Approximate inference in astronomy

Philipp Florian Frank

München 2021

Approximate inference in astronomy

Philipp Florian Frank

Dissertation
der Fakultät für Physik
der Ludwig-Maximilians-Universität
München

vorgelegt von
Philipp Florian Frank
aus München

München, den 23.08.2021

Erstgutachter: PD. Dr. Torsten Enßlin
Zweitgutachter: PD. Dr. Martin Kerscher
Tag der mündlichen Prüfung: 27.10.2021

Abstract

This thesis utilizes the rules of probability theory and Bayesian reasoning to perform inference about astrophysical quantities from observational data, with a main focus on the inference of dynamical systems extended in space and time. The necessary assumptions to successfully solve such inference problems in practice are discussed and the resulting methods are applied to real world data. These assumptions range from the simplifying prior assumptions that enter the inference process up to the development of a novel approximation method for resulting posterior distributions.

The prior models developed in this work follow a maximum entropy principle by solely constraining those physical properties of a system that appear most relevant to inference, while remaining uninformative regarding all other properties. To this end, prior models that only constrain the statistically homogeneous space-time correlation structure of a physical observable are developed. The constraints placed on these correlations are based on generic physical principles, which makes the resulting models quite flexible and allows for a wide range of applications. This flexibility is verified and explored using multiple numerical examples, as well as an application to data provided by the Event Horizon Telescope about the center of the galaxy M87. Furthermore, as an advanced and extended form of application, a variant of these priors is utilized within the context of simulating partial differential equations. Here, the prior is used in order to quantify the physical plausibility of an associated numerical solution, which in turn improves the accuracy of the simulation. The applicability and implications of this probabilistic approach to simulation are discussed and studied using numerical examples.

Finally, utilizing such prior models paired with the vast amount of observational data provided by modern telescopes, results in Bayesian inference problems that are typically too complex to be fully solvable analytically. Specifically, most resulting posterior probability distributions become too complex, and therefore require a numerical approximation via a simplified distribution. To improve upon existing methods, this work proposes a novel approximation method for posterior probability distributions: the geometric Variational Inference (geoVI) method. The approximation capacities of geoVI are theoretically established and demonstrated using numerous numerical examples. These results suggest a broad range of applicability as the method provides a decrease in approximation errors compared to state of the art methods at a moderate level of computational costs.

Zusammenfassung

Diese Dissertation verwendet die Regeln der Wahrscheinlichkeitstheorie und Bayes'scher Logik, um astrophysikalische Größen aus Beobachtungsdaten zu rekonstruieren, mit einem Schwerpunkt auf der Rekonstruktion von dynamischen Systemen, die in Raum und Zeit definiert sind. Es werden die Annahmen, die notwendig sind um solche Inferenz-Probleme in der Praxis erfolgreich zu lösen, diskutiert, und die resultierenden Methoden auf reale Daten angewendet. Diese Annahmen reichen von vereinfachenden Prior-Annahmen, die in den Inferenzprozess eingehen, bis hin zur Entwicklung eines neuartigen Approximationsverfahrens für resultierende Posterior-Verteilungen.

Die in dieser Arbeit entwickelten Prior-Modelle folgen einem Prinzip der maximalen Entropie, indem sie nur die physikalischen Eigenschaften eines Systems einschränken, die für die Inferenz am relevantesten erscheinen, während sie bezüglich aller anderen Eigenschaften agnostisch bleiben. Zu diesem Zweck werden Prior-Modelle entwickelt, die nur die statistisch homogene Raum-Zeit-Korrelationsstruktur einer physikalischen Observablen einschränken. Die gewählten Bedingungen an diese Korrelationen basieren auf generischen physikalischen Prinzipien, was die resultierenden Modelle sehr flexibel macht und ein breites Anwendungsspektrum ermöglicht. Dies wird anhand mehrerer numerischer Beispiele sowie einer Anwendung auf Daten des Event Horizon Telescope über das Zentrum der Galaxie M87 verifiziert und erforscht. Darüber hinaus wird als erweiterte Anwendungsform eine Variante dieser Modelle zur Simulation partieller Differentialgleichungen verwendet. Hier wird der Prior als Vorwissen benutzt, um die physikalische Plausibilität einer zugehörigen numerischen Lösung zu quantifizieren, was wiederum die Genauigkeit der Simulation verbessert. Die Anwendbarkeit und Implikationen dieses probabilistischen Simulationsansatzes werden diskutiert und anhand von numerischen Beispielen untersucht.

Die Verwendung solcher Prior-Modelle, gepaart mit der riesigen Menge an Beobachtungsdaten moderner Teleskope, führt typischerweise zu Inferenzproblemen die zu komplex sind um vollständig analytisch lösbar zu sein. Insbesondere ist für die meisten resultierenden Posterior-Wahrscheinlichkeitsverteilungen eine numerische Näherung durch eine vereinfachte Verteilung notwendig. Um bestehende Methoden zu verbessern, schlägt diese Arbeit eine neuartige Näherungsmethode für Wahrscheinlichkeitsverteilungen vor: Geometric Variational Inference (geoVI). Die Approximationsfähigkeiten von geoVI werden theoretisch ermittelt und anhand numerischer Beispiele demonstriert. Diese Ergebnisse legen einen breiten Anwendungsbereich nahe, da das Verfahren bei moderaten Rechenkosten eine Verringerung des Näherungsfehlers im Vergleich zum Stand der Technik liefert.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The advancements in modern observational technologies and experimental designs have brought a sheer flood of high quality and extremely informative data to many areas of astronomy, physics, and the natural sciences in general. In particular in astrophysical imaging, modern telescopes provide an incredibly detailed view on many different aspects of the sky, ranging from large and extensive surveys covering the entire sky such as the Sloan Digital Sky survey [15] or the Gaia mission [24], over extremely narrow and high resolution images via radio interferometers such as the Event Horizon Telescope (EHT) [36], up to the novel detection mechanisms for gravitational waves provided by the LIGO [2] and VIRGO [3] experiments. Recovering the physically relevant information contained in the measurements conducted by these telescopes is a challenging task. Traditional approaches often stem from the era of observational astronomy where the concept of a camera and an associated image are taken literal, and a simple back-projection of the measured data onto the sky provided a satisfactory result that could be used for comparisons against predictions made by theoretical models. While modern telescopes such as the EHT still have the virtual concept of a camera, it physically consists of a collection of radio telescopes that are located all over the world and the "camera" only emerges once the data-sets of all telescopes get shipped to a central location where they get combined in a computer. Furthermore, such telescopes often also observe the sky at multiple frequencies simultaneously which virtually results in a three-dimensional image, where the third axis represents the observed range of frequencies. Finally, as the 2017 observations of the center of the neighboring galaxy M87 carried out by EHT [39] have impressively validated, the extreme conditions around the black hole located in the center of M87 result in a temporal variability on scales below the time scales of the observational cycle. Therefore treating the entire cycle as exposure time, i.E. by co-adding photon flux gathered throughout the cycle to increase the quality of the image is not justified any more and the observations actually give rise to a time variable video of the galactic center which constitutes an additional dimension to the observed "image", in addition to the one introduced by the extended frequency range. These settings, commonly referred to as multi-frequency imaging and imaging of time variable sources also appear in the context of other observational missions and in combination with the increasingly complex measurement setup due to the combination

of information gathered from multiple telescopes has lead to a demand for concepts and ultimately computational algorithms that can solve imaging problems in such complex scenarios. To properly extend and adapt the concept of a camera and an image to these problems required a more formal approach towards astrophysical imaging, in particular it requires the ability to perform inference in the context of astronomy in a formal way.

Inference, in particular the ability to reason about aspects of our World in light of novel observations, has always been an integral part of science. It enters the scientific process of making deductions about the World in a variety of ways, ranging from verifying/falsifying models and hypotheses, over the process of constraining open aspects of a model given observations, up to the advanced task of inventing/extending a model in light of novel, unexplained observations. Despite its undeniable importance to science, or maybe even because of that, there has been a recurring debate about how precisely inference should best be carried out in practice. In particular in the regime of reasoning under uncertainty, i.E. in cases where aspects of a hypothetical model (often loosely referred to as model parameters) cannot be determined from the data to a degree of absolute certainty, it has been debated how inference, or more traditionally phrased, parameter estimation, is to be performed. One of the main reasons for this debate is the fact that the process of parameter estimation is ultimately *subjective*, as it depends on various assumptions regarding the underlying model, the data, and what is assumed to be known about the parameters prior to conducting the measurement. This subjective nature has often led to a variety of seemingly disagreeing recipes to estimate parameters given a data-set, which resulted in debates about whether one recipe is superior to the other. Fortunately, the development of modern probability theory, and in particular the works of Jeffreys [65] and Cox [26], opened up a path towards a consistent and less ambiguous treatment of inference. The key idea is to constrain the state of knowledge about the parameters using the more abstract concept of *information*, rather then data alone, and to describe this state using probabilities. In particular the inference process includes all background information I that summarizes all assumptions that are made, in addition to the data d that has been observed. This inclusion is done in a formal way, i.E. by defining a probability distribution over the parameters that are considered to be plausible given the data as well as the background information. It may seem like a purely academic distinction at first, whether this additional information is included as background information in a probability or via specification of a different recipe. In fact, in many practical cases, solely utilizing probability theory to obtain the most probable parameter in light of d and I, often agrees with the output of a recipe that has the associated background information implicitly built in. If however, the inclusion of I into probability theory is done properly, it opens up the possibility to answer novel questions that were previously inaccessible. For example it is possible to use the rules of probability theory directly to compute how likely I (or a part of it) is given the data, and thus instead of debating about the superiority of one recipe, it may simply be tested which background assumptions are more likely compared to others, in light of the observations.[1]

[1]Note that in general, not all aspects of I may be tested for in this way as there exist assumptions for which no experiment exists (yet) that may yield a data-set to provide a conclusive answer.

In addition, I appears on the same level as d in the sense that both represent information on which the parameters get constrained on and therefore formally, changing (parts of) I to include novel/different ideas is done in the same way as exchanging the data once a new data-set becomes available. More abstractly speaking, the initially subjective task of determining the parameters of a model given d becomes *objective*, as using the rules of probability theory given all information available (i.E. d and I) always leads to the same result regarding the state of knowledge about the parameters, irrespective of how precisely this information is encoded. This consistent treatment of information is so fundamental to probability theory that it became one of the desiderata that were initially formulated by Cox.

The formal inclusion of I into probabilities turned out to be a mathematically non-trivial step because traditional probability theory as formulated via the Kolmogorov axioms [74] is based on frequencies, i.E. a probability is interpreted as the frequency of a recurring event. Typical background information, however, also includes general statements about the system which cannot be described as a recurring event and therefore the rules of probability theory had to be extended. This extension ultimately leads to a re-interpretation of probabilities as *degrees of plausibility* which describe how plausible a specific statement appears to be, given the information at hand. It is noteworthy that the approach is really an extension, not a modification, as the rules of probability theory and their consequences remain the same, they simply become applicable to generic statements, rather then solely recurring events. This is also reflected in later results [64] which showed that the probability axioms of Kolmogorov can be derived from Cox axioms and therefore give rise to the same logical system, when restricted to frequencies. Nevertheless, this fundamental philosophical shift regarding the interpretation of probabilities has led to an extensive debate on the validity of this novel approach. In particular its implications to inference, know known as *Bayesian inference*, were heavily debated for a long time. Fortunately today, the results of Cox and in particular Bayesian inference have become largely accepted by most people in the scientific community, and debates have returned to the much more constructive discussions regarding the validity of specific model assumptions.

1.1 Probability theory

The rules of probability theory can be regarded as a formal language to perform reasoning under uncertainty. They are an extension to Boolean's algebraic system used to describe formal logic in which a statement A may either be true or false. Instead, probability theory assigns a degree of plausibility to A, which represents the state of belief that A is a true. These plausibilities incorporate three desiderata, initially formulated by Cox (sometimes also referred to as Cox' axioms):

(I) Degrees of plausibility are represented by real numbers;

(II) They are qualitatively correspondent with common sense;

(III) Reasoning should always be consistent.

These desiderata require some further specifications to derive a mathematically consistent theory of probabilities, and a detailed derivation is beyond the scope of this introduction. The interested reader may refer to the work of Jaynes [64], where a detailed introduction of the desiderata, some small extensions, and an extensive discussion of the resulting theory of probability is given.

The nomenclature regarding probabilities adapted in this work is such that the probability of the statement A being true is denoted as $P(A)$. It is defined such that certainty about A being true is denoted as $P(A) = 1$, whereas certainty that A is false is given via $P(A) = 0$. All intermediate states of belief about A are assigned a numerical value between 0 and 1, where an increase in belief is associated with an increasing value. There is no constraint on the exact numerical values of $P(A)$ aside from normalization, that is

$$P(A) + P(\bar{A}) \stackrel{!}{=} 1 \ , \tag{1.1}$$

where $\bar{A}$ stands for not A being true (i.E. A is false).

If instead of a single Boolean statement, a categorical random variable x that takes N distinct values (specifically $x \in \{x_i\}_{i\in\{1,...,N\}}$) is considered, it is possible to assign a truth statement to every possible outcome

$$A_i \equiv \text{“}x = x_i\text{”} \quad \forall i \in \{1, ..., N\} \ . \tag{1.2}$$

To keep notation short, however, the associated probabilities are typically abbreviated by assigning a probability to each category, i.E.

$$P(x_i) \equiv P(A_i) \ , \tag{1.3}$$

which may be understood as the probability of a random variable taking the value x_i. In this case, normalization is required via a summation over all possible outcomes

$$\sum_{i=1}^{N} P(x_i) \stackrel{!}{=} 1 \ . \tag{1.4}$$

Qualitatively, it seems straightforward to imagine how to extend this assignment to continuous random variables such as scalars $x \in \mathbb{R}$ (or vectors in N dimensional spaces $\vec{x} \in \mathbb{R}^N$). By assigning a truth statement to every location in $\mathbb{R}$, it is possible to set up an infinite number of probabilities, one for every location in space. As it turns out, however, as soon as an arbitrarily small (but nonzero) interval of $\mathbb{R}$ is plausible for x, all probabilities assigned to individual locations become infinitely small due to the fact that there are infinitely many plausible values within the interval and their probabilities have to add up to one. A more meaningful description may thus be obtained via “dividing” the probabilities by the associated volume element $\mathrm{d}x$, and to consider this density instead. If done in a formal way, this gives rise to what is called a *probability density* $\mathcal{P}(x)$, a function that assigns positive real values to every x such that its integral is normalized

$$\int_{-\infty}^{\infty} \mathcal{P}(x) \ \mathrm{d}x \stackrel{!}{=} 1 \ . \tag{1.5}$$

Probabilities can then be assigned in a meaningful way to x being within a certain interval, e.g. $A \equiv$ "$x \in [a, b]$", as

$$P(A) = \int_a^b \mathcal{P}(x)\ \mathrm{d}x = \int_{-\infty}^{\infty} \Theta(x-a)\Theta(b-x)\ \mathcal{P}(x)\ \mathrm{d}x\ , \tag{1.6}$$

where Θ denotes the step function with $\Theta(x) = 0$ if $x \leq 0$ and $\Theta(x) = 1$ if $x \geq 0$. Thus a truth statement regarding a continuous variable is best represented by a density, and a probability arises again when integrating this density over a volume.

