be bounded below using Assumption 3.17.

The Existence of a d-Small Set

Both the Lyapunov function and the d-contraction property are persevered when considering the n-step transition kernel P^n instead of the one-step kernel P. For a fixed bounded set S, we can choose n large enough so that the set is a d-small set. The simplest way to see this is to bound the probability that two pCN algorithms, one started at x and the other one at y, both accept n-times in a row. We denote this event by A and bound its probability using the basic coupling. In this case, the distance $\|X_n - Y_n\|$ decreases at each step by a factor of $(1 - 2\delta)^{\frac{1}{2}}$. If we choose n large enough, $\|X_n - Y_n\| \leq \frac{\epsilon}{2}$ regardless of x and y in S. This results in the following upper bound

$$d\left(P(x, \cdot), P(y, \cdot)\right) \leq \frac{1}{2}\mathbb{P}(A) + 1\mathbb{P}(A^c) < 1$$

for all x and y in S.

Dimension Independence

In order to make the above discussion rigorous, we have to bound the probability of appropriate events. These bounds have to be dimension independent in order to obtain a result that is dimension independent, too. This is possible as most quantities can be written as a monotonic function of the norm for which the following result holds.

Lemma 3.20. *(From Article **I**) Let $f : \mathbb{R} \to \mathbb{R}$ be monotonically increasing, then*

$$\int f(\|\xi\|) d\gamma_m(\xi) \leq \int f(\|\xi\|) d\gamma(\xi)$$

and in particular,

$$\gamma_m(B_R(0)) \geq \gamma(B_R(0)). \tag{3.34}$$

Proof. The truncated Karhunen-Loeve expansion relates γ_m to γ and yields

$$\sum_{i=1}^{m} \lambda_i \xi_i^2 \leq \sum_{i=1}^{\infty} \lambda_i \xi_i^2.$$

Hence the result follows by monotonicity of the integral and of the function f

$$\int f\left(\|\xi\|\right) d\gamma_m(\xi) = \mathbb{E}\left(\sqrt{f\left(\sum_{i=1}^{m} \lambda_i \xi_i^2\right)}\right) \leq \mathbb{E}\left(\sqrt{f\left(\sum_{i=1}^{\infty} \lambda_i \xi_i^2\right)}\right) = \int f\left(\|\xi\|\right) d\gamma(\xi).$$

Inserting $f = \chi_{B_R(0)^c}$, this yields Equation (3.34). □

The lemma above is crucial as it implies that the bounds for the conditions of the weak Harris theorem are dimension independent. In this way, we obtain our main result for the pCN algorithm. For a concluding section on Article **I** and directions of further research, we refer the reader to Section 3.4.

3.3.2 Article II

In the previous section, we have discussed the results of Article **I** and in particular, the dimension-independent performance of the pCN based on a quickly converging proposal Markov chain for the Gaussian reference measure. In contrast, Article **II** considers how a quickly converging Markov chain for non-Gaussian reference measures can be used to obtain quickly converging Metropolis-Hastings algorithms for the corresponding target measures. If the density of the target measure is bounded from above and away from zero, the Independence Sampler (IS) is uniformly ergodic as the state space is a small set [135]. For Metropolis-Hastings algorithms with local proposals, this theory does not apply which is one reason why the independence sampler is the only MCMC

algorithm considered in [92]. In Article **II**, we address this problem by considering local proposal kernels Q that are reversible and have an $L^2_{\mu_0}$-spectral gap with respect to a given reference probability measure μ_0. If the density of the target measure μ is bounded above and below, then the lazy version of the resulting Metropolis-Hastings algorithm has an L^2_μ-spectral gap. A similar fact has been proved for the Gibbs sampler for perturbations of Gaussian measures in [4]. However, it is not clear how it could be generalised to the random walk proposal for arbitrary reference measures.

For the elliptic inverse problem with a prior based on a series expansion with uniformly distributed coefficients introduced in Section 2.2, we have constructed a class of appropriate proposals. We call the resulting algorithm Reflection Random Walk Metropolis (RRWM) algorithm. Subsequently, we present our main result and introduce this class of algorithms. For the corresponding simulations, we refer to Article **II**.

3.3.2.1 L^2_μ-Spectral Gaps of Lazy Metropolis-Hastings Algorithms

In the following, we present our main result proved in Article **II** stating that the L^2_μ-spectral gap of the lazy version of the Metropolis-Hastings algorithm can be bounded in terms of the $L^2_{\mu_0}$-spectral gap of the proposal chain. In contrast to Section 3.3.1.1, we use the characterisation of the L^2-spectral gap in terms of the associated Dirichletform in order to obtain a lower bound thereof. However, before stating and proving our main theorem, we introduce the lazy version associated to a Markov chain and explain the relation between their L^2_μ-spectra.

This construction is necessary because the our methods only yields a lower bound on the upper L^2_μ-spectral gap $1-\Lambda$ of P. In order to obtain a lower bound on the L^2_μ-spectral gap also a lower bound of the lower L^2_μ-spectral gap is required. This problem can be circumvented by considering the lazy version of a Markov chain with transition kernel $\tilde{P}$ which is given by $\tilde{P} = \frac{1}{2}(I+P)$. This transition can be interpreted by throwing a coin and

- if it comes up heads, the Markov chain performs a step according to P;
- if it comes up tails, the Markov chain does not make a transition.

Subsequently, we show that $\tilde{P}$ is positive definite and we derive a lower bound on its L^2_μ-spectral gap. The construction of the lazy Markov chain is well-known in the literature and goes at least back to [127]. If -1 is in the spectrum $\sigma(P)$, this corresponds to a period 2 behaviour which can be broken through considering the lazy version.

Following Section 3.2.1.2, we consider the Markov kernel P as an operator on either L^2_μ or $L^2_{\mu,0}$, where the latter denotes the orthogonal complement of the subspace of constant functions in L^2_μ. The spectrum $\sigma_{L^2_{\mu,0}}(P)$ is then contained in $[\lambda, \Lambda]$ where

$$\lambda = \inf_{\|f\|_{L^2_{\mu,0}}=1} \langle Pf, f\rangle \text{ and } \Lambda = \sup_{\|f\|_{L^2_{\mu,0}}=1} \langle Pf, f\rangle$$

which is the characterisation of the smallest and largest eigenvalue of a self-adjoint operator, respectively. More generally, for a self-adjoint operator $A : H \to H$ the smallest and largest eigenvalue are characterised by

$$\lambda^H_{\min}(A) = \inf_{f\in H} \frac{\langle Af, f\rangle}{|f|^2} \text{ and } \lambda^H_{\max}(A) = \sup_{f\in H} \frac{\langle Af, f\rangle}{|f|^2}, \tag{3.35}$$

respectively. The L^2_μ-spectral gap can then be represented as

$$1 - \beta = \min\{1 - \lambda, 1\text{-}\Lambda\}. \tag{3.36}$$

This motivates the following notions.

Definition 3.21. *For a Markov kernel P, we refer to the quantities $1 - \lambda$ and $1 - \Lambda$ as the lower and upper L^2_μ-spectral gap, respectively.*

These notions are introduced because we will only be able to obtain a lower bound for the upper L^2_μ-spectral gap. In some sense, an upper L^2_μ-spectral gap is sufficient as the lazy version of the chain has an L^2_μ-spectral gap of almost the same size. The spectrum of the transition kernel $\tilde{P}$ of the lazy version can be derived easily from that of

P. The smallest eigenvalue $\tilde{\lambda}$ and the largest eigenvalue $\tilde{\Lambda}$ of $\tilde{P}$ restricted to $L^2_{\mu,0}$ satisfy

$$\begin{aligned}\tilde{\lambda} &= \frac{1+\lambda}{2} \\ \tilde{\Lambda} &= \frac{1+\Lambda}{2}.\end{aligned}$$

In particular, $\sigma_{L^2_{\mu,0}}(\tilde{P}) \subseteq [\frac{1+\lambda}{2}, \frac{1+\Lambda}{2}]$. Thus, $\tilde{P}$ has an L^2_μ-spectral gap if $\Lambda < 1$ because this implies that $0 \leq \tilde{\lambda}$ and $\tilde{\Lambda} < 1$. The L^2_μ-spectral gap $1 - \tilde{\beta}$ of the lazy chain with transition kernel $\tilde{P}$ is given by

$$1 - \tilde{\beta} = \frac{1 - \tilde{\Lambda}}{2}.$$

Depending on the bound on Λ, there is $p \neq \frac{1}{2}$ such that $P' = pI + (1-p)P$ has a larger L^2_μ-spectral gap than $\tilde{P}$. However, we stick to the choice $p = \frac{1}{2}$ as our bounds are not sharp at any rate. Instead, we are interested in proving the existence of a dimension independent lower bound on the spectral gap.

3.3.2.2 Lower Bounds on the Upper L^2-Spectral Gap of the Metropolis-Hastings Kernel

We will first rewrite the characterisation of the upper L^2-spectral gap in order to use the lower bound on the upper L^2-spectral gap for the Metropolis-Hastings kernel. The upper spectral gap $1 - \lambda^{L^2_0(\mu)}_{\max}(P)$ is given by the smallest eigenvalue $\lambda^{L^2_0(\mu)}_{\min}$ of $I - P$ on $L^2_0(\mu)$ and can be characterised as

$$1 - \lambda^{L^2_0(\mu)}_{\max}(P) = \inf_{f \in L^2_0(\mu)} \frac{\langle (I-P)f, f\rangle}{|f|^2} = \inf_{f \in L^2(\mu)} \frac{\langle (I-P)\Pi f, \Pi f\rangle}{|\Pi f|^2} \tag{3.37}$$

where $\Pi : L^2_0(\mu) \to L^2(\mu)$ is the orthogonal projection onto $L^2_0(\mu)$ given by

$$\Pi f = f - \mu(f).$$

The denominator can be rewritten as

$$\begin{aligned} |\Pi f|^2 &= Var_\mu(f) = \int (f-\mu(f))^2 \, d\mu \\ &= \int f^2 d\mu - \mu(f)^2 = \frac{1}{2}\int \mu(dx)\mu(dy)\,(f(x)-f(y))^2\,. \end{aligned} \tag{3.38}$$

Moreover, the nominator in (3.37) can be rewritten as

$$\begin{aligned} \langle (I-P)(f-\mu(f)), f-\mu(f)\rangle &= \langle (I-P)f, f-\mu(f)\rangle = \langle (I-P)f, f\rangle \\ &= \int \mu(dx)P(x,dy)\left(f(x)^2 - f(x)f(y)\right) dy \\ &= \frac{1}{2}\int \mu(dx)P(x,dy)\,(f(x)-f(y))^2\, dy =: \mathcal{E}_\mu^P(f,f). \end{aligned}$$

The bilinear form $\mathcal{E}(f,f)$ is called Dirichlet form associated with the Markov kernel P. Reversible Markov processes can be studied in terms of their Dirichlet form. We refer the reader to [169] for a short survey for time-continuous Markov processes, to [73] for generalities of the theory and to [126] for a review for discrete Markov chains. However, we will not make use of this theory. Instead the characterisation of the upper L^2-spectral gap

$$1-\lambda_{\max}^{L_0^2(\mu)} = \inf_{f\in L^2(\mu)} \frac{\mathcal{E}_\mu^P(f,f)}{\mathrm{Var}(f)} \tag{3.39}$$

is sufficient.

Having disposed of this preliminary step, we are now in the position to state and prove our main theorem presented in Article **II**. The considerations above show that it is enough to bound the second largest eigenvalue Λ in order to bound the spectral gap of the lazy chain. The following theorem uses the characterisation in Equation (3.39) and is an adaptation of the results from [56] for continuous state spaces and the Metropolis-Hastings kernel.

Theorem 3.22. *Suppose that the proposal kernel Q satisfies a lower bound on the upper $L^2_{\mu_0}$-spectral gap $1-\lambda_{max}^{L_0^2(\mu)}(P) > 0$ and the target measure takes the form*

$$\mu = \frac{L}{Z}\mu_0.$$

Then the upper L^2_μ-spectral gap satisfies

$$\left(1-\lambda_{max}^{L^2_0(\mu_0)}(Q)\right)\frac{L^{\star 3}}{L_\star^3}\geq 1-\lambda_{max}^{L^2_0(\mu)}(P)\geq \frac{L_\star^4}{L^{\star 4}}\left(1-\lambda_{max}^{L^2_0(\mu_0)}(Q)\right)$$

where $L_\star := \inf L \leq L \leq \sup L = L^\star$. In particular, the lazy version $\tilde{P}$ has an L^2_μ-spectral gap $1-\beta_{lazy}$ satisfying

$$\frac{1}{2}\left(1-\lambda_{max}^{L^2_0(\mu_0)}(Q)\right)\frac{L^{\star 3}}{L_\star^3}\geq 1-\beta_{lazy}\geq \frac{1}{2}\frac{L_\star^4}{L^{\star 4}}\left(1-\lambda_{max}^{L^2_0(\mu_0)}(Q)\right).$$

Proof. From Equation (3.38) follows that

$$\frac{L_\star^2}{Z^2}\mathrm{Var}_\mu(f)\leq \mathrm{Var}_{\mu_0}(f)\leq \frac{L^{\star 2}}{Z^2}\mathrm{Var}_\mu(f).$$

Similarly, we notice that

$$\begin{aligned}
\mathcal{E}_\mu^P(f,f) &= \frac{1}{2}\int \mu_0(dx)Q(x,dy)\frac{L}{Z}\alpha(x,y)\left(f(x)-f(y)\right)^2\\
&\geq \frac{L_\star}{Z}\alpha_\star\frac{1}{2}\int \mu_0(dx)Q(x,dy)\left(f(x)-f(y)\right)^2\\
&\geq \frac{L_\star^2}{ZL^\star}\left(1-\lambda_{\max}^{L^2_0(\mu_0)}(Q)\right)\mathrm{Var}_{\mu_0}(f)\\
&\geq \frac{L_\star^4}{Z^3L^\star}\left(1-\lambda_{\max}^{L^2_0(\mu_0)}(Q)\right)\mathrm{Var}_{\mu}(f)\\
&\geq \frac{L_\star^4}{L^{\star 4}}\left(1-\lambda_{\max}^{L^2_0(\mu_0)}(Q)\right)\mathrm{Var}_{\mu}(f).
\end{aligned}$$

Thus, we can conclude that

$$1-\lambda_{\max}^{L^2_0(\mu)}(P)=\inf_{f\in L^2(\mu)}\frac{\mathcal{E}_\mu^P(f,f)}{\mathrm{Var}(f)}\geq \frac{L_\star^4}{L^{\star 4}}\left(1-\lambda_{\max}^{L^2_0(\mu_0)}(Q)\right).$$

The other inequality is obtained in the following way

$$\begin{aligned}
\mathcal{E}_\mu^Q(f,f) &= \frac{1}{2}\int \mu_0(dx)Q(x,dy)\frac{L}{Z}\alpha(x,y)\left(f(x)-f(y)\right)^2\\
&\geq \frac{L_\star}{Z}\frac{1}{2}\int \mu(dx)P(x,dy)\left(f(x)-f(y)\right)^2
\end{aligned}$$

$$\geq \frac{L_\star}{Z}\left(1-\lambda_{\max}^{L_0^2(\mu)}(P)\right)\mathrm{Var}_\mu(f)$$
$$\geq \frac{L_\star^3}{Z^3}\left(1-\lambda_{\max}^{L_0^2(\mu)}(P)\right)\mathrm{Var}_{\mu_0}(f).$$

The result for the lazy chain can be derived by an application of the theory presented at the beginning of this section. □

The statement of Theorem 3.22 highlights the insight that the choice of the reference measure is crucial for designing efficient sampling algorithms on function spaces. A typical example is the use of a Markov chain that has an $L^2_{\mu_0}$-spectral gap where μ_0 is the prior of a Bayesian inverse problem. If the likelihood is bounded, then the Metropolis-Hastings algorithm with this chain as the proposal has an $L^2_{\mu^y}$-spectral gap with μ^y being the posterior. However, the result is not limited to this situation.

3.3.2.3 The Reflection Random Walk Metropolis Algorithm for Uniform Series Priors

We describe the construction of reversible proposals for posteriors arising from uniform series priors satisfying the conditions of Theorem 3.22. These priors are commonly used for the particular Bayesian inverse problem (EIP) of reconstructing the diffusion coefficient from the pressure as introduced in Section 2.2.2. We recall that this prior is given by

$$a_u(x) = \bar{a}(x) + \sum_{j\geq 1} u_j \psi_j(x), \quad x \in D \tag{2.7}$$
$$\mu_0 = \bigotimes_{j=1}^{\infty} \mathcal{U}(-1,1). \tag{2.8}$$

Constructing a reversible proposal with an $L^2_{\mu_0}$-spectral gap can be reduced to constructing such a proposal for $\mathcal{U}(-1,1)$ because of the tensorisation property of the L^2-spectral gap, for which we refer the reader to [80] and [12].

Such a proposal can also be obtained from a Metropolis-Hastings chain which is explained in more detail in Article **I**. However, we follow the more straightforward approach of considering a proposal based on a repeated reflection of random walks. A

repeated reflection at the boundary -1 and 1 can be represented by the following map

$$R(x) = \begin{cases} y & y \leq 1 \\ 2 - y & 1 < y < 3 \\ -4 + y & 3 \leq y \leq 4 \end{cases}, \text{ where } y = x \bmod 4.$$

Therefore the proposals take the form

$$Q^{\mathrm{RRWM}}(x, dy) = \mathcal{L}\left(R(x + \xi)\right), \quad \text{with } \xi \sim \tilde{q}$$

where $\mathcal{L}$ denotes the law of a random variable. The resulting density q^{RRWM} of Q^{RRWM} can be written using the change of variable formula

$$q^{\mathrm{RRWM}}(x, dy) = \sum_{k \in \mathbb{Z}} \tilde{q}(x - y + 4k) + \tilde{q}(x + y + 4k + 2). \tag{3.40}$$

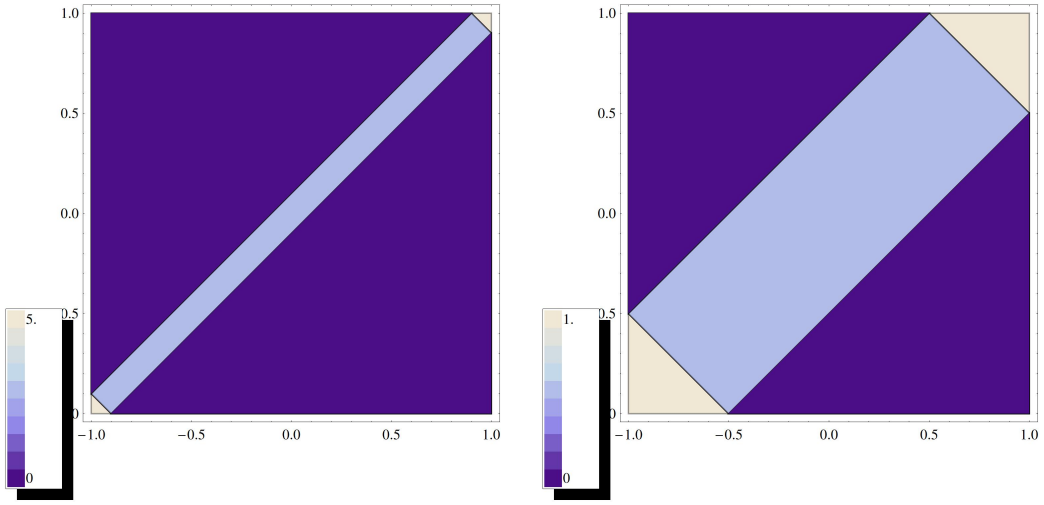
We focus on the choices

$$\begin{aligned} \tilde{q}_\epsilon^{\mathrm{RURWM}} &= \mathcal{U}(-\epsilon, \epsilon) \text{ and} \\ \tilde{q}_\epsilon^{\mathrm{RSRWM}} &= \mathcal{N}(0, \epsilon^2) \end{aligned}$$

which we call Reflection Uniform Random Walk Metropolis (RURWM) and Reflection Standard Random Walk Metropolis (RSRWM) algorithm, respectively. In the Figures 3.1 and 3.2, we plot the density of the one dimensional transition kernel of the RURWM and the RSRWM algorithm based on Equation (3.40). Higher dimensional proposals are obtained by applying this proposal to each component independently.

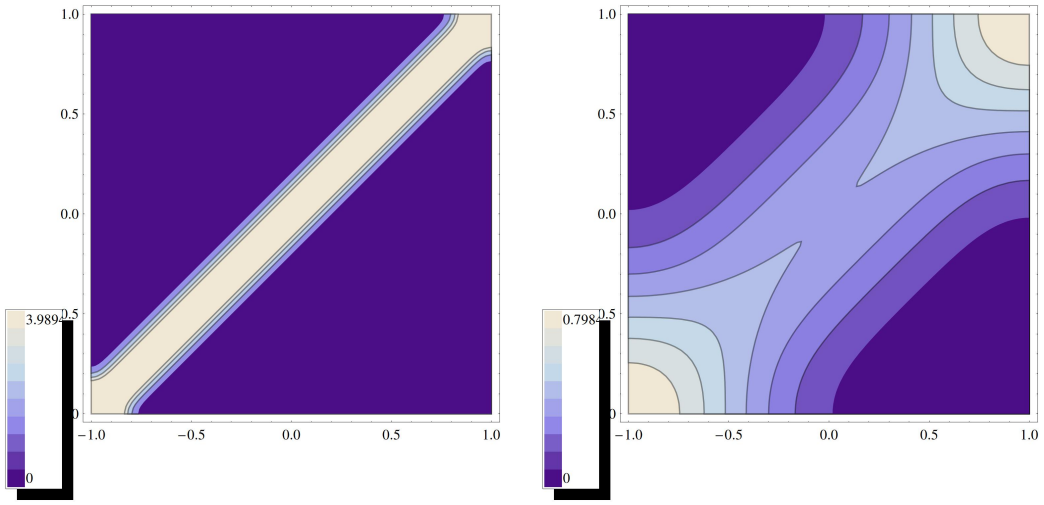
In Article **II**, we show that the $L^2_{\mu_0}$-spectral gaps for both Markov chains are bounded away from zero and are of order ϵ. Using our main theorem, we can conclude that the lazy version of the resulting Metropolis-Hastings algorithm has L^2_μ -spectral gaps.

The lower bound on the L^2_μ-spectral gap of the lazy chain obtained from Theorem 3.22 is best for the independence sampler based on proposing independent samples form μ_0 because its L^2_μ-spectral gap is given by $1 - \beta = 1$ and therefore it is as large as



(a) $\epsilon = 0.1$ (b) $\epsilon = 0.5$

Figure 3.1: Transition density for the RURWM



(a) $\epsilon = 0.1$ (b) $\epsilon = 0.5$

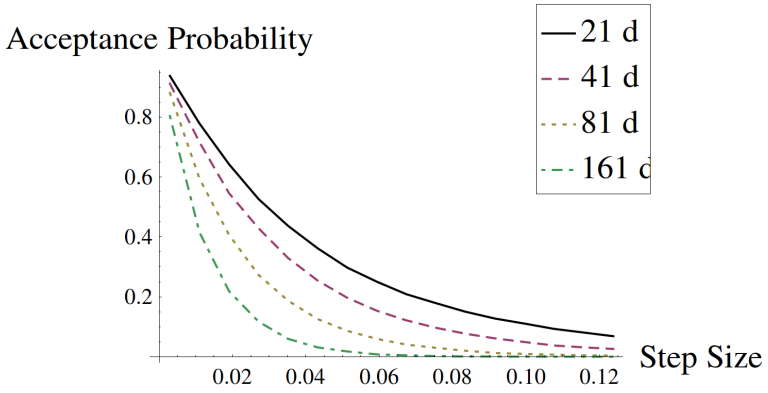
Figure 3.2: Transition density for the RSRWM

possible. However, this result only concerns lower bounds. In contrast, the simulations in Section 5 of Article **II** suggest that the RURWM and the RSRWM algorithm converge to equilibrium more quickly than the IS for peaked measures. More precisely, we implement the inverse problem introduced in Section 2.2.2 in one spatial dimension. Notice that the representation of the diffusion coefficient through a series expansion leads to a high-dimensional inference problem.

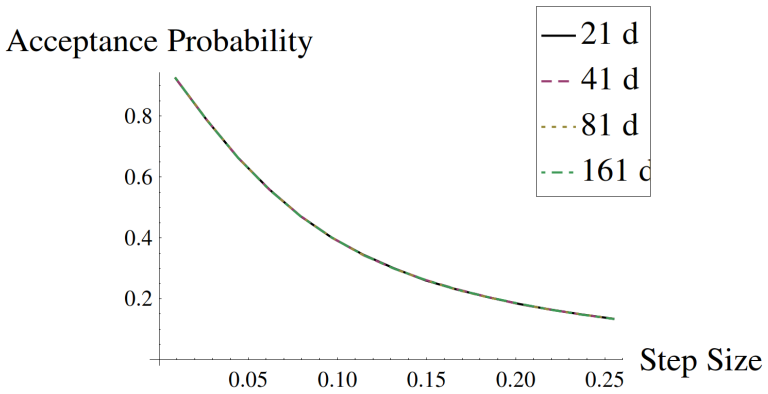
This Bayesian inverse problem is precisely set up in Section 5.1 of Article **II**. In the simulations, we compare the IS, RWM, RURWM and RSRWM algorithm in terms of the acceptance probability and autocorrelation. Note that there is in general a relation between the average acceptance probability and the step-size. If the target density is continuous, it is plausible that small step sizes lead to large acceptance probabilities and large step sizes lead to low acceptance probabilities because it is very likely for the proposal to lie in the tails. For the RWM algorithm, this can be seen from the definition of the acceptance probability in Equation (3.6). In Figure 3.3, it has been demonstrated that for the RURWM and the RSRWM algorithm this relationship does not depend on the dimension in contrast to the RWM algorithm.

A more direct quantification of the performance is given through the autocorrelation. It is well-known that the integrated autocorrelation agrees with the asymptotic variance of the Markov chain CLT. For different representations of the asymptotic variance, we refer the reader to [81]. In Section 5.3 of Article **II**, we compare the autocorrelation of the IS, RWM, RURWM and RSRWM algorithm.

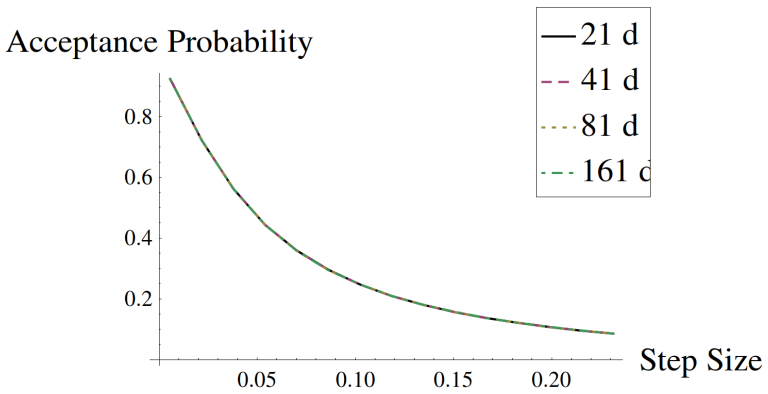
In this section, we summarised the construction of the Reflection Random Walk algorithm which is particularly suited for sampling the posterior of the elliptic inverse problem with a prior based on a series expansion with uniformly distributed coefficients introduced in Section 2.2.2. For this example, we have constructed a proposal which satisfies the conditions of our main theorem indicating its importance for applications. Subsequently, we draw an overall conclusion about this chapter and Articles **I** and **II**.



(a) Acceptance rate vs. step size for the RWM algorithm



(b) Acceptance rate vs. step size for the RURWM algorithm



(c) Acceptance rate vs. step size for the RSRWM algorithm

Figure 3.3: Dependence of the acceptance probability on the dimension

3.4 Conclusion and Avenues of Further Research

Subsequently, we draw a conclusion of this chapter and Articles **I** and **II** reviewed in Section 3.3. The chapter has provided an exposition and literature review of Metropolis-Hastings algorithms both on finite and infinite dimensional state spaces in the Sections 3.1 and 3.2. Our research fills a gap by providing one of the first dimension independent rigorous convergence results for locally moving MCMC chains.

3.4.1 Article I

We have provided the first dimension independent convergence result for a locally moving Metropolis-Hastings algorithm, the preconditioned Crank-Nicolson (pCN). In contrast, we also quantified how the standard RWM algorithm deteriorates in terms of the L^2_μ-spectral gap. For the pCN, we considered target measures given as densities with respect to a Gaussian measure on separable Hilbert spaces. The results for the pCN algorithm can also be verified on separable Banach spaces. The finite dimensional approximations in Article **I** have been based on the Karhunen-Loeve expansion. For target measures defined on Banach spaces, we would use a general Gaussian series (c.f. Section 3.5 in [26]) instead. The dimension independence in the present form is based on an explicit calculation based on the Karhunen-Loeve expansion, see Lemma 3.20. For more general target measures, the same result is due to Theorem 3.3.6 in [26].

Moreover, we have demonstrated that the theory based on small sets and ϕ-irreducibility, as developed in [137], is not well-suited for dealing with convergence to equilibrium for stochastic processes on infinite dimensional state spaces. Instead, we use the theory based on the weak Harris theorem which has been developed in [84].

A natural direction for further investigation is to prove convergence results for other MCMC algorithms that are well-defined on function spaces. Natural candidates are the preconditioned Crank-Nicolson Langevin (pCNL) and Hybrid Monte Carlo algorithm.

For the pCNL, which is also called MALA algorithm with semi-implicit proposal, results of this type have been obtained in [64]. However, the main result of the article only shows a contraction of the Wasserstein distance in bounded sets under restricted

log-concavity assumptions. This could be a starting point of an analysis based on the weak Harris theorem because this result yields a 'large' d-small set.

The function space version Hybrid Monte Carlo algorithm has originally been introduced in [20] and revised in [18]. It is based on geometric integrators for Hamiltonian dynamics. An optimal scaling result in terms of the ESJD, see Section 3.2.2, has been obtained in [19]. More recently, a non-reversible version of this algorithm has been constructed in [147] along with a scaling result that suggests a dimension-independent performance. The non-reversibility introduces additional technical difficulties because many methods, such as L^2_μ-spectral gaps, depend on the reversibility. Studying rigorous convergence bounds for this algorithm seems to be an interesting direction for further investigation.

However, a Wasserstein spectral gap would at least imply bounds on the MSE and a CLT for Lipschitz functionals, the corresponding results can be found in [98] and [115], respectively. Moreover, it seems worthwhile to develop a theory of polynomial Wasserstein convergence to match the theory of polynomial ergodicity, for the latter we refer the reader to [94] and [69]. Subgeometric ergodicity is known to give rise to CLTs and bounds on the sample average, consult [94] and [120]. In contrast, such results are not known for Wasserstein convergence.

3.4.2 Article II

We have shown that if the proposal kernel Q is reversible and has an $L^2_{\mu_0}$-spectral gap with respect to a given reference probability measure μ_0 and the density of the target measure μ is bounded from above and below, then the lazy version of the resulting Metropolis-Hastings algorithm has an L^2_μ-spectral gap. It is therefore natural to ask the question whether the same is true for the Metropolis-Hastings algorithm chain or just for its lazy version.

We constructed appropriate proposals for the elliptic inverse problem introduced in Section 2.2 based on the tensorisation property of L^2_μ-spectral gaps and a reflection argument which we called Reflection Random Walk Metropolis (RRWM) algorithm.

As it is quite natural to assume that the density is bounded above and below on

bounded sets, the global bound can be viewed locally as an additional tail assumption. Even for Markov chains on $\mathbb{R}^n$, it is quite difficult to pose general sufficient conditions on the tail behaviour of the density implying geometric ergodicity of he Metropolis-Hastings Markov chain for a given proposal. One of the most general sufficient conditions are formulated in Equation (3.17) and has been derived in [93]. This condition is non-trivial to verify. It is of interest to derive similar conditions in the infinite dimensional setting and therefore to weaken the assumptions in Articles **I** and **II**.

Even though our bounds for the spectral gap are worse for the RRWM algorithm, compared to the IS algorithm numerical simulations suggest that the former has a better performance for concentrated measures. Posteriors arising in Bayesian inverse problems often become more concentrated if the observational noise is turned to zero. This suggests that it would be interesting to rigorously quantify the complexity of sampling as the posterior in a Bayesian inverse problem goes to zero.

Chapter 4

Posterior Consistency for Bayesian Inverse Problems

The quality of the Bayesian method described in Section 2.1 can be evaluated by considering the posteriors arising from artificially generated data $y = \mathcal{G}(u^\dagger) + \eta$ for a fixed 'truth' $u^\dagger$. Because the aim is the reconstruction of the truth, it is desirable that the posterior contracts to the 'truth' as, for example, the noise goes to zero or the amount of data to infinity. This property of the statistical model is called posterior consistency and is the subject of this chapter. In particular, we present the results of Article **III** constituting one of the first posterior consistency result for a nonlinear inverse problem in infinite dimensions.

In the following, we define posterior consistency for a sequence of inverse problems $(\mu_0^n, \mathcal{G}_n, \mathbb{Q}_0^n)$. In order to quantify the posterior concentration, we denote by B_ϵ^d the ball of radius ϵ with respect to a metric d.

Definition 4.1. *A sequence of Bayesian inverse problems $(\mu_0^n, \mathcal{G}_n, {\mathbb{Q}_0}^n)$ is posterior consistent for $u^\dagger$ with rate $\epsilon_n \downarrow 0$ with respect to a metric d if for*

$$y_n = \mathcal{G}_n(u^\dagger) + \eta_n \text{ with } \eta_n \sim \mathbb{Q}_0^n, \tag{4.1}$$

there exists a constant C and a sequence $l_n \to 1$ such that

$$\mathbb{Q}_0^n \left(\mu^{y_n} \left(B^d_{C\epsilon_n}(u^\dagger) \right) \geq l_n \right) \to 1. \tag{4.2}$$

If we do not consider the rate of posterior consistency, we simply say that $(\mu_0^n, \mathcal{G}_n, {\mathbb{Q}_0}^n)$ is posterior consistent if the above holds with ϵ_n replaced by ϵ, for any fixed $\epsilon > 0$.

In the subsequent discussion, we concentrate on the following two special cases of this definition:

- Posterior consistency in the small noise limit corresponds to

$$\mathbb{Q}_0^n = S_{\frac{1}{\sqrt{n}}\star}\mathbb{Q}_0 \text{ and } \mathcal{G}_n = \mathcal{G}$$

  where $S_a(x) = ax$;

- Posterior consistency in the large data limit corresponds to

$$\mathbb{Q}_0^n = \otimes_{i=1}^n \mathbb{Q}_0 \text{ and } \mathcal{G}_n = \prod_{i=1}^n \mathcal{G}^i = (\mathcal{G}^1, \ldots, \mathcal{G}^n).$$

It is important to note that posterior consistency implies the existence of a consistent estimator. More precisely, for the choice

$$\hat{u}_n = \arg\max_u \mu^{y_n} \left(B^d_{C\epsilon_n}(u) \right),$$

it follows that $d(\hat{u}_n, u^\dagger) \leq 2C\epsilon_n$ if $l_n > 0.5$ because in this case $B^d_{C\epsilon_n}(u^\dagger)$ and $B^d_{C\epsilon_n}(\hat{u}_n)$ cannot be disjoint. More generally, posterior consistency properties evaluate Bayesian methods from the frequentist perspective by studying how the posterior depends on the realisation of the noise. The posterior is always a trade off between the prior and the data. In this context, a lack of posterior consistency can be interpreted as a bias introduced by the prior that cannot be overcome by the data. Posterior inconsistency does not only occur in pathological cases as described in the following. Priors based on Dirichlet processes have been used for the location problem in [43, 41] and [42]. Nevertheless,

in [50] it has been shown for the location problem that these priors lead to inconsistent posteriors. The example provided in article [50] illustrates the importance of posterior consistency and its dependence on the choice of the prior. For this reason, it seems plausible to choose priors that have the best possible posterior consistency properties for a very large class of possible truths. However, this contradicts the philosophical idea behind the Bayesian approach because the prior is only supposed to represent a priori knowledge. A philosophical justification for studying posterior consistency can be seen in the fact that posterior consistency is equivalent to the property that the posteriors arising from different priors merge for a broad class of models as considered in [51]. For an in-depth discussion, we refer the reader to [76]. In practice, priors are often chosen based on their computational performance and some of their parameters are adapted to represent subjective knowledge. This is the case for the choice of base measures and the intensity of a Dirichlet process prior in clustering [91]. It is worth noting that the bulk of the field has moved towards establishing consistency, after the initial works like [50] indicated the care needed in selecting the priors.

This chapter is structured as follows. Section 4.1 contains a review of the development of posterior consistency in the literature of parametric and non-parametric statistics as well as for inverse problems. We use this background to present our contributions summarised in the research Article **III** in Section 4.2. The key idea is to use inverse stability results which can be combined with posterior consistency for regression problems in order to prove posterior consistency for non-linear inverse problems. This idea is presented in more detail in Section 4.2.

4.1 The Development of Posterior Consistency - A Literature Review

In the following, we review the literature on posterior consistency for parametric and non-parametric statistics as well as for inverse problems. We first simplify the problem of posterior consistency for a Bayesian inverse problems to posterior consistency for the problem of identifying a probability distribution from samples. Whereas the main

research question is to address posterior consistency of nonlinear inverse problems in infinite dimensions, we would like show first why it is difficult to adapt the techniques that work for finite dimensions. In this way, we illustrate the need for new methods. The approach taken in Article **III** is presented in Section 4.2 and constitutes an important step in that direction.

Reformulation of Posterior Consistency for Bayesian Inverse Problems

Many statistical models and indeed Bayesian inverse problems can be reduced to the problem of identifying a probability distribution from samples. In the following, we show that this is also the case for inverse problems if the forward operators are injective. In this case, u can be identified with $P_u = \mathsf{T}_{\mathcal{G}_n(u)\star}\mathbb{Q}_0^n$ where $\mathsf{T}_y(x) := x + y$ denotes the translation operator. Moreover, we define by

$$\mathcal{P}_n := \left\{ P_u \middle| u \in X \right\},$$

a set of probability measures on Y. Both prior and posterior on u yield a prior Π_n and a posterior $\Pi_n(\cdot|y_n)$ on $\mathcal{P}_n$ given by

$$\begin{aligned} \Pi_n &= \mu_0^m(du)\mathsf{T}_{\mathcal{G}_n(u)\star}\mathbb{Q}_0^n \\ \Pi_n(\cdot|y_n) &= \mu^y(du)P_u. \end{aligned} \tag{4.3}$$

This statistical model is related to the following problem.

Problem 4.2. *Let $\mathcal{P}$ be a set of probability measures on E and $\theta \mapsto P_\theta$ be a parametrisation for $\theta \in \Theta$. For a fixed $P_{\theta_0} \in \mathcal{P}$, the data is generated from n samples $y_n = (X_1, \ldots, X_n)$ where $X_i \overset{i.i.d.}{\sim} P_{\theta_0}$.*

It is worth mentioning that Problem 4.2 and the reformulation of posterior consistency for inverse problems in Equation (4.3) are not equivalent. However, they are closely related as the discussion in Section 4.1.2 shows. The main question arising for this problem is whether the sequence of posteriors $\Pi_n(\cdot|y_n)$ corresponding to a sequence of priors Π_n and the data y_n concentrates around θ_0.

We would like to point out that Definition 4.1 is posed in analogy to the following definition of posterior consistency for Problem 4.2 presented in [183, 76] and [110].

Definition 4.3. *(Posterior rate of contraction). The posterior distribution* $\Pi_n(\cdot|Y^n)$ *is said to contract at* θ_0 *at rate* $\epsilon_n \downarrow 0$ *if*

$$\Pi_n(\theta : d(\theta, \theta_0) \geq M_n \epsilon_n | Y^n) \to 0 \tag{4.4}$$

in $P_{\theta_0}^n$*-probability for every* $M_n \to \infty$ *as* $n \to \infty$*. If the above holds for every constant* ϵ*, we say that the posterior is consistent.*

In Definition 4.3, d denotes a metric on $\mathcal{P}$ which might be the total variation, the Hellinger distance of the corresponding distribution P_θ or a distance induced by a metric on Θ. The relation between the Definitions 4.1 and 4.3 becomes apparent through the reformulation of the underlying problem in Equation (4.3). The additional sequence M_n makes Definition 4.3 a bit more cumbersome at the benefit of including borderline cases. In particular, the multiplicative constant of the convergence rate can be hidden in M_n.

Bernstein-von-Mises - Posterior Consistency in Finite Dimensions

For finite dimensional $\Theta \subseteq \mathbb{R}^d$ and $\Theta \ni \theta \mapsto P_\theta$ sufficiently regular, the Bernstein-von-Mises theorem implies that under appropriate assumptions on $\theta \mapsto P_\theta$

$$\sup_A \left| \Pi_n \left(\theta \in A \middle| X_1, \ldots, X_n \right) - \mathcal{N}(\hat{\theta}_n, n^{-1} I_\theta^{-1})(A) \right| \to 0 \tag{4.5}$$

for any asymptotically efficient estimator $\hat{\theta}_n$. Due to the well-known properties of $\theta - \hat{\theta}_n$ for asymptotically efficient estimators and $\mathcal{N}(\hat{\theta}_n, n^{-1} I_\theta^{-1})(A)$, Equation (4.5) implies that the posterior contracts at the rate $\epsilon_n = n^{-\frac{1}{2}}$. The Bernstein-von-Mises theorem consolidates Bayesian and frequentists' statistics by providing a comparison of asymptotic confidence sets and credible sets. For more details on confidence sets and their relation, we refer the reader to [39] and [112]. A good overview of the Bernstein-von-Mises theorem can be found in [183] and [91]. It is an interesting historical fact that this theorem can be traced back to Laplace in 1810 as discussed in [125].

It is important to note that the Bernstein-von-Mises theorem does not imply consistency for the MAP estimator. However, consistency of the MAP-estimator is clearly a related question which has been considered for Bayesian inverse problems with Gaussian priors in [45].

However, even though the Bernstein-von-Mises theorem provides very strong results for the finite dimensional case, it is not straightforward to generalise it to the non-parametric or infinite dimensional setting as has been shown in [39]. More precisely, in this article, the author considers a regression problem in the large data limit and proves that there are sets with posterior probability arbitrarily close to one that constitute confidence sets to an arbitrary low level. Nonetheless, for some special cases Bernstein-von-Mises type results hold, for example, it has been proved in [105] that a non-parametric version of the Bernstein-von-Mises theorem applies to the large class of Lévy process priors for survival models. Bernstein-von-Mises type theorems seem to hold in more generality in the semi-parametric setting, that is for a finite dimensional parameter with an infinite dimensional nuisance parameter. One of the first of these results has been presented in [104] for the hazard model. For a recent account of the theory of semi-parametric inference and an extension to more irregular models, we refer the reader to [108]. In general, the problem of posterior consistency for non-linear inverse problems is for this reason treated differently.

4.1.1 Non-Parametric Statistics

We review posterior consistency results for a non-parametric version of Problem 4.2 following [91] and [76]. The difference to the parametric approach is that the parameter space Θ is infinite dimensional or the dimension is increasing with the number of samples. Without any restrictive assumptions, Doob has shown in [60] that if θ is a measurable function of $\sigma(X_1, \dots)$, then the posterior is consistent for $\prod$-.a.e. θ for Problem 4.2 with respect to the Prokohorv metric metrising weak convergence. However, Doob's result does not establish a rate, nor does it guarantee consistency for particular choices of θ_0.

The first example of posterior inconsistency has been provided in [71] for the Problem 4.2 for probability measures on $\mathbb{N}$. In a similar setting, the same author has proved in

the research article [72] that this behaviour is topologically generic.

Article [171] contains a general theory for posterior consistency based on the well-developed theory of tests in order to bound $\Pi_n(U_{\theta_0}^c|Y_n)$ for any neighbourhood $U_{\theta_0}^c$ of the truth P_{θ_0}. The basic idea is to use the existence of tests Φ_n satisfying

$$P_{\theta_0}(\Phi_n) \leq B\exp(-bn) \quad \sup_{\theta \in U_{\theta_0}^c} P_\theta(1-\Phi_n).$$

The tests Φ_n can be used to bound the following expression with respect to the realisation of the noise

$$\Pi_n\left(U_{\theta_0}^c|Y_n\right) \leq \Phi_n + \frac{(1-\Phi_n)\int_{U_{\theta_0}^c}\prod_{i=1}^n \frac{p(X_i)}{p_0(X_i)}d\Pi_n(p)}{\int_{U_{\theta_0}^c}\prod_{i=1}^n \frac{p(X_i)}{p_0(X_i)}d\Pi_n(p)}$$

under appropriate additional assumptions.

The test-driven approach has been developed into a quantitative approach in [74] by reducing the existence of appropriate tests to metric entropy numbers with respect to, for example, the Hellinger metric. The existence of such tests is then guaranteed by results presented in [124] and [23]. This method gives rise to conclusions of the following form.

Theorem 4.4. *(Theorem 2.1 in [74]) Suppose there exist $\epsilon_n \to 0$ with $n\epsilon_n^2 \to \infty$, a constant $C > 0$ and sets $\mathscr{P}_n \subseteq \mathcal{P}$ such that*

$$\log D\left(\epsilon_n, \mathscr{P}_n, d\right) \leq n\epsilon_n^2 \tag{4.6}$$

$$\Pi_n\left(\mathscr{P} \setminus \mathscr{P}_n\right) \leq \exp\left(-n\epsilon_n^2\left(C+4\right)\right) \tag{4.7}$$

$$\Pi_n\left(P : -P_0\left(\log\frac{p}{p_0}\right) \leq \epsilon_n^2, P_0\left(\log\frac{p}{p_0}\right)^2 \leq \epsilon_n^2\right) \geq \exp\left(-n\epsilon_n^2 C\right) \tag{4.8}$$

where d is a distance on the set of measures $\mathcal{P}$ and $D\left(\epsilon_n, \mathscr{P}_n, d\right)$ is the ϵ-packing number. Then for sufficiently large M

$$\Pi_n(P : d(P, P_0) \geq M\epsilon_n | X_1, \ldots, X_n) \to 0.$$

The assumptions stated in Theorem 4.4 have explicitly been verified for Gaussian priors in [182]. In the same article, the authors show that also other problems in non-

parametric statistics, such as regression and density estimation, can be transformed into the Model 4.2. For regression problems of the form

$$Y_i = f(X_i) + \epsilon_i$$

with $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ and $X_i \sim Q$ for a known Q, this transformation works as follows

$$f \mapsto P_f = \int \mathsf{T}_{f(x)\star} Q(dx)$$

where $\mathsf{T}_h(x) = x + h$ is the translation operator. It is then left to relate distances on the measures P_f to distances on the function f. For a more detailed discussion about posterior consistency, we refer the reader to [182], [91] and [76]. More details about posterior consistency for non-parametric regression are presented in Section 4.2.

4.1.2 Bayesian Inverse Problems

In the following, we review posterior consistency results for Bayesian inverse problems available in the literature stressing that most articles address linear inverse problems under restrictive assumptions on the prior. In contrast, we present our contributions addressing posterior consistency for nonlinear inverse problems in Section 4.2.

Whereas we focus on posterior consistency of the infinite dimensional problem, posterior consistency for discretisations of linear models has been considered in [141] and in references therein.

At first sight, the literature suggests that the notions of posterior consistency for Bayesian inverse problems and for non-parametric statistics, do not have much in common. The methods from non-parametric statistics were deemed as 'not suitable to deal with ill-posed inverse problems'[1]. This is not quite true as has been demonstrated in [153] by a modification of the proof of Theorem 4.4. The author has been able to obtain sharp rates in the linear case with this method and he has also established results for some classes of non-Gaussian priors.

[1] p. 22 in B. Knapik. Bayesian Asymptotics - Inverse Problems and Irregular Models, PhD thesis, Vrije Universiteit Amsterdam, 2013.

Linear forward operators have also been considered in articles [112, 113, 1] and [2] under the assumption that both the prior and the noise are Gaussian. Whereas articles [112] and [113] treat the case when prior and noise covariances are jointly diagonalisable, [1] and [2] use PDE methods that weaken this assumption. However, articles [112, 113, 1, 2] make use of the explicit Gaussian structure of the posterior. For this problem, Gaussian priors are conjugate, that is both the prior and the posterior are of the same class, in this case they are Gaussian measures. In contrast, [153] is able to treat a much larger class of non-conjugate priors.

As far as the author is aware, posterior consistency results for nonlinear inverse problems have only been considered in finite dimensions in articles [25] and [24]. Both articles consider generalised linear models with non-Gaussian noise. However, it seems that the appropriate assumptions are not satisfied for many non-linear problems such as the (EIP). Posterior consistency with respect to the Ky-Fan metric has been proved in [24]. This result also holds if the dimension increases sufficiently slowly. For a fixed finite dimension, a version of the Bernstein-von-Mises theorem has been derived in [25].

It becomes apparent from the above literature review that so far no posterior consistency results have been established for nonlinear infinite dimensional Bayesian inverse problems with non-Gaussian priors. In the following section, we present results from Article **III** filling this gap.

4.2 Contributions of Article III

In Article **III**, we develop a method to prove posterior consistency for non-linear inverse problems with additive Gaussian noise under weak assumptions on the prior and apply it to the elliptic inverse problem (EIP) as considered in Section 2.2. Our method can be summarised by reducing the problem of posterior consistency of the non-linear inverse problem to posterior consistency of a Bayesian regression problem using inverse stability results. These inverse stability results are readily available for many problems in the literature as they often form the basis for convergence results for regularisation problems (see Theorem 10.4 in [67]).

Besides establishing this general method, the contribution of this paper naturally falls into four parts:

1. Assuming the behaviour of small ball probabilities and the tail behaviour of the prior, we establish new results of posterior consistency with rate for the Bayesian regression problem with Gaussian noise both for Hilbert spaces and point-wise observations (consider Section 3 in Article **III**).

2. We discuss an interesting example of posterior inconsistency for a particular regression problem.

3. We apply the method in detail to elliptic inverse problems as considered in Section 2.2 verifying the assumptions for the posterior consistency results for the Bayesian regression problem (see Section 4 in Article **III**).

4. We derive small ball probability asymptotics for a prior based on a series expansion with independent and uniformly distributed coefficients in order to obtain a rate of posterior consistency.

In the remaining part of this chapter, we describe our method and the novel posterior consistency results for the Bayesian regression problem in more detail. Suppose we are given a sequence of inverse problems with prior μ_0, forward operator $\mathcal{G}_n$ and noise $\eta \sim \mathcal{N}(0, \Gamma_n)$. The Bayesian framework for this inverse problem (IP), as described in Section 2.1, can be summarised as follows

Prior	μ_0 on u
Data	$y = \mathcal{G}_n(u) + \eta_n,\ \eta_n \sim \mathcal{N}(0, \Gamma_n)$
Posterior	$\frac{d\mu^n}{d\mu_0}(u) \propto \exp\left(-\frac{1}{2}\left\|\mathcal{G}_n(u)\right\|_{\Gamma_n}^2 + \langle y, \mathcal{G}_n(u)\rangle_{\Gamma_n}\right).$

(IP)

We assume that the forward operator $\mathcal{G}_n$ can be written as a composition of the model operator G and an observation operator $\mathcal{O}_n$ such that

$$\mathcal{G}_n = \mathcal{O}_n(G(u)).$$

The model operator can be thought of as the solution operator to the underlying continuum model and the observation operator can be interpreted as evaluations of the solution. We consider a related Bayesian regression problem with the following push-forward prior

$$\boxed{\begin{array}{cc} \text{Prior} & \tilde{\mu}_0 = G_\star \mu_0 \text{ on } v \\ \text{Data} & y = \mathcal{O}_n(v) + \eta_n,\ \eta_n \sim \mathcal{N}(0, \Gamma_n) \\ \text{Posterior} & \frac{d\tilde{\mu}^{y_n}}{d\tilde{\mu}_0}(v) \propto \exp\left(-\frac{1}{2}\left\|\mathcal{O}(v)\right\|_{\Gamma_n}^2 + \langle y, \mathcal{O}(v)\rangle_{\Gamma_n}\right). \end{array}} \quad \text{(BRP)}$$

The (IP) is related to the (BRP) by a change of variable formula (consider Theorem Appendix B.1. in Article **III**) implying that

$$\mu^{y_n}\left(B_\epsilon^d(u^\dagger)\right) = \tilde{\mu}^{y_n}\left(G\left(B_\epsilon^d(u^\dagger)\right)\right).$$

Recall that $B_\epsilon^d(x)$ denotes a ball of radius ϵ with respect to the metric d. Inverse stability results yield a statement of the form $G(B_\epsilon^d(u^\dagger)) \supseteq B_{b(\epsilon)}^{d'}\left(G(v^\dagger)\right)$ which then gives rise to the following theorem which is the main result of Section 2 in Article **III**.

Theorem 4.5. *Suppose $\mathcal{G}_n = \mathcal{O}_n \circ G$ with $G : (X, d_X) \to (Y, d_Y)$ and $\mathcal{O}_n : (Y, d_Y) \to (Z, d_Z)$. Moreover, we assume that*

- *there exists an inverse stability result of the form*

$$d_X(a_1, a_2) \leq b\left(d_Y\left(G(u_1), G(u_2)\right)\right)$$

  where $b : \mathbb{R}^+ \to \mathbb{R}^+$is increasing and, $b(0) = 0$;

- *there is a set A such that the sequence of Bayesian inverse problems $(G_\star\mu_0, \mathcal{O}_n, \mathcal{L}(\xi_n))$ is posterior consistent with respect to d_Y for all $v^\dagger \in A$ with rate ϵ_n.*

Then $(\mu_0, \mathcal{G}_n, \mathcal{L}(\xi_n))$ is posterior consistent with respect to d_X for all $u^\dagger \in G^{-1}(A)$ with rate $b(\epsilon_n)$.

In the following, we consider $\mathcal{O}_n = \text{Id}$ or $\mathcal{O}_n = (e_{x_i})_{i=1}^n$, where e_{x_i} denotes the evaluation at x_i. For the latter, the (BRP) becomes a standard regression problem.

However, as this problem is equipped with the push-forward prior, the prior does not belong to a particular class. In the case of $\mathcal{O}_n = (e_{x_i})_{i=1}^n$, this problem has mostly been dealt with in the case of Gaussian process priors and/or random covariates as described in [91] and [76]. The only exceptions are [34] and [33] containing posterior consistency results under conditions similar to ours, however, without a rate. In contrast, we obtain a rate of posterior consistency under assumptions posed in [34] and are able to conclude posterior consistency under much weaker assumptions. The corresponding results are contained in Section 3.2 in Article **III**.

For $\mathcal{O}_n = \mathrm{Id}$, the only posterior consistency results that apply to the (BRP) are those for linear Bayesian inverse problems. However, these results only apply to particular classes of priors as described in Section 4.1. We derive a result for general priors satisfying assumptions on the small ball probabilities as well as the tail behaviour. We compare our results in the conjugate Gaussian setting with jointly diagonalisable covariances to the optimal rates obtained in [112]. Even though our result applies in far greater generality, our rates are suboptimal but quite close to the optimal rates in the special case where [112] is applicable (for details see Section 3.1 in Article **III**, in particular, Figure 1).

In order to give a flavour of the results presented in Article **III**, we state a special case that applies to the (EIP) as considered in Section (2.2.2).

Theorem 4.6. *Consider the sequence of posteriors μ^{y_n} arising from the (EIP) as set up in Equation (2.10) for $\Gamma_n = \frac{1}{\sqrt{n}}(-\Delta_{Dirichlet})^{-r}$. Suppose that the parametrisation of the diffusion coefficient in Equation (2.7) satisfies*

$$\left(\sum_{i=1}^{\infty} \|\psi_i\|_{C^\beta}\right)^{\frac{1}{\nu}} < \infty$$

for $\beta > r + \frac{d}{2} - 1$. Then the (EIP) is posterior consistent in the small noise limit with respect to the $C^{\tilde{\beta}}$-norm for any $\tilde{\beta} < \beta$. Additionally, the (EIP) is posterior consistent in the small noise limit with respect to the L^∞-norm with rate $n^{-\kappa}$ for any κ such that

$$\kappa < \left(\frac{\alpha}{\alpha + 2 + \frac{d}{2} - r} \wedge 1\right)\left(\frac{1}{2 + \frac{1}{\nu} - 1} \wedge \frac{\alpha}{2\left(\alpha + 1 + \frac{d}{2r}\right)}\right).$$

The above theorem is crucial as it shows that the bias due to the prior choice vanishes if the noise in the data is scaled to zero. Moreover, a lower bound on the rate has been provided.

4.3 Conclusion and Further Directions

In Article **III**, we have provided one of the first posterior consistency results for non-linear inverse problems in infinite dimensions. Besides providing a general method to approach this problem, the article contains new results for Bayesian regression problems as well as small ball probabilities. We have applied our method to the (EIP) with appropriate priors. These do not include log-Gaussian priors on the diffusion coefficient as the resulting push forward prior on the pressure does not have any exponential moments. It is therefore of interest that posterior consistency for the (BRP) can be proved under weaker assumptions on the tails.

We would like to stress that it is straightforward to apply the method to any inverse problem for which the appropriate inverse stability results are available. One example is the Calderon problem (electrical impedance tomography) for which the inverse stability results are given in [3]. As the problem is severely ill-posed, the stability result is very weak and therefore our method would also yield a slow rate of posterior consistency on the log-scale.

In Article **III**, we have proved posterior consistency in the large data limit assuming the data is i.i.d. distributed. A possible extension is to consider dependent data which has been considered in [75] for the problem of identifying a distribution from samples (c.f. Model 4.2). It seems to be possible to use the ideas from [33] in order to prove such a result.

As described just before Section 4.1.1, the consistency of the MAP estimator is a related question that neither implies, nor is implied by posterior consistency. However, there might be additional assumptions under which consistency of the MAP estimator implies posterior consistency. If this is the case, it is possible to use results from [45] in order to establish posterior consistency.

Chapter 5

An Under-Determined Elliptic Multiscale Inverse Problem

Many physical phenomena happening around us are in fact occurring on many different spatial and temporal scales. Therefore all of these different scales have to be taken into account in order to model the phenomena accurately. A very descriptive example is the global climate as it is influenced by effects that happen on small spatial scales, as for example, the forming of clouds, and short temporal scales, like volcano eruptions. If all effects are taken into account explicitly, the computer model would have to resolve up to finest temporal and spatial scales leading to a prohibitively expensive model. Therefore an effective model has to be constructed that only involves large temporal and spatial scales. However, in case of the climate model, it is not possible to derive such an effective model rigorously. In contrast, more idealised multiscale phenomena can be treated in a rigorous way. One example is the phenomenon of heat conduction for a composite material consisting of alternating layers of different materials. For this problem, it is possible to describe the composite material as being effectively homogenous. In doing so, we obtain a bound on the difference between the two descriptions depending on the layer-size. Consequently, the mathematical theory behind this fact is called homogenisation.

Solving the effective equation has usually a much smaller computational cost than solving the multiscale equation while still being accurate. For this reason, it is of interest to study the consequences of using the effective model for data obtained from the

physical model. This approach has, for example, been considered in [144]. However, in the following, we study a different aspect. We would like to investigate the following questions: How much information about the multiscale structure is contained in the measurements and what effect does homogenisation have on the inverse problem? In the simple picture of homogenisation of a two layered material, one can already expect that there are many two layered materials corresponding to one homogeneous material. This makes it apparent that the reconstruction of the complete two-scale structure is in fact a heavily underdetermined problem.

In this last chapter, we would like present our work in progress regarding the behaviour of the posterior in a multiscale inverse problem. We have summarised our initial results and observations in draft Article **IV**. More precisely, in Article **IV**, we study the reconstruction of the two-scale diffusion coefficient from measurements of the pressure. This is an extension of the problem studied in Section 2.2. The relation is again modelled as a linear second order elliptic PDE of the form

$$-\nabla \cdot \left(a\left(x, \frac{x}{\epsilon}\right) \nabla p^{\epsilon} \right) = g \text{ for } x \in D \tag{5.1}$$

$$p^{\epsilon} = 0 \text{ for } x \in \partial D \tag{5.2}$$

on a domain $D \subseteq \mathbb{R}^d$. The basic assumptions are that $a : D \times \mathbb{R}^d \to \mathbb{R}^d$ is periodic in its second argument and twice differentiable in both arguments and g is an element of $C^2(D)$. Moreover, we assume that a is uniformly elliptic, that is

$$\langle a(x,y)\xi, \xi \rangle \geq \alpha \left\| \xi \right\|^2 .$$

We would like to point out that Equation (5.1) also describes the heat conduction phenomenon described above.

A scale separation occurs as x and $\frac{x}{\epsilon}$ vary on the scale of order 1 and ϵ, respectively. Because of the periodicity of a, the theory of periodic homogenisation applies which we present in a nutshell in Section 5.1 following [149]. This implies that there is an effective

equation of the form

$$\begin{aligned} -\nabla \cdot (\bar{a}(x)\nabla \bar{p}) &= g \text{ for } x \in D \\ \bar{p} &= 0 \in \partial D. \end{aligned} \tag{5.3}$$

In this setting, it is possible to show that $\|p^\epsilon - \bar{p}\|_\infty$ is of order ϵ [17, p.19].

In order to present our ideas, we introduce the inverse problem of interest. We consider a two-scale diffusion coefficient of the form

$$a(b,c)(x,y) : = a_0(x) + b(x) + c(y) \tag{5.4}$$

where $a_0, b \in C^2(D, \mathbb{R}^d)$ and $c \in C^2(\mathbb{R}^d, \mathbb{R}^d)$. We call b and c the coarse and fine diffusion coefficient, respectively and denote the corresponding homogenised diffusion coefficient by $\bar{a}(b,c)$.

We assume that a_0, ϵ and g are known and consider the inverse problem for b and c given measurements of the pressure p. We summarise the homogenisation theory and set up the inverse problem in more detail before presenting the results contained in Article **IV**. In this article, we study the inverse problem on the one dimensional domain $D = [0,1]$. Our initial findings are the following:

1. We consider the set $\bar{a}^{-1}\left(\bar{a}(b^\dagger, c^\dagger)\right)$ of different coarse and fine diffusion coefficients b and c that homogenise to the same effective diffusion coefficient corresponding to $\bar{a}(b^\dagger, c^\dagger)$. This set has a manifold structure and can be represented as a graph over c. In order study the form of the level set, we investigate the dependence of $\bar{a}$ on b and c.

2. We study the Bayesian approach to the inverse problem by generating artificial data corresponding to $b^\dagger$and $c^\dagger$. Using simulations, we demonstrate that for small ϵ and small observational noise the posterior concentrates around the level set $\bar{a}^{-1}\left(\bar{a}(b^\dagger, c^\dagger)\right)$.

3. We prove that the posterior based on the homogenised model is close to that of the

multiscale model. Moreover, we use disintegration to illustrate why the posterior concentrates around the level set.

Subsequently, we set up the homogenised problem first before introducing the corresponding inverse problem in Section 5.1. In Section 5.2, we present a brief literature review. The considerations in Article **IV** are interesting in their own right but are in fact meant as a starting point of an investigation of the following two aspects

1. The study of heavily underdetermined problems with posteriors concentrating around manifolds and the development of efficient MCMC methods for them.

2. The consideration of inference for multiscale models, in particular, considering the size of the fast scale as unknown, too.

Because it is a common phenomenon that the posterior concentrates in different directions at different rates, appropriate MCMC algorithms are needed.

5.1 Periodic Homogenisation and the Multiscale Inverse Problem

We present the main ideas of the periodic homogenisation theory of elliptic PDEs before setting up the corresponding inverse problem and the Bayesian approach to it more explicitly.

5.1.1 Periodic Homogenisation of Elliptic PDEs

We follow [149] to present the main ideas of homogenisation and refer to [17] and [35] for further reading. The effective Equation (5.3) can be derived by treating $\frac{x}{\epsilon}$ as an independent variable y and performing an asymptotic expansion of the form

$$p^{\epsilon}(x,y) = p_0(x,y) + \epsilon p_1(x,y) + \epsilon^2(x,y)\dots$$

for which we refer the reader to [149]. In particular, this expansion implies that $p_0(x,y) = p_0(x) =: \bar{p}(x)$. The homogenised diffusion coefficient is given by

$$\bar{A}(x) = \int_{\mathbb{T}^d} A(x,y) + A(x,y)\nabla_y \chi(x,y)^T dy,$$

where $\chi : D \times \mathbb{R}^d \to \mathbb{R}^d$ is the corrector corresponding to the solution of the cell problem

$$-\nabla_y \cdot \left(\nabla_y \chi A^T\right) = \nabla_y \cdot A^T \text{ with periodic boundary conditions.}$$

The formal derivation can be justified by appropriate convergence results of the form $p^\epsilon \to p$. In the case of Equation (5.1), it is possible to derive an explicit error bound such that

$$\|p^\epsilon - \bar{p}\| \leq K(A)\epsilon. \tag{5.5}$$

This estimate can be obtained using the maximum principle as described in [17, p. 19]. The appropriate maximum principle can be found in [178]. Bounds of this form in stronger topologies have been obtained in [138]. Moreover, we have obtained bounds with an explicit constant in the appendix of Article **IV** for $D = (0,1)$.