The integral in equation (1.6) can be understood as a weighted average of the function $f(x) \equiv \Theta(x-a)\Theta(b-x)$ where the weighting is given by $\mathcal{P}$. This average can not only be performed for step functions, but also for arbitrary functions $\tilde{f}(x)$. In some sense it describes what value is to be expected for $\tilde{f}$, as values of $\tilde{f}(x)$ get multiplied by a higher weight in case the density of x is larger. This weighted average is therefore known as an *expectation value* and is denoted as

$$\left\langle \tilde{f}(x) \right\rangle_{\mathcal{P}} \equiv \int_{-\infty}^{\infty} \tilde{f}(x)\ \mathcal{P}(x)\ \mathrm{d}x\ . \tag{1.7}$$

In case of probabilities P there also exists a notion of expectation values which may be stated in the generic form

$$\left\langle \tilde{f}(x) \right\rangle_{P} \equiv \sum_{x \in \Omega} \tilde{f}(x)\ P(x)\ , \tag{1.8}$$

where Ω denotes the set of all possible values for x. It can be proven that the rules of probability theory derived for probabilities P, equally apply to probability densities $\mathcal{P}$ and therefore, in this work, the symbols P for probabilities and $\mathcal{P}$ for probability densities are used interchangeably, where $P(x)$ denotes a probability in case the domain on which x is defined is discrete, and a probability density in case this domain is continuous.

To alter the state of knowledge described by a probability (density), the so-called conditional probabilities are introduced as $P(A|B)$, which describe the probability of A given that the statement B is true. In addition, it is possible to assign a joint probability to two (or multiple) statements A and B being true simultaneously given some background information I, which is denoted as $P(A, B|I)$. Finally, it is also possible to assign a probability $P(A + B|I)$ to the statement that A or B are true, given I. Here $+$ denotes the logical "or". Joint probability distributions and conditional distributions can be related via the product rule of probability, given as

$$P(A, B|I) = P(A|B, I)\ P(B|I) = P(B|A, I)\ P(A|I)\ . \tag{1.9}$$

Furthermore $P(A + B|I)$ can be decomposed using the sum rule

$$P(A + B|I) = P(A|I) + P(B|I) - P(A, B|I)\ . \tag{1.10}$$

The sum and the product rule are one of the most basic, but also most important results of probability theory as they, together with normalization, are sufficient to derive every statement within probability theory.

One fundamental derived rule is *marginalization*, given in formula via

$$P(B|I) = \sum_{A \in \Lambda} P(A, B|I) \ , \tag{1.11}$$

where Λ is a placeholder for the domain of A, e.g. $\Lambda = \left\{A, \bar{A}\right\}$ in case of a simple truth statement, or $\Lambda = \mathbb{R}$ for a continuous random variable (in which case the sum is replaced with an integral). Qualitatively it states that the probability of B can be obtained from the joint distribution of A and B by adding up all possible outcomes for A. Furthermore, the product rule opens up a formal way towards deductive reasoning, via what is known as *Bayes theorem*, given as

$$P(A|B, I) = \frac{P(B|A, I) \ P(A|I)}{P(B|I)} = \frac{P(B|A, I) \ P(A|I)}{\sum_{A \in \Lambda} P(B|A, I) \ P(A|I)} \ . \tag{1.12}$$

It allows to update the state of knowledge regarding A in light of B in a formal way by solely using the probability of A without knowing B, and the influence of A on B as encoded in $P(B|A, I)$. While Bayes theorem is a simple and direct consequence of the product rule, its implications regarding inference, in particular combined with the ability to reason about arbitrary truth statements, are exceptional.

1.2 Bayesian inference for physical systems

In the context of inferring properties of physical systems, Bayes theorem becomes very important as it enables reasoning in scenarios that are basically inaccessible without it. To give an example, consider an experimental setup measuring data d that is assumed to depend in a structured way on some physical parameters θ. This dependency describes the essential aspects of the measurement device but also covers possible nuisance effects. Therefore, in general, it is of probabilistic nature and may be denoted as the likelihood $P(d|\theta, I)$ of a possible outcome d of the experiment, given a specific parameter configuration θ. I stands for all the physical knowledge required to describes the measurement process that leads from θ to d, and is referred to as background information. To reason about θ, however, the quantity of interest is the probability $P(\theta|d, I)$, the so-called *posterior*. While a descriptive physical model relating θ to d determines the functional form of the likelihood, it is usually not possible to directly construct the posterior, as there typically exists no physical model that directly describes θ as a function of d. Therefore, in general only the combination of physics with probability theory, specifically by employing Bayes theorem, enables the computation of the posterior

$$P(\theta|d, I) = \frac{P(d|\theta, I) \ P(\theta|I)}{P(d|I)} \ . \tag{1.13}$$

In this sense, despite being an immediate consequence of the product rule for probabilities, the theorem has had a great impact on the natural sciences as it opens up the possibility to

perform reasoning in the context of deducting parameters of physical models from observations. One interesting consequence of the theorem is that in addition to the likelihood, a second probability $P(\theta|I)$ is required to fully specify the posterior. It states the plausibility of θ disregarding the information gained via d and thus, in some sense, describes the state of knowledge prior to the measurement.

1.2.1 Prior probability distribution

The so-called prior distribution $P(\theta|I)$ can be used to encode direct additional physical knowledge about the properties of the system described by θ and as such constitutes additional background information to I. It is often described as "everything that is known to be true about the system prior to the measurement". While this may be a valid introduction to give an idea about the role of a prior, it is a somewhat incomplete picture as there might very well also be prior knowledge encoded in the likelihood, such as the physical relation between θ and the observables and therefore whether or not "prior" information enters the prior or the likelihood distribution remains ambiguous. In addition, it is also possible to deliberately include assumptions that may or may not be true and use the inference results to compute how likely these assumptions are compared to alternative hypotheses. Finally, there usually exist many aspects of the system that are true but simply irrelevant for the inference problem regarding θ and thus can be disregarded. Therefore, to keep a generic notation, all knowledge regarding the system that enters the inference in addition to the data d, is simply summarized as background information I and both, the prior as well as the likelihood distribution are conditioned on (parts of) I.

While this general treatment of background information is the most flexible one, it also remains the least specific one, leaving much open room for debates regarding different model choices. To reduce the amount of subjective knowledge that has to enter the inference process, multiple attempts have been made to remove (parts of) the prior choices a hypothetical user has to make to arrive at a well defined inference problem. One approach is the idea to set up an "uninformative" prior distribution $P(\theta|I)$, a distribution that avoids specifying any subjective preference for a set of parameter configurations prior to the measurement. Achieving this desired goal of being uninformative, however, turns out to be a nontrivial task. The most straightforward attempt is what is known as a flat prior, a distribution that assigns equal probability to all possible values of θ. Specifically,

$$P(\theta|I) \propto \text{cst.} \quad \forall\theta \in \Theta \,, \tag{1.14}$$

where Θ denotes the space in which θ is defined. In many ways this flat prior can be regarded as a maximally uninformative choice, in particular in case the values of θ have a direct physical interpretation and describe distinct states of a system. In case θ merely denotes a parametrization of some physical quantity, however, this prior choice is only apparently uninformative due to the fact that while probability theory is invariant under reparametrizations, a flat prior is not. To see this, consider a reparametrization $\theta = f(\theta')$ with $f : \Theta' \to \Theta$ being an invertible transformation. Reparametrization invariance, roughly

speaking, requires that the integration measure associated with the probability density remains invariant

$$P(\theta|I)\ \mathrm{d}\theta \stackrel{!}{=} P(\theta'|I)\ \mathrm{d}\theta' \ . \tag{1.15}$$

Therefore the initially flat distribution using the coordinates θ may become arbitrarily structured in coordinates θ' as

$$P(\theta'|I) = P(\theta|I) \left|\left| \frac{\mathrm{d}\theta}{\mathrm{d}\theta'} \right|\right| \propto \left|\left| \frac{\partial f}{\partial \theta'} \right|\right| \ , \tag{1.16}$$

where $||\bullet||$ denotes the absolute value of the determinant. Interestingly, the inverse statement is also true: given some arbitrarily structured prior distribution, it is possible to construct a coordinate system in which this prior becomes flat, a process known as inverse transform sampling [27]. This implies that solely the choice of a coordinate system is prior information, as it is sufficient to encode any desired prior knowledge into the inference problem. Therefore, in case the coordinates θ are not special in a way that justifies an equal treatment of all possible values of θ, different prior choices are more appropriate to be uninformative.

In particular, attempts to construct prior distributions that are flat in coordinate systems that are in some sense "natural" candidates (in the absence of additional physical prior knowledge) have been made in the past. One of the most prominent example is given by Jeffreys' prior, a distribution that is proportional to the square root of the Fisher information metric (FIM) [42]. The FIM plays a central role not only in prior construction, but also in the field of *information geometry* and will be introduced in detail in chapter 5. For the moment, it is sufficient to understand that the FIM $\mathcal{M}(\theta)$ is a metric tensor that introduces a distance measure between two infinitesimally close likelihood distributions $P(d|\theta, I)$ and $P(d|\theta + \mathrm{d}\theta, I)$. Qualitatively this distance measure describes the difference in information content between the two likelihood distributions parameterized by θ and $\theta + \mathrm{d}\theta$. The FIM gives rise to what is known as a statistical manifold, a Riemannian manifold where each point on the manifold can be associated with a probability distribution. The functional form of the FIM is uniquely determined given a likelihood distribution and Jeffreys' prior may be written using $\mathcal{M}$ as

$$P(\theta|I) \propto \sqrt{|\mathcal{M}(\theta)|} \ . \tag{1.17}$$

This prior assigns equal probability to each location on the manifold induced via $\mathcal{M}$, while respecting the notion of distance as given by the metric. Therefore, it is a flat prior in the coordinate system that appears natural from the perspective of the manifold, which then gets transformed into the coordinates θ. Note, however, that this prior choice depends on the likelihood, which enters $\mathcal{M}(\theta)$. This implies that the form of Jeffreys' prior is not epistemologically precise, as how a quantity is measured enters its prior measurement assumptions. Nevertheless, in absence of any physically motivated choice of an "equal probability" coordinate system, the FIM and Jeffreys' prior can provide a valuable alternative.

While either encoding everything that is know about a system, by employing a full physical theory, or nothing, via uninformative prior choices, into a prior probability distribution may both be justified and valuable in a given context, in practice a somewhat intermediate route is often taken. In particular it might be useful to constrain some, but not all, aspects of the system using physical knowledge, and remain fully agnostic with regard to all other unconstrained aspects. A formal way to do so in practice has been developed by what is known as the construction of *maximum entropy* distributions.

1.2.2 Maximum Entropy for prior construction

A convenient way to introduce such partial background information into a probability distribution is given by constraining an expectation value (or multiple values) of a distribution to a specific numerical value. As an illustrative example, consider a probability distribution $P(x)$ with $x \in \mathcal{X}$, and furthermore assume that the following statement is valid for the system

$$\left\langle x^2 \right\rangle_{P(x)} \stackrel{!}{=} \sigma^2 \ , \tag{1.18}$$

i.E. the second moment of $P(x)$, the so-called variance, is known to take the value σ^2. It is clear that there are many possible distributions $P(x)$ that satisfy equation (1.18) and solely using this variance constraint does not yield a unique choice of P. To select a prior, the additional requirement of being maximally uninformative comes into play again as the least informative distribution that satisfies equation (1.18) appears to be an appropriate prior candidate. This requirement can be formalized using the concept of *entropy*, which for continuous probability distributions is defined as

$$\mathcal{S}[P] = -\int_{\mathcal{X}} P(x) \ \log\left(\frac{P(x)}{Q(x)}\right) \ \mathrm{d}x \ , \tag{1.19}$$

where Q is a so-called reference distribution. It represents a maximally uninformative distribution for x, which, for all practical considerations here, may simply be set to the flat distribution. The entropy is a measure for the difference in information content between P and Q, and maximizing the entropy is associated with a minimal difference of P compared to the maximally uninformative reference measure Q. Given this formal definition of the entropy and the constraint in equation (1.18), obtaining P is described via a constrained optimization problem where in addition to (1.18) the function P also has to satisfy the normalization constraint (1.5) to describe a probability density. It is a well defined optimization problem, and can be solved using e.g. Lagrange multipliers. In the particular example chosen above and for x being a real valued random variable (i.E. $\mathcal{X} \equiv \mathbb{R}$) this yields

$$P(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \ e^{-\frac{1}{2}\left(\frac{x}{\sigma}\right)^2} \quad \forall x \in \mathbb{R} \ . \tag{1.20}$$

This distribution is what is called a maximum entropy distribution, and equals a specific instance of the famous Gaussian (or normal) distribution which plays a huge role in almost

all areas of statistics. Analogous to the discussion above, it can be shown that it is also the maximum entropy distribution in case the first moment is known to take the value m (i.E. $\langle x \rangle_P = m$) and σ^2 measures the variance of x around m. This yields the usual definition of the normal distribution $\mathcal{N}$, given as

$$\mathcal{N}(x; m, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \, e^{-\frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2} \, . \tag{1.21}$$

It is an extremely powerful distribution to summarize the knowledge about a parameter x, as it encodes an expected value m and a degree of (un-)certainty encoded via the variance σ^2. In addition, its maximum entropy nature ensures that no additional hidden information enters the state of knowledge described by P, it remains maximally uninformative regarding everything not determined via m and σ. A normal distribution can also be derived for a multi-dimensional quantity such as an N-dimensional vector $\vec{x} \in \mathbb{R}^N$. Given a set of constraints for each entry x_i of $\vec{x}$ of the form

$$m_i = \langle x_i \rangle_{P(\vec{x})} \quad \text{and} \quad D_{ij} = \langle (x-m)_i (x-m)_j \rangle_{P(\vec{x})} \, , \tag{1.22}$$

the maximum entropy distribution is the *multivariate* normal distribution which takes the form

$$\mathcal{N}(\vec{x}; \vec{m}, \mathbf{D}) = \frac{1}{\sqrt{|2\pi\mathbf{D}|}} \, e^{-\frac{1}{2}(\vec{x}-\vec{m})^\dagger \mathbf{D}^{-1} (\vec{x}-\vec{m})} \, , \tag{1.23}$$

where $\mathbf{D}$ is a symmetric and positive definite matrix that has entries D_{ij}, $|\bullet|$ denotes the determinant of a matrix, and $\dagger$ denotes the adjoint of a vector (i.E. its transposed in case $\vec{x}$ is real, or the transposed and conjugated values in case of complex variables). In analogy to the scalar case x, this maximum entropy distribution also encodes an expected value as well as a notion of uncertainty for $\vec{x}$. In addition, however, the multivariate case also has a notion of correlation between different entries x_i and x_j with $i \neq j$ built in, via the corresponding entry in the covariance D_{ij}. This makes it an extremely powerful and popular distribution for prior construction as it summarizes some fundamental aspects of any system described by multiple continuous variables, while being maximally uninformative regarding all other aspects.

While summarizing a state of knowledge in terms of expectation values is a very simple and powerful method to approximate information, in some cases there exists a more direct approach to summarize knowledge. In terms of probability theory, knowledge is always represented in terms of a probability, and instead of summarizing this information in terms of expectation values which are ultimately cast back again into a maximum entropy distribution, it is possible to directly approximate one distribution in terms of another one which summarizes the information contained in the former distribution as best as possible. While closely related to entropy principles, this approximation problem poses an information theoretical question in itself.

1.3 Approximation of probabilities

From an information theoretical point of view, one immediate question that arises is why such an approximation might even be desirable. In the context of maximum entropy, it is clear that introducing some information into a distribution via constraints in order to deviate from a maximally uninformative state can be useful for prior construction, but doing the opposite and removing information seems to be an undesirable goal. As an example, assume that there exists an informative distribution $P(x|d_1, I)$ which was obtained by performing inference given data d_1 and furthermore assume that the background information I is known to be valid. In case a new measurement d_2 becomes available, it is clear that the best prior to choose for the second inference problem is the posterior $P(x|d_1, I)$ of the first one. Deviating from this prior choice by approximating $P(x|d_1, I)$ in terms of another distribution seems like a bad idea, since using an approximation inevitably removes (possibly) valuable information. At best, the approximation is perfect where in this case the inference results will simply be equally good compared to using $P(x|d_1, I)$ directly.

Even though there exists no direct theoretical reason to do approximation, it becomes a highly relevant topic in most real world applications, in particular as soon as information theory meets finite resources. Initially, the theory of approximation has been developed around the idea of communicating knowledge through a channel with limited bandwidth. This is certainly a very relevant issue, as any real world communication channel is necessarily limited. In the context of inference, however, the necessity of approximation already arises once inference problems have to be solved using limited computational resources. In particular the computation of expectation values, one of the most fundamental ways of extracting information from a probability distribution, requires integration over the distribution which, for most practically relevant cases, cannot be computed analytically but requires an approximation. A direct numerical approximation of such integrals can be computationally very challenging, and therefore an approximation of P with another distribution that can easily be integrated is desired. Therefore, accepting that approximations have to be made in practice, a desirable goal is to at least keep these approximations as close as possible to the true distributions. In direct analogy to the entropy principles, the difference in information between two distributions becomes the relevant quantity of interest, and minimizing the loss of information when moving from the true distribution to an approximation appears to be a desirable criterion. Indeed, this is also the formal solution to the approximation problem, where the optimal approximation $Q(x)$ is chosen such that the loss of information w.r.t. the true distribution $P(x)$ gets minimized. This information loss is described by the Kullback-Leibler divergence (KL) [76][2], defined as

$$\mathrm{KL}\left[P;Q\right] \equiv \int P(x) \log\left(\frac{P(x)}{Q(x)}\right) \,\mathrm{d}x \ . \tag{1.24}$$

It is noteworthy that the formula for the KL only differs from the definition of the entropy

[2]The mathematical form of the KL was already introduced by Gibbs in 1902 [52], who therefore should deserve credit for it. Interestingly, Kullback and Leibler actually did not propose the formula referred to as KL today, but a symmetrized version of it, whereas Gibbs got it right in the first place.

(1.19) by a minus sign, where for the entropy, however, Q stood for the uninformative reference measure, while here it stands for the approximation Q of P. Without further constraints, minimizing the KL w.r.t. Q simply yields $Q = P$ as if possible, using the true distribution as an approximation is clearly the optimal choice. Information loss only emerges once additional constraints come into play. Typically the approximation is chosen out of a parametric family of distribution $Q = Q_\phi$, where ϕ represents a set of parameters, and this family ideally solely consists of distributions that can easily be integrated. Finding the optimal approximation $Q^* = Q_{\phi^*}$ translates into an optimal choice ϕ^* of the parameters, which is obtained by minimizing the KL

$$\phi^* = \underset{\phi}{\operatorname{argmin}} \left(\mathrm{KL}\left[P; Q_\phi\right] \right) \ . \tag{1.25}$$

Note that in case of the normal distribution discussed above, where $\phi = (m, \sigma)$, the optimal approximation obtained from minimizing the KL coincides with the maximum entropy result. In particular computing the first and second moments from P and then determining the maximum entropy distribution given these moments agrees with the distribution Q_{ϕ^*}. The applicability of the KL, however, extends to arbitrary parameter families including those that cannot easily be related to moments.

While it appears that given these results the problem of approximating probabilities is fully solved, it turns out that there remains a very practical challenge that unfortunately renders this optimal form of approximation infeasible in many cases. As initially motivated, one of the reasons for approximation is that computing integrals involving P may computationally be very challenging. To replace P by an optimal Q^* in those integrals, however, requires minimization of the KL which itself contains an integral involving P. Thus in case it is computationally infeasible to compute those integrals, it also becomes infeasible to do optimal approximation in practice. Therefore, and alternative approximation method known as variational approximation is often used. This method also relies on minimizing the KL, but the input arguments P and Q get swapped, i.E. the true distribution becomes the second input and the approximation Q the first one:

$$\mathrm{KL}\left[Q; P\right] \equiv \int Q(x) \log\left(\frac{Q(x)}{P(x)}\right) \, \mathrm{d}x \ . \tag{1.26}$$

This swapped definition is often also referred to as variational KL, or reverse KL. While it is clear that in general the results obtained using the variational approach differ from the optimal approximation due to the fact that the KL is non-symmetric in its arguments, it at least becomes a computationally less demanding quantity, as integrals involving Q should, by choice of Q, be simpler to compute. Fortunately, in addition, in case Q is close to P the results obtained by minimization become similar. In particular assuming $Q = P + \mathrm{d}P$ yields

$$\mathrm{KL}\left[P; P + \mathrm{d}P\right] = \mathrm{KL}\left[P + \mathrm{d}P; P\right] + \mathcal{O}\left(\mathrm{d}P^3\right) \ . \tag{1.27}$$

Therefore, in practice, given a parametric family Q_ϕ that contains a distribution that is close enough to P, it is possible to obtain a near optimal variational approximation $Q_{\tilde{\phi}}$

which is close to the optimal choice Q_{ϕ^*} via minimization of the variational KL. Care must be taken, however, once a close match of P cannot be found within Q_ϕ. While optimal approximation guarantees that the statistical properties that can be captured by Q_ϕ (may it be some specific moments or other properties) are accurately approximated, i.E. precisely match the respective properties given by P, a variational approximation does not give those guarantees. In particular assuming again that ϕ represents the mean and variance of a normal distribution and furthermore assume that P is not close to a normal distribution in any way, the resulting variational approximation $\tilde{\phi}$ is not guaranteed to represent the mean and variance values of P any more, whereas an optimal approximation ϕ^* guarantees to be equal to those moments. Therefore, pairing variational approximation with a highly expressive family Q_ϕ to increase the chances that P can be well approximated in this family has proven to be crucial to the success of the variational approach. In practice, it has been demonstrated that this pairing can result in a very powerful approximation tool, even for highly structured distributions P.

1.4 Imaging via the inference of fields

Having established some aspects of probability theory and Bayesian inference, a few recipes for prior construction, and a way of approximating the information contained in probability distributions, it is finally possible to return to the task of imaging and set it up in a formal way. While the concepts introduced above are valid and used in a much more generic context then imaging, in particular in astrophysical imaging problems they are all relevant and constitute integral parts in order to solve those problems. First of all, astrophysical observations are inherently subject to multiple nuisance effects that, unlike in a laboratory setting, often cannot be removed as they are out of reach. These effects partially obscure the information content of an observation or even make it inaccessible. As an example, consider the task of imaging a distant galaxy, by measuring the light arriving at a telescope on Earth coming from the galaxy. During its journey, this light is subject to many distortions, may it be the effects of the atmosphere of the Earth, or e.g. gas in our Galaxy that partially absorbs the light. While it is nowadays possible at great expenses to remove atmospheric effects by placing telescopes on satellites that orbit the Earth, for all what is known today, it remains impossible to place a telescope outside of our Galaxy. Therefore, while some nuisance effects may be dealt with explicitly nowadays, there always remain some effects which introduce uncertainty that cannot be removed and therefore have to be dealt with in a probabilistic fashion. In addition, even though modern telescopes are capable of obtaining an incredible sensitivity and resolution, when facing such complex objects as e.g. a galaxy, it is obvious that the information about the image gathered by the telescope inevitably has to be limited, and therefore additional prior knowledge regarding the physical plausibility of a resulting image needs to be incorporated. At the same time, however, directly employing physical theories regarding astrophysical objects to construct a prior distribution of plausible images can become quite complex, as our understanding of the Universe has become quite elaborate. Instead, extracting

the prior information most relevant to the imaging task at hand from a physical theory and summarizing it in a maximum entropy distribution has proven to be a much simpler alternative which in many cases yields (almost) equally detailed posterior results. Finally, while the information gathered by telescopes may be limited viewed from the perspective of obtaining a complete picture of our Cosmos, it is in almost all cases at the very limit of computational feasibility, and often those limits have to be pushed further to cope with every new generation of telescopes. Therefore, approximations of the posterior information are usually inevitable in order to access the information content in a way that can be used to draw conclusions in practice, which makes a theory for approximation and in particular variational approximation very valuable to astrophysical imaging.

1.4.1 Images and field like objects

The traditional view of imaging in observational astronomy by taking the word image literal, i.E. via mounting a camera at the end of a telescope, was historically the way to gather information about the sky. Even in this historic setting, in principle, there already exists a distinction between an image, the quantity of interest of the observation, and the output of the camera, the data, which is the film that had to be developed in order to retrieve an image. Nowadays, telescopes are of course much more complex, and while some are still build around the concept of a telescope and a camera, some employ entirely new observation techniques. Nevertheless, the idea of data remains, i.E. any measurement device outputs data that contains information regarding some quantities of interest. As discussed in the beginning of this introduction, however, these quantities often deviate from a literal image and, in many cases, are better described as a physical observable that is as closely related as possible to (or even equal to) the fundamental quantities that describe the physical system that has been observed. As an example in [79] a tomographic reconstruction of the three dimensional density of the galactic dust extinction has been carried out using the data obtained from the Gaia satellite. This dust extinction can be regarded as a three dimensional scalar field, i.E. a quantity that takes a value at every location in space, and therefore has little to do with a literal image. In addition, even though it is certainly closely related to the distribution of Galactic dust itself, from the point of view of a physical theory it is a derived quantity, as the dust density has to be accompanied with an absorption mechanism in order to give rise to the extinction density. Nevertheless, in order to constrain open parameters or to test the validity of a model for Galactic dust, the extinction density provides a much more direct and easily accessible description of the information gathered by Gaia compared to the measurement data itself. In this sense, what used to be an image has become a physical observable that ideally carries all information contained in the measurements, while being easily accessible in order to test theories about our Universe against it. This ability to test theories has, aside from the obvious desire to display the Universe as detailed as possible, always been at the very core of observational astronomy.

1.4.2 Fields

What precisely defines an observable depends on the context of a given astrophysical inference problem, but in many cases may be described as a field like quantity that is extended in space, time, frequency, or a combination of thereof. Furthermore it is assumed to be free from all contributions that obscure the observable, may it be the Galaxy acting as a foreground for extra-galactic observations, the effects of the atmosphere, or contributions directly attributed to the telescope such as effects due to the detector geometry, incomplete coverage of the observed field of view, and the various nuisance effects that every detection mechanism experiences. While the field may be free of those effects, they are present in the observations, and therefore the goal of field inference is to recover all plausible field configurations that may have led to the observed data. To do so, the concepts of probability theory that have been established for scalar and vector valued parameters, have to be extended to fields. Performing this extension is subject of Information Field Theory (IFT) [34] which applies the rules of probability theory to the task of reasoning about field like quantities.

Mathematically, a field is defined as a function $s(\vec{x})$ which maps a position $\vec{x}$ in an N-dimensional space $\vec{x} \in \mathcal{X} \subset \mathbb{R}^N$ onto a scalar value, a vector or a tensor. For the sake of simplicity only scalar fields are considered here, specifically

$$\begin{aligned} s \; &: \mathcal{X} \to \mathbb{R} \\ &\vec{x} \to s(\vec{x}) \; . \end{aligned} \tag{1.28}$$

Furthermore s is restricted to be an element of a space of functions, often it is the space of square integrable functions $s \in L^2(\mathcal{X})$, which implies that

$$s^\dagger s \equiv \int_{\mathcal{X}} s^*(\vec{x})\, s(\vec{x}) \; \mathrm{d}^N x < \infty \; , \tag{1.29}$$

where $*$ denotes complex conjugation in case the fields map onto complex values, and $a^\dagger b$ is a short notation for the integral and represents a scalar product on L^2. Mathematically, assigning probabilities to fields in a formal way is a non-trivial task and therefore this approach is solely motivated in this introduction. The interested reader may refer to e.g. [32]. It can be shown that L^2 fulfills all axioms of a vector space and therefore may be regarded as an infinite dimensional space with a distance measure given via the scalar product defined in equation (1.29). Therefore, setting aside the issue of infinite dimensions for a moment, it may be argued that distributions such as the multivariate normal distribution given in (1.23), which are valid for finite dimensional vector spaces, may also apply to infinite dimensional ones. In fact, an instructive view on spaces such as L^2 is given as the limiting case of a finite dimensional space. In particular given a space $\mathcal{X}$ it is possible to define a pixelization of this space, e.g. by dividing the space into a regularly pixelated grid, where all edges of a pixel have length Δx. A value s_i may be assigned to all pixels that represents the average value of the field s in this pixel. Collecting all those values results in a finite dimensional vector $\vec{s}$ for which a probability distribution $P(\vec{s})$ can be set up. Assuming that $s(\vec{x})$ does not vary arbitrarily between neighbouring locations $\vec{x}$ there

exists a pixelization scale which, if fine enough, results in a discrete representation $\vec{s}$ which becomes basically indistinguishable of the field s. Therefore by taking what is known as the *continuum* limit, i.E. increasing the number of pixels while simultaneously decreasing the size of Δx such that the area covered by the pixelization does not change, the vector $\vec{s}$ converges towards what is defined as the field s in case s does not vary too extremely. As probabilities are designed to describe states of knowledge about an object, having an object converging to another object in the limit, also implies that the probabilities have to become equal, specifically $P(\vec{s}) \to P(s)$ in the limit. Care must be taken mathematically to ensure that this limiting process converges which translates to some restrictions that s must uphold, which are discussed in further detail in chapter 2.

Assuming for the moment that those requirements are met and that the limit is indeed well defined, the exemplary case of a normal distribution can be extended to fields, where its distribution takes the form

$$P(s) = \mathcal{N}(s; m, D) \equiv \frac{1}{\sqrt{|2\pi D|}}\, e^{-\frac{1}{2}(s-m)^\dagger D^{-1}(s-m)} \,. \tag{1.30}$$

While the symbols are deliberately chosen such that the distribution resembles the definition of the multivariate case (1.23) to emphasize its relation to it as a limiting case, their precise meaning differs in the above formula. Note, however, that every mathematical object as well as every operation involved in equation (1.30) has a corresponding discrete counterpart which defines it via taking the continuum limit thereof. While the meaning of m as being the expected field configuration $m(\vec{x}) \in L^2(\mathcal{X})$ of s is somewhat intuitive, the precise definition of the covariance D requires some more explanation. It arises as the second moment of s and is defined as

$$\langle s(\vec{x}) s(\vec{y}) \rangle_{P(s)} \equiv D(\vec{x}, \vec{y}) \,, \tag{1.31}$$

and therefore is a function that maps two positions $\vec{x}, \vec{y} \in \mathcal{X}$ onto a real number,

$$D(\vec{x}, \vec{y}) \;:\; \mathcal{X} \times \mathcal{X} \to \mathbb{R} \,. \tag{1.32}$$

It describes the correlation between the values of s at two locations $\vec{x}$ and $\vec{y}$. At the same time D can also be understood as a linear operator that acts on fields, i.E. $D : L^2(\mathcal{X}) \to L^2(\mathcal{X})$ with an action defined via

$$(Ds)(\vec{x}) \equiv \int_{\mathcal{X}} D(\vec{x}, \vec{y})\; s(\vec{y})\; \mathrm{d}^N y \,. \tag{1.33}$$

This may seem surprising at first, but may best be understood using its discrete counterpart. Specifically a discrete covariance also describes the correlations between different entries of their corresponding random vector via its entries, while simultaneously represents a matrix, i.E. an operator that acts on elements of the associated discrete vector space and also outputs a vector. In this sense, a covariance function D is no different, its actions are merely carried out in the continuum.