5.1.2 The Multiscale Elliptic Inverse Problem

We introduce the inverse problem of reconstructing the multiscale diffusion coefficient from measurements of the pressure. We assume that the magnitude of the fine scale ϵ is known, therefore this can be viewed as a slight generalisation of the problem studied in Section 2.2. We assume that the data is given by

$$y = \mathcal{O}(p) + \xi = \{p(i\Delta y)\}_{i=0}^{\left\lfloor \frac{1}{\Delta y} \right\rfloor} + \xi,$$

where $\mathcal{O}$ denotes the observation operator and ξ is the additive observational noise. By introducing the solution operators G_ϵ and $\bar{G}$ corresponding to the Equations (5.1) and (5.3), respectively, we may construct the appropriate forward operators as follows

$$y = \mathcal{G}_\epsilon(a(b,c)) + \eta := \mathcal{O}(G_\epsilon(a(b,c))) + \eta = \mathcal{O}(p^\epsilon) + \eta. \tag{5.6}$$

We also consider the inverse problem of reconstructing the multiscale diffusion coefficient using the solution operator $\bar{G}$ to the homogenised problem given by

$$\bar{y} = \bar{\mathcal{G}}(b, c) + \eta := \mathcal{O}(\bar{G}(\bar{a}(b, c))) + \eta = \mathcal{O}(\bar{p}) + \eta. \tag{5.7}$$

Similar to Section 2.2, we take the Bayesian approach by specifying the observational noise η to be a finite dimensional Gaussian of the form $\mathcal{N}(0, \Gamma)$ and placing a prior on b and c.

Already the inverse problem in $d = 1$ suffers from the lack of identifiability as different parameters b's and c's give rise to the same homogenised diffusion coefficient $\bar{a}$ and even to the same a. Studying the problem in dimensions $d > 1$ yields additional difficulties which we describe briefly in the following. Given an equation of the form

$$-\nabla \cdot (a(x)\nabla p) = g \text{ for } x \in D, \tag{5.8}$$

we can add any $n : D \times \mathbb{R}^d \to \mathbb{R}^d$ such that $n(x)\nabla p = 0$ for all $x \in D$. Thus,

$$-\nabla \cdot (a(x)\nabla p + n(x)) = g \text{ for } x \in D.$$

In article [47], this problem is circumvented by considering among all matrix fields a satisfying Equation (5.8) the one with the smallest L^2-norm as the solution to the inverse problem.

The contributions of Article **IV** concern the two-scale inverse problem on $D = [0, 1]$. Under appropriate assumptions on the prior, the posterior takes the form

$$\frac{d\mu^y}{d\mu_0}(a^\epsilon) \propto \exp\left(-\frac{1}{2}\left\|y - \mathcal{G}_\epsilon(a)\right\|_\Gamma^2\right). \tag{5.9}$$

In Article **IV**, the influence of homogenisation to the behaviour of the posterior in identical twin experiments for a known small ϵ is considered.

5.2 Literature Review

Articles [143] and [144] consider the same multiscale inverse problem from a different angle. The authors fix a true multiscale diffusion coefficient $a^\dagger$ and are interested in reconstructing the corresponding homogenised diffusion coefficient $\bar{a}^\dagger$ based on the data

$$y = \mathcal{G}(a^\dagger) + \eta.$$

A natural way would be to feed this data into the inverse problem

$$y = \bar{G}(h) + \eta.$$

However, they demonstrate that a better result is obtained when the homogenised inverse problem is considered with a different noise distribution which can be derived on the basis of homogenisation theory. This idea has been extended to a semi-parametric drift estimation of a multiscale diffusion coefficient, for which we refer the reader to [116].

The Bayesian inverse method has also directly been applied to discretisations such as multi-scale finite element methods in [61] and [65]. They speed up MCMC simulations by rejecting proposals first on the basis of the coarse scale dynamics before any fine scale simulations are performed. It is worth mentioning the work of [148] representing the inverse problem using a multiscale finite elements method. In this article, a prior and posterior are formulated as joint probability distribution. Moreover, the assumption is posed that the diffusion coefficient is conditionally independent given the coarse stiffness matrix. On this basis an appropriate MCMC algorithm is developed.

So far, we are concerned with a manifold of different fine and coarse diffusion coefficients giving rise to same homogenised diffusion coefficient. One of our future goals is to develop efficient MCMC algorithms for this particular case and similar problems. We point the reader to [180] for an optimal scaling study of the performance of the RWM as the target measure concentrates around simple manifolds.

5.3 Contributions of Article IV

We summarise some of our initial findings contained in Article **IV** using a toy model. In the article, we expand both the coarse diffusion coefficient b and the fine diffusion coefficient c in terms of a Fourier series. We consider the toy model with $D = (0, 1)$ and we assume that the diffusion coefficient is parametrised by

$$a_0(x) = 0, \quad b(x) = b_0, \quad c(y) = c_1 \sin(2\pi y)$$

with a corresponding homogenised diffusion coefficient given by

$$\bar{a}(b_0, c_1) = \left(\int (b_0 + c_1 \sin(2\pi y))^{-1} \right)^{-1}$$

which in this case is just a constant function. Our contributions for the general case are as follows:

1. Using the implicit function theorem, we show that $\bar{a}^{-1}(b^\dagger, c^\dagger) \cap \{(b, c) \big| a(b, c) > 0\}$ form a manifold with graph structure in a function space. For the toy model, the coarse diffusion coefficient b and the homogenised diffusion coefficient $\bar{a}$ are constants leading to finite dimensional level sets. These do also have a graph structure as demonstrated in Figure 5.1 illustrating the contours of $\bar{a}$ in the b_0 and c_1 space.

2. The manifold structure in higher dimensions turned out be not as rich as expected. We have investigated the impact of a change in the fine diffusion coefficient c, given as a Fourier expansion of the homogenised diffusion coefficient $\bar{a}$. Changes in the Fourier modes of c almost only effect the constant mode of the homogenised diffusion coefficient as shown in Figure 5.2. Even in the toy model depicted in Figure 5.1, it is apparent that c_1 has less impact on the homogenised diffusion coefficient than b_0. In general, this can be made quantitative by taking the derivative of the Fourier coefficients of $\bar{a}$ with respect to those of b and c. The corresponding simulation is explained in more detail in Section 3 of Article **IV**.

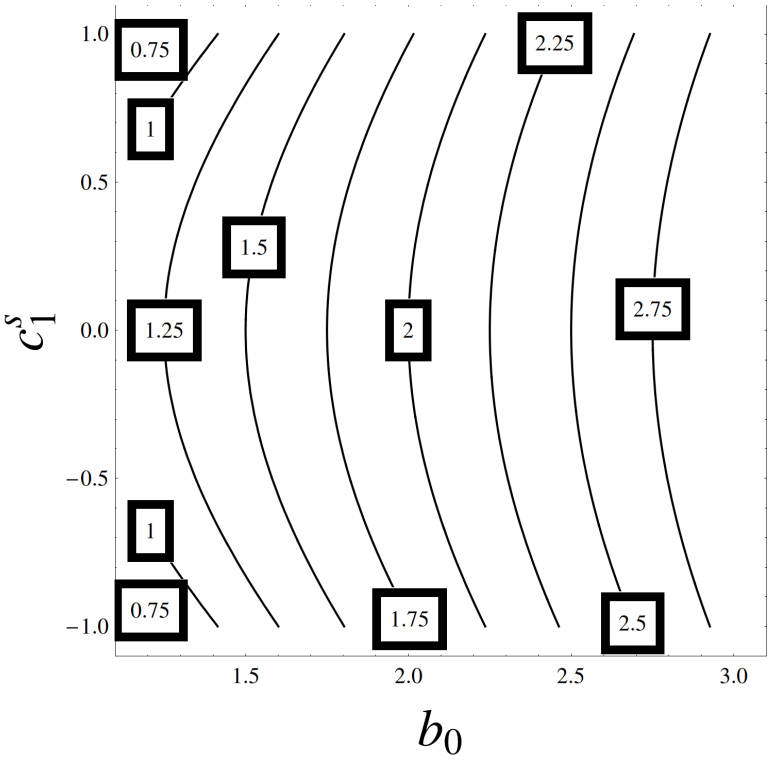


Figure 5.1: Level sets of $\bar{a}$

3. We study the Bayesian inverse problem using artificial data corresponding to $y = \mathcal{G}_\epsilon(b^\dagger, c^\dagger) + \eta$. In this case, the posterior consistency property, see Chapter 4, does not hold. The reason for this is that for small ϵ it is difficult to distinguish between different coarse and fine diffusion coefficients b and c giving rise to the same homogenised diffusion coefficient $\bar{a}$. Using MCMC samples, we confirm that instead the posterior concentrates around the level set $\bar{a}^{-1}(b^\dagger, c^\dagger)$. This is illustrated for an extended toy model in Figure 5.3

4. We consider an approximate posterior based on the homogenised forward model. We derive an appropriate bound on K in Equation (5.5) such that we can show that both posteriors are ϵ close in the Hellinger and the total variation distance. Moreover, we use disintegration to show that the homogenised inverse problem determines the distribution of the posterior of different level sets and that the prior determines the distribution on these level sets.

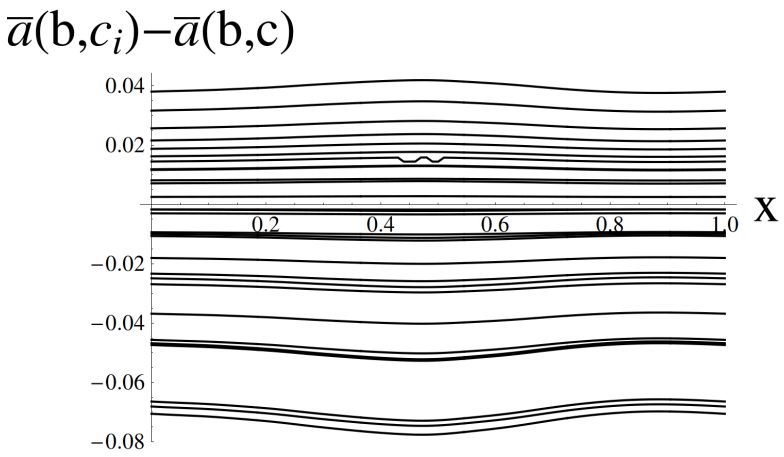


Figure 5.2: Influence of the fine diffusion coefficient c on the homogenised diffusion coefficient $\bar{a}$

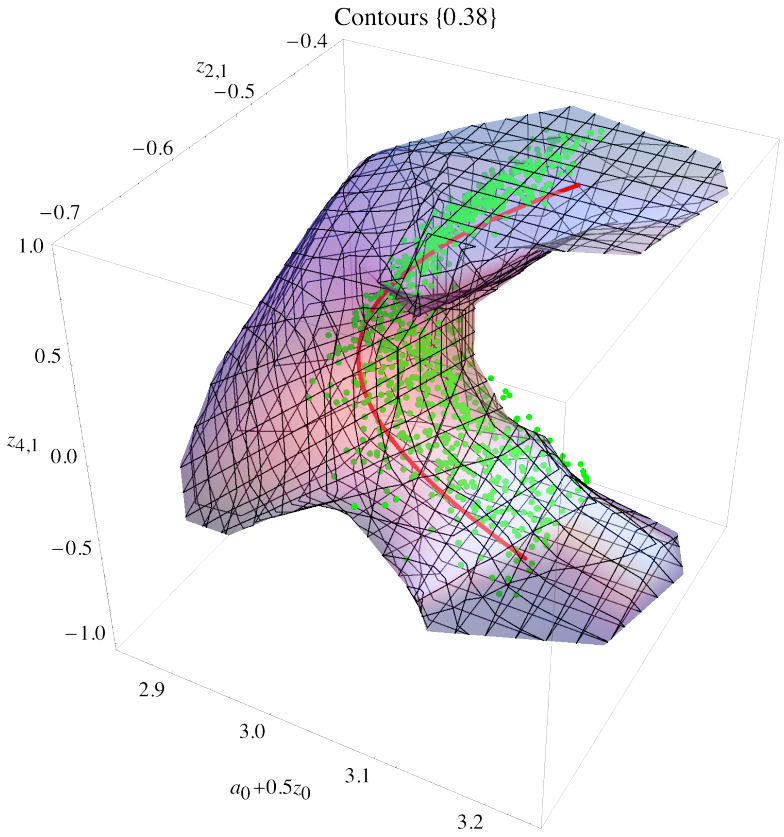


Figure 5.3: MCMC points (green), manifold (red), level sets of the L^2-distance to $p^\dagger$

5.4 Future Goals

One of our major goals is to develop MCMC algorithms that make efficient use of the a priori knowledge of the geometry of the forward problem. More precisely, the level sets of the homogenised diffusion coefficients are a priori known and the data only determines on which level sets the posterior concentrates as the noise goes to zero. The simplest way to make use of this knowledge is to construct a proposal with large steps in the direction tangent to the level set and smaller steps normal to the manifold. In this way, the derivative of the mapping to the homogenised diffusion coefficient is used in order to explore the state space faster in directions that have less impact on the effective problem. Another objective is to use and maybe adapt the approach based on Riemannian manifolds developed in [79]. Moreover, we would like to investigate how particle methods, as described in [48] and [49], can be used in order to explore concentrated posteriors; the benefit of this approach being that several areas can be explored at once.

Part B

The Original Reserach Articles

Research Article I

Spectral Gaps for a Metropolis-Hastings Algorithm in Infinite Dimensions.

Martin Hairer, Andrew M. Stuart and Sebastian J. Vollmer, 2011. *Accepted with minor revisions by the Annals of Applied Probability, 39 pages.*

SPECTRAL GAPS FOR A METROPOLIS-HASTINGS ALGORITHM IN INFINITE DIMENSIONS

By Martin Hairer *, Andrew M. Stuart †
and Sebastian J. Vollmer‡

Mathematical Institute, University of Warwick, Coventry, CV4 7AL, UK.

Abstract. We study the problem of sampling high and infinite dimensional target measures arising in applications such as conditioned diffusions and inverse problems. We focus on those that arise from approximating measures on Hilbert spaces defined via a density with respect to a Gaussian reference measure. We consider the Metropolis-Hastings algorithm that adds an accept-reject mechanism to a Markov chain proposal in order to make the chain reversible with respect to the the target measure. We focus on cases where the proposal is either a Gaussian random walk (RWM) with covariance equal to that of the reference measure or an Ornstein-Uhlenbeck proposal (pCN) for which the reference measure is invariant.

Previous results in terms of scaling and diffusion limits suggested that the pCN has a convergence rate that is independent of the dimension while the RWM method has undesirable dimension-dependent behaviour. We confirm this claim by exhibiting a dimension-independent Wasserstein spectral gap for pCN algorithm for a large class of target measures. In our setting this Wasserstein spectral gap implies an L^2-spectral gap. We use both spectral gaps to show that the ergodic average satisfies a strong law of large numbers, the central limit theorem and non-asymptotic bounds on the mean square error, all dimension independent. In contrast we show that the spectral gap of the RWM algorithm applied to the reference measures degenerates as the dimension tends to infinity.

1. Introduction. The aim of this article is to study the complexity of certain sampling algorithms in high dimensions. Creating samples from a high dimensional probability distribution is an essential tool in Bayesian inverse problems Stuart (2010), Bayesian statistics Lee (2004), Bayesian nonparametrics Hjort et al. (2010) and conditioned diffusions Hairer, Stuart and Voss (2007). For example, in inverse problems, some input data such as initial conditions or parameters for a forward mathematical model have to be determined from observations of noisy output. In the Bayesian approach, assuming a prior on the unknown input, and conditioning on the data, results in the posterior distribution, a natural target for sampling algorithms. In fact these sampling algorithms are also used in optimisation in form of simulated annealing Geyer and Thompson (1995); Pillai, Stuart and Thiéry (2011).

The most widely used method for general target measures are Markov chain Monte Carlo (MCMC) algorithms which use a Markov chain that in stationarity yields dependent samples from the target. Moreover, under weak conditions, a law of large numbers holds for the empirical average of a function f (observable) applied to the steps of the Markov chain. We quantify the computational cost

*Supported by EPSRC, the Royal Society, and the Leverhulme Trust.
†Supported by EPSRC and ERC.
‡Supported by ERC.
AMS 2000 subject classifications: 65C40, 60B10, 60J05, 60J22
Keywords and phrases: Wasserstein spectral gaps, L^2-spectral gaps, Markov Chain Monte Carlo in infinite dimensions, Weak Harris theorem, Random-walk metropolis

of such an algorithm as

$$\text{number of necessary steps} \;\times\; \text{cost of a step.}$$

While for most algorithms the cost of one step grows with the dimension, a major result of this article is to exhibit an algorithm which, when applied to measures defined via a finite-dimensional approximation of a measure defined by a density with respect to a Gaussian random field, requires a number of steps independent of the dimension in order to achieve a given level of accuracy.

For ease of presentation we work on a separable Hilbert space $(\mathcal{H}, \langle\cdot,\cdot\rangle)$ equipped with a mean-zero Gaussian reference measure γ with covariance operator $\mathcal{C}$. Let $\{\varphi_n\}_{n\in\mathbb{N}}$ be an orthonormal basis of eigenvectors of $\mathcal{C}$ corresponding to the eigenvalues $\{\lambda_n^2\}_{n\in\mathbb{N}}$. Thus γ can be written as its Karhunen-Loeve Expansion Adler (1990)

$$\gamma = \mathcal{L}(\sum_{i=1}^{\infty} \lambda_i e_i \xi_i), \qquad \text{where } \xi_i \overset{\text{i.i.d}}{\sim} \mathcal{N}(0,1)$$

and where $\mathcal{L}(\cdot)$ denotes the law of a random variable. The target measure μ is assumed to have a density with respect to γ of the form

$$\mu(dx) = M \exp(-\Phi(x))\gamma(dx). \tag{1.1}$$

With P_m being the projection onto the first m basis elements, we consider the following m-dimensional approximations to γ and μ

$$\begin{aligned} \gamma_m(dx) &= \mathcal{L}(\sum_{i=1}^{m} \lambda_i e_i \xi_i)(dx) \\ \mu_m(dx) &= M_m \exp(-\Phi(P_m x))\gamma_m(dx). \end{aligned} \tag{1.2}$$

The approximation error, namely the difference between μ and μ_m, is already well studied Dashti and Stuart (2011); Mattingly, Pillai and Stuart (2012b) and can be estimated in terms of the closeness between $\Phi \circ P_m$ and Φ.

In this article we consider Metropolis-Hastings MCMC methods Metropolis et al. (1953); Hastings (1970). For an overview of other MCMC methods, which have been developed and analysed, we refer the reader to Robert and Casella (2004); Liu (2008). The idea of the Metropolis-Hastings algorithm is to add an accept-reject mechanism to a Markov chain proposal in order to make the resulting Markov chain reversible with respect to the target measure. We denote the transition kernel of the underlying Markov chain by $Q(x, dy)$ and the acceptance probability for a proposed move from x to y by $\alpha(x, y)$. The transition kernel of the Metropolis-Hastings algorithm reads

$$\mathcal{P}(x, dz) = Q(x, dz)\alpha(x, z) + \delta_x(dz) \int (1 - \alpha(x, u))Q(x, du) \tag{1.3}$$

where $\alpha(x, y)$ is chosen such that $\mathcal{P}(x, dy)$ is reversible with respect to μ Tierney (1998). For the Random Walk Metropolis algorithm (RWM) the proposal kernel corresponds to

$$Q(x, dy) = \mathcal{L}(x + \sqrt{2\delta}\xi)(dy) \;,$$

with $\xi \sim \gamma_m$ which leads to the following acceptance probability

$$\alpha(x, y) = 1 \wedge \left(\Phi(x) - \Phi(y) + \frac{1}{2}\langle x, \mathcal{C}^{-1}x\rangle - \frac{1}{2}\langle y, \mathcal{C}^{-1}y\rangle \right). \tag{1.4}$$

Notice that the quadratic form $\frac{1}{2}\langle y, \mathcal{C}^{-1} y\rangle$ is almost surely infinite with respect to the proposal because it corresponds to the Cameron-Martin norm of y. For this reason the RWM algorithm is not defined on the infinite dimensional Hilbert space $\mathcal{H}$ (consult Cotter et al. (2011) for a discussion) and we will study it only on $m-$dimensional approximating spaces. In this article we will demonstrate that the RWM can be considerably improved by using the preconditioned Crank-Nicolson (pCN) algorithm which is defined via

$$Q(x, dy) = \mathcal{L}((1-2\delta)^{\frac{1}{2}} x + \sqrt{2\delta}\xi) \tag{1.5}$$

$$\alpha(x, y) = 1 \wedge \exp(\Phi(x) - \Phi(y)) \tag{1.6}$$

with $\xi \sim \gamma$. The pCN was introduced in Beskos et al. (2008) as the PIA algorithm, in the case $\alpha = 0$. Numerical experiments in Cotter et al. (2011) demonstrate its favourable properties in comparison with the RWM algorithm. In contrast to RWM, the acceptance probability is well-defined on a Hilbert space and this fact gives an intuitive explanation for the theoretical results derived in this paper in which we develop a theory explaining the superiority of pCN over RWM when applied on sequences of approximating spaces of increasing dimension. Our main positive results about pCN can be summarised as (rigorous statements in Theorems 2.14, 2.15, 4.3 and 4.4):

Claim. Suppose that both Φ and its local Lipschitz constant satisfy a growth assumption at infinity. Then for a fixed $0 < \delta \leq \frac{1}{2}$ the pCN algorithm applied to $\mu_m(\mu)$

I. has a unique invariant measure μ_m (μ);
II. has a Wasserstein spectral gap uniformly in m;
III. has an L^2-spectral gap $1-\beta$ uniform in m.

The corresponding sample average $S_n(f) = \frac{1}{n}\sum_{i=1}^{n} f(X_i)$

IV. satisfies a strong law of large numbers and a central limit theorem (CLT) for a class of locally Lipschitz functionals for every initial condition;
V. satisfies a CLT for μ (μ_m)-almost every initial condition with asymptotic variance uniformly bounded in m for $f \in L^2_\mu$ $\left(L^2_{\mu_m}\right)$;
VI. has an explicit bound on the mean square error (MSE) between itself and $\mu(f)$ for certain initial distributions ν.

These positive results about pCN clearly apply to $\Phi = 0$ which corresponds to the target measures γ and γ_m respectively; in this case the acceptance probability of pCN is always one and the theorems mentioned are simply statements about a discretely sampled Ornstein-Uhlenbeck (OU) process on $\mathcal{H}$ in this case. On the other hand the RWM algorithm applied to a specific Gaussian target measure γ_m has an L^2_μ-spectral gap which converges to 0 as $m \to \infty$ as fast as any negative power of m, see Theorem 2.17.

While it is a major contribution of this article to establish the results I, II and IV for pCN and to establish the negative results for RWM, the statements III, V and VI follow by verification of the conditions of known results.

In addition to the significance of these results in their own right for the understanding of MCMC methods, we would also like to highlight the techniques that we use in the proofs. We apply recently developed tools for the study of Markov chains on infinite dimensional spaces Hairer, Mattingly and Scheutzow (2011). The weak Harris theorem makes a Wasserstein spectral gap verifiable in practice and for reversible Markov processes it even implies an L^2-spectral gap.

1.1. *Literature Review.* The results in the literature can broadly be classified as follows Rudolf (2012); Meyn and Tweedie (2009):

1. For a metric on the space of measures such as the total variation or the Wasserstein metric the rate of convergence to equilibrium can be characterised through the decay of $d(\nu\mathcal{P}^n, \mu)$ where ν is the initial distribution of the Markov chain.
2. For the Markov operator $\mathcal{P}$ the convergence rate is given as the operator norm of $\mathcal{P}$ on a space of functions from X to $\mathbb{R}$ modulo constants. The most prominent example here is the L^2-spectral gap.
3. Direct methods like regeneration and the so-called split-chain which use the dynamics of the algorithm to introduce independence. The independence can be used to prove central limit theorem. Previous results have been formulated in terms of the following three main types of convergence:

Between these notions of convergence, there are many fruitful relations, for details consult Rudolf (2012). All these convergence types have been used to study MCMC algorithms.

The first systematic approach to prove L^2-spectral gaps for Markov chains was developed in Lawler and Sokal (1988) using the conductance concept due to Cheeger Cheeger (1970). These results were extended and applied to the Metropolis-Hastings algorithm with uniform proposal and a log-concave target distribution on a bounded convex subset of $\mathbb{R}^n$ in Lovász and Simonovits (1993). The consequences of a spectral gap for the ergodic average in terms of a CLT and the MSE have been investigated in Kipnis and Varadhan (1986); Cuny and Lin (2009) and Rudolf (2012) respectively and were first brought up in the MCMC literature in Geyer (1992); Chan and Geyer (1994).

For finite state Markov chains the spectral gap can be bounded in terms of quantities associated with its graph Diaconis and Stroock (1991). This idea has also been applied to the Metropolis-Algorithm in Sinclair and Jerrum (1989) and Frigessi et al. (1993).

A different approach using the splitting chain technique mentioned above was independently developed in Nummelin (1978) and Athreya and Ney (1978) to bound the total variation distance between the n-step kernel and the invariant measure. Small and petite sets are used in order to split the trajectory of a Markov chain into independent blocks. This theory was fully developed in Meyn and Tweedie (2009) and again adapted and applied to the Metropolis-Hastings algorithm in Roberts and Tweedie (1996) resulting in a criterion for geometric ergodicity

$$\|\mathcal{P}(x,\cdot)^n - \mu\|_{\mathrm{TV}} \leq C(x)c^n \qquad \text{for some } c < 1.$$

Moreover, they established a criterion for a CLT. Extending this method, it was also possible to derive rigorous confidence intervals in Łatuszyński and Niemiro (2011).

In most infinite dimensional settings the splitting chain method cannot be applied since measures tend to be mutually singular. The method is hence not well-adapted to the high-dimensional setting. Even Gaussian measures with the same covariance operator are only equivalent if the difference between their means lies in the Cameron-Martin space. As a consequence the pCN algorithm is not irreducible in the sense of Meyn and Tweedie (2009), hence there is no non trivial measure φ such that $\varphi(A) > 0$ implies $\mathcal{P}(x, A) > 0$ for all x. By inspecting the Metropolis-Hastings transition kernel (1.3), the pCN algorithm is not irreducible. More precisely if $x - y$ is not in the Cameron-Martin space $Q(x, dz)$ and $Q(y, dz)$ are mutually singular, consequently the same is true for $P(x, dz)$ and $P(y, dz)$. This may also be shown to be true for the n-step kernel by expressing it as a sum of densities times Gaussian measures and applying the Feldman-Hajek Theorem Da Prato and Zabczyk (1992).

For these reasons, existing theoretical results concerning RWM and pCN in high dimensions have been confined to scaling results and derivations of diffusion limits. In Beskos, Roberts and Stuart

(2009a) the RWM algorithm with a target that is absolutely continuous with respect to a product measure has been analysed for its dependence on the dimension. The proposal distribution is a centred normal random variable with covariance matrix $\sigma_n I_n$. The main result there is that δ has to be chosen as a constant times a particular negative power of n to prevent the expected acceptance probability to go to one or to zero. In a similar setup it was recently shown that there is a μ-reversible SPDE limit if the product law is a truncated Karhunen-Loeve expansion Mattingly, Pillai and Stuart (2012a). This SPDE limit suggests that the number of steps necessary for a certain level of accuracy grows like $\mathcal{O}(m)$ because $\mathcal{O}(m)$ steps are necessary in order to approximate the SPDE limit on $[0,T]$. A similar result in Pillai, Stuart and Thiéry (2011) suggests that the pCN algorithm only needs $\mathcal{O}(1)$ steps.

Uniform contraction in a Wasserstein distance was first applied to MCMC in Joulin and Ollivier (2010) in order to get bounds on the variance and bias of the sample average of Lipschitz functionals. We use the weak Harris theorem to verify this contraction and, by using the results from Rudolf (2012), we obtain non-asymptotic bounds on the sample average of L^2_μ-functionals. In Eberle (2012) exponential convergence for a Wasserstein distance is proved for the Metropolis-adjusted-Langevin (MALA) and pCN algorithm for log-concave measures having a density with respect to a Gaussian measure. The rates obtained in this article are explicit in terms of additional bounds on the derivates of the density. In our proofs we do not assume log-concavity. However, the rate obtained here is less explicit.

Similarly, approaches based on the Bakery-Emery criterion Bakry and Émery (1985) seem to be only applicable if the measure is log-concave.

1.2. *Outline.* In this paper we substantiate these ideas by using spectral gaps derived by an application of the weak Harris theory Hairer, Mattingly and Scheutzow (2011). Section 2 contains the statements of our main results, namely Theorems 2.12, 2.14 and 2.15 concerning the desirable dimension-independence properties of the pCN method and Theorem 2.17 dealing with the undesirable dimension dependence of the RWM method. Section 2 starts by specifying the RWM and pCN algorithms as Markov chains, the statement of the weak Harris theorem and a discussion of the relationship between exponential convergence in a Wasserstein distance and L^2_μ-spectral gaps. The proofs of the theorems in Section 2 are given in Section 3. We highlight that the key steps can be found in the Sections 3.1.2 and 3.2.2 where we dealt with the cases of global and local Lipschitz Φ respectively. In Section 4 we exploit the Wasserstein and L^2_μ-spectral gaps in order to derive a law of large numbers (LLN), central limit theorems (CLTs) and mean square error (MSE) bounds for sample-path ergodic averages of the pCN method, again emphasising the dimension independence of these results. We draw overall conclusions in Section 5.

Acknowledgement. We thank Feng-Yu Wang for pointing out the connection between Wasserstein and L^2-spectral gaps and Professor Andreas Eberle for many fruitful discussions about this topic.

2. Main Results. In Section 2.1 we specify the RWM and pCN algorithms before summarising the weak Harris theorem in Section 2.2. Subsequently, we describe how a Wasserstein spectral gap implies an L^2_μ-spectral gap. Based on the weak Harris theorem, we give necessary conditions on the target measure for the pCN algorithm in order to have a dimension independent spectral gap in a Wasserstein distance in Section 2.3. In Section 2.4 we highlight one of the disadvantages of the RWM by giving an example satisfying our assumptions for the pCN algorithm for which the spectral gap of the RWM algorithm converges to zero as fast as any negative power of m as $m \to \infty$.

Algorithm 1 Preconditioned Crank-Nicolson

Initialize X_0.
For $n \geq 0$ do:

1. Generate $\xi \sim \gamma$ and set $p_{X_n}(\xi) = (1-2\delta)^{\frac{1}{2}} X_n + \sqrt{2\delta}\,\xi$.
2. Set
$$X_{n+1} = \begin{cases} p_{X_n} & \text{with probability } \alpha(X_n, p_{X_n}) \\ X_n & \text{otherwise} \end{cases}$$
Here, $\alpha(x,y) = 1 \wedge \exp(\Phi(x) - \Phi(y))$.

Algorithm 2 Random Walk Metropolis

Initialise X_0.
For $n \geq 0$ do:

1. Generate $\xi \sim \gamma_m$ and set $p_{X_n}(\xi) = X_n + \sqrt{2\delta}\,\xi$.
2. Set
$$X_{n+1} = \begin{cases} p_{X_n} & \text{with probability } \alpha(X_n, p_{X_n}) \\ X_n & \text{otherwise} \end{cases}$$
Here, $\alpha(x,y) = 1 \wedge \exp(\Phi(x) - \Phi(y) + \frac{1}{2}\langle x, \mathcal{C}^{-1}x\rangle - \frac{1}{2}\langle y, \mathcal{C}^{-1}y\rangle)$.

2.1. *Algorithms.* We focus on convergence results for the pCN algorithm (Algorithm 1) which generates a Markov chain $\{X^n\}_{n\in\mathbb{N}}$ with $X^n \in H$ and $\{X_m^n\}_{n\in\mathbb{N}}$ when it is applied to the measures μ and μ_m respectively. The corresponding transition Markov kernels are called $\mathcal{P}$ and $\mathcal{P}_m$ respectively. We use the same notation for the Markov chain generated by the RWM (Algorithm 2). This should not cause confusion as the statements concerning the pCN and RWM algorithms occur in separate sections.

2.2. *Preliminaries.* In this section we review Lyapunov functions, Wasserstein distances, d-small sets and d-contracting Markov kernels in order to state a weak Harris theorem which has been recently proved by Hairer et al. Hairer, Mattingly and Scheutzow (2011). By weakening the notion of small sets, this theorem gives a sufficient condition for exponential convergence in a Wasserstein distance. Moreover, we explain how this implies an L^2-spectral gap.

2.2.1. *Weak Harris Theorem.*

DEFINITION 2.1. Given a Polish space $\mathbf{E}$, a function $d : \mathbf{E}\times\mathbf{E} \to \mathbb{R}_+$ is a *distance-like* function if it is symmetric, lower semi-continuous and $d(x,y) = 0$ is equivalent to $x = y$.

This induces the 1-Wasserstein "distance" associated with d for the measures ν_1, ν_2

$$d(\nu_1,\nu_2) = \inf_{\pi\in\Gamma(\nu_1,\nu_2)} \int_{\mathbf{E}\times\mathbf{E}} d(x,y)\pi(dx,dy) \tag{2.1}$$

where $\Gamma(\nu_1,\nu_2)$ is the set of couplings of ν_1 and ν_2 (all measures on $\mathbf{E}\times\mathbf{E}$ with marginals ν_1 and ν_2). If d is a metric, the Monge-Kantorovich duality states that

$$d(\nu_1,\nu_2) = \sup_{\|f\|_{Lip(d)}=1} \int f d\nu_1 - \int f d\nu_2.$$

We use the same notation for the distance and the associated Wasserstein distance; we hope that this does not lead to any confusion.

DEFINITION 2.2. A Markov kernel $\mathcal{P}$ is *d-contracting* if there is $0 < c < 1$ such that $d(x, y) < 1$ implies

$$d(\mathcal{P}(x, \cdot), \mathcal{P}(y, \cdot)) \leq c \cdot d(x, y).$$

DEFINITION 2.3. Let $\mathcal{P}$ be a Markov operator on a Polish space $\mathbf{E}$ endowed with a distance-like function $d : \mathbf{E} \times \mathbf{E} \to [0, 1]$. A set $S \subset \mathbf{E}$ is said to be *d-small* if there exists $0 < s < 1$ such that for every $x, y \in S$

$$d(\mathcal{P}(x, \cdot), \mathcal{P}(y, \cdot)) \leq s.$$

REMARK. The d-Wasserstein distance associated with

$$d(x, y) = \chi_{\{x \neq y\}}(x, y)$$

coincides with the total variation distance (up to a factor 2). If S is a small set Meyn and Tweedie (2009), then there exists a probability measure ν such that $\mathcal{P}$ can be decomposed into

$$\mathcal{P}(x, dz) = s\tilde{\mathcal{P}}(x, dz) + (1 - s)\nu(dz) \qquad \text{for } x \in S.$$

This implies that $d_{\mathrm{TV}}(\mathcal{P}(x, \cdot), \mathcal{P}(y, \cdot)) \leq s$ and hence S is d-small, too.

DEFINITION 2.4. A Markov kernel $\mathcal{P}$ has a Wasserstein spectral gap if there is a $\lambda > 0$ and a $C > 0$ such that

$$d(\nu_1\mathcal{P}^n, \nu_2\mathcal{P}^n) \leq C \exp(-\lambda n) d(\nu_1, \nu_2) \text{ for all } n \in \mathbb{N}.$$

DEFINITION 2.5. V is a *Lyapunov* function for the Markov operator $\mathcal{P}$ if there exist $K > 0$ and $0 \leq l < 1$ such that

$$\mathcal{P}^n V(x) \leq l^n V(x) + K \text{ for all } x \in \mathbf{E} \text{ and all } n \in \mathbb{N}. \tag{2.2}$$

(Note that the bound for $n = 1$ implies all other bounds but with a different constant K.)

PROPOSITION 2.6. *(**Weak Harris Theorem** Hairer, Mattingly and Scheutzow (2011)) Let $\mathcal{P}$ be a Markov kernel over a Polish space $\mathbf{E}$ and assume that*

1. *$\mathcal{P}$ has a Lyapunov function V such that (2.2) holds;*
2. *$\mathcal{P}$ is d-contracting for a distance-like function $d : \mathbf{E} \times \mathbf{E} \to [0, 1]$;*
3. *the set $S = \{x \in \mathbf{E} : V(x) \leq 4K\}$ is d-small.*

Then there exists $\tilde{n}$ such that for any two probability measures ν_1, ν_2 on $\mathbf{E}$ we have

$$\tilde{d}(\nu_1\mathcal{P}^{\tilde{n}}, \nu_2\mathcal{P}^{\tilde{n}}) \leq \frac{1}{2}\tilde{d}(\nu_1, \nu_2)$$

where $\tilde{d}(x, y) = \sqrt{d(x, y)(1 + V(x) + V(y))}$ and $\tilde{n}(l, K, c, s)$ is increasing in l, K,c and s. In particular there is at most one invariant measure. Moreover, if there exists a complete metric d_0 on $\mathbf{E}$ such that $d_0 \leq \sqrt{d}$ and such that $\mathcal{P}$ is Feller on $\mathbf{E}$, then there exists a unique invariant measure μ for $\mathcal{P}$.

REMARK. Setting $\nu_2 = \mu$ we obtain the convergence rate to the invariant measure.

2.2.2. *The Wasserstein spectral gap implies an L^2-spectral gap.* In this section we give reasons why a Wasserstein spectral gap implies an L^2_μ- spectral gap under mild assumptions for a Markov kernel $\mathcal{P}$. The proof is based on a comparison of different powers of $\mathcal{P}$ using the spectral theorem.

DEFINITION 2.7. (L^2_μ-spectral gap) A Markov operator $\mathcal{P}$ with invariant measure μ has an L^2_μ-spectral gap $1-\beta$ if

$$\beta = \|\mathcal{P}\|_{L^2_0 \to L^2_0} = \sup_{f \in L^2_\mu} \frac{\|\mathcal{P}f - \mu(f)\|_2}{\|f - \mu(f)\|_2} < 1.$$

The following proposition is due to F.-Y. Wang and is a discrete-time version of Theorem 2.1(2) in Wang (2003). The proof given below is from private communication with F.-Y. Wang and is presented because of its beauty and the tremendous consequences in combination with the weak Harris theorem.

PROPOSITION 2.8. *(Private Communication Röckner and Wang (2001)) Let $\mathcal{P}$ be a Markov transition operator which is reversible with respect to μ and suppose that $Lip(\tilde{d}) \cap L^\infty_\mu$ is dense in L^2_μ. If for every such f there exists a constant $C(f)$ such that*

$$\tilde{d}((\mathcal{P}^n f)\mu, \mu) \le C(f) \exp(-\lambda n) \tilde{d}(f\mu, \mu),$$

then this implies the L^2_μ-spectral gap

$$\|\mathcal{P}^n f - \mu(f)\|_2^2 \le \|f - \mu(f)\|_2^2 \exp(-\lambda n). \tag{2.3}$$

PROOF. First assume that $0 \le f \in Lip(\tilde{d}) \cap L^\infty(\mu)$ with $\mu(f) = 1$ and π being the optimal coupling between $(\mathcal{P}^{2n} f)\mu$ and μ for the Wasserstein distance associated with d. Reversibility implies $\int (\mathcal{P}^n f)^2 d\mu = \int (\mathcal{P}^{2n} f) f d\mu$ which leads to

$$\begin{aligned} \|\mathcal{P}^n f - \mu(f)\|_2^2 &= \mu\left((\mathcal{P}^n f)^2\right) - 1 = \int (f(x) - f(y)) d\pi \\ &\le Lip(f) \int \tilde{d}(x,y) d\pi \le Lip(f) \tilde{d}(\mathcal{P}^{2n} f\mu, \mu) \\ &= Lip(f) \tilde{d}((f\mu)\mathcal{P}^{2n}, \mu) \le C Lip(f) \exp(-2\lambda n). \end{aligned}$$

Since the above extends to $a \cdot f$, we note that for general $f \in L^\infty \cap Lip(\tilde{d})$

$$\|P_t f - \mu(f)\|_2^2 \le 2\left\|P_t f^+ - \mu(f^+)\right\|_2^2 + 2\left\|P_t f^- - \mu(f^-)\right\|_2^2.$$

By Lemma 2.9, the bound (2.3) holds for functions in $Lip \cap L^\infty(\mu)$. Hence the result follows by taking limits of such functions. □

LEMMA 2.9. *Let $\mathcal{P}$ be a Markov transition operator which is reversible with respect to μ. If the following relationship holds for some $f \in L^2(\mu)$, the constants $C(f)$ and $\lambda > 0$*

$$\|\mathcal{P}^n f - \mu(f)\|_2^2 \le C(f) \exp(-\lambda n) \text{ for all } n,$$

then for the same f

$$\|\mathcal{P}^n f - \mu(f)\|_2^2 \le \|f - \mu(f)\|_2^2 \exp(-\lambda n) \text{ for all } n.$$

PROOF. Without loss of generality we assume that $\mu(\hat{f}^2) = 1$ where $\hat{f} = f - \mu(f)$. Applying the spectral theorem to $\mathcal{P}$ yields the existence of a unitary map $U : L^2(\mu) \mapsto L^2(X, \nu)$ such that UPU^{-1} is a multiplication operator by m. Moreover, $\mu(\hat{f}^2) = 1$ implies that $(U\hat{f})^2\nu$ is a probability measure. Thus for $k \in \mathbb{N}$

$$\begin{aligned}\int (\mathcal{P}^n \hat{f}(x))^2 d\mu &= \int m(x)^{2n}(U\hat{f})^2(x)d\nu = \int m(x)^{(2n+k)\frac{2n}{2n+k}} d(U\hat{f})^2\nu \\ &\leq \left(\int m(x)^{2n+k} d(U\hat{f})^2\nu\right)^{\frac{2n}{2n+k}} \leq C^{\frac{2n}{2n+k}} \exp(-\lambda 2n).\end{aligned}$$

Letting $k \to \infty$ yields the required claim. □

2.3. *Dimension-Independent Spectral Gaps for the pCN-Algorithm.* Using the weak Harris theorem, we give necessary conditions on μ (see (1.1)) in terms of regularity and growth of Φ to have a uniform spectral gap in a Wasserstein distance for $\mathcal{P}$ and $\mathcal{P}^m$. We need Φ to be at least locally Lipschitz; the case where it is globally Lipschitz is more straightforward and is presented first. Using the notation $\rho = 1 - (1 - 2\delta)^{\frac{1}{2}}$ we can express the proposal of the pCN algorithm as

$$p_{X^n}(\xi) = (1 - \rho)X^n + \sqrt{2\delta}\,\xi.$$

The following results do all hold for δ in $(0, \frac{1}{2}]$:

The mean of the proposal $(1-\rho)X^n$ suggests that we can prove that $f(\|\cdot\|)$ is a Lyapunov function for certain f and that $\mathcal{P}$ is d-contracting (for a suitable metric). This relies on having a lower bound on the probability of X_{n+1} being in a ball around the mean. In fact, our assumptions are stronger because we assume a uniform lower bound on $\mathbb{P}(p_x \text{ is accepted} | p_x = z)$ for z in $B_{r(\|x\|)}((1 - \rho)x)$.

ASSUMPTION 2.10. *There is $R > 0$, $\alpha_l > -\infty$. and a function $r : \mathbb{R}^+ \mapsto \mathbb{R}^+$ satisfying $r(s) \leq \frac{\rho}{2}s$ for all $|s| \geq R$ such that for all $x \in B_R(0)^c$*

$$\inf_{z \in B_{r(\|x\|)}((1-\rho)x)} \alpha(x, z) = \inf_{z \in B_{r(\|x\|)}((1-\rho)x)} \exp\left(-\Phi(z) + \Phi(x)\right) > \exp(\alpha_l). \tag{2.4}$$

ASSUMPTION 2.11. *Let Φ in (1.1) have global Lipschitz constant L and assume that $\exp(-\Phi)$ is γ-integrable.*

THEOREM 2.12. *Let Assumption 2.10 and 2.11 be satisfied with either*

1. *$r(\|x\|) = r\|x\|^a$ where $r \in \mathbb{R}^+$ for any $a \in (\frac{1}{2}, 1)$ then we consider $V = \|x\|^i$ with $i \in \mathbb{N}$ or $V = \exp(v\|x\|)$, or*
2. *$r(\|x\|) = r \in R$ for $r \in \mathbb{R}^+$ then we take $V = \|x\|^i$ with $i \in \mathbb{N}$.*

Under these assumptions μ_m (μ) is the unique invariant measure for the Markov chain associated with the pCN algorithm applied to μ_m (μ). Moreover, define

$$\begin{aligned}\tilde{d}(x, y) &= \sqrt{d(x, y)(1 + V(x) + V(y))} \text{ with} \\ d(x, y) &= 1 \wedge \frac{\|x - y\|}{\epsilon}.\end{aligned}$$

Then for ϵ small enough there exists an $\tilde{n}$ such that for all probability measures ν_1 and ν_2 on $\mathcal{H}$ and $P_m\mathcal{H}$ respectively

$$\begin{aligned}\tilde{d}(\nu_1\mathcal{P}^{\tilde{n}}, \nu_2\mathcal{P}^{\tilde{n}}) &\leq \frac{1}{2}\tilde{d}(\nu_1, \nu_2), \\ \tilde{d}(\nu_1\mathcal{P}_m^{\tilde{n}}, \nu_2\mathcal{P}_m^{\tilde{n}}) &\leq \frac{1}{2}\tilde{d}(\nu_1, \nu_2)\end{aligned}$$

for all $m \in \mathbb{N}$.

PROOF. The conditions of the weak Harris theorem (Proposition 2.6) are satisfied by the Lemmata 3.2, 3.3 and 3.4. □

A key step in the proof is to verify the d-contraction. In order to obtain an upper bound on $d(\mathcal{P}(x,\cdot),\mathcal{P}(y,\cdot))$ (see (2.1)), we choose a particular coupling between the algorithm started at x and y and distinguish between the cases when both proposals are accepted, both are rejected and only one is accepted. The case when only one of them is accepted is the most difficult to tackle. By choosing $d = 1 \wedge \frac{\|x-y\|}{\epsilon}$ with ϵ small enough, it turns out that the Lipschitz constant of $\alpha(x,y)$ can be brought under control.

By changing the distance function d, we can also handle the case when Φ is locally Lipschitz provided that the local Lipschitz constant does not grow too fast.

ASSUMPTION 2.13. *Let $\exp(-\Phi)$ be integrable with respect to γ and assume that for any $\kappa > 0$ there is an M_κ such that*

$$\phi(r) = \sup_{x \neq y \in B_r(0)} \frac{|\Phi(x) - \Phi(y)|}{\|x-y\|} \leq M_\kappa e^{\kappa r}.$$

THEOREM 2.14. *Let the Assumptions 2.10 and 2.13 be satisfied with $r(\|x\|) = r\|x\|^a$ where $r \in \mathbb{R}$, $a \in (\frac{1}{2}, 1)$ and either $V = \|x\|^i$ with $i \in \mathbb{N}$ or $V = \exp(v\|x\|)$.*

Then μ_m (μ) is the unique invariant measure for the Markov chain associated with the pCN algorithm applied to μ_m (μ).

For $\mathsf{A}(T,x,y) := \{\psi \in C^1([0,T],\mathcal{H}), \psi(0) = x, \psi(T) = y, \|\dot\psi\| = 1\}$,

$$\begin{aligned} \tilde d(x,y) &= \sqrt{d(x,y)(1+V(x)+V(y))} \text{ with} \\ d(x,y) &= 1 \wedge \inf_{T,\psi\in\mathsf{A}(T,x,y)} \frac{1}{\epsilon}\int_0^T \exp(\eta\|\psi\|)dt \end{aligned}$$

and η and ϵ small enough there exists an $\tilde n$ such that for all ν_1, ν_2 probability measures on $\mathcal{H}$ and on $P_m\mathcal{H}$ respectively and $m \in \mathbb{N}$

$$\begin{aligned} \tilde d(\nu_1\mathcal{P}^{\tilde n}, \nu_2\mathcal{P}^{\tilde n}) &\leq \frac{1}{2}\tilde d(\nu_1,\nu_2) \\ \tilde d(\nu_1\mathcal{P}_m^{\tilde n}, \nu_2\mathcal{P}_m^{\tilde n}) &\leq \frac{1}{2}\tilde d(\nu_1,\nu_2). \end{aligned}$$

PROOF. This time the Lemmata 3.2, 3.6 and 3.7 verify the conditions of the weak Harris theorem (Proposition 2.6). □

REMARK. Our arguments work for $\delta \in (0, \frac{1}{2}]$; for $\delta = \frac{1}{2}$ the pCN algorithm becomes the independence sampler and the Markov transition kernel becomes irreducible so that this case can be dealt by with the theory of Meyn and Tweedie Meyn and Tweedie (2009).

In order to get the same lower bound for the L^2_μ-spectral gap, we just have to verify that $Lip(\tilde\delta) \cap L^\infty_\mu$ is dense in L^2_μ.

THEOREM 2.15. *If the conditions of Theorem 2.12 or 2.14 are satisfied, then we have the same lower bound on the L^2_μ-spectral gap of $\mathcal{P}$ and $\mathcal{P}_m$ uniformly in m.*

PROOF. By Proposition 2.8 we only have to show that $Lip(\tilde{d}) \cap L^\infty(\mu)$ is dense in $L^2(H, \mathcal{B}, \mu)$. Since $\tilde{d}(x, y) \geq C(1 \wedge \|x - y\|)$, one has $Lip(\|\cdot\|) \cap L^\infty(\mu) \subseteq Lip(\tilde{d})$, so that it is enough to show that $Lip(\|\cdot\|) \cap L^\infty(\mu)$ is dense in $L^2(H, \mathcal{B}, \mu)$. Suppose not, then there is $0 \neq g \in L^2(\mu)$ such that

$$\int f g d\mu = 0 \quad \text{for all } f \in Lip \cap L^\infty(\mu).$$

Since all Borel probability measures on a separable Banach space are characterised by their Fourier transform (Bochner's theorem, see for example Bogachev (2007)), they are characterised by integrals against bounded Lipschitz functions. Hence $g d\mu$ is the zero measure and hence $g \equiv 0$ in L^2_μ. □

2.4. *Dimension-Dependent Spectral Gaps for RWM.* So far we have shown convergence results for the pCN. Therefore we present an example subsequently where these results apply but the spectral gap of the RWM goes to 0 as m tends to infinity. We consider the target measures μ_m on

$$\mathcal{H}^\sigma_m := \left\{ x \,\middle|\, \|x\|_\sigma = \sum_{i=1}^m i^{2\sigma} x_i^2 < \infty \right\}$$

with $0 < \sigma < \frac{1}{2}$ given by

$$\mu_m = \gamma_m = \mathcal{L}\left(\sum_{i=1}^m \frac{1}{i} \xi_i e_i \right) \qquad \xi \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, 1). \tag{2.5}$$

In the setting of (1.1) this corresponds to $\Phi = 0$. Hence the assumptions of Theorem 2.14 are satisfied and we obtain a uniform lower bound on the L^2_μ-spectral gap for the pCN. For the RWM algorithm we show that the spectral gap converges to zero faster than any negative power of m if we scale $\delta = s\, m^{-a}$ for any $a \in [0, 1)$.

Using the notion of conductance

$$\mathsf{C} = \inf_{\mu(A) \leq \frac{1}{2}} \frac{\int_A \mathcal{P}(x, A^c) d\mu(x)}{\mu(A)}, \tag{2.6}$$

we obtain an upper bound on the spectral gap by Cheeger's inequality Lawler and Sokal (1988); Sinclair and Jerrum (1989)

$$1 - \beta \leq 2\mathsf{C}. \tag{2.7}$$

Our main observation is that there is a simple upper bound for the conductance of a Metropolis-Hastings algorithm because it can only move from a set A if

- the proposed move lies in A^c and
- the proposed move is accepted.

Just considering either event gives rise to simple upper bounds that can be used to make many results from the scaling analysis rigorous. We denote the expected acceptance probability for a proposal from x as

$$\alpha(x) = \int_{\mathcal{H}} \alpha(x, y) dQ(x, dy).$$

Considering only the acceptance of the proposal gives rise to

$$C \leq \inf_{\mu(A) \leq \frac{1}{2}} \frac{\int_A \alpha(x) \mu(dx)}{\mu(A)}.$$

In particular, for any set B such that $\mu(B) \leq \frac{1}{2}$, it follows that

$$C \leq \sup_{x \in B} \alpha(x)$$

and also that

$$C \leq 2\mathbb{E}_\mu \alpha(x).$$

The last result allows us to make scaling results like those in Beskos, Roberts and Stuart (2009b) rigorous. Similarly, just supposing that the Metropolis-Hastings algorithm accepts all proposals gives rise to the following bound

$$C \leq \inf_{\mu(A) \leq \frac{1}{2}} \frac{\int_A Q(x, A^c) d\mu(x)}{\mu(A)}.$$

We summarise these observations in the subsequent proposition.

PROPOSITION 2.16. *Let $\mathcal{P}$ be a Metropolis-Hastings transition kernel for a target measure μ with proposal kernel $Q(x, dy)$ and acceptance probability $\alpha(x, y)$. The L^2_μ-spectral gap can be bounded by*

$$1 - \beta \leq 1 - \Lambda \leq 2C \leq 2 \begin{cases} \sup_{x \in B} \alpha(x) & \text{for any } \mu(B) \leq \frac{1}{2} \\ 2\mathbb{E}_\mu \alpha(x) \end{cases} \tag{2.8}$$

and

$$1 - \beta \leq 1 - \Lambda \leq 2C \leq 2 \inf_{\mu(A) \leq \frac{1}{2}} \frac{\int_A Q(x, A^c) d\mu(x)}{\mu(A)}. \tag{2.9}$$

In the following Theorem we use the Proposition 2.16 for the RWM algorithm applied to μ_m as in Equation (2.5) in order to quantify the behaviour of the spectral gap as m goes to ∞. We consider polynomial scaling of the step size parameter of the form $\delta_m \sim m^{-a}$ to zero. For $a < 1$ the bound in Equation 2.8 is most useful as the acceptance behaviour is the determining quantity. For $a \geq 1$ the bound in Equation 2.9 is most useful as the properties of the proposal kernel are determining in this regime.

THEOREM 2.17. *Let $\mathcal{P}_m$ be the Markov kernel and α be the acceptance probability associated with the RWM algorithm applied to μ_m as in Equation (2.5).*

1. *For $\delta_m \sim m^{-a}$, $a \in [0, 1)$ and any p there exists a $K(p, a)$ such that the spectral gap of $\mathcal{P}_m$ satisfies*
$$1 - \beta_m \leq K(p, a) m^{-p}.$$
2. *For $\delta_m \sim m^{-a}$, $a \in [1, \infty)$ there exists a $K(a)$ such that the spectral gap of $\mathcal{P}_m$ satisfies*
$$1 - \beta_m \leq K(a) m^{-\frac{a}{2}}.$$

PROOF. For the first part of this proof we work on the space H_σ with $\sigma \in [0, \frac{1}{2})$ where we determine σ later. We choose $B_r(0)$ such that $\mu(B_r(0)) \leq \frac{1}{4}$ and by (3.1) below we know that $\mu_m(B^m_r(0))$ is decreasing towards $\mu(B_r(0))$. Hence for all m larger than some M we know that $\mu(B^m_r(0)) \leq \frac{1}{2}$. In order to apply Proposition 2.16, we have to gain an upper bound on $\alpha(x)$ in $B^m_r(0)$. Thus we use $u \wedge v \leq u^\lambda v^{1-\lambda}$ to bound

$$\alpha(x, y) = 1 \wedge \exp\left(-\sum_{i=1}^m \frac{i^2}{2}(y_i^2 - x_i^2) \right) \leq \exp\left(-\sum_{i=1}^m \frac{i^2}{2}(y_i^2 - x_i^2)\lambda \right).$$

Using this inequality, we can find an upper bound on the acceptance probability $\alpha(x)$.

$$\int \alpha Q(x,dy) \leq \int \frac{m!}{(4\delta\pi)^{\frac{m}{2}}} \exp\left(-\sum_{i=1}^{m} \frac{i^2}{2}\left[(y_i^2-x_i^2)\lambda + \frac{(x_i-y_i)^2}{2\delta}\right]\right) dy.$$

Completing the square and using the normalisation constant yields

$$\begin{aligned}
&\leq \int \frac{m!}{(4\delta\pi)^{\frac{m}{2}}} \exp\left(-\sum_{i=1}^{m} \frac{i^2}{2}\left[(\lambda+\frac{1}{2\delta})\left(y_i - \frac{x_i}{2\delta\lambda+1}\right)^2 - \frac{2\delta\lambda^2 x_i^2}{(2\delta\lambda+1)}\right]\right) dy \\
&\leq (1+2\lambda\delta)^{-\frac{m}{2}} \exp\left(\sum_{i=1}^{m} \frac{\delta\lambda^2 i^2 x_i^2}{(2\delta\lambda+1)}\right).
\end{aligned}$$

For $x \in B_r^m(0)$ in $\mathcal{H}_\sigma$, using $\delta = m^{-a}$ and setting $\lambda = m^{-b}$

$$\alpha(x) \quad \leq \quad (1+2m^{-(a+b)})^{-\frac{m}{2}} \exp\left(\frac{r m^{2-2\sigma-a-2b}}{3}\right).$$

We want to choose a and b in the above equation such that the RHS goes to zero as $m \to \infty$. In order to obtain decay from the first factor, we need that $a+b<1$ and to prevent growth from the second $a+2b > 2-2\sigma$ which corresponds to $a+2b>1$ for σ sufficiently close to $\frac{1}{2}$. This can be satisfied with $b = \frac{2(1-a)}{3}$ and $\sigma = \frac{2+a}{6} < \frac{1}{2}$. In this case the first factor decays faster than any negative power of m since

$$(1+2m^{-(a+b)})^{-\frac{m}{2}} = \exp\left(-\frac{m}{2}\log(1+2m^{-(a+b)})\right) \leq \exp(-Cm^{1-(a+b)}).$$

For the second part of the poof we use $\alpha(x,y) \leq 1$ and $A = \{x \in \mathbb{R}^n | x_1 \geq 0\}$ which by using a symmetry argument satisfies $\gamma_m(A) = \frac{1}{2}$ to bound the conductance

$$\begin{aligned}
\frac{\mathsf{C}}{2} &\leq \int_A P(x,A^c) d\mu \\
&\leq \int_A \int_{A^c} \frac{\alpha(x,y) n!^2}{(2\pi)^n (2\delta)^{\frac{n}{2}}} \exp\left(-\frac{1}{2}\sum_{i=1}^{m} i^2 (x_i^2 + (x_i-y_i)^2/(2\delta)\right) dxdy \\
&\leq .t_0^\infty \int_{-\infty}^{0} \frac{\exp(-\frac{1}{2}\frac{(y_1-x_1)^2}{2\delta})}{2\pi\sqrt{2\delta}} dy_1 \exp\left(-\frac{1}{2}x_1^2\right) dx_1 \\
&= \int_0^\infty \int_{-\infty}^{-\frac{x_1}{\sqrt{2\delta}}} \frac{\exp(-\frac{1}{2}z^2)}{2\pi} dy_1 \exp\left(-\frac{1}{2}x_1^2\right) dx_1.
\end{aligned}$$

Combining Fernique's theorem and Markov's inequality yields

$$\mathsf{C} \leq K \int_0^\infty \exp(-\frac{1}{2}(\frac{\delta+1}{\delta})x_1^2) dx \leq K\sqrt{2\pi\frac{\delta}{\delta+1}} \leq \tilde{K} m^{-\frac{a}{2}},$$

so that the claim follows again by an application of Cheeger's inequality. □

3. Spectral Gap: Proofs. We check the three conditions of the weak Harris theorem (Proposition 2.6) for globally and locally Lipschitz Φ (see (1.1)) in the Sections 3.1 and 3.2 respectively. For each condition we use the following lemma for the dependence of the constants l, K, c and s in the weak Harris theorem on m. This allows us to conclude that there exists $\tilde{n}(m) \leq \tilde{n}$ such that

$$\begin{aligned} \tilde{d}(\nu_1 \mathcal{P}^{\tilde{n}}, \nu_2 \mathcal{P}^{\tilde{n}}) &\leq \frac{1}{2}\tilde{d}(\nu_1, \nu_2) \\ \tilde{d}(\nu_1 \mathcal{P}_m^{\tilde{n}(m)}, \nu_2 \mathcal{P}_m^{\tilde{n}(m)}) &\leq \frac{1}{2}\tilde{d}(\nu_1, \nu_2) \end{aligned}$$

for all measures ν_1, ν_2 probability measures on $\mathcal{H}$ and $P_m\mathcal{H}$ respectively.

Replacing $r(s) \wedge \frac{\rho}{2}s$ only weakens the condition (2.4) so we can and will assume that $r(s) \leq \rho s/2$.

LEMMA 3.1. *Let $f : \mathbb{R} \to \mathbb{R}$ be monotone increasing, then*

$$\int f(\|\xi\|) d\gamma_m(\xi) \leq \int f(\|\xi\|) d\gamma(\xi)$$

and in particular

$$\gamma_m(B_R(0)) \geq \gamma(B_R(0)). \tag{3.1}$$

PROOF. The truncated Karhunen-Loeve expansion relates γ_m to γ and yields

$$\sum_{i=1}^{m} \lambda_i \xi_i^2 \leq \sum_{i=1}^{\infty} \lambda_i \xi_i^2.$$

Hence the result follows by monotonicity of the integral and of the function f

$$\int f(\|\xi\|) d\gamma_m(\xi) = \mathbb{E}(\sqrt{f(\sum_{i=1}^{m} \lambda_i \xi_i^2)}) \leq \mathbb{E}(\sqrt{f(\sum_{i=1}^{\infty} \lambda_i \xi_i^2)}) = \int f(\|\xi\|) d\gamma(\xi).$$

This yields Equation (3.1) by inserting $f = \chi_{B_R(0)^c}$. □

3.1. *Global log-Lipschitz density.* In this section we will prove Theorem 2.12 by checking the three conditions of the weak Harris Theorem 2.6 for the distance-like functions

$$d(x, y) = 1 \wedge \frac{\|x - y\|}{\epsilon}. \tag{3.2}$$

3.1.1. *Lyapunov Functions.* Under Assumption 2.10 we show the existence of a Lyapunov function V. This follows from two facts: First, the decay of V on $B_{r(\|x\|)}((1-\rho)x)$ and second the probability of the next step of the algorithm lying in that ball can be bounded below by Fernique's theorem, see Proposition A.1. Similarly, we will use the second part of Proposition A.1 to deal with proposals outside $B_{r(\|x\|)}((1-\rho)x)$.

LEMMA 3.2. *If Assumption 2.10 is satisfied with*

1. *$r(\|x\|) = r \in \mathbb{R}$; or*
2. *$r(\|x\|) = r\|x\|^a$, $\kappa > 0$ and $a \in (\frac{1}{2}, 1)$.*

Then the function $V(x) = \|x\|^i$ with $i \in \mathbb{N}$ in the first case and additionally $V(x) = \exp(\ell\|x\|)$ in the second case are Lyapunov functions for both $\mathcal{P}$ and $\mathcal{P}_m$ with constants l and K uniform in m.

PROOF. In both cases we choose R as in Assumption 2.10. Then there exists a constant K_1 such that

$$\sup_{x\in B_R(0)} \mathcal{P}V(x) \leq \sup_{x\in B_R(0)} \int \left(\|x\| + \sqrt{2\delta}\,\|\xi\|\right)^i d\gamma(\xi) =: K_1 < \infty\,.$$

On the other hand, there exists $0 < \tilde{l} < 1$ such that for all $x \in B_R(0)^c$

$$\sup_{y\in B_{r(\|x\|)}((1-\rho)x)} V(y) \leq \tilde{l}V(x). \tag{3.3}$$

We denote by $A = \{\omega | \sqrt{2\delta}\,\|\xi\| \leq r(\|x\|)\}$ the event that the proposal lies in a ball with a lower bound on the acceptance probability due to Assumption 2.10. This yields the bound

$$\begin{aligned}
\mathcal{P}V &\leq \mathbb{P}(A)\left[\mathbb{P}(accept|A)\tilde{l}V(x) + \mathbb{P}(reject|A)V(x)\right] + \mathbb{E}(V(p_x)\vee V(x); A^c)\\
&\leq \mathbb{P}(A)\left[(1-\mathbb{P}(accept|A)(1-\tilde{l})\right]V(x) + \mathbb{E}(V(p_x)\vee V(x); A^c)\\
&\leq \theta\mathbb{P}(A)V(x) + \mathbb{E}(V(p_x)\vee V(x); A^c)
\end{aligned}$$

for some $\theta < 1$. It remains to consider $\mathbb{E}(V(p_x)\vee V(x); A^c)$ where the differences will arise between the cases 1 and 2. For the first case we know that by an application of Fernique's theorem

$$\begin{aligned}
\mathbb{E}(V(p_x)\vee V(x); A^c) &\leq \int_{\sqrt{2\delta}\|\xi\|\geq c} \|x\|^i \vee \left((1-\rho)\,\|x\| + \sqrt{2\delta}\,\|\xi\|)\right)^i d\gamma(\xi)\\
&\leq \int_{\|\xi\|\geq \frac{c}{\sqrt{2\delta}}} \left(\|x\|^i + K\,\|\xi\|^p\right) d\gamma(\xi)\\
&\leq \mathbb{P}(A^c)V(x) + K_2.
\end{aligned}$$

Because a ball around the mean of a Gaussian measure on a separable space always has positive mass (Theorem 3.6.1 in Bogachev (1998)), we note that

$$\mathcal{P}V(x) \leq V(x)(\mathbb{P}(A)\theta + \mathbb{P}(A^c)) + K_2 \leq lV(x) + K_2,$$

for some constant $l < 1$.