1.4.3 Inference

Having established a notion of probabilities in the context of fields, it is finally possible to set up probabilistic inference models for fields. Returning to the problem of inferring astrophysical objects from data obtained from a telescope, one integral part that remains missing is how to properly set up the likelihood, i.E. the distribution that relates the observable fields s that are of interest, to data d. As said initially, an observable is ideally free from any obscuring contamination, to make the information as accessible as possible. This contamination, however, is present in reality and also imprints onto the data, and therefore if it is not captured by s, it has to be modeled by another quantity. This quantity ultimately defines the relation to data and is known as an *instrument response*, a map that takes a possible configuration of s as an input, and outputs the average response of the instrument to this input. Specifically, keeping the simplifying assumption of scalar fields $s \in L^2(\mathcal{X})$, and furthermore assuming that the obtained data is given as an M dimensional and real valued vector $d \in \mathbb{R}^M$, a response R may be defined via

$$\begin{aligned} R \;:\; L^2(\mathcal{X}) &\to \mathbb{R}^M \\ s &\to R\left[s\right] \;. \end{aligned} \tag{1.34}$$

In addition to R encoding the various distortions that the information contained in s experiences on its way to (and in some sense also through) the telescope, it necessarily also has to encode some sort of discretization operation. While the field s is defined to be an infinite dimensional quantity, data hast to be finite dimensional as it is simply impossible to collect an infinite amount of data. Such discretization operations are often determined by the type of the detector, may it be a pixelated camera that measures the incoming flux integrated over the pixels, or a device that probes the field of view at multiple sparse locations. The output of the response, however, is usually not yet what is precieved and stored as data, as there are typically also noise contributions, i.E. effects caused by other processes independent of the process related to s, that imprint on the data. One simple example may be additive noise n which allows to determine the relation between s and d via

$$d = R\left[s\right] + n \;. \tag{1.35}$$

A relation of this form is also referred to as data model since, together with a known distribution $P(n|I)$ of the noise, allows to set up a likelihood distribution $P(d|s,I)$ via

$$P(d|s,I) = \int \delta\left(d - \left(R\left[s\right] + n\right)\right)\; P(n|I)\; \mathrm{d}n = P(n = d - R\left[s\right]|I) \;, \tag{1.36}$$

where I denotes all background information regrading the form of $P(n)$ and R that is necessary to make this formula explicit.

While the procedure of translating a data model into a likelihood can identically also be used in a discrete setting by simply changing the space on which R acts, having an infinite dimensional quantity of interest has some special implications when using such a likelihood for inference. In particular assuming a discrete example with K statistically

independent entries in $\vec{s}$ and furthermore assuming that the measurement data probes $M < K$ of those without any measurement error, it is clear that after the measurement only $K - M$ entries remain that are not fully constrained by the data. In case of an infinite dimensional s, however, obtaining M measurements of the field, e.g. by probing it at M distinct locations, still leaves an infinite number of field configurations that fulfill those M constraints. If in addition the prior for s is chosen uninformative, i.E. assigns equal plausibility to every possible field configuration, it is evident that measuring those M locations does not constrain the values at other locations in $\mathcal{X}$ at all. In fact, only prior knowledge, e.g. regarding the correlations between field values at different locations, enables a propagation of the information gathered by knowing the field values at those M locations to other locations of the space. Therefore, while in a discrete case a prior may also be chosen uninformative, to infer fields in a way that constraints the configuration globally, an informative prior becomes a necessity.

1.4.4 Prior correlations

The construction of a prior for fields usually follows the same maximum entropy principles as discussed in 1.2.2. Which information about the field s is best included a priori, however, in general depends on the measurement setup as this setup determines which information will be constrained by the data and which additional information may be necessary to propagate it. In many cases, however, being able to constrain the covariance structure D a priori is extremely valuable, as it gives a simple and direct notion of dependency between different field values. The expectation value m may, of course, also provide valuable information. In many cases, however, such an expected configuration is not readily available a priori. Therefore, it has become a simple and powerful approach to imaging to use a Normal distribution for s where the covariance D is set using physical knowledge regarding s.

Fully determining $D(\vec{x}, \vec{y})$ from theory, however, may be a challenging task as it requires a detailed knowledge regarding the dependencies of s at all locations. Therefore sometimes simpler, less restrictive constraints are used as a prior. One prominent example is to reduce the constraints to the shifted average of s, given as

$$b(\vec{x}) = \frac{1}{V} \int_{\mathcal{X}} s(\vec{x} + \vec{y}) \; s(\vec{y}) \; \mathrm{d}^N y \; , \tag{1.37}$$

where V denotes the volume of $\mathcal{X}$. The maximum entropy constraint is then given by setting the expected value of b as

$$B(\vec{x}) = \langle b(\vec{x}) \rangle_{P(s)} \; . \tag{1.38}$$

The resulting distribution for s remains a Normal distribution where D takes the form

$$D(\vec{x}, \vec{y}) = B(\vec{x} - \vec{y}) \; , \tag{1.39}$$

i.E. the covariance solely depends on the vector distance between two locations rather then both locations itself. Due to this fact, such a prior is also known as a statistically

homogeneous prior distribution and B is often referred to as the kernel of $P(s)$. From the maximum entropy perspective, it is clear that this is a less restrictive distribution compared to a generic normal distribution, as additional constraints not covered by B have to be made in order to arrive at a general (in-homogeneous) covariance structure $D(\vec{x}, \vec{y})$. Unfortunately, whether or not a prior distribution is statistically homogeneous is often falsely attributed as an assumption regarding the prior, which might suggest that homogeneity is a restricting property that may not be true and should be tested for. From the perspective of prior construction, however, it becomes clear that this is not the case. Considering a system that is assumed to follow an arbitrary in-homogeneous, or even non-normal distribution $\tilde{P}(s)$, it is clear that the kernel B, as defined via the expectation value of b, can also be computed for $\tilde{P}(s)$. While in general this kernel is certainly no complete description of the system defined via $\tilde{P}$, it is nevertheless a quantity that also exists for it and might carry partial information of it. Therefore, using a homogeneous normal prior distribution $P(s)$ that matches the kernel obtained from $\tilde{P}(s)$ in an inference problem, simply implies that the least informative prior distribution that matches the kernel of the system is used during inference. The resulting posterior distribution cannot be "false" in the sense that it is inconsistent with the posterior distribution obtained using $\tilde{P}(s)$, the only benefit from using $\tilde{P}$ instead of P can be an increase in the certainty regarding the results, as $\tilde{P}$ has to be at least as informative as P by construction. The misconception of properties such as homogeneity being assumptions stems from the perspective of a prior being a model for the system. In case of a distribution being a model, as it is the case for $\tilde{P}$, statistical homogeneity is certainly an assumption regarding the system that may or may not be true. In fact hypothesizing and testing for such assumptions is a fundamental aspect of modeling nature with physical theories. A prior distribution, on the other hand, does not necessarily have to be a complete description of the system, in fact, to be useful in practice, it is often sufficient if it encodes some important, but partial information regarding the system under consideration.

To conclude the discussion regarding priors of fields, a second, valuable reduction in the number of constraints, namely statistical *isotropy* is considered. It often appears in combination with homogeneity and, for the sake of simplicity, is solely considered in this joint context here. Its constraints may be defined via the expectation value of the spherical average of b. Specifically

$$C(r) = \frac{1}{(2\pi)^{N-1}} \left\langle \int b(\vec{x}) \; \mathrm{d}\Omega^{N-1} \right\rangle_{P(s)} \quad , \tag{1.40}$$

where $r \equiv |\vec{x}|$ denotes the length of $\vec{x}$ and Ω denotes a parametrization of the hypersphere over which the integration is performed. Its maximum entropy distribution is a normal distribution where the covariance takes the form

$$D(\vec{x}, \vec{y}) = C\left(|\vec{x} - \vec{y}|\right) \quad , \tag{1.41}$$

and therefore C is also often referred to as a spherical symmetric kernel. This prior can be a simple and powerful distribution in case $\mathcal{X}$ represents a space where the dimensions share

the same meaning, e.g. in case $\mathcal{X}$ is a subset of the three dimensional physical space, or is the two-dimensional sphere on which the sky may be defined. In such cases, even solely the specification of the spherically averaged correlations of a process can provide sufficient additional prior information such that inference can be performed successfully.

1.5 Work presented in this thesis

To enable the usage of statistically homogeneous and/or isotropic prior distributions for the inference of a physical observable requires to correctly determine the kernels from the underlying physical system describing the observable. In many cases, however, aspects of the theory may be the subject of active research, and no definite choice of a kernel can be provided by theory. To circumvent the need of directly specifying a prior kernel for inference, the first part of this thesis focuses on the development and application of an alternative approach. In particular, by constructing prior distributions for kernels themselves, a joint inference problem is set up and solved for in order to determine the appropriate prior kernel together with the original quantity of interest. Roughly speaking, the construction of priors for kernels evolves around identifying shared properties of classes of physical systems and their impact on prior kernels. This ultimately leads to prior distributions of physically plausible kernels which, while developed around the context of astrophysical imaging, also remain valid and applicable in a broader context of inference of field like quantities. Therefore, in addition to astrophysical imaging tasks, also non-astrophysical examples are discussed, some of them even outside the scope of imaging.

Chapter 2 establishes a relation between statistically homogeneous correlation kernels for fields defined in space and time and the dynamic response of the underlying system to external influences. Utilizing fundamental properties valid for any physical response function, such as e.g. the requirement of a causal propagation of the response in space-time, allows to define a prior distribution of plausible kernels respecting those constraints. Several realizations of this prior together with different measurement configurations are set up to demonstrate and establish the range of applicability. The content of this chapter contains work that has been peer-reviewed and published in *Annalen der Physik* [46].

In chapter 3, a model for physically plausible, statistically homogeneous and isotropic correlation structures is discussed. In addition, a way of combining multiple of such kernels is described which determines the prior correlations of fields defined over a space that is given as the product of multiple sub-spaces. The results are used in order to obtain the first space, time and frequency resolved reconstruction of the galactic center of the galaxy Messier 87 using data obtained in the 2017 observational cycle of the *Event Horizon Telescope*. The chapter contains parts of work that has been carried out in a joint effort together with colleagues of mine and has been submitted for publication to *Nature Astronomy*, where it is currently under review [8].

The homogeneous and isotropic model discussed in chapter 3 is based on an idea initially formulated in [88] which has been further developed throughout the course of this work [9] with its latest version being the one discussed in chapter 3. This latest incarnation of the

model has proven to be successful for field inference in a variety of additional astrophysical inference problems in the domains of radio astronomy [7], galactic all-sky imaging [62, 63], three dimensional galactic tomography of dust [79, 78], and astroparticle physics [110], but also non-astrophysical related problems [56, 95].

In addition, in chapter 4 an advanced application of a variant of this model is discussed in the context of simulating partial differential equations (PDEs). First, the task of PDE simulation is cast into an inference problem where the constraint of satisfying the PDE at a location in space-time is regarded as an observation which, together with a prior model for the solution of the PDE, is cast into a posterior distribution regarding plausible field configurations that satisfy the PDE constraints. The content of this chapter contains work that has been submitted for publication to *Physical Review E* [45].

The remaining part of this thesis, chapter 5, focuses on the numerical realization and approximation of inference problems, where the main focus is on variational approximations of posterior distributions that are too complex and high dimensional to be handled differently in a reasonable way. The issue of selecting a family of approximating distributions that are guaranteed to contain a close match for the true posterior in order to ensure the validity of the entire variational approach is addressed. In particular a family of distributions is developed that matches the geometric properties of the posterior distribution when interpreted as a function of the quantities of interest. This geometric match ensures that by construction, the family contains a distribution that provides a close match for one mode of the posterior and therefore guarantees the validity of a variational approximation regarding this mode. One main result of this chapter is the development of an associated algorithm called *geometric Variational Inference* (geoVI), which allows for an efficient solution of the related inference problems. The content of this chapter contains work that has been peer-reviewed and published in *Entropy* [47].

The implementations of the kernel models discussed in chapter 2 and 3, as well as an implementation of geoVI have been deployed in version 7 of the software package *Numerical Information Field Theory* (NIFTy). In this way, the kernel models provide integral parts to the construction of inference models, while geoVI serves as a central inference engine to approximately solve those inference problems.

Finally, chapter 6 concludes this thesis and reflects on some of the findings of this work and their implications regarding astrophysical imaging, but also probabilistic inference in general. Furthermore, an outlook on possible future research goals is given which may help to improve and extend the existing described methods, or to remove some of its discussed limitations.

1.6 Additional Work

During the PhD: Aside from the work presented in this thesis, I contributed to various additional research projects throughout the course of my PhD. In [83], a decomposition of the Galactic multi-frequency sky was performed using a Bayesian and extended variant of a Variational Autoencoder, where I was involved in the development of the utilized method

and contributed to the theoretical model. This work utilizes an improvement version of Galactic all-sky maps provided by [84], where I was involved in the development of the Gaussian Mixture Model variant used to perform the analysis. In addition, in [110] I contributed a prototype of the inference method used to perform the reconstruction of detectable radio pulses that arise from the particle showers of high energy cosmic rays. A Bayesian unification of the imaging and calibration process in radio interferometry was achieved in [9], where I contributed to a model for statistically homogeneous and isotropic correlation structures, its implementation, and the associated text. This model can be seen as a predecessor of the model discussed in chapter 3. In [111], the connections between the inference of dynamical systems and a supersymmetric theory of stochastic dynamics are established, where I helped to illustrate how the spontaneous breakdown of supersymmetry, in which case a system evolves chaotic, affects the uncertainty of associated inference problems. To enable the study about the causal correspondence between the age and the SARS-CoV-2 viral load of a patient done in [56], I contributed to the development of a non-parametric estimator for probability densities. Finally, all numerical implementations discussed in this thesis, as well as most applications in these related works, utilize the software package NIFTy, where I have been involved in the development of versions 3 - 7 [102, 6].

Before the PhD: In [48], I developed a predecessor to the method discussed in chapter 2, which features a simplified model for the correlation structure of a dynamical system. In addition, I contributed to the development of a method to separate the diffuse emission from point sources in astrophysical images [72, 73]. Finally, in [49] I developed a method to study the relations between galaxy properties and the properties of their surrounding cosmological large-scale-structure.

Chapter 2

Field dynamics inference for local and causal interactions

The following chapter has first been published in Annalen der Physik *with me as the first author [46]. All authors read, commented, and approved the final manuscript.*

Abstract

Inference of fields defined in space and time from observational data is a core discipline in many scientific areas. This work approaches the problem in a Bayesian framework. The proposed method is based on statistically homogeneous random fields defined in space and time and demonstrates how to reconstruct the field together with its prior correlation structure from data. The prior model of the correlation structure is described in a non-parametric fashion and solely builds on fundamental physical assumptions such as space-time homogeneity, locality, and causality. These assumptions are sufficient to successfully infer the field and its prior correlation structure from noisy and incomplete data of a single realization of the process as demonstrated via multiple numerical examples.

2.1 Introduction

Modeling as well as inferring quantities defined in space and time on the basis of observational data thereof has always been at the very core of many scientific areas. In recent years, astrophysical imaging began to become sensitive to the temporal dimension, in addition to the spatial ones. This is due to the fact that although large astrophysical objects such as galaxies appear to be static on observational timescales, small objects such as stars and binary black holes exhibit transient, periodic, and quasi-periodic modulations of the emission on observable timescales.

In addition, the spatio-temporal correlation structure of non-astronomical systems plays a central role in the calibration of modern telescopes. In particular the temporal variability

of these systems is used in order to identify and distinguish them from the typically static astronomical object of interest. As a prominent example, modern radio telescopes such as LOFAR [109] or the upcoming SKA are limited in resolution by the deflection of incoming radio signals due to the ionosphere. The strength of these distortions ultimately depends on the electron density of the ionosphere. As this density is not known for all observed locations x at time t, it has to be inferred along with the incoming flux. Typically, the electron density is probed via observing a calibration target with known flux at location x' and time t'. Therefore, it is also necessary to make a statement about the correlation structure of the electron density in order to extrapolate the information gained at (t', x') to the space-time location (t, x) where the actual observation is made.

In order to tackle these as well as other inference problems of this kind in space and time, we have to perform inference of continuous quantities, or fields, from a finite set of measurement data. This problem is in general ill-posed, as we aim to constrain infinite degrees of freedoms (dofs) on finite, usually also noisy, measurements. Consequently we rely on Bayesian inference, more precisely on Information Field Theory (IFT) [32], and use this language to encode prior knowledge about the system under consideration. Typically, this prior knowledge is incomplete, and there exist a set of unknown hyper-parameters, such as the spatio-temporal prior correlation structure, which have to be inferred along with the field of interest. We outline in this paper how physically motivated concepts such as spatio-temporal homogeneity, locality as well as causality can be encoded into the prior correlation structure. Furthermore, we demonstrate that the resulting hierarchical prior model is restrictive enough to perform inference on the basis of noisy and incomplete data of a single realization of the process, while still being flexible enough to capture complicated correlation structures.

Traditionally there exist two different pictures on random fields in space-time, one results in space and time being treated separately [89, 55], while the other models the field as defined over a single space, namely space-time [28, 97] (see e.g. [5] for an extensive discussion). In this work we rely on the latter picture. Consequently the corresponding inference problems can be regarded as the task of inferring a field defined on a single space, given a finite amount of measurements in this space.

To this end, in section 2.2 we start with a brief introduction to IFT and the notation used in this work. In section 2.3 we discuss how to encode our prior knowledge about the field and its correlation structure into a joint prior distribution thereof. This prior is then used in section 2.4 in order to solve the corresponding inference problem and the performance of the resulting algorithm is demonstrated in section 2.5. Ultimately, in section 2.6, we conclude the paper with a brief summary of the proposed concepts.

2.2 Information Field Theory and Gaussian processes

Information Field Theory is a statistical field theory that aims to describe Bayesian inference of fields defined over some continuous space, or space-time. For simplicity, we first consider a one dimensional random process, and provide an extension to space-time at the

end of this section. To this end, consider a zero mean square-integrable random process s defined over a closed interval $I = [0, B]$, i.E. $s^x \in L^2(I)$ with probability $P(s)$. We define the covariance function S as

$$S(x, y) \equiv \langle s^x \, (s^y)^* \rangle_{P(s)} \ , \tag{2.1}$$

where * denotes complex conjugation. If we associate with S a linear operator $O_S : L^2(I) \to L^2(I)$ and define its application via

$$(O_S s)^x \equiv \int_I S(x, y) s^y \ \mathrm{d}y \ , \tag{2.2}$$

we may define the eigenvalues λ_k and eigenfunctions e_k of the linear operator O_S via

$$(O_S \ e_k)^x = \int_I S(x, y) \ e_k^y \ \mathrm{d}y = \lambda_k \ e_k^x \ . \tag{2.3}$$

Since the eigenfunctions of O_S form an orthonormal basis and the random process s lies within the span of e_k, the Karhunen-Loève theorem [67, 80] states that s may be represented in this basis as

$$s^x = \sum_{k=-\infty}^{\infty} e_k^x \ \tilde{s}^k \ , \tag{2.4}$$

with the modes $\tilde{s}^k$ defined via

$$\tilde{s}^k \equiv \int_I (e_k^x)^* \ s^x \ \mathrm{d}x \quad \text{with} \quad \left\langle \tilde{s}^k \right\rangle = 0 \ , \left\langle \tilde{s}^k \ \tilde{s}^q \right\rangle = \delta_{kq} \ \lambda_k \quad \forall k, q \in \mathbb{Z} \ . \tag{2.5}$$

Specifically all $\tilde{s}^k$ become zero mean and uncorrelated random variables with variance λ_k. Consequently, an inference problem of s^x can be reduced to an inference problem in $\tilde{s}^k$

2.2.1 Statistically homogeneous Gaussian processes

Statistically homogeneous Gaussian processes are a special, but very useful, process for prior modeling of physical processes as the statistical homogeneity implies that a priori no specific location in I is singled out. We may again define a zero mean random process s^x with the additional requirement that the covariance takes the form

$$S(x, y) = S(x - y) \ \forall x, y \in I \ . \tag{2.6}$$

If we additionally require the space to obey periodic boundary conditions such that

$$s^{x+B} = s^x \quad \text{and} \quad S(x + B, y) = S(x, y) \ , \tag{2.7}$$

the Wiener-Khinchin theorem [112] implies that the eigenbasis of the linear operator associated with this covariance function is the Fourier basis and its spectrum is the Fourier power spectrum. This allows for a representation of s as

$$s^x = \sum_{k=-\infty}^{\infty} e^{\frac{2\pi}{B} i k x} \ g^k \xi^k \quad \text{with} \quad \left| g^k \right|^2 = \lambda_k \ , \ \xi^k \sim \mathcal{G}(\xi^k, 1) \ \forall k \in \mathbb{Z} \ , \tag{2.8}$$

where i denotes the imaginary unit and ξ^k are independent and identically distributed Gaussian random variables with mean zero and variance one. For a compact notation we may define the Fourier transformation $\mathcal{F}_x^k = e^{\frac{2\pi}{B}ikx}$ and an infinite dimensional diagonal matrix $\widehat{G}_q^k = \delta_{kq} g^k$ to write

$$s^x = (\mathcal{F}\widehat{G}\xi)^x \equiv \sum_{k=-\infty}^{\infty} \sum_{q=-\infty}^{\infty} \mathcal{F}_k^x \widehat{G}_q^k \xi^q \ . \tag{2.9}$$

The application of its adjoint, abbreviated as $\mathcal{F}^\dagger$, is defined as

$$\tilde{s}^k = \left(\mathcal{F}^\dagger s\right)^k \equiv \int (\mathcal{F}_k^x)^* \, s^x \ \mathrm{d}x \ . \tag{2.10}$$

2.2.2 Linear Measurements and the Wiener Filter

In order to perform Bayesian inference of s given some observational data d, we require a data generating model (or data-model) that describes how d is obtained from s and possibly additional nuisance parameters n that describe the measurement noise. A simple but very powerful idealized model is a linear measurement of s with additive Gaussian distributed noise n, independent of s, defined as

$$d^i = (Rs)^i + n^i = \int_I R_x^i s^x \ \mathrm{d}x \quad \forall i \in \{1, ..., M\} \quad \text{with} \quad n \sim \mathcal{G}(n, N) \ , \tag{2.11}$$

where $R : L^2(I) \to \mathbb{R}^M$ is a linear operator that maps onto a discrete M-dimensional space, called data space, and $\mathcal{G}(n, N)$ denotes a Gaussian distribution with zero mean and covariance N. A linear measurement operator may represent common scenarios such as measurements of values at single locations, integrated measurements over a specific area, partially masked areas, convolution with a point spread function, and linear combinations thereof.

We can represent the data-model (Eq. (2.11)) as a generating process in terms of ξ by inserting the Fourier basis representation of s (Eq. (2.9)) to get

$$d = R\mathcal{F}\widehat{G} \ \xi + n \ . \tag{2.12}$$

The inference problem may thus be regarded as the task of constructing the posterior distribution of ξ, given d and the background information i.E. the specific form of R and the prior spectrum g which in turn defines $\widehat{G}$. This problem allows for a closed form solution by means of quadratic completion (see e.g. [32]) and the posterior remains a Gaussian distribution with mean m_ξ and covariance D_ξ given as

$$\begin{aligned} D_\xi &= \left(\widehat{G}^\dagger \mathcal{F}^\dagger R^\dagger N^{-1} R \mathcal{F} \widehat{G} + \mathbb{1}\right)^{-1} &&= \mathbb{1} - \widehat{G}^\dagger \mathcal{F}^\dagger R^\dagger \left(R\mathcal{F}\widehat{G}\widehat{G}^\dagger \mathcal{F}^\dagger R^\dagger + N\right)^{-1} R\mathcal{F}\widehat{G} \\ m_\xi &= D_\xi \widehat{G}^\dagger \mathcal{F}^\dagger R^\dagger N^{-1} d &&= \widehat{G}^\dagger \mathcal{F}^\dagger R^\dagger \left(R\mathcal{F}\widehat{G}\widehat{G}^\dagger \mathcal{F}^\dagger R^\dagger + N\right)^{-1} d \end{aligned} \tag{2.13}$$

where the first part of the equations is the common representation of the Wiener Filter, now for a infinite number random variables ξ which are the coefficients of the random

process s in the eigenbasis of its prior. The right hand side can be obtained by straightforward manipulation of the expressions. It has the convenient property that the only matrix inversion involved appears in the finite dimensional data-space and thus entirely avoids inversion of infinite dimensional matrices.

The posterior of the coefficients ξ can be used to construct the posterior of s by insertion of the modes into the expansion of s (Eq. (2.9)). Therefore the posterior mean m and the covariance D of s are denoted as

$$m = \mathcal{F}\widehat{G}\; m_\xi \quad \text{and} \quad D = \mathcal{F}\widehat{G} D_\xi \widehat{G}^\dagger \mathcal{F}^\dagger \; . \tag{2.14}$$

This concludes the description of the Wiener Filter theory applied to square integrable random processes in terms of the eigenbasis of the linear operator associated with the prior covariance. A mathematically more rigorous and coordinate free discussion of these concepts is beyond the scope of this work, but is described in great detail by e.g. [105].

2.2.3 Consistent discretization

For many physically relevant choices of R and $\widehat{G}$ the Fourier integrals involved in Eqs. (2.13) and (2.14) may be difficult to solve or may not have a closed form representation analytically. Therefore, for practical applications, the inference problem is often discretized and the discrete problem is solved instead. However, as shown by e.g. [82], care must be taken when defining a discretization in order to ensure that the finite dimensional approximation is consistent with the infinite dimensional inference problem. In this work, we achieve a discrete representation by truncating the Fourier series at a maximal / minimal value $\pm {}^K\!/_2$. Specifically

$$\bar{s}^x \equiv \sum_{k \in W} g^k \xi^k \; e^{\frac{2\pi}{B} ikx} \quad \text{with} \quad W \equiv \{-{}^K\!/_2, ..., {}^K\!/_2\} \; . \tag{2.15}$$

A measure for the discretization error may be defined by means of the expected squared difference between $\bar{s}$ and s as

$$(\epsilon^x)^2 \equiv \left\langle \left| s^x - \bar{s}^x \right|^2 \right\rangle = \sum_{k \in \mathbb{Z} \setminus W} \left\langle \left| g^k \xi^k \; e^{\frac{2\pi}{B} ikx} \right|^2 \right\rangle = \sum_{k \in \mathbb{Z} \setminus W} \left| g^k \right|^2 \; \forall x \in I \; , \tag{2.16}$$

which quantifies the difference between the infinite dimensional process and the finite dimensional approximation of the quantity of interest s. In order to ensure that inference is consistent, the discretization error ϵ_R of the observed quantity Rs has to be considered. It is given as

$$\left(\epsilon_R^j\right)^2 \equiv \left\langle \left| (Rs)^j - (R\bar{s})^j \right|^2 \right\rangle = \sum_{k \in \mathbb{Z} \setminus W} \left| g^k \right|^2 \left| (Re_k)^i \right|^2 \quad \text{with} \quad e_k^x = e^{\frac{2\pi}{B} ikx} \; . \tag{2.17}$$

A small ϵ_R is sufficient to ensure that the discrete approximation of the inference problem is close to the continuous one as it ensures that the contribution of modes not in W to

the observed quantity is small and therefore the information gained about these modes via the observation is also small compared to the information gain about the modes in W. For posterior analysis of s it is also relevant to have a good discrete approximation and therefore in general also ϵ should be small.

A minimal requirement is that the Gaussian process is continuous, which implies that $\left|g^k\right|^2$ decays asymptotically at least with $1/|k|^{2+\gamma}$, $\gamma \geq 0$. This ensures that the series expansion of s converges and that there exists a K such that ϵ becomes small. Fortunately, the assumption of continuity is met by most physically relevant processes.

In general, the magnitude of ϵ_R can only be specified for a given measurement scenario as it depends on the specific form of the measurement operator R. However, as the properties of R can fully be defined via its action on the Fourier basis e_k, we may qualitatively discuss three distinct cases. First, consider the case where $|Re_k|^2 \sim \mathcal{O}(1)$. Typical examples are the measurements of individual locations or sub intervals of I. In this case ϵ_R is comparable to ϵ. The second case are measurements that suppress small scales, as for example integration over an interval or convolution with a spatially extended kernel such as a point spread function. In case of integration, we get that $|Re_k|^2 \propto 1/|k|^2$ and thus ϵ_R becomes smaller then ϵ. In these two cases the discretization error of the observable is comparable or smaller then ϵ and thus a small discretization error for the field is sufficient for a consistent reconstruction. The third case are measurement operators which amplify small scale structures. One important special example are measurement operations involving spatial derivatives. The action of the derivative on e_k leads to a multiplicative factor proportional to k and therefore $|Re_k|^2 \propto |k|^2$. Care must be taken in this case since ϵ_R may become large or even infinite even though ϵ is small. This also shows that not all combinations of g and R lead to an inference problem that allows for a consistent finite dimensional representation.

Nevertheless, given a consistent combination of g and R, there always exists a cutoff K, that can be chosen prior to the reconstruction, for which the error between the discrete representation and the continuous inference problem becomes small.

Unknown prior spectrum

In this work, however, we aim to infer the prior correlation structure, specifically the form of g, in addition to ξ, from the observed data. This poses a problem, since we cannot deduce a sensible choice for K a priori. In some cases, the measurement setup allows to provide an estimate for the small scale / asymptotic behaviour of the true spectrum from the observed data, and therefore allows to set K accordingly. In many cases, however, such an estimate is not feasible without performing the full reconstruction. One approach to resolve this issue is that after a reconstruction with a chosen K, a new reconstruction using a larger cutoff $K' > K$ is performed, and both results are compared. If the two results obtained this way are similar, specifically if the difference between the large and small reconstruction for s as well as the observable Rs are small, we may conclude that the chosen discretization sufficiently resolves the true underlying process. If this is not the case, the reconstruction has to be re-run with an even larger cutoff and the procedure has to be

repeated, until the deviations become small enough. This ensures that the reconstruction is consistent with the infinite dimensional problem, assuming that the true observed process has a further decaying spectrum for modes above the cutoff.

2.2.4 Higher dimensional representation in space-time

In order perform inference in a space-time setting we define a $d+1$ dimensional space Ω, where d is the number of spatial dimensions. Specifically we define

$$\Omega \equiv [0,T] \otimes \left(\bigotimes_{i=1}^{d} [0, B_i] \right) \ , \tag{2.18}$$

and impose periodic boundary conditions along each axis of Ω. This allows for a multi-dimensional Fourier series expansion of random processes $s \in L^2(\Omega)$, that are statistically homogeneous in space and time, in direct analogy to the one dimensional setting. Furthermore we may label coordinates on Ω via $x = (t, \mathbf{x})$, and additionally $k = (\omega, \mathbf{k})$ labels the associated Fourier modes.

The periodic boundary conditions introduce possibly unwanted correlations between the boundaries, in particular along the time axis. To avoid this, we extend the time domain to $2T$, to be twice the size of the observed domain, and ensure by construction of the prior of g that the process becomes uncorrelated for moments in time with a distance greater than T. Similar to the procedure of finding an appropriate discretization, we may ensure that the posterior of g has sufficiently decreased within the interval $[0, T]$ after the reconstruction. If this requirement is not met, the time domain has to be enlarged even further and the reconstruction is performed again. For the sake of simplicity, in all examples of this work, we keep the periodic boundaries in the spatial dimensions. However, these may be omitted as well by an analogous extension of the space.