For the second case we estimate

$$\mathbb{E}(V(p_x)\vee V(x); A^c) \leq M_v \int_{\|\eta\|>r\|x\|^a} e^{v(\|x\|+\sqrt{2\delta}\|\xi\|)} d\gamma(\xi).$$

The right hand side of the above is uniformly bounded in $x \in B_R(0)^c$ by some K_2 due to Proposition A.1. Hence in both cases there exists an $l < 1$ such that

$$\mathcal{P}V(x) \leq lV(x) + \max(K_1, K_2) \quad \forall x.$$

For the m-dimensional approximation $\mathcal{P}_m$ the probability of the event A is larger than for $\mathcal{P}$ by Lemma 3.1. Since there is a common lower bound for $\mathbb{P}(accept|A)$ $l(m)$ is smaller than or equal to l. Similarly, $K_i(m)$ is smaller than K_i by Lemma 3.1. □

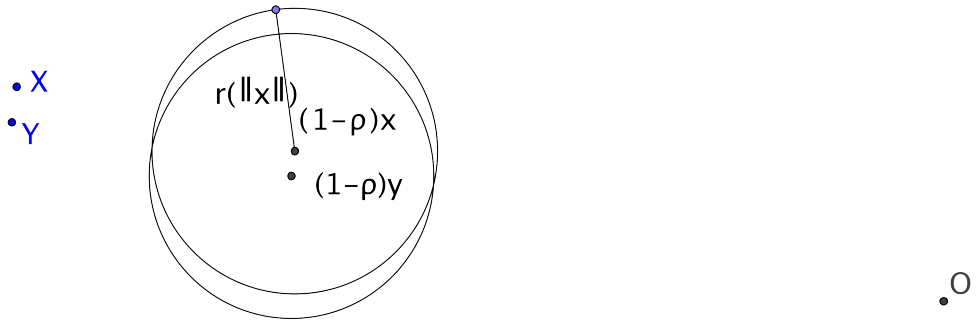


FIGURE 1. *Contraction*

3.1.2. *The d-Contraction.* In this section we show that $\mathcal{P}$ is d-contracting for $d(x,y) = 1 \wedge \frac{\|x-y\|}{\epsilon}$ by bounding $d(\mathcal{P}(x,\cdot),\mathcal{P}(y,\cdot))$ (see (2.1)) with a particular coupling. For x and y we choose the same noise ξ giving rise to the proposals $p_x(\xi)$ and $p_y(\xi)$ and the same uniform random variable for acceptance. Subsequently, we will refer to this coupling as the basic coupling and bound the expectation of d under this coupling by inspecting the following cases:

1. The proposals for the algorithm started at x and y are both accepted.
2. Both proposals are rejected.
3. One of the proposals is accepted and the other rejected.

LEMMA 3.3. *If Φ in (1.1) satisfies Assumption 2.10 and 2.11, then $\mathcal{P}$ and $\mathcal{P}_m$ are d-contracting for d as in (3.2) with a contraction constant uniform in m.*

PROOF. By Definition 2.2 we only need to consider x and y such that $d(x,y) < 1$ which implies that $\|x-y\| < \epsilon$. Later we will choose $\epsilon \ll 1$ so that if $\|x-y\| < \epsilon$, then either $x, y \in B_R(0)$ or $x, y \in B^c_{\tilde{R}}(0)$ with $\tilde{R} = R - 1$ and we will treat both cases separately. We assume without loss of generality that $\|y\| \geq \|x\|$.

For $x, y \in B_R(0)$ and $A = \{\omega | \sqrt{2\delta}\,\|\xi\| \leq R\}$ the basic coupling yields

$$\begin{aligned} d(\mathcal{P}(x,\cdot),\mathcal{P}(y,\cdot)) \quad \leq \quad & \mathbb{P}(A)\,[\mathbb{P}(\text{both accept}|A)(1-\rho)d(x,y) + \\ & \mathbb{P}\,(\text{both reject}|\text{A})d(x,y))] + \mathbb{P}(A^c)d(x,y) + \\ & \int_{\mathcal{H}} |\alpha(x,p_x)(\xi) - \alpha(y,p_y)(\xi)|\,d\gamma(\xi) \end{aligned} \tag{3.4}$$

where the last term bounds the case that only one of the proposals is accepted. Using the bound $\mathbb{P}(\text{both reject}|\text{A}) \leq 1 - \mathbb{P}(\text{both accept}|A)$ yields a non-trivial convex combination of d and $(1-\rho)d$ because the probability $\mathbb{P}(\text{both accept}|A)$ is bounded below by $\exp(-\sup\{\Phi(z)|\,\|z\| \leq 2R\} + \inf\{\Phi(z)|\,\|z\| \leq 2R\})$ due to (1.5). The first two summands in (3.4) form again a non-trivial convex combination since $\mathbb{P}(A) > 0$ so that there is $\tilde{c} < 1$ with

$$d(\mathcal{P}(x,\cdot),\mathcal{P}(y,\cdot)) \leq \tilde{c}d(x,y) + \int_{\mathcal{H}} |\alpha(x,p_x)(\xi) - \alpha(y,p_y)(\xi)|\,d\gamma(\xi).$$

Note that $\tilde{c}$ is independent of ϵ. For the last term we use that $1 \wedge \exp(\cdot)$ has Lipschitz constant 1

$$\begin{aligned}\int_X |\alpha(x,p_x)(\xi) - \alpha(y,p_y)(\xi)| d\gamma(\xi) \\ \leq \int_{\mathcal{H}} |\Phi(p_x) - \Phi(p_y)| + |\Phi(x) - \Phi(y)| \, d\gamma(\xi) \\ \leq 2L \, |x-y| \leq 2L\epsilon d(x,y)\end{aligned} \tag{3.5}$$

which yields an overall contraction for ϵ small enough.

Similarly, we get for $x, y \in B_{\tilde{R}}(0)^c$ and $B = \{\omega | \sqrt{2\delta} \, \|\zeta\| \leq r(\|x\| \wedge \|y\|)\}$

$$\begin{aligned}d(\mathcal{P}(x,\cdot), \mathcal{P}(y,\cdot)) \leq & \mathbb{P}(B)[\mathbb{P}(\text{both accept}|B)(1-\rho) + \mathbb{P}(\text{both reject}|\text{B})] d(x,y) \\ & \mathbb{P}(B^c) d(x,y) + \int_{\mathcal{H}} |\alpha(x,p_x)(\xi) - \alpha(y,p_y)(\xi)| \, d\gamma(\xi).\end{aligned}$$

The lower bound for $\mathbb{P}(\text{both accept}|B)$ follows this time from Assumption 2.10.

All occurring ball probabilities are larger in the m-dimensional approximation due to Lemma 3.1 and the acceptance probability is larger since inf and sup are applied to smaller sets. Thus the contraction constant is uniform in m. □

3.1.3. *The d-Smallness.* The d-smallness of the level sets of V is achieved by replacing the Markov kernel by the n-step one. This preserves the d-contraction and the Lyapunov function. The variable n is chosen large enough so that if the algorithms started at x and y both accept n times in a row then d drops below $\frac{1}{2}$. Hence

$$d(\mathcal{P}^n(x,\cdot), \mathcal{P}^n(y,\cdot)) \leq 1 - \frac{1}{2}\mathbb{P}(\text{accept n-times}).$$

LEMMA 3.4. *If S is bounded, then there exists an n and $0 < s < 1$ such that for all $x, y \in S$, $m \in \mathbb{N}$ and for d as in (3.2)*

$$d\left(\mathcal{P}_m^n(x,\cdot), \mathcal{P}_m^n(y,\cdot)\right) \leq s \quad \text{and} \quad d\left(\mathcal{P}^n(x,\cdot), \mathcal{P}^n(y,\cdot)\right) \leq s \ .$$

PROOF. In order to obtain an upper bound for $d\left(\mathcal{P}^n(x,\cdot), \mathcal{P}^n(y,\cdot)\right)$, we choose the basic coupling (see Section 3.1.2) as before. Let R_S be such that $S \subset B_{R_S}(0)$ and B be the event that both instances of the algorithm accept n times in a row. In the event of B it follows by the definition of d (c.f. (3.2)) that

$$d(X_n, Y_n) \leq \frac{1}{\epsilon} \|X_n - Y_n\| \leq \frac{1}{\epsilon}(1-\rho)^n \|X_0 - Y_0\| \leq \frac{1}{\epsilon}(1-\rho)^n \text{diam } S \leq \frac{1}{2}$$

which implies that if X_0 and Y_0 are in S, then $d(X_n, Y_n) \leq \frac{1}{2}$. Hence

$$d(\mathcal{P}^n(x,\cdot), \mathcal{P}^n(y,\cdot)) \leq \mathbb{P}(B)\frac{1}{2} + (1 - \mathbb{P}(B)) \cdot 1 < 1.$$

Writing ξ^i for the noise in the i-th step, we bound

$$\begin{aligned}\mathbb{P}(B) &\geq \mathbb{P}\left(\left\|\sqrt{2\delta}\xi^i\right\| \leq \frac{R}{n} \text{ for } i = 1 \dots n\right) \mathbb{P}\Big(\text{both accept } n\text{times} \,\Big|\, \left\|\xi^i\right\| \leq \frac{R}{n}\Big) \\ &\geq \mathbb{P}\Big(\|\zeta\| \leq \frac{R}{n}\Big)^n \exp\Big(-\sup_{z \in B_{2R}(0)} \Phi(z) + \inf_{z \in B_{2R}(0)} \Phi(z)\Big)^n > 0,\end{aligned}$$

uniformly for all $X_0, Y_0 \in B_R(0)$. For the m-dimensional approximation the lower bound exceeds that in the infinite dimensional case due to Lemma 3.1 and the fact that

$$-\sup_{z\in B_{2R}(0)}\Phi(z)+\inf_{z\in B_{2R}(0)}\Phi(z)\le -\sup_{z\in B_{2R}(0)}\Phi(P_nz)+\inf_{z\in B_{2R}(0)}\Phi(P_nz).$$

Hence the claim follows. □

3.2. *Local log-Lipschitz density.* Now we allow the local Lipschitz constant

$$\phi(r)=\sup_{x\neq y\in B_r(0)}\frac{|\Phi(x)-\Phi(y)|}{\|x-y\|}$$

to grow in r. We used that Φ is globally Lipschitz to prove that $\mathcal{P}$ and $\mathcal{P}_m$ is d-contracting (c.f. Equation (3.5)). Now there is no one fixed ϵ that makes $\mathcal{P}$ d-contracting. Instead the idea is to change the metric in a way such that two points far out have to be closer in $\|\cdot\|_{\mathcal{H}}$ in order to be considered "close" i.e. $d(x,y)<1$. This is inspired by constructions in Hairer and Majda (2010); Hairer, Mattingly and Scheutzow (2011). Setting

$$\mathsf{A}(T,x,y):=\{\psi\in C^1([0,T],\mathcal{H}),\psi(0)=x,\psi(T)=y,\|\dot\psi\|=1\},$$

we define the two metrics d and $\bar d$ by

$$d(x,y)=1\wedge\bar d(x,y)\qquad \bar d(x,y)=\inf_{T,\psi\in\mathsf{A}(T,x,y)}\frac{1}{\epsilon}\int_0^T\exp(\eta\,\|\psi\|)dt \tag{3.6}$$

where ϵ and η will be chosen depending on Φ and γ in the subsequent proof. The situation is different from before because even in the case when "both accept" the distance can increase because of the weight. In order to control this, we notice that

LEMMA 3.5. *Let ψ be a path connecting x,y with $\|\dot\psi\|=1$, then for $\bar d$ as in (3.6)*

1. $\frac{1}{\epsilon}\int_0^T\exp(\eta\,\|\psi\|)dt<1$ *implies*
$$T\le J:=\epsilon\exp\left(-\eta(\|x\|\vee\|y\|-\epsilon)\vee 0\right)\le\epsilon.$$
2. $\bar d(x,y)\le\frac{\|x-y\|}{\epsilon}\exp\left(\eta(\|x\|\vee\|y\|)\right)$ *and*
$$\frac{\|x-y\|}{\epsilon}\exp\left(\eta(\|x\|\vee\|y\|-J)\vee 0\right)\le\bar d(x,y)$$
for all points such that $\bar d(x,y)<1$.
3. *For points such that $\bar d(x,y)<1$*
$$\frac{\bar d(p_x,p_y)}{\bar d(x,y)}\le(1-2\delta)^{\frac12}e^{-\eta\rho\left[\|x\|\vee\|y\|+\eta(\left\|\sqrt{2\delta}\xi\right\|+J)\right]}.$$

PROOF. In order to prove the first statement, we observe that

$$\epsilon\ge\int_0^T e^{\eta|\|x\|\vee\|y\|-t|}dt\ge Te^{\eta(\|x\|\vee\|y\|-T)\vee0)}\ge Te^{\eta(\|x\|\vee\|y\|-\epsilon)\vee0)}.$$

For the second part we denote by ψ the line segment connecting x and y in order to obtain an upper bound $d(x,y)$. For the lower bound we use $\|\psi\| \geq (\|x\| \vee \|y\| - J) \vee 0$ from the first part combined with the fact that $T \leq \epsilon$. Using the second part we get

$$\begin{aligned}\bar{d}(p_x,p_y) &\leq \frac{1}{\epsilon}(1-2\delta)^{\frac{1}{2}} \|x-y\| e^{\eta\left[(\|x\|\vee\|y\|)-\rho(\|x\|\vee\|y\|)+\sqrt{2\delta}\|\xi\|\right]} \\ &\leq (1-2\delta)^{\frac{1}{2}} e^{\eta\left[-\rho(\|x\|\vee\|y\|)+\sqrt{2\delta}\|\xi\|+J\right]} \frac{1}{\epsilon}\|x-y\| e^{\eta(\|x\|\vee\|y\|-J)} \\ &\leq (1-2\delta)^{\frac{1}{2}} e^{\eta\left[-\rho(\|x\|\vee\|y\|)+\sqrt{2\delta}\|\xi\|+J\right]} \bar{d}(x,y)\end{aligned}$$

which is precisely the required bound. □

3.2.1. *Lyapunov Functions.* This condition neither depends on the distance function d nor on the Lipschitz properties of Φ. Hence Lemma 3.2 applies.

3.2.2. *The d-Contraction.* The main difference between local and global Lipschitz Φ is proving that $\mathcal{P}$ and $\mathcal{P}_m$ is d-contracting.

LEMMA 3.6. *If Φ satisfies Assumption 2.10 and 2.13, then $\mathcal{P}$ and $\mathcal{P}_m$ are d-contracting for d as in (3.6) with a contraction constant uniform in m.*

PROOF. First suppose $x, y \in B_R(0)$ with $d(x,y) < 1$ and denote the event $A = \left\{\omega | \, \|\xi\| \leq \frac{2R}{\sqrt{2\delta}}\right\}$. First we choose R large, before dealing with the case when η is small and when ϵ is small. We have

$$\begin{aligned}d(\mathcal{P}(x,\cdot),\mathcal{P}(y,\cdot)) \leq\ & \mathbb{P}(A)\left[\mathbb{P}(\text{both accept}|A)(1-\tilde{\rho})d(x,y)\right. \\ & \left.+\left[\mathbb{P}(\text{both reject}|A)d(x,y)\right]\right. \\ & +\mathbb{E}((\alpha(x,p_x)\wedge\alpha(y,p_y))d(p_x,p_y);A^c) \qquad (3.7) \\ & +\mathbb{E}((1-\alpha(x,p_x)\vee\alpha(y,p_y))d(x,y);A^c) \qquad (3.8) \\ & +\mathbb{P}(\text{only one accepts})\cdot 1\end{aligned}$$

where the first two lines deal with both accept and both reject in the case of A, the third and fourth line consider the same case in the event of A^c. The last line deals with the case when only one accepts. For the first two lines of Equation (3.7) we argue that

$$\mathbb{P}(\text{both accept}|A) \geq \inf_{x,z\in B_{3R}(0)} \mathbb{P}(\text{accepts}|p_x = z) = \exp(-\Phi^+(3R)+\Phi^-(3R)).$$

If both are accepted, we know from Lemma 3.5 that

$$\begin{aligned}\frac{\bar{d}(p_x,p_y)}{\bar{d}(x,y)} &\leq (1-2\delta)^{\frac{1}{2}} \exp\left(-\eta\rho\left(\|x\|\vee\|y\|\right)+\eta(\left\|\sqrt{2\delta}\xi\right\|+J)\right) \\ &\leq (1-2\delta)^{\frac{1}{2}} e^{\eta(3R+J)} \leq (1-\tilde{\rho})\end{aligned}$$

where the last step follows for η small enough. Using the complementary probability, we obtain the following estimate

$$\mathbb{P}(\text{both reject}|A) \leq 1 - \mathbb{P}(\text{both accept}|A)\ .$$

Combining both estimates, it follows that $\mathbb{P}(A)\,(1-\mathbb{P}(\text{both accept}|A)(1-\tilde{\rho}))$ as coefficient in front of $d(x,y)$. In order to show that $\mathcal{P}$ is d-contracting, we have to prove that the expression in the

third and fourth line of Equation (3.7) is close to $\mathbb{P}(A^c)\cdot d(x,y)$. We notice that

$$\begin{aligned}
&\mathbb{E}\left((1-\alpha(x,p_x)\vee\alpha(y,p_y))d(x,y);A^c\right)+\mathbb{E}\left((\alpha(x,p_x)\wedge\alpha(y,p_y))d(p_x,p_y);A^c\right)\\
&\quad\le \mathbb{E}\left(d(p_x,p_y)\vee d(x,y);A^c\right)\le \bar d(x,y)\mathbb{E}\frac{\bar d(p_x,p_y)}{\bar d(x,y)}\vee 1\\
&\quad\le d(x,y)\int\limits_{\sqrt{2\delta}\|\xi\|>2R} 1\vee e^{\eta(\sqrt{2\delta}\|\xi\|+J)}d\gamma(\xi),
\end{aligned}$$

where the last step followed by Lemma 3.5. For small η the above is arbitrarily close to $\mathbb{P}(A^c)\cdot d(x,y)$ by the Dominated Convergence theorem. By writing the integrand as $\chi_{\sqrt{2\delta}\|\xi\|>2R}\left(1\vee\exp(\eta(\sqrt{2\delta}\,\|\xi\|+J)\right)$ and applying Lemma 3.1, we conclude that this estimate holds uniformly in m. Combining the first four lines, the coefficient in front of $d(x,y)$ is less than 1 independently of ϵ. Only $\mathbb{P}(\text{only one accepts})\cdot 1$ is left to bound in terms of $d(x,y)$:

$$\begin{aligned}
\mathbb{P}(\text{only one accepts}) &= \int|\alpha(x,p_x)-\alpha(y,p_y)|\,d\gamma(\xi)\\
&\le \int(|\Phi(p_x)-\Phi(p_y)|+|\Phi(x)-\Phi(y)|)d\gamma(\xi)\\
&\le \epsilon d(x,y)\int(\phi((1-\rho)R+\sqrt{2\delta}\,\|\xi\|)+\phi(R))d\gamma(\xi).
\end{aligned}$$

The integral above is bounded by Fernique's theorem. Hence for ϵ small enough, we get an overall contraction when we combine this with the result above.

Now let $x,y\in B_R^c(0)$ with $d(x,y)<1$ and without loss of generality we assume that $\|y\|\ge\|x\|$. Similar to the first case we bound with $A=\{\omega|\|\sqrt{2\delta}\zeta\|\le r(\|x\|)\}$, we have

$$\begin{aligned}
d(\mathcal{P}(x,\cdot),\mathcal{P}(y,\cdot)) \le\ & \mathbb{P}(A)\left[\mathbb{P}(\text{both accept}|A)(1-\rho)d(x,y)+\right.\\
&\left.\mathbb{P}(\text{both reject}|A)d(x,y)\right]+\mathbb{E}\left(d(x,y)\vee d(p_x,p_y);A^c\right)\\
&+\mathbb{P}(\text{only one accepts})\cdot 1.
\end{aligned}$$

If "both accept", then the contraction factor associated to the event of A is smaller than $(1-\rho)$ because $r(\|x\|)\le\frac{\rho}{2}\|x\|$ and by an application of Lemma 3.5. For the next term it follows that

$$\begin{aligned}
\mathbb{E}\left(d(p_x,p_y)\vee d(x,y);A^c\right) &\le \bar d(x,y)\mathbb{E}\frac{\bar d(p_x,p_y)}{\bar d(x,y)}\vee 1\\
&\le \bar d(x,y)\int\limits_{A^c} 1\vee e^{-\rho\eta(\|y\|)+\eta(\left\|\sqrt{2\delta}\xi\right\|+J)}d\gamma(\xi).
\end{aligned}$$

Denoting the integral above by I, its integrand by $f(\xi)$ and $F>0$, this yields

$$I\le I_1+I_2=\int\limits_{\rho(\|y\|-J)+F\ge\left\|\sqrt{2\delta}\xi\right\|\ge r(\|x\|\wedge\|y\|)} f(\xi)d\gamma(\xi)\quad+\int\limits_{\left\|\sqrt{2\delta}\xi\right\|\ge\rho(\|y\|-J)+F} f(\xi)d\gamma(\xi)\quad.$$

For the first part we have the upper bound $\mathbb{P}(A^c)e^{\sqrt{2\delta}\eta F}$ and for the second part we take $g\in X^\star$ with $\|g\|=1$. We note that $\{x|g(x)>R\}\subseteq B_R(0)^c$ and hence

$$\gamma(B_R(0)^c)\ge\gamma(\{x|g(x)>R\})\ge\exp(-\tilde\beta R^2+\zeta)$$

using the one dimensional lower bound. For the uniformity in m we choose $g = e_1^\star$. We incorporate all occurring constants into ζ and use Proposition A.1 to bound

$$\begin{aligned} I_2 &\leq \mathbb{P}(A^c)\exp\Big(\tilde{\beta}\frac{r(\|x\|)^2}{2\delta} - \rho\eta(\|y\| - J) \\ &\qquad \eta\sqrt{2\delta}(\rho(\|y\| - J) + F) - \beta\sqrt{2\delta}(\rho(\|y\| - J) + F)^2 + \zeta\Big). \end{aligned}$$

For any $\tau > 0$ we choose F large enough and then η small enough so that $I \leq (1+\tau)\mathbb{P}(A^c)d(x,y)$. Again the estimates above are independent of ϵ which we choose small in order to bound $\mathbb{P}(\text{only one accepts}|A^c)$ in terms of $d(x,y)$. Calculating as above yields

$$\begin{aligned} &\int |\alpha(x,p_x) - \alpha(y,p_y)|\, d\gamma(\xi) \\ \leq\ &\int |\Phi(x) - \Phi(y)| + |\Phi(p_x) - \Phi(p_y)|\, d\gamma(\xi) \\ \leq\ &\int (\phi(\|y\|) + \phi(\|p_x\| \vee \|p_y\|)\, d\gamma(\xi)\, \|x - y\| \\ \leq\ &\left(M_\kappa e^{\kappa\|y\|} + \int \phi((1-\rho)\|y\| + \sqrt{2\delta}\,\|\xi\|)d\gamma(\xi)\right)\|x - y\| \\ \leq\ &CM_\kappa\epsilon e^{-\eta(\|x\|\vee\|y\| - \epsilon)\vee 0 + \kappa\|y\|}\bar{d}(x,y) \end{aligned}$$

where the last step follows using the upper bound for $\|x - y\|$ from Lemma 3.5. Choosing $\kappa = \frac{\eta}{2}$ and ϵ small enough, we can guarantee a uniform contraction. Checking line by line, the same is true for the m-dimensional approximation. □

3.2.3. *The d-Smallness.* Analogous to the globally Lipschitz case, we have

LEMMA 3.7. *If S is bounded, then $\exists n \in \mathbb{N}$ and $0 < s < 1$ such that for all $x, y \in S$, $m \in \mathbb{N}$ and for d as in (3.6)*

$$d(\mathcal{P}_m^n(x,\cdot), \mathcal{P}_m^n(y,\cdot)) \leq s \quad \text{and} \quad d(\mathcal{P}^n(x,\cdot), \mathcal{P}^n(y,\cdot)) \leq s\ .$$

PROOF. By Lemma 3.4 d and $\|\cdot\|$ are comparable on bounded sets. If $X_0, Y_0 \in B_R(0)$ and both algorithms accept n proposals in a row which are all elements of $B_{2R}(0)$, then for n large enough

$$d(X_n, Y_n) \leq \frac{\exp(\eta(2R + J))}{\epsilon}\text{diam}(S)(1 - 2\delta)^{n/2} \leq \frac{1}{2}.$$

Hence the result follows analogue to Lemma 3.4. □

4. Results Concerning the Sample-Path Average. In this section we focus on sample path properties of the pCN algorithm which can be derived from the Wasserstein and the L^2_μ-spectral gap. We prove a strong law of large numbers, a CLT and a bound on the MSE. This allows us to quantify the approximation of $\mu(f)$ by

$$S_{n,n_0}(f) = \frac{1}{n}\sum_{i=1}^{n} f(X_{i+n_0}).$$

4.1. *Consequences of the Wasserstein Spectral Gap.* The immediate consequences of a Wasserstein spectral gap are weaker than the results from the L^2-spectral gap because they apply to a smaller class of observables but they hold for the algorithm started at any deterministic point.

4.1.1. *Change to a Proper Metric and Implications for Lipschitz Functionals.* For the Wasserstein CLT Komorowski and Walczuk (2012) we need a Wasserstein spectral gap with respect to a metric. The reason for this is that the Monge-Kantorovich duality is used for its proof Komorowski and Walczuk (2012). Recall that Theorem 2.14 yields a Wasserstein spectral gap for the distance

$$\tilde{d} = \sqrt{(1 + \|x\|^i + \|y\|^i)(1 \wedge d)} \qquad \text{where}$$
$$d = \inf_{T,\psi \in \mathsf{A}(T,x,y)} \frac{1}{\epsilon} \int_0^T \exp(\eta \|\psi\|).$$

Because $\tilde{d}$ does not necessarily satisfy the triangle inequality, we introduce

$$d'(x,y) = \sqrt{\inf_{\substack{x = z_1, \dots, z_n = y \\ n \geq 2,}} \sum_{j=1}^{n-1} d_0(z_j, z_{j+1})} \tag{4.1}$$
$$d_0(x,y) = d_1(x,y) \wedge d_2(x,y)$$
$$d_1(x,y) = \begin{cases} 0 & x = y \\ (1 + \|x\|^i + \|y\|^i) & \text{otherwise} \end{cases}$$
$$d_2(x,y) = \inf_{T,\psi \in \mathsf{A}(T,x,y)} F(\psi)$$
$$F(\psi) = \frac{1}{\epsilon} \int_0^T \exp(\eta \|\psi\|)(1 + \|\psi\|^i) dt.$$

It is straightforward to verify that d' is a metric by first showing that the expression inside the square root is a metric (triangle inequality is satisfied because of the infimum) and using that a square root of a metric is again a metric.

Moreover, $\mathcal{P}$ and $\mathcal{P}^m$ have a Wasserstein spectral gap with respect to d' because of the following lemma

LEMMA 4.1. *Provided that ϵ is small enough, there exists a constant $C > 0$ such that*

$$d'(x,y) \leq \tilde{d}(x,y) \leq C d'(x,y)$$

for all pairs of points x, y in $\mathcal{H}$.

PROOF. Without loss of generality we assume that $\|y\| \geq \|x\|$. The inequality $d' \leq \tilde{d}$ follows from Lemma 4.2 since $d' \leq \sqrt{d_0}$ by definition.

In order show that $\tilde{d} \leq C d'$, we will use Lemma 4.2 and reduce the number of summands appearing in Equation (4.1) for d'. We can certainly assume that there is at most one index j in (4.1) such that $d_0(z_j, z_{j+1}) = d_1(z_j, z_{j+1})$ because otherwise there are $1 \leq j < k \leq n$ such that

$$d_0(z_j, z_{j+1}) = d_1(z_j, z_{j+1}), \quad d_0(z_k, z_{k+1}) = d_1(z_k, z_{k+1})$$

which would lead to

$$d_0(z_j, z_{j+1}) + \dots + d_0(z_k, z_{k+1}) \geq 2 + \|z_j\|^i + \|z_{k+1}\|^i > d_1(z_j, z_{k+1}).$$

Hence the expression could be made smaller by removing all intermediate points between z_j and z_{k+1}, contradiction.

Because d_2 is a Riemannian metric, it satisfies the triangle inequality in a sharp way in the sense that $d_2(x,y) = \inf_z \big(d_2(x,z) + d_2(z,y)\big)$. As a consequence, the infimum is not changed by assuming that in Equation (4.1) there is no index j such that

$$d_0(z_j, z_{j+1}) = d_2(z_j, z_{j+1}), \quad d_0(z_{j+1}, z_{j+2}) = d_2(z_{j+1}, z_{j+2}).$$

Combining these two facts, Equation (4.1) thus reduces to

$$\begin{aligned}(d'(x,y))^2 = \min \Big\{ & d_0(x,y), \inf_{z_2,z_3} d_2(x,z_2) + d_1(z_2,z_3) + d_2(z_3,y) \\ & \inf_{z_2} d_2(x,z_2) + d_1(z_2,y), \inf_{z_2} d_1(x,z_2) + d_2(z_2,y) \Big\}. \end{aligned} \tag{4.2}$$

Recalling Lemma 4.2, it remains to show that $d' \geq C\sqrt{d_0}$ with d' given by (4.2). This is of course non-trivial only if (x,y) is such that $d'(x,y) < \sqrt{d_0(x,y)}$. Therefore we assume this fact from now on.

Suppose first that $\|y\| \leq Q$, for some constant $Q > 0$ to be determined later. Since $d'(x,y) \neq \sqrt{d_0(x,y)}$, there is at least one j such that $d_0(z_j, z_{j+1}) = d_1(z_j, z_{j+1})$ which leads to

$$1 + 2Q^i \geq d_0(x,y) \geq \big(d'(x,y)\big)^2 \geq 1 \,,$$

so that the bound $(1 + 2Q^i) d'(x,y) \geq \sqrt{d_0(x,y)}$ indeed follows in this case.

Suppose now that $\|y\| \geq Q$. Again, one summand $d_0(z_j, z_{j+1})$ in Equation (4.2) satisfies

$$d_0(z_j, z_{j+1}) = d_1(z_j, z_{j+1}) \,,$$

thus giving rise to a simple lower bound on d':

$$d'(x,y) \geq \sqrt{1 + \|z_j\|^i} \,. \tag{4.3}$$

Because of (4.2), z_{j+1} is either equal to y or connected to y through a path $\psi_y \in \mathsf{A}(T, z_{j+1}, y)$ which is such that

$$F(\psi_y) \leq 1 + 2\,\|y\|^i \tag{4.4}$$

where $F(\psi)$ is as in the definition of d_2. By the same reasoning as in the proof of Lemma 4.2, for Q large enough it is sufficient to consider paths starting in y and such that $\|\psi(t)\| \geq \|y\|/2$. The bound (4.4) thus yields an upper bound on $\|z_{j+1} - y\|$ by

$$1 + 2\,\|y\|^i \geq F(\psi_y) \geq \frac{1}{\epsilon} \|z_{j+1} - y\| \, \exp(\eta\,\|y\|/2) \,. \tag{4.5}$$

Combining this with (4.3), we have

$$\begin{aligned} d'(x,y)^2 &\geq 1 + (\|y\| - \|z_{j+1} - y\|)^i \geq 1 + \|y\|^i - i\,\|y\|^{i-1}\,\|z_{j+1} - y\| \\ &\geq 1 + \frac{\|y\|^i}{2} + \left(\frac{\|y\|^i}{2} - \epsilon(1 + 2\,\|y\|^i) \exp(-\eta\,\|y\|/2) \right) \,, \end{aligned}$$

Provided that $\epsilon < 1/4$ and Q is large enough, the third summand is positive so that $d'(x,y)^2 \geq \frac{1}{4} d_1(x,y) \geq \frac{1}{4} d_0(x,y)$ concluding the proof. □

LEMMA 4.2. *There is a $C > 0$ such that d_0 as defined in Equation (4.1) satisfies*

$$d_0(x,y) \le \tilde{d}(x,y)^2 \le C d_0(x,y) \text{ for all } x, y.$$

PROOF. We assume again that $\|y\| \ge \|x\|$. In order to prove that $d_0(x,y) \le \tilde{d}(x,y)^2$, we only have to show that

$$\inf_{T,\psi\in\mathsf{A}(T,x,y)} F(\psi) \le \inf_{T,\psi\in\mathsf{A}(T,x,y)} \frac{1}{\epsilon}\int_0^T \exp(\eta\,\|\psi\|)dt(1+\|x\|^i+\|y\|^i)\,.$$

Replacing $\psi(t)$ by

$$(1 \wedge \|y\|/\|\psi(t)\|)\,\psi(t)\,,$$

in the expressions above does not cause an increase. Hence it is sufficient to consider paths ψ which satisfy

$$\|\psi(t)\| \le \|y\|\,, \quad t \in [0,T]. \tag{4.6}$$

The bound $d_0 \le \tilde{d}^2$ then follows at once from

$$1 + \|\psi\|^i \le 1 + \|x\|^i + \|y\|^i\,.$$

We proceed now to show that $\tilde{d}(x,y)^2 \le C d_0(x,y)$ for which we only have to consider

$$d_2(x,y) = \inf_{T,\psi\in\mathsf{A}(T,x,y)} \frac{1}{\epsilon}\int_0^T \exp(\eta\,\|\psi\|)(1+\|\psi\|^i)dt \le (1+\|x\|^i+\|y\|^i) \tag{4.7}$$

since the minimum expressions in $\tilde{d}^2$ and d_0 have $(1+\|x\|^i+\|y\|^i)$ in common.

We will first use this to show that x and y have to be close if $\|y\|$ is large. We will show that any path ψ for which the expression in d_2 is close to the infimum has to satisfy $\|y\| \ge \psi \ge \frac{\|y\|}{2}$. Hence $1+\|\psi\|^i$ and $(1+\|x\|^i+\|y\|^i)$ are comparable. In order to gain a lower bound on $d_2(x,y)$, we distinguish between paths ψ which intersects or do not intersect $B_R(0)$. If the path lies completely outside the ball, we have

$$d_2(x,y) \ge \frac{1}{\epsilon}\,\|x-y\|\exp(\eta R)(1+R^i).$$

If ψ and $B_R(0)$ intersect, then $d_2(x,y)$ is larger than $d_2(B_R(0),y)$ which by symmetry corresponds to

$$\begin{aligned} d_2(x,y) &\ge \frac{1}{\epsilon}\int_0^{\|y\|-R} \exp(\eta(\|y\|-t))(1+(\|y\|-t)^i)dt \\ &\ge \frac{1}{\epsilon}(\|y\|-R)\exp(\eta(\|y\|-R)dt(1+(\|y\|-R)^i). \end{aligned}$$

We choose $R = \frac{\|y\|}{2}$ and note that $\frac{\|y\|}{2} \ge \frac{\|x-y\|}{4}$, leading in both cases to

$$d_2(x,y) \ge \frac{1}{4\epsilon}\,\|x-y\|\exp(\eta\,\|y\|/2)(1+\frac{\|y\|^i}{2}).$$

By (4.7) this implies

$$\|x-y\| \le \frac{4\epsilon\exp(-\eta\frac{\|y\|}{2})}{1+\|y\|^i/2}\left(1+2\,\|y\|^i\right). \tag{4.8}$$

For x and y in $B_Q(0)$ we have that $(\tilde{d})^2 \leq (2Q^i+1)d_0$ because we can assume $\|\psi(t)\| \leq \|y\|$ as above . It is only left to consider $x, y \in B_{\tilde{Q}}(0)^c$ for $\tilde{Q} = Q - 4\epsilon\exp(-\eta\frac{Q}{2})(1+2Q^i)$ because of Equation (4.8). Subsequently, we will show that for Q and hence $\tilde{Q}$ large enough, it is sufficient for the infimum expression for d_2 to consider paths ψ that do not intersect $B_R(0)$ for $R = \frac{\|y\|}{2}$.

Suppose that the path ψ would intersect $B_R(0)$. Then the functional is larger than the shortest path $\hat{\psi}$ to the boundary of the ball and hence

$$\begin{aligned} d_2 &\geq F(\hat{\psi}) \geq \frac{1}{\epsilon}\int_0^{\|y\|-R} e^{\eta(\|y\|-t)}(1+(\|y\|-t)^i)dt \\ &= \frac{1}{\epsilon}\Bigg[\exp(\eta\,\|y\|)(\eta^{-1}(1+\|y\|^i)+\sum_{j=1}^{i}\eta^{-1-j}\frac{i!}{(i-j)!}\,\|y\|^{i-j}) \\ &\qquad -\exp(\eta R)(\eta^{-1}(1+R^i)+\sum_{j=1}^{i}\eta^{-1-j}\frac{i!}{(i-j)!}R^{i-j})\Bigg] \end{aligned} \tag{4.9}$$

by $i+1$ integrations by parts. Let l be the line connecting x and y, then using (4.8) yields

$$F(l) \leq \frac{1}{\epsilon}\,\|x-y\|\,e^{\eta\|y\|}(1+\|y\|^i) \leq 4\exp(\eta\frac{\|y\|}{2})(1+2\,\|y\|^i)^2.$$

For $R = \frac{\|y\|}{2}$ and $\tilde{Q}$ large enough we have $F(\psi) > F(l)$. Therefore for all $t \in [0,T]$ $\|y\| \geq \psi \geq \|y\|/2$ and thus

$$2^{i+1}(1+\|\psi\|^i) \geq (1+\|x\|^i+\|y\|^i)$$

which yields that $\max(2Q^i+1, 2^{i+1})d_0 \geq \tilde{d}^2$. □

4.1.2. *Strong Law of Large Numbers.* In this section we will prove a strong law of large numbers for Lipschitz functions. Since μ_m (μ) are the unique invariant measures for $\mathcal{P}$ ($\mathcal{P}_m$) (respectively), μ_m (μ) is ergodic and Birkhoff's ergodic theorem applies. However, this theorem only applies to almost every initial condition but we are able to extend it to every initial condition in this case which yields a strong law of large numbers.

THEOREM 4.3. *In the setting of Theorem 2.12 or 2.14, suppose that supp $\mu = \mathcal{H}$ and $h : \mathcal{H} \to \mathbb{R}$ has Lipschitz constant L with respect to $\tilde{d}$, then for arbitrary $X_0 \in \mathcal{H}$*

$$\left|\frac{1}{n}\sum_{i=1}^{n}h(X^i) - \mathbb{E}_\mu h\right| \overset{a.s}{\to} 0.$$

PROOF. By Birkhoff's ergodic theorem, we know that this is true for measurable h and every initial condition in some set of full measure A. Because μ has full support, for any $t > 0$ we can choose $Y_0 \in A$ with $\tilde{d}(X_0, Y_0) \leq t^2$. Hence

$$\begin{aligned} \left|\frac{1}{n}\sum_{i=1}^{n}h(X^i) - \mathbb{E}_\mu h\right| &\leq \left|\frac{1}{n}\sum_{i=1}^{n}h(Y^i) - \mathbb{E}_\mu h\right| + \left|\frac{1}{n}\sum_{i=1}^{n}(h(X^i)-h(Y^i))\right| \\ &\leq \left|\frac{1}{n}\sum_{i=1}^{n}h(Y^i) - \mathbb{E}_\mu h\right| + \frac{1}{n}\sum_{i=1}^{n}L\tilde{d}(X^i, Y^i). \end{aligned}$$

By the Wasserstein spectral gap, we can couple X^n and Y^n such that

$$\mathbb{E}\tilde{d}(X^n, Y^n) \leq Cr^n \tilde{d}(X^0, Y^0)$$

for some $0 < r < 1$. An application of Markov's inequality yields

$$\mathbb{P}\left(\tilde{d}(X^n, Y^n) \geq c\right) \leq C\frac{r^n \tilde{d}(X^0, Y^0)}{c}.$$

Since Birkhoff's theorem applies to the Markov process started at Y_0, we have

$$\begin{aligned}\mathbb{P}\left[\limsup \left|\frac{1}{n}\sum_{i=1}^{n} h(X^i - \mathbb{E}_\mu h)\right| \geq c\right] &= \mathbb{P}\left[\limsup \frac{1}{n}\sum_{i=1}^{n} \left|h(X^i) - h(Y^i)\right| \geq c\right] \\ &\leq C\frac{L}{c(1-r)}\tilde{d}(X^0, Y^0).\end{aligned}$$

Setting $c = \frac{t}{L}$ yields

$$\mathbb{P}\left(\limsup \left|\frac{1}{n}\sum_{i=1}^{n} h(X^i - \mathbb{E}_\mu h)\right| \leq t\right) \geq 1 - t\frac{C}{1-r}$$

and because t was chosen arbitrarily, the result follows. □

4.1.3. *Central Limit Theorem.* The result above does not give any rate of convergence. With a CLT on the other hand it is possible to derive (asymptotic) confidence intervals and to estimate the error for finite n. Because of Lemma 4.1 and arguments from Lemma 3.2, it is straightforward to verify that our assumptions imply those needed for the Wasserstein CLT in Komorowski and Walczuk (2012). This results in the following theorem

THEOREM 4.4. *If the conditions of Theorem 2.12 or 2.14 are satisfied, then there exists* $\sigma \in [0, +\infty)$ *such that*

$$\lim_{n\to+\infty} \frac{1}{n}\mathbb{E}\left(\sum_{i=1}^{n} \tilde{f}(X_s)\right)^2 = \sigma^2$$

where $\tilde{f} := f - \mu(f)$ *and* f *is Lipschitz with respect to* d'*. Moreover, we have*

$$\lim_{T\to\infty} \mathbb{P}(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \tilde{f}(X_s) < \xi) = \Phi_\sigma(\xi), \quad \forall \xi \in \mathbb{R}$$

where $\Phi_\sigma(\cdot)$ *is the distribution function of* $\mathcal{N}(0, \sigma^2)$ *a zero mean normal law whose variance equals* σ^2*.*

4.2. *Consequences of* L^2_μ*-Spectral Gap.* Under the assumptions of Theorem (2.12) or (2.14), we have proved the existence of an L^2_μ-spectral gap in Section 2.2.2. Now we may use all existing consequences for the ergodic average with and without burn in ($n_0 = 0$)

$$S_{n,n_0}(f) = \frac{1}{n}\sum_{j=1}^{n} f(X_{j+n_0}) \qquad S_n = S_{n,0}.$$

The following result of Kipnis and Varadhan (1986) (see also Łatuszyński and Roberts (2011) whence the statement was adapted) then yields a CLT:

PROPOSITION 4.5. *Consider an ergodic Markov chain with transition operator P which is reversible with respect to a probability measure μ. Let $f \in L^2$ be such that*

$$\sigma_{f,P}^2 = \left\langle \frac{1+P}{1-P} f, f \right\rangle < \infty,$$

then for $X_0 \sim \mu$ the expression $\sqrt{n}(S_n - \mu(f))$ converges weakly to $\mathcal{N}(0, \sigma_{f,P}^2)$.

In our case, provided that f is mean-zero, it follows from the L^2-spectral gap that

$$\sigma_{f,\mathcal{P}}^2 \leq \frac{2\mu(f^2)}{1-\beta}.$$

Due to Theorem 2.14, we have a lower bound on the spectral gap $1-\beta$ of $\mathcal{P}$ and $1-\beta_m$ of $\mathcal{P}_m$ which is uniform in m. Thus the ergodic average of the pCN algorithm applied to the target measures μ and μ_m has an m-independent upper bound on the asymptotic variance.

The result of Proposition 4.5 has been extended to μ for almost every initial condition in Cuny and Lin (2009) which also applies to our case.

A different approach due to Rudolf (2012) is to consider the MSE

$$e_\nu(S_{n,n_0}, f) = (\mathbb{E}_{\nu,K} \left| S_{n,n_0}(f) - \mu(f)^2 \right|)^{1/2}.$$

Using Tschebyscheff's inequality, this results in a confidence interval for $S(f)$. We can bound it by using the following proposition from Rudolf (2012):

PROPOSITION 4.6. *Suppose that we have a Markov chain with Markov operator $\mathcal{P}$ which has an L^2_μ-spectral gap $1-\beta$. For $p \in (2, \infty]$ let $n_0(p)$ be the smallest natural number which is greater or equal to*

$$\frac{1}{\log(\beta^{-1})} \begin{cases} \frac{p}{2(p-2)} \log(\frac{32p}{p-2}) \left\| \frac{d\nu}{d\mu} - 1 \right\|_{\frac{p}{p-2}} & p \in (2,4) \\ \log(64) \left\| \frac{d\nu}{d\mu} - 1 \right\|_{\frac{p}{p-2}} & p \in [4, \infty]. \end{cases} \tag{4.10}$$

Then

$$\sup_{\|f\|_p \leq 1} e_\nu(S_{n,n_0}, f) \leq \frac{2}{n(1-\beta)} + \frac{2}{n^2(1-\beta)^2}.$$

In our setting $n_0(p)$ is finite for $\nu = \gamma$ under the additional assumption that for all $u_1 > 0$ there is a u_2 such that

$$\Phi(\|x\|) \leq u_1 \|x\|^2 + u_2.$$

Using Fernique's theorem, this implies that $\frac{d\gamma}{d\mu} - 1$ has moments of all orders.

5. Conclusion. From an applications perspective, the primary thrust of this paper is to develop an understanding of MCMC methods in high dimension. Our work has concentrated on identifying the (possibly lack of) dimension dependence of spectral gaps for the standard random walk method RWM and a recently developed variant pCN adapted to measures defined via a density with respect to a Gaussian. It is also possible to show that the function space version of the MALA Beskos, Kalogeropoulos and Pazos (2013) has a spectral gap if, in addition to the assumptions in this article, the gradient of Φ satisfies strong assumptions and the gradient step is very small. There is

also a variant of the Hybrid Monte-Carlo methods Beskos et al. (2011) adapted to the sampling of measures defined via a density with respect to a Gaussian and it would be interesting to employ the weak Harris theory to study this algorithm.

Other classes of target measures, such as those arising from Besov prior measures Lassas, Saksman and Siltanen (2009); Dashti, Harris and Stuart (2012) or an infinite product of uniform measures in Schwab and Stuart (2012) would also provide interesting applications. The proposal of the pCN is reversible and has a spectral gap with respect to the Gaussian reference measure. For arbitrary reference and target measures, the third author has recently proved that for bounded Φ the Metropolis-Hastings algorithm has a spectral gap if the proposal is reversible and has a spectral gap with respect to the reference measure Vollmer (2013).

More generally, we expect that the weak Harris theory will be well-suited to the study of many MCMC methods in high dimensions because of its roots in the study of Markov processes in infinite dimensional spaces Hairer, Mattingly and Scheutzow (2011). In contrast, the theory developed in Meyn and Tweedie (2009) does not work well for the kind of high dimensional problems that are studied here.

From a methodological perspective, we have demonstrated a particular application of the theory developed in Hairer, Mattingly and Scheutzow (2011), demonstrating its versatility for the analysis of rates of convergence in Markov chains. We have also shown how that theory, whose cornerstone is a Wasserstein spectral gap, may usefully be extended to study L^2-spectral gaps and resulting sample path properties. These observations will be useful in a variety of applications not just those arising in the study of MCMC.

All our results were presented for separable Hilbert spaces but in fact they do also hold on an arbitrary Banach space. This can be shown by using a Gaussian series (c.f. Section 3.5 in Bogachev (1998)) instead of the Karhunen-Loeve expansion and the m-independence is due to Theorem 3.3.6 in Bogachev (1998).

APPENDIX A: GAUSSIAN MEASURES

As a consequence of Fernique's Theorem, we have the following explicit bound on exponential moments of the norm of a Gaussian random variable, which is needed to show that $\mathcal{P}$ and $\mathcal{P}_m$ are d-contracting (see Section 3.2.2).

PROPOSITION A.1. *For β small enough, there exists a constant F_β such that*

$$\int_X \exp(\beta \|u\|^2) d\gamma(u) = F_\beta \,.$$

Furthermore, for any $\alpha \in \mathbb{R}^+$ there is a constant $C_{\alpha,\beta}$ such that for $K > \frac{\alpha}{2\beta}$

$$\int_{\{\|u\| \geq K\}} \exp(\alpha \|u\|) d\gamma(u) \leq C_{\alpha,\beta} e^{-\beta K^2 + \alpha K}.$$

PROOF. The first claim is just Fernique's theorem, see for example Bogachev (1998); Da Prato and Zabczyk (1992); Hairer (2010). Using integration by parts and Fubini, we get

$$\int_{\|x\| \geq K} f(\|x\|) d\gamma = f(K)\gamma(\|x\| \geq K) + \int_K^\infty \gamma(\|x\| \geq t) f'(t)\, dt.$$

Setting $f(x) = \exp(\alpha x)$ and applying Fernique's Theorem yields

$$\int_{\|x\| \geq K} \exp(\alpha \|x\|) d\gamma \leq F_\beta \exp(-\beta K^2 + \alpha K) + F_\beta \alpha \int_K^\infty \exp(-\beta t^2 + \alpha t)\, dt \,.$$

Since, for K as in the statement, one verifies that

$$\beta t^2 - \alpha t \geq \beta K^2 - \alpha K + \beta (t - K)^2 \,,$$

the required bound follows at once. □

REFERENCES

Adler, R. J. (1990). *An introduction to continuity, extrema, and related topics for general Gaussian processes. Institute of Mathematical Statistics Lecture Notes—Monograph Series, 12.* Institute of Mathematical Statistics, Hayward, CA. MR1088478 (92g:60053)

Athreya, K. B. and Ney, P. (1978). A new approach to the limit theory of recurrent Markov chains. *Trans. Amer. Math. Soc.* **245** 493–501. MR511425 (80i:60092)

Bakry, D. and Émery, M. (1985). Diffusions hypercontractives. In *Séminaire de probabilités, XIX, 1983/84. Lecture Notes in Math.* **1123** 177–206. Springer, Berlin. MR889476 (88j:60131)

Beskos, A., Kalogeropoulos, K. and Pazos, E. (2013). Advanced MCMC Methods for Sampling on Diffusion Pathspace. *Stochastic Process. Appl.* **123** 1415-1453. MR3016228

Beskos, A., Roberts, G. O. and Stuart, A. M. (2009a). Optimal Scalings for Local Metropolis-Hastings Chains on Nonproduct Targets in High Dimensions. *Ann. Appl. Probab.* **19** 863–898. MR2537193 (2010j:60183)

Beskos, A., Roberts, G. O. and Stuart, A. M. (2009b). Optimal Scalings for Local Metropolis-Hastings Chains on Nonproduct Targets in High Dimensions. *Ann. Appl. Probab.* **19** 863-898.

Beskos, A., Roberts, G. O., Stuart, A. M. and Voss, J. (2008). MCMC Methods for Diffusion Bridges. *Stoch. Dyn.* **8** 319–350.

Beskos, A., Pinski, F., Sanz-Serna, J. M. and Stuart, A. M. (2011). Hybrid Monte-Carlo on Hilbert Spaces. *Stoch. Proc. Appl.*

Bogachev, V. I. (1998). *Gaussian Measures. Mathematical Surveys and Monographs* **62**. Amer. Math. Soc., Providence, RI. MR1642391 (2000a:60004)

Bogachev, V. I. (2007). *Measure Theory. Vol. I, II.* Springer-Verlag, Berlin. MR2267655 (2008g:28002)

Chan, K. S. and Geyer, C. J. (1994). Discussion: Markov chains for exploring posterior distributions. *The Annals of Statistics* 1747–1758.

Cheeger, J. (1970). A Lower Bound for the Smallest Eigenvalue of the Laplacian. In *Problems in analysis* 195–199. Princeton Univ. Press, Princeton, N. J. MR0402831 (53 ##6645)

Cotter, S. L., Roberts, G. O., Stuart, A. M. and White, D. (2011). MCMC Methods for Functions: Modifying Old Algorithms to Make Them Faster. *ArXiv preprint 1202.0709.* to appear Stat. Sci.

Cuny, C. and Lin, M. (2009). Pointwise ergodic theorems with rate and application to the CLT for Markov chains. In *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques* **45** 710–733. Institut Henri Poincaré.

Da Prato, G. and Zabczyk, J. (1992). *Stochastic Equations in Infinite Dimensions. Encyclopedia of Mathematics and its Applications* **44**. Cambridge University Press, Cambridge. MR1207136 (95g:60073)

Dashti, M., Harris, S. and Stuart, A. M. (2012). Besov Priors for Bayesian Inverse Problems. *Inverse Probl. Imaging* **6** 183-200.

Dashti, M. and Stuart, A. M. (2011). Uncertainty Quantification and Weak Approximation of an Elliptic Inverse Problem. *SIAM J. Numer. Anal.* **49** 2524–2542.

Diaconis, P. and Stroock, D. (1991). Geometric Bounds for Eigenvalues of Markov Chains. *Ann. Appl. Probab.* **1** 36–61. MR1097463 (92h:60103)

Eberle, A. (2012). Metropolis-Hastings Algorithms for Perturbations of Gaussian Measures in High Dimensions: Contraction Properties and Error Bounds in the Logconcave Case. *ArXiv e-prints.*

Frigessi, A., di Stefano, P., Hwang, C.-R. and Sheu, S. J. (1993). Convergence rates of the Gibbs sampler, the Metropolis algorithm and other single-site updating dynamics. *J. Roy. Statist. Soc. Ser. B* **55** 205–219. MR1210432 (94a:60071)

Geyer, C. J. (1992). Practical Markov Chain Monte Carlo. *Statistical Science* 473–483.

Geyer, C. J. and Thompson, E. A. (1995). Annealing Markov chain Monte Carlo with applications to ancestral inference. *Journal of the American Statistical Association* **90** 909–920.

Hairer, M. (2010). *An Introduction to Stochastic PDEs.* Lecture Notes.

Hairer, M. and Majda, A. J. (2010). A simple framework to justify linear response theory. *Nonlinearity* **23** 909–922. MR2602020 (2011d:82048)

Hairer, M., Mattingly, J. C. and Scheutzow, M. (2011). Asymptotic Coupling and a General Form of Harris' Theorem with Applications to Stochastic Delay Equations. *Probab. Theory Related Fields* **149** 223–259. MR2773030

Hairer, M., Stuart, A. M. and Voss, J. (2007). Analysis of SPDEs Arising in Path Sampling. Part II: The Nonlinear Case. *Ann. Appl. Probab.* 1657–1706.

Hastings, W. K. (1970). Monte-Carlo Sampling Methods Using Markov Chains and their Applications. *Biometrika* **57** 97.

Hjort, N. L., Holmes, C., Müller, P. and Walker, S. G., eds. (2010). *Bayesian Nonparametrics. Cambridge Series in Statistical and Probabilistic Mathematics* **28**. Cambridge University Press, Cambridge. MR2722987 (2011f:62004)

Joulin, A. and Ollivier, Y. (2010). Curvature, Concentration and Error Estimates for Markov Chain Monte Carlo. *Ann. Probab.* **38** 2418–2442. MR2683634 (2011j:60229)

Kipnis, C. and Varadhan, S. R. S. (1986). Central Limit Theorem for Additive Functionals of Reversible Markov Processes and Applications to Simple Exclusions. *Comm. Math. Phys.* **104** 1–19.

Komorowski, T. and Walczuk, A. (2012). Central Limit Theorem for Markov Processes with Spectral Gap in the Wasserstein Metric. *Stoch. Proc. Appl.* **122** 2155–2184. MR2921976

Lassas, M., Saksman, E. and Siltanen, S. (2009). Discretization invariant Bayesian inversion and Besov space priors. *Inverse Problems and Imaging* 87-122.

Łatuszyński, K. and Niemiro, W. (2011). Rigorous Confidence Bounds for MCMC under a Geometric Drift Condition. *J. Complexity* **27** 23–38. MR2745298 (2011k:65009)

Łatuszyński, K. and Roberts, G. O. (2011). CLTs and Asymptotic Variance of Time-Sampled Markov Chains. *Methodol. Comput. Appl. Probab.* 1–11.

Lawler, G. F. and Sokal, A. D. (1988). Bounds on the L^2-Spectrum for Markov Chains and Markov Processes: A Generalization of Cheeger's Inequality. *Amer. Math. Society* **309**.

Lee, P. M. (2004). *Bayesian statistics*, Third ed. Arnold, London. An introduction. MR2069264 (2005a:62005)

Liu, J. S. (2008). *Monte Carlo strategies in scientific computing*. Springer Verlag.

Lovász, L. and Simonovits, M. (1993). Random Walks in a Convex Body and an Improved Volume Algorithm. *Random Structures Algorithms* **4** 359–412. MR1238906 (94m:90091)

Mattingly, J. C., Pillai, N. S. and Stuart, A. M. (2012a). Diffusion Limits of the Random Walk Metropolis Algorithm in High Dimensions. *Ann. Appl. Probab.* **22** 881–930. MR2977981

Mattingly, J. C., Pillai, N. and Stuart, A. M. (2012b). Diffusion Limits of Random Walk Metropolis Algorithms in High Dimensions. *Ann. Appl. Probab.* **22** 881-930.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., Teller, E. et al. (1953). Equation of State Calculations by Fast Computing Machines. *J. Chem. Phys.* **21** 1087.

Meyn, S. and Tweedie, R. L. (2009). *Markov Chains and Stochastic Stability*, Second ed. Cambridge University Press, Cambridge. With a prologue by Peter W. Glynn. MR2509253 (2010h:60206)

Nummelin, E. (1978). A splitting technique for Harris recurrent Markov chains. *Probability Theory and Related Fields* **43** 309–318.

Pillai, N. S., Stuart, A. M. and Thiéry, A. H. (2011). Optimal Proposal Design for Random Walk Type Metropolis Algorithms with Gaussian Random Field Priors. *ArXiv e-prints*.

Robert, C. P. and Casella, G. (2004). *Monte Carlo Statistical Methods*, second ed. *Springer Texts in Statistics*. Springer-Verlag, New York. MR2080278 (2005d:62006)

Roberts, G. O. and Tweedie, R. L. (1996). Geometric Convergence and Central Limit Theorems for Multidimensional Hastings and Metropolis Algorithms. *Biometrika* **83** 95.

Röckner, M. and Wang, F. Y. (2001). Weak Poincaré inequalities and L^2-convergence rates of Markov semigroups. *J. Funct. Anal.* **185** 564–603. MR1856277 (2002j:47075)

Rudolf, D. (2012). Explicit Error Bounds for Markov Chain Monte Carlo. *Dissertationes Math. (Rozprawy Mat.)* **485** 1–93. MR2977521

Schwab, C. and Stuart, A. M. (2012). Sparse Deterministic Approximation of Bayesian Inverse Problems. *Inverse Probl.* **28** 045003, 32. MR2903278

Sinclair, A. and Jerrum, M. (1989). Approximate Counting, Uniform Generation and Rapidly Mixing Markov Chains. *Inform. and Comput.* **82** 93–133.

Stuart, A. M. (2010). Inverse Problems: A Bayesian Perspective. *Acta Numer.* **19** 451–559. MR2652785 (2011i:65093)

Tierney, L. (1998). A Note on Metropolis-Hastings Kernels for General State Spaces. *Ann. Appl. Probab.* **8** 1–9. MR1620401 (99a:60066)

Vollmer, S. (2013). Reversible Proposal for MCMC and the Preservation of Spectral Gaps. in preparation.

Wang, F. Y. (2003). Functional Inequalities for the Decay of Sub-Markov Semi-Groups. *Potential Anal.* **18** 1–23. MR1953493 (2004a:47051)

Research Article II

Dimension-Independent MCMC Sampling for Inverse Problems with Non-Gaussian Priors

Sebastian J. Vollmer, 2013. *Submitted to the SIAM/ASA Journal on Uncertainty Quantification, 22 pages.*

DIMENSION-INDEPENDENT MCMC SAMPLING FOR INVERSE PROBLEMS WITH NON-GAUSSIAN PRIORS

SEBASTIAN J. VOLLMER*

Abstract. The computational complexity of MCMC methods for the exploration of complex probability measures is a challenging and important problem. A challenge of particular importance arises in Bayesian inverse problems where the target distribution may be supported on an infinite dimensional space. In practice this involves the approximation of measures defined on sequences of spaces of increasing dimension. Motivated by an elliptic inverse problem with non-Gaussian prior, we study the design of proposal chains for the Metropolis-Hastings algorithm with dimension independent performance. Dimension-independent bounds on the Monte-Carlo error of MCMC sampling for Gaussian prior measures have already been established. In this paper we provide a simple recipe to obtain these bounds for non-Gaussian prior measures. To illustrate the theory we consider an elliptic inverse problem arising in groundwater flow. We explicitly construct an efficient Metropolis-Hastings proposal based on local proposals, and we provide numerical evidence which supports the theory.

Key words. MCMC, inverse problems, Bayesian, spectral gaps, non-Gaussian

AMS subject classifications. 65C40, 60J22, 60J05, 35R30, 62F15

1. Introduction. In many applications in science and technology the main unknowns cannot be observed directly or a direct observation would be destructive. A prime example for this is computed tomography where the aim is the reconstruction of the properties of a human body given measurements at the rim of the X-ray tube. Often it is possible to model the data as the output of a mathematical model taking the unknowns as parameters. The area of inverse problems is concerned with the reconstruction of these parameters from data. Classically, this is achieved by choosing the parameter which minimises a regularized least squares functional. Whereas it is difficult to quantify the uncertainty for this method, it is straightforward in the Bayesian approach to inverse problems. The Bayesian method is based on the idea that not all parameter choices are a priori equally likely. Instead, the a priori knowledge about these parameters is modelled as a probability distribution - called the prior. By specifying the distribution of the noise, the parameters and the observed data can then be treated as jointly distributed random variables. Under certain conditions on the prior, model and noise, there exists a unique conditional distribution of the parameters given the data. This distribution is called the posterior and is an update of the prior using the data. In this way uncertainty can be quantified using the posterior variance or the posterior probability of a set in the parameter space. Usually, the posterior is only expressed implicitly as an unnormalised density with respect to the prior. For this reason sampling algorithms are used to approximate posterior expectations by the sample average. The most famous sampling algorithms are those of Metropolis-Hastings type which were introduced by N. Metropolis [39] and generalised by W. K. Hastings [23]. The idea of the Metropolis-Hastings algorithm is to add an accept-reject step to a Markov chain proposal so that the resulting Markov chain converges to the target measure. For a recent review of Markov Chain Monte Carlo (MCMC) algorithms we refer the reader to [8]. In this article, we consider target measures arising from Bayesian inverse problems. In this case the underlying mathematical models are often based on PDEs or integral operators that have to be approximated for computations. Appropriate reviews are contained in [26] and

*University of Warwick, Mathematics Institute, Zeeman Building, Coventry CV4 7AL

[51]. The former is a key reference as the Bayesian approach is applied to real world applications using MCMC and optimisation techniques. This reference shows that the resulting methods can compete with state-of-the regularisation techniques in, for example, dental X-ray imaging. Whereas this reference applies the Bayesian method to a discretised version of the inverse problem, the survey article [26] concerns the Bayesian approach to the full infinite dimensional problem which was originally developed in [10]. This approach was also taken independently in [32, 29]. We design efficient proposals for Metropolis-Hastings algorithms for target measures arising from the Bayesian approach to inverse problems.

The overall error in estimating posterior expectation using MCMC can be reduced by a better approximation to the underlying continuum model or by increasing the number of samples used in the average. Under limited computing power, this results in a trade-off between approximation and Monte-Carlo error which has quantitatively been investigated in [24]. This trade-off is influenced by the fact that for many sampling algorithms the Monte-Carlo error increases with the dimension of the state space. Thus, even if the number of samples stays fixed, a finer approximation can lead to a worse Monte-Carlo error. For the Bayesian approach to inverse problems, the prior, the posterior and some Metropolis-Hastings algorithms can be formulated on appropriate function spaces. The Monte-Carlo error of these algorithms for a fixed number of MCMC steps is only effected up to a point by the dimensionality of the state space. This insight was first properly stated by A. M. Stuart, J. Voss and P. Wiberg in [53]. These three authors pointed out the need for dimension independent sampling algorithms and constructed such methods for conditioned diffusions in the additive noise setting. This aim has been achieved by constructing well-defined sampling algorithms for measures defined via a density with respect to Gaussian measures on function spaces. This subject has then been developed further, both for conditioned diffusion and Bayesian inverse problems, and is surveyed in [5] and [11]. Recently, we have made this insight rigorous for the preconditioned Crank-Nicolson sampling algorithm by considering the convergence rate of underlying Markov chains in terms of the L^2-spectral gaps [21]. In the inverse problem setting, this corresponds to posteriors arising from priors given by a density with respect to a Gaussian measure. However, having a density with respect to a Gaussian measure is not a natural assumption for all applications. In image processing for example, Bayesian methods are used to recover and reconstruct images and Gaussian priors that tend to remove or blur the edges which are supposed to be recovered. This effect is described in detail in [7]. Recovering the sharp interfaces between different rocks is also important for applications in geophysics. This has lead to the investigation of non-Gaussian priors for example in [13, 22, 30, 31]. In this paper we extend the idea of dimension independent sampling to Bayesian inverse problems with non-Gaussian priors.

In general, we assume that the target measure μ is a Borel measure on the Banach space X. It is given by

$$\mu \propto L\,\mu_0 \tag{1.1}$$

where μ_0 is the reference Borel probability measure on X. The main result in this article is based on the observation that the proposal kernels for function space sampling algorithms in [11] are all reversible and have an L^2_γ-spectral gap when applied to the Gaussian reference measure γ. If an MCMC algorithm for a target measure μ has an L^2_μ-spectral gap, then asymptotic and non-asymptotic confidence intervals can be derived for the Monte-Carlo error for any L^2-function f in terms of the variance

$\mathrm{Var}_\mu(f)$ and a lower bound on the L^2-spectral gap. More details on this can be found in Section 3.1. This motivates our main result which can be summarized as follows:

THEOREM 1.1. *Suppose that $\mu \propto L\,\mu_0$, L is bounded above and away from zero and the proposal Markov kernel Q has an $L^2_{\mu_0}$-spectral gap. Then the lazy versions of the resulting Metropolis-Hastings Markov kernel have an L^2_μ-spectral gaps.* The lazy version of a Markov chain follows the transition with probability $\frac{1}{2}$ and does not make a transition with probability $\frac{1}{2}$. The resulting Markov chain has a positive spectrum and therefore a bound on the second largest eigenvalue is enough to obtain an $L^2_{\mu_0}$-spectral gap. This fact is well-known in the literature and goes at least back to [36]. Thus, our strategy for Bayesian inverse problems will be to design proposals that are reversible with respect to the prior and that have an $L^2_{\mu_0}$-spectral gap. The Metropolis-Hastings algorithm will then perform an accept-reject step according to the likelihood in order to produce samples from the posterior. This result should be viewed in context of our recent results in [21] demonstrating that the L^2-spectral gap of the preconditioned Crank-Nicolson (pCN) algorithm with respect to the Gaussian reference measure is preserved for the corresponding Metropolis-Hastings kernel. In the same article, we used the Ornstein-Uhlenbeck proposal and assumed that L in (1.1) is log-Lipschitz. However, no global bounds on L were needed in order to prove the preservation of the L^2_μ-spectral gap. In this way the main result here can be viewed as an extension to a much larger class of proposals and reference measures under partially stronger assumptions.

A related result has been proved for the Gibbs sampler applied to perturbations of Gaussian measures in [1]. However, it is not clear how it could be generalized for arbitrary reference measure.

Our main result stated above is proved by expressing the L^2_μ-spectral gap in terms of the associated Dirichlet form in Section 3.2. The the proof is similar to that of the comparison Theorem [16]. With this method we only obtain a bound on the upper L^2_μ-spectral gap but not the L^2_μ-spectral gap (see Section 3.2). This problem can be circumvented by considering the lazy chain.

As guiding application, we consider the posterior arising from the inverse problem of reconstructing the diffusion coefficient from noisy measurements of the pressure in a Darcy model of groundwater flow. The underlying continuum model then corresponds to a linear elliptic PDE in divergence form. The Bayesian approach to the inverse problem is taken by placing a prior based on a series expansion with uniformly distributed coefficients. In [50], well-definedness of the Bayesian inverse problem and a general Polynomial Chaos (gPC) method for approximating posterior expectations are established for this inverse problem. For a full comparison of the gPC method to MCMC algorithms, we refer the reader to [24]. In this research article, the Monte-Carlo error of the Metropolis-Hastings algorithms is bounded using convergence results for Markov chains from [40]. However, these results assume the Markov chain associated with the Metropolis-Hastings algorithm is ϕ-irreducible. On function spaces, this condition seems only to be verifiable in special cases such as the independence sampler (IS) algorithm. The IS is an MCMC algorithm making independent proposals from one distribution. This choice of proposal leads to a poor performance especially if the posterior is concentrated. The ergodicity properties of the algorithm are investigated in [38]. Our main result allows us to extend bounds on the Monte-Carlo error in [24] to a large class of locally moving algorithms. In particular we design Reflection Random Walk Metropolis (RRWM) algorithms and show that it has the same asymptotic complexity as the IS algorithm using the main

result of this article. Finally, we provide numerical evidence that the RRWM and the IS algorithms are robust with respect to an increase in dimension. Furthermore, the simulations show that the RRWM algorithm is a substantial improvement over the IS algorithm especially for concentrated measures.

We give a brief exposition of Bayesian inverse problems and Metropolis-Hastings algorithms on general state spaces in Section 2. In Section 3, we introduce L^2-spectral gaps and the consequences for the sample average before we prove our main theorem. Section 4 focuses on elliptic inverse problems. We construct the RRWM sampling algorithm which satisfies the conditions of our main theorem for the previously introduced elliptic inverse problem. In Section 5, we compare the RRWM, the standard Random Walk Metropolis (RWM) and the IS algorithms using numerical simulations for the posterior arising from this particular inverse problem.

2. Review of Bayesian Inverse Problems and Metropolis-Hastings Algorithms. This section is devoted to giving a brief summary on the relevant material on Bayesian inverse problems and to giving an introduction to Metropolis-Hastings algorithms on general state spaces. For more details we refer the reader to [51, 52] and [54, 8] respectively. The main idea of the Bayesian approach is to treat the parameters, the output of the mathematical model and the data as jointly distributed random variables. The randomness of the parameters is introduced artificially to subjectively model the uncertainty based on the a priori knowledge. The distribution of the parameters is called the prior. In the Bayesian framework the conditional probability distribution of the parameters given the noisy data is called the posterior. It is an update of the prior using the data and can be viewed as the solution to the inverse problem because it describes the a posteriori uncertainty about the parameters. The posterior is a very important tool because it can be used to

- obtain point estimates for the unknown in an inverse problem such as the posterior mean or the MAP estimator which can be related to the Tikhonov regularisation, see [14].
- quantify the uncertainty through the posterior variance or the posterior probability of a set in the parameter space.

We concentrate on the latter and note that both quantities can be written as posterior expectations. However, the calculation of posterior expectations is difficult to establish because the normalization constant is unknown. This is where Metropolis-Hastings algorithms come into the play. They can be used to approximate expectations without using the normalization constant because only ratios of the densities are needed. In order to implement a Metropolis-Hastings algorithm, the parameter space and the forward problem have to be discretised leading to a high dimensional state space. Therefore it is crucial that the algorithm performs well as the dimension increases which might be due to a a finer discretisation of the underlying continuum model. The performance of the algorithm can be measured by convergence of the underlying stochastic process to equilibrium. We survey different ways how the convergence rate is measured and provide them at the end of Section 2.2.

2.1. Bayesian Inverse Problems. In the following we consider a general inverse problem for which the data is generated by

$$y = \mathcal{G}(a) + \eta \in Y.$$

Here η is the observational noise, $a \in X$ is the input of the mathematical model, for example the initial condition or coefficients for a PDE, and $\mathcal{G}$ is the forward operator,

a mapping between the Hilbert spaces X and Y. In this setting the inverse problem is concerned with the reconstruction of the input a to the model $\mathcal{G}$ given its noisy output, the data y. The problem has typically to be regularised in some way because $\mathcal{G}$ can be non-injective and η is unknown. Classically, this is done by choosing a as the minimiser of a regularised least squares functional. Regularisation can also be approached by placing a prior $\mu_0(da)$ probability measure on a containing all the a priori information. If, in addition, the forward operator $\mathcal{G}$ and the distributions of η is given, then (a, y) can be treated as jointly varying random variables. Under mild assumptions, there exists a conditional probability measure on a which is called the posterior, an update of the prior using the data. In contrast to the minimiser of a least squares functional, the posterior is continuous in the data with respect to the total variation and the Hellinger distance. The posterior is also continuous with respect to approximations of the forward model. For the precise statements of these results we refer the reader to the surveys [51] and [52]. Due to the latter result, it is possible to bound the difference between expectations calculated with respect to the posterior associated with the infinite dimensional and the discretised forward model. In Sections 2.2 and 3, we explain how the Metropolis-Hastings algorithm can be used to approximate expectations with respect to the posterior associated with the discretised forward model and how the resulting Monte-Carlo error can be bounded. In order to use Metropolis-Hastings algorithm we specify the posterior more explicitly. For finite dimensional distributions given as probability densities Bayes' rule yields

$$\text{posterior} \propto \text{likelihood} \times \text{prior}, \tag{2.1}$$

more details for classical Bayesian statistics can be found in [4].

We consider a generalisation of Bayes' rule to infinite dimension. In this article, we only consider finite dimensional data, that is $Y = \mathbb{R}^N$, but the results in [51] and [52] allow the data to be infinite dimensional as well. In the case of finite data, where the observational noise has a Lebesgue density ρ, the Bayesian framework can be summarised as follows

$$\begin{aligned} \text{Prior} \quad & a \sim \mu_0 \\ \text{Noise} \quad & \eta \text{ with pdf } \rho(\eta) \\ \text{Likelihood} \quad & y|a \text{ r.v. with pdf } \rho(y - \mathcal{G}(a)) \\ & L(a) = \rho(y - \mathcal{G}(a)) \\ \text{Posterior} \quad & \frac{d\mu^y}{d\mu_0}(a) \propto L(a) \end{aligned} \tag{2.2}$$

Subsequently, we drop the y and hope that this does not cause any confusion for the reader. The important point to note here is that the Equations (2.1) and (2.2) are of the same form as the general target measure for the Metropolis-Hastings algorithm as in Equation (1.1) which will be reviewed in the next section.

2.2. The Metropolis-Hastings Algorithm on General State Spaces. The common idea of MCMC algorithms is to create a Markov chain with a prescribed invariant measure, called the target measure. Samples of this Markov chain under (mild) conditions satisfy a law of large numbers and can thus be used to approximate expectations with respect to the target measure. Under stronger conditions it is possible to control the resulting random error using a central limit theorem (CLT) or to establish bounds on the mean square error.

In this article, we consider the application of Metropolis-Hastings algorithms to Bayesian inverse problems previously introduced in Section 2.1. In order to implement Metropolis-Hastings algorithms to approximate posterior expectations such as the mean or the variance we have to discretise and approximate $\mathcal{G}$. Nevertheless some Metropolis-Hastings algorithms can be formulated on function spaces and it is conceivable that those perform better as the dimension of the approximation increases than those that cannot be formulated on function spaces. We have made this rigorous for the preconditioned Crank-Nicolson and the standard Random Walk Metropolis (RWM) algorithms for Gaussian prior in [21]. In the present article, we present this problem for non-Gaussian priors. For this reason we formulate the Metropolis-Hastings algorithm on general state spaces following [54].

The idea of the Metropolis-Hastings kernel is to add an independent accept-reject step to a proposal Markov kernel $Q(x, dy)$ in order to produce a Markov kernel $P(x, dy)$ with μ as an invariant measure, that is

$$\mu P = \int_X \mu(dx)P(x, dy) = \mu(dy).$$

Subsequently, we will discuss a choice of the acceptance probability such that μ is invariant for P. Thereafter we consider the reversibility of both the proposal and Metropolis-Hastings kernel. This property is important because it yields error bounds on the sample average in combination with an L^2-spectral gap (c.f. Section 3). We will close this section by reviewing convergence results for Metropolis-Hastings algorithms.

The Metropolis-Hastings algorithm accepts a move from x to y proposed by the kernel $Q(x, dy)$ with acceptance probability $\alpha(x, y)$. Thus, the algorithm takes the following form

Algorithm Initialise X_0. For i=0,...,n do:
Generate $Y \sim Q(X_i, \cdot)$, $U \sim \mathcal{U}(\prime, \infty)$ independently and set

$$X_{i+1} = \begin{cases} Y & \text{if } \alpha(X_i, Y) > U \\ X_i & \text{otherwise} \end{cases}.$$

The transition kernel $P(x, dy)$ associated with the Metropolis-Hastings algorithm can be written as

$$P(x, dy) = \alpha(x, y)Q(x, dy) + \delta_x(dy)\left(1 - \int_E Q(x, dy)\alpha(x, y)\right).$$

If the Radon-Nikodym derivative $\frac{d\mu(dy)Q(y,dx)}{d\mu(dx)Q(x,dy)}$ exists, then μ is invariant for P for the choice

$$\alpha(x, y) := \min\left(1, \frac{d\mu(dy)Q(y, dx)}{d\mu(dx)Q(x, dy)}\right). \tag{2.3}$$

In finite dimensions a common dominating measure is the Lebesgue measure. However, in infinite dimensions there is no equivalent of the Lebesgue measure. Furthermore, the Feldmann-Hajek theorem (c.f. [12]) implies that $Q(x, dy)$ corresponding to a Gaussian random walk is mutually singular for different x. Nevertheless, it is instructive to consider the case if there is a common dominating measure λ, that is

$$\mu \propto L\lambda \text{ and } Q(x, dy) = q(x, y)\lambda(dy),$$

then

$$\alpha(x,y) = \frac{L(y)q(y,x)}{L(y)q(x,y)} \wedge 1. \tag{2.4}$$

Note that we can work with the unnormalised density because the acceptance probability is based on the ratio of L at x and y. In fact, μ is not only invariant for the Metropolis-Hastings kernel P but but the kernel P is also reversible with respect to μ (see [54]), which is defined subsequently.

DEFINITION 2.1. *A Markov kernel P is reversible with respect to a measure μ if*

$$\mu(dx)P(x,dy) = \mu(dy)P(y,dx).$$

If the proposal Q is reversible with respect to the prior μ_0, then (2.3) reduces to

$$\alpha(x,y) = \min\left(1, \frac{d\, L(y)\mu_0(dy)Q(y,dx)}{d\, L(x)\mu_0(dx)Q(x,dy)}\right) = \frac{L(y)}{L(x)} \wedge 1. \tag{2.5}$$

The problem in designing (efficient) proposals on function spaces is that the Radon-Nikodym derivative in Equation (2.3) does often not exist. This follows from different almost sure properties of $\mu(dx)$ and $\int Q(y,dx)d\mu(y)$ such as quadratic variation or regularity properties. The simplest proposal which preserves these properties is to pick ν with the same almost sure properties and use the proposal kernel

$$Q(x,dy) = \nu(dy).$$

The resulting algorithm is called independence sampler (IS) because the proposal does not depend on the current state x.

For Bayesian inverse problems it is natural to design proposals that are reversible for the prior because this preserves the almost sure properties and leads to a simple acceptance rule only involving the likelihood (c.f. Equation (2.5)). In particular this is the case for the IS algorithm with $\nu = \mu_0$.

In general, Metropolis-Hastings algorithms are run in order to approximate $\int \mu(dx)f(x)$ by

$$S_{n,n_0}(f) = \frac{1}{n}\sum_{i=n_0}^{n_0+n} f(X_i) \tag{2.6}$$

where n_0 is the burn-in corresponding to throwing away the first n_0 samples in order to reduce the bias. The resulting error takes the form.

$$e_{n,n_0}(f) = \mu(f) - S_{n,n_0}(f).$$

The complexity of Metropolis-Hastings algorithms can be quantified as

$$\text{number of necessary steps} \times \text{cost of one step.}$$

The cost of one step is usually easy to quantify and depends on the problem at hand. The number of necessary steps depends on the prescribed error level (for example fixed width (asymptotic) confidence interval see [25], [33] and [47]) and the convergence properties of the Markov chain. If X_i were i.i.d. samples, the central

limit theorem yields that the error is of order $\mathcal{O}(n^{-\frac{1}{2}})$. A large part of the literature is concerned with proving that this is still the case for the correlated samples of an algorithm or even bounding the leading constant in $\mathcal{O}(n^{-\frac{1}{2}})$. The methods which are used in the literature are related to different notions of convergence of the Markov chain associated with an Metropolis-Hastings algorithm to its equilibrium. These can broadly be classified as follows [40, 47]:

1. For a metric d on the space of measures, such as the total variation or the Wasserstein metric, the rate of convergence to equilibrium can be characterised through the decay of $d(\nu P^n, \mu)$ where ν is the initial distribution of the Markov chain.
2. For the Markov operator P the convergence rate is given as the operator norm of P on a space of functions from X to $\mathbb{R}$ modulo constants. The most prominent example here is the L^2-spectral gap (see Section 3).
3. In the regeneration and the so-called split-chain approach the evolution of the algorithm is split into independent pieces. In this case the CLT of the Markov chain follows from the CLT for i.i.d random variables.

The regeneration and total variation methods have been very successful in obtaining rates for finite dimensional problems. An excellent review of this is given in [45]. However, in that article it is assumed that the algorithm is ψ-irreducible, that is the existence of a positive measure ϕ such that

$$\phi(A) > 0 \Rightarrow P(x, A) > 0 \quad \forall x.$$

This property often fails for infinite dimensional problems because the transition probabilities tend to be mutually singular for different starting points (this is even the case for a Gaussian random walk). One exception is the IS algorithm [24].

Having introduced Bayesian inverse problems and Metropolis-Hastings algorithms on general state spaces, we are now in the position to formulate and prove the main result of this article.

3. L^2-Spectral Gaps for Metropolis-Hastings algorithms. Metropolis-Hastings algorithms play an important role for the approximation of $\mu(f)$ by $S_{n,n_0}(f)$ given by Equation (2.6) with error $e_{n,n_0}(f)$. Therefore much theory is aimed at establishing (asymptotic) bounds on the error. We will first define () L^2-spectral gaps and state the appropriate theorems form the literature that allow us to bound the error in terms of an L^2-spectral gap. Here lies the importance of our main theorem because it establishes an L^2_μ-spectral gap for a lazy versions of the Metropolis-Hastings chains for the posterior in terms of the $L^2_{\mu_0}$-spectral gap of the corresponding proposal chain for the prior.

3.1. The L^2-Spectral Gap and its Implications. In order to define L^2-spectral gaps, we recall how a Markov kernel P with invariant measure μ acts on $L^2_\mu(X)$. Recall that $L^2_\mu(X)$ is the set of all Borel-measurable functions on X such that

$$\|f\|^2_{L^2_\mu} = \int_X f^2(x) d\mu(x) < \infty.$$

The Markov kernel P acts naturally on $L^2_\mu(X)$ as

$$Pf(x) = \int_X P(x, dy) f(y).$$

Jensen's inequality implies that the spectrum $\sigma(P)$ of P is contained in the unit disk. If P is reversible such as the transition operator of the Metropolis-Hastings algorithm then the spectrum is real valued, thus $\sigma(P) \subseteq [-1,1]$. Moreover P always has an eigenvalue 1 as $P1 = 1$. The L^2_μ-spectral gap is given by the difference between 1 and the modulus of the second largest the largest eigenvalue (in terms of the modulus) of Markov operator restricted to $L^2\mu$.The following definition is based on the variational characterisation of the largest eigenvalue of the linear operator P modulo constants. This can be represented as the difference between 1 and the spectral radius of the operator P restricted to orthogonal complement of the space of constant functions which we denote by $L^2_0(\mu)$.

DEFINITION 3.1. *(L^2_μ-spectral gap) A Markov operator P with invariant measure μ has an L^2_μ-spectral gap $1-\beta$ if*

$$\beta = \sup_{f \in L^2_\mu} \frac{\|P(f-\mu(f))\|_2}{\|f-\mu(f)\|_2} = \sup_{f \in L^2_0(\mu)} \frac{\|Pf\|_2}{\|f\|_2} < 1.$$

3.1.1. Characterisation of L^2_μ-Spectral Gaps. For a self-adjoint operator $A : H \to H$ the smallest and largest eigenvalue are charcterised by

$$\lambda^H_{\min}(A) = \inf_{f \in H} \frac{\langle Af, f\rangle}{|f|^2} \text{ and } \lambda^H_{\max}(A) = \sup_{f \in H} \frac{\langle Af, f\rangle}{|f|^2}, \tag{3.1}$$

respectively. We will write $\lambda := \lambda^{L^2_0(\mu)}_{\min}(P)$ and $\Lambda := \lambda^{L^2_0(\mu)}_{\min}(P)$. In this way the L^2_μ-spectral gap takes the form

$$1-\beta = \min\{1-\lambda, 1\text{-}\Lambda\}. \tag{3.2}$$

This motivates the following notions.

DEFINITION 3.2. *For a Markov kernel P, we refer to the quantities $1-\lambda$ and 1-Λ as the lower and upper L^2_μ-spectral gap, respectively.*

These notions are introduced because we will only be able to obtain a lower bound for the upper L^2_μ-spectral gap. In some sense, an upper L^2_μ-spectral gap is sufficient as it is possible to modify the chain resulting in an L^2_μ-spectral gap of almost the same size. The modification consists of adding an additional rejection step resulting in the so-called lazy-chain which we review at the beginning of Section 3.2.

The upper spectral gap $1-\lambda^{L^2_0(\mu)}_{\max}(P)$ is given by the smallest eigenvalue $\lambda^{L^2_0(\mu)}_{\min}$ of $I-P$ on $L^2_0(\mu)$ can be characterised as

$$1-\lambda^{L^2_0(\mu)}_{\max}(P) = \inf_{f \in L^2_0(\mu)} \frac{\langle (I-P)f, f\rangle}{|f|^2} = \inf_{f \in L^2(\mu)} \frac{\langle (I-P)\Pi f, \Pi f\rangle}{|\Pi f|^2} \tag{3.3}$$

where $\Pi : L^2_0(\mu) \to L^2(\mu)$ is the orthogonal projection onto $L^2_0(\mu)$ given by

$$\Pi f = f - \mu(f).$$

The denominator can be rewritten as

$$\begin{aligned} |\Pi f|^2 &= Var_\mu(f) = \int (f-\mu(f))^2 \, d\mu \\ &= \int f^2 d\mu - \mu(f)^2 = \frac{1}{2}\int \mu(dx)\mu(dy)\,(f(x)-f(y))^2 . \end{aligned} \tag{3.4}$$

The nominator in (3.3) can be rewritten as

$$\begin{aligned}\langle (I-P)(f-\mu(f)), f-\mu(f)\rangle &= \langle (I-P)f, f-\mu(f)\rangle = \langle (I-P)f, f\rangle \\ &= \int \mu(dx)P(x,dy)\left(f(x)^2 - f(x)f(y)\right)dy \\ &= \frac{1}{2}\int \mu(dx)P(x,dy)\left(f(x)-f(y)\right)^2 dy =: \mathcal{E}_\mu^P(f,f).\end{aligned}$$

The bilinear form $\mathcal{E}(f,f)$ is the Dirichlet form associated with the Markov chain given through the transition kernel P. There is a large literature on studying Markov processes through their Dirichlet form. We refer the reader to [48] for a short survey for time-continuous Markov processes, to [17] for generalities of the theory and to [35] for an review for discrete Markov chains. We only need the characterisation of the upper spectral gap

$$1-\lambda_{\max}^{L_0^2(\mu)} = \inf_{f\in L^2(\mu)} \frac{\mathcal{E}_\mu^P(f,f)}{\mathrm{Var}(f)} \tag{3.5}$$

that we have just derived and which will be used to derive our main theorem in Section 3.2.

3.1.2. Implications of L^2_μ-Spectral Gaps. The two main implications of an L^2_μ-spectral gap are a CLT for $S_{n,n_0}(f)$ which implies an asymptotic bound on the error of size $\mathcal{O}(\frac{1}{\sqrt{n}})$ and a non-asymptotic bound on the mean square error. The latter yields non-asymptotic confidence intervals using Chebyshev's inequality.

In the following we present the the precise statement of the CLT due to Kipnis and Varadhan [27]. The following version is taken from [34].

Proposition 3.3. *(Kipnis-Varadhan) Consider an ergodic Markov chain with transition operator P which is reversible with respect to a probability measure μ and which has an L^2_μ-spectral gap $1-\beta$. For $f\in L^2$ we define*

$$\sigma^2_{f,P} = \left\langle \frac{1+P}{1-P} f, f \right\rangle.$$

Then for $X_0\sim\mu$ the expression $\sqrt{n}(S_n-\mu(f))$ converges weakly to $\mathcal{N}(0,\sigma^2_{f,P})$. Moreover, the following inequality holds

$$\sigma^2_{f,P} \le \frac{2\mu((f^2-\mu(f)^2))}{(1-\beta)} < \infty.$$

The non-asymptotic bounds on the mean square error is due to Rudolf [47] and take the following form

Proposition 3.4. *Suppose that we have a Markov chain with Markov operator P having an L^2_μ-spectral gap $1-\beta$. For $p\in(2,\infty]$ let $n_0(p)$ be defined by*

$$n_0(p) \ge \frac{1}{\log(\beta^{-1})} \begin{cases} \frac{p}{2(p-2)}\log(\frac{32p}{p-2}) \left\| \frac{d\nu}{d\mu} - 1 \right\|_{\frac{p}{p-2}} & p\in(2,4) \\ \log(64) \left\| \frac{d\nu}{d\mu} - 1 \right\|_{\frac{p}{p-2}} & p\in[4,\infty]. \end{cases} \tag{3.6}$$

Then for S_{n,n_0} as in Equation (3.6) and $f\in L^2_\mu$

$$\sup_{\|f\|_2\le 1} \mathbb{E}\left[\left(\mu(f)) - \frac{1}{n}\sum_{i=n_0}^{n_0+n} f(X_i)\right)^2\right] \le \frac{2}{n(1-\beta)} + \frac{2}{n^2(1-\beta)^2}.$$

If a Metropolis-Hastings algorithm has an L^2_μ-spectral gap, then the two results above can be used to derive asymptotic and non-asymptotic confidence intervals and levels for the Monte-Carlo error $e_{n,n_0}(f) = \mu(f) - S_{n,n_0}(f)$. The CLT only provides asymptotic confidence intervals. In contrast, bounds on the MSE implies non-asymptotic confidence intervals using Chebyshev's inequality. Moreover, the size of the confidence intervals can be shrank using the 'median trick' which estimates $\mathbb{E}f$ through the median of multiple shorter runs leading to exponential tight bounds. This trick was developed for MCMC algorithms in[42] and another good reference is given by [33].

3.2. Main Result . In order to bound the L^2_μ-spectral gap, we need an upper bound on Λ and a lower bound on λ. However, due to the construction of the lazy chain, an upper bound on Λ is enough to a obtain a L^2_μ-spectral gap for the lazy chain. For a Markov chain with transition kernel P we define the associated lazy Markov chain as $\tilde{P} = \frac{1}{2}(I + P)$. This transition can be interpreted as a two step procedure. We are throwing a coin and

- if it comes up heads, the Markov chain makes a transition according to P.
- if it come up tails, the Markov chain does not make a transition.

. It is straightforward to see that

$$\sigma(\tilde{P}) \subseteq [\frac{1+\lambda}{2}, \frac{1+\Lambda}{2}].$$

Thus, $\tilde{P}$ has a spectral gap if $\Lambda < 1$. Its size can be optimized be choosing the acceptance probability dependent on Λ and if a bound is available on λ instead of $\frac{1}{2}$.

The following theorem provides an explicit lower bound on the L^2_μ-spectral gap of the lazy version of the Metropolis-Hastings chain in terms of the $L^2_{\mu_0}$-spectral gap of the proposal chain and in terms of the bounds on the density of the posterior with respect to the prior.

The following result is close in spirit to the comparison theorem for discrete Markov chains obtained in [16].

THEOREM 3.5. *Suppose that the proposal kernel Q satisfies a lower bound on the upper $L^2_{\mu_0}$-spectral gap $1 - \lambda_{max}^{L^2_0(\mu)}(P) > 0$ and the target measure takes the form*

$$\mu = \frac{L}{Z}\mu_0.$$

Then the upper L^2_μ-spectral gap satisfies

$$\left(1 - \lambda_{max}^{L^2_0(\mu_0)}(Q)\right) \frac{L^{\star 3}}{L^3_\star} \geq 1 - \lambda_{max}^{L^2_0(\mu)}(P) \geq \frac{L^4_\star}{L^{\star 4}} \left(1 - \lambda_{max}^{L^2_0(\mu_0)}(Q)\right)$$

where $L_\star := \inf L \leq L \leq \sup L = L^\star$. In particular the lazy version $\tilde{P}$ has an L^2_μ-spectral gap $1 - \beta_{lazy}$ satisfying

$$\frac{1}{2}\left(1 - \lambda_{max}^{L^2_0(\mu_0)}(Q)\right) \frac{L^{\star 3}}{L^3_\star} \geq 1 - \beta_{lazy} \geq \frac{1}{2}\frac{L^4_\star}{L^{\star 4}} \left(1 - \lambda_{max}^{L^2_0(\mu_0)}(Q)\right).$$

Proof. From Equation (3.4) follows that

$$\frac{L^2_\star}{Z^2}\mathrm{Var}_\mu(f) \leq \mathrm{Var}_{\mu_0}(f) \leq \frac{L^{\star 2}}{Z^2}\mathrm{Var}_\mu(f).$$

Similarly we notice that

$$\begin{aligned}
\mathcal{E}_\mu^P(f,f) &= \frac{1}{2}\int \mu_0(dx)Q(x,dy)\frac{L}{Z}\alpha(x,y)\left(f(x)-f(y)\right)^2 \\
&\geq \frac{L_\star}{Z}\alpha_\star\frac{1}{2}\int \mu_0(dx)Q(x,dy)\left(f(x)-f(y)\right)^2 \\
&\geq \frac{L_\star^2}{ZL^\star}\left(1-\lambda_{\max}^{L_0^2(\mu_0)}(Q)\right)\mathrm{Var}_{\mu_0}(f) \\
&\geq \frac{L_\star^4}{Z^3L^\star}\left(1-\lambda_{\max}^{L_0^2(\mu_0)}(Q)\right)\mathrm{Var}_{\mu}(f) \\
&\geq \frac{L_\star^4}{L^{\star 4}}\left(1-\lambda_{\max}^{L_0^2(\mu_0)}(Q)\right)\mathrm{Var}_{\mu}(f).
\end{aligned}$$

Thus we can conclude that

$$1-\lambda_{\max}^{L_0^2(\mu)}(P) = \inf_{f\in L^2(\mu)}\frac{\mathcal{E}_\mu^P(f,f)}{\mathrm{Var}(f)} \geq \frac{L_\star^4}{L^{\star 4}}\left(1-\lambda_{\max}^{L_0^2(\mu_0)}(Q)\right).$$

The other inequality is obtained in the following way

$$\begin{aligned}
\mathcal{E}_\mu^Q(f,f) &= \frac{1}{2}\int \mu_0(dx)Q(x,dy)\frac{L}{Z}\alpha(x,y)\left(f(x)-f(y)\right)^2 \\
&\geq \frac{L_\star}{Z}\frac{1}{2}\int \mu(dx)P(x,dy)\left(f(x)-f(y)\right)^2 \\
&\geq \frac{L_\star}{Z}\left(1-\lambda_{\max}^{L_0^2(\mu)}(P)\right)\mathrm{Var}_{\mu}(f) \\
&\geq \frac{L_\star^3}{Z^3}\left(1-\lambda_{\max}^{L_0^2(\mu)}(P)\right)\mathrm{Var}_{\mu_0}(f).
\end{aligned}$$

The result for the lazy chain follows from the discussion at the beginning of this Section. □

This result highlights the insight that the reference measure is crucial for designing efficient sampling algorithms on function spaces. A typical example would be the use of a Markov chain that has an $L^2_{\mu_0}$-spectral gap where μ_0 is the prior of a Bayesian problem. If the likelihood is bounded, then the lazy version of the resulting Metropolis-Hastings algorithm with this chain as the proposal has an L^2_μ-spectral gap with μ being the posterior. However, the result is not limited to this situation because μ_0 and μ can be arbitrary measures such that the density of μ with respect to μ_0 is bounded.

Remark 1. *For a fixed target measure a larger $L^2_{\mu_0}$-spectral gap of Q implies a larger lower bound on the L^2_μ-spectral gap of P. In particular the largest lower bound is achieved for the IS algorithm. It is import to note that this does not imply that this choice leads to the largest spectral gap for P. In fact, the simulations in Section 5 suggest otherwise.*

Remark 2. *The results obtained in [1] for the Gibbs sampler applied to a perturbation of a Gaussian measure suggest that the sharper inequalities*

$$\left(\frac{L_\star}{L^\star}\right)(1-\beta_{prop}) \leq 1-\beta \leq \left(\frac{L^\star}{L_\star}\right)(1-\beta_{prop}).$$

might hold. This seems to be an interesting question for further investigation. Moreover, it is worth mentioning that these inequalities might also apply directly to the chain and not only to a lazy version of it.

4. Application to an Elliptic Inverse Problem. The theoretical result of the previous section was motivated by studying the reconstruction of the diffusion coefficient a given noisy observations of the pressure p. We approach this inverse problem in the Bayesian framework by imposing a prior based on a series expansion with uniform coefficients.

Firstly, we will set up the forward problem and review the literature on the resulting inverse problem focusing on the Bayesian approach. Secondly, we will describe our prior, impose Gaussian observational noise and then show that the resulting posterior has a bounded density with respect to this prior. The rest of the section is devoted to constructing appropriate proposal kernels and proving a lower bound on their $L^2_{\mu_0}$-spectral gap. Thus, our main theorem implies a lower bound on the L^2_μ- spectral gaps of the corresponding Metropolis-Hastings algorithms, in particular the RRWM algorithms. Whereas the simulations in Section 5 suggest that the RRWM outperforms the IS algorithm, our main result guarantees a lower bound on the spectral gap that is of the same order. Moreover, the construction of the RRWM is important in its own right because it constitutes an efficient sampling algorithm for elliptic inverse problems.

4.1. The Underlying PDE and Well-Definedness of the Forward Model. The forward problem is based on the relation between p and a modelled by the following elliptic PDE with Dirichlet boundary conditions

$$\begin{cases} -\nabla \cdot (a \nabla p) &= g(x) \quad \text{in } D \\ p &= 0 \qquad \text{on } \partial D \end{cases} \tag{4.1}$$

where D is a bounded domain in $\mathbb{R}^d$ and p and a are scalar functions on D. We assume that $a^\star \geq a(x) \geq a_\star > 0$ for all almost every $x \in D$. The subset of $L^\infty(D)$-functions that satisfy this condition is denoted by

$$L^\infty_+ := \left\{ u \in L^\infty \,\middle|\, \operatorname*{ess\,inf}_D u > 0 \right\}.$$

If, additionally, g is in the Sobolev space H^{-1}, then the solution operator $p(x;a) : L^\infty_+ \to H^1$, mapping to the unique weak solution of (4.1), is well-defined (for details we refer the reader to [52]). We suppose that the forward operator $\mathcal{G}$, giving rise to the data, is based on the solution operator as follows

$$\mathcal{G}(a) = \mathcal{O}\left(p(\cdot; a)\right) \tag{4.2}$$

where $\mathcal{O}$ is called the observation operator. Additionally, we suppose that is is equal to $\mathcal{O} = (l_1, \ldots, l_N)$ with $l_i \in H^{-1}$.

The inverse problem associated with the above forward problem is well-known and it is particularly relevant in oil reservoir simulations and the modelling of groundwater flow, see for example [37]. A survey of classical least squares approaches to this inverse problem can be found in [28] for which recently error estimates have been obtained in [55]. A rigorous Bayesian formulation of this inverse problem with log-Gaussian priors and Besov priors is given in [15] and [13] respectively, both are reviewed in [52]. There is also an extensive literature in the uncertainty quantification community studying how uncertainty propagates through the forward model. This can be investigated by considering different realisations of the input. This approach can be combined with the finite element [18] and Galerkin methods [2] used to approximate the underlying

equation. For the elliptic problem under consideration, this has been studied in [9]. In fact, it can be more efficient to use generalised Polynomial Chaos (gPC) [49] instead of Monte Carlo methods. Recently, gPC methods also have been applied to the elliptic inverse problem considered in this article [50, 24]. Since gPC often suffers from a large constant and has only been developed for a few inverse problems, it is important to construct efficient samplers tailored for the prior and likelihood at hand. Moreover, also MCMC can be speeded up using the multi level approach. The expectation of interest is written as difference corresponding to a finer and finer discretization such that more MCMC samples are used for coarser discretisations [24].

4.2. Prior on an expansion of the Diffusion Coefficient . Following [24, 50] we choose a prior on the coefficients $(u_1, \dots, u_J)$ for $J \in \mathbb{N} \cup \{\infty\}$ giving rise to the diffusion coefficient

$$a(u)(x) = \bar{a}(x) + \sum_{j \in \mathbb{J}} \gamma_j u_j \psi_j(x) \tag{4.3}$$

where $\|\psi_i\|_{L^\infty} = 1$. We suppose that $u_i \overset{\text{i.i.d.}}{\sim} \mathcal{U}(-1, 1)$ which corresponds to a prior given by

$$\mu_0{}^J = \bigotimes_{j=1}^{J} \mathcal{U}(-1, 1.)$$

Additionally, the choice of γ_i is supposed to satisfy $a_\star = \inf \bar{a} - \sum_{j=1}^{J} \gamma_j > 0$ for all choices of J. In particular $\{\gamma_i\}$ have to be summable, $a \in L^\infty_+$ μ_0-a.s. and the solution operator p is well-defined for μ_0 almost every $a(u)$.

We would like to note that similar probability measures have been studied for the propagation of uncertainty in [9].

4.3. Bounds on the Density of the Posterior. We suppose the data is given by

$$y = \mathcal{G}(a(u)) + \eta$$

where $\eta \sim \mathcal{N}(0, \Gamma)$. The well-definedness of the corresponding posterior for $J \in \mathbb{N} \cup \{\infty\}$ has been proven in [50] and [52]. It takes the form

$$\frac{d\mu}{d\mu_0} \propto \exp\left(-\frac{1}{2} \|y - \mathcal{G}(a)\|^2_\Gamma\right).$$

We also know that

$$\|\mathcal{G}(a)\|_\Gamma \leq \|\Gamma\|_2 N \max_i \|l_i\|_{H^{-1}} \sup_{-1 \leq a_i \leq 1} \|p(a)\|_{H^1} \leq C \|\Gamma\|_2 N \max_i \|l_i\|_{H^{-1}} a_\star^{-1}$$

where $a_\star = \operatorname*{ess\,inf}_D a$. Note that C depends on N (see Equation (4.2)) but can be chosen uniformly in J. This gives rise to the following upper and lower bounds on the likelihood

$$\begin{aligned} L^\star &= 1 \\ L_\star &= \exp\left(-2C^2 \left\|\Gamma^{-1} 1\right\|_2^1 N^2 \left(\max_i \|l_i\|_{H^{-1}}\right)^2 a_\star^{-2}\right). \end{aligned}$$

4.4. Spectral Gaps for the Prior and the Posterior. In order to apply our main result, we have to choose a proposal kernel Q that is reversible and has an $L^2_{\mu_0}$-spectral gap with respect to $\mu_0 = \mathcal{U}(-1,1)^J$.

Given any kernel that has an $L^2_{\mathcal{U}(-1,1)}$-spectral gap we may apply the tensorisation property of L^2-spectral gaps (see e.g. [3, 19]) to conclude that applying this kernel to each component yields a kernel with the same spectral gap for $\mathcal{U}(-1,1)^J$. Whereas we construct the one dimensional proposal distributions explicitly below, it is worth pointing out that it is possible to obtain an appropriate one-dimensional proposal using the Metropolis-Hastings kernel for $\mathcal{U}(-1,1)$ with a one-dimensional proposal distribution. Then the resulting Markov kernel is uniformly ergodic under mild assumptions [45] implying an $L^2_{\mathcal{U}(-1,1)}$-spectral gap [44]. Note that the resulting proposal on $[-1,1]^J$ can be accepted even if some of the one-dimensional Metropolis-Hastings algorithms have rejected their proposal.

Alternatively, a Markov kernel with $L^2_{\mathcal{U}(-1,1)}$ -spectral gap can be obtained by considering a random walk with symmetric proposal

$$\begin{aligned} Q_{\mathrm{RW}}(x, dy) &= q(x-y)dy \\ Q_{\mathrm{RW}}(x, dy) &= \mathcal{L}(x+\xi), \quad \text{where } \xi \sim \tilde{q} \end{aligned}$$

and repeatedly reflecting y at the boundaries -1 and 1. The reflection can be represented according to the following function

$$R(x) = \begin{cases} y & y \leq 1 \\ 2-y & 1 < y < 3 \\ -4+y & 3 \leq y \leq 4 \end{cases}, \text{ where } y = x \bmod 4.$$

We call the Metropolis-Hastings algorithm based on tensorisations of this proposal Reflection Random Walk Metropolis (RRWM) algorithm. In this way we can write the proposal kernel as

$$Q^{\mathrm{RRWM}}(x, dy) = \mathcal{L}\left(R(x+\xi)\right) \tag{4.4}$$

where $\xi \sim \tilde{q}$. Its density with respect to the Lebesgue measure takes the form

$$q^{\mathrm{RRWM}}(x, dy) = \sum_{k \in \mathbb{Z}} \tilde{q}(x - y + 4k) + \tilde{q}(x + y + 4k + 2). \tag{4.5}$$

The proposal kernel Q_{RRWM} is reversible with respect to $\mathcal{U}(-1,1)$ because $q^{\mathrm{RRWM}}(x,y) = q^{\mathrm{RRWM}}(y,x)$. In the following we consider the RRWM with uniform random walk ($\xi \sim \mathcal{U}(-\epsilon, \epsilon)$) and with standard random walk ($\xi \sim \mathcal{N}(0, \epsilon^2)$) which we call Reflection Uniform Random Walk Metropolis (RURWM) and Reflection Standard Random Walk Metropolis (RSRWM), respectively. In contrast to the RSRWM, the proposal of the RURWM has a density $q_\epsilon^{\mathrm{RURWM}}$ with closed form. For $\epsilon < 1$ it is given by

$$q_\epsilon^{\mathrm{RURWM}}(x,y) \propto \begin{cases} 1 & -1 \leq x, y \leq 1,\ |x-y| \leq \epsilon, y > -x-2+\epsilon, y < -x+2-\epsilon \\ 2 & -1 \leq x, y \leq 1,\ y \leq -x-2+\epsilon \text{ or } y \geq -x+2-\epsilon \\ 0 & \text{otherwise} \end{cases}. \tag{4.6}$$

The following result shows that the lazy versions of the RURWM have an L^2-spectral gap of order ϵ^2.

THEOREM 4.1. *There is $c > 0$ such that the $L^2_{\mathcal{U}(-1,1)}$-spectral gap $1-\beta_\epsilon$ of $Q_\epsilon^{\mathrm{RURWM}}$ for $\epsilon \leq 1$ satisfies*

$$1 - \beta_\epsilon \geq c\epsilon^2.$$

Proof. We will first prove that for ϵ small enough there is an ϵ-independent lower bound on the $L^2_{\mu_0}$-spectral gap of $\left(Q_\epsilon^{\mathrm{RURWM}}\right)^n$ with $n(\epsilon) = \left\lceil \frac{1}{\epsilon^2} \right\rceil$. This is achieved by showing that $[-1,1]$ is a small set for $\left(Q_\epsilon^{\mathrm{RURWM}}\right)^n$. This implies uniform ergodicity and in turn a lower bound on the $L^2_{\mathcal{U}(-1,1)}$-spectral gap of $\left(Q_\epsilon^{\mathrm{RURWM}}\right)^n$. A lower bound on the $L^2_{\mathcal{U}(-1,1)}$-spectral gap of $Q_\epsilon^{\mathrm{RURWM}}$ can then be obtained using the spectral theorem. By $\tilde{q}_\epsilon(x,y) = \mathbb{1}_{(x-\epsilon,x+\epsilon)}(y)$ we denote the density of the unreflected random walk. The density $q_{\epsilon,n}^{\mathrm{RURWM}}$ of $\left(Q_\epsilon^{\mathrm{RURWM}}\right)^n$ is point-wise larger than $\tilde{q}_{\epsilon,n}$ because each y might have several preimages under R (c.f. Equation (4.6)). In order to show that $[-1,1]$ is a small set we need to obtain a uniformly lower bound on the transition density $\tilde{q}_{\epsilon,n}$. This is achieved using a local limit theorem from [43].

THEOREM 4.2. *(Theorem 13 in Section 7 of [43]) Let $\{X_n\}$ be a sequence of independent random variables having a common distribution with zero mean, non-zero variance, and finite absolute moment $\mathbb{E}\left|X_1\right|^k$ of some integer order $k \geq 3$. Let the random variable $\frac{1}{\sigma\sqrt{n}}\sum_{j=1}^n X_j$ for some $n = N$ a bounded density $p_N(x)$. Then*

$$p_n(x) = \phi(x) + \sum_{v=1}^{k-2} \frac{q_\nu(x)}{n^{\nu/2}} + o\left(\frac{1}{n^{(k-2)/2}}\right).$$

where $\phi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$ is the density of $\mathcal{N}(0,1)$.

By calculating $q_1 = 0$ we may conclude that

COROLLARY 4.3. *Let $U_n \overset{i.i.d.}{\sim} \mathcal{U}(-1,1)$ then the density of p_n of $\frac{1}{\sqrt{n/3}}\sum_{j=1}^n U_j$ satisfies*

$$p_n(x) = \phi(x) + O(\frac{1}{n})$$

We denote the probability density of $\epsilon\sum_{i=1}^n U_i$ by $\tilde{p}_n^\epsilon(x)$ which is related to p_n through

$$\tilde{p}_n^\epsilon(x) = p_n^\epsilon\left(\frac{x}{\epsilon\sqrt{\frac{1}{3}n}}\right)\frac{1}{\sqrt{\frac{1}{3}n\epsilon}}.$$

Using $n(\epsilon) = \left\lceil \frac{1}{\epsilon^2} \right\rceil$ and Corollary 4.3 we know that

$$\begin{aligned}\left|\tilde{p}_{n(\epsilon)}^\epsilon(x) - \phi\left(\frac{x}{\epsilon\sqrt{\frac{1}{3}n(\epsilon)}}\right)\frac{1}{\epsilon\sqrt{\frac{1}{3}n(\epsilon)}}\right| &\leq \frac{1}{\epsilon\sqrt{\frac{1}{3}n(\epsilon)}}\left|p_n\left(\frac{x}{\epsilon\sqrt{\frac{1}{3}n(\epsilon)}}\right) - \phi\left(\frac{x}{\epsilon\sqrt{\frac{1}{3}n(\epsilon)}}\right)\right| \\ &\leq \sqrt{3}C\frac{1}{n(\epsilon)}.\end{aligned}$$

Since p_n is symmetric and log-concave $\inf_{x\in[-2,2]} \tilde{p}_{n(\epsilon)}(x) = \tilde{p}^{\epsilon}_{n(\epsilon)}(2)$ the aim is to obtain a lower bound on $\tilde{p}^{\epsilon}_{n}(2)$. This achieved by noting

$$\phi\left(\frac{2}{\epsilon\sqrt{\frac{1}{3}n(\epsilon)}}\right)\frac{1}{\epsilon\sqrt{\frac{1}{3}n(\epsilon)}} \geq \phi\left(2\sqrt{3}\right)\frac{\sqrt{3}}{2} =: 2l.$$

For all $\epsilon \leq \epsilon_0$ small enough and hence $n(\epsilon)$ large enough

$$\left|\tilde{p}^{\epsilon}_{n(\epsilon)}(x) - \phi\left(\frac{x}{\epsilon\sqrt{\frac{1}{3}n(\epsilon)}}\right)\frac{1}{\epsilon\sqrt{\frac{1}{3}n(\epsilon)}}\right| \leq l \;\forall x.$$

Using the triangle inequality this yields a uniform lower bound on the transition kernel

$$\tilde{q}_{\epsilon,n}(x,y) \geq \tilde{p}^{\epsilon}_{n(\epsilon)}(2) \geq \phi\left(\frac{2}{\epsilon\sqrt{\frac{1}{3}n(\epsilon)}}\right)\frac{1}{\epsilon\sqrt{\frac{1}{3}n(\epsilon)}} - l \geq l \;\forall x,y \in [-1,1].$$

Therefore $q^{\mathrm{RURWM}}_{\epsilon,n}$ also satisfies

$$q^{\mathrm{RURWM}}_{\epsilon,n}(x,y) \geq \tilde{q}_{\epsilon,n}(x,y) \geq l \quad \forall x,y \in [-1,1].$$

Thus, the state space $[-1,1]$ is a small set and Theorem 8 in [45] implies that

$$\left\|\left(Q^{\mathrm{RURWM}}_{\epsilon}\right)^{n\cdot k}(d,dy) - \mathcal{U}(-1,1)\right\|_{\mathrm{TV}} \leq (1-l)^k.$$

For reversible Markov processes uniform ergodicity implies an $L^2_{\mathcal{U}(-1,1)}$-spectral gap of the same size, see for example [47]. Hence $\left(Q^{\mathrm{RURWM}}_{\epsilon}\right)^n$ has an $L^2_{\mathcal{U}(-1,1)}$-spectral gap of size 1-$\hat{\beta} = l$. The $L^2_{\mathcal{U}(-1,1)}$-spectral theorem for self-adjoint operators now implies that the $L^2_{\mathcal{U}(-1,1)}$-spectral gap of $Q^{\mathrm{RURWM}}_{\epsilon}$

$$1 - \beta_\epsilon = 1 - (1-l)^{\frac{1}{n}} \geq \frac{l}{n} \geq \frac{l}{2}\epsilon^2.$$

It is left to treat $1 \geq \epsilon > \epsilon_0$, for those ϵ we choose $n = n(\epsilon_0) = \left\lceil \frac{1}{\epsilon_0^2} \right\rceil$

$$\left|\tilde{p}^{\epsilon}_{n(\epsilon_0)}(x) - \phi\left(\frac{x}{\epsilon\sqrt{\frac{1}{3}n(\epsilon_0)}}\right)\frac{1}{\epsilon\sqrt{\frac{1}{3}n(\epsilon_0)}}\right| =$$

$$\left|p_{n(\epsilon_0)}\left(\frac{x}{\epsilon\sqrt{\frac{1}{3}n(\epsilon_0)}}\right)\frac{1}{\epsilon\sqrt{\frac{1}{3}n(\epsilon_0)}} - \phi\left(\frac{x}{\epsilon\sqrt{\frac{1}{3}n(\epsilon_0)}}\right)\frac{1}{\epsilon\sqrt{\frac{1}{3}n(\epsilon_0)}}\right| \leq \frac{\sqrt{3}\epsilon_0}{\epsilon} C \frac{1}{n(\epsilon_0)}$$

$$\leq \frac{\epsilon_0}{\epsilon} l.$$

On the other hand

$$\phi\left(\frac{2}{\epsilon\sqrt{\frac{1}{3}n(\epsilon_0)}}\right)\frac{1}{\epsilon\sqrt{\frac{1}{3}n(\epsilon_0)}} \geq \phi\left(\sqrt{3}2\frac{\epsilon_0}{\epsilon}\right)\frac{\epsilon_0}{\epsilon}\frac{\sqrt{3}}{2} \geq \frac{\epsilon_0}{\epsilon}2l.$$

Thus we know similar to the above that

$$q_{\epsilon,n}^{\mathrm{RURWM}}(x,y) \geq \tilde{q}_{\epsilon,n}(x,y) \geq \frac{\epsilon_0}{\epsilon} l \quad \forall x,y \in [-1,1].$$

And therefore know that the $L^2_{\mathcal{U}(-1,1)}$-spectral gap of $(Q_\epsilon^{\mathrm{RURWM}})^{n(\epsilon_0)}$ is bounded below by $\frac{\epsilon_0}{\epsilon} l$. Using the spectral theorem we know that the $L^2_{\mathcal{U}(-1,1)}$-spectral gap $1-\beta_\epsilon$ of $Q_\epsilon^{\mathrm{RURWM}}$ satisfies

$$1-\beta_\epsilon \geq 1 - \left(1 - \frac{\epsilon_0}{\epsilon} l\right)^{\frac{1}{n(\epsilon)}} \geq \frac{\epsilon_0}{\epsilon} l \frac{1}{n(\epsilon)} \geq \frac{\epsilon_0^3}{2\epsilon} l = \frac{\epsilon_0^3}{\epsilon^3} \frac{l}{2} \epsilon^2 \geq \epsilon_0^3 \frac{l}{2} \epsilon^2$$

□

In a similar manner it can also be shown that the proposal of the RSRWM has an $L^2_{\mu_0}$-spectral gap of order ϵ. In particular, the lower bound on the transition density of the random walk is much more straightforward.

The lower bound on the spectral gap of the resulting lazy versions of Metropolis-Hastings algorithms follows now from our main theorem.

COROLLARY 4.4. *Let Q be a Markov kernel that has an $L^2_{\mathcal{U}_{(-1,1)}}$-spectral gap $1-\beta_{prop}$, $J \in \mathbb{N} \cup \{\infty\}$ and $Q_J = \bigotimes_{j=1}^J Q(a_j, d\tilde{a}_j)$. Then the lazy version of the Metropolis-Hastings transition kernel P_J for μ_J with proposal Q_J has an $L^2_{\mu_J}$-spectral gap $1-\beta_J$ and there is a J-independent lower bound of the form*

$$1-\beta \geq \frac{1}{2} \exp\left(-8C^2 \left\|\Gamma^{-}1\right\| N^2 \left(\max_i \|l_i\|_{H^{-1}}\right)^2 a_\star^{-2}\right) (1-\beta_{prop})^2.$$

In this section, we have constructed the RRWM algorithm for the elliptic inverse problem with prior based on a series expansion with uniformly distributed coefficients. In the next section, we will compare this algorithm to the IS and RWM algorithms using simulations.

5. Numerical Comparison of Different MCMC Algorithms for a particular Elliptic Inverse Problem. In this section, we apply the Random Walk Metropolis (RWM) algorithm, the Importance Sampling (IS) and the Reflection Random Walk Metropolis (RRWM) algorithms to the posterior arising from the elliptic inverse problem considered in Section 4. We use the resulting simulations to illustrate the following two aspects:

- On the one hand the acceptance probability of the standard RWM algorithm decreases quickly as the dimension of the state space increases. On the other hand, the relation between the step size and the acceptance probability of the RRWM algorithm is not affected by the dimension.
- The performance of the IS algorithm is only affected up to a point by the dimension J of the state space. However, it does not perform well for concentrated target measures. However, choosing an appropriate step size for the RRWM algorithm leads to a good performance.

We first describe the implementation of the forward model, the choice of prior and the implementation of the IS, the RWM and the RRWM algorithms. Even-though our result only applies to the lazy version of the Metropolis-Hastings algorithm, we believe that this is artificial and present simulations for the non-lazy versions.

The remaining part of the section is then divided into presenting the dependence of the relationship between step size and acceptance rate on the dimension as well as the decay of the autocorrelation.

5.1. The Setup. We consider the elliptic inverse problem as described in Section 4 on the domain $D = [0, 1]$. In this case there is an explicit formula linking the pressure p and the diffusion coefficient a which has been implemented using a trapezoidal rule. We choose the prior as in Equation (4.3) on the coefficients u_i, that is

$$\mu_0{}^J = \bigotimes_{j=0}^{J} \mathcal{U}(-1, 1).$$

These coefficients give rise to the diffusion coefficient

$$a(u)(x) = \bar{a}(x) + \sum_{j=0}^{J} \gamma_j u_j \psi_j(x) \text{ where } a_j \overset{\text{i.i.d.}}{\sim} \mathcal{U}(-1, 1). \tag{5.1}$$

For our simulations we set

$$\begin{aligned} \bar{a}(x) &= 4.38. \\ \psi_{2j-1}(x) &= \cos(2\pi j x),\ \gamma_{2j} = \frac{1}{j^2},\ K \geq j \geq 1 \\ \psi_{2j}(x) &= \sin(2\pi j x),\ \gamma_{2j-1} = \frac{1}{j^2},\ K \geq j \geq 1 \\ \psi_0(x) &= 1,\ \gamma_0 = 1. \end{aligned}$$

Note that the lower bound $a(x) \geq 1$ is independent of $J = 2K$. Data corresponds to evaluations of the pressure uniformly spaced at distance d apart

$$y = \mathcal{G}(a^\dagger) + \eta = (p(id) + \eta_i)_{i=0}^{\lfloor 1/d \rfloor}$$

where $\eta \sim \mathcal{N}(0, \sigma^2 I)$ and $a^\dagger$ is fixed input which is generated a draw from the prior.

Subsequently, we consider the IS, RWM, RURWM and RSRWM algorithms with the following proposal kernels

$$Q^{\text{IS}}(x, dy) = \mu_0(dy) \tag{5.2}$$

$$Q_\epsilon^{\text{RWM}}(x, dy) = \mathcal{N}\left(x, \epsilon I_{d\times d}\right)(dy) \tag{5.3}$$

$$Q_\epsilon^{\text{RURWM}}(x, dy) = \otimes_{i=1}^{d} \mathcal{L}\left(R(x + \epsilon\xi)\right),\ \xi \sim \mathcal{U}(-1, 1) \tag{5.4}$$

$$Q_\epsilon^{\text{RSRWM}}(x, dy) = \otimes_{i=1}^{d} \mathcal{L}\left(R(x + \epsilon\xi)\right),\ \xi \sim \mathcal{N}(0, 1). \tag{4.4}$$

Note that the Metropolis-Hastings acceptance ratio, as described in Section 2, implies that the RWM algorithm simply rejects any proposal outside the unit cube.

5.2. Acceptance Probabilities for the RWM and RRWM Algorithms. In Figure 5.1, we have plotted the acceptance probability against the step size for the RWM, RURWM and RSRWM algorithms for different choices of K. The target for both is the posterior arising from 33 measurements with $\sigma = 0.05$ with artificial data.

The step size parameter ϵ affects the performance of all three algorithm. On the one hand large step sizes are beneficial because the algorithm can explore the state space more quickly whereas they lead to a small acceptance ratio (see Figure 5.1). On the other hand small step sizes lead to a high acceptance ratio but to highly correlated samples. The IS algorithm does not have a step size parameter and its average acceptance probability does not depend on the dimension. For this choice of parameters it is approximately 4.4%.

Figure 5.1 clearly illustrates that the acceptance probability of the RWM algorithm for a fixed step size deteriorates as the dimension increases. One reason for the decay of the acceptance probability of the RWM algorithm is that the probability of the proposal lying outside $[0,1]^d$ increases to 1 as $d \to \infty$. Moreover, there is no visible impact of the dimension on the acceptance probability for the RURWM and RSRWM algorithms.

5.3. Autocorrelation of the IS, the RWM, the RURWM and the RSRWM Algorithms. Even though our lower bound on the L^2_μ-spectral gap is smaller for the RRWM algorithms than for the IS algorithm (cf. Remark 1), the numerical results in this section suggest that the RRWM algorithms outperforms the IS algorithm especially if μ is peaked. The peakedness of μ is achieved by observing p on a fine mesh with small noise ($dx = 0.03$ and $\sigma = 0.03$).

The computational cost of both algorithms is nearly the same because the cost of computing the likelihood is more expensive than generating the proposal, which is slightly more expensive for the RRWM algorithm. Subsequently, we compare the RWM, the IS, the RURWM and the RSRWM algorithm by plotting the autocorrelation. We consider $K = 25$ ($K = 250$) corresponding to an expansion with 25 sine and 25 (250) cosine coefficients and a constant term thus giving rise to a 51 (501) dimensional problem.

In order to compare the RWM and RRWM algorithms in a fair way we choose the step size ϵ for in a way to get an acceptance rate close to 0.135. This is motivated for the RWM algorithm by the optimal scaling results in [41]. The optimality of this acceptance rate is indicated by proving that the properly rescaled samples converge to a Langevin diffusion whose time scale depends on the acceptance rate of the RWM algorithm. An acceptance rate of 0.135 corresponds to the largest time scale and thus to a quicker convergence to equilibrium of the Langevin diffusion. For the RRWM algorithms the acceptance rate is not affected by the choice of J. However, it is reasonable to choose a step size with acceptance probability bounded away from one and zero.

For the lazy version of the RRWM algorithm we know that the L^2_μ-spectral gap is bounded below and thus the asymptotic variance of the CLT (c.f. Proposition 3.3) for $f \in L^2_\mu$ is bounded above. The asymptotic variance can be related to the autocorrelation which is given by

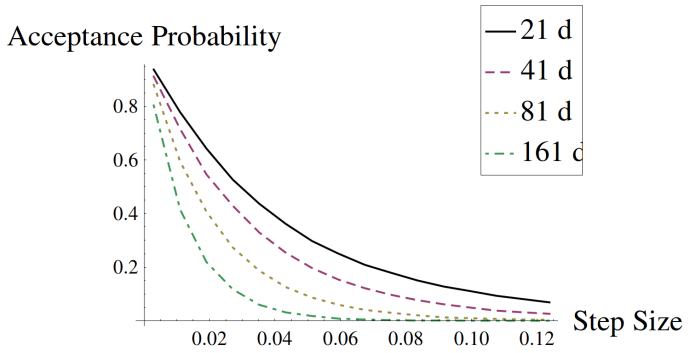
$$c_i = \mathrm{C}ov(f(X_0), f(X_i))$$

where X_i is the evolution of the corresponding Markov chain. It is well known that the asymptotic variance is equal to the integrated autocorrelation [20, 40] which is given by

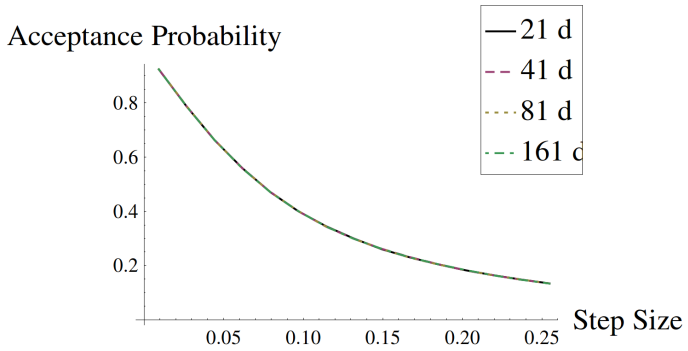
$$\sigma^2 = c_0 + 2\sum_{i=0}^{\infty} c_i.$$

We consider the Markov chain resulting from the IS, the RWM, the RURWM and the RSRWM algorithm on the state space $[-1,1]^{J+1}$. We denote by u_i the $i = 0, \dots, J$ projections onto the $i+1$-th coordinate. In the following we consider the autocorrelation for u_0 (c.f. Equation 5.1) for the algorithms mentioned above.

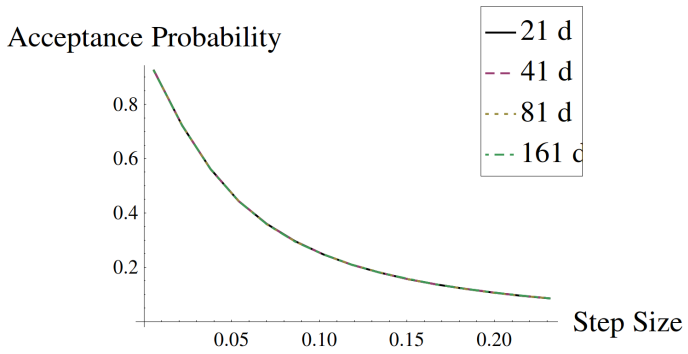
Simulations for $d = 0.1$ and $\sigma = 0.1$ are presented in Figure 5.2 which shows that the autocorrelation of the RURWM, the RSRWM and the IS algorithm is only affected up to a point by the dimension of the state space. In contrast, the autocorrelation



(a) Acceptance rate vs. step size for the RWM algorithm

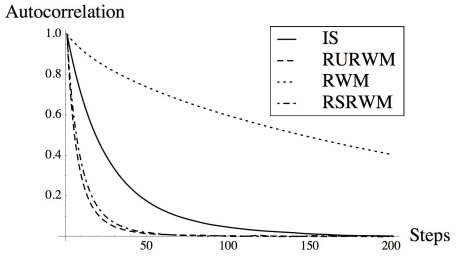


(b) Acceptance rate vs. step size for the RURWM algorithm

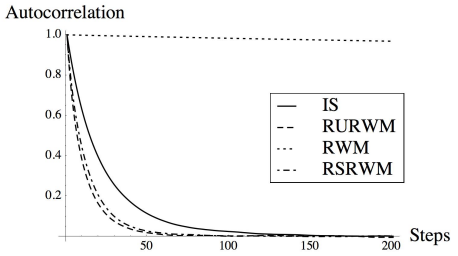


(c) Acceptance rate vs. step size for the RSRWM algorithm

Figure 5.1: Dependence of the acceptance probability on the dimension

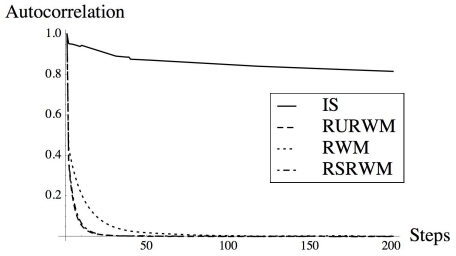


(a) 51-dimensional state space, acceptance rate: RWM 14.1%, IS 3.9%, RURWM 24.8% RSRWM 27.7%

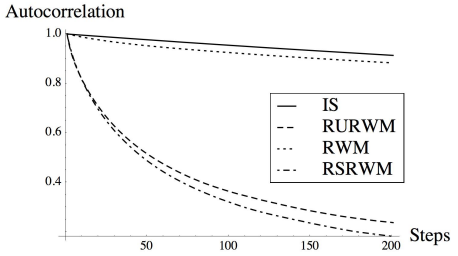


(b) 501-dimensional state space, acceptance rate: IS 4.6%, RWM 14.4%, RURWM 26.6%, RSRWM 28.7%

Figure 5.2: Autocorrelation arising from posterior for σ=0.1 and $d = 0.1$



(a) 51-dimensional state space, acceptance rate: IS 0.0001%, RWM 25.4%, RURWM 21%,RSRWM 24.6%



(b) 501-dimensional state space, acceptance rate: IS 0.04%, RWM 14.2%, RURWM 30%, RSRWM 25.5%

Figure 5.3: Autocorrelation arising from posterior for σ=0.05 and $d = 0.05$

of the RWM decays much slower for the 501 dimensional state space as for the 51 dimensional state space. In Figure 5.3, we consider the decay of the autocorrelation of the IS, the RWM, the RURWM and the RSRWM algorithms for more observations and lower observational noise ($d = 0.04$ and $\sigma = 0.03$). This has the effect that the measure concentrates in smaller regions of the state space making it harder to sample from. Figure 5.3 illustrates that the RURWM and the RSRWM algorithms can be tuned to work well for concentrated target measures such as this measure whereas the IS algorithm, even though dimension independent, behaves poorly.