2.3 Prior

In the introduction we proposed three different concepts we aim to encode into our prior model, particularly into the prior of the correlation structure, defined via g. These are

- Statistical space-time homogeneity
- (Spatio-) temporal causality
- Locality.

The first concept is already satisfied by construction as we assume that the correlation structure is diagonal in the Fourier representation and thus fully specified via the Fourier spectrum g. We may associate a Linear operator G with g, by Fourier transforming the diagonal Matrix $\widehat{G}$ defined in Eq. (2.9) as

$$G(x, x') = G(x - x') \equiv \left(\mathcal{F} \widehat{G} \mathcal{F}^\dagger \right)(x, x') = \sum_{\omega=-\infty}^{\infty} \sum_{\mathbf{x}=-\infty}^{\infty} g^{\omega \mathbf{k}} \ e^{2\pi i \left(\omega \frac{t-t'}{2T} + \mathbf{k} \cdot \frac{\mathbf{x}-\mathbf{x}'}{\mathbf{B}} \right)} \ , \tag{2.19}$$

where $\mathbf{B} = (B_1, ..., B_d)$ denotes the length of the spatial domain along each axis and $\cdot$ denotes the scalar product. This is the representation of G in space-time coordinates. Indeed we find that

$$s^x = \left(\mathcal{F}\widehat{G}\xi\right)^x = \int_\Omega G(x - x')\ \xi^{x'}\ \mathrm{d}x' \quad \text{with} \quad \xi^{x'} \equiv (\mathcal{F}\xi)^{x'} \ . \tag{2.20}$$

This equivalence allows for a convenient physical interpretation of G and ξ. Assume that s models the deviations of a physical system from its steady state, induced via an external force, an excitation. Furthermore assume that the response of the system to such an excitation is stationary in space and time, i.E. is the same irrespective of where the excitation happened. This implies that ξ plays the role of the external excitation, while G models the stationary response (also called Green's function) of the system to ξ. This interpretation is purely artificial at this point, however it motivates the prior concepts that we aim to include into G as they are fundamental for the response of a physical system.

The second concept, causality, is introduced via an additional constraint on G. As G should model the response of the system to the excitations ξ, it should not contain a response at times before an excitation happens. This can be formulated mathematically as

$$G_\mathrm{c}(x - x') = \Theta(t - t')\ G(x - x') \ , \tag{2.21}$$

where Θ denotes a step function in time. This trivially ensures that no response happens before an excitation. Technically, as we imposed periodic boundary conditions in time, the constraint has to be modified such that

$$G_\mathrm{c}(x - x') = \Theta(t - t')\ \Theta(T - (t - t'))\ G(x - x') \ . \tag{2.22}$$

The additional step function ensures that an excitation at t' can only cause a response within the interval $[t', T + t']$ but leaves later times unaffected. Together with the expansion of the space by a factor of 2 as discussed in section 2.2.4 it ensures that excitations at T do not wrap around the space to cause a response at $t = 0$.

In space-time, the physical concept of causality also implies a maximal finite propagation speed of interactions. This leads to propagation within a light cone, as depicted in figure 2.1. Therefore, given a maximal propagation speed c, we can restrict the reconstruction to Greens functions that are non-zero only within the light cone. We can implement this constraint by extending eq. 2.21 as

$$G_\mathrm{lc}(x - x') = \Theta(l^2)\ G_\mathrm{c}(x - x') \ , \tag{2.23}$$

where

$$l^2 = (t - t')^2 - (\mathbf{x} - \mathbf{x}')^\dagger \mathbf{C}^{-1} (\mathbf{x} - \mathbf{x}') \ . \tag{2.24}$$

This ensures that the propagator is only non-zero within the light cone. In general we might expect different maximal propagation speeds in different directions, and consequently $\mathbf{C}$ becomes a symmetric tensor. In case propagation is isotropic in space with speed c we have $\mathbf{C}_{ij} = c^2 \delta_{ij}$. Note that even in cases where we do not know $\mathbf{C}$, for example if the

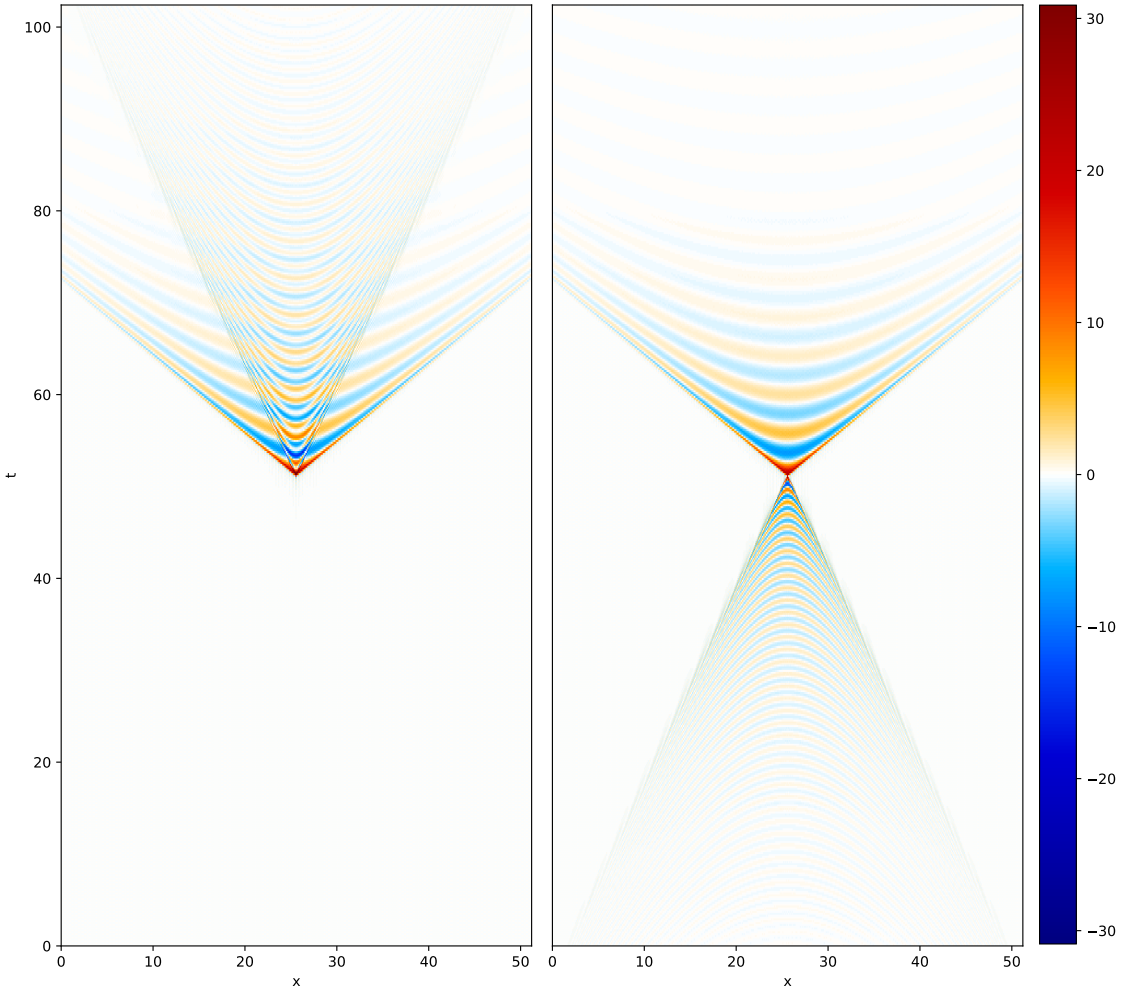

Figure 2.1: **Left:** Causal response of a system defined as a combination of two damped harmonic oscillators with different masses. **Right:** Anti-causal response of the system where one of the oscillations travels backwards in time.

medium in which the interaction is realized is unknown, such a constraint can also be useful if we elevate $\mathbf{C}$ to be an additional unknown parameter that has to be inferred. Since $\mathbf{C}$ is assumed to be the same for all scales, an inference algorithm can use large scale information, where the signal to noise ration (SNR) is usually higher, to determine $\mathbf{C}$, which then effectively increases the SNR for smaller scales. A detailed description how $\mathbf{C}$ enters the reconstruction can be found in appendix 2.A.

In order to encode the last concept, locality, we first have to revisit some properties of Green's functions. Consider a system, undisturbed by external excitation, that can be described via

$$\mathcal{L}s = 0 \ , \tag{2.25}$$

where $\mathcal{L}$ is a linear differential operator. The response G of such a system to an external force has to fulfill

$$(\mathcal{L}G)\,(x - x') = \delta(x - x') \ . \tag{2.26}$$

In order to give rise to a homogeneous G, $\mathcal{L}$ also has to be homogeneous and therefore is diagonal in Fourier space

$$\mathcal{L}(x - x') = \left(\mathcal{F}\widehat{f}\mathcal{F}^\dagger\right) \ , \tag{2.27}$$

where we also introduced the diagonal Fourier space representation of $\mathcal{L}$ denoted via the complex Fourier coefficients f^k. Consequently G takes the form

$$G^k_{k'} = \frac{1}{f^k}\delta^k_{k'} \ , \tag{2.28}$$

and therefore the eigen-spectrum of G reads

$$g^k = \frac{1}{f^k} \ . \tag{2.29}$$

It turns out that locality can be encoded more intuitively in terms of a prior for f. As we can see in Figure 2.2, the locality of a response is related to the order in the derivatives of the differential operator $\mathcal{L}$. Low order derivatives result in an almost instantaneous response of the system while higher order derivatives lead to an apparent non-local response. Therefore we seek to formulate a prior for f such that lower order derivatives are a priori favoured. Nevertheless it should be possible that f can deviate from this assumption if there is enough evidence in the data to support this. We impose a certain degree of smoothness for f, i.E. we want that two modes f^k, $f^{k'}$ are correlated, where the correlation decays as $|k - k'|$ increases. To this end we define the complex modes f as

$$f^k = f_r(k) + i f_i(k) \quad \text{with} \quad k \in \mathbb{Z} \ , \quad f_r, f_i \in L^2(\mathbb{R}) \ . \tag{2.30}$$

In words, the real and imaginary part of the modes f^k are defined via two square integrable functions $f_{r/i}$ which get evaluated at the integer spaced Fourier locations k. We place a Gaussian process prior on both functions f_r and f_i of the form

$$P(f_{r/i}) = \mathcal{G}(f_{r/i}, T) = \frac{1}{|2\pi T|^{\frac{1}{2}}} e^{-\frac{1}{2} f^\dagger_{r/i} T^{-1} f_{r/i}} \ , \tag{2.31}$$

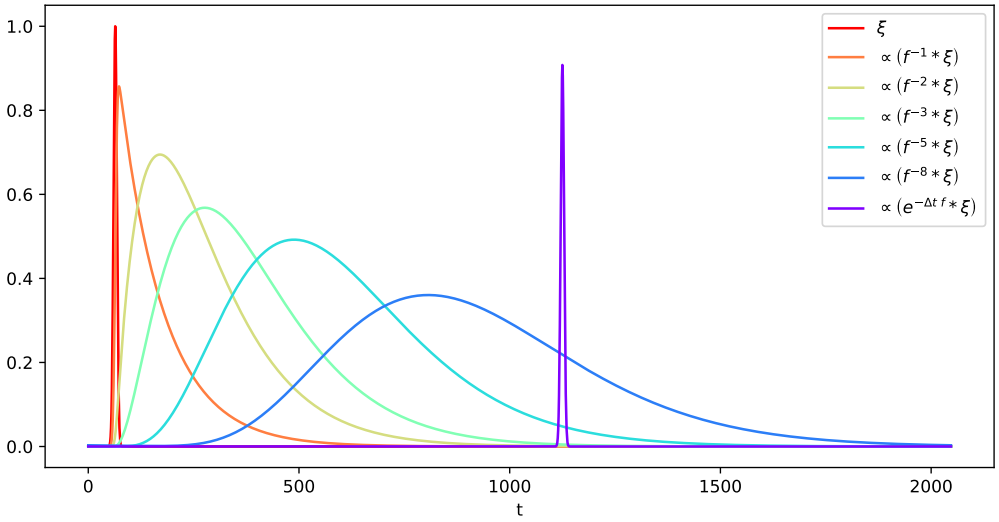

Figure 2.2: Excitation ξ and responses of various systems defined in terms of powers of f, with $f_t = \partial_t + \gamma$ and γ small. We see that higher order derivatives result in apparent non-local responses. Since time derivatives are the generators of temporal translation, the case where the response is the exponential of $-\Delta t\ f$ leads to translations by Δt.

where the exponential factor reads

$$-\frac{1}{2} f_{r/i}^{\dagger} T^{-1} f_{r/i} = -\frac{1}{2\sigma^2} \int \left| (\gamma - \triangle_k)\, f_{r/i}(k) \right|^2 dk \quad \gamma, \sigma > 0 \ , \tag{2.32}$$

where Δ_k denotes the Laplace operator, σ is an overall scaling parameter that steers the strength of this prior, and γ is a low frequency cutoff to ensure that the prior is proper. For further details see [88, 48]. The covariance function associated with T takes the form

$$T(k, k') = \frac{\sigma^2}{2} \sqrt{\frac{\pi}{2\gamma^3}}\ (1 + \sqrt{\gamma}\ \ |k - k'|)\ \ e^{-\sqrt{\gamma}\ |k - k'|} \ . \tag{2.33}$$

Therefore the real and imaginary part of the Fourier modes f^k, which are just the functions $f_{r/i}$ evaluated at $k \in \mathbb{Z}$, are defined to be two independent, infinite dimensional Gaussian random vectors with zero mean and covariance

$$T_{kk'} = T(k, k') \ , \quad k, k' \in \mathbb{Z} \ . \tag{2.34}$$

All in all, we can combine the above concepts to end up with a generative description of our prior which takes the form

$$s(a) = G_{\mathrm{lc}}\ \xi = \mathcal{F}\ \widehat{g_{\mathrm{lc}}}\ \mathcal{F}^{\dagger}\ \xi \ , \tag{2.35}$$

where $\widehat{g_{\mathrm{lc}}}$ denotes a diagonal matrix in Fourier space with $(g_{\mathrm{lc}})^k$ on its diagonal and $a = (f, \xi, \mathbf{C})$ denote the quantities of interest in the generative process. Furthermore

$$g_{\mathrm{lc}} = \mathcal{F}^\dagger \, \widehat{X} \, \mathcal{F} \frac{1}{f} \quad \text{with} \quad X^x = \Theta(T-t)\Theta(t) \; \Theta(l^2) \; , \tag{2.36}$$

with f being a priori distributed according to eq. (2.31).

2.3.1 Comparison to Matèrn type and other parametric kernels

There exists a vast literature about Gaussian process priors with a stationary covariance [50, 94] which discuss a great variety of different forms of covariance functions. Two important classes are squared exponential, and the Matèrn class of covariance Functions. As a stationary covariance allows for a diagonal representation in Fourier space, it makes sense to compare the spectra of the associated operators. Squared exponential kernels imply that the spectrum takes the form

$$g^k \propto e^{-\gamma |k|^2} \quad \gamma > 0 \; . \tag{2.37}$$

While such kernels are very popular as they are particularly easy to implement and use in practice, the quadratic-exponential suppression of small scale structures often appears to be non-physical. A physically better motivated type of spectrum is provided by Matèrn covariances which give rise to spectra of the form

$$g^k \propto \frac{\alpha}{(\beta + |k|^2)^\gamma} \quad \alpha, \beta, \gamma > 0 \; . \tag{2.38}$$

As many physical processes can be well approximated as a power-law, this parametrization provides a more sensible statistical structure. In addition the large-scale cutoff ensures that the process is well defined as the variance remains finite for all k. We notice that the denominator of this spectrum is very well represented by the prior process we impose on f and thus the Matèrn class of covariances is represented in our prior assumptions. In contrast to a fixed Matèrn covariance function, however, the form of f remains unknown prior to the reconstruction and thus is inferred to match the observed data. In addition, a non-parametric process for f appears to be more flexible in modeling deviations from a (possibly idealized) power-law shape of the spectrum.