For a fixed step size the RWM algorithm deteriorates as the dimension increases because the probability that one component steps outside $[-1, 1]$ converges to one. If the step size is scaled to zero appropriately the performance of the RWM algorithm deteriorates slower but for a large enough state space even the IS algorithm outperforms the RWM algorithm in any case. The reason for this is that Corollary 4.4 yields a dimension independent lower bound on the performance of the IS, RURWM and RSRWM algorithms.

6. Conclusion and Avenues of Further Research. In this article, we have shown that it is possible to transfer L^2-spectral gaps from the proposal Markov kernel to the lazy version of the Metropolis-Hastings Markov kernel. This yields theoretical

bounds for a large class of proposals for non-Gaussian measures on function spaces. Our main assumption is that the density with respect to the reference measure is bounded above and below. This is a very restrictive condition but it is difficult to prove any results in great generality under weaker assumptions. The assumption that the density is bounded above and below on bounded sets seems weak enough. Both assumptions only differ in the tails and restricting the problem to a large enough set decreases the probability of a sampling algorithm leaving it in the duration of a simulation to almost zero. But it is often the tail behaviour which prevents algorithms from satisfying the desired convergence properties, see for example [46] which describes the phenomenon for the Langevin diffusion. This effect is also described in [6], but it is not clear what impact this behaviour has on the sample average.

Our main result justifies the use of sampling methods other than the IS algorithm for the Bayesian elliptic inverse problem considered above. However, our bounds do not show that locally moving algorithms, as the RURWM and RSRWM algorithms designed in Section 4, are asymptotically better than the IS algorithm. Comparing two sampling algorithms is difficult since it depends on the specific target. Moreover, the performance also depends on the choice of the parameters for example the step size of the algorithms. Nonetheless, rigorously showing that the RURWM and RSRWM algorithms outperforms the IS algorithm, even in a special case, would be an interesting result.

Moreover, the range of the posterior density goes to infinity as the variance of the noise goes to zero. This suggests that sampling methods perform worse and worse as the observational noise goes to zero. Getting precise asymptotics of this behaviour would lead to a better understanding of the performance of sampling algorithms for Bayesian inverse problems.

As mentioned in Section 4, the proposal kernels of the RURWM and RSRWM algorithms are based on a tensorisation of Markov kernels for the uniform distribution on $[-1, 1]$. It is also interesting to consider tensorisation of Metropolis-Hastings kernels for the uniform distribution on $[-1, 1]$. Whereas we used the explicit structure of the prior, an interesting direction for more complicated priors is to use Metropolis-Hastings chains or combinations, such as tensorisation. This can lead to good proposals for another Metropolis-Hastings chain. Note that even if some of the Metropolis-Hastings algorithms in the tensorisation reject, the overall proposal can still be accepted. A deeper investigation of this approach can lead to a better understanding and guidelines for the design of efficient proposals. An interesting special case are MCMC algorithms for Bayesian inverse problems formulated on the coefficients of a Fourier series expansion. Usually the coefficients corresponding to high frequencies have only little impact on forward problem and hence the inverse problem. Developing proposals that exploit this phenomenon should also be pursued.

In this article, we considered the application of Metropolis-Hastings algorithms to the Bayesian approach to an elliptic inverse problem. A particular interesting extension would be to consider a multi-scale diffusion coefficient because there is interest in the fine and coarse scale properties of the permeability for example in subsurface geophysics. Homogenization results imply that different combinations of fine and coarse scales lead to effectively the same homogenized problem thus leading to a lack of identifiability. This also seems to be a very interesting idea.

Acknowledgements. The author would like to thank Professor Andreas Eberle, Professor Martin Hairer and Professor Andrew Stuart for helpful discussions. SJV is grateful for the support of an ERC scholarship.

REFERENCES

[1] Y. Amit. Convergence Properties of the Gibbs Sampler for Perturbations of Gaussians. *Ann. Statist.*, 24(1):122–140, 1996.

[2] I. Babuska, R. Tempone, and G. E. Zouraris. Galerkin Finite Element Approximations of Stochastic Elliptic Partial Differential Equations. *SIAM J. Numer. Anal.*, 42(2):800–825, 2004.

[3] Dominique Bakry. Functional Inequalities for Markov Semigroups. In *Probability measures on groups: recent directions and trends*, pages 91–147. Tata Inst. Fund. Res., Mumbai, 2006.

[4] J.M. Bernardo and A.F.M. Smith. *Bayesian Theory*. Wiley, 1994.

[5] A. Beskos, K. Kalogeropoulos, and E. Pazos. Advanced MCMC Methods for Sampling on Diffusion Pathspace. *Stochastic Process. Appl.*, 123(4):1415–1453, 2013.

[6] N. Bou-Rabee and M. Hairer. Non-asymptotic mixing of the MALA algorithm. *IMA J. Numer. Anal.*, (33):pp. 80–110, 2013.

[7] C. Bouman and K. Sauer. A generalized Gaussian image model for edge-preserving MAP estimation. *IEEE Trans. Image Process.*, 2(3):296–310, 1993.

[8] Steve Brooks, Andrew Gelman, Galin Jones, and Xiao-Li Meng. *Handbook of Markov Chain Monte Carlo*. Chapman and Hall/CRC, 2011.

[9] A. Cohen, R. A. Devore, and C. Schwab. Convergence rates of best N-term Galerkin approximations for a class of elliptic sPDEs. *Found. Comput. Math.*, 10(6):615–646, 2010.

[10] S. L. Cotter, M. Dashti, J. C. Robinson, and A. M. Stuart. Bayesian Inverse Problems for Functions and Applications to Fluid Mechanics. *Inverse Probl.*, 25(11):115008, 43, 2009.

[11] S. L. Cotter, G. O. Roberts, A. M. Stuart, and D. White. MCMC Methods for Functions: Modifying Old Algorithms to Make Them Faster. *ArXiv preprint 1202.0709*, 2011. to appear Stat. Sci.

[12] Giuseppe Da Prato and Jerzy Zabczyk. *Stochastic Equations in Infinite Dimensions*, volume 44 of *Encyclopedia of Mathematics and its Applications*. Cambridge University Press, Cambridge, 1992.

[13] M. Dashti, S. Harris, and A. M. Stuart. Besov Priors for Bayesian Inverse Problems. *Inverse Probl. Imaging*, 6:183–200, 2012.

[14] M. Dashti, K. J. H. Law, A. M. Stuart, and J. Voss. MAP Estimators and Posterior Consistency in Bayesian Nonparametric Inverse Problems. *ArXiv preprint 1303.4795*, 2013.

[15] M. Dashti and A. M. Stuart. Uncertainty Quantification and Weak Approximation of an Elliptic Inverse Problem. *SIAM J. Numer. Anal.*, 49:2524–2542, 2011.

[16] Persi Diaconis and Laurent Saloff-Coste. Comparison theorems for reversible Markov chains. *Ann. Appl. Probab.*, 3(3):696–730, 1993.

[17] Masatoshi Fukushima, Yoichi Oshima, and Masayoshi Takeda. *Dirichlet forms and symmetric Markov processes*, volume 19 of *de Gruyter Studies in Mathematics*. Walter de Gruyter & Co., Berlin, extended edition, 2011.

[18] Roger G. Ghanem and Pol D. Spanos. *Stochastic finite elements: a spectral approach*. Courier Dover Publications, 2003.

[19] Alice Guionnet and Boguslaw Zegarlinski. *Lectures on Logarithmic Sobolev Inequalities*, volume 1801 of *Séminaire de Probabilités, XXXVI*. Springer, 2002.

[20] O. Häggström and J. S. Rosenthal. On Variance Conditions for Markov Chain CLTs. *Electron. Comm. Probab.*, 12:454–464 (electronic), 2007.

[21] M. Hairer, A. M. Stuart, and S. J. Vollmer. Spectral Gaps for a Metropolis-Hastings Algorithm in Infinite Dimensions. *ArXiv preprint 1112.1392*, 2011.

[22] T. M. Hansen, K. S. Cordua, and K. Mosegaard. Inverse problems with non-trivial priors: Efficient solution through sequential Gibbs sampling. *Comput. Geosci.*, pages 1–19, 2012.

[23] W. K. Hastings. Monte-Carlo Sampling Methods Using Markov Chains and their Applications. *Biometrika*, 57(1):97, 1970.

[24] V. H. Hoang, C. Schwab, and A. M. Stuart. Complexity Analysis of Accelerated MCMC Methods for Bayesian Inversion. *ArXiv e-prints*, July 2012.

[25] G. L. Jones, M. Haran, B. S. Caffo, and R. Neath. Fixed-width output analysis for Markov chain Monte Carlo. *J. Amer. Statist. Assoc.*, 101(476):1537–1547, 2006. confidenceMCMC.

[26] Jari Kaipio and Erkki Somersalo. *Statistical and Computational Inverse Problems*, volume 160 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 2005.

[27] C. Kipnis and S. R. S. Varadhan. Central Limit Theorem for Additive Functionals of Reversible Markov Processes and Applications to Simple Exclusions. *Comm. Math. Phys.*, 104(1):1–19, 1986.

[28] K. Kunisch. Numerical Methods for Parameter Estimation Problems Inverse Problems in Diffusion Processes. In *Proc. GAMM-SIAM Symp. (Philadelphia, PA: SIAM)*, pages 199–

216, 1995.

[29] S Lasanen. Measurements and infinite-dimensional statistical inverse theory. *PAMM*, 1080102:1080101–1080102, 2007.

[30] S. Lasanen. Non-Gaussian Statistical Inverse Problems. Part I: Posterior Distributions. *Inverse Probl. Imaging*, 6(2):215–266, 2012.

[31] S. Lasanen. Non-Gaussian Statistical Inverse Problems. Part II: Posterior Convergence for Approximated Unknowns. *Inverse Probl. Imaging*, 6(2):267–287, 2012.

[32] Sari Lasanen. Discretizations of generalized random variables with applications to inverse problems. *Ann. Acad. Sci. Fenn. Math. Diss.*, (130):64, 2002. Dissertation, University of Oulu, Oulu, 2002.

[33] K. Łatuszyński and W. Niemiro. Rigorous Confidence Bounds for MCMC under a Geometric Drift Condition. *J. Complexity*, 27(1):23–38, 2011.

[34] K. Łatuszyński and G. O. Roberts. CLTs and Asymptotic Variance of Time-Sampled Markov Chains. *Methodol. Comput. Appl. Probab.*, pages 1–11, 2011.

[35] David Asher Levin, Yuval Peres, and Elizabeth Lee Wilmer. *Markov Chains and Mixing Times*. AMS Bookstore, 2009.

[36] László Lovász and Peter Winkler. Mixing times. *Microsurveys in discrete probability*, 41:85–134, 1998.

[37] D. McLaughlin and L. R. Townley. A Reassessment of the Groundwater Inverse Problem. *Water Resour. Res.*, 32(5):1131–1161, 1996.

[38] K. L. Mengersen and R. L. Tweedie. Rates of Convergence of the Hastings and Metropolis Algorithms. *Ann. Statist.*, 24(1):101–121, 1996.

[39] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, E. Teller, et al. Equation of State Calculations by Fast Computing Machines. *J. Chem. Phys.*, 21(6):1087, 1953.

[40] Sean Meyn and Richard L. Tweedie. *Markov Chains and Stochastic Stability*. Cambridge University Press, Cambridge, second edition, 2009. With a prologue by Peter W. Glynn.

[41] P. Neal, G. O. Roberts, and W. K. Yuen. Optimal scaling of random walk Metropolis algorithms with discontinuous target densities. *Ann. Appl. Probab.*, 22(5):1880–1927, 2012.

[42] W. Niemiro and P. Pokarowski. Fixed Precision MCMC Estimation by Median of Products of Averages. *J. Appl. Probab.*, 46(2):309–329, 2009.

[43] V. V. Petrov. *Sums of independent random variables*. Springer-Verlag, New York, 1975. Translated from the Russian by A. A. Brown, Ergebnisse der Mathematik und ihrer Grenzgebiete, Band 82.

[44] G. O. Roberts and J. S. Rosenthal. Geometric Ergodicity and Hybrid Markov Chains. *Electron. Comm. Probab*, 2:13–25, 1997.

[45] G. O. Roberts and J. S. Rosenthal. General State Space Markov Chains and MCMC Algorithms. *Probab. Surv.*, 1:20–71, 2004.

[46] G. O. Roberts and R. L. Tweedie. Exponential Convergence of Langevin Distributions and their Discrete Approximations. *Bernoulli*, pages 341–363, 1996.

[47] D. Rudolf. Explicit Error Bounds for Markov Chain Monte Carlo. *Dissertationes Math. (Rozprawy Mat.)*, 485:1–93, 2012.

[48] Byron Schmuland. Dirichlet forms: some infinite-dimensional examples. *Canad. J. Statist.*, 27(4):683–700, 1999.

[49] C. Schwab and C. J. Gittelson. Sparse Tensor Discretizations of High-Dimensional Parametric and Stochastic PDEs. *Acta Numer.*, 20:291–467, 2011.

[50] C. Schwab and A. M. Stuart. Sparse Deterministic Approximation of Bayesian Inverse Problems. *Inverse Probl.*, 28(4):045003, 32, 2012.

[51] A. M. Stuart. Inverse Problems: A Bayesian Perspective. *Acta Numer.*, 19:451–559, 2010.

[52] A. M. Stuart. The Bayesian Approach to Inverse Problems. *ArXiv preprint 1302.6989*, 2013.

[53] A. M. Stuart, P. P. Wiberg, and J. Voss. Conditional path sampling of SDEs and the Langevin MCMC method. *Commun. Math. Sci.*, 2:685–697, 2004.

[54] L. Tierney. A Note on Metropolis-Hastings Kernels for General State Spaces. *Ann. Appl. Probab.*, 8(1):1–9, 1998.

[55] L. Wang and J. Zou. Error Estimates of Finite Element Methods for Parameter Identification Problems in Elliptic and Parabolic Systems. *Discrete Contin. Dyn. Syst. Ser. B*, 14:1641–1670, 2010.

Research Article III

Posterior Consistency for Bayesian Inverse Problems through Stability and Regression Results.

Sebastian J. Vollmer, 2013. *Submitted to Inverse Problems, 38 pages.*

Posterior Consistency for Bayesian Inverse Problems through Stability and Regression Results

Sebastian J. Vollmer ‡

Mathematics Institute, Zeeman Building, University of Warwick, Coventry CV4 7AL, UK.

E-mail: Sebastian.Vollmer@stats.ox.ac.uk

Abstract. In the Bayesian approach, the a priori knowledge about the input of a mathematical model is described via a probability measure. The joint distribution of the unknown input and the data is then conditioned, using Bayes' formula, giving rise to the posterior distribution on the unknown input. In this setting we prove posterior consistency for nonlinear inverse problems: a sequence of data is considered, with diminishing fluctuations around a single truth and it is then of interest to show that the resulting sequence of posterior measures arising from this sequence of data concentrates around the truth used to generate the data. Posterior consistency justifies the use of the Bayesian approach very much in the same way as error bounds and convergence results for regularisation techniques do. As a guiding example, we consider the inverse problem of reconstructing the diffusion coefficient from noisy observations of the solution to an elliptic PDE in divergence form. This problem is approached by splitting the forward operator into the underlying continuum model and a simpler observation operator based on the output of the model.

In general, these splittings allow us to conclude posterior consistency provided a deterministic stability result for the underlying inverse problem and a posterior consistency result for the Bayesian regression problem with the push-forward prior. Moreover, we prove posterior consistency for the Bayesian regression problem based on the regularity, the tail behaviour and the small ball probabilities of the prior.

AMS classification scheme numbers: 35R30, 62C10, 62G20

Submitted to: *Inverse Problems*

‡ Present address: Department of Statistics, University of Oxford, 1 South Parks Road, Oxford OX1 3TG.

1. Introduction

Many mathematical models used in science and technology contain parameters for which a direct observation is very difficult. A good example is subsurface geophysics. The aim in subsurface geophysics is the reconstruction of subsurface properties such as density and permeability given measurements on the surface. Based on the laws of physics a forward model is designed. The forward model maps the input parameters to measurements which we will call the data. Inverting such a relationship is non-trivial and lies in the focus of the area of inverse problems. Classically, these parameters are estimated by minimisation of a regularised least squares functional which is based on the data output mismatch (Tikhonov). The idea of this approach is to use optimisation techniques aiming at parameters that produce nearly the same noiseless output as the given noisy data while being not too irregular. However, it is difficult to quantify how the noise in the data translates into the uncertainty of the reconstructed parameters for this method. The reason being that the solution is a point estimate which depends among others on the choice of the norms used. On the contrary, uncertainty quantification is much more straightforward in the Bayesian approach. The basic idea of the Bayesian method is that not all parameter choices are a priori equally likely. Instead, the parameters are artificially treated as random variables by modelling their distribution using a priori knowledge. This distribution is accordingly called the prior. For a specific forward model and given the distribution of the observational noise, the parameters and the data can be treated as jointly varying random variables. Under mild conditions, the prior can then be updated by conditioning the parameters on the data.

The posterior is one of the main tools for making inference about the parameters. Possible estimates include approximation of the posterior mean or the maximum a posteriori (MAP) estimator. Moreover, it is possible to quantify the uncertainty of the reconstructed parameter by posterior variance or posterior probability of a set around for example an estimate of the parameters under consideration.

The main focus of this article lies on posterior consistency which quantifies the quality of the resulting posterior in a thought experiment. As for any evaluation for an approach to inverse problems, an identical twin experiment is performed, that is for a fixed set of parameters and artificial data is generated. It is conceivable to expect that, under appropriate conditions, the posterior concentrates around this set of 'true' parameters. Results of this type are called posterior consistency. It justifies the Bayesian method by establishing that this method recovers the 'true' parameters sometimes with a specific rate.

So far, there are only posterior consistency results available for linear forward models and mainly Gaussian priors [30, 1, 35, 20]. In this article, we prove posterior consistency of nonlinear inverse problems with explicit bounds on the rate. The main idea behind our posterior consistency results is to use stability properties of the deterministic inverse problem to reduce posterior consistency of a nonlinear inverse problem to posterior consistency of a Bayesian non-parametric regression problem. Our

guiding example is the inverse problem of reconstructing the the diffusion coefficient from measurements in an elliptic boundary value problem. More precisely, we consider an elliptic partial differential equation (PDE) with Dirichlet boundary conditions

$$\begin{cases} -\nabla \cdot (a \nabla p) = f(x) & \text{in } D \\ p = 0 & \text{on } \partial D \end{cases} \tag{1}$$

where D is a bounded smooth domain in $\mathbb{R}^d$. In the following, we will refer to p as the pressure because Equation (1) models for instance the relationship between the permeability an pressure. The forward problem associated with Equation (1) consists in computing p given a and the source term f. The inverse problem is concerned with estimating the diffusion coefficient a, given f and noisy values or functionals of p and introduced with more detail in Section . s For this guiding example the required stability results are due to [36]. However, our methods are generally applicable to inverse problems with deterministic stability results. These are often available in the literature because they are also needed for convergence results of the Tikhonov regularisation (consider for example Theorem 10.4. in [17]). Finally, we complete our reasoning by proving appropriate posterior consistency results for the corresponding Bayesian non-parametric regression problem.

Structure of this Article

In Section 2, we both review preliminary material and give a detailed exposition of our main ideas, steps and results. In Section 3, we provide novel posterior consistency results for Bayesian non-parametric regression. In order to evaluate the rate for the regression problem, we compare our rates to those for Gaussian priors for which optimal rates are known. These results are needed in order to obtain posterior consistency for the elliptic inverse problem in Section 4. We obtain explicit rates for priors based on a series expansion with uniformly distributed coefficients. In Section 5, we draw conclusions and mention other inverse problems to which this approach is applicable. The appendix contains a detailed summary of relevant technical tools such as Gaussian measures and Hilbert scales which are used in the proofs of our main results.

Acknowledgments

The author would like to thank Professor Martin Hairer, Professor Andrew Stuart, Dr. Hendrik Weber and Sergios Agapiou for helpful discussions. SJV is grateful for the support of an ERC scholarship.

2. Preliminaries and Exposition to Posterior Consistency for Nonlinear Inverse Problems

Our crucial idea for proving posterior consistency for a nonlinear Bayesian inverse problem is the use of stability results which allow us to break it down to posterior

consistency of a Bayesian regression problem. Because the proofs are quite technical, it is worth becoming familiar with the outline of our main ideas first. Therefore this section is intended to motivate, review and summarise our investigation of posterior consistency for a nonlinear inverse problem leaving technical details to the Sections 3 and 4. For the convenience of the reader we also repeat the relevant material on Bayesian inverse problems in Section 2.1 without proofs, thus making our exposition self-contained. In Section 2.2, we precisely define posterior consistency in this setting and place it within the literature. Subsequently, we introduce an elliptic inverse problem as guiding example for which we apply our method using stability results from [36].

Finally, we conclude our exposition by giving a general abstract theorem of posterior consistency for nonlinear inverse problems with stability results in Section 2.4.

2.1. Summary of the Bayesian Approach to Inverse Problems on Hilbert Spaces

The key idea of Bayesian inverse problems is to model the input $a \in X$ of a mathematical model, for example an initial condition or a coefficient of a PDE, as random variable with distribution $\mu_0(da)$ based on a priori knowledge. This distribution is called the prior which is updated based on the observed data y. The resulting distribution μ^y is called posterior and lies in the focus of the Bayesian approach.

We assume that the data is modelled as

$$y = \mathcal{G}(a) + \xi \tag{2}$$

with $\mathcal{G}$ being the forward operator, a mapping between the Hilbert spaces X and Y, and with the observational noise ξ. The aim of the inverse problem is the reconstruction of a given the data y. Because $\mathcal{G}$ might be non-injective and ξ is unknown, the problem is not exactly solvable as stated. If the distribution of the noise ξ is known, then a and y can be treated as jointly varying random variables. Under mild assumptions on the prior, the distribution of the noise and the forward operator, there exists a conditional probability measure on a, called the posterior μ^y. It is an update of the prior using the data and models the a posteriori uncertainty. Therefore it can be viewed as the solution to the inverse problem itself. In this way it is possible to obtain different explanations of the data corresponding to different modes of the posterior.

In this article, we assume that the law of the observational noise $\mu_\xi = \mathcal{N}(0, \Gamma)$ is a mean-zero Gaussian with covariance Γ. In this case Bayes' rule can be generalised for any $\mathcal{G}$ mapping into a finite dimensional space Y. It follows that

$$\begin{aligned} \frac{d\mu^y}{d\mu_0}(a) \quad &\propto \exp\left(-\frac{1}{2}\|\mathcal{G}(a) - y\|_\Gamma^2\right) \propto \exp\left(-\frac{1}{2}\|\mathcal{G}(a)\|_\Gamma^2 + \langle y, \mathcal{G}(a)\rangle_\Gamma - \|y\|_\Gamma^2\right) \\ &\propto \exp\left(-\frac{1}{2}\|\mathcal{G}(a)\|_\Gamma^2 + \langle y, \mathcal{G}(a)\rangle_\Gamma\right). \end{aligned} \tag{3}$$

By $\|\cdot\|_\Gamma$ we denote the norm of the Cameron-Martin space $(H_{\mu_\xi}, \langle\cdot,\cdot\rangle_\Gamma)$ of μ_ξ that is the closure of Y with respect to $\langle\cdot,\cdot\rangle_\Gamma = \langle\Gamma^{-1}\cdot,\cdot\rangle$ (see Appendix A for more details).

A proper derivation of Equation (3), including the fact that its last line is also valid for functional data, and an appropriate introduction to Bayesian inverse problems can be found in [40] and [41]. All in all the Bayesian approach can be summarised as

$$\begin{aligned} &\text{Prior} && a \sim \mu_0 \\ &\text{Noise} && \xi \sim \mathcal{N}(0, \Gamma) \\ &\text{Posterior} && \frac{d\mu^y}{d\mu_0} \propto \exp\left(-\tfrac{1}{2}\|\mathcal{G}(a)\|_\Gamma^2 + \langle y, \mathcal{G}(a)\rangle_\Gamma\right). \end{aligned} \tag{4}$$

As one can see in this example, the posterior can usually only be expressed implicitly as an unnormalised density with respect to the prior. Thus, in order to estimate the input parameters or perform inference using the posterior, it has to be probed using either

- sampling methods, such as MCMC which aim at generating draws from the posterior or
- variational methods for determining the location of an infinitesimal ball with maximal posterior probability.

The second approach is also called the maximum a posteriori probability (MAP) estimator. It can be viewed as an extension to many classical methods for inverse problems. For example, it can be linked to the L^2-Tikhonov regularisation by considering a Gaussian prior and noise [11]. This relates the choice of norms in the Tikhonov regularisation to the choice of the covariance of the prior and the noise.

These regularisation techniques can be justified by convergence results. Similarly, inference methods based on the posterior can be justified by posterior consistency, a concept which we introduce in the next section.

2.2. Posterior Consistency for Bayesian Inverse Problems

As for any approach to inverse problems, the Bayesian method can be evaluated by considering an identical twin experiment. Therefore a fixed input $a^\dagger$, called the *'truth'*, is considered and data is generated using a sequence of forward models

$$y_n = \mathcal{G}_n(a^\dagger) + \xi_n$$

which might correspond to the increasing amount of data or diminishing noise. For each n we denote the posterior corresponding to the prior μ_0, the noise distribution μ_{ξ_n} and the forward operator $\mathcal{G}_n$ by μ_n^y. Under appropriate assumptions, the posterior μ_n^y is well-defined for $y = \mathcal{G}(a) + \xi$ given by Bayes' rule in Equation (4) for μ_0-a.e. a and μ_{ξ_n}-a.e. ξ_n (c.f. [41]). This Bayes' rule does not give rise to a well-defined measure for arbitrary y. However, we will pose assumptions such that the normalising constant in the Bayes' rule will be bounded above and below for every $a^\dagger$ belonging to a particular set and μ_{ξ_n}-a.e. $y = y_n = \mathcal{G}_n(a^\dagger) + \xi_n$. We will denote these posteriors by μ^{y_n}. This sequence of inverse problems is called posterior consistent if the posteriors μ^{y_n} concentrate around the 'truth' $a^\dagger$. We quantify the concentration by the posterior probability assigned to the ball B_ϵ^d. Here B_ϵ^d denotes a ball of radius ϵ with respect to a metric d.

In the following we define this concept precisely and place it within the literature before closing this section by relating posterior consistency to small ball probabilities for the prior.

Definition 1. *(Analogue to [22]) A sequence of Bayesian inverse problems* $(\mu_0, \mathcal{G}_n, \mathcal{L}(\xi_n))$ *is posterior consistent for* $a^\dagger$ *with rate* $\epsilon_n \downarrow 0$ *and with respect to a metric* d *if for*

$$y_n = \mathcal{G}_n(a^\dagger) + \xi_n,$$

there exists a constant M *and a sequence* $l_n \to 1$ *such that*

$$\mathbb{P}_{\xi_n}\left(\mu^{y_n}\left(B^d_{M\epsilon_n}(a^\dagger)\right) \geq l_n\right) \to 1. \tag{5}$$

We simply say that $(\mu_0, \mathcal{G}_n, \mathcal{L}(\xi_n))$ *is posterior consistent if the above holds for any fixed constant* $\epsilon_n = \epsilon > 0$.

Two important special cases of this definition are

- posterior consistency in the small noise limit:

$$\mathcal{L}(\xi_n) = \mathcal{L}\left(\frac{1}{\sqrt{n}}\xi\right) \text{ and } \mathcal{G}_n = \mathcal{G}$$

- posterior consistency in the large data limit:

$$\mathcal{L}(\xi_n) = \otimes_{i=1}^n \mathcal{L}(\xi) \text{ and } \mathcal{G}_n = \prod_{i=1}^n \mathcal{G}^i = (\mathcal{G}^1, \ldots, \mathcal{G}^n).$$

In the above formulation $\mathcal{G}^i$ corresponds to different measurements while $\mathcal{L}(\cdot)$ denotes the law of a random variable.

There exists a variety of results for posterior consistency and inconsistency for statistical problems. Two important examples are the identification of a distribution from (often i.i.d.) samples or density estimation [14, 22, 43, 28]. The former is concerned with considering a prior distribution on a set of probability distributions and the resulting posterior based on n samples of one of these probability distributions. In [16], Doob proved that if a countable collection of samples almost surely allows the identification of the generating distribution, then the posterior is consistent for almost every probability distribution with respect to the prior. This very general result is not completely satisfactory because it does not provide a rate and the interest may lie in showing posterior consistency for every possible truth in a certain class. Moreover, some surprisingly simple examples of posterior inconsistency have been provided for example by considering distributions on $\mathbb{N}$ [21]. The necessary bounds for posterior consistency (c.f. Equation (5)) can be obtained using the existence of appropriate statistical tests which are due to bounds on entropy numbers. These methods are used in a series of articles, for example in [22, 39, 43, 23]. This idea has also recently been applied to the Bayesian approach to linear inverse problems in [35].

In general, posterior consistency for infinite dimensional inverse problems has mostly been studied for linear inverse problems in the small noise limit where the prior is either a sieve prior, a Gaussian or a wavelet expansion with uniform distributed

coefficients [30, 1, 35, 20]. Except for [35], all these articles exploit the explicit structure of the posterior in the conjugate Gaussian setting, that means that we have a Gaussian prior as well as a Gaussian posterior.

In contrast, we consider general priors, general forward operators and Gaussian noise in this article. Usually, the posterior has a density with respect to the prior as in Equation (4). However, it is possible to provide examples where both the prior and posterior are Gaussian but not absolutely continuous. This can be achieved using for example Proposition 3.3 in [2].

Subsequently, we assume that the posterior has a density with respect to the prior implying that the posterior probability of a set is zero whenever the prior probability of this set is zero. Therefore it is necessary that $a^\dagger$ is in the support of the prior giving rise to the following definition.

Definition 2. *The support of a measure μ in a metric space (X, d) is given by*

$$supp_d(\mu) = \left\{ x \middle| \mu\left(B_\epsilon^d(x) > 0 \; \forall \epsilon > 0\right) \right\}.$$

It is natural to expect that the posterior consistency rate depends on the behaviour of $\mu_0\left(B_\epsilon^d(a^\dagger)\right)$ as $\epsilon \to 0$. Asymptotics of this type are called small ball probabilities. We recommend [32] as a good survey and refer the reader to [34] for an up-to-date list of references. In this article, we consider algebraic rates of posterior consistency, that means we take $\epsilon_n = n^{-\kappa}$ in Definition 1. In order to establish these rates of posterior consistency, we consider small ball asymptotics of the following form

$$\log(\mu_0(B_\epsilon^d(a^\dagger)) \gtrsim -\epsilon^{-\rho},$$

where $\rho > 0$ and with the notation as in Appendix A.

Both posterior consistency and the contraction rate depend on properties of the prior. This suggests that we should choose a prior with favourable posterior consistency properties. From a dogmatic point of view the prior is only supposed to be chosen to match the subjective a priori knowledge. In practice priors are often picked based on their computational performance whereas some of their parameters are adapted to represent the subjective knowledge. An example for this is the choice of the base measure and the intensity for a Dirichlet process [28].

Finally, we would like to conclude this Section by mentioning that it has been shown in [15] that posterior consistency is equivalent to the property that the posteriors corresponding to two different priors merge. The yet unpublished book [23] contains a more detailed discussion about the justification of posterior consistency studies for dogmatic Bayesians.

2.3. An Elliptic Inverse Problem as an Application of our Theory

The aim of this section is to set up the elliptic inverse problem for which we will prove posterior consistency (c.f. Section 2.2) both in the small noise and the large data limit. In a second step we describe the available stability results and how they can be used to

reduce the problem of posterior consistency of a nonlinear inverse problem to that of a linear regression problem. We end this section by stating a special case of our posterior consistency results in Section 4.

Our results do not only apply to this particular elliptic inverse problem but to any nonlinear inverse problem with appropriate stability results (c.f. Section 2.4). However, the results for the elliptic inverse problem are of particular interest because it is used in oil reservoir simulations and the reconstruction of the groundwater flow [46, 36, 27].

The forward model corresponding to our elliptic inverse problem is based on the relation between p and a given by the elliptic PDE in 1.

We would like to highlight that the relation between a and p is nonlinear. Under the following assumptions, the solution operator $p(x;a)$ to the above PDE is well-defined [24].

Assumption 1. *(Forward conditions) Suppose that*

(i) *D is compact, satisfies the exterior sphere condition (see [24]) and has a smooth boundary;*

(ii) *$a \in C^1(D) \cap C(\bar{D})$ and f is smooth in $\bar{D}$;*

(iii) *$a > a_{min} > 0$ and $f > f_{min} > 0$ in Equation (1).*

Under these assumptions, the regularity results from [24] yield the following forward stability result.

Proposition 2.1. *If a_1 and a_2 satisfy Assumption 1 and are elements of C^α for $\alpha \geq 1$, then*

$$\|p(\cdot;a_1) - p(\cdot;a_2)\|_{C^{\alpha+1}} \leq M \|a_1 - a_2\|_{C^\alpha}. \tag{6}$$

The inverse problem is concerned with the reconstruction of a given the data

$$y_n = \mathcal{G}_n(a) + \xi,$$

which is related to p in the following way.

Assumption 2. *The forward operator $\mathcal{G}$ can be split into a composition of the solution operator p and an observation operator $\mathcal{O}$, that is*

$$\mathcal{G}_n(a) = \mathcal{O}_n\left(p(\cdot;a)\right). \tag{7}$$

The Bayesian approach to the Elliptic Inverse Problem (EIP) summarises as

Model	$-\nabla \cdot (a \nabla p(\cdot;a)) = f(x)$ in D, $\quad p = 0$ on ∂D
Prior	μ_0 on a
Data	$y = \mathcal{G}_n(a) + \xi_n = \mathcal{O}_n(p(\cdot,a)) + \xi_n$, $\xi_n \sim \mathcal{N}(0, \Gamma_n)$
Posterior	$\frac{d\mu^n}{d\mu_0}(a) \propto \exp\left(-\frac{1}{2}\|\mathcal{G}_n(a)\|^2_{\Gamma_n} + \langle y, \mathcal{G}(a)\rangle_{\Gamma_n}\right)$.

(EIP)

A rigorous Bayesian formulation of this inverse problem, with log-Gaussian priors and Besov priors has been given in [12] and [10] respectively. In [38] the problem is considered

with a prior based on a series expansion with uniformly distributed coefficients (see Section 4.1.1). In the same article, a generalised Polynomial Chaos (gPC) method is derived in order to approximate posterior expectations.

We consider posterior consistency as set up in Definition 1 in the following cases:

- the small noise limit with $\mathcal{O}_n = \mathrm{Id}$ corresponding to a functional observation and an additive Gaussian random field as noise such that
$$y_n = p(\cdot; u) + \frac{1}{\sqrt{n}}\xi;$$
- the large data limit with $\mathcal{O}_n = (e_{x_i})_{i=1}^n$ where e_{x_i} are evaluations at $x_i \in D$. In this case the data takes the form
$$y_n = \{p(x_i; a)\}_{i=1}^n + \xi_n.$$

Posterior consistency in both cases are based on a stability result which can be derived by taking a as the unknown in Equation (1). This leads to the following hyperbolic PDE

$$-\nabla a \cdot \nabla p - a\Delta p = f. \tag{8}$$

Imposing Assumption 1, it has been established that there exists a unique solution a to this PDE without any additional boundary conditions:

Proposition 1 (Corollary 2 on page 220 in [36])**.** *Suppose p arises as a solution to Equation (1) with a as diffusion coefficient satisfying Assumption 1. Then Equation (8) is uniquely solvable for any $f \in L^\infty(D)$ and a such that*

$$\|a\|_\infty \leq D\left(a_{min}, f_{min}, \|\nabla a\|_\infty\right) \|p\|_\infty.$$

Moreover, if a_1 and a_2 satisfy these assumptions, then

$$\|a_1 - a_2\|_\infty \leq M \|a_1\|_{C^1} \cdot \|p(\cdot, a_1) - p(\cdot, a_2)\|_{C^2}.$$

The stability result above and a change of variables (Theorem Appendix B.1) implies

$$\mu^{y_n}(B_\epsilon^{L^\infty}(a^\dagger)) = \tilde{\mu}^{y_n}(p(B_\epsilon^{L^\infty}(a^\dagger)) \geq \tilde{\mu}^{y_n}\left(B_{\frac{\epsilon}{M}}^{C^2}(p^\dagger)\right).$$

This statement reduces posterior consistency of the EIP in L^∞ to posterior consistency of the following Bayesian Regression Problem (BRP) in C^2

Prior	$\tilde{\mu}_0 = p_\star \mu_0$ on p
Data	$y = \mathcal{O}_n(p) + \xi_n$, $\xi_n \sim \mathcal{N}(0, \Gamma_n)$
Posterior	$\frac{d\tilde{\mu}^{y_n}}{d\tilde{\mu}_0}(p) \propto \exp\left(-\frac{1}{2}\|\mathcal{O}(p)\|_{\Gamma_n}^2 + \langle y, \mathcal{O}(p)\rangle_{\Gamma_n}\right)$
	with $\mathcal{O}_n = \mathrm{Id}$ or $\mathcal{O}_n = (e_{x_i})_{i=1}^n$

(BRP)

where p is now treated as the unkown, that is the prior and the posterior are now formulated on the pressure space. Moreover, $p_\star \mu_0$ denotes the push forward of the prior

under p. Note that for $\mathcal{O}_n = \mathrm{Id}$ the BRP can also be viewed as the simplest linear inverse problem.

The required posterior consistency results for the BRP can be derived from those in Section 3 using interpolation inequalities. In this way we obtain posterior consistency results in Section 4 a special case of which is the following theorem:

Theorem 2.2 (4.1)**.** *Suppose that the prior μ_0 satisfies*

$$a(x) \geq \lambda > 0 \quad \forall x \in D \text{ and } \|a\|_{C^\alpha} \leq \Lambda \qquad \text{for } \mu_0\text{-a.e. } a \text{ and for } \alpha > 1$$

Let the noise be given by $\xi \sim \mathcal{N}\left(0,(\text{-}\Delta_{Dirichlet})^{-r}\right)$. If $\alpha > r + \frac{d}{2} - 2$ and $\alpha > r - 1$, then (EIP) is posterior consistent for any $a^\dagger \in supp_{C^\alpha}\mu_0$ in the small noise limit with respect to the $C^{\tilde{\alpha}}$-norm for any $\tilde{\alpha} < \alpha$.

This approach is not limited to the EIP as the following section shows.

2.4. Posterior Consistency through Stability Results

In Section 2.3, we present our main idea, that is the reduction of the problem of posterior consistency of the EIP to that of the BRP. The main ingredients of this reduction are the stability result that was summarised in Proposition 1 and the posterior consistency results for the BRP. This approach is not limited to the EIP but it is applicable to any inverse problem for which appropriate stability results are available. This is the case for many inverse problems such as the inverse scattering problem in [31] or the Calderon problem in [3]. We would like to point out that these stability results are also crucial for proving the convergence of regularisation methods (see Theorem 10.4 in [17]).

Theorem 2.3. *Suppose $\mathcal{G}_n = \mathcal{O}_n \circ G$ with $G : (X, \|\cdot\|_X) \to (Y, \|\cdot\|_Y)$ and $\mathcal{O}_n : (Y, \|\cdot\|_Y) \to (Z, \|\cdot\|_Z)$. Moreover, we assume that*

- *there exists a stability result of the form*

$$\|a_1 - a_2\|_X \leq b(\|G(a_1) - G(a_2)\|_Y)$$
$$\text{where } b : \mathbb{R}^+ \to \mathbb{R}^+ \text{is increasing and, } b(0) = 0;$$

- *the sequence of Bayesian inverse problems $(G_\star\mu_0, \mathcal{O}_n, \mathcal{L}(\xi_n))$ is posterior consistent with respect to $\|\cdot\|_Y$ for all $p^\dagger \in A$ with rate ϵ_n.*

Then $(\mu_0, \mathcal{G}_n, \mathcal{L}(\xi_n))$ is posterior consistent with respect to $\|\cdot\|_X$ for all $a^\dagger \in G^{-1}(A)$ with rate $b(\epsilon_n)$.

Proof. Using the notation of Section 2.3, we denote the posteriors for the Bayesian inverse problems $(\mu_0, \mathcal{G}_n, \mathcal{L}(\xi_n))$ and $(G_\star\mu_0, \mathcal{O}_n, \mathcal{L}(\xi_n))$ by μ^{y_n} and $\tilde{\mu}^y$, respectively. Then a change of variables (c.f. Theorem Appendix B.1) implies

$$\mu^y(B^X_{b(\epsilon_n)}(a^\dagger)) \geq \tilde{\mu}^y(B^Y_{\epsilon_n}(G(a^\dagger))).$$

□

3. Posterior Consistency for Bayesian Regression

As described in the previous section, for many inverse problems posterior consistency can be reduced to posterior consistency of a BRP (c.f. Section 2.4) using stability results. Thus, with the results obtained in this section we may conclude posterior consistency for apparently harder nonlinear inverse problems. For the EIP this is achieved by an application of the results in Theorem 3.3 and 3.7. Because the derivation of these two results is quite technical, we first give a summary and we recommend the reader to become familiar with both theorems but to skip the technical details on the first read.

It is classical to model the response as

$$y_n = \mathcal{O}_n(p) + \xi_n.$$

In the following we consider two Bayesian regression models with

- $\mathcal{O}_n = \mathrm{Id}$ and the noise is a Gaussian random field that is scaled to zero like $\xi_n = n^{-\frac{1}{2}}\xi$ or
- $\mathcal{O}_n = (e_{x_i})_{i=1}^n$ and $\mathcal{L}(\xi_n) = \otimes_{i=1}^n \mathcal{N}(0, \sigma^2)$ corresponding to evaluations of a function with additive i.i.d. Gaussian noise.

These models represent the large data and the small noise limit, respectively.

We prove posterior consistency for both problems under weak assumptions on the prior. This is necessary because the BRPs resulting from nonlinear inverse problems are usually only given in an implicit form. For both cases we are able to obtain a rate assuming appropriate asymptotic lower bounds on the small ball probabilities of the prior around $a^\dagger$ (see Section 2.2). Moreover, posterior consistency with respect to stronger norms can be obtained using prior or posterior regularity in combination with interpolation inequalities which is the subject of Section 3.3.

For the large data limit, that is $\mathcal{O}_n = (e_{x_i})_{i=1}^n$, we obtain posterior consistency with respect to the L^∞-norm in Section 3.2. We assume an almost sure upper bound on a Hölder norm for the prior and an additional condition on the locations of the observations. The latter is justified by construction of a counterexample.

For the small noise limit, that is $\mathcal{O}_n = \mathrm{Id}$, we prove posterior consistency with respect to the Cameron-Martin norm of the noise in Section 3.1. This norm corresponds to the $\|\cdot\|_1$-norm in the Hilbert scale with respect to the covariance operator Γ. Both the Cameron-Martin norm and Hilbert scales are introduced in Appendix A. If an appropriate $\|\cdot\|_s$-norm is μ_0-a.s. bounded, we obtain an explicit rate of posterior consistency. Otherwise, the rate is implicitly given as a low-dimensional optimisation problem. However, the condition for mere posterior consistency takes a simple form.

Corollary 3.1. *(See Corollary 3.5 for the case of general noise)*
Suppose that the noise is given by $\xi \sim \mathcal{N}(0, (-\Delta)^{-r})$ *and* $\mu_0\left(\exp(f\|p\|_{H^s}^e)\right) < \infty$ *for* $s > r + \frac{d}{2}$ *and* $f > 0$. *Then the posterior is consistent in* H^r *for any* $a^\dagger \in supp_{\mathcal{H}^r}$ *if* e *and* $\lambda = \frac{s-r-\frac{d}{2}}{s-r}$ *satisfy the following conditions*

$$e > -1 + \sqrt{8 - 8\lambda} \quad \text{if } \lambda \in \left[0, \tfrac{1}{2}\right]$$
$$e > 2 - 2\lambda \qquad \text{if } \lambda \in \tfrac{1}{2}, 1.$$

Remark 1. *If the prior is Gaussian, then the above inequality is satisfied because $e = 2$ and the RHS is less than 2 for any $\lambda \in (0, 1)$. Thus, the only remaining condition is $s > r + \frac{d}{2}$.*

Remark 2. *It is worth pointing out that for the large class of log-concave measures it is known that $e \geq 1$, for details consult [5].*

In the statistics literature regression models are mainly concerned with pointwise observations. Despite its name this is also true for *functional data analysis* (see [19]). However, the regression problem associated with $\mathcal{O}_n = \mathrm{Id}$ can be viewed as a particular linear inverse problem. As described in the introduction, this has been studied for Gaussian priors in [30] and [1]. Although our focus lies on establishing posterior consistency for general priors and non-linear models, we also obtain rates which in the special case of Gaussian priors are close to the optimal rates given in the references above.

3.1. The Small Noise Limit for Functional Response

In the following we study posterior consistency for a Bayesian regression problem assuming that the data takes values in the Hilbert space H. In particular we deal with the regression model

$$y = a + \frac{1}{\sqrt{n}}\xi \tag{9}$$

with y, a and ξ all being elements of H. Moreover, we suppose that the observational noise ξ is a Gaussian random field $\mu_\xi = \mathcal{N}(0, \Gamma)$ on H and we assume that it satisfies the following assumption.

Assumption 3. *Suppose there is $\sigma_0 \geq 0$ such that Γ^σ is trace-class for all $\sigma > \sigma_0$, that is*

$$\sum_{k=1}^{\infty} \lambda_k^{2\sigma} < \infty.$$

Imposing this assumption, it becomes possible to quantify the regularity of the observational noise in terms of the Hilbert scale defined with respect to the covariance operator (c.f. Appendix A). More precisely, this is possible due to Lemma Appendix A.2. from [1].

The regression model in Equation (9) is a special case of a general inverse problem as considered in Equation (2). Hence the corresponding posterior takes the following form (c.f. Equation (4)).

$$\frac{d\mu^y}{d\mu_0} = Z(n, \xi) \exp\left(-\frac{1}{2} n \|a\|_1^2 + n \langle a, y \rangle_1\right). \tag{10}$$

Remark 3. *We note that the above formula cannot be derived by a direct application of the Cameron-Martin formula to μ_0. Instead, it follows from writing the joint distribution as density with respect to a product measure of the prior and the noise distribution using*

the Cameron-Martin formula. Equation (10) then follows from a conditioning result, see [41] for more details.

Assuming that the data takes values in the Hilbert space H, Equation (10) can simply be derived by an application of the Cameron-Martin lemma in combination with the conditioning lemma (Lemma 5.3 in [26]). We generate data for a fixed 'truth' $a^\dagger$

$$y = a^\dagger + \frac{1}{\sqrt{n}}\xi. \tag{11}$$

By changing the normalising constant, we may rewrite the posterior in the following way

$$\frac{d\mu^y}{d\mu_0} = Z(n,\xi)\exp\left(-\frac{n}{2}\left\|a-a^\dagger\right\|_1^2 + \sqrt{n}\left\langle a-a^\dagger,\xi\right\rangle_1\right). \tag{12}$$

The normalising constant is bounded above and below for $y_n = \mathcal{G}_n(a^\dagger) + \xi_n$ for μ_{ξ_n}-a.e. ξ. In fact, this holds under weaker assumptions than needed for our results.

Lemma 3.2. *Suppose $\mu_0\left(\exp\left(f\left\|a\right\|_s^e\right)\right) < \infty$ for $s > 1+\sigma_0$ and $e > \frac{2\sigma_0}{s-1+\sigma_0}$. Then the normalising constant in Equation (12) is bounded for μ_{ξ_n}-a.s. and every $a^\dagger \in \mathcal{H}^s$ above and away form zero.*

Proof. See Appendix D. □

The expression above suggests that the posterior concentrates in balls around the truth in the Cameron-Martin norm. First, we make this fact rigorous for priors which are a.s. uniformly bounded with respect to the $\|\cdot\|_s$-norm. In a second step, we assume that the prior has higher exponential moments. Considering Gaussian priors, we show that our rate is close to the optimal rate obtained in [30].

3.1.1. Posterior Consistency for Uniformly Bounded Priors The following theorem can be viewed as a preliminary step towards Theorem 3.4 which contains our most general posterior consistency result for the Bayesian regression problem in the small noise limit. While containing our main ideas, the following result also establishes an explicit rate for posterior consistency which will be used for the EIP in Section 4.

Theorem 3.3. *Suppose that the noise satisfies Assumption 3 and*

$$\left\|a\right\|_s \leq U \ \mu_0\text{-a.s.} \tag{13}$$

for $s > 1+\sigma_0$. If $a^\dagger \in supp_{\mathcal{H}^1}(\mu_0)$ and $a^\dagger \in \mathcal{H}^s$, then μ^{y_n} is consistent in $\mathcal{H}^1$. Additionally, if the following small ball asymptotic is satisfied

$$\log(\mu_0(B_\epsilon^1(a^\dagger)) \gtrsim -\epsilon^{-\rho}, \tag{14}$$

then this holds with rate $Mn^{-\kappa}$ for any $\kappa < \min\left\{\frac{1}{2(2-\lambda)}, \frac{1}{2+\rho}\right\}$ with $\lambda = \frac{s-1-\sigma_0}{s-1}$.

Proof. Our proof is based on the observation that posterior consistency is implied by the existence of a sequence of subsets S_n such that $\mu_\xi(S_n) \to 1$ and

$$\sup_{\xi \in S_n} \frac{\mu^{y_n}(B^1_{\epsilon n^{-\kappa}}(a^\dagger)^c)}{\mu^{y_n}\left(B^1_{\epsilon n^{-\kappa}}(a^\dagger)\right)} \to 0 \text{ for } n \to \infty \tag{15}$$

where $y_n = a^\dagger + \frac{1}{\sqrt{n}}\xi$. This implication holds because

$$\mu^{y_n}(B^1_{\epsilon n^{-\kappa}}(a^\dagger)) + \mu^{y_n}(B^1_{\epsilon n^{-\kappa}}(a^\dagger)^c) = 1$$

and thus

$$\sup_{\xi \in S_n} \frac{\mu^{y_n}(B^1_{\epsilon n^{-\kappa}}(a^\dagger)^c)}{\mu^{y_n}\left(B^1_{\epsilon n^{-\kappa}}(a^\dagger)\right)} \leq \delta \quad \Rightarrow \quad \frac{1}{1+\delta} \leq \sup_{\xi \in S_n} \mu^{y_n}\left(B^1_{\epsilon n^{-\kappa}}(a^\dagger)\right) \tag{16}$$

which together with $\mu_\xi(S_n) \to 1$ implies posterior consistency, for details see Equation (5).

Fix $\gamma > 0$. Then $S_n = B^{1-\sigma_0-\gamma}_{K'_n}(0)$ with $K'_n \uparrow \infty$ as $n \to \infty$ sufficiently slow. We notice that Lemma Appendix A.2 implies that $\mathbb{P}_\xi(\xi \in B^{1-\sigma_0-\gamma}_{K'_n}(0)) \to 1$ as $n \to \infty$. The remainder of the proof will be devoted to showing that Equation (15) holds. We bound $\left\langle a - a^\dagger, \xi \right\rangle_1$ by smoothing ξ at the expense of $a - a^\dagger$

$$\begin{aligned}
\left|\left\langle a - a^\dagger, \xi \right\rangle_1\right| &\leq \left|\left\langle \Gamma^{-1+\frac{1-\sigma_0-\gamma}{2}}(a - a^\dagger), \Gamma^{\frac{\sigma_0-1+\gamma}{2}}\xi \right\rangle_1\right| \\
&\leq \left\|a - a^\dagger\right\|_{1+\sigma_0+\gamma} \|\xi\|_{1-\sigma_0-\gamma} \\
&\leq \left\|a - a^\dagger\right\|_{1+\sigma_0+\gamma} K'_n \; \forall \xi \in B^{1-\sigma_0}_{K'_n}(0).
\end{aligned}$$

Interpolating between $\mathcal{H}^1$ and $\mathcal{H}^s$ for s (c.f. Lemma Appendix A.1) yields

$$\left|\left\langle a - a^\dagger, \xi \right\rangle_1\right| \leq K'_n \left\|a - a^\dagger\right\|_1^\lambda \left\|a - a^\dagger\right\|_s^{1-\lambda} \leq K_n \left\|a - a^\dagger\right\|_1^\lambda \tag{17}$$

with $\lambda = \frac{s-1-\sigma_0-\gamma}{s-1}$. An application of Equation (12) yields the following upper bound

$$\begin{aligned}
\mu^y(B^1_{\frac{\epsilon}{n^\kappa}}(a^\dagger)) &\geq Z(n,\xi) \inf_{a \in B^1_{\frac{\epsilon}{2}n^{-\kappa}}} \exp\left(-n\|a - a^\dagger\|_1^2 - \sqrt{n}\left\langle a - a^\dagger, \xi \right\rangle_1\right) \mu_0\Big[B^1_{\frac{\epsilon}{2}n^{-\kappa}}\left(a^\dagger\right)\Big] \\
&\geq Z(n,\xi) \exp\left[-n^{1-2\kappa}\left[\frac{\epsilon\|a - a^\dagger\|_1}{2}\right]^2 - K_n n^{\frac{1}{2}-\lambda\kappa}\left(\frac{\epsilon}{2}\right)^\lambda\right] \mu_0\Big[B^1_{\frac{\epsilon}{2}n^{-\kappa}}\left(a^\dagger\right)\Big]. \qquad (18)
\end{aligned}$$

Similarly, we obtain the following upper bound

$$\mu^y\left(B^1_{\epsilon n^{-\kappa}}\left(a^\dagger\right)\right) \leq Z(n,\xi) \sup_{a \in B^1_{\epsilon n^{-\kappa}}(a^\dagger)} \exp\left(-n\|a - a^\dagger\|_1^2 + K_n\sqrt{n}\left\|a - a^\dagger\right\|_1^\lambda\right).$$

The expression in the exponential in Equation (12) can be rewritten as a function $f(d) = -nd^2 + K_n n^{\frac{1}{2}} d^\lambda$ of $d = \|a - a^\dagger\|$ which is decreasing on $[(K_n\lambda n^{-\frac{1}{2}}/2)^{\frac{1}{2-\lambda}}, \infty)$. If

$$-\frac{1}{2(2-\lambda)} < -\kappa, \tag{19}$$

then $\left\|a-a^\dagger\right\|_1^2 \in [(K_n\lambda n^{-\frac{1}{2}}/2)^{\frac{1}{2-\lambda}},\infty)$ for $a \in B^1_{\epsilon n^{-\kappa}}(a^\dagger)$ and n large enough leading to

$$\mu^y\left(B^1_{\epsilon n^{-\kappa}}(a^\dagger)\right) \leq Z(n,\xi)\exp\left(-\epsilon^2 n^{1-2\kappa} + n^{\frac{1}{2}-\kappa\lambda}\epsilon^\lambda K_n\right). \tag{20}$$

We now derive sufficient conditions for $n^{1-2\kappa}$ to be the dominant term in the exponential in the Equations (18) and (20) implying Equation (15). This is the case if, in addition to Inequality (19),

$$\begin{aligned} 1-2\kappa &> \frac{1}{2}-\kappa\lambda \text{ and} \\ \log\mu_0\left(B^1_{\frac{\epsilon n^{-\kappa}}{2}}(a^\dagger)\right) &\gtrsim -n^{1-2\kappa} \end{aligned}$$

hold. The first line is equivalent to Inequality (19) and using Inequality (14) the second line is implied by

$$1-2\kappa > \kappa\rho. \tag{21}$$

Thus, the Inequalities (19) and (21) imply that $-n^{1-2\kappa}$ is the dominant term in the Inequalities (18) and (20) establishing Equation (15). Letting $\gamma \to 0$ concludes the proof. □

3.1.2. Extension to the Case of Unbounded Priors In the following we weaken the assumptions of Theorem 3.3 by assuming that the prior has exponential moments of $\|\cdot\|_s^e$. The price we pay is that the algebraic rate of convergence is implicitly given as a low-dimensional optimisation problem.

Theorem 3.4. *Suppose that the noise satisfies Assumption 3, the prior satisfies the small ball asymptotic*

$$\log(\mu_0(B^1_\epsilon(a^\dagger)) \gtrsim -\epsilon^{-\rho}$$

and $\int \exp(3f\,\|a\|_s^e)d\mu_0(a) < \infty$ *for* $f>0$ *and* $e>0$ *for* $s > 1+\sigma_0$. *If the following optimisation problem has a solution* $\kappa^\star > 0$, *then for any* $\kappa < \kappa^\star$ *the posterior* μ^{y_n} *is consistent in* $\mathcal{H}^1$ *for* $a^\dagger$ *in* $\mathcal{H}^s$ *with rate* $n^{-\kappa}$.

Maximize κ with respect to $\kappa, p \geq 1, \eta, \theta \geq 0$ subject to

$$\begin{aligned} \frac{1}{2}+\eta\frac{p}{q}-\kappa\lambda p &< 1-2\kappa && \textit{(C.3)} \\ \frac{1}{2}-\eta+(1-\lambda)q\theta &< 1-2\kappa && \textit{(C.4)} \\ \rho\kappa &< e\theta && \textit{(C.6)} \\ \rho\kappa &< 1-2\kappa && \textit{(C.7)} \\ \lambda p &< 2 && \textit{(C.8)} \\ \left(\eta\frac{p}{q}-\frac{1}{2}\right)\frac{1}{2-\lambda p} &< -\kappa && \textit{(C.11)} \\ (1-\lambda)q &< e && \textit{(C.13)} \end{aligned}$$

$$\left(\frac{1}{2}-\eta\right)\left(1+\frac{1}{e-(1-\lambda)q}\right) < \max(1-2\kappa,\theta e) \tag{C.16}$$

where $\lambda := \frac{s-1+\sigma_0}{s-1}$.

Proof. See Appendix C. □

Remark 4. *In general,* $e(s)$ *might depend on* s *for* $\int \exp(3f\,\|a\|_s^e)d\mu_0(a) < \infty$ *to hold. Therefore the rate might be improved by optimising over different* $s > 1+\sigma_0$.

Whereas the algebraic rate in Theorem 3.4 is implicit, the following corollary yields a simple condition implying posterior consistency.

Corollary 3.5. *Suppose that the noise satisfies Assumption 3,* $a^\dagger \in \mathrm{supp}_{\mathcal{H}^1}(\mu_0)$ *and* $\int \exp(3f\,\|a\|_s^e)d\mu_0(a) < \infty$ *for* $f > 0$, $e > 0$ *and* $s > 1+\sigma_0$. *If one of the following two conditions holds*

$$0 < \lambda \le \frac{1}{2} \text{ and } \quad e > -1+2\sqrt{2}\sqrt{1-\lambda} \quad \text{or}$$
$$\frac{1}{2} < \lambda < 1 \text{and} \quad e > 2-2\lambda,$$

then μ^{y_n} *is posterior consistent for* $a^\dagger$ *in* $\mathcal{H}^s$.

Proof. It follows from the proof of Theorem 3.4 that we only have to find $\eta, \theta \ge 0$, $p \ge 1$ and s such that the Inequalities (C.3), (C.4), (C.8), (C.13) and (C.16) are satisfied. Choosing η as large as Inequality (C.3) permits, that is $\eta := \frac{1}{2(p-1)} - \epsilon$, extends the range of solutions of the other inequalities ((C.4) and (C.16)) containing η. Similarly, choosing θ as large as (C.4) permits, that is $\theta := \frac{0.5+\eta}{(1-\lambda)q} - \epsilon$, extends the range of solutions of Inequality (C.16). Letting $\epsilon \to 0$ in (C.16) yields

$$p \ge 1$$
$$\lambda p < 2 \tag{C.8}$$
$$(1-\lambda)q < e \tag{C.13}$$
$$\frac{(p-2)\left(\frac{p-1}{e(p-1)+(\lambda-1)p}+1\right)}{2(p-1)} < \max\left(1, \frac{e}{2-2\lambda}\right). \tag{22}$$

Now it is left to perform a case-by-case analysis. Starting from Inequality (22), the first two cases are $\frac{e}{2-\lambda} < 1$ and $\frac{e}{2-\lambda} \ge 1$. For these cases we have to treat $e(-1+p)+p(-1+\lambda) < 0$ and $e(-1+p)+p(-1+\lambda) \ge 0$ separately in order to rearrange Equation (22) to a quadratic inequality in p. The details are tedious but straightforward algebra. □

Remark 5. *We would like to point out that the Remarks 1 and 2 are also valid for this more general Corollary 3.5.*

3.1.3. Comparison for the Special Case of Gaussian Priors In the special case of a Gaussian prior with covariance operator that is jointly diagonalisable with noise covariance, we evaluate the consistency rate in Theorem 3.4 by comparing it with the optimal rates obtained in [30]. By numerically solving the optimisation problem in Theorem 3.4, we indicate that our rates are close to the optimal rate.

First, we introduce the assumptions on the prior and the noise covariance and state our result in this setting. In a second step, we reformulate the problem in the language of [30] and present their result. We close this section by comparing both posterior consistency rates.

We suppose that the prior is Gaussian $\mu_0 = \mathcal{N}(0, \mathcal{C}_0)$ and that the covariance operators $\mathcal{C}_0$ of the prior and Γ of the noise are jointly diagonalisable over $\{e_i\}$ denoting an orthonormal basis of eigenvectors. Furthermore, we assume that the eigenvalues μ_j^2 and λ_j^2 of $\mathcal{C}_0$ and Γ satisfy

$$\mu_j = j^{-t} \tag{23}$$

$$\lambda_j = j^{-r}, \tag{24}$$

where t and r, are the rates of the exponential decay for μ and λ, respectively. The inner product of the Hilbert scale with respect to Γ can now explicitly be written as

$$\langle x, y \rangle_r = \sum_{j=1}^{\infty} \mu_j^{-2r} x_j y_j, \quad \|x\|_r^2 = \sum_{j=1}^{\infty} \mu_j^{-2r} x_j^2.$$

Moreover, we remark that Assumption 3 is satisfied with $\sigma_0 = \frac{1}{2r}$. The covariance operator $\tilde{\mathcal{C}}_0$ of μ_0 on $\mathcal{H}^s$ has eigenvalues $\mu_j|_{\mathcal{H}^s} = j^{-t+rs}$ which can be seen by denoting $S^a e_k := k^a e_k$ and calculating

$$\begin{aligned} \mathbb{E}_{\mu_0} \langle x, u \rangle_{\mathcal{H}^s} \langle x, v \rangle_{\mathcal{H}^s} &= \mathbb{E}_{\mu_0} \left\langle x, S^{2sr} u \right\rangle_{\mathcal{H}} \left\langle x, S^{2sr} v \right\rangle_{\mathcal{H}} \\ &= \left\langle \mathcal{C}_0 S^{2sr} u, S^{2sr} u \right\rangle_{\mathcal{H}} = \left\langle S^{2sr} \mathcal{C}_0 u, v \right\rangle_{\mathcal{H}^s}. \end{aligned} \tag{25}$$

In order to conclude that $\tilde{\mathcal{C}}_0$ is trace-class on $\mathcal{H}_s$, we need to impose that $t > rs + \frac{1}{2}$. In this case, we know from Example 2 and Proposition 3 in Section 18 of [33] that the small balls asymptotic

$$\log(\mu_0(B_\epsilon^1(a^\dagger)) \gtrsim -\epsilon^{-\rho}$$

is satisfied for μ_0 with $\rho = \frac{-1}{t-r-1}$.

For this problem we adapt Theorem 3.4 by optimising over s in the appropriate range as described in Remark 4. Moreover, Fernique's theorem [4] for Gaussian measures motivates us setting $e = 2$ and $\rho = \frac{-1}{t-r-1}$ as discussed above.

Corollary 3.6. *Let the prior and the observational noise be specified as in Equation (23) and (24). If the following optimisation problem has a solution $\kappa^\star > 0$, then for any $\kappa < \kappa^\star$ the posterior μ^{y_n} is consistent in $\mathcal{H}^1$ for $a^\dagger$ in $\mathcal{H}^s$ with rate $n^{-\kappa}$.*
Maximize κ with respect to κ, $p \geq 1, \eta, \theta \geq 0, 1 + \frac{1}{2r} < s < \frac{t-\frac{1}{2}}{r}$ subject to

$$\frac{1}{2} + \eta\frac{p}{q} - \kappa\lambda p \qquad\qquad < 1 - 2\kappa$$

$$\begin{aligned}
&\frac{1}{2} - \eta + (1-\lambda)q\theta && < 1 - 2\kappa \\
&\frac{1}{t-r-1}\kappa && < 2\theta \\
&\frac{1}{t-r-1}\kappa && < 1 - 2\kappa \\
&\left(\eta\frac{p}{q} - \frac{1}{2}\right)\lambda p && < -\kappa \\
&\lambda p && < 2 \\
&(1-\lambda)q && < 2 \\
&\left(\frac{1}{2} - \eta\right)\left(1 + \frac{1}{2-(1-\lambda)q}\right) && < \max(1-2\kappa, \theta 2)
\end{aligned} \tag{26}$$

where $\lambda := \frac{s-1-\sigma_0}{s-1}$.