2.3.2 Prior distributions for excitations

So far we restricted the discussion to independent Gaussian distributed excitations ξ which, for a given fixed G, give rise to a Gaussian process for s.

From the physical perspective of s being the result of a dynamic response G to external excitations ξ, it is not necessary that ξ is Gaussian distributed. In fact, in many applications it might be more realistic to define a different prior for the excitations. In order to demonstrate the implications of a non-Gaussian prior on ξ, in section 2.5, we show results

for the inference problem in cases where ξ is distributed according to an inverse-gamma distribution at each location in space-time. This prior is typically used as a sparsity prior, and in our case results in a system that is subject to sparse external excitations. Of course, many other prior distributions are also reasonable and important, but for the sake of simplicity we stick to ξ being either Gaussian or inverse-gamma distributed in the examples. Note that even though physically motivated, exchanging the prior of ξ to be an inverse-gamma distribution is non-trivial in the continuum limit. The goal of the examples is to demonstrate the applicability of this method also for non-Gaussian excitations, and therefore a rigorous mathematical treatment is beyond the scope of this work. In the case of inverse Gamma excitations, we therefore revert to the discrete representation of the process and leave the continuous treatment to future research.

2.4 Inference

In section 2.2.2 we already discussed the inference problem in the case of a linear measurement and a Gaussian prior with known covariance. Now, with the appropriate prior for the covariance at hand, we can set up the task of inferring the covariance together with the field, given observational data about the field. To this end, consider again a linear measurement as defined in eq. (2.11) which in terms of a takes the form

$$d = R\ s(a) + n\ , \tag{2.39}$$

where $s(a)$ is given via Eq. (2.35). If we assume n being Gaussian distributed with zero mean and covariance N we get that the joint distribution of (d, a) reads

$$P(d, a) = \mathcal{G}(d - R\ s(a), N)\ \mathcal{G}(f, T)\ P(\xi)\ . \tag{2.40}$$

The posterior $P(a|d)$ is proportional to the joint distribution up to a factor that only depends on d since

$$P(a, d) = P(a|d)\ P(d) \propto P(a|d)\ . \tag{2.41}$$

This posterior is intractable, due to the non-linear dependency of s on a. Consequently the corresponding inference problem cannot be solved analytically and we have to rely on a numerical approximation.

There exist a variety of different approximation techniques for Bayesian inverse problems ranging from point estimations such as the maximum a posterior (MAP) estimate, over variational approximations, to posterior sampling techniques such as Markov-Chain Monte Carlo (MCMC) [19, 25] and Hybrid Monte Carlo (HMC) [29]. As shown in [71], MAP tend to perform poorly in the task of reconstructing the excitations together with the prior correlation structure, as uncertainty information is vital to correctly estimate the prior statistics, which are missing in point estimates. While MCMC and HMC algorithms are very attractive due to their theoretical guarantees to converge to the true posterior statistics, they tend to become expensive for many astrophysical field inference applications compared to simpler, less expressive approaches. Therefore, in this work, we use a

variational approximation algorithm called Metric Gaussian Variational Inference (MGVI) [71] where we approximate the true posterior with a Gaussian distribution in order to get an estimate for the mean and the covariance of the posterior. As shown in [71], MGVI provides an accurate estimate of the first moment (i.E. the posterior mean) as well as a tight lower bound on the second moment (posterior covariance) when compared against HMC techniques, while being substantially faster.

2.4.1 Variational Inference

In general, variational inference can be described as the task of approximating one probability distribution $P(x)$ for some quantity x with another distribution $Q_\sigma(x)$ which, in addition, is defined up to a set of parameters σ. Approximation is then achieved via minimizing the Forward Kullbach-Leibler divergence (KL) with respect to σ. The KL is defined as

$$\begin{aligned} \mathrm{KL}(Q_\sigma, P) &\equiv \int Q_\sigma(x) \log\left(\frac{Q_\sigma(x)}{P(x)}\right) \mathrm{d}x \\ &= \langle H_P \rangle_{Q_\sigma} - \langle H_{Q_\sigma} \rangle_{Q_\sigma} \ , \end{aligned} \tag{2.42}$$

where we also introduced the so called Information Hamiltonian H defined as

$$H_P = -\log(P) \ . \tag{2.43}$$

In our case, the approximate distribution Q is chosen to be a Gaussian distribution in a with mean m and covariance A. For many relevant inference problems, and also for the one studied in this work, this approximation cannot be performed analytically as typically $\langle H_P \rangle_{Q_\sigma}$ cannot be calculated analytically. Therefore, as discussed in section 2.2.3, we choose an appropriate discretization for the space-time domain Ω, and perform a variational approximation of the corresponding discrete problem. It turns out that in order to achieve a reasonable resolution in space-time, the inference problems can become very high dimensional. As an example, in section 2.5, we show an application for a discretized 1+1 dimensional space-time with a resolution of 256×200 pixels. This yields that the number of dofs in a is $\approx 2 \times 256 \times 200 \propto 10^5$ and consequently the number of entries in A are $\propto 10^{10}$ which renders an explicit representation of A on a computer to be inefficient generally. Therefore, following [71], we avoid an explicit representation by setting it to be equal to the inverse Fisher Metric $\mathcal{M}^{-1}$ [4] of the posterior, evaluated at m. For the posterior distribution as defined in eqs. (2.40) and (2.41), together with a Gaussian prior for ξ with zero mean and unit covariance, we get that $\mathcal{M}$ takes the form

$$\mathcal{M} = \left[\left(\frac{\partial s}{\partial a}\right)^\dagger R^\dagger N^{-1} R \frac{\partial s}{\partial a} + \begin{pmatrix} \mathbb{1} & 0 \\ 0 & T^{-1} \end{pmatrix} \right]_{a=m} \equiv \left(\tilde{R}_m\right)^\dagger N^{-1} \tilde{R}_m + S^{-1} \ . \tag{2.44}$$

Using the definition of s (eq. (2.35)) we get that

$$\frac{\partial s}{\partial \xi} = \mathcal{F}^{-1}\, \widehat{g}\, \mathcal{F}\ , \tag{2.45}$$

$$\frac{\partial s}{\partial f} = \mathcal{F}^{-1}\widehat{\mathcal{F}\xi}\mathcal{F}\ \widehat{X}\ \mathcal{F}^{-1}\widehat{-1/f^2}\ . \tag{2.46}$$

Therefore, together with the definition of T (eq. (2.32)) we see that $\mathcal{M}$ has an implicit representation, i.e. it can be applied solely using Fourier transformations and diagonal operations, avoiding the explicit storage of this matrix at any point. Consequently the application of $A = \mathcal{M}^{-1}$ is achieved via linear solvers such as the conjugate gradient method. In addition, the structure of $\mathcal{M}$ allows for an efficient sampling of the approximate distribution Q. Specifically we may draw a random realization of Q as

$$a^* = m + \mathcal{M}^{-1}\left(\left(\tilde{R}_m\right)^\dagger n^* + b^*\right)\ , \tag{2.47}$$

where n^* and b^* are independent samples drawn from the noise statistics and the joint prior distribution with covariance S, respectively.

This is another important property since the KL involves expectation values w.r.t Q, which can be approximated via samples from Q. Ultimately, the Fisher metric is also a measure for the local curvature of the KL and therefore enables us to use second order optimization schemes to solve the corresponding optimization problem in m.

As discussed in the previous section, in some cases space-time causality can only be imposed if we also infer the propagation speed $\mathbf{C}$. To do so, we notice that given a prior for $\mathbf{C}$, the above still applies with the extension that s is now also a function of $\mathbf{C}$. Therefore the Fisher metric gets an additional entry with the same structure as in eq. (2.44), for the corresponding gradient $\partial s/\partial \mathbf{C}$ (See appendix 2.A for further details).

All in all, the solution strategy as defined in MGVI, starting from a random initialized m, can be summarized as:

- Using $\mathcal{M}$, evaluated at the current approximate mean m, draw a set of samples $\{a^*\}$ from the approximate distribution Q.
- With these samples, calculate an estimation for the current value of the KL and its gradient.
- Together with the metric $\mathcal{M}$ perform a second order Newton minimization in order to get a new estimate for m.
- Repeat this procedure by re-evaluating everything using the updated m, until convergence.

2.5 Application

To demonstrate the applicability of our method we apply it to several synthetic data examples. First, we demonstrate the performance of the algorithm in a one dimensional setting, where we only aim to infer the Green's function G of the system, given the excitations and data. In the second example we perform a reconstruction of excitations that are distributed according to a unit Gaussian, as well as the Green's function from data alone, in a 1+1 dimensional setting. In the last example we add another level of complexity via non-Gaussian statistics in the excitations. Specifically we assume the excitations to follow an inverse-gamma distribution at each location in discretized space-time. Hereby we perform source detection, the task of inferring sparse excitations at various locations in space-time, in a case where also the Green's function is unknown. Throughout all examples, the prior model for G follows the one described in section 2.3. Specifically, we describe G in terms of f and the propagation speed $\mathbf{C}$. We place a flat prior on the logarithm of $\mathbf{C}$ to ensure positivity of the propagation speed.

2.5.1 Implementation details

All examples were implemented using the NIFTy software package in its version 5 [6]. We included an implementation of the here introduced prior structure for the Green's function into this publicly available package.

Throughout all examples, to solve the optimization problem associated with the MGVI algorithm, we realize the algorithm described in 2.4.1 with 30 total steps of the entire loop, where we draw 10 samples from the approximation Q to estimate the KL during optimization. For posterior analysis, we use 50 approximate posterior samples.

For the first, one dimensional example, the optimization can be performed within less then a minute, while for the other two examples the total runtime is around 10 minutes on a standard laptop. The runtime and scaling of MGVI with model complexity as well as dimensionality is described in great detail in [71].

2.5.2 Temporal evolution

In our first application, shown in figure 2.3, we aim to demonstrate the inference of the dynamics encoding field f alone in a one dimensional setting where there is only a temporal evolution to be reconstructed. The excitation field ξ is known during inference. The hyperparameters of the prior for f (see Eq. (2.31)) are set to $(\sigma, \gamma) = (2.9, 1.7)$. We generate synthetic data according to eq. 2.39, with R being the identity. The excitations were drawn from an inverse gamma distribution to model known, sparse excitations of the system (e.g. in a laboratory setting where the unknown system is driven via sparse excitations). The synthetic signal s as well as corresponding data d is shown in the top-left panel of figure 2.3. The dynamic operator used to generate signal and data is of the form:

$$\mathcal{L} = (\partial_t^2 + m^2 + \gamma\partial_t)^5 \ (\partial_t^2 + \tilde{m}^2 + \tilde{\gamma}\partial_t) \ , \tag{2.48}$$

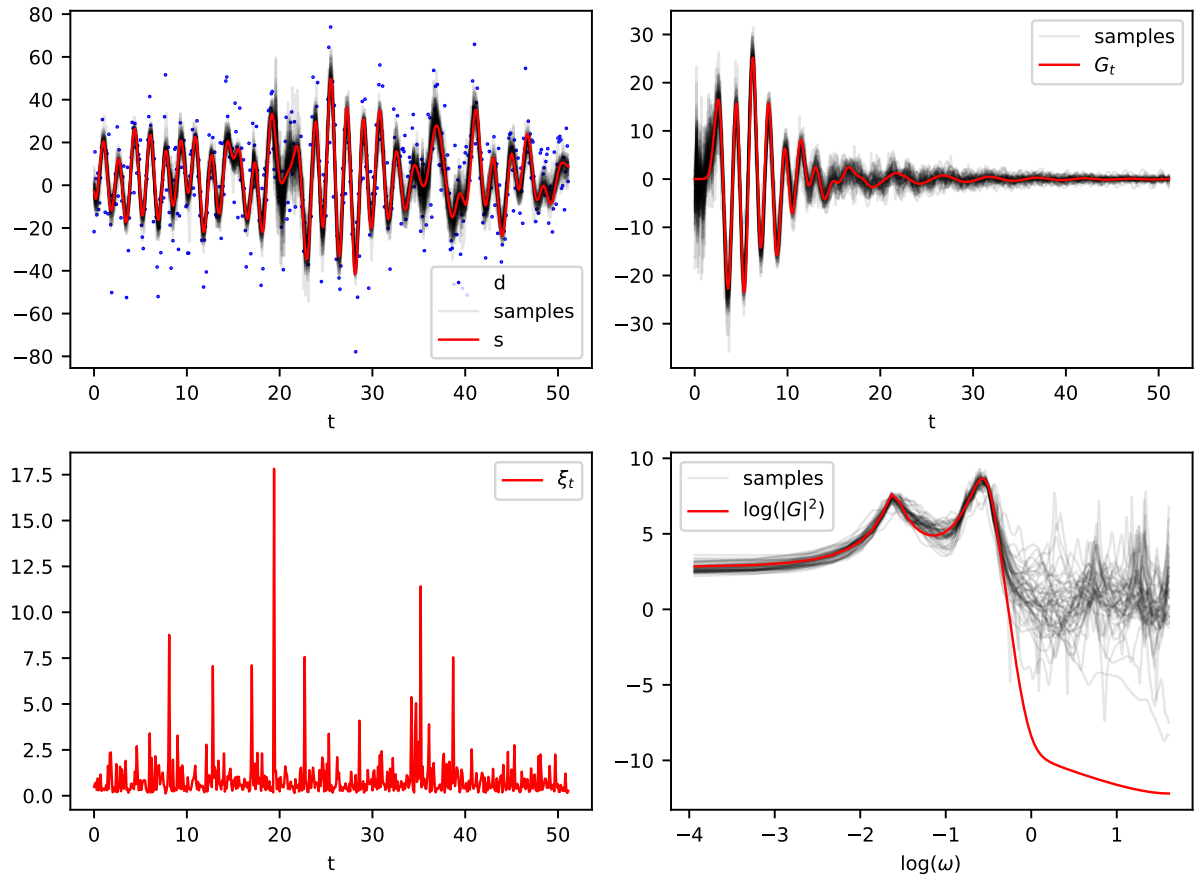

Figure 2.3: **Top:** On the left we depict the signal (red line), the data (blue dots) as well as 50 over-plotted posterior samples (gray). The right panel shows the synthetic propagator (Greens function) in the temporal domain (red) as well as corresponding posterior samples. **Bottom:** Left: Excitation field used to generate the signal. Right: Natural logarithmic spectrum of the synthetic propagator (red) and corresponding posterior samples.

with $(m, \gamma, \tilde{m}, \tilde{\gamma}) = (0.6, 0.23, 0.2, 0.03)$. The results of the reconstruction are shown in figure 2.3.

We see that the reconstruction of the Green's function is in agreement with the ground truth in the temporal domain, within uncertainties. Due to the fact that the reconstructed dynamics is uncertain, the recovered signal also has uncertainty although the excitations are known. The reconstructed Green's function indicates that it is indeed possible to reconstruct an apparent non-local response of the system (due to higher order derivatives in this setup) since the true propagator as well as its reconstruction show oscillations that grow in the beginning of the propagator before decaying exponentially. We also notice that there is relatively high uncertainty in the first timesteps of the response G_t. This is caused by the low initial response to excitations of the true propagator. The initial part of the reconstructed propagator is purely dominated by noise and thus only constrained

up to the noise level (the standard deviation of the noise σ_n is set to be $\sigma_n = 15$ and the temporal domain is discretized via 512 equidistant pixels).