We now recast our problem reformulating it in the setting and notation of [30]. Letting ζ be H-valued white noise, our problem corresponds to recovering a from

$$y = a + \frac{1}{\sqrt{n}}\Gamma^{\frac{1}{2}}\zeta.$$

This problem is equivalent to

$$\tilde{Y} = Ka + \frac{1}{\sqrt{n}}\zeta \tag{27}$$

where $K = \Gamma^{\frac{1}{2}}$. Let $\{f_n\}$ be an orthonormal basis of eigenvectors of Γ on H. In order to adapt the notation of [30], we write $H_2 := H$ and note that H_1 will be equivalent to the Cameron-Martin space which takes the form

$$H_1 = S^r_{\mathcal{H}_2} := \left\{v \in \mathcal{H}_2 | v = \sum v_i f_i \text{ s.t. } \sum v_i^2 i^{2r} < \infty\right\}$$

with orthonormal basis $e_k = f_k/k^r$. Moreover, let $K : H_1 \to H_2$ be defined as

$$Ke_k := \Gamma^{-\frac{1}{2}}e_k = \frac{\lambda_k}{k^r}f_k.$$

In order to match Assumption 3.1 in [30], we have to bound the eigenvalues κ_i of K^TK as follows

$$M^{-1}i^{-p} \leq \kappa_i \leq Mi^{-p}.$$

We determine these eigenvalues by noting that

$$\left\langle K^T f_k, e_j\right\rangle_{H_2} = \langle f_k, Ke_j\rangle_{H_2} = \delta_{jk}\frac{\lambda_k}{k^r}.$$

The calculation above yields

$$K^TKf_k = \left(\frac{\lambda_k}{k^r}\right)^2 f_k$$

and thus

$$\kappa_k = \left(\frac{\lambda_k}{k^r}\right)^2 \asymp 1 = n^0 \Rightarrow p = 0.$$

As in Equation (25), we identify the covariance operator of μ_0 on H_1 through its eigenvalues

$$\tilde{\lambda}_k \asymp k^{-2t+2r}.$$

By Theorem 4.1 in [30] the posterior contraction rate is given by

$$n^{-\frac{\alpha\wedge\beta}{1+2\alpha+2p}}$$

where $-1-2\alpha = -2t+2r$ (compare Equation (3.5) in [30]) and β is the regularity of the truth. As above, we suppose that $\beta \geq \alpha$ resulting in

$$\kappa_{\text{opt}} = \frac{t-r-\frac{1}{2}}{2(t-r)-1}.$$

In Figure 1, we use numerical optimisation to compare our rate to the optimal one for $r=1$ with varying t.

Just considering Inequality (26) (essential to our approach since this implies that the Cameron-Martin term dominates the prior measure c.f. Equation (C.2)) yields an explicitly solvable optimisation problem giving rise to

$$\kappa_{\text{bound}} = \frac{t-r-1}{2(t-r)-1}.$$

The term κ_{bound} is an analytic upper bound on the solution κ_{Cor} to the optimisation problem in Corollary 3.6 because the other inequality constraints in Equation (26) are not necessarily satisfied. Thus, even if we are able to improve our bounds, there is a genuine gap between our rate and the optimal rate in the case of Gaussian priors. The reason for this gap is that Theorem 3.4 is applicable to any prior satisfying the stated regularity and small ball assumptions. Nevertheless, Figure 1 indicates that the obtained rates are quite close. In contrast, [30] is only applicable to Gaussian priors for which the Gaussian stucture of the prior and the posterior are explicitly used.

3.2. Pointwise Observations in the Large Data Limit

We consider the following non-parametric Bayesian regression problem

$$y_i = a(x_i) + \xi_i \quad i := 1, \dots, n \tag{28}$$

with $a : D \to \mathbb{R}$, D a bounded domain and $\xi_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$ representing i.i.d. mean-zero Gaussian noise with standard deviation σ. We assume that a prior μ_0 is supported on $C(D, \mathbb{R})$ resulting in a posterior of the form

$$\frac{d\mu^{y_n}}{d\mu_0} \propto \exp\left(-\sum_{i=1}^{n} \frac{(a(x_i)-y_i)^2}{2\sigma^2}\right).$$

Subsequently, we will prove posterior consistency for this problem for the case $D=[0,1]$. However, the same reasoning applies to any bounded domain $D \subseteq \mathbb{R}^d$ but the actual posterior consistency rate depends on d.

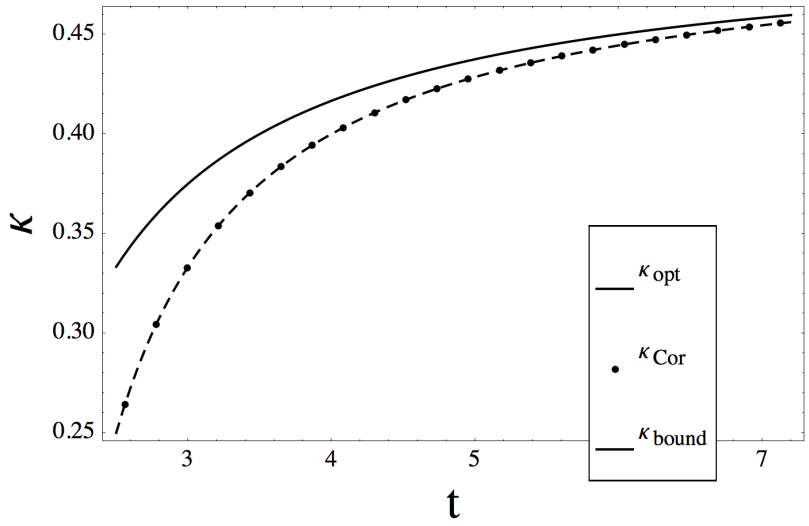


Figure 1. Posterior consistency rate for the Bayesian regression model with the noise and prior given in Equations (23) and (24). We denote the rate obtained in [30] and the one based on Corollary 3.6 as κ_{opt} and κ_{Cor}, respectively. We also plot $\kappa_{\text{bound}} = \frac{t-r-1}{2(t-r)-1}$ an upper bound on the rate that is obtainable with our method which is based on the small ball asymptotics of the prior.

As in the previous section, we suppose that the data y_i in Equation (28) is generated for a fixed 'truth' $a^\dagger$. Hence

$$\begin{aligned} y_i \quad &= a^\dagger(x_i) + \xi_i \\ \frac{d\mu^{y_{1:n}}}{d\mu_0} &\propto \exp\left(-\sum_{i=1}^{n} \frac{\left(a(x_i) - a^\dagger(x_i)\right)^2 + 2\left(a(x_i) - a^\dagger(x_i)\right)\xi_i}{2\sigma^2}\right). \end{aligned} \tag{29}$$

In this setup posterior consistency depends on the properties of the prior as well as on the sequence $\{x_i\}_{i\in\mathbb{N}}$. In the following, we discuss appropriate assumptions on both giving rise to Theorem 3.7. Moreover, we relate this result with its assumptions to the literature.

Assumption 4. *There exist $\beta \in (0,1]$ and $L > 0$ such that*

$$\|a\|_\beta \leq L \text{ and } \|a\|_\infty \leq L \quad \mu_0\text{-a.s.}.$$

As n increases, we gain more and more information about the function a. In particular, if $\{x_i\}_{i\in\mathbb{N}}$ is dense in $[0,1]$ it is even possible to reconstruct the value of $a^\dagger(x)$ from y_i. More precisely, let $x \in [0,1]$ be arbitrary, then there are $\left|x_{n_j} - x\right| \leq \frac{1}{j^{\frac{2}{\beta}}}$ such that

$$a(x) = \lim_{J\to\infty} \frac{1}{J} \sum_{j=1}^{J} a(x_{n_j}).$$

However, we will see that this is not sufficient for posterior consistency. In fact, we will give an example of posterior inconsistency for this case. So far, the problem of

posterior consistency for this type of regression problems has mainly been investigated for random evaluation points x_i which are known as random covariates. Appropriate results of this type can be found in [39, 44]. An exception is [6] where posterior consistency without a rate with respect to the L^1-norm for deterministic x_i is shown. The result of [6] is obtained under the following assumption.

Assumption 5. *Suppose that there exists a constant K such that whenever $b-a \geq \frac{1}{Kn}$ for $0 < a < b < 1$ there is at least one i such that $x_i \in (a, b)$.*

The above condition guarantees that the number of observations in each interval satisfies a lower bound proportional to its size. More precisely, an interval of size $n^{-\kappa}$ has at least order $n^{1-\kappa}$ for n being large enough. This can be seen by chopping the interval into intervals of size $\frac{1}{Kn}$. In contrast to [6], we are also able to obtain a posterior consistency rate under this assumption. Posterior consistency without a rate can be concluded under the following weaker Assumption.

Assumption 6. *We suppose that for $\{x_i\}_{i\in\mathbb{N}}$ there exists a $K > 0$ such that for any $a < b \in [0, 1]$ there is an $N(a, b)$ such that*

$$F_n(b) - F_n(a) \geq K(b - a) \quad \forall n > N(a, b)$$

where F_n denotes the empirical distribution of $\{x_i\}_{i=1}^n$.

Theorem 3.7. *Suppose that the Assumptions 4 and 6 are satisfied. Then $\mu^{y_{1:n}}$ is posterior consistent with respect to the L^∞-norm for any $a^\dagger \in supp_{C^\beta}(\mu_0)$. Moreover, if Assumption 5 and the small ball asymptotic*

$$\log(\mu_0(B_\epsilon^{L^\infty}(a^\dagger)) \gtrsim -\epsilon^{-\rho}$$

are satisfied, then $\mu^{y_{1:n}}$ is posterior consistent with respect to the L^∞-norm with any rate $n^{-\kappa}$ and

$$\kappa < \min\left\{\frac{1}{2(2+\frac{1}{\beta})}, \frac{2\beta}{(2\beta+1)(2+\rho)}\right\}.$$

Proof. As in Theorem 3.3, posterior consistency is implied by

$$\sup_{(\xi_1,\dots,\xi_n)\in S_n} \frac{\mu^{y_{1:n}}(B_{\epsilon n^{-\kappa}}^{L^\infty}(a^\dagger)^c)}{\mu^{y_{1:n}}\left(B_{\epsilon n^{-\kappa}}^{L^\infty}(a^\dagger)\right)} \to 0 \text{ for } n \to \infty$$

for increasing sets S_n such that $\mu_\xi((\xi_1, \dots, \xi_n) \in S_n) \to 1$. For notational convenience we write $h := a - a^\dagger$, $S := \sqrt{\sum_{i=1}^n h(x_i)^2}$ and we denote by η a generic $\mathcal{N}(0, \sigma^2)$ random variable. This allows us to rewrite the posterior in Equation (29) as

$$\frac{d\mu^{y_{1:n}}}{d\mu_0} \propto Z(n, \eta) \exp\left(-\frac{1}{2\sigma^2}(S^2 + 2S\eta)\right). \tag{30}$$

Since $y_{1:n}$ is finite dimensional, it is easy to see that $Z(n, \eta)$ is bounded from above and below. Again, fixing $\gamma > 0$, we only need to consider $\eta \in B_{n^\gamma}(0)$. Thus, for $0 < l_\epsilon < 1$ we have a lower bound on

$$\mu^{y_{1:n}}\left(B_\epsilon^{L^\infty}\left(a^\dagger\right)\right) \geq Z(n, \eta) \exp\left(-n\frac{l_\epsilon^2\epsilon^2}{2\sigma^2} - \frac{l_\epsilon \epsilon n^{\frac{1}{2}} n^\gamma}{\sigma^2}\right) \mu_0\left(B_{l_\epsilon\epsilon}^{L^\infty}\left(a^\dagger\right)\right).$$

In order to derive an upper bound on $\mu^{y_{1:n}}(B_\epsilon^{L^\infty}(a^\dagger)^c)$, let $a \in B_\epsilon^{L^\infty}(a^\dagger)^c$ be chosen arbitrarily and notice that $f(S) = -S^2 + Sn^\gamma$ is decreasing for $S > n^\gamma$. The upper bound on $\mu^{y_{1:n}}(B_\epsilon^{L^\infty}(a^\dagger)^c)$ therefore boils down to a lower bound on S that is larger than n^γ. In fact, there is $\hat{x}$ such that $\left|a^\dagger(\hat{x}) - a(\hat{x})\right| \geq \epsilon$. Applying Hölder continuity to $a \in B_\epsilon^{L^\infty}(a^\dagger)^c$ yields

$$\left|a^\dagger(x) - a(x)\right| \geq \epsilon/2 \quad \text{for } x \in (\hat{x} - \Delta x, \hat{x} + \Delta x]$$

for $\Delta x = \left(\frac{\epsilon}{4L}\right)^{\frac{1}{\beta}}$. Let I be the following index set

$$I = \{i | x_i \in (\hat{x} - \Delta x, \hat{x} + \Delta x]\}.$$

For n larger than $N_\epsilon = \max\{N(i\Delta x, (i+1)\Delta x) | i = 0 \dots \lfloor 1/\Delta x \rfloor - 1\}$ it follows that

$$K\frac{1}{2}\Delta x \leq F_n(\hat{x} + \Delta x) - F_n(\hat{x} - \Delta x) = \frac{|I|}{n}.$$

If we only consider x_i with $i \in I$, we obtain that

$$S \geq \sqrt{\frac{\epsilon^2}{4} n K \frac{\Delta x}{2}}$$

which gives rise to the following upper bound

$$\mu^{y_{1:n}}\left(B_\epsilon^{L^\infty}\left(a^\dagger\right)^c\right) \leq Z(n,\xi) \exp\left[-\frac{n\epsilon^2 K \Delta x}{8\sigma^2} + \left(\frac{K}{2}\Delta x\right)^{\frac{1}{2}} \frac{\epsilon}{4\sigma^2} n^{\frac{1}{2}+\gamma}\right]. \quad (31)$$

By choosing l_ϵ small enough, we also know that

$$\frac{\mu^{y_{1:n}}\left(B_\epsilon^{L^\infty}(a^\dagger)^c\right)}{\mu^{y_{1:n}}\left(B_\epsilon^{L^\infty}(a^\dagger)\right)} \to 0 \quad \text{as } n \to \infty.$$

In order to obtain a rate of posterior consistency, we use $\tilde{\kappa} > \kappa$ and hence

$$\mu^{y_n}\left(B_{n^{-\kappa}}^{L^\infty}\left(a^\dagger\right)\right) \geq \mu^{y_n}\left(B_{n^{-\tilde{\kappa}}}^{L^\infty}\left(a^\dagger\right)\right) \geq Z(n,\xi) \exp\left[-\frac{1}{2\sigma^2} n^{1-2\tilde{\kappa}} - \frac{1}{2\sigma^2} n^{\frac{1-\tilde{\kappa}}{2}+\gamma} - c n^{\tilde{\kappa}\rho}\right] (32)$$

Thus, Equation (31) implies that

$$\mu^{y_{1:n}}\left(B_{n^{-\kappa}}^{L^\infty}\left(a^\dagger\right)^c\right) \leq Z(n,\xi) \exp\left(-\frac{K}{8\sigma^2(4L)^{\frac{1}{\beta}}} n^{1-(2+\frac{1}{\beta})\kappa} + \frac{K^{\frac{1}{2}} n^{\frac{1}{2}-\kappa(1+\frac{1}{2\beta})}}{2\sigma^2\sqrt{2}(4L)^{\frac{1}{\beta}}}\right). \quad (33)$$

The first term in the exponential in Equation (32) is dominant over the corresponding term in Equation (33) by choosing

$$\tilde{\kappa} := \kappa(1 + \frac{1}{2\beta}) + \gamma.$$

Moreover, the first term in the Equations (32) and (33) is dominant over the other terms respectively if

$$\begin{aligned} 1 - 2\tilde{\kappa} &> \tilde{\kappa}\rho \\ 1 - 2\tilde{\kappa} &> \frac{1-\tilde{\kappa}}{2} + \gamma \\ 1 - \left(2 + \frac{1}{\beta}\right)\kappa &> \frac{1}{2} - \kappa\left(1 + \frac{1}{2\beta}\right). \end{aligned}$$

These three inequalities are respectively implied by

$$\begin{aligned}\frac{1-\rho\gamma-2\gamma}{\left(1+\frac{1}{2\beta}\right)(2+\rho)} &> \kappa\\ \frac{2-10\gamma}{6\left(1+\frac{1}{2\beta}\right)} &> \kappa\\ \frac{1}{2+\frac{1}{\beta}} &> \kappa.\end{aligned}$$

Choosing γ small enough, we see that μ^{y_n} is consistent in L^∞ with any rate

$$\kappa < \min\left\{\frac{1}{3(1+\frac{1}{2\beta})}, \frac{1}{\left(1+\frac{1}{2\beta}\right)(2+\rho)}\right\}.$$

□

The assumptions in the theorem above can be justified because a slight violation leads to the example of posterior inconsistency in the next section.

3.2.1. Example of Posterior Inconsistency In this section, we construct a counterexample to illustrate that despite the strong Assumption 4 it is not sufficient for $\{x_i\}_{i=1}^n$ to be dense in order to establish posterior consistency. Given such a sequence it is always possible to extract a subsequence satisfying Assumption 6. Even though all the other observations can be viewed as additional, we will choose a prior so that the posterior sequence is not consistent.

In the following, we choose the prior concentrated on functions g which are continuous, satisfy $g(\frac{1}{2}) = 0$ and are linear on $[0, \frac{1}{2}]$ and $[\frac{1}{2}, 1]$. By identifying $g(0)$ and $g(1)$ with the first and second component respectively the following two-dimensional example can be extended to the setting of Equation (28). This extension is an example of posterior inconsistency with respect to the L^p-norm for $1 \le p \le \infty$ because any of these norms is equivalent to $\|(g(0), g(1))\|$ for an arbitrary norm on $\mathbb{R}^2$.

Example 1. *We consider the following prior on* $\mathbb{R}^2$

$$\mu_0 = M\sum_{k=1}^{\infty} \delta_{\left(\frac{1}{\sqrt{k}},0\right)} \exp\left(-2k^2\right) + \delta_{\left(\frac{1}{2\sqrt{k}},1\right)} \exp\left(-k^2\right)$$

and we choose $a^\dagger = (0,0)$ *as 'truth'. The data consists of* n *and* n^θ *with* $0 < \theta < 1$ *being measurements of the form*

$$y_i = a_1^\dagger + \xi_i \text{ and } \tilde{y}_i = a_2^\dagger + \tilde{\xi}_i \text{ with } \xi_i, \tilde{\xi}_i \overset{i.i.d.}{\sim} \mathcal{N}(0,1),$$

respectively. Consequently, the posterior takes the form

$$\mu^{(y_{1:n}, \tilde{y}_{1:n^\theta})} \propto \mu_0(da_1, da_2) \exp\left(-\frac{1}{2}\sum_{j=1}^{n}(a_1 - \xi_j)^2 - \frac{1}{2}\sum_{j=1}^{n^\theta}\left(a_2 - \tilde{\xi}_j\right)^2\right).$$

Here, posterior consistency of $\mu^{(y_{1:n},\tilde{y}_{1:n^\theta})}$ *is equivalent to the statement that for any* K *there is* $l_n \uparrow 1$ *such that*

$$\mathbb{P}_\eta\left(\mu^{(y_{1:n},\tilde{y}_{1:n^\theta})}\left(\bigcup_{k=K}^{\infty}(\frac{1}{\sqrt{k}},0)\right)\geq l_n\right)\to 1 \text{ for } n\to\infty.$$

We will not only show that $\mu^{(y_{1:n},\tilde{y}_{1:n^\theta})}$ *is posterior inconsistent but also that there is* $l_n \downarrow 0$ *such that for* $A=\bigcup_{k=1}^{\infty}(\frac{1}{k},0)$

$$\mathbb{P}_\eta\left(\mu^{(y_{1:n},\tilde{y}_{1:n^\theta})}(A)\leq l_n\right)\to 1 \text{ for } n\to\infty.$$

Because of $\mu^{(y_{1:n},\tilde{y}_{1:n^\theta})}(A)+\mu^{(y_{1:n},\tilde{y}_{1:n^\theta})}(A^c)=1$*, we may proceed as in the proofs of the Theorems 3.3 and 3.7 and thus it is enough to construct sets of increasing* $\mathbb{P}_\xi$*-probability such that on these sets*

$$\begin{aligned}\frac{\mu^{(y_{1:n},\tilde{y}_{1:n^\theta})}(A)}{\mu^{(y_{1:n},\tilde{y}_{1:n^\theta})}(A^c)}&=\frac{\sum_k\exp\left(-\frac{1}{2}\sum_{j=1}^{n}\left(\frac{1}{\sqrt{k}}-\xi_j\right)^2-\frac{1}{2}\sum_{j=1}^{n^\theta}\tilde{\xi}^2-2k^2\right)}{\sum_k\exp\left(-\frac{1}{2}\sum_{j=1}^{n}\left(\frac{1}{2\sqrt{k}}-\xi_j\right)^2-\frac{1}{2}\sum_{j=1}^{n^\theta}\left(\tilde{\xi}_j-1\right)^2-k^2\right)}\\&=\frac{\sum_k\exp\left(-\frac{1}{2}\sum_{j=1}^{n}\frac{1}{k}-\frac{2}{k}\xi_j-2k^2\right)}{\sum_k\exp\left(-\frac{1}{2}\sum_{j=1}^{n}\frac{1}{4k}-\frac{1}{\sqrt{k}}\xi_j-\frac{1}{2}\sum_{j=1}^{n^\theta}1-2\tilde{\xi}_k-k^2\right)}\to 0.\end{aligned}$$

The $\mathbb{P}_\xi$*-probabilities of* $|\sum_{j=1}^{n}\xi_j|>Mn^{\frac{1}{2}+\gamma}$ *and* $|\sum_{j=1}^{n^\theta}\tilde{\xi}_j|>Mn^{\frac{1}{2}+\gamma}$ *are exponentially small in* n*. Thus, it is enough to consider*

$$\begin{aligned}\frac{\sum_k\exp\left(-\frac{1}{2}n\frac{1}{k}-2k^2\right)}{\sum_k\exp\left(-\frac{1}{2}n\frac{1}{4k}-k^2\right)}&=\frac{\sum_k^{\sqrt{n}}\exp\left(-\frac{1}{2}n\frac{1}{k}-2k^2\right)+\sum_{\sqrt{n}}^{\infty}\exp\left(-\frac{1}{2}n\frac{1}{k}-2k^2\right)}{\sum_k^{\sqrt{n}}\exp\left(-\frac{1}{2}n\frac{1}{4k}-k^2\right)+\sum_{\sqrt{n}}^{\infty}\exp\left(-\frac{1}{2}n\frac{1}{4k}-k^2\right)}\\&\leq\max\left\{\exp\left(-\frac{3}{4}\sqrt{n}\right),\exp(-n)\right\}\to 0 \text{ as } n\to\infty.\end{aligned}$$

Hence we have shown that μ^{y_n} *is not posterior consistent.*

This example relies on the prior having strong correlations between its two components. Therefore it seems an interesting question how the assumptions on μ_0 can be strengthened in order to relax those on $\{x_i\}$.

3.3. Convergence in Stronger Norms

We conclude this section by showing that interpolation inequalities can be used in order to strengthen the norm in which the posterior concentrates. In particular we consider the small noise limit as described in Section 3.1.

Suppose we know that the posterior concentrates around the truth $a^\dagger$ in the Cameron-Martin norm $\|\cdot\|_1$. In order to show consistency in $\|\cdot\|_r$, we write

$$\begin{aligned}\left\{\left\|a-a^\dagger\right\|_r>\epsilon\right\}&\subset\left\{\left\|a-a^\dagger\right\|_1^{\lambda}\left\|a-a^\dagger\right\|_s^{1-\lambda}>\epsilon\right\}\\&\subset\left\{\left\|a-a^\dagger\right\|_1^{\lambda}>\frac{\epsilon}{K}\right\}\cup\left\{\left\|a-a^\dagger\right\|_s^{1-\lambda}>K\right\}.\end{aligned}$$

The posterior probability of the first set is small due to the posterior consistency in $\mathcal{H}^1$. The posterior probability of the second set is small due to the tails of the prior and the posterior. Obtaining estimates of this type can be done similarly to the steps subsequent to Equation (C.12) in the proof of Theorem 3.4. Using this technique, it is also possible to apply the results of this section to the EIP in the next section. A similar technique based on interpolation inequalities between Hlder spaces applies to the large data limit and is also used for the EIP.

4. Posterior Consistency for an Elliptic Inverse Problem

In Section 2.3, we introduced the idea of reducing posterior consistency of the EIP to that of the BRP. For this example we demonstrate our method for both the small noise and the large data limit. We start by giving the proof for the small noise limit in detail before sketching the same steps for the large data limit. We emphasise the case of posterior consistency in the small noise limit because of its analogy with convergence results for regularisation methods.

4.1. Posterior Consistency in the Small Noise Limit

Using Theorem 3.3 from Section 3 to conclude posterior consistency of the EIP (c.f. Section 2.3) is not entirely straightforward because we have to lift the posterior consistency for the BRP to C^2. Moreover, we have to find appropriate assumptions on the prior μ_0 so that the push forward prior $p_\star\mu_0$ satisfies the assumptions of the Theorem 3.3. Again, a rate of posterior consistency is obtained if the prior satisfies appropriate small ball asymptotics. In a second step we verify those for the so-called uniform priors which are based on a series expansion with uniformly distributed coefficients, for details see below or consider [38, 29, 41].

In order to formulate assumptions on μ_0 implying that $p_\star\mu_0$ satisfies the assumptions of Theorem 3.3, we assume for simplicity that $\xi \sim \mathcal{N}(0, (-\Delta_{\text{Dirichlet}})^{-r})$ where $\Delta_{\text{Dirichlet}}$ denotes the Laplacian with homogeneous Dirichlet conditions. In this case the abstract Hilbert scale $\mathcal{H}^s$ (c.f. Appendix A) corresponds to the standard Sobolev space H^{rs}. Thus, the almost sure bounds in Theorem 3.3 are implied by the appropriate assumptions on the prior and classical results from [18, 24].

Moreover, the choice $\xi \sim \mathcal{N}(0, (-\Delta_{\text{Dirichlet}})^{-r})$ also implies that Assumption 3 holds for $\sigma_0 = \frac{d}{2r}$. This is due to the fact that the operator $(-\Delta_{\text{Dirichlet}})^{-r}$ has eigenvalues λ_k^2 with $\lambda_k \asymp k^{-2r/d}$ (see Section Appendix A for notation) where d denotes the dimension of the domain D. These results are called Weyl asymptotics and further details can be found in [45] and [37].

The following theorem summarises the consequences for the posterior consistency of the EIP.

Theorem 4.1. *Suppose that the noise is given by* $\xi \sim \mathcal{N}(0, (-\Delta_{Dirichlet})^{-r})$ *and that the*

prior μ_0 satisfies

$$a(x) \geq \lambda > 0 \, \forall x \in D \text{ and } \|a\|_{C^\alpha} \leq \Lambda \qquad \text{for } \mu_0\text{-a.e. } a \quad \text{and for } \alpha > 1.$$

If $\alpha > r + \frac{d}{2} - 2$, $\beta + 1 > r$ and $a^\dagger \in supp_{C^\beta} \mu_0$, then the EIP is posterior consistent with respect to the $C^{\tilde{\alpha}}$-norm for any $\tilde{\alpha} < \alpha$. Additionally, if

$$\log(\mu_0(B_\epsilon^{C^\beta}(a^\dagger)) \gtrsim -\epsilon^{-\rho}$$

then the EIP is posterior consistent with respect to the L^∞-norm with rate $n^{-\kappa}$ for any κ such that

$$\kappa < \left(\frac{\alpha}{\alpha + 2 + \frac{d}{2} - r} \wedge 1 \right) \left(\frac{1}{2 + \rho} \wedge \frac{\alpha}{2\left(\alpha + 1 + \frac{d}{2r}\right)} \right).$$

Before proving Theorem 4.1, we notice that forward stability results (as Proposition 2.1) can be used to transfer small ball asymptotics from μ_0 to $\tilde{\mu_0} = p_\star \mu_0$.

Lemma 4.2. *If the prior satisfies the small ball asymptotic*

$$\log\left(\mu_0(B_\epsilon^{C^\beta}(a^\dagger)\right) \gtrsim -\epsilon^{-\rho},$$

then

$$\log\left(\tilde{\mu_0}(B_\epsilon^{C^{\beta+1}}(p^\dagger)\right) \gtrsim -\epsilon^{-\rho}.$$

Proof. Proposition 2.1 implies that

$$\begin{aligned} \left\|a - a^\dagger\right\|_{C^\beta} \qquad &\leq \epsilon \Rightarrow \left\|p - p^\dagger\right\|_{C^\beta} \leq M\epsilon \\ \text{and } p\left(B_\epsilon^{C^\beta}\left(a^\dagger\right)\right) &\subseteq B_{C\epsilon}^{C^{\beta+1}}\left(p^\dagger\right). \end{aligned}$$

Hence the statement follows. □

Having established Lemma 4.2, we are now in the position to prove the main theorem of this section.

Proof of Theorem 4.1. Subsequently, M will denote a generic constant in different contexts that may change form line to line. We will first prove posterior consistency in L^∞ before we use an interpolation inequality to bootstrap it to $C^{\tilde{\alpha}}$. In order to prove posterior consistency in the L^∞-norm, it is enough to show posterior consistency of the BRP in the C^2-norm because

$$\mu^{y_n}(B_\epsilon^{L^\infty}(a^\dagger)) = \tilde{\mu}^{y_n}(p(B_\epsilon^{L^\infty}(a^\dagger)) \geq \tilde{\mu}^{y_n}\left(B_{\frac{\epsilon}{M}}^{C^2}(p^\dagger)\right) \tag{34}$$

which follows by an application of Proposition 1 and a change of variables (see Theorem Appendix B.1). Using Theorem 6.19 from [24], we may conclude that

$$\|p\|_{H^{\alpha+2}} \lesssim \|p\|_{C^{\alpha+2}} \leq K \ \tilde{\mu}_0\text{-a.s.}.$$

Since $\alpha + 2 > r$, p is $\tilde{\mu}_0$-a.s. an element of the Cameron-Martin space of μ_ξ as it corresponds to H^r. Posterior consistency of the BRP with respect to the H^r-norm is now implied by Theorem 3.3. Its conditions are satisfied because

$$\|p\|_{\mathcal{H}^s} \leq M\Lambda \quad \tilde{\mu}_0\text{-a.s.}$$

with

$$s = \frac{\alpha+2}{r} > 1 + \frac{d}{2r} = 1 + \sigma_0.$$

Furthermore, Proposition 2.1 and the fact that $a^\dagger \in \text{supp}_{C^\beta}\mu_0$ imply that $p(a^\dagger) \in \text{supp}_{H^r}\tilde{\mu_0}$. In order to bootstrap to posterior consistency in the C^2-norm, we use a generalisation of the Sobolev embedding theorem for Besov spaces and an interpolation inequality between Besov spaces on domains (for details consult [42]). We first note that $B^\tau_{22} = H^\tau$ and $C^\tau = B^\tau_{\infty\infty}$ for $\tau \notin \mathbb{Z}$. In particular Theorem 4.33 in [42] implies that

$$\|g\|_{B^{r-\frac{d}{2}-\gamma}_{\infty\infty}} \leq M\|g\|_{H^r}$$

for $\gamma > 0$ being small. If $r > \frac{d}{2}+2$, we can conclude posterior consistency in the C^2-norm because

$$B^{C^2}_{\frac{\epsilon}{M}}(p^\dagger) \supseteq \left\{p | \|p-p^\dagger\|_{B^{r-\frac{d}{2}-\gamma}_{\infty\infty}} \leq \frac{\epsilon}{M}\right\} \supseteq \left\{p | \|p-p^\dagger\|_{H^r} \leq \frac{\epsilon}{M}\right\} \tag{35}$$

holds for γ small enough. Otherwise, we use the interpolation inequality between Besov spaces subject of Theorem 4.17 in [42]

$$\|g\|_{C^{2+\gamma}} \leq \|g\|^\theta_{B^{r-\frac{d}{2}-\gamma}_{\infty\infty}} \|g\|^{1-\theta}_{B^{\alpha+2}_{\infty\infty}} \tag{36}$$

for γ small enough and with $\theta = \frac{\alpha}{\alpha+2+\frac{d}{2}-r+\gamma}$. Similar to Equation (35), it follows that

$$B^{C^2}_{\frac{\epsilon}{M}}(p^\dagger) \supseteq \left\{p | \|p-p^\dagger\|^\theta_{B^{r-\frac{d}{2}-\gamma}_{\infty\infty}} \leq \frac{\epsilon}{KM}\right\} \supseteq \left\{p | \|p-p^\dagger\|_{H^r} \leq \frac{\epsilon^{\theta^{-1}}}{M}\right\}.$$

The Equations (35) and (37) allow us to bootstrap the posterior consistency of $\tilde{\mu}^{y_n}$ to C^2. Equation (34) implies posterior consistency of μ^{y_n} in the L^∞-norm. Similarly, we bootstrap to posterior consistency in $C^{\tilde{\alpha}}$ for $\tilde{\alpha} < \alpha$ using the same interpolation technique as above.

In order to obtain a rate for posterior consistency, we first note that $\log(\mu_0(B^{C^\beta}_\epsilon(a^\dagger)) \gtrsim -\epsilon^{-\rho}$ implies

$$\log(\tilde{\mu_0}(B^r_\epsilon(a^\dagger)) \gtrsim -\epsilon^{-\rho}$$

due to Lemma 4.2. Now Theorem 3.3 implies posterior consistency of the sequence of posteriors $\tilde{\mu}^{y_n}$ in H^r with any rate κ such that

$$\kappa < \frac{1}{2+\rho} \wedge \frac{\alpha+1}{2\left(\alpha+1+\frac{d}{2r}\right)}.$$

Using the interpolation inequality as above gives rise to posterior consistency for$\tilde{\mu}^{y_n}$ in $C^{2+\alpha}$ with rate $n^{-\kappa}$ for any κ such that

$$\kappa < \left(\frac{\alpha}{\alpha+2+\frac{d}{2}-r} \wedge 1\right)\left(\frac{1}{2+\rho} \wedge \frac{\alpha}{2\left(\alpha+1+\frac{d}{2r}\right)}\right).$$

As above, this implies the same rate of posterior consistency for μ^{y_n} in L^∞.

□

4.1.1. Uniform Prior In this section, we establish a rate of posterior consistency for the EIP with the so-called uniform prior introduced in [41, 29, 38]. This choice of the prior was motivated by the preceding analysis in the uncertainty quantification literature, see for instance [7, 8]. It is given by

$$\mu_0 = \mathcal{L}\left(a_0(x) + \sum_{i=1}^{\infty} \gamma_i z_i \psi_i(x)\right) \quad z_i \overset{\text{i.i.d.}}{\sim} \mathcal{U}[-1,1] \tag{37}$$

where $\mathcal{L}$ denotes the law of a random variable. Moreover, we suppose that $\|\psi_i(x)\|_{C^\beta} = 1$, $\gamma_i > 0$ and $S = \sum_{i=1}^{\infty} \gamma_i < \infty$ such that

$$0 < a_{\min} \leq a \leq a_{\max} \quad \mu_0\text{-a.s..}$$

In order to obtain a rate for the EIP with this prior, we derive a small ball asymptotic under an appropriate assumption on the decay of $\{\gamma_i\}$.

Assumption 7. *There exists $\nu^\star \in (0,1)$ such that for all $\nu > \nu^\star$*

$$S_\nu = \left(\sum_{i=1}^{\infty} \gamma_i^\nu\right)^{\frac{1}{\nu}} < \infty.$$

Since the series in Equation (37) is absolutely convergent, we assume without loss of generality that γ_i is decreasing. This allows us to use the following classical inequality from approximation theory [13]

$$\left(\sum_{n>N} \gamma_n\right) \leq N^{1-\frac{1}{\nu}} S_\nu. \tag{38}$$

Lemma 4.3. *Suppose that μ_0 is given as in Equation (37), Assumption 7 is satisfied with $\nu^\star$ and*

$$a^\dagger = \sum_{i=1}^{\infty} \gamma_i z_i^\dagger \psi_i(x) \quad \textit{where } z_i^\dagger \in [-1,1].$$

Then for any $\nu > \nu^\star$

$$\log \mu_0(B_\epsilon^{C^\beta}(a^\dagger)) \gtrsim -\epsilon^{-\frac{1}{\frac{1}{\nu}-1}}.$$

Proof. We obtain an asymptotic lower bound on the small ball probability by choosing an appropriate subset $D_\epsilon(a^\dagger)$ of $B_\epsilon^{C^\beta}(a^\dagger)$. We denote a generic element of this set by

$$a = \sum_{i=1}^{\infty} \gamma_i z_i \psi_i(x).$$

Choosing N_ϵ such that $\sum_{i=N_\epsilon}^{\infty} \gamma_i \leq \frac{\epsilon}{2}$, the corresponding terms contribute at most $\frac{\epsilon}{2}$ to the difference $\left\|a^\dagger - a\right\|$. The subset $D_\epsilon(a^\dagger)$ prescribes intervals for z_i $i = 1 \dots N_\epsilon$ such

that this contribution is at most $\frac{\epsilon}{2}$, too. More precisely, let $\nu^\star < \tilde{\nu} < \nu$, then Equation (38) implies

$$\sum_{n>N_\epsilon} \gamma_n \leq N_\epsilon^{1-\frac{1}{\nu}} S_{\tilde{\nu}} \leq \frac{\epsilon}{2}$$

for $N_\epsilon \geq \left(\frac{2S_{\tilde{\nu}}}{\epsilon}\right)^{\frac{1}{\frac{1}{\nu}-1}}$. Let the subset $D_\epsilon(a^\dagger) \subseteq B_\epsilon^{C^\beta}(a^\dagger)$ be given by

$$\left\{a \middle| \left(z_i^\dagger > 0 \wedge z_i^\dagger - \frac{\epsilon}{2S} \leq z_i \leq z_i^\dagger\right) \vee \left(z_i^\dagger \leq 0 \wedge z_i^\dagger + \frac{\epsilon}{2S} \geq z_i \geq z_i^\dagger\right)\ 1 \leq i \leq N_\epsilon\right\}.$$

Then

$$\begin{aligned} \mu_0\left(B_\epsilon^{C^\beta}(a^\dagger)\right) &\geq \left(\frac{\epsilon}{2S}\right)^{N_\epsilon} \\ \log \mu_0\left(B_\epsilon^{C^\beta}(a^\dagger)\right) &\gtrsim N_\epsilon \log \epsilon \gtrsim -\epsilon^{-\frac{1}{\nu-1}}. \end{aligned}$$

□

Combining Lemma 4.3 and Theorem 4.1 results in the following theorem which characertises posterior consistency for this class of priors.

Theorem 4.4. *Let the prior μ_0 be defined as in Equation (37) and let Assumption 7 be satisfied. Additionally, we assume that $\alpha \geq \beta \geq r+1$, $\alpha > r + \frac{d}{2} - 2$ and $\|a\|_\alpha \leq K$ $\mu_0 - a.s.$. Then the posterior μ^{y_n} is consistent for any*

$$a^\dagger = \sum_{i=1}^{\infty} \gamma_i z_i^\dagger \psi_i(x) \quad \textit{where } z_i^\dagger \in [-1,1]$$

with respect to the L^∞-norm *with rate $\epsilon_n = M(\kappa) n^{-\kappa}$ for any κ such that*

$$\kappa < \left(\frac{\alpha}{\alpha + 2 + \frac{d}{2} - r} \wedge 1\right)\left(\frac{1-\nu}{2-\nu} \wedge \frac{\alpha - r + 2}{2\alpha + d - 2r + 4}\right).$$

4.2. Posterior Consistency in the Large Data Limit

In the following we show that the results for the BRP can be transferred to posterior consistency results in the large data limit for the EIP. We consider only the case $d = 1$ with $D = [0,1]$ as the general case is similar. Furthermore, assuming that the observations are of the form

$$y_i = p(x_i; a) + \xi \quad i = 1 \dots n,$$

the sequence of posteriors is given by

$$\frac{d\mu^{y_n}}{d\mu_0}(a) \propto \exp\left(-\sum_{i=1}^{n} \frac{(p(a)(x_i) - y_i)^2}{2\sigma^2}\right).$$

Posterior consistency of the EIP in L^∞ can then be derived on the basis of Theorem 3.7.

Theorem 4.5. *Suppose that the sequence $\{x_i\}$ satisfies Assumption 6, $\|a\|_{C^\gamma} \leq L$ μ_0-a.s. with $\gamma > 1$ and $a \geq a_{min}$ μ_0-a.s.. If $a^\dagger \in supp_{C^\gamma}\mu_0$, then the EIP is posterior consistent in the large data limit with respect to $C^{\tilde{\gamma}}$ for any $\tilde{\gamma} < \gamma$.*

Proof. An application of Theorem 6.13 in [24] yields the existence of $M(D, \gamma)$ so that for all a satisfying

$$\|a\|_{C^\gamma} \leq S \text{ and } a \geq a_{min}$$

there is a unique solution p such that $\|p\|_{C^{2+\gamma}} \leq M$. Thus, $\tilde{\mu}_0 = p_\star \mu_0$ satisfies the assumptions of Theorem 3.7 implying that $\tilde{\mu}^{y_{1:n}}$ is posterior consistent in L^∞. Using the interpolation inequality between L^∞ and $C^{2+\gamma}$, we also obtain consistency in C^2. As in Theorem 4.1, Proposition 1 can be used in order to conclude posterior consistency of $\mu^{y_{1:n}}$ in L^∞. We can bootstrap from $L^\infty(D)$ to $C^{\tilde{\gamma}}$ by interpolating between L^∞ and C^γ. □

5. Concluding Remarks

In this article, we have established a novel link between stability results for an inverse problem and posterior consistency for the Bayesian approach to it. We have explicitly shown this link for an elliptic inverse problem (c.f. EIP) but the same method is also applicable for the general case. An instance is electrical impedance tomography (Caldern problem) for which stability results are available [3]. This example would lead to a very slow posterior consistency rate since its stability results are weak. Essentially, we would have to redo all the calculations on a log-scale instead of an algebraic scale.

So far, we need exponential moments of the prior for the Bayesian regression of functional response and for pointwise observations (see also Section 4.2.2 in [6]). For this reason it is harder to prove posterior consistency for example for log-Gaussian priors. Log-Gaussian measures have moments of arbitrary order but no exponential moments. This is a problem that we would like to pursue further in the future.

Appendix A. Notation and Review of Technical Tools

Appendix A.1. Asymptotic Inequalities

We use the following notation for asymptotic inequalities:

Let a_n and b_n be sequences in $\mathbb{R}$. We denote by $\mathbb{R}$ $a_n \lesssim b_n$ that there are $N \in \mathbb{N}$ and $M \in \mathbb{R}$ such that $a_n \leq M b_n$ for $n \geq N$. Moreover, if $a_n \lesssim b_n \lesssim a_n$, we write $a_n \asymp b_n$.

Appendix A.1.1. Hilbert Scales In order to measure the smoothness of the noise and samples of the prior, we introduce Hilbert scales following [17]. Let Γ be a self-adjoint, positive-definite, trace-class linear operator with eigensystem (λ_k^2, ϕ_k). We know that Γ^{-1} is a densely defined, unbounded, symmetric and positive-definite operator because

$$H = \overline{\mathcal{R}(\Gamma)} \oplus \mathrm{Ker}(\Gamma)^\perp = \overline{\mathcal{R}(\Gamma)}.$$

We define the *Hilbert scale* by $((\mathcal{H}^t, \langle\cdot,\cdot\rangle_t))_{t\in\mathbb{R}}$ with $\mathcal{H}^t := \overline{\mathcal{M}}^{\|\cdot\|_t}$ for

$$\begin{aligned}\mathcal{M} &:= \bigcap_{n=0}^{\infty} \mathcal{D}(\Gamma^{-n}) \\ \langle u, v\rangle_t &:= \left\langle \Gamma^{-\frac{t}{2}} u, \Gamma^{-\frac{t}{2}} v \right\rangle \\ \|u\|_t &:= \left\| \Gamma^{-\frac{t}{2}} u \right\|.\end{aligned}$$

We will denote balls with respect to the $\|\cdot\|_t$-norm by

$$B_R^t(u) = \{x | \, \|u - x\|_t \leq R\}.$$

Moreover, these collection of norms satisfies an interpolation inequality

Proposition Appendix A.1. *(Proposition 8.19 in [17]) Let $q < r < s$ then the following interpolation inequality holds*

$$\|x\|_r \leq \|x\|_q^{\frac{s-r}{s-q}} \|x\|_s^{\frac{r-q}{s-q}}.$$

Remark. *Our definition here is slightly different from the literature in order to match it to the Sobolev spaces for $\Gamma = (-\Delta_{Dirichlet})^{-1}$*

Appendix A.2. Gaussian Measures

In this section, we set out our notation for some standard results about infinite dimensional Gaussian measures which can be found in the following textbooks and lecture notes [4, 9, 25]. Let γ be a Gaussian measure on a Hilbert space $(H, \langle\cdot,\cdot\rangle)$. It is characterised by its *mean* given by the Bochner integral

$$m = \int_H x \, d\gamma(x)$$

and the *covariance operator* $\Gamma : H \to H$ characterised by the relation

$$\langle Cu, v\rangle = \int \langle u - m, x\rangle \langle v - m, x\rangle \, d\gamma(x).$$

From this it is clear that the covariance operator is positive-definite and self-adjoint. Moreover, we note that Γ is necessarily trace-class and the Gaussian can be expressed through eigenvalues λ_k^2 and the corresponding eigenbasis ϕ_k

$$\gamma = \mathcal{L}\left(m + \sum_{i=1}^{\infty} \lambda_k \phi_k \xi_k\right) \text{ with } \quad \xi_k \overset{\text{i.i.d}}{\sim} \mathcal{N}(0,1).$$

The *Cameron-Martin space* associated with γ is

$$H_\gamma = \left\{ x | x = \sum x_i \phi_i \text{ s.t. } \sum \frac{1}{\lambda_i^2} x_i^2 < \infty \right\} \subset H$$

equipped with the inner product

$$\langle x, y\rangle_\gamma = \sum \frac{1}{\lambda_i^2} x_i y_i$$

where $x = \sum x_i \phi_i$ and $y = \sum y_i \phi_i$.This space characterises the support as well as the direction such that

$$T_{h\star}\gamma \ll \gamma$$

where T_h is the translation operator $T_h(x) = x + h$.

We also consider the Hilbert scale $(\mathcal{H}^s, \|\cdot\|_s)$ generated by γ and the regularity of a draw $\zeta \sim \gamma$ can be expressed as follows.

Lemma Appendix A.2. *([1]) Imposing Assumption 3 the following statements hold:*

(i) Let ζ be a white noise, then $\mathbb{E}\|\Gamma^{\frac{\sigma}{2}}\zeta\| < \infty$ for all $\sigma > \sigma_0$.

(ii) Let $u \sim \mu_0$, then $u \in \mathcal{H}^{1-\sigma}$ μ_0-a.s. for every $\sigma > \sigma_0$.

Appendix B. Change of Variables for the Posterior

The state of a model can be described in several ways. In this section, we present the resulting relationship between two different descriptions of the same model.

Theorem Appendix B.1. *Suppose $\mathcal{G}_n = \mathcal{O}_n \circ G$ with $G : (X, \|\cdot\|_X) \to (Y, \|\cdot\|_Y)$ and $\mathcal{O} : (Y, \|\cdot\|_Y) \to (Z, \|\cdot\|_Z)$. Furthermore, assume that the posterior $\mu^{y_n}(\tilde{\mu}^y)$ is well-defined for the forward operator $\mathcal{G}_n(\mathcal{O}_n)$, the prior $\mu_0(da)$ $(\tilde{\mu}_0(dp))$ and the noise $\xi \sim \mathcal{N}(0, \Gamma)$. It is given by*

$$\begin{aligned}\frac{d\mu^{y_n}}{d\mu_0}(a) &\propto \exp\left(-\frac{1}{2}\|\mathcal{G}(a)\|_\Gamma^2 + \langle y, \mathcal{G}(a)\rangle_\Gamma\right)\\ \frac{d\tilde{\mu}^y}{d\tilde{\mu}_0}(p) &\propto \exp\left(-\frac{1}{2}\|\mathcal{O}(p)\|_\Gamma^2 + \langle y, \mathcal{O}(p)\rangle_\Gamma\right).\end{aligned}$$

In this case $G_\star \mu^{y_n} = \tilde{\mu}^y$.

Proof. It is sufficient to show that both measures agree on all sets $A \in \mathcal{B}(Y)$

$$(G_*\mu^{y_n})(A) = \int_A 1 dG_*\mu^{y_n}(da).$$

By the transformation rule

$$\begin{aligned}(G_*\mu^{y_n})(A) = \int_{G^{-1}(A)} 1 d\mu^{y_n}(a) &= \int_{G^{-1}(A)} \mathrm{c}\cdot \exp\left(-\frac{1}{2}\|\mathcal{O}(G(va)) - y\|_\Gamma^2\right)\mathrm{d}\mu_0(a)\\ &= \int_A c\cdot \exp\left(-\frac{1}{2}\|\mathcal{O}(v) - y\|_\Gamma^2\right) dG_*\mu_0(v).\end{aligned}$$

□

Appendix C. Proof of Theorem 3.4

Proof of Theorem 3.4. We follow the same steps as in the proof of Theorem 3.3 up to Equation (17) reading

$$\left|\left\langle a-a^\dagger,\xi\right\rangle_1\right| \leq K_n \left\|a-a^\dagger\right\|_1^\lambda \left\|a-a^\dagger\right\|_s^{1-\lambda} 17$$

with $\lambda = \frac{s-1-\sigma_0-\gamma}{s-1}$. We now separate the product using Young's inequality with $\frac{1}{p}+\frac{1}{q}=1$

$$\begin{aligned}\left|\left\langle a-a^\dagger,\xi\right\rangle_1\right| &\leq \left(K_n\left(2^{(\lambda-1)q}qn^{-\eta}\right)^{-\frac{1}{q}}\left\|a-a^\dagger\right\|_1^\lambda\right)\left(\left(2^{(\lambda-1)q}qn^{-\eta}\right)^{\frac{1}{q}}\left\|a-a^\dagger\right\|_s^{1-\lambda}\right)\\ &\leq \tilde{K_n}\left\|a-a^\dagger\right\|_1^{\lambda p} + n^{-\eta}\left(\frac{1}{2}\left\|a-a^\dagger\right\|_s\right)^{(1-\lambda)q} \qquad \text{(C.1)}\end{aligned}$$

where, for simplicity of notation, we used $\tilde{K}_n := \frac{K_n^p(2^{-(1-\lambda)q}qn^{-\eta})^{-\frac{p}{q}}}{p}$.

Lower bound on $\mu^{y_n}(B^1_{\epsilon n^{-\kappa}}(a^\dagger))$: The following lower bound on $\mu^{y_n}(B^1_{\epsilon n^{-\kappa}}(a^\dagger))$ is based on Equation (C.1)

$$\begin{aligned}\mu^{y_n}\left(B^1_{\epsilon n^{-\kappa}}(a^\dagger)\right) &\geq \mu^{y_n}\left(B^1_{\frac{\epsilon}{2}n^{-\kappa}}(a^\dagger)\cap B^s_R(0)\right) \geq Z(n,\xi)\mu_0\left(B^1_{\frac{\epsilon n^{-\kappa}}{2}}\left(a^\dagger\right)\cap B^s_R(0)\right)\\ &\quad\cdot\exp\left[-\frac{n^{1-2\kappa}}{2}\frac{\epsilon^2}{4} - n^{\frac{1}{2}-\lambda p\kappa}\tilde{K}_n\left(\frac{\epsilon}{2}\right)^{\lambda p} - n^{\frac{1}{2}-\eta}\left[R^{(1-\lambda)q}+\|a^\dagger\|_s^{(1-\lambda)q}\right]\right] \qquad \text{(C.2)}\end{aligned}$$

The term $n^{1-2\kappa}$ has to be dominant in Equation (C.2) because the same exponent is appearing in Equation (C.10) except for a larger coefficient. Choosing $R = n^\theta$ and substituting the expression for $\tilde{K}_n$, this is the case if

$$1-2\kappa > \frac{1}{2}+\eta\frac{p}{q}-\kappa\lambda p \qquad \text{(C.3)}$$

$$1-2\kappa > \frac{1}{2}-\eta+(1-\lambda)q\theta \qquad \text{(C.4)}$$

$$\log\mu_0\left(B^1_{\frac{\epsilon n^{-\kappa}}{2}}\left(a^\dagger\right)\cap B^s_R(0)\right) \gtrsim n^{1-2\kappa}. \qquad \text{(C.5)}$$

We need small ball probabilities and the exponential moments of μ_0 in order to obtain explicit sufficient conditions on κ. We first note that

$$\mu_0\left(B^1_{\frac{\epsilon n^{-\kappa}}{2}}\left(a^\dagger\right)\cap B^s_R(0)\right) \geq \mu_0\left(B^1_{\frac{\epsilon n^{-\kappa}}{2}}\left(a^\dagger\right)\right) - \mu_0\left(B^s_R(0)^c\right).$$

Equation (C.5) holds if

$$\rho\kappa < e\theta \qquad \text{(C.6)}$$

$$\rho\kappa < 1-2\kappa. \qquad \text{(C.7)}$$

Upper bound on $\mu^{y_n}\left(B^1_{\epsilon n^{-\kappa}}(a^\dagger)^c\right)$*:* We bound $\mu^{y_n}\left(B^1_{\epsilon n^{-\kappa}}\left(a^\dagger\right)^c\right)$ by

$$\mu^{y_n}\left(B^1_{\epsilon n^{-\kappa}}(a^\dagger)^c\right) \leq \mu^{y_n}\left(B^1_{\epsilon n^{-\kappa}}(a^\dagger)^c \cap B^s_R(0)\right) + \mu^{y_n}\left(B^1_{\epsilon n^{-\kappa}}(a^\dagger)^c \cap B^s_R(0)^c\right).$$

Upper bound on $\mu^{y_n}\left(B^1_{\epsilon n^{-\kappa}}(a^\dagger)^c \cap B^s_R(0)\right)$*:* We denote by $M_{B^1_{\epsilon n^{-\kappa}}(a^\dagger)^c \cap B^s_R(0)}$ the following supremum

$$\sup_{B^1_{\epsilon n^{-\kappa}}(a^\dagger)^c \cap B^s_R(0)} -\frac{n}{2}\left\|a-a^\dagger\right\|_1^2 + \sqrt{n}\tilde{K}_n\left\|a-a^\dagger\right\|_1^{\lambda p} + n^{\frac{1}{2}-\eta}\left(\frac{1}{2}\left\|a-a^\dagger\right\|_s\right)^{(1-\lambda)q}$$

which is finite if

$$\lambda p < 2. \tag{C.8}$$

The first two summands above can be rewritten as a function f of $\left\|a-a^\dagger\right\|_1$ where

$$f(d) = -\frac{n}{2}d^2 + \sqrt{n}\tilde{K}_n d^{\lambda p}.$$

By considering f', we see that f is decreasing for $d \geq (\tilde{K}_n \lambda p n^{-\frac{1}{2}})^{\lambda p}$. Thus, for

$$\epsilon n^{-\kappa} \geq (\tilde{K}_n \lambda p n^{-\frac{1}{2}})^{\lambda p} \tag{C.9}$$

the following inequality holds

$$\begin{aligned}&\mu^{y_n}\left(B^1_{\epsilon n^{-\kappa}}(a^\dagger)^c \cap B^s_R(0)\right)\\ &\leq Z(n,\xi)\exp\left[-\frac{n^{1-2\kappa}}{2}\epsilon^2 + n^{\frac{1}{2}-\lambda p\kappa}\tilde{K}_n\epsilon^{\lambda p} + \quad n^{\frac{1}{2}-\eta}\left(R^{(1-\lambda)q} + \|a^\dagger\|_s^{(1-\lambda)q}\right)\right].\end{aligned} \tag{C.10}$$

Then for large n, Equation (C.9) is implied by

$$\left(\eta\frac{p}{q} - \frac{1}{2}\right)\lambda p < -\kappa. \tag{C.11}$$

Upper bound on $\mu^{y_n}\left(B^1_{\epsilon n^{-\kappa}}(a^\dagger)^c \cap B^s_R(0)^c\right)$*:* In this section, we bound $\mu^{y_n}(B^1_\epsilon(a^\dagger)^c \cap B^s_R(0)^c)$ using Markov's inequality in combination with the exponential moments of the prior

$$\begin{aligned}\mu^{y_n}\left(\exp(f\,\|\cdot\|_s^e)\chi_{B^1_{\epsilon n^{-\kappa}}(a^\dagger)^c}\right) \leq &\int\limits_{B^1_{\epsilon n^{-\kappa}}(a^\dagger)^c} C(n,\xi)\exp\left(n^{\frac{1}{2}-\eta}\left\|a\right\|_s{}^{(1-\lambda)q}\right)\\ &\exp\left(-\frac{n}{2}\left\|a-a^\dagger\right\|_1^2 + \sqrt{n}\tilde{K}_n\left\|a-a^\dagger\right\|_1^{\lambda p} + n^{\frac{1}{2}-\eta}\left\|a^\dagger\right\|_s^{(1-\lambda)q}\right)d\mu_0(a).\end{aligned} \tag{C.12}$$

We denote the term appearing in the exponential in the second line by T_0. It can be bounded similar to the upper bound on $\mu^{y_n}\left(B^1_{\epsilon n^{-\kappa}}(a^\dagger)^c \cap B^s_R(0)\right)$

$$T_0 \leq \mathfrak{U}_{T_0} := -\frac{n^{1-2\kappa}}{2}\epsilon^2 + n^{\frac{1}{2}-\lambda p\kappa}\tilde{K}_n\epsilon^{\lambda p} + n^{\frac{1}{2}-\eta}\left\|a^\dagger\right\|_s^{(1-\lambda)q}.$$

We denote by $\lesssim$ an inequality with a multiplicative constant not involving n or κ. In order to get an upper bound for Equation (C.12), we bound the exponential moment by

$$\mu^{y_n}\left(\exp(f\left\|\cdot\right\|_s^e)\chi_{B^1_{\epsilon n^{-\kappa}}(a^\dagger)^c}\right) \lesssim C(n,\xi)\int \exp\left(n^{\frac{1}{2}-\eta}\left\|a\right\|_s^{(1-\lambda)q} + f\left\|a\right\|_s^e + \mathfrak{U}_{T_0}\right) d\mu_0(d).$$

Introducing

$$\begin{aligned} g(r) &= n^{\frac{1}{2}-\eta}r^{(1-\lambda)q} + fr^e, \\ g'(r) &= n^{\frac{1}{2}-\eta}(1-\lambda)qr^{(1-\lambda)q-1} + efr^{e-1} \end{aligned}$$

and performing an integration by parts, it follows that

$$\begin{aligned} \frac{\mu^{y_n}(\exp(f\left\|\cdot\right\|_s^e)\chi_{B^1(a^\dagger)^c})}{C(n,\xi)} &\lesssim \int \exp\left(g(\left\|a\right\|_s) + \mathfrak{U}_{T_0}\right) d\mu_0(a) \\ &\lesssim \exp(\mathfrak{U}_{T_0})\int\left[\int_0^{\left\|a\right\|_s} g'(r)\exp(g(r))dr\right] + 1 d\mu_0(a) \\ &\lesssim \int_0^\infty g'(r)\exp\left(g(r)+\mathfrak{U}_{T_0}\right) d\mu_0\left(\left\|a\right\|_s > r\right) dr \\ &\lesssim \int_0^\infty g'(R)\exp\left(n^{\frac{1}{2}-\eta}r^{(1-\lambda)q} - 2fr^e\right) dr. \end{aligned}$$

The above can only be expected to be finite if

$$(1-\lambda)q < e. \tag{C.13}$$

Moreover, we assume that $\eta < \frac{1}{2}$ since otherwise

$$\int \exp\left(n^{\frac{1}{2}-\eta}\left\|a\right\|_s^{(1-\lambda)q} + f\left\|a\right\|_s^e\right) d\mu_0(a) \lesssim \int \exp\left(2f\left\|a\right\|_s^e\right) d\mu_0(a).$$

In order to achieve an upper bound, we split the term in the exponential into $T_1 := n^{\frac{1}{2}-\eta}r^{(1-\lambda)q} - fr^e$ and $T_2 := -fr^e$. The first term is negative whenever

$$r \geq r_z := \left(n^{\frac{1}{2}-\eta}f^{-1}\right)^{\frac{1}{e-(1-\lambda)q}}.$$

For n large enough $r_z \geq 1$ holds. On the interval $[0, s_z]$ an upper bound $\mathfrak{U}_{T_1}$ on the maximum value of T_1 can be derived as follows

$$\begin{aligned} T_1' &= 0 \Rightarrow r = \left((1-\lambda)qn^{\frac{1}{2}-\eta}e^{-1}f^{-1}\right)^{\frac{1}{e-(1-\lambda)q}} \\ \mathfrak{U}_{T_1} &:= \left(\frac{(1-\lambda)q}{ef}\right)^{\frac{1}{e-(1-\lambda)q}}\left(n^{\frac{1}{2}-\eta}\right)^{1+\frac{1}{e-(1-\lambda)q}}. \end{aligned} \tag{C.14}$$

Putting everything together gives rise to

$$\begin{aligned} \frac{\mu^{y_n}\left(\exp(f\left\|\cdot\right\|_s^e)\chi_{B^1_{n^{-\kappa}}(a^\dagger)^c}\right)}{C(n,\xi)} &\lesssim \int\left(n^{\frac{1}{2}-\eta}(1-\lambda)qr^{(1-\lambda)q-1} + efr^{e-1}\right)\exp\left(\mathfrak{U}_{T_1}+\mathfrak{U}_{T_0}\right) dr \\ &\quad + \int_{r_z}^\infty\left(n^{\frac{1}{2}-\eta}(1-\lambda)qr^{(1-\lambda)q-1} + efr^{e-1}\right)\exp\left(\mathfrak{U}_{T_0} - fr^e\right) dr \end{aligned}$$

$$\lesssim n^a \exp\left(\mathfrak{U}_{T_1} + \mathfrak{U}_{T_0}\right)$$

for some a. Using Markov's inequality, this yields

$$\mu^{y_n}\left(B^1(a^\dagger)^c \cap B_R^s(0)^c\right) \lesssim C(n,\xi) n^{\frac{1}{2}-\eta} \exp\left(\mathfrak{U}_{T_0} + \mathfrak{U}_{T_1} - fR^e\right). \tag{C.15}$$

Again substituting $R = n^\theta$, this is asymptotically smaller than $\exp\left(-\frac{n^{1-2\kappa}}{2}\frac{2}{4}\right)$ if

$$\left(\frac{1}{2} - \eta\right)\left(1 + \frac{1}{e - (1-\lambda)q}\right) < \max\left(1 - 2\kappa, \theta e\right). \tag{C.16}$$

Collecting the inequalities from above, we see that the results follow by letting $\gamma \to 0$.

□

Appendix D. Normalising Constant of the BRP

Proof of Lemma 3.2. In order to bound $Z(n,\xi)$ in Equation (12), we rewrite it as

$$\mu^{y_n} = Z(n,\xi)\exp\left(-\Phi\right)\mu_0,$$

where $Z(n,\xi) = \mu_0(\exp\left(-\Phi\right))$. We bound $-\Phi$ using the Cauchy-Schwarz inequality

$$-\Phi \leq -\frac{1}{2}n\left\|a\right\|_1^2 + n\left\|a^\dagger\right\|_1\left\|a\right\|_1 + n^{\frac{1}{2}}\left\langle a,\xi\right\rangle_1.$$

The following steps are quite similar to the steps in the proof of the Theorems 3.3 and 3.4. We treat $\langle a,\xi\rangle_1$ by smoothing ξ at the expense of a

$$\begin{aligned}\left|\langle a,\xi\rangle_1\right| &\leq \left|\left\langle \Gamma^{-1+\frac{1-\sigma_0-\gamma}{2}}a, \Gamma^{\frac{\sigma_0-1+\gamma}{2}}\xi\right\rangle\right| \\ &\leq \left\|a\right\|_{1+\sigma_0+\gamma}\left\|\xi\right\|_{1-\sigma_0-\gamma}.\end{aligned}$$

We use the interpolation inequality for Hilbert scales with $\lambda = \frac{s-1-\sigma_0-\gamma}{s-1}$ (see Lemma Appendix A.1) and Hölder's inequality with $\frac{1}{p}+\frac{1}{q}=1$ to obtain

$$\left\|a\right\|_{1+\sigma_0+\gamma} \leq \left\|a\right\|_1^\lambda\left\|a\right\|_s^{1-\lambda} \leq \frac{\left\|a\right\|_1^{p\lambda}}{p} + \frac{\left\|a\right\|_1^{q(1-\lambda)}}{q}.$$

Combining these bounds yields

$$-\Phi \leq -\frac{1}{2}n\left\|a\right\|_1^2 + n\left\|a^\dagger\right\|_1\left\|a\right\|_1 + \frac{\left\|a\right\|_1^{p\lambda}}{p}\left\|\xi\right\|_{1-\sigma_0-\gamma} + \frac{\left\|a\right\|_1^{q(1-\lambda)}}{q}\left\|\xi\right\|_{1-\sigma_0-\gamma}.$$

The first three terms are bounded in a because they are dominated by the first if $\lambda p < 2$. This is implied by choosing $q = \frac{2+\gamma}{2-\lambda}$. Note that $\left\|\xi\right\|_{1-\sigma_0-\gamma}$ is μ_{ξ_n}-a.s. bounded due to Lemma Appendix A.2. Thus $Z(n,q)$ is bounded below if $e > q$. Letting $\gamma \downarrow 0$ in q we see that this is the case for

$$e > \frac{2\sigma_0}{s-1+\sigma_0}.$$

An upper bound on $Z(n,q)$ follows from a simple lower bound on $-\Phi$ on $B_M^{1+\sigma+\gamma}(0)$ and the prior measure of this set.

□

[1] S. Agapiou, S. Larsson, and A. M. Stuart. Posterior Contraction Rates for the Bayesian Approach to Linear Ill-Posed Inverse Problems. *Stochastic Process. Appl.*, 123(10):3828 – 3860, 2013.

[2] S. Agapiou, A. M. Stuart, and Y.-X. Zhang. Bayesian Posterior Contraction Rates for Linear Severely Ill-posed Inverse Problems. *ArXiv e-prints*, October 2012.

[3] G. Alessandrini. Stable Determination of Conductivity by Boundary Measurements. *Appl. Anal.*, 27(1-3):153–172, 1988.

[4] Vladimir I. Bogachev. *Gaussian Measures*, volume 62 of *Mathematical Surveys and Monographs*. Amer. Math. Soc., Providence, RI, 1998.

[5] C. Borell. Convex measures on locally convex spaces. *Ark. Mat.*, 12:239–252, 1974.

[6] T. Choi and M. J. Schervish. On Posterior Consistency in Nonparametric Regression Problems. *J. Multivariate Anal.*, 98(10):1969–1987, 2007.

[7] A. Cohen, R. A. Devore, and C. Schwab. Convergence rates of best N-term Galerkin approximations for a class of elliptic sPDEs. *Found. Comput. Math.*, 10(6):615–646, 2010.

[8] A. Cohen, R. A. Devore, and C. Schwab. Analytic regularity and polynomial approximation of parametric and stochastic elliptic PDE's. *Anal. Appl. (Singap.)*, 9(1):11–47, 2011.

[9] Giuseppe Da Prato and Jerzy Zabczyk. *Stochastic Equations in Infinite Dimensions*, volume 44 of *Encyclopedia of Mathematics and its Applications*. Cambridge University Press, Cambridge, 1992.

[10] M. Dashti, S. Harris, and A. M. Stuart. Besov Priors for Bayesian Inverse Problems. *Inverse Probl. Imaging*, 6:183–200, 2012.

[11] M. Dashti, K. J. H. Law, A. M. Stuart, and J. Voss. MAP Estimators and Posterior Consistency in Bayesian Nonparametric Inverse Problems. *ArXiv preprint 1303.4795*, 2013.

[12] M. Dashti and A. M. Stuart. Uncertainty Quantification and Weak Approximation of an Elliptic Inverse Problem. *SIAM J. Numer. Anal.*, 49:2524–2542, 2011.

[13] R. A. DeVore. Nonlinear approximation. In *Acta numerica, 1998*, volume 7 of *Acta Numer.*, pages 51–150. Cambridge Univ. Press, Cambridge, 1998.

[14] P. Diaconis and D. A. Freedman. On Inconsistent Bayes Estimates of Location. *Ann. Statist.*, 14(1):68–87, 1986.

[15] P. Diaconis and D. A. Freedman. On the Consistency of Bayes Estimates. *Ann. Statist.*, 14(1):1–67, 1986. With a discussion and a rejoinder by the authors.

[16] Joseph L. Doob. Application of the Theory of Martingales. In *Le Calcul des Probabilités et ses Applications*, Colloques Internationaux du Centre National de la Recherche Scientifique, no. 13, pages 23–27. Centre National de la Recherche Scientifique, Paris, 1949.

[17] Heinz W. Engl, Martin Hanke, and Andreas Neubauer. *Regularization of Inverse Problems*, volume 375 of *Mathematics and its Applications*. Kluwer Academic Publishers Group, Dordrecht, 1996.

[18] Lawrence C. Evans. *Partial Differential Equations*, volume 19 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, second edition, 2010.

[19] Frédéric Ferraty and Philippe Vieu. *Nonparametric functional data analysis*. Springer Series in Statistics. Springer, New York, 2006. Theory and practice.

[20] J.-P. Florens and A. Simoni. Regularized posteriors in linear ill-posed inverse problems. *Scand. J. Stat.*, 2012.

[21] D. A. Freedman. On the Asymptotic Behavior of Bayes' Estimates in the Discrete Case. *Ann. Math. Statist.*, 34:1386–1403, 1963.

[22] S. Ghosal, J. K. Ghosh, and A. W. van der Vaart. Convergence Rates of Posterior Distributions. *Ann. Statist.*, 28(2):500–531, 2000.

[23] S. Ghosal and A. van der Vaart. Fundamentals of Nonparametric Bayesian Inference. unpublished, 2012.

[24] David Gilbarg and Neil S. Trudinger. *Elliptic Partial Differential Equations of Second Order*. Classics in Mathematics. Springer-Verlag, Berlin, 2001. Reprint of the 1998 edition.

[25] M. Hairer. An Introduction to Stochastic PDEs. Lecture Notes, 2009.

[26] M. Hairer, A. M. Stuart, and J. Voss. Analysis of SPDEs Arising in Path Sampling. Part ii: The

Nonlinear Case. *Ann. Appl. Probab.*, pages 1657–1706, 2007.

[27] T. M. Hansen, K. S. Cordua, and K. Mosegaard. Inverse problems with non-trivial priors: Efficient solution through sequential Gibbs sampling. *Comput. Geosci.*, pages 1–19, 2012.

[28] Nils Lid Hjort, Chris Holmes, Peter Müller, and Stephen G. Walker, editors. *Bayesian Nonparametrics*, volume 28 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 2010.

[29] V. H. Hoang, C. Schwab, and A. M. Stuart. Complexity Analysis of Accelerated MCMC Methods for Bayesian Inversion. *ArXiv e-prints*, July 2012.

[30] B. T. Knapik, A. W. van der Vaart, and J. H. van Zanten. Bayesian Inverse Problems with Gaussian Priors. *Ann. Statist.*, 39(5):2626–2657, 2011.

[31] P. Kuchment and G. Uhlmann. The Radon and X-Ray Transforms. 2012.

[32] W. V. Li and Q.-M. Shao. Gaussian processes: inequalities, small ball probabilities and applications. In *Stochastic processes: theory and methods*, volume 19 of *Handbook of Statist.*, pages 533–597. North-Holland, Amsterdam, 2001.

[33] M. A. Lifshits. *Gaussian random functions*, volume 322 of *Mathematics and its Applications*. Kluwer Academic Publishers, Dordrecht, 1995.

[34] M. A. Lifshits. Bibliography of small deviation probabilities. `http://www.proba.jussieu.fr/pageperso/smalldev/biblio.pdf`, July 2012. accessed 01.09.2013.

[35] K. Ray. Bayesian Inverse Problems with Non-Conjugate Priors. *arXiv preprint arXiv:1209.6156*, 2012.

[36] G. R. Richter. An Inverse Problem for the Steady State Diffusion Equation. *SIAM J. Appl. Math.*, 41(2):210–221, 1981.

[37] John Roe. *Elliptic operators, topology and asymptotic methods*, volume 395 of *Pitman Research Notes in Mathematics Series*. Longman, Harlow, second edition, 1998.

[38] C. Schwab and A. M. Stuart. Sparse Deterministic Approximation of Bayesian Inverse Problems. *Inverse Probl.*, 28(4):045003, 32, 2012.

[39] X. Shen and L. Wasserman. Rates of convergence of posterior distributions. *Ann. Statist.*, 29(3):687–714, 2001.

[40] A. M. Stuart. Inverse Problems: A Bayesian Perspective. *Acta Numer.*, 19:451–559, 2010.

[41] A. M. Stuart. The Bayesian Approach to Inverse Problems. *ArXiv preprint 1302.6989*, 2013.

[42] Hans Triebel. *Function spaces and wavelets on domains*, volume 7 of *EMS Tracts in Mathematics*. European Mathematical Society (EMS), Zürich, 2008.

[43] A. W. van der Vaart and J. H. van Zanten. Rates of Contraction of Posterior Distributions Based on Gaussian Process Priors. *Ann. Statist.*, 36(3):1435–1463, 2008.

[44] S. Walker. New approaches to Bayesian consistency. *Ann. Statist.*, 32(5):2028–2043, 2004.

[45] H. Weyl. Das asymptotische Verteilungsgesetz der Eigenwerte linearer partieller Differentialgleichungen (mit einer Anwendung auf die Theorie der Hohlraumstrahlung). *Math. Ann.*, 71(4):441–479, 1912.

[46] William WG Yeh. Review of parameter identification procedures in groundwater hydrology: The inverse problem. *Water Resources Research*, 22(2):95–108, 1986.

Research Article IV

Notes on a Bayesian Elliptic Multiscale Inverse Problem.

Andrew M. Stuart and Sebastian J. Vollmer, 2013. *Article in preparation, 28 pages.*

Notes on a Bayesian Elliptic Multiscale Inverse Problem

A. M. Stuart and S. J. Vollmer

6th September 2013

Abstract

We consider the inverse problem of reconstructing the diffusion coefficient of a linear second order elliptic PDE from measurements of its solution. The diffusion coefficient is supposed to vary across two scales as a sum of a function depending on coarse variables and a periodic function depending on the fine variables. As the fine scale becomes finer, homogenisation occurs and the forward problem is well-approximated by a single scale homogenised model. However, different choices of the fine and coarse functions can lead to the same homogenised problem. We make this phenomenon precise by proving the existence of a function space manifold relating the fine and coarse scales of the diffusion coefficient leading to the same homogenised problem. If the observational noise is large compared to the homogenisation error, then the inverse problem of recovering fine and coarse scales is thus highly ill-posed due to the resulting lack of identifiability.

We consider this problem from a Bayesian perspective by specifying a prior on the fine and coarse scales by means of an expansion with random coefficients. The size of the fine scale as well as the forcing are assumed to be known. In this case, we demonstrate that the posterior distribution concentrates close to this manifold and that the concentration along this manifold is dominated by the prior. This is made apparent by an appropriate disintegration. We confirm this numerically in an identical twin experiment by generating the data using a fixed set of parameters and considering the distance of MCMC samples of the posterior to the corresponding manifold.

1 Introduction

Fitting mathematical models to data is a fundamental problem for many industries and sciences. It is used in order to explain and predict certain scenarios. In particular, these problems can be tackled using the theory of inverse problems. With the Bayesian approach, it is possible to use a priori information and to quantify the uncertainty arising from the noise in the data which is described by a probability distribution of the unknown input parameter.

For some forward problems, such as simulations of biological cells, dynamics occur on many different temporal and spatial scales. Resolving all these scales is computationally very expensive. However, another approach to some of these problems is to use the technique of homogenisation. This results in a homogenised problem which is more tractable and has nearly the same output as the multiscale problem on coarse scales. We consider the inverse problem of recovering a multiscale diffusion coefficient from observations of the solution to a linear second order multiscale elliptic PDE. Since efficient evaluations of the forward problem are beneficial for the inverse problem, we are motivated to investigate the exploitation of homogenisation. A related problem is to study how uncertainty propagates through the homogenised forward model. This has been studied with standard Monte Carlo techniques in [2] and more efficiently using generalised polynomial chaos (gPC) in [13]. Recently, the gPC approach has also been extended to multiscale diffusion coefficients in [9].

The Bayesian approach to inverse problems for the homogenised diffusion coefficient has been investigated for a log-Gaussian prior in [7], for log-Besov priors in [6] and for a prior based on series expansions with uniformly distributed coefficients in [8] and [14].

In [11], the single-scale inverse problem was fed with the data from the multiscale problem instead of solving the complete multiscale inverse problem. The authors have considered the question how the homogenisation error can be treated as noise leading to an enhanced estimation procedure. In that case, the noise and homogenisation error are of the same order.

In contrast to the ideas exploited in [11], we are interested in the case when the homogenisation error is small compared to the noise. Moreover, we treat the multiscale inverse problem motivated by the aim to identify fine scales. These can actually matter because they have physical meaning and may be required for certain predictions in problems such as subsurface geophysics.

1.1 An Elliptic PDE and its Homogenisation

The relation between a diffusion coefficient a^ϵ and the pressure p^ϵ is modelled as a second order linear elliptic PDE in divergence form

$$\begin{cases} -\nabla \cdot (a^\epsilon(x)\nabla p^\epsilon) & = f(x) \text{ in D} \\ p^\epsilon & = 0 \text{ on } \partial D \end{cases} \tag{1}$$

where the domain D is assumed to be $D = (0,1)$ and $f \in C(D)$ denotes the forcing throughout this article. The same equation is also used to describe heat conductivity in a composite material [12]. The multiscale structure of a^ϵ in Equation (1) is supposed to be periodic and of the form

$$a^\epsilon(x) = a\left(x, \frac{x}{\epsilon}\right)$$

with $a \in C^2([0,1] \times \mathbb{R})$ being 1-periodic in the second argument. We impose the following ellipticity assumption

$$a_{\text{inf}} = \inf_{x,y} a(x,y) > 0$$

in order to guarantee the existence of a solution to Equation (1). We call the arguments x and y of $a(x,y)$ the coarse and the fine variable, respectively. In this setting, the problem stated in Equation (1) is called the multiscale problem.

The aim of homogenisation is to obtain an approximation $\bar{p}$ to the pressure p^ϵ for small ϵ. This approximation satisfies an effective equation of the form

$$\begin{cases} -\nabla \cdot (\bar{a}(x)\nabla \bar{p}) & = f(x) \text{ in } D \\ \bar{p} & = 0 \text{ on } \partial D \end{cases}, \tag{2}$$

where $\bar{a}$ is the homogenised diffusion coefficient. In general, the homogenised diffusion coefficient $\bar{a}$ is given in terms of the corrector which is a solution to the cell problem, another second-order linear elliptic PDE [3]. The periodic homogenisation of elliptic PDEs is a well established mathematical theory. For a classical work, we refer the reader to [3] and we recommend [5] and [12] for introductory material. For the technical details and an account of the theory, we refer the reader to these references.

For the one dimensional problem with $D = (0,1)$, there exists an explicit formula for the homogenised diffusion coefficient $\bar{a}$ given by

$$\bar{a}(x) := \left(\int_0^1 \frac{1}{a(x,y)} dy \right)^{-1}. \tag{3}$$

The relation between the Equations (1) and (2) can be justified through the following convergence result.

Proposition 1. *Let p^ϵ be the solution of the multiscale problem in Equation (1) and let $\bar{p}$ denote the solution of the homogenised problem in Equation (2). Then the following bound holds*

$$\|\bar{p} - p_\epsilon\|_{L^\infty([0,1])} \leq C\epsilon.$$

This bound is a well-known result in the literature on homogenisation, see for example [3, p. 19]. The constant C can be derived by an application of the maximum principle [16] and depends in an exponential way on a_{inf}^{-1}. For our approximation results in Section 4, we need sharper bounds which are explicitely established in Theorem 6 in the Appendix of this article.

This convergence phenomenon affects the corresponding inverse problem introduced in the next section.

1.2 The Inverse Problem of Reconstructing the Multiscale Diffusion Coefficient

The main focus of this article is to study the heavily under-determined inverse problem of reconstructing the multiscale diffusion coefficient $a(x, y)$ from measurements of the pressure for a known fine scale ϵ and the forcing f. Throughout this article, we assume that $a(x, y)$ exhibits an additive form which reduces the under-determinedness of the inverse problem.

Assumption 2. *The multiscale diffusion coefficient can be written as the sum of the coarse diffusion coefficient, a function b in terms of the coarse variable x, and the fine diffusion coefficient, a function c depending on the fine variable y, as follows*

$$a(b, c)(x, y) := a(x, y) \quad = \quad a_0(x) + b(x) + c(y) \tag{4}$$

where a_0 and b are elements of $C^2([0, 1])$ and $c \in C^2_{per}(\mathbb{R})$ denoting the 1-periodic two times continuously differentiable functions on $\mathbb{R}$.

In the following, we consider both the forward model based on the multiscale Equation (1) and the corresponding version based on the homogenised Equation (2). The inverse problems are then concerned with the reconstruction of the multiscale structure (b, c) from measurments y and $\bar{y}$ of the multiscale pressure p^ϵ and the homogenised pressure $\bar{p}$, respectively. In both cases, the data is modelled as

$$y = \mathcal{G}_\epsilon(b, c) + \xi := \mathcal{O}(G_\epsilon(a(b, c))) + \xi = \mathcal{O}(p^\epsilon) + \xi \tag{5}$$

and

$$\bar{y} = \bar{\mathcal{G}}(b, c) + \xi := \mathcal{O}(\bar{G}(\bar{a}(b, c))) + \xi = \mathcal{O}(\bar{p}) + \xi \tag{6}$$

where $\bar{G}$ and G_ϵ are solution operators to the multiscale and the homogenised problem, respectively. Moreover, we write ξ for the observational noise and we denote by $\mathcal{O}$ the observation operator. In this article, we restrict our attention to observation operators of the form

$$\mathcal{O}(p) = \{p(i\Delta y)\}_{i=0}^{\lfloor \frac{1}{\Delta y} \rfloor}, \tag{7}$$

which corresponds to observing p on a grid with grid size Δy.

Our main attention concerns the multiscale inverse problem which is influenced by The homogenisation effect quantified through Proposition 1. As $\epsilon \to 0$, it becomes increasingly difficult to distinguish between (b_1, c_1) and (b_2, c_2) from the data y giving rise to the same homogenised diffusion coefficient, that is

$$\bar{a}(b_1, c_1) = \bar{a}(b_2, c_2).$$

This is the case because $\mathcal{G}_\epsilon(b_1, c_1)$ and $\mathcal{G}_\epsilon(b_2, c_2)$ are ϵ close to each other due to Proposition 1. We fix $b^\dagger$ and $c^\dagger$ and study the level set $\bar{a}^{-1}(\bar{a}(b^\dagger, c^\dagger))$ containing (b, c) giving rise to the same diffusion coefficient as $(b^\dagger, c^\dagger)$ does, that is

$$\bar{a}(b, c) = \bar{a}(b^\dagger, c^\dagger). \tag{8}$$

In this case (b, c) and $(b^\dagger, c^\dagger)$ correspond also to the same homogenised problem. We use a function space version of the implicit function theorem to show that the level sets form a manifold given by a graph. The details of this fact are contained in Section 2. We introduce a toy problem which allows us to illustrate our ideas in Section 1.4.

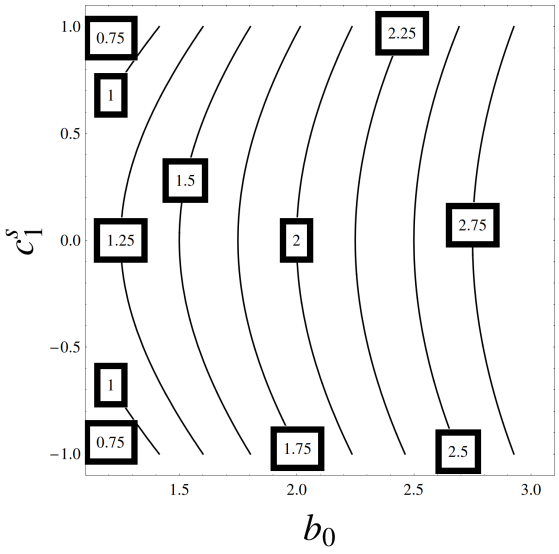


Figure 1: Level sets of $\bar{a}$

A Toy Problem
We consider the following toy model

$$a_0(x) = 0, \quad b(x) = b_0, \quad c(y) = c_1 \sin(2\pi y).$$

In this setting, the homogenised diffusion coefficient is just a constant function given by

$$\bar{a}(b_0, c_1) = \left(\int (b_0 + c_1 \sin(2\pi y))^{-1} dy \right)^{-1}.$$

In Figure 1, we have depicted different level sets of $\bar{a}$. These level sets form graphs over the c_1 coordinate because $\partial_{c_1^s} \bar{a} > 0$. We consider an identical twin experiment and assume that the data is generated according to Equation (5) for the parameters $b_0^\dagger$ and $c_1^\dagger$. Because of homogenisation effects, we expect that the data contains much information about the question on which branch the pair of parameters $\left(b_0^\dagger, c_1^\dagger\right)$ lies but only little about its exact position. It is worth pointing out that this phenomenon has an impact on any approach to the inverse problem.

1.3 The Bayesian Approach

The Bayesian approach is based on the idea that the uncertainty of the unknown parameters of a mathematical model can be modelled as a probability distribution. For an inverse problem, the a priori uncertainty has to be specified as a probability distribution, called the prior μ_0, on the input of the mathematical model, here the functions b and c. For a given distribution of the observational noise ξ and the forward operator $\mathcal{G}$, this gives rise to a joint probability distribution on (a, y). Given the data y, this results in a unique conditional distribution μ^y on a called the posterior given via its unnormalised density with respect to the prior. We suppose that the observational noise is Gaussian $\xi \sim \mathcal{N}(0, \Gamma)$. Under mild additional assumptions on the prior and the forward operator $\mathcal{G}$, the posterior takes the following form

$$\frac{d\mu^y}{d\mu_0}((b, c)) \propto \exp\left(-\frac{1}{2} \left\| y - \mathcal{G}((b, c)) \right\|_\Gamma^2 \right). \tag{9}$$

In Section 4, we study the inverse problems associated with the forward operators $\mathcal{G}$ and $\bar{\mathcal{G}}$ corresponding to the multiscale and the homogenised problem stated in Equations (5) and (6), respectively.

1.4 Main Results and Outline

We generalise the toy model introduced in Section 1.2 by considering (truncated) Fourier series expansions of the additive components b and c of the multiscale diffusion coefficient a. The main difference to the toy model is that the resulting homogenised diffusion coefficient $\bar{a}$ is no longer a constant function. In particular, we consider the following aspects:

1. We show that the level sets of the homogenised diffusion coefficient mapping $\bar{a}$, mapping the parameters b and c to the homogenised diffusion coefficient $\bar{a}(b, c)$, form manifolds that are given as graphs over c as in the toy problem. This result is proved using a functional version of the implicit function theorem in Section 2. For the toy model, we have illustrated the level set in Figure 1.

2. We investigate both the structure of the homogenised diffusion coefficient and its level sets using simulations in Section 3.

   (a) We study the influence of b and c on the homogenised diffusion coefficient $\bar{a}$. We calculate the derivative of the Fourier coefficients of $\bar{a}$ in terms of those of b and c. Whereas the derivative with respect to the Fourier coefficients of b is close to the identity, the Fourier coefficients of c have little impact on those of $\bar{a}$ except for the constant mode.

   (b) The consequences of (2a) on the level sets of $\bar{a}$ is that they are well-represented in the Fourier space. This allows us to consider an extended toy model based on three Fourier coefficients.

3. We consider the Bayesian approach to the multiscale inverse problem in Section 4 and introduce uniform series priors. This is a natural form of the prior because of the parametrisation in Equation (10). If the observational error is small and the homogenisation error is in comparison even smaller, it is conceivable to expect that the posterior concentrates around the level set of the homogenised diffusion coefficient corresponding to the parameters used to generate the data. In particular, we establish the following facts:

   (a) We confirm by MCMC simulations that the posterior concentrates around the level set corresponding to the truth in Section 5. We first present simulations for the extended toy model before considering the general case.

   (b) We show that the posterior based on the multiscale problem can be well represented using the homogenised model and disintegration in Section 4. By bounding their difference in the total variation or the Hellinger distance, the resulting bound holds for appropriate uniform series priors as well as for log-Gaussian priors.

In the following section, we set up the notation in order to establish the corresponding results in detail.

1.5 Parametrisation of the Diffusion Coefficient

In the following, we introduce a Fourier series parametrisation of the fine and coarse diffusion coefficient that are used throughout this article. We impose the following assumption.

Assumption 3. *(Fine scales in the inverse problem) For c as in Equation (4),*

$$\int_D c(y, z)dy = 0.$$

Imposing this assumption, we reduce the under-determinedness of the inverse problem because the constants cannot be exchanged between b and c in Equation (4) anymore. Throughout the article, we consider the following Fourier series representation of both b and c

$$
\begin{aligned}
a(x,y;z) &:= a(x,y;z) := a_0(x) + b(x,\mathbf{z}) + c(y,\mathbf{z}) \\
b(x,z) &= z_0 b_0 + \sum_{k=1}^{\infty} z_{1,k} b_k^c \cos(2\pi k x) + z_{2,k} b_k^s \sin(2\pi k x) \\
c(y,z) &= \sum_{k=1}^{\infty} z_{3,k} c_k^c \cos(2\pi k y) + z_{4,k} c_k^s \sin(2\pi k y). \qquad (10)
\end{aligned}
$$

This parametrisation is particularly natural for the uniform series prior considered in Section 4 corresponding to $z_{i,j} \overset{\text{i.i.d.}}{\sim} \mathcal{U}(-1,1)$.

Notation

For a suitable function $f : [0,1] \to \mathbb{R}$, we denote its Fourier coefficients by

$$
\begin{aligned}
\mathcal{F}_{\mathrm{c}}^k(f) &= 2\int_0^1 \cos(2\pi k x) f(x) dx \\
\mathcal{F}_{\mathrm{s}}^k(f) &= 2\int_0^1 \sin(2\pi k x) f(x) dx \\
\mathcal{F}^0(f) &= \int_0^1 f(x) dx.
\end{aligned}
$$

We denote by $f_{\inf}$ and $f_{\sup}$ the infimum and supremum over the relevant arguments, respectively.

2 Homogenisation of Additive Multiscale Diffusion Coefficients

Subsequently, we study the effect of homogenisation through the homogenised diffusion coefficient mapping. In particular, we restrict our attention to

$$
\mathrm{Adm} := \left\{ (b,c) \in C^2([0,1]) \times C^2_{\mathrm{per}}(\mathbb{R}) \middle| a(b,c) > 0 \right\}
$$

and view $\bar{a}(b,c)$ as a mapping $\bar{a} : \mathrm{Adm} \to C^2([0,1])$ and call it homogenised diffusion coefficient mapping. We study the level set of $\bar{a}^{-1}(\bar{a}(b^\dagger, c^\dagger))$ for fixed $(b^\dagger, c^\dagger)$ in Adm given by the solution to the following equation

$$
\left(\int_0^1 \frac{1}{a_\dagger(\cdot, y)} dy \right)^{-1} = \bar{a}(b_\dagger, c_\dagger)(\cdot) = \bar{a}(b,c)(\cdot) = \left(\int_0^1 \frac{1}{a(\cdot, y)} dy \right)^{-1} \qquad (11)
$$

using the formula for the homogenised diffusion coefficient in Equation (3). We prove that these level sets have the structure of a manifold. This can be seen as generalisation of the level sets corresponding to the toy problem in Section 1.2 which are depicted in Figure 1.

Subsequently, we use the following version of the implicit function theorem.

Theorem. *(Implicit Function Theorem [1]) Let $F \in C^l(\Lambda \times U, Y)$, $k \geq 1$ where Y is a Banach space and Λ (respectively U) is an open subset of the Banach space T (respectively X). Suppose that $F(\lambda^\star, u^\star) = 0$ and $F_u(\lambda^\star, u^\star) \in Inv(X,Y)$. Then there exist neighbourhoods Θ of $\lambda^\star$ and $U^\star$ of $u^\star$ in X and a map $g \in C^l(\Theta, X)$ such that*

1. *$F(\lambda, g(\lambda)) = 0$ for all $\lambda \in \Theta$,*
2. *$F(\lambda, g(\lambda)) = 0$, $(\lambda, u) \in \Theta \times U^\star$ implies $u = g(\lambda)$ and*
3. *$g'(\lambda) = -F_u(p)^{-1} \circ F_\lambda(p)$ where $p = (\lambda, g(\lambda))$ and $\lambda \in \Theta$.*

It is easy to see that $\bar{a} : \text{Adm} \to C^2([0,1])$ is infinitely often Fréchet-differentiable. We summarise the application of the implicit function theorem with an additional uniqueness result as follows.

Theorem 4. *If $(b^\dagger, c^\dagger) \in Adm$, then there exists an open subset $S \subseteq \Pi_2 Adm$, where Π_2 denotes the projection onto the second component, such that $\bar{a}^{-1}\left(\bar{a}(b^\dagger, c^\dagger)\right)$ can be written as*

$$\{(c, b(c)) \,|c \in S\} \tag{12}$$

with $b : C^2_{per}(\mathbb{R}) \to C^2([0,1])$ being a C^l- function for every $l \in \mathbb{N}$. Moreover, defining $b_\star(c) = \inf\left\{b \middle| \bar{a}(b,c) \geq 0 \,\forall y\right\}$, we may characterise the set S as follows. If

$$\bar{a}(b_\star, c) < \bar{a}(b_\dagger, c_\dagger), \tag{13}$$

then

$$c \in S.$$

Proof. The implicit function theorem applies to this case because

$$D_b \bar{a}(b,c)(\delta_b) = \left(\int \frac{1}{a_0(x) + b(x) + c(y)} dy\right)^{-2} \cdot \delta_b(x)$$

is an invertible linear map as $a_0(x) + b(x) + c(y) > 0$ for $(b,c) \in \text{Adm}$. Thus,

$$b(x) = b(c(\cdot))(x)$$

in a neighbourhood around any solution. Moreover, for each c there exists at most one b. If we assume the converse, then there is a c such that $(b_1, c), (b_2, c) \in \text{Adm}$ and

$$\bar{a}(b_1, c) = \bar{a}(b_2, c) = \bar{a}_\dagger.$$

This yields a contradiction because

$$0 = \bar{a}^{-1}(b_1, c)(x) - \bar{a}^{-1}(b_2, c)(x) = \int \frac{1}{(a_0 + b_1 + c)(a_0 + b_2 + c)} dy (b_1(x) - b_2(x)).$$

Let c be an arbitrary element of $C^2_{\text{per}}(\mathbb{R})$ satisfying Equation (13). Then the intermediate value theorem can be used to construct an appropriate $b \in C^2([0,1])$ such that (b,c) is a solution to Equation (8). □

We first present our numerical investigation of these level sets before studying the consequences for the Bayesian inverse problem in Section 4.

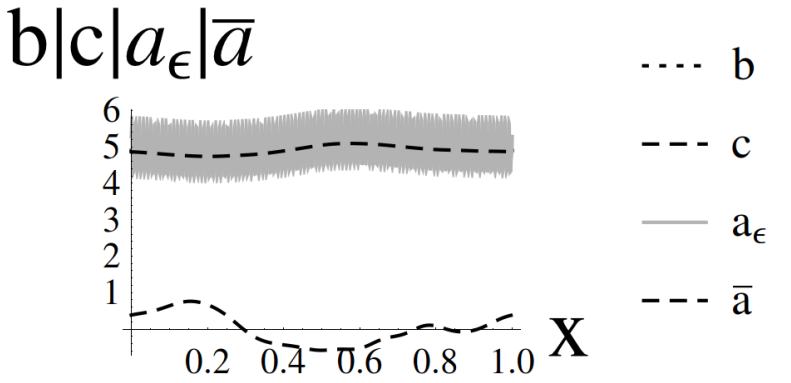
3 Numerical Investigations of the Homogenised Diffusion Coefficient Mapping and its Level Sets

We present our numerical investigations of both the homogenised diffusion coefficient mapping and its level sets in the setting of Equation (10). A crucial aspect studied in this section is the dependence of $\bar{a}$ on both b and c. In fact, the simulations demonstrate that c has little impact on $\bar{a}$.

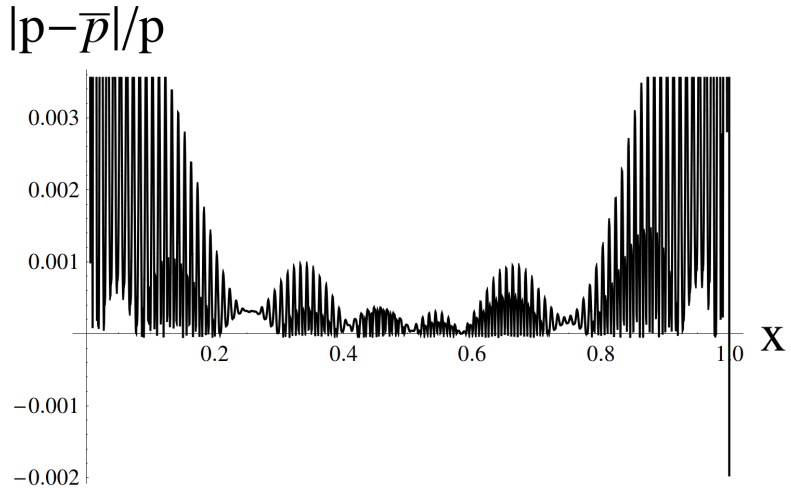
3.1 The Multiscale Phenomenon

In the following simulation, we illustrate the homogenisation for a particular instance of the general parametrisation in Equation (10) and the effect of the fine diffusion coefficient c on the homogenised diffusion coefficient. We pick the following parameters in Equation (10)

$$\begin{aligned} b_k^c = b_k^s = \frac{1}{2k^2}, k = 1, \ldots, 10 \quad &= c_k^c = c_k^s = \frac{1}{k^3}, k = 1, \ldots, 10 \\ b_0 = 1 \text{ and } a_0(x) = 5.4 & \end{aligned} \tag{14}$$



(a) Coarse scale b, fine scale c, diffusion coefficient a, homogenised diffusion coefficient $\bar{a}$



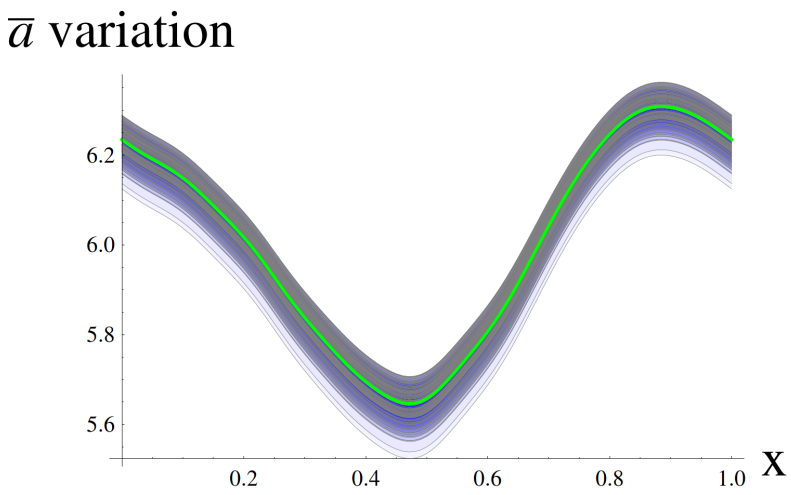
(b) Homogenisation error

Figure 2: Homogenisation Phenomenon

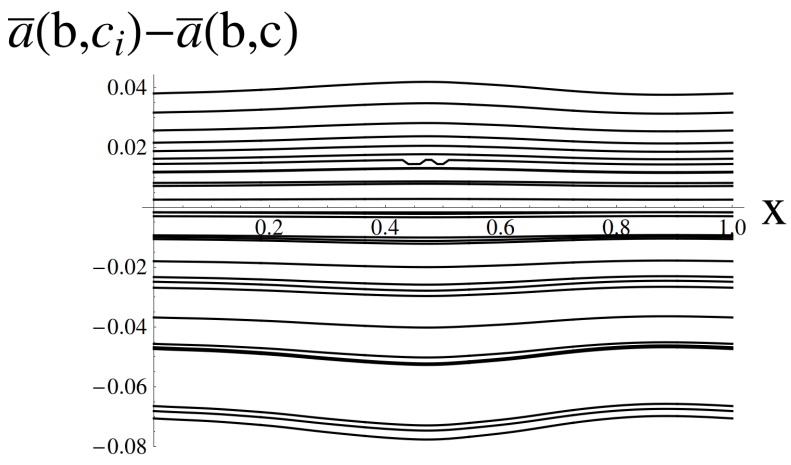
and $z_{i,j} \overset{\text{i.i.d.}}{\sim} \mathcal{U}(-1,1)$. The particular realisation has in fact little impact on the qualitative phenomenon described in this section. In Figure 2a, the functions b, c, a and $\bar{a}$ are illustrated. The relative difference between p^ϵ and $\bar{p}$ is small and is shown in Figure 2b. Even though we have chosen the Fourier coefficients of c in a much larger range, the effect of c onto $\bar{a}$ is small. We visualise the variation of $\bar{a}$ due to random samples of c in Figure 3a. The influence of c on $\bar{a}$ becomes much more apparent in Figure 3b. This figure illustrates $\bar{a}(b, c_i) - \bar{a}(b, c)$ for different realisations of c_i. We see that a change in c has the greatest impact on the constant of $\bar{a}$ and only little on the higher modes. This effect has also an impact on the manifold and its tangent direction. Before concentrating on the level sets of $\bar{a}$ in Sections 3.3 to 3.4, we study the effect of b and c on $\bar{a}$ in a quantitative way in the next section.

3.2 The Influence of b and c on $\bar{a}$

In the previous section, we illustrated qualitatively that the influence of c on $\bar{a}$ is small. In this section, we study the dependence of $\bar{a}$ on b and c quantitatively by calculating the derivative of the Fourier coefficient of $\bar{a}$ with respect to those of b and c. The corresponding Jacobian matrices can be derived using Equation

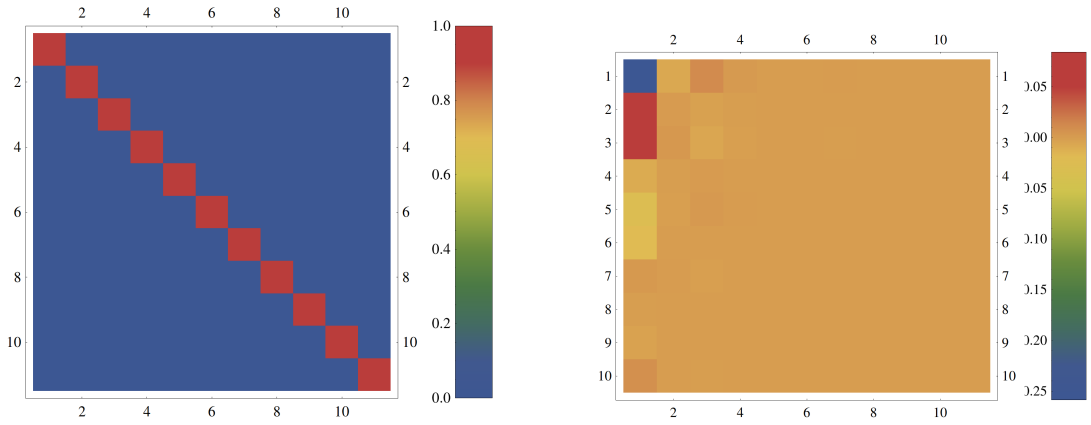


(a) Variation of $\bar{a}$ due to c



(b) Influence of c on$\bar{a}$

Figure 3: Influence of the parameter c on the homogenised diffusion coefficient $\bar{a}$



(a) Jacobian $J_b\bar{a}$ (b) Jacobian $J_c\bar{a}$

Figure 4: Derivative of Fourier coefficients of $\bar{a}$ with respect to those of b and c

(3) and we arrange the sine and cosine modes in an alternating fashion

$$
\begin{aligned}
(J_b\bar{a})_{2i,2j} &= 2\int_{[0,1]^2} \frac{\bar{a}(x)^2\cos(2\pi ix)\cos(2\pi jx)}{\left(a_0(x)+b(x)+c(y)\right)^2}dxdy, \quad i=0,...,k, j=0,...,k,\\
(J_b\bar{a})_{2i-1,2j} &= 2\int_{[0,1]^2} \frac{\bar{a}(x)^2\sin(2\pi ix)\cos(2\pi jx)}{\left(a_0(x)+b(x)+c(y)\right)^2}dxdy, \quad i=1,...,k, j=0,...,k,\\
(J_b\bar{a})_{2i,2j-1} &= 2\int_{[0,1]^2} \frac{\bar{a}(x)^2\cos(2\pi ix)\sin(2\pi jx)}{\left(a_0(x)+b(x)+c(y)\right)^2}dxdy, \quad i=0,...,k, j=1,...,k,\\
(J_b\bar{a})_{2i-1,2j-1} &= 2\int_{[0,1]^2} \frac{\bar{a}(x)^2\sin(2\pi ix)\sin(2\pi jx)}{\left(a_0(x)+b(x)+c(y)\right)^2}dxdy, \quad i=0,...,k, j=0,...,k,
\end{aligned}
$$

and

$$
\begin{aligned}
(J_c\bar{a})_{2i,2j} &= 2\int_{[0,1]^2} \frac{\bar{a}(x)^2\cos(2\pi ix)\cos(2\pi jy)}{\left(a_0(x)+b(x)+c(y)\right)^2}dxdy, \quad i=0,...,k, j=0,...,k,\\
(J_c\bar{a})_{2i-1,2j} &= 2\int_{[0,1]^2} \frac{\bar{a}(x)^2\sin(2\pi ix)\cos(2\pi jy)}{\left(a_0(x)+b(x)+c(y)\right)^2}dxdy, \quad i=1,...,k, j=0,...,k,\\
(J_c\bar{a})_{2i,2j-1} &= 2\int_{[0,1]^2} \frac{\bar{a}(x)^2\cos(2\pi ix)\sin(2\pi jy)}{\left(a_0(x)+b(x)+c(y)\right)^2}dxdy, \quad i=0,...,k, j=1,...,k,\\
(J_c\bar{a})_{2i-1,2j-1} &= 2\int_{[0,1]^2} \frac{\bar{a}(x)^2\sin(2\pi ix)\sin(2\pi jy)}{\left(a_0(x)+b(x)+c(y)\right)^2}dxdy, \quad i=0,...,k, j=0,...,k,
\end{aligned}
$$

These matrices are visualised in Figures 4a and 4b, respectively. It is clear from the figures that the Fourier coefficients of c have little influence on those of $\bar{a}$ except for the constant mode. The structure of the $J_b\bar{a}$ does not depend much on the choice of (b,c). The same is true for $J_c\bar{a}$ except for the magnitude sign of the entries that are not too close to zero.

The fact that $J_b\bar{a}$ is close to the identity justifies the study of discretised versions of the level sets of $\bar{a}$ in the truncated Fourier space. We start with an extended toy model based on three Fourier coefficients in the next section.

3.3 Fourier Series with Three Coefficients - An Extended Toy Model

We investigate the level set $\bar{a}^{-1}(\bar{a}(b,c))$ numerically by considering a simple parametrisation, a special instance of Equation (10), only based on the following three coefficients

$$\begin{aligned} a(x,y) = &\quad a_0 + b(x) + c(y) \quad \text{with} \\ b(x) = &\quad z_0 + z_{2,1}\sin(2\pi x) \\ c(z_{4,1})(y) = &\quad z_{4,1}\sin(2\pi\frac{y}{\epsilon}) \end{aligned} \tag{15}$$

where z_0, $z_{2,1}$ and $z_{4,1}$ are real numbers. We fix $a^\dagger$ through

$$a_0 = 3, \quad z_0^\dagger = 0, \quad z_{2,1}^\dagger = -0.5, \quad z_{4,1}^\dagger = 0.5.$$

For all numerical simulations in this section, we set $\epsilon = 0.005$. The solution to Equation (1) and the point-wise homogenisation error are plotted in Figures 5 and 6, respectively.

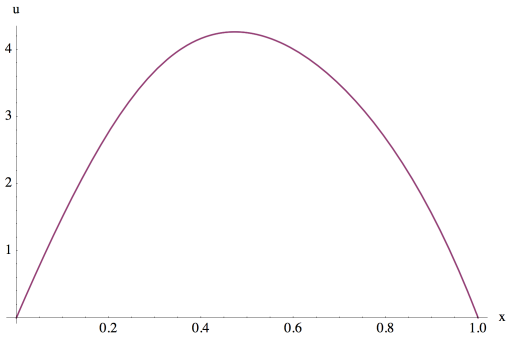


Figure 5: Pressure and homogenisation for the true input

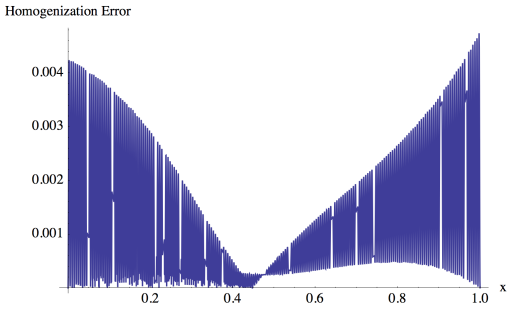


Figure 6: Point-wise homogenisation error

From the previous section, we know that all solutions of

$$\bar{a}(b,c)(x) = \bar{a}^\dagger(x)$$

form a manifold such that $b = b(c(z_{4,1}))$. We solve for $b(x)$

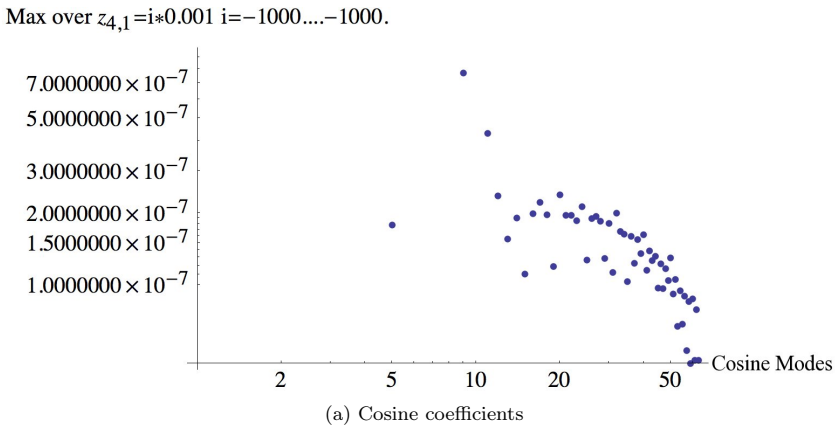
$$\bar{A}_\dagger^\epsilon(x) = \left(\int \frac{1}{a_0 + b(x) + c(y,\omega)} dy\right)^{-1}$$

using Newton's method for both x and $z_{4,1}$ on a grid dividing $[-1,1]$. We calculate the Fourier coefficients for the resulting function $b(z_{4,1})$.

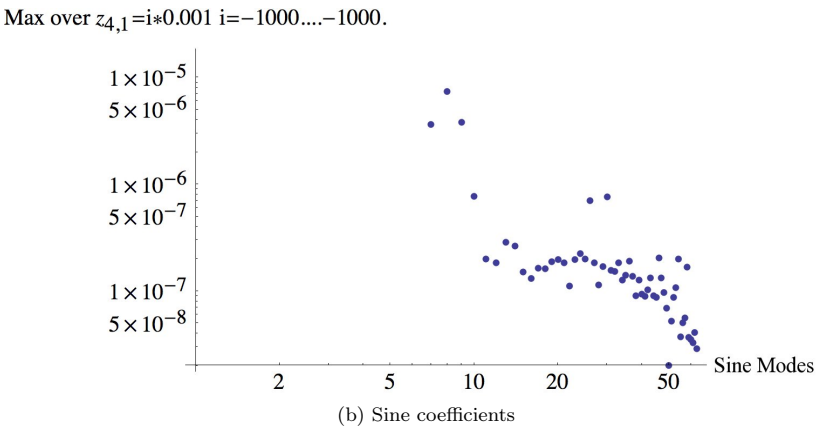
The level sets of $\bar{a}$ are very well-represented in Fourier space as Figure 7. Numerical simulations suggest that only the constant and the first sine coefficients are non-negligible. Note that due to Equation (15), this is the case for $b^\dagger$. All Fourier coefficients except for the constant term and the first sine-coefficients are close to zero. Hence b is approximately of the form stated in Equation (15). The investigation of $J_b\bar{a}$ and $J_c\bar{a}$ in the previous section can be seen as an explanation. Setting the other Fourier coefficients equal to zero, we present the L^2-differences in the pressure and the homogenised diffusion coefficient in Table 1. The level set of $\bar{a}$ is illustrated by the red curve in Figure 8. The figure also contains the level sets of $\left\|p(a_\dagger^\epsilon) - p(a_c^\epsilon)\right\|_{L^2}$ and $\|\bar{a_\dagger} - \bar{a_c}\|_{L^2}$. Both level sets wrap nicely around the manifold corresponding to the level set $\bar{a}^{-1}(\bar{a}(b^\dagger, c^\dagger))$.

$z_{4,1}$	$a_0 + \mathcal{F}_0(b)$	$\mathcal{F}^1_{\sin}(b)$	$\max_{\substack{i \in \{2,\dots 64\} \\ j \in \{1,\dots 64\}}} \left\{ \begin{array}{c} \lvert \mathcal{F}^j_{\mathrm{c}}(b_c) \rvert, \\ \lvert \mathcal{F}^i_{\mathrm{s}}(b_c) \rvert \end{array} \right\}$	$\left\lVert p(a^\epsilon_\dagger) - p(a^\epsilon) \right\rVert_{L^2}$	$\left\lVert \bar{a^\epsilon_\dagger} - \bar{a^\epsilon} \right\rVert_{L^2}$
-0.9	3.09313	-0.48478	0.00121174	0.0010831	0.00091151
-0.8	3.06517	-0.48922	0.00086781	0.0010123	0.00064758
-0.7	3.04028	-0.49326	0.00054762	0.0009415	0.00040704
-0.6	3.01853	-0.49683	0.00025661	0.0008710	0.00019447
-0.5	3.	-0.49992	0.00002446	0.0007931	0.00005090
-0.4	2.98475	-0.50250	0.00021737	0.0009917	0.00016562
-0.3	2.97282	-0.50453	0.00039124	0.0006368	0.00028683
-0.2	2.96428	-0.50599	0.00051810	0.0005578	0.00037606
-0.1	2.95914	-0.50688	0.00059529	0.0004772	0.00043032
0	2.95742	-0.50717	0.00062121	0.0003981	0.00044852
0.1	2.95914	-0.50688	0.00059529	0.0003156	0.00043032
0.2	2.96428	-0.50599	0.00051810	0.0002365	0.00037606
0.3	2.97282	-0.50453	0.00039124	0.0001566	0.00028683
0.4	2.98475	-0.50250	0.00021737	0.0003908	0.00016562
0.5	3.	-0.49992	0.00002446	0.0000124	0.00005090
0.6	3.01853	-0.49683	0.00025661	0.0000739	0.00019447
0.7	3.04028	-0.49326	0.00054762	0.0001511	0.00040704
0.8	3.06517	-0.48922	0.00086781	0.0002211	0.00064758
0.9	3.09313	-0.48478	0.00121174	0.0002914	0.00091151
1.	3.12407	-0.47995	0.00157394	0.0003655	0.00119558

Table 1: Properties of $b(c)$

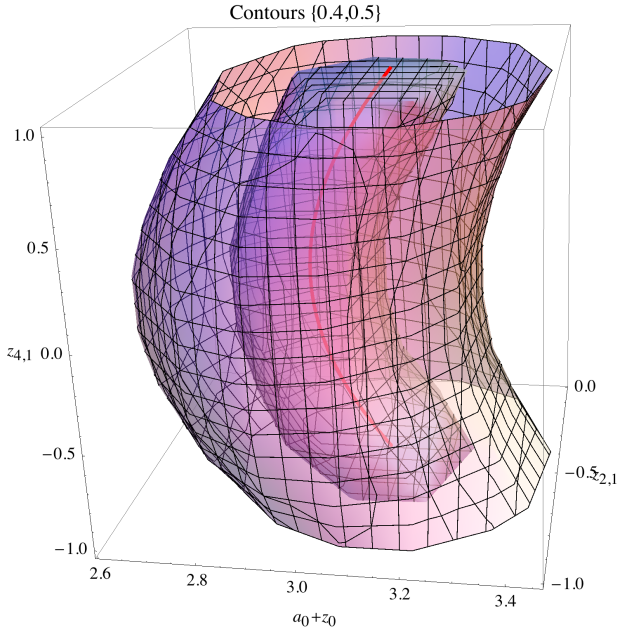


(a) Cosine coefficients

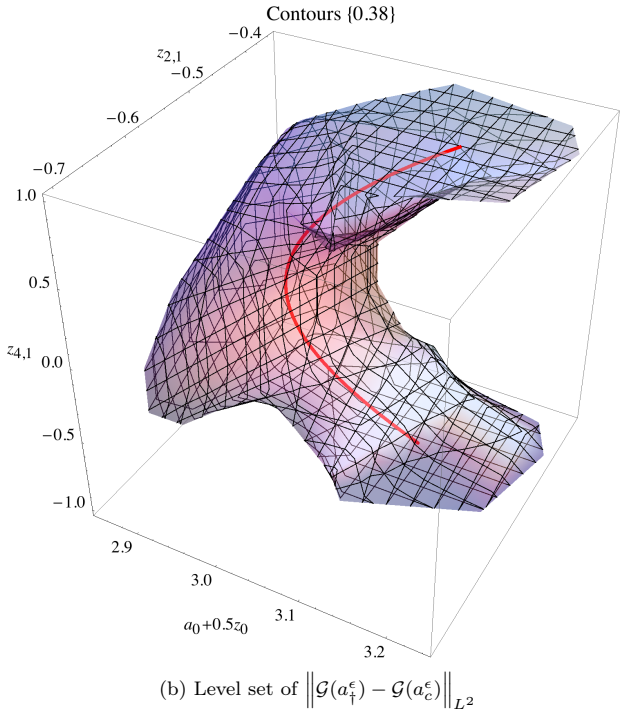


(b) Sine coefficients

Figure 7: Maximum of the Fourier coefficients



(a) Level sets of $\left\|\bar{a}_\dagger^\epsilon - \bar{a}_c^\epsilon\right\|_{L^2}$



(b) Level set of $\left\|\mathcal{G}(a_\dagger^\epsilon) - \mathcal{G}(a_c^\epsilon)\right\|_{L^2}$

Figure 8: Contour plots

3.4 Fourier Representation of the Manifold for Higher Order Expansions

We show that for higher order Fourier series expansions of $b^\dagger$ and $c^\dagger$ as in Equation (10), the level sets of $\bar{a}$ are also well represented with the same number of non-zero Fourier coefficients. We take

$$b_k^c = b_k^s, k = 1, \ldots, 5 \quad = c_k^c = c_k^s = \frac{1}{k^2}, k = 1, \ldots, 10$$

in

$$\begin{aligned} a(x, y; \omega) &:= a(x, y; \mathbf{z}(\omega)) := a_0(x) + b(x, \mathbf{z}) + c(y, \mathbf{z}). \\ b(x, \mathbf{z}) &= z_0 b_0 + \sum_{k=1}^{\infty} z_{1,k} b_k^c \cos(2\pi k x) + z_{2,k} b_k^s \sin(2\pi k x) \\ c(y, \mathbf{z}) &= \sum_{k=1}^{\infty} z_{3,k} c_k^c \cos(2\pi k y) + z_{4,k} c_k^s \sin(2\pi k y), \end{aligned}$$

with all other coefficients being zero and

$$a(x) = a_0 = 7.52676.$$

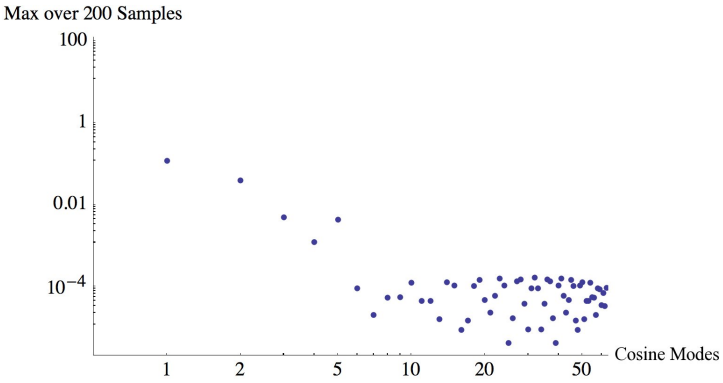
Moreover, we fix $\epsilon = 0.005$ and

$$z_0^\dagger = 0, z_{i,k}^\dagger = U_{i,k}, \quad k = 1, \ldots, 5$$

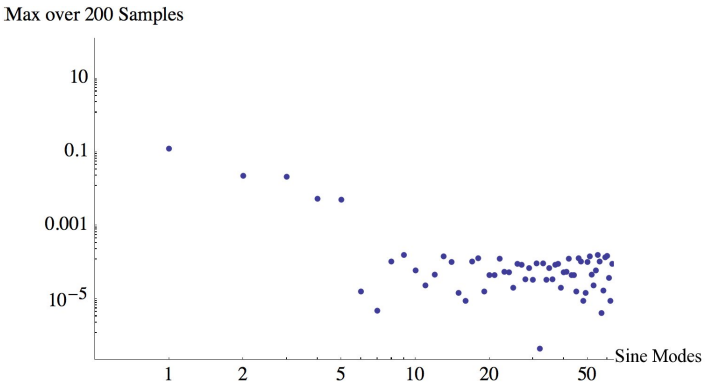
with a particular realisation of $U_{i,k} \overset{\text{i.i.d.}}{\sim} \mathcal{U}(-1, 1)$.

We calculate the Fourier series of b(c) for $c = c_1, \ldots, c_{200}$ corresponding to random independent samples of $z_{3,i}$ and $z_{4,i}$. In Figure 9, we plot the maximum over the Fourier coefficients of $b(c)$ for $c = c_1, \ldots, c_{200}$.

Similar to the low dimensional example considered in Section 3, we see a sharp cut off in magnitude of the Fourier coefficients. This means that the level set $\bar{a}^{-1}(\bar{a}(b^\dagger, c^\dagger))$ is well-represented by a truncated Fourier series. Again the form of $J_b \bar{a}$ and $J_c \bar{a}$ as considered in Section3.3 can be seen as an explanation.



(a) Cosine coefficients



(b) Sine coefficients

Figure 9: Maximum of the Fourier coefficients

4 The Bayesian Approach and Approximation Results

We apply the Bayesian approach to the inverse problem of reconstructing the additive multiscale diffusion coefficient as presented in Section 1.2. We introduce the prior and state the formulae for the posterior before deriving an approximation result based on the homogenised forward operator. Using MCMC simulations, we demonstrate that the posterior concentrates around a level set of the homogenised diffusion coefficient mapping in Section 5.

4.1 The Multiscale Inverse Problem from the Bayesian Perspective

Following the Equations (5) and (6), we use the forward operators $\mathcal{G}_\epsilon$ and $\bar{\mathcal{G}}$ and assume that the data can be modelled as

$$y = \mathcal{G}_\epsilon(b, c) + \eta \tag{16}$$
$$y = \bar{\mathcal{G}}(b, c) + \eta. \tag{17}$$

We set up the Bayesian inverse problem by assuming that the observational noise $\eta \sim \mathcal{N}(0, \Gamma)$ is Gaussian and by specifying the prior. As described in [14] and [8], we choose a prior that is based on a series expansion with uniformly distributed coefficients. This prior comes naturally in our setting as it corresponds to randomising the coefficients z in the parametrisation given in Equation (10) recalled in the following

$$\begin{aligned} a(x, y; z) &:= a(x, y; z) := a_0(x) + b(x, \mathbf{z}) + c(y, \mathbf{z}). \\ b(x, z) &= z_0 b_0 + \sum_{k=1}^{\infty} z_{1,k} b_k^c \cos(2\pi k x) + z_{2,k} b_k^s \sin(2\pi k x) \\ c(y, z) &= \sum_{k=1}^{\infty} z_{3,k} c_k^c \cos(2\pi k y) + z_{4,k} c_k^s \sin(2\pi k y). \end{aligned} \tag{10}$$

Drawing $z_{i,j}$ independently from the uniform distribution corresponds to

$$z \sim \mathcal{U}(-1, 1)^{4k+1}$$

where $k \in \mathbb{N} \cup \{\infty\}$ and $\mathcal{U}(-1, 1)$ denotes the uniform distribution on $[-1, 1]$. A finite k represents a discretised model whereas an infinite k is linked to the ideal infinite dimensional model. Moreover, we suppose that a_0, b_k^c, b_k^s, c_k^c and c_k^s in Equation (10) are chosen such that

$$(b(z), c(z)) \in \text{Adm for all } z \in [-, 1]^{4k+1}$$

with $\text{Adm} := \left\{ (b, c) \in C^2([0, 1]) \times C^2_{\text{per}}(\mathbb{R}) \middle| a(b, c) > 0 \right\}$. In particular, this holds if there exists $\kappa > 0$ such that

$$b_0 + \sum_{k=1}^{\infty} b_k^s + b_k^c + c_k^s + c_k^c \leq \frac{\kappa}{1 + \kappa} a_0^{\min} \tag{18}$$

where

$$a_0^{\min} = \inf_x a_0(x).$$

For this problem, the prior and posterior can either be formulated in terms of the coefficients $z = \{z_{i,k}\}$ or on some function space for the diffusion coefficient through $\mu_0 = \mathcal{L}(a^\epsilon)$. We consider the prior on the coefficients to avoid technicalities.

For finte dimensional noise ξ with distribution given as a density ρ with respect to the Lebesgue measue, the likelihood $L((b, c)|y)$ of the data can be represented as

$$L((b, c)|y) = \rho(y - \mathcal{G}_\epsilon(b, c)).$$

Since $\xi \sim \mathcal{N}(0, \Gamma)$, the likelihood is proportional to

$$L((b,c)|y) \propto \exp(-\mathbf{\Phi}(y;(b,c)))$$

where the potential $\mathbf{\Phi}$ is given by

$$\mathbf{\Phi}(y;(b,c)) = \frac{1}{2}\left\|y - \mathcal{G}_\epsilon((b,c))\right\|_\Gamma^2.$$

For the homogenised inverse problem, we denote by $\bar{L}$ and $\bar{\Phi}$ the corresponding formulae with $\mathcal{G}_\epsilon$ replaced by $\bar{\mathcal{G}}$, respectively. The prior and the likelihood determine the posterior as

$$\frac{d\mu^y}{d\mu_0}((b,c)) \propto \exp\left(-\frac{1}{2}\left\|y - \mathcal{G}_\epsilon((b,c))\right\|_\Gamma^2\right) \tag{19}$$

which corresponds to the regular conditional probability measure of (b,c) given y. A derivation of this fact can be found in [15]. Again, the analogue formula holds for the inverse problem for reconstruction of the homogenised diffusion coefficient

$$\frac{d\bar{\mu}^y}{d\bar{\mu}_0}((b,c)) \propto \exp\left(-\bar{\mathbf{\Phi}}\left(y;(b,c)\right)\right).$$

4.2 An Approximation to the Posterior Based on Disintegration

In this section, we derive an approximation of the full posterior using disintegration of the prior and the posterior of the homogenised problem. Using the disintegration property of regular conditional probability distributions (Theorem 5.3 in [10] or Theorem 10.4.14 in [4]), we know that

$$\mu_0(b,c) = \int \bar{a}_\star\mu_0(dh)\mu_{0|\bar{a}=h}(db,dc)$$

such that the full posterior of μ^y takes the form

$$\int f(b,c)\mu(b,c) = \int \left(\bar{a}_\star\mu_0\right)(dh)\mu_{0|\bar{a}=h}(db,dc)f(b,c)\rho_\xi(\mathcal{G}_\epsilon(b,c)-)Z_\epsilon^{-1}$$

where Z_ϵ denotes the normalisation constant. We consider the following approximation μ_h^y such that

$$\int f(b,c)\mu_h(b,c) = \int \left(\bar{a}_\star\mu_0\right)(dh)\mu_{0|\bar{a}=h}(db,dc)f(b,c)\rho_\xi(\bar{\mathcal{G}}(\bar{a})-)Z_h^{-1}$$

where Z_h denotes again the normalisation constant. In this setting, we obtain a bound on $d_{TV}(\mu^y,\mu_h^y)$ and $d_{\text{Hell}}(\mu^y,\mu_h^y)$ in terms of the forward difference.

Theorem 5. *Suppose that the posterior μ^y and the measure μ_h^y are well-defined, $\|\rho_\xi\|_\infty$ is finite and*

$$\left\|\mathcal{G}_\epsilon(b,c) - \bar{\mathcal{G}}(h)\right\|_\Gamma \leq \phi(\epsilon)K(b,c)$$

with $\bar{a}(b,c) = h$ and K which is integrable with respect to the prior. Then the total variation distance between μ_h and μ can be bounded as follows

$$d_{TV}(\mu^y,\mu_h^y) \leq C\phi(\epsilon).$$

Moreover, if additionally $\left\|\frac{(\rho_\xi)^2}{\rho_\xi}\right\|_\infty$ is bounded and K^2 is integrable with respect to the prior, then also

$$d_{\text{Hell}}(\mu^y,\mu_h^y) \leq C\phi(\epsilon).$$

Proof. First, we bound the difference in the normalising constant using disintegration

$$\begin{aligned}
|Z_\epsilon - Z_h| &\leq \int (\bar{a}_\star \mu_0)(dh) \int \mu_{0|\bar{a}=h}(db, dc) \left| \rho_\xi(\mathcal{G}_\epsilon(b,c)-) - \rho_\xi(\bar{\mathcal{G}}(h) - y) \right| \\
&\leq \int (\bar{a}_\star \mu_0)(dh) \int \mu_{0|\bar{a}=h}(db, dc) \left\| \rho_\xi \right\|_\infty \left\| \mathcal{G}_\epsilon(b,c) - \bar{\mathcal{G}}(h) \right\| \\
&\leq C \int (\bar{a}_\star \mu_0)(dh) \int \mu_{0|\bar{a}=h}(db, dc) \left\| \rho_\xi \right\|_\infty \left\| \mathcal{G}_\epsilon(b,c) - \bar{\mathcal{G}}(h) \right\| \\
&\leq C\phi(\epsilon).
\end{aligned}$$

This allows us to bound the total variation distance as follows

$$\begin{aligned}
d_{\mathrm{TV}}(\mu^y, \mu_h^y) &= \int (\bar{a}_\star \mu_0)(dh) \int \mu_{0|\bar{a}=h}(db, dc) \left| Z_\epsilon^{-1} \rho_\xi(\mathcal{G}_\epsilon(b,c) - d) - Z_h^{-1} \rho_\xi(\bar{\mathcal{G}}(h) - y) \right| \\
&\leq \int (\bar{a}_\star \mu_0)(dh) \int \mu_{0|\bar{a}=h}(db, dc) Z_h^{-1} \left| \rho_\xi(\mathcal{G}_\epsilon(b,c) - d) - \rho_\xi(\bar{\mathcal{G}}(h) - d) \right| \\
&\quad + \left| Z_h^{-1} - Z_\epsilon^{-1} \right| Z_h^{-1} \int (\bar{a}_\star \mu_0)(dh) \int \mu_{0|\bar{a}=h}(db, dc) \rho_\xi(\mathcal{G}_\epsilon(b,c) - d) \\
&\leq CC\phi(\epsilon).
\end{aligned}$$

Similarly, we calculate the Hellinger distance

$$\begin{aligned}
2d_{\mathrm{Hell}}\left(\mu^y, \mu_h^y\right)^2 &\leq \int (\bar{a}_\star \mu_0)(dh) \int \mu_{0|\bar{a}=h}(db, dc) \left| Z^{-\frac{1}{2}} \rho_\xi(\mathcal{G}_\epsilon(b,c) - d)^{\frac{1}{2}} - Z_h^{-\frac{1}{2}} \rho_\xi(\bar{\mathcal{G}}(h) - d)^{\frac{1}{2}} \right|^2 \\
&\leq I_1 + I_2
\end{aligned}$$

with

$$\begin{aligned}
I_1 &\leq \frac{2}{Z} \int (\bar{a}_\star \mu_0)(dh) \int \mu_{0|\bar{a}=h}(db, dc) \left| \rho_\xi(\mathcal{G}_\epsilon(b,c) - d)^{\frac{1}{2}} - \rho_\xi(\bar{\mathcal{G}}(h) - y)^{\frac{1}{2}} \right|^2 \\
I_2 &\leq 2(Z^{\frac{1}{2}} - Z_h^{\frac{1}{2}})^2 \int (\bar{a}_\star \mu_0)(dh) \int \mu_{0|\bar{a}=h}(db, dc) \rho_\xi(\bar{\mathcal{G}}(h) - d).
\end{aligned}$$

We bound the first term using the assumption that $K(b,c)$ has a finite second moment with respect to the prior as follows

$$I_1 \leq \frac{2}{Z} \int (\bar{a}_\star \mu_0)(dh) \int \mu_{0|\bar{a}=h}(db, dc) \left\| \frac{(D\rho_\xi)^2}{\rho_\xi} \right\|_\infty \left\| \mathcal{G}_\epsilon(b,c) - \bar{\mathcal{G}}(h) \right\|^2 \leq C\phi(\epsilon)^2.$$

Moreover, we know that $I_2 \leq C\phi(\epsilon)^2$ because the integral inside I_2 is bounded and

$$\left| Z^{\frac{1}{2}} - Z_h^{\frac{1}{2}} \right|^2 \leq C \left(Z^{-3} \vee Z_h^{-3} \right) |Z - Z_h|^2 \leq C\phi(\epsilon)^2.$$

□

Similar results hold for $d_{\mathrm{TV}}(\bar{\mu}^y, \bar{a}_\star \mu^y)$ and $d_{\mathrm{Hell}}(\bar{\mu}^d, \bar{a}_\star \mu^y)$, that is the distance between the push forward of the multiscale posterior on the homogenised diffusion coefficient and the posterior of the homogenised problem.

Moreover, an appropriate bound on the homogenisation error is obtained in Theorem 6. Note that this result then also holds for log-Gaussian priors because the corresponding factor $K(b,c)$ in front of ϵ in Equation (21) in Theorem 6 has moments of all order.

5 MCMC Simulations for the Multiscale Inverse Problem

MCMC simulations for Bayesian elliptic inverse problems, as introduced in the previous section, are presented. We perform again an identical twin experiment by generating data corresponding to fixed parameters $b^\dagger$ and $c^\dagger$ in order to study the consequences of homogenisation and especially the level set structure of $\bar{a}$ to the posterior in the Bayesian approach. We show both visually and by a histogram that the samples of the MCMC chain are close to the level set $\bar{a}^{-1}(\bar{a}(b,c))$. As in Section 3, we study both the three dimensional toy model and higher order expansions. For all simulations, we set $\epsilon = 0.005$ and $a_0 = 3$.

5.1 MCMC Simulations for the Extended Toy Model

We revisit the toy model of Section 3.3, this time considering the corresponding inverse problem. The data is generated using the multiscale equation. We follow Section 3.3 by visualising the level set of $\bar{a}^{-1}(\bar{a}(b^\dagger, c^\dagger))$ for a three term expansion of the coarse and fine diffusion coefficients. Additionally, we plot MCMC samples from the posterior of the corresponding inverse problem in Figure 10. We choose a prior according to Equation (10) with

$$b_0 = 0.5, b_k^s = 1, c_k^s = 1$$

all other coefficients are zero,

and an observation operator $\mathcal{O}$ corresponding to 50 equally spaced observations of the pressure taking the form

$$\mathcal{O}(p) = \{p(\Delta y i)\}_{i=0,\dots,\lfloor 1/\Delta y \rfloor} \quad \Delta y = 0.02.$$

We generate artificial data corresponding to

$$y = \mathcal{O}(G(a^\epsilon(z^\dagger))) + \eta$$

with $\eta \sim \mathcal{N}(0, \sigma^2 I)$ and $z_0^\dagger = 0, z_{2,1}^\dagger = -0.5$, $z_{4,1}^\dagger = 0.5$ and choose the forcing $f = 2 + x^2$ in Equation (1). As long as

observational noise $\gg$ homogenization error,

it is natural to expect that the posterior concentrates along the level set $\bar{a}^{-1}(\bar{a}(b^\dagger, c^\dagger))$ as the observational noise decreases (compare Figure 10).

5.1.1 Distance to the Level Set $\bar{a}^{-1}(\bar{a}(b^\dagger, c^\dagger))$

We are interested in the distribution of the distance of (b, c) to $\bar{a}^{-1}(\bar{a}(b^\dagger, c^\dagger))$ according to the posterior μ^y. We approximate this by considering the MCMC chain $(z_0^n, z_{2,1}^n, z_{4,1}^n)$. Naturally, the smaller the observational noise is, the closer are the samples to the level set. This can be seen in Figure 11 depicting MCMC samples for different choices of the observation noise corresponding to $\sigma = 0.05$, $\sigma = 0.01$ and $\sigma = 0.05$. We quantify the distance by bounding the distance of each sample of the MCMC chain to the level set

$$\inf_{c \in S} \left\| b(c) - \left(z_0^n + z_{2,1}^n \sin(2\pi x)\right) \right\|_{L^2}^2 + \left\| c - \left(z_{4,1}^n \sin(2\pi y)\right) \right\|_{L^2}^2 .$$

This can be bounded from above by the vertical distance

$$\left\| b(z_{4,1}^n \sin(2\pi y)) - \left(z_0^n + z_{2,1}^n \sin(2\pi x)\right) \right\|_{L^2}^2 .$$

We present a histogram of this quantity for $\sigma = 0.05$ and $\sigma = 0.01$ in Figure 12. As expected, the distribution of the vertical distance to the level sets of the MCMC samples from the posterior corresponding to $\sigma = 0.01$ is much closer to zero than those corresponding to $\sigma = 0.05$.

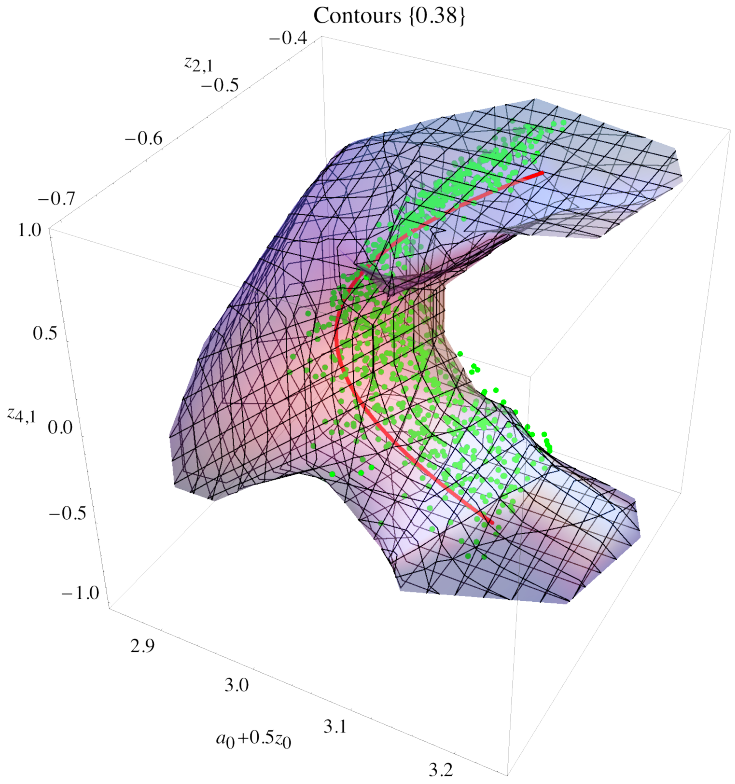
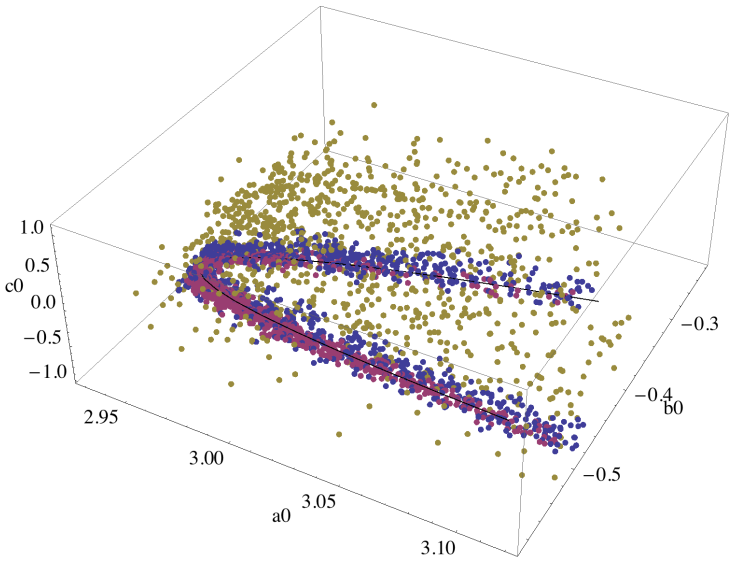
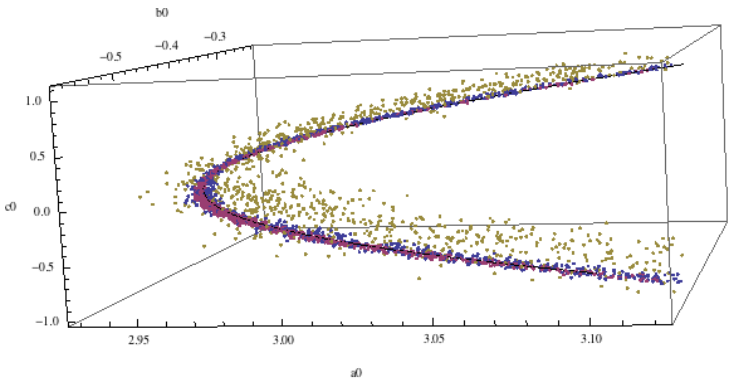


Figure 10: MCMC points (green), manifold (red), level sets of the L^2-distance to $p^\dagger$ for $\sigma = 0.05$ (blue)

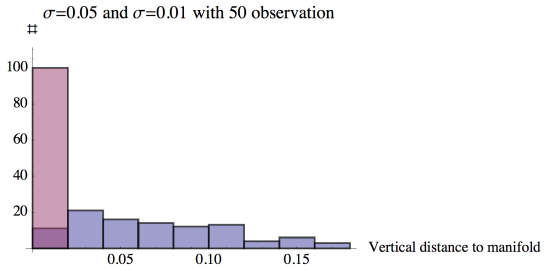


(a) A viewpoint

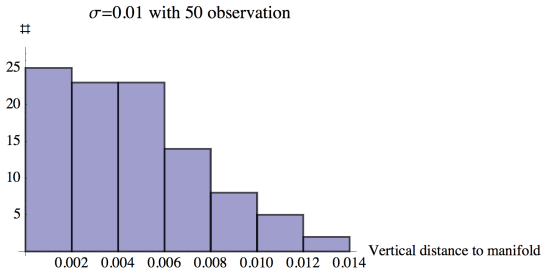


(b) Another viewpoint

Figure 11: MCMC samples of the posterior μ^y for different magnitudes of observational noise $\sigma = 0.05$, $\sigma = 0.01$ and $\sigma = 0.05$



(a) Observational noise with $\sigma = 0.05$ and $\sigma = 0.01$



(b) Observational noise $\sigma = 0.01$

Figure 12: Histogram of the distance of MCMC samples to the level set $\bar{a}^{-1}(\bar{a}(b^\dagger, c^\dagger))$

5.2 Distance of MCMC Samples to the Manifold for Higher Order Expansions

We consider the elliptic multiscale inverse problem with 20 Fourier coefficients for each the fast and coarse scales similar to Section 3.4. In this case, the MCMC samples cannot be visualised alongside the level set $\bar{a}^{-1}(\bar{a}(b^\dagger, c^\dagger))$ any more. Instead, similar to the last section, we show a histogram of vertical distances to the manifold for different magnitudes of observational noise. The parametrisation in Equation (10) is used with

$$b_k^c = b_k^s = c_k^c = c_k^s = \frac{1}{k^2}, \ldots, k = 1, \ldots, 5$$
$$b_k^c = b_k^s = c_k^c = c_k^s = 0, \ldots, k = 6, \ldots, \infty$$

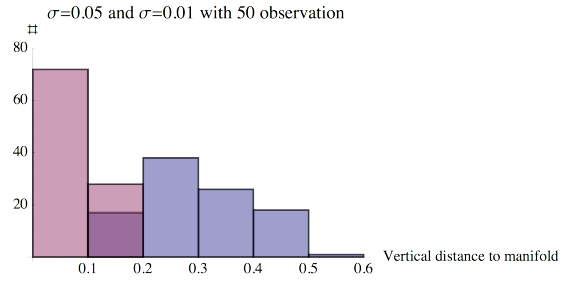
$$a(x) = a_0 = 7.15444.$$

We take the forcing $f = 2 + x^2$ in Equation (1). Moreover, we fix $z^\dagger$ as one sample of the prior and create artificial data. Again, we use an MCMC chain applied to the posterior corresponding to 50 equally spaced observations, that is $\Delta y = 0.02$ in Equation (7), with i.i.d. one-dimensional normal noise and standard deviations $\sigma = 0.05$ and $\sigma = 0.01$. We take 100 random samples from a chain of 1000000 and calculate the vertical distance to the level set. The result is presented in Figure 13.

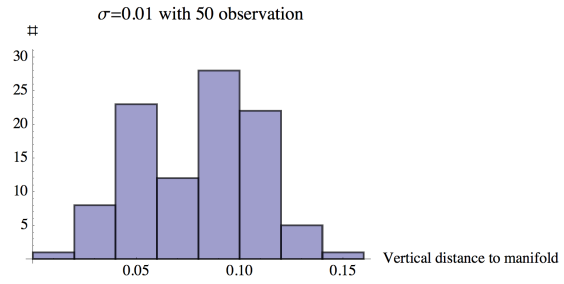
Remark. The reason that there are only a few samples in the histograms in Figure 13, which have nearly zero distance, is that the set of these points has a very small volume.

6 Conclusion and Avenues of Further Research

We have investigated an elliptic inverse problem with an additive multiscale structure. It was shown that there exists a manifold of coarse and fine variables that homogenise to the same effective problem,



(a) $\sigma = 0.05$ and $\sigma = 0.01$



(b) $\sigma = 0.01$

Figure 13: Histogram of the distance of MCMC samples to the level set $\bar{a}^{-1}(\bar{a}(b^\dagger, c^\dagger))$

the level sets of the homogenised diffusion coefficient $\bar{a}$. Starting from the simple toy problem in Section 1.2, for which the homogenised diffusion coefficient corresponds to a constant function, we investigated generalisations of the former.

Moreover, we considered higher Fourier expansions of the functions b and c. Through simulations, we concluded that the Fourier coefficients of c influence almost only the constant mode of $\bar{a}$. Moreover, the Jacobian of the Fourier coefficients of $\bar{a}$ with respect to those of b is close to the identity. A possible direction for further investigation is to justify this structure analytically using the Riemann-Lebesgue lemma. In particular, the manifold structure of the level sets of $\bar{a}$ is not as rich as expected. However, this result justifies the consideration of an extended toy model based on two Fourier coefficients for b and one coefficient for c.

Particular attention has been paid to the extended toy model both for representing the level sets of $\bar{a}$ in Section 3.3 and for MCMC simulations in Section 5.1.

Nevertheless, this research addresses the fundamental phenomenon in inverse problems that uncertainty in some directions disappears and persists in others as the amount of data goes to infinity. For the Bayesian approach, this often results in the posterior concentrating around a manifold. On the one hand, this leads to the deterioration of MCMC methods as they often propose to points off the manifold. On the other hand, we showed for the inverse problem at hand that it is sometimes possible to identify an effective problem all of whose parameters can be identified as the amount of data goes to infinity. For this example, we have clearly illustrated in Section 4.2 that the distributions along the level sets of $\bar{a}$ are dominated by the prior.

For this approximation result, we need a bound on the difference between the homogenised solution $\bar{p}$ and the multiscale solution p_ϵ. The resulting bound must be sharp enough in order to be integrable with respect to the prior. This does not constitute a problem for the uniform series prior. However, we have provided a sharp enough bound that also applies to log-Gaussian priors

It might be worth developing these initial findings by considering in more detail what the effect of the dependence of $\bar{a}$ on c has on the posterior. In particular, the dependence has an impact on both the

curvature in the c-directions tangential to the level set and concentration behaviour in the c-directions that are normal to the level set.

The location of the level sets is informed by the data such that methods for solving this Bayesian inverse problem, which do not take this into account, are expected to degenerate quickly as the fine scales become finer and the observational noise goes to zero. An interesting direction to pursue further is to use this information in order to construct more efficient MCMC algorithms. It is of particular interest to generate informed proposals. A promising direction is to generate large steps along the current level set of $\bar{a}$ and only smaller steps for the transversal. This should lead to a large improvement over proposing steps that are isotropically small.

In these notes, we have assumed that both the fine scale ϵ and the forcing are known. Especially assuming ϵ to be unknown constitutes a challenging problem. It would be an interesting direction to develop MCMC algorithms to sample the marginal posterior on ϵ.

A Bound of the Homogenisation Error in L^∞

The theory of homogenisation of elliptic PDEs is a well established field with many excellent textbooks [3, 5, 12]. In this appendix, we derive an error bound of the form

$$\|p_\epsilon - \bar{p}\|_\infty \leq K\epsilon$$

for the pressure p^ϵ and and the homogenised pressure $\bar{p}$ as defined in the Equations (1) and (2) for $D = (0, 1)$. This bound is needed in order to verify the integrability assumptions in Theorem 5 which depends on the prior. A bound on K is derived with explicit dependence on

$$a_{\inf}^{-1}, a_{\max}, \|f\|_\infty, \|\partial_x a\|_\infty, \|\partial_y a\|_\infty .$$

The standard approach to prove the error bound for the Dirichlet problem uses the maximum principle and the asymptotic expansion as described in [3, p. 19]. However, an apprioriate maximum princple can be found in [16] leading to constant K which is exponentially large in $a_{\inf}$. For this reason, we use a more direct approach by considering explicit formulae for the solution

$$-\left(a(x, \frac{x}{\epsilon})p'_\epsilon\right)' = f \quad p(0) = p(1) = 0. \tag{20}$$

Integrating Equation 20 twice, we conclude that the solution is given by

$$\begin{aligned} p_\epsilon(x) &= \int_0^x -\frac{F(s) + C_\epsilon}{a(s, \frac{s}{\epsilon})} ds \text{ with} \\ C_\epsilon &= \int_0^1 -\frac{F(s)}{a(s, \frac{s}{\epsilon})} ds \left(\int_0^1 \frac{1}{a(s, \frac{s}{\epsilon})} ds\right)^{-1}. \end{aligned}$$

The solution to the homogenised equation

$$\begin{aligned} -\left(\bar{a}(x)\bar{p}'\right)' &= f \quad p(0) = p(1) = 0 \\ \bar{a}(x) &= \left(\int a(x, y)^{-1} dy\right)^{-1} \end{aligned}$$

takes the form

$$\begin{aligned} \bar{p}(x) &= \int_0^x -\frac{F(s) + \bar{C}}{\bar{a}(s)} ds \text{ with} \\ \bar{C} &= \int_0^1 -\frac{F(s)}{\bar{a}(s)} ds \left(\int_0^1 \frac{1}{\bar{a}(s)} ds\right)^{-1} \leq \frac{a_{\sup}}{a_{\inf}} \|f\|_\infty . \end{aligned}$$

Our aim is to obtain a bound of the form

$$\|\bar{p} - p_\epsilon\|_{L^\infty([0,1])} \leq K(a_{\sup}, a_{\inf}, \|\partial_x a\|_\infty, \|\partial_y a\|_\infty, \|F\|_\infty)\epsilon$$

where K is a polynomial.

Theorem 6. *The homogenisation error between p_ϵ and $\bar{p}$ satisfies*

$$\|\bar{p} - p_\epsilon\|_{L^\infty([0,1])} \leq C_{univ}(\|f\|_\infty \vee 1)^2(a_{max}^2 \vee 1)(a_{min}^{-5} \vee 1)(\|\partial_x a\|_\infty \vee 1)\epsilon. \tag{21}$$

Proof. Using the explicit formulae, we obtain

$$\begin{aligned}\bar{p} - p_\epsilon &= \int_0^x -\frac{(F(s)+\bar{C})a(s,\frac{s}{\epsilon}) - (F(s)+C_\epsilon)\bar{a}(s)}{\bar{a}(s)a(s,\frac{s}{\epsilon})}ds \\ &= \int_0^x -\frac{F(s)\left(\mathrm{a}(\mathrm{s},\frac{s}{\epsilon})\text{-}\bar{a}(\mathrm{s})\right)}{\bar{a}(s)a(s,\frac{s}{\epsilon})}ds + \int_0^x -\frac{\bar{C}\left(a(s,\frac{s}{\epsilon}) - \bar{a}(s)\right)}{\bar{a}(s)a(s,\frac{s}{\epsilon})}ds + \int_0^x -\frac{\bar{a}(s)(\bar{C}-C_\epsilon)}{\bar{a}(s)a(s,\frac{s}{\epsilon})}ds.\end{aligned}$$

We call the terms in the last line Term A, B and C, respectively. We bound Term A using $k\epsilon \leq x < (k+1)\epsilon$ so that

$$A = \int_0^x \frac{F(s)\left(\frac{\bar{a}(s)}{a(s,\frac{s}{\epsilon})}\text{-}1\right)}{\bar{a}(s)}ds = \sum_{i=1}^{k}\int_{(i-1)\epsilon}^{i\epsilon}\frac{F(s)\left(\frac{\bar{a}(s)}{a(s,\frac{s}{\epsilon})}\text{-}1\right)}{\bar{a}(s)}ds + \int_{k\epsilon}^{x}\frac{F(s)\left(\frac{\bar{a}(s)}{a(s,\frac{s}{\epsilon})}\text{-}1\right)}{\bar{a}(s)}ds. \tag{22}$$

We introduce the function $h(z_1,z_2,z_3) = \frac{z_1\left(\frac{z_2}{z_3}-1\right)}{x_2}$ and rewrite

$$\left|\int_{i\epsilon}^{(i+1)\epsilon}\frac{F(s)\left(\frac{\bar{a}(s)}{a(s,\frac{s}{\epsilon})}\text{-}1\right)}{\bar{a}(s)}ds\right| = \left|\int_{i\epsilon}^{(i+1)\epsilon} h(F(s),\bar{a}(s),a(s,\frac{s}{\epsilon}))ds\right|.$$

We note that $\nabla h(z_1,z_2,z_3) = \left(z_3^{-1} - z_2^{-1}, z_1z_2^{-2}, -z_1z_3^{-2}\right)$. Thus, we can rewrite the above for some $\xi(s)$ on the line between $(F(i\epsilon),\bar{a}(i\epsilon),a(i\epsilon,\frac{s}{\epsilon}))$ and $(F(s),\bar{a}(s),a(s,\frac{s}{\epsilon}))$ as follows

$$\left|\int_{i\epsilon}^{(i+1)\epsilon} h(F(i\epsilon),\bar{a}(i\epsilon),a(i\epsilon,\frac{s}{\epsilon}))ds + \int_{i\epsilon}^{(i+1)\epsilon}\nabla h(\xi(s))\underbrace{\left(F(s)-F(i\epsilon),\bar{a}(s)-\bar{a}(i\epsilon),a(s,\frac{s}{\epsilon})-a(i\epsilon,\frac{s}{\epsilon})\right)}_{\Delta z}ds\right|. \tag{23}$$

We note that

$$\int_{i\epsilon}^{(i+1)\epsilon} h(F(i\epsilon),\bar{a}(i\epsilon),a(i\epsilon,\frac{s}{\epsilon}))ds = F(i\epsilon)\bar{a}(i\epsilon)^{-1}\underbrace{\int_{i\epsilon}^{(i+1)\epsilon}\left(\frac{\bar{a}(i\epsilon)}{a(i\epsilon,\frac{s}{\epsilon})}\text{-}1\right)ds}_{0}.$$

Hence, it is left to bound the second summand in Equation (23). We note that

$$\begin{aligned}\|\nabla h\|_{L^\infty} &\leq C_{\mathrm{univ}}a_{\min}^{-1} \vee a_{\min}^{-2} \vee a_{\max}^{-2}\|f\|_\infty \leq C_{\mathrm{univ}}\left(a_{\min}^{-2}\vee 1\right)(\|f\|_\infty \vee 1) \\ \|\Delta z\|_{L^\infty} &\leq \epsilon\left(\|f\|_\infty \vee a_{\max}^2 a_{\min}^{-2}\|\partial_x a\|_\infty \vee \|\partial_x a\|_\infty\right).\end{aligned}$$

In this way, we obtain that

$$\left|\int_{i\epsilon}^{(i+1)\epsilon}\frac{F(s)\left(\frac{\bar{a}(s)}{a(s,\frac{s}{\epsilon})}\text{-}1\right)}{\bar{a}(s)}ds\right| \leq C_{\mathrm{univ}}\epsilon^2(\|f\|_\infty \vee 1)^2(a_{\max}^2\vee 1)(a_{\min}^{-2}\vee 1)(\|\partial_x a\|_\infty \vee 1).$$

Summing the terms, we have an error of size ϵ with explicit constant. The second summand for Term A in Equation (22) can be bounded by ϵ times the supremum of the integrand. Combining both bounds gives rise to

$$A \leq C_{\text{univ}}\epsilon(\|f\|_\infty \vee 1)^2(a_{\max}^2 \vee 1)(a_{\min}^{-4} \vee 1)(\|\partial_x a\|_\infty \vee 1).$$

The Term B and C can be bounded in a similar fashion giving rise to

$$\begin{aligned} B &\leq C_{\text{univ}}\epsilon(\|f\|_\infty \vee 1)^3(a_{\max}^3 \vee 1)(a_{\min}^{-5} \vee 1)(\|\partial_x a\|_\infty \vee 1) \\ C &\leq C_{\text{univ}}\epsilon(\|f\|_\infty \vee 1)^2(a_{\max}^2 \vee 1)(a_{\min}^{-5} \vee 1)(\|\partial_x a\|_\infty \vee 1). \end{aligned}$$

Combining the bounds for the Terms A, B and C the result follows.

References

[1] Antonio Ambrosetti and Giovanni Prodi. *A Primer of Nonlinear Analysis*, volume 34 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 1995. Corrected reprint of the 1993 original.

[2] I. Babuska, R. Tempone, and G. E. Zouraris. Galerkin Finite Element Approximations of Stochastic Elliptic Partial Differential Equations. *SIAM J. Numer. Anal.*, 42(2):800–825, 2004.

[3] Alain Bensoussan, Jacques-Louis Lions, and George Papanicolaou. *Asymptotic Analysis for Periodic Structures*, volume 5 of *Studies in Mathematics and its Applications*. North-Holland Publishing Co., Amsterdam, 1978.

[4] Vladimir I. Bogachev. *Measure Theory. Vol. I, II*. Springer-Verlag, Berlin, 2007.

[5] Doina Cioranescu and Patrizia Donato. *An Introduction to Homogenization*, volume 17 of *Oxford Lecture Series in Mathematics and its Applications*. The Clarendon Press Oxford University Press, New York, 1999.

[6] M. Dashti, S. Harris, and A. M. Stuart. Besov Priors for Bayesian Inverse Problems. *Inverse Probl. Imaging*, 6(2):183–200, 2012.

[7] M. Dashti and A. M. Stuart. Uncertainty Quantification and Weak Approximation of an Elliptic Inverse Problem. *SIAM J. Numer. Anal.*, 49:2524–2542, 2011.

[8] V. H. Hoang, C. Schwab, and A. M. Stuart. Complexity Analysis of Accelerated MCMC Methods for Bayesian Inversion. *ArXiv e-prints*, July 2012.

[9] V. H. Hoang and Ch. Schwab. Analytic Regularity and Polynomial Approximation of Stochastic, Parametric Elliptic Multiscale PDEs. 2012.

[10] Olav Kallenberg. *Foundations of Modern Probability*. Springer Verlag, 2002.

[11] James Nolen, Gregoris Pavliotis, and Andrew M. Stuart. Multiscale Modelling and Inverse Problems. In I.G. Graham, Th. Y. Hou, O. Lakkis, and R. Scheichl, editors, *Lecture Notes in Computational Science and Engineering*, volume 83. Springer.

[12] Grigorios A. Pavliotis and Andrew M. Stuart. *Multiscale Methods*, volume 53 of *Texts in Applied Mathematics*. Springer-Verlag, New York, 2008.

[13] C. Schwab and C. J. Gittelson. Sparse Tensor Discretizations of High-Dimensional Parametric and Stochastic PDEs. *Acta Numer.*, 20:291–467, 2011.

[14] C. Schwab and A. M. Stuart. Sparse Deterministic Approximation of Bayesian Inverse Problems. *Inverse Probl.*, 28(4):045003, 32, 2012.

[15] A. M. Stuart. The Bayesian Approach to Inverse Problems. *ArXiv preprint 1302.6989*, 2013.

[16] G. Sweers. Maximum Principles, a Start. `http://aw.twi.tudelft.nl/~sweers/maxpr/maxprinc.pdf`, 2000. accessed 01.09.2013.

□

Bibliography

[1] S. Agapiou, S. Larsson, and A. M. Stuart. Posterior Contraction Rates for the Bayesian Approach to Linear Ill-Posed Inverse Problems. *Stochastic Process. Appl.*, 123(10):3828 – 3860, 2013.

[2] S. Agapiou, A. M. Stuart, and Y.-X. Zhang. Bayesian Posterior Contraction Rates for Linear Severely Ill-posed Inverse Problems. *ArXiv e-prints*, October 2012.

[3] G. Alessandrini. Stable Determination of Conductivity by Boundary Measurements. *Appl. Anal.*, 27(1-3):153–172, 1988.

[4] Y. Amit. Convergence Properties of the Gibbs Sampler for Perturbations of Gaussians. *Ann. Statist.*, 24(1):122–140, 1996.

[5] C. Andrieu, E. Moulines, and P. Priouret. Stability of Stochastic Approximation under Verifiable Conditions. *SIAM J. Control Optim.*, 44(1):283–312, 2005.

[6] A. Apte, C. K. R. T. Jones, A. M. Stuart, and J. Voss. Data Assimilation: Mathematical and Statistical Perspectives. *Int. J. Num. Meth. Fluids*, 56:1033–1046, 2008.

[7] Søren Asmussen and Peter W. Glynn. *Stochastic Simulation: Algorithms and Analysis*. Springer, 2007.

[8] George Backus. Inference from Inadequate and Inaccurate Data. I, II. *Proc. Nat. Acad. Sci. U.S.A. 65 (1970), 1–7; ibid.*, 65:281–287, 1970.

[9] George Backus. Inference from Inadequate and Inaccurate Data. III. *Proc. Nat. Acad. Sci. U.S.A.*, 67:282–289, 1970.

[10] George Backus. Inference from Inadequate and Inaccurate Data. In *Mathematical problems in the geophysical sciences, Vol. 2. Inverse problems, dynamo theory, and tides*, pages 1–105. Lectures in Appl. Math., Vol. 14. Amer. Math. Soc., Providence, R. I., 1971.

[11] D. Bakry, P. Cattiaux, and A. Guillin. Rate of Convergence for Ergodic Continuous Markov Processes: Lyapunov versus Poincaré. *J. Funct. Anal.*, 254(3):727–759, February 2008.

[12] Dominique Bakry. Functional Inequalities for Markov Semigroups. In *Probability measures on groups: recent directions and trends*, pages 91–147. Tata Inst. Fund. Res., Mumbai, 2006.

[13] P. H. Baxendale. Renewal Theory and Computable Convergence Rates for Geometrically Ergodic Markov Chains. *Ann. Appl. Probab.*, 15(1B):700–738, 2005.

[14] Jacob Bear and Alexander H-D Cheng. *Modeling Groundwater Flow and Contaminant Transport*. Springer Dordrecht, Heidelberg, London, New York, 2010.

[15] M. Bédard. Weak Convergence of Metropolis Algorithms for Non-I.I.D. Target Distributions. *Ann. Appl. Probab.*, 17(4):1222–1244, 2007.

[16] W. Bednorz. The Kendall's Theorem and its Application to the Geometric Ergodicity of Markov Chains. *arXiv preprint arXiv:1301.1481*, 2013.

[17] Alain Bensoussan, Jacques-Louis Lions, and George Papanicolaou. *Asymptotic Analysis for Periodic Structures*, volume 5 of *Studies in Mathematics and its Applications*. North-Holland Publishing Co., Amsterdam, 1978.

[18] A. Beskos, K. Kalogeropoulos, and E. Pazos. Advanced MCMC Methods for Sampling on Diffusion Pathspace. *Stoch. Proc. Appl.*, 123(4):1415 – 1453, 2013.

[19] A. Beskos, N. Pillai, G. O. Roberts, J.-M. Sanz-Serna, and A. M. Stuart. Optimal Tuning of Hybrid Monte-Carlo. to appear.

[20] A. Beskos, F. Pinski, J.-M. Sanz-Serna, and A. M.Stuart. Hybrid Monte-Carlo on Hilbert Spaces. *Stoch. Proc. Appl.*, 121:2201–2230, 2011.

[21] A. Beskos, G. O. Roberts, and A. M. Stuart. Optimal Scalings for Local Metropolis-Hastings Chains on Nonproduct Targets in High Dimensions. *Ann. Appl. Probab.*, 19:863–898, 2009.

[22] A. Beskos, G. O. Roberts, A. M. Stuart, and J. Voss. MCMC Methods for Diffusion Bridges. *Stoch. Dyn.*, 8(3):319–350, 2008.

[23] L. Birgé. Approximation dans les espaces métriques et théorie de l'estimation. *Z. Wahrsch. Verw. Gebiete*, 65(2):181–237, 1983.

[24] N. Bochkina. Consistency of the Posterior Distribution in Generalised Linear Inverse Problems. *Inverse Probl.*, 29(9), 2013.

[25] N. Bochkina and P. J. Green. The Bernstein-von Mises Theorem for Non-Regular Generalised Linear Inverse Problems. *submitted to Ann. Statist.*, pages 1–47, 2012.

[26] Vladimir I. Bogachev. *Gaussian Measures*, volume 62 of *Mathematical Surveys and Monographs*. Amer. Math. Soc., Providence, RI, 1998.

[27] Vladimir I. Bogachev. *Measure Theory. Vol. I, II.* Springer-Verlag, Berlin, 2007.

[28] Vivek S. Borkar. *Stochastic Approximation: A Dynamical Systems Viewpoint.* Cambridge University Press Cambridge, 2008.

[29] Steve Brooks, Andrew Gelman, Galin Jones, and Xiao-Li Meng. *Handbook of Markov Chain Monte Carlo.* Chapman and Hall/CRC, 2011.

[30] O. Butkovsky. Subgeometric Rates of Convergence of Markov Processes in the Wasserstein Metric. *submitted to Ann. Appl. Probab.*, pages 1–23, 2013.

[31] L. Cavalier. Nonparametric Statistical Inverse Problems. *Inverse Probl.*, 24(3):034004, June 2008.

[32] J. Cheeger. A Lower Bound for the Smallest Eigenvalue of the Laplacian. In *Problems in analysis*, pages 195–199. Princeton Univ. Press, Princeton, N. J., 1970.

[33] T. Choi. Asymptotic Properties of Posterior Distributions in Nonparametric Regression with Non-Gaussian Errors. *Ann. I. Stat. Math.*, 98(4):835–859, February 2008.

[34] T. Choi and M. J. Schervish. On Posterior Consistency in Nonparametric Regression Problems. *J. Multivariate Anal.*, 98(10):1969–1987, 2007.

[35] Doina Cioranescu and Patrizia Donato. *An Introduction to Homogenization*, volume 17 of *Oxford Lecture Series in Mathematics and its Applications*. The Clarendon Press Oxford University Press, New York, 1999.

[36] S. L. Cotter, M. Dashti, J. C. Robinson, and A. M Stuart. Bayesian Inverse Problems for Functions and Applications to Fluid Mechanics. *Inverse Probl.*, 25:115008, 2009.

[37] S. L. Cotter, M. Dashti, and A. M. Stuart. Approximation of Bayesian Inverse Problems for PDEs. *SIAM J. Numer. Anal.*, 48(1):322–345, 2010.

[38] S. L. Cotter, G. O. Roberts, A. M. Stuart, and D. White. MCMC Methods for Functions: Modifying Old Algorithms to Make Them Faster. *ArXiv preprint 1202.0709*, 2011. to appear Stat. Sci.

[39] D. D. Cox. An Analysis of Bayesian Inference for Nonparametric Regression. *Ann. Statist.*, 21(2):903–923, 1993.

[40] Giuseppe Da Prato and Jerzy Zabczyk. *Stochastic Equations in Infinite Dimensions*, volume 44 of *Encyclopedia of Mathematics and its Applications*. Cambridge University Press, Cambridge, 1992.

[41] S. R. Dalal. Dirichlet Invariant Processes and Applications to Nonparametric Estimation of Symmetric Distribution Functions. *Stoch. Proc. Appl.*, 9(1):99–107, 1979.

[42] S. R. Dalal, Jr. Hall, and J. Gaineford. On Approximating Parametric Bayes' Models by Nonparametric Bayes Models. *Ann. Statist.*, 8(3):664–672, 1980.

[43] Siddhartha R. Dalal. Nonparametric and Robust Bayes Estimation of Location. In *Optimizing Methods in Statistics (Proc. Internat. Conf., Indian Inst. Tech., Bombay, 1977)*, pages 141–166. Academic Press, New York, 1979.

[44] M. Dashti, S. Harris, and A. M. Stuart. Besov Priors for Bayesian Inverse Problems. *Inverse Probl. Imaging*, 6(2):183–200, 2012.

[45] M. Dashti, K. J. H. Law, A. M. Stuart, and J. Voss. MAP Estimators and Posterior Consistency in Bayesian Nonparametric Inverse Problems. *ArXiv preprint 1303.4795*, 2013.

[46] M. Dashti and A. M. Stuart. Uncertainty Quantification and Weak Approximation of an Elliptic Inverse Problem. *SIAM J. Numer. Anal.*, 49:2524–2542, 2011.

[47] K. Deckelnick and M. Hinze. Convergence and Error Analysis of a Numerical Method for the Identification of Matrix Parameters in Elliptic PDEs. *Inverse Probl.*, 28(11):115015, 15, 2012.

[48] Pierre Del Moral. *Feynman-Kac formulae*. Probability and its Applications (New York). Springer-Verlag, New York, 2004. Genealogical and interacting particle systems with applications.

[49] Pierre Del Moral. *Mean field simulation for Monte Carlo integration*, volume 126 of *Monographs on Statistics and Applied Probability*. CRC Press, Boca Raton, FL, 2013.

[50] P. Diaconis and D. A. Freedman. On Inconsistent Bayes Estimates of Location. *Ann. Statist.*, 14(1):68–87, 1986.

[51] P. Diaconis and D. A. Freedman. On the Consistency of Bayes Estimates. *Ann. Statist.*, 14(1):1–67, 1986. With a discussion and a rejoinder by the authors.

[52] P. Diaconis and L. Saloff-Coste. Logarithmic Sobolev Inequalities for Finite Markov Chains. *Ann. Appl. Probab.*, 6(3):695–750, 1996.

[53] P. Diaconis and M. Shahshahani. Generating a Random Permutation with Random Transpositions. *Z. Wahrsch. Verw. Gebiete*, 57(2):159–179, 1981.

[54] P. Diaconis and D. Stroock. Geometric Bounds for Eigenvalues of Markov Chains. *Ann. Appl. Probab.*, 1(1):36–61, 1991.

[55] Persi Diaconis. Mathematical Developments from the Analysis of Riffle Shuffling. In *Groups, combinatorics & geometry (Durham, 2001)*, pages 73–97. World Sci. Publ., River Edge, NJ, 2003.

[56] Persi Diaconis and Laurent Saloff-Coste. Comparison theorems for reversible Markov chains. *Ann. Appl. Probab.*, 3(3):696–730, 1993.

[57] J. Dick, D. Rudolf, and H. Zhu. Discrepancy Bounds for Uniformly Ergodic Markov Chain Quasi-Monte Carlo. *ArXiv e-prints*, 2013.

[58] Josef Dick and Friedrich Pillichshammer. *Digital Nets and Sequences: Discrepancy Theory and Quasi-Monte Carlo Integration*. Cambridge University Press, 2010.

[59] W. Doeblin. Sur deux problèmes de M. Kolmogoroff concernant les chaînes dénombrables. *Bull. Soc. Math. France*, 66:210–220, 1938.

[60] Joseph L. Doob. Application of the Theory of Martingales. In *Le Calcul des Probabilités et ses Applications*, Colloques Internationaux du Centre National de la Recherche Scientifique, no. 13, pages 23–27. Centre National de la Recherche Scientifique, Paris, 1949.

[61] P. Dostert, Y. Efendiev, T. Y. Hou, and W. Luo. Coarse-Gradient Langevin Algorithms for Dynamic Data Integration and Uncertainty Quantification. *J. Comput. Phys.*, 217(1):123–142, 2006.

[62] Arnaud Doucet, Nando De Freitas, Neil Gordon, et al. *Sequential Monte Carlo Methods in Practice*, volume 1. Springer New York, 2001.

[63] Richard M. Dudley. *Real Analysis and Probability*, volume 74. Cambridge University Press, 2002.

[64] A. Eberle. Metropolis-Hastings Algorithms for Perturbations of Gaussian Measures in High Dimensions: Contraction Properties and Error Bounds in the Log-concave Case. *ArXiv e-prints*, October 2012.

[65] Y. Efendiev, T. Hou, and W. Luo. Preconditioning Markov Chain Monte Carlo Simulations Using Coarse-Scale Models. *SIAM J. Sci. Comput.*, 28(2):776–803 (electronic), 2006.

[66] T. A. El Moselhy and Y. M. Marzouk. Bayesian Inference with Optimal Maps. *J. Comput. Phys.*, 231(23):7815–7850, 2012.

[67] Heinz W. Engl, Martin Hanke, and Andreas Neubauer. *Regularization of Inverse Problems*, volume 375 of *Mathematics and its Applications*. Kluwer Academic Publishers Group, Dordrecht, 1996.

[68] Lawrence C. Evans. *Partial Differential Equations*, volume 19 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, second edition, 2010.

[69] G. Fort and G. O. Roberts. Subgeometric Ergodicity of Strong Markov Processes. *Ann. Appl. Probab.*, 15(2):1565–1589, 2005.

[70] J. N. Franklin. Well-Posed Stochastic Extensions of Ill-Posed Linear Problems. *J. Math. Anal. Appl.*, 31:682–716, 1970.

[71] D. A. Freedman. On the Asymptotic Behavior of Bayes' Estimates in the Discrete Case. *Ann. Math. Statist.*, 34:1386–1403, 1963.

[72] D. A. Freedman. On the Asymptotic Behavior of Bayes' Estimates in the Discrete Case. II. *Ann. Math. Statist.*, 36:454–456, 1965.

[73] Masatoshi Fukushima, Yoichi Oshima, and Masayoshi Takeda. *Dirichlet forms and symmetric Markov processes*, volume 19 of *de Gruyter Studies in Mathematics*. Walter de Gruyter & Co., Berlin, extended edition, 2011.

[74] S. Ghosal, J. K. Ghosh, and A. W. van der Vaart. Convergence Rates of Posterior Distributions. *Ann. Statist.*, 28(2):500–531, 2000.

[75] S. Ghosal and A. van der Vaart. Convergence Rates of Posterior Distributions for Non-i.i.d. Observations. *Ann. Statist.*, 35(1):192–223, 2007.

[76] S. Ghosal and A. van der Vaart. Fundamentals of Nonparametric Bayesian Inference. unpublished, 2012.

[77] David Gilbarg and Neil S. Trudinger. *Elliptic Partial Differential Equations of Second Order*. Classics in Mathematics. Springer-Verlag, Berlin, 2001. Reprint of the 1998 edition.

[78] M. B. Giles. Multilevel Monte Carlo Path Simulation. *Oper. Res.*, 56(3):607–617, 2008.

[79] Mark Girolami and Ben Calderhead. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 73(2):123–214, 2011. With discussion and a reply by the authors.

[80] Alice Guionnet and Boguslaw Zegarlinski. *Lectures on Logarithmic Sobolev Inequalities*, volume 1801 of *Séminaire de Probabilités, XXXVI*. Springer, 2002.

[81] O. Häggström and J. S. Rosenthal. On Variance Conditions for Markov Chain CLTs. *Electron. Comm. Probab.*, 12:454–464 (electronic), 2007.

[82] M. Hairer. An Introduction to Stochastic PDEs. Lecture Notes, 2009.

[83] M. Hairer and J. C. Mattingly. Spectral Gaps in Wasserstein Distances and the 2D Stochastic Navier-Stokes Equations. *Ann. Probab.*, 36(6):2050–2091, 2008.

[84] M. Hairer, J. C. Mattingly, and M. Scheutzow. Asymptotic Coupling and a General Form of Harris' Theorem with Applications to Stochastic Delay Equations. *Probab. Theory Related Fields*, 149(1-2):223–259, 2011.

[85] M. Hairer, A. M. Stuart, and J. Voss. Analysis of SPDEs Arising in Path Sampling. Part ii: The Nonlinear Case. *Ann. Appl. Probab.*, pages 1657–1706, 2007.

[86] Martin Hairer, Andrew M. Stuart, and Jochen Voss. Sampling Conditioned Diffusions. In *Trends in stochastic analysis*, volume 353 of *London Math. Soc. Lecture Note Ser.*, pages 159–185. Cambridge Univ. Press, Cambridge, 2009.

[87] Martin Hairer, Andrew M. Stuart, and Jochen Voss. Signal Processing Problems on Function Space: Bayesian Formulation, Stochastic PDEs and Effective MCMC Methods. In *The Oxford handbook of nonlinear filtering*, pages 833–873. Oxford Univ. Press, Oxford, 2011.

[88] Theodore Edward Harris. The Existence of Stationary Measures for Certain Markov Processes. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, volume 2, pages 113–124. University of California Press Berkeley, 1956.

[89] W. K. Hastings. Monte-Carlo Sampling Methods Using Markov Chains and their Applications. *Biometrika*, 57(1):97, 1970.

[90] Michael Hinze, Rene Pinnau, Michael Ulbrich, and Stefan Ulbrich. *Optimization with PDE Constraints. Mathematical Modelling: Theory and Applications*, volume 23. Springer Verlag Berlin, 2009.

[91] Nils Lid Hjort, Chris Holmes, Peter Müller, and Stephen G. Walker, editors. *Bayesian Nonparametrics*, volume 28 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 2010.

[92] V. H. Hoang, C. Schwab, and A. M. Stuart. Complexity Analysis of Accelerated MCMC Methods for Bayesian Inversion. *ArXiv e-prints*, July 2012.

[93] S. F. Jarner and E. Hansen. Geometric Ergodicity of Metropolis Algorithms. *Stoch. Proc. Appl.*, 85(2):341 – 361, 2000.

[94] S. F. Jarner and G. O. Roberts. Polynomial Convergence Rates of Markov Chains. *Ann. Appl. Probab.*, 12(1):224–247, 2002.

[95] S. F. Jarner and R. L. Tweedie. Necessary Conditions for Geometric and Polynomial Ergodicity of Random-Walk-Type Markov Chains. *Bernoulli*, 9(4):559–578, 2003.

[96] S. F. Jarner and W. K. Yuen. Conductance Bounds on the L^2-Convergence Rate of Metropolis Algorithms on Unbounded State Spaces. *Adv. in Appl. Probab.*, 36(1):243–266, 2004.

[97] L. T. Johnson and C. J. Geyer. Variable Transformation to Obtain Geometric Ergodicity in the Random-Walk Metropolis Algorithm. *Ann. Statist.*, 40(6):3050–3076, February 2012 2012 2012.

[98] A. Joulin and Y. Ollivier. Curvature, Concentration and Error Estimates for Markov Chain Monte Carlo. *Ann. Probab.*, 38(6):2418–2442, 2010.

[99] Jari Kaipio and Erkki Somersalo. *Statistical and Computational Inverse Problems*, volume 160 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 2005.

[100] Olav Kallenberg. *Foundations of Modern Probability*. Springer Verlag, 2002.

[101] N. Kantas, A. Beskos, and A. Jasra. Sequential Monte Carlo Methods for High-Dimensional Inverse Problems: A Case Study for the Navier-Stokes Equations. *ArXiv e-prints*, July 2013.

[102] David G. Kendall. Unitary Dilations of Markov Transition Operators, and the Corresponding Integral Representations for Transition-Probability Matrices. In *Probability and statistics: The Harald Cramér volume (edited by Ulf Grenander)*, pages 139–161. Almqvist & Wiksell, Stockholm, 1959.

[103] C. Ketelsen, R. Scheichl, and A. L. Teckentrup. A Hierarchical Multilevel Markov Chain Monte Carlo Algorithm with Applications to Uncertainty Quantification in Subsurface Flow. *ArXiv e-prints*, March 2013.

[104] Y. Kim. The Bernstein-von Mises Theorem for the Proportional Hazard Model. *Ann. Statist.*, 34(4):1678–1700, 2006.

[105] Y. Kim and J. Lee. A Bernstein-von Mises Theorem in the Nonparametric Right-Censoring Model. *Ann. Statist.*, 32(4):1492–1512, 2004.

[106] C. Kipnis and S. R. S. Varadhan. Central Limit Theorem for Additive Functionals of Reversible Markov Processes and Applications to Simple Exclusions. *Comm. Math. Phys.*, 104(1):1–19, 1986.

[107] Alexandre A. Kirillov and Alexej D. Gvishiani. *Theorems and Problems in Functional Analysis*. Problem Books in Mathematics. Springer-Verlag, New York, 1982. Translated from the Russian by Harold H. McFaden.

[108] B. J. K. Kleijn. Semiparametric Posterior Limits. *arXiv preprint arXiv:1305.4836*, pages 1–47, May 2013.

[109] Peter E. Kloeden and Eckhard Platen. *Numerical Solution of Stochastic Differential Equations*. Springer, 1992.

[110] B. Knapik. *Bayesian Asymptotics Inverse Problems and Irregular Models*. PhD thesis, Vrije Universiteit Amsterdam, 2013.

[111] B. T. Knapik, B. T. Szabó, A. W. van der Vaart, and J. H. van Zanten. Bayes Procedures for Adaptive Inference in Nonparametric Inverse Problems. *arXiv preprint arXiv:1209.3628*, 2012.

[112] B. T. Knapik, A. W. van der Vaart, and J. H. van Zanten. Bayesian Inverse Problems with Gaussian Priors. *Ann. Statist.*, 39(5):2626–2657, 2011.

[113] B. T. Knapik, A. W. van der Vaart, and J. H. van Zanten. Bayesian Recovery of the Initial Condition for the Heat Equation. *Arxiv preprint arXiv:1111.5876*, 2011.

[114] T. Komorowski, S. Peszat, and T. Szarek. On Ergodicity of some Markov Processes. *Ann. Appl. Probab.*, 38(4):1401–1443, 2010.

[115] T. Komorowski and A. Walczuk. Central Limit Theorem for Markov Processes with Spectral Gap in the Wasserstein Metric. *Stoch. Proc. Appl.*, 122(5):2155–2184, 2012.

[116] S. Krumscheid, G. A. Pavliotis, and S. Kalliadasis. Semiparametric Drift and Diffusion Estimation for Multiscale Diffusions. *Multiscale Model. Simul.*, 11(2):442–473, 2013.

[117] K. Kunisch. Numerical Methods for Parameter Estimation Problems Inverse Problems in Diffusion Processes. In *Proc. GAMM-SIAM Symp. (Philadelphia, PA: SIAM)*, pages 199–216, 1995.

[118] S. Lasanen. Non-Gaussian Statistical Inverse Problems. Part I: Posterior Distributions. *Inverse Probl. Imaging*, 6(2):215–266, 2012.

[119] S. Lasanen. Non-Gaussian Statistical Inverse Problems. Part II: Posterior Convergence for Approximated Unknowns. *Inverse Probl. Imaging*, 6(2):267–287, 2012.

[120] K. Łatuszyński, B. Miasojedow, and W. Niemiro. Nonasymptotic Bounds on the Estimation Error of MCMC Algorithms. *Bernoulli to appear*, 2013 2013 2013.

[121] K. Łatuszyński and W. Niemiro. Rigorous Confidence Bounds for MCMC under a Geometric Drift Condition. *J. Complexity*, 27(1):23–38, 2011.

[122] K. Łatuszyński and G. O. Roberts. CLTs and Asymptotic Variance of Time-Sampled Markov Chains. *Methodol. Comput. Appl. Probab.*, 2011.

[123] G. F. Lawler and A. D. Sokal. Bounds on the L^2-Spectrum for Markov Chains and Markov Processes: A Generalization of Cheeger's Inequality. *Amer. Math. Society*, 309(2), 1988.

[124] Lucien Le Cam. *Asymptotic Methods in Statistical Decision Theory.* Springer Series in Statistics. Springer-Verlag, New York, 1986.

[125] Lucien Le Cam and Grace Lo Yang. *Asymptotics in Statistics.* Springer Series in Statistics. Springer-Verlag, New York, 1990. Some basic concepts.

[126] David Asher Levin, Yuval Peres, and Elizabeth Lee Wilmer. *Markov Chains and Mixing Times.* AMS Bookstore, 2009.

[127] L. Lovász and M. Simonovits. Random Walks in a Convex Body and an Improved Volume Algorithm. *Random Structures Algorithms*, 4(4):359–412, 1993.

[128] K. J. H. Law M. A. Iglesias and A. M. Stuart. Evaluation of Gaussian Approximations for Data Assimilation in Reservoir Models. Submitted.

[129] Neal Madras and Dana Randall. Factoring Graphs to Bound Mixing Rates. In *37th Annual Symposium on Foundations of Computer Science (Burlington, VT, 1996)*, pages 194–203. IEEE Comput. Soc. Press, Los Alamitos, CA, 1996.

[130] P. Mathé and E. Novak. Simple Monte Carlo and the Metropolis Algorithm. *J. Complexity*, 23(4-6):673–696, 2007.

[131] J. C. Mattingly, N. S. Pillai, and A. M. Stuart. Diffusion Limits of the Random Walk Metropolis Algorithm in High Dimensions. *Ann. Appl. Probab.*, 22(3):881–930, 2012.

[132] J. C. Mattingly, A. M. Stuart, and D. J. Higham. Ergodicity for SDEs and Approximations: Locally Lipschitz Vector Fields and Degenerate Noise. *Stoch. Proc. Appl.*, 101(2):185–232, 2002.

[133] J. C. Mattingly, A. M. Stuart, and M. V. Tretyakov. Convergence of Numerical Time-Averaging and Stationary Measures via Poisson Equations. *SIAM J. Numer. Anal.*, 48(2):552–577, 2010.

[134] D. McLaughlin and L. R. Townley. A Reassessment of the Groundwater Inverse Problem. *Water Resour. Res.*, 32(5):1131–1161, 1996.

[135] K. L. Mengersen and R. L. Tweedie. Rates of Convergence of the Hastings and Metropolis Algorithms. *Ann. Statist.*, 24(1):101–121, 1996.

[136] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, E. Teller, et al. Equation of State Calculations by Fast Computing Machines. *J. Chem. Phys.*, 21(6):1087, 1953.

[137] Sean Meyn and Richard L. Tweedie. *Markov Chains and Stochastic Stability*. Cambridge University Press, Cambridge, second edition, 2009. With a prologue by Peter W. Glynn.

[138] S. Nazarov. Homogenization of Elliptic Systems with Periodic Coefficients: Weighted l^p and l^∞Estimates for Aymptotic Remainders. *St. Petersburg Math. J.*, 18(2):269–304, 2007.

[139] R. M. Neal. Improving Asymptotic Variance of MCMC Estimators: Non-Reversible Chains are Better. Technical report, Dept. of Stat., University of Toronto., 2004.

[140] Radford M. Neal. Regression and Classification using Gaussian Process Priors. In *Bayesian statistics, 6 (Alcoceber, 1998)*, pages 475–501. Oxford Univ. Press, New York, 1999.

[141] A. Neubauer and H. K. Pikkarainen. Convergence Results for the Bayesian Inversion Theory. *J. Inverse Ill-Posed Probl.*, 16(6):601–613, 2008.

[142] W. Niemiro and P. Pokarowski. Fixed Precision MCMC Estimation by Median of Products of Averages. *J. Appl. Probab.*, 46(2):309–329, 2009.

[143] J. Nolen and G. Papanicolaou. Fine Scale Uncertainty in Parameter Estimation for Elliptic Equations. *Inverse Probl.*, 25(11):115021, 2009.

[144] James Nolen, Gregoris Pavliotis, and Andrew M. Stuart. Multiscale Modelling and Inverse Problems. In I.G. Graham, Th. Y. Hou, O. Lakkis, and R. Scheichl, editors, *Lecture Notes in Computational Science and Engineering*, volume 83. Springer, 2010.

[145] E. Nummelin and R. L. Tweedie. Geometric Ergodicity and R-Positivity for General Markov Chains. *Ann. Probability*, 6(3):404–420, 1978.

[146] Y. Ollivier. Ricci Curvature of Markov Chains on Metric Spaces. *J. Funct. Anal.*, 256(3):810–864, 2009.

[147] M. Ottobre, N. S. Pillai, F. J. Pinski, and A. M. Stuart. A Function Space HMC Algorithm with Second Order Langevin Diffusion Limit. *ArXiv e-prints*, August 2013.

[148] Matthew David Parno. A Multiscale Framework for Bayesian Inference in Elliptic Problems. Master's thesis, Massachusetts Institute of Technology, 2011.

[149] Grigorios A. Pavliotis and Andrew M. Stuart. *Multiscale Methods*, volume 53 of *Texts in Applied Mathematics*. Springer-Verlag, New York, 2008.

[150] N. S. Pillai, A. M. Stuart, and A. H. Thiéry. Gradient Flow from a Random Walk in Hilbert Space. *ArXiv e-prints*, August 2011.

[151] N. S. Pillai, A. M. Stuart, and A. H. Thiéry. On the Random Walk Metropolis Algorithm for Gaussian Random Field Priors and the Gradient Flow. submitted, 2011.

[152] N. S. Pillai, A. M. Stuart, and A. H. Thiéry. Optimal Scaling and Diffusion Limits for the Langevin Algorithm in High Dimensions. *Ann. Appl. Probab.*, 22(6):2320–2356, 2012.

[153] K. Ray. Bayesian Inverse Problems with Non-Conjugate Priors. *arXiv preprint arXiv:1209.6156*, 2012.

[154] G. R. Richter. An Inverse Problem for the Steady State Diffusion Equation. *SIAM J. Appl. Math.*, 41(2):210–221, 1981.

[155] H. Robbins and S. Monro. A Stochastic Approximation Method. *Ann. Math. Statistics*, 22:400–407, 1951.

[156] Christian P. Robert and George Casella. *Monte Carlo Statistical Methods*. Springer Texts in Statistics. Springer-Verlag, New York, second edition, 2004.

[157] G. O. Roberts, A. Gelman, and W. R. Gilks. Weak Convergence and Optimal Scaling of Random Walk Metropolis Algorithms. *Ann. Appl. Probab.*, 7(1):110–120, 1997.

[158] G. O. Roberts and J. S. Rosenthal. Geometric Ergodicity and Hybrid Markov Chains. *Electron. Comm. Probab*, 2:13–25, 1997.

[159] G. O. Roberts and J. S. Rosenthal. Optimal Scaling for Various Metropolis-Hastings Algorithms. *Statist. Sci.*, 16(4):351–367, 2001.

[160] G. O. Roberts and J. S. Rosenthal. General State Space Markov Chains and MCMC Algorithms. *Probab. Surv.*, 1:20–71, 2004.

[161] G. O Roberts and O. Stramer. On Inference for Partially Observed Nonlinear Diffusion Models Using the Metropolis-Hastings Algorithm. *Biometrika*, 88(3):603–621, 2001.

[162] G. O. Roberts and R. L. Tweedie. Exponential Convergence of Langevin Distributions and their Discrete Approximations. *Bernoulli*, pages 341–363, 1996.

[163] G. O. Roberts and R. L. Tweedie. Geometric Convergence and Central Limit Theorems for Multidimensional Hastings and Metropolis Algorithms. *Biometrika*, 83(1):95, 1996.

[164] G. O. Roberts and R. L. Tweedie. Geometric L2 and L1 Convergence are Equivalent for Reversible Markov Chains. *J. Appl. Probab.*, 38(2001):37–41, 2001.

[165] Jeffrey S. Rosenthal. Minorization conditions and convergence rates for Markov chain Monte Carlo. *J. Amer. Statist. Assoc.*, 90(430):558–566, 1995.

[166] Jeffrey S. Rosenthal. Quantitative convergence rates of Markov chains: a simple account. *Electron. Comm. Probab.*, 7:123–128 (electronic), 2002.

[167] D. Rudolf. Explicit Error Bounds for Markov Chain Monte Carlo. *Dissertationes Math. (Rozprawy Mat.)*, 485:1–93, 2012.

[168] C. Schillings and C. Schwab. Sparse, Adaptive Smolyak Quadratures for Bayesian Inverse Problems. *Inverse Probl.*, 29(6):065011, 2013.

[169] Byron Schmuland. Dirichlet forms: some infinite-dimensional examples. *Canad. J. Statist.*, 27(4):683–700, 1999.

[170] C. Schwab and A. M. Stuart. Sparse Deterministic Approximation of Bayesian Inverse Problems. *Inverse Probl.*, 28(4):045003, 32, 2012.

[171] L. Schwartz. On Bayes Procedures. *Z. Wahrsch. Verw. Gebiete*, 4:10–26, 1965.

[172] C. Sherlock. Optimal Scaling of the Random Walk Metropolis: General Criteria for the 0.234 Acceptance Rule. *J. Appl. Probab.*, 50(1):1–15, 2013.

[173] C. Sherlock, P. Fearnhead, and G. O. Roberts. The Random Walk metropolis: Linking Theory and Practice through a Case Study. *Stat. Sci.*, 25(2):172–190, 2010.

[174] C. Sherlock and G. O. Roberts. Optimal Scaling of the Random Walk Metropolis on Elliptically Symmetric Unimodal Targets. *Bernoulli*, 15(3):774–798, August 2009.

[175] A. Sinclair and M. Jerrum. Approximate Counting, Uniform Generation and Rapidly Mixing Markov Chains. *Inform. and Comput.*, 82(1):93–133, 1989.

[176] A. M. Stuart. Inverse Problems: A Bayesian Perspective. *Acta Numer.*, 19:451–559, 2010.

[177] A. M. Stuart. The Bayesian Approach to Inverse Problems. *ArXiv preprint 1302.6989*, 2013.

[178] G. Sweers. Maximum Principles, a Start. http://aw.twi.tudelft.nl/~sweers/maxpr/maxprinc.pdf, 2000. accessed 01.09.2013.

[179] Albert Tarantola. *Inverse Problem Theory and Methods for Model Parameter Estimation*. SIAM, Philadelphia, PA, 2005.

[180] Alexandre H. Thiéry. *Scaling Analysis of MCMC Algorithms*. PhD thesis, University of Warwick, 2013.

[181] L. Tierney. A Note on Metropolis-Hastings Kernels for General State Spaces. *Ann. Appl. Probab.*, 8(1):1–9, 1998.

[182] A. W. Van der Vaart and J. H. Van Zanten. Rates of Contraction of Posterior Distributions Based on Gaussian Process Priors. *Ann. Statist.*, 36(3):1435–1463, 2008.

[183] Aad W. van der Vaart. *Asymptotic Statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 1998.

[184] Santosh Vempala. Geometric Random Walks: A Survey. In *Combinatorial and Computational Geometry*, volume 52 of *Math. Sci. Res. Inst. Publ.*, pages 577–616. Cambridge Univ. Press, Cambridge, 2005.

[185] Cédric Villani. *Optimal Transport: Old and New*, volume 338. Springer-Verlag, 2009.

[186] F.-Y. Wang. Functional Inequalities for the Decay of Sub-Markov Semi-Groups. *Potential Anal.*, 18(1):1–23, 2003.

[187] Fengyu Wang. *Functional Inequalities, Markov Semigroups and Spectral Theory.* Elsevier, 2006.

[188] L. Wang and J. Zou. Error Estimates of Finite Element Methods for Parameter Identification Problems in Elliptic and Parabolic Systems. *Discrete Contin. Dyn. Syst. Ser. B*, 14:1641–1670, 2010.