We also notice that the posterior solution for the spectrum levels out for high-frequency modes (large ω) below the signal to noise ratio, while the true spectrum continues to decay (ultimately also the true spectrum levels out in the numerical example due to the finite size of the considered space). The uncertainty increases in this region, but not enough to capture the true solution. Here we notice the limitations of the variational inference, which provides a local approximation of the posterior with a Gaussian. Consequently the true uncertainty might be underestimated, as in this case. However this deviation occurs orders of magnitudes below the peak of the spectrum and therefore has only a barely visible effect on the reconstruction of the signal. One way to allow for a better extrapolation to higher frequencies would be to provide a more restrictive prior for the dynamics encoding field e.g. by defining it on a polynomial basis which is a more suitable basis for this particular setup as the true dynamics is also described in terms of a polynomial. However we aim to provide a general and less restrictive approach here capable of also reconstructing non-polynomial dynamics encoding functions and consequently being less able to extrapolate to regions where we have no information provided by data.

In addition, in figure 2.4, we depict the reconstructed real and imaginary part of the inverse of the propagator spectrum and compare it to the spectrum associated to the differential operator $\mathcal{L}$ (Eq. (2.48)) used to generate the mock data. We see that the true spectrum is in agreement with the reconstruction, within uncertainty. Furthermore we notice that the posterior uncertainty is small close to the two resonant frequencies ω_r (see figure 2.4) corresponding to the two peaks in the propagator spectrum depicted in figure 2.3. This is due to the fact that at these frequencies both, the real and imaginary part of the spectrum, are close to zero and thus the magnitude of its inverse is large. Therefore small deviations around these values result in large changes in the corresponding realization of the random process s and therefore the posterior uncertainty has to be small in order to stay consistent with the data.

To further quantify the reconstruction error, we also investigate the residual as well as the corresponding posterior uncertainty for the signal s_t and the time representation of the Greens function G_t (see figure 2.5). The residual is defined as the difference between the true solution and the posterior mean of the reconstruction. In addition to the case of $\sigma_n = 15$, we also show the residuals and uncertainties for various other noise levels. For a better comparison, no other changes where made during reconstruction. In particular also the random number generator used to generate the synthetic data as well as the approximate posterior samples during reconstruction was seeded with the same random seed for all runs. We notice that the posterior uncertainty appears to be on a reasonable scale as the residual is within the one or two sigma confidence interval for almost all cases. In addition, we notice that the posterior uncertainty of the signal s is particularly high in regions right after a strong excitation happened. This is due to the fact that the reconstruction of the Green's function G, which is the response to these excitations, is most uncertain for the first timesteps. This uncertainty propagates into the uncertainty of s.

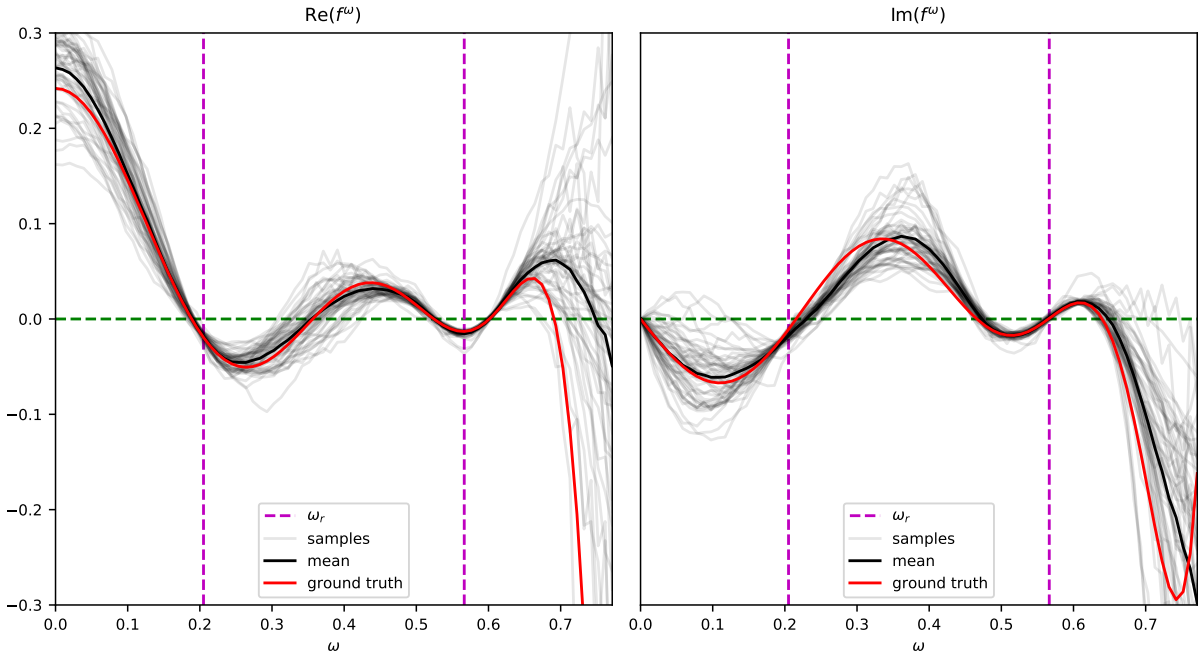

Figure 2.4: Posterior mean (black line) and posterior samples (gray lines) of the real part (left) and imaginary part (right) of the inverse of the propagator spectrum $f^\omega = 1/g^\omega$. The red lines indicate the real and imaginary part of the differential operator $\mathcal{L}$ (Eq. (2.48)) used to generate the data of this example. The purple dashed lines indicate the values of the two resonant frequencies ω_r corresponding to $\mathcal{L}$ where the magnitude of f is smallest and thus the contribution to the observed process s is largest.

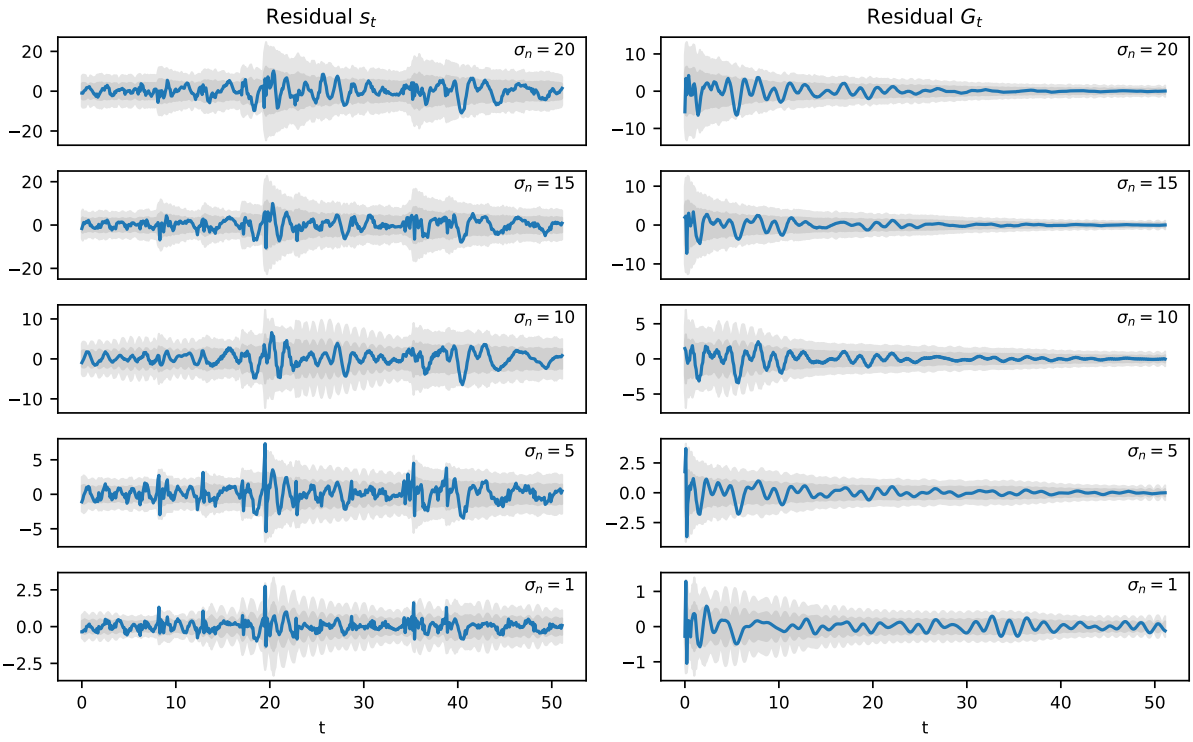

Figure 2.5: **Left:** Residuals between the true signal s_t and the posterior mean of the reconstruction (blue line) as well as the one and two sigma confidence intervals of the corresponding posterior uncertainty. We show those for various different noise standard deviations σ_n starting with the highest noise level at the top to the lowest at the bottom. **Right:** Corresponding residuals and confidence intervals for the temporal representation of the dynamic Green's function G_t again for various noise levels. In all inference runs, we seeded the random number generator used for data generation and during reconstruction with the same random seed such that the only difference in these reconstructions is a different σ_n.

2.5.3 Space-time evolution

In our second example, shown in figure 2.6, we aim to reconstruct the dynamics as well as the excitations in a spatio-temporal (1+1 dimensional) setting from incomplete and noisy observations. We aim to infer the dynamical field f, the propagation speed c, and the excitations ξ from noisy and incomplete measurements d of the field s alone. The hyper-parameters of the prior for f (see Eq. (2.31)) are set to $(\sigma, \gamma) = (1.8, 0.5)$. The dynamical system used for the generation of synthetic data is a product of a damped harmonic oscillator and an advection-diffusion generating term. This reads

$$\mathcal{L} = (\partial_t^2 - c^2\partial_x^2 + m^2 + \alpha\partial_t - \beta\partial_x)\ (\partial_t + \tilde{m}^2 - \gamma\partial_x^2 + \tilde{\beta}\partial_x)\ . \tag{2.49}$$

Furthermore, the excitations are Gaussian distributed with zero mean and unit covariance from which a single realization was drawn and convolved with the synthetic Green's function corresponding to $\mathcal{L}$. The signal s, the data d and the reconstruction of s are shown in figure 2.6 for a case with $(c, m, \alpha, \beta, \tilde{m}, \gamma, \tilde{\beta}) = (0.4, 0.16, -0.19, -0.05, 0.1, 0.5, 0.2)$. We generate the data according to eq. 2.39 with a linear measurement response, which partially ($\approx 25\%$) masks the observed region, and Gaussian distributed noise with $\sigma_n = 10$. The space-time is discretized via a regular grid with 256×200 pixels, respectively.

The reconstruction algorithm is capable of reconstructing the signal in regions where we have observations thereof, while being relatively blind in unobserved regions. Consequently the posterior uncertainty is higher there. In addition, we notice that unlike the posterior mean, posterior samples consistently fill unobserved regions. Although in these regions the samples deviate strongly from the true signal, the information on the statistical properties, inferred from the observed regions, propagates into the unobserved regions due to the assumed statistical homogeneity. Therefore the posterior samples are statistically consistent throughout the entire space-time interval, which is important for posterior analysis. Furthermore we notice that there also exists variance in the statistical properties of the posterior, as can be seen for example in the difference between small scale structures of the posterior mean and the sample displayed in figure 2.6. This is due to the fact that also the reconstruction of the statistical properties (described via f) is imperfect due to the noisy data and thus subject to uncertainty. This posterior uncertainty about the small scale properties of f results in a variation between different posterior samples of f, which ultimately propagates into the statistical properties of the corresponding sample of s.

In figures 2.7 and 2.8 we study the posterior properties of the Green's function in more detail. In particular we compare the reconstructed Green's function as well as the corresponding spectrum with the underlying ground truth. The spectrum is comparable to the ground truth in regions with sufficient SNR while it levels out in regions where we have no information given via data. In addition, the reconstructed propagator also shows oscillations consistent with the true propagator. However we notice that modes that propagate "downwards" are reconstructed well while the weaker "upwards" propagating modes are not reconstructed due to the fact that they are below the noise level. In addition, we see that deviations in posterior samples of the propagator only occur within a "cone" and remain zero outside. This is due to the fact that we also reconstruct the maximal

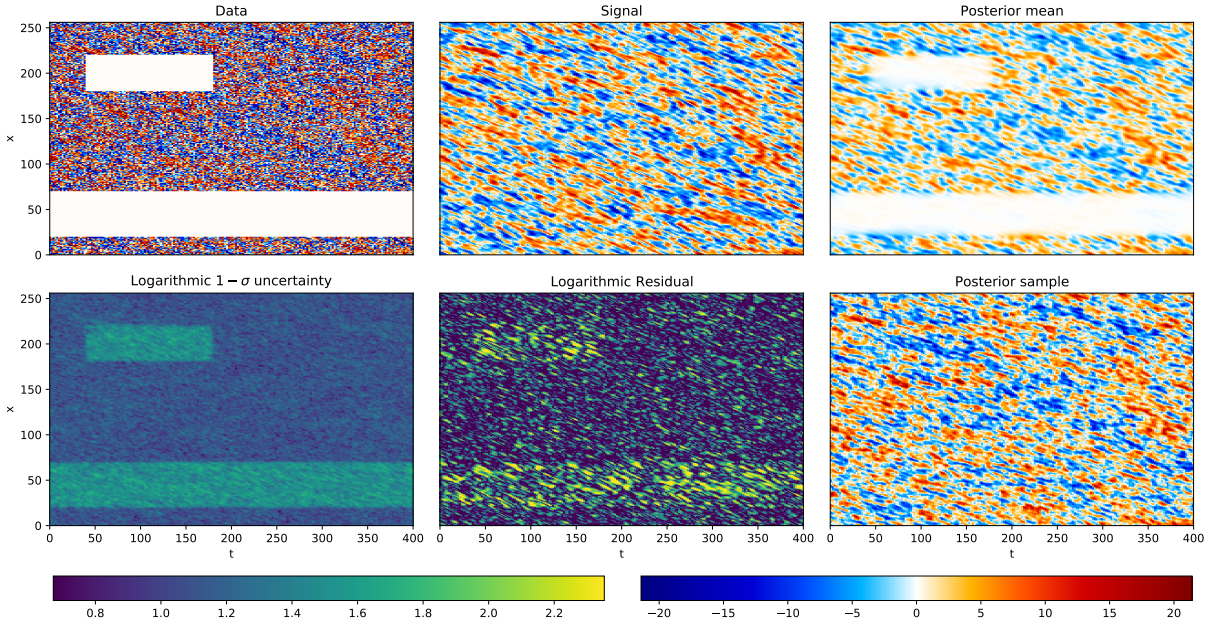

Figure 2.6: **Top:** The left panel shows the spatio-temporal masked and noisy data, drawn from the synthetic signal (middle panel) and the corresponding signal reconstruction (right panel). **Bottom:** Natural logarithmic one-sigma posterior uncertainty (left panel), natural logarithmic residual between the true signal and the posterior mean (middle panel), as well as an approximate posterior sample (right panel).

propagation speed of the process, which is reconstructed to be $c \approx 0.45$ (± 0.08) enclosing the correct value of $c = 0.4$.

2.5.4 Source detection

In our last example we aim to perform source detection in the excitation field, in a case where also the dynamic response is unknown. To demonstrate this scenario we again generate a 1+1-dimensional synthetic example where in this case we assume that we are only able to measure the temporal evolution of the system at several locations. In particular we measure the temporal evolution at 50 randomly selected locations of the space under consideration. This results in $\approx$ 80% of the discrete space-time being unobserved, as the resolution is the same as in the previous example. As before, we assume that the measurements are subject to additive Gaussian noise with $\sigma_n = 0.3$ and also assume the system to be at rest at $t = 0$. The resulting data is shown in the top-left panel of figure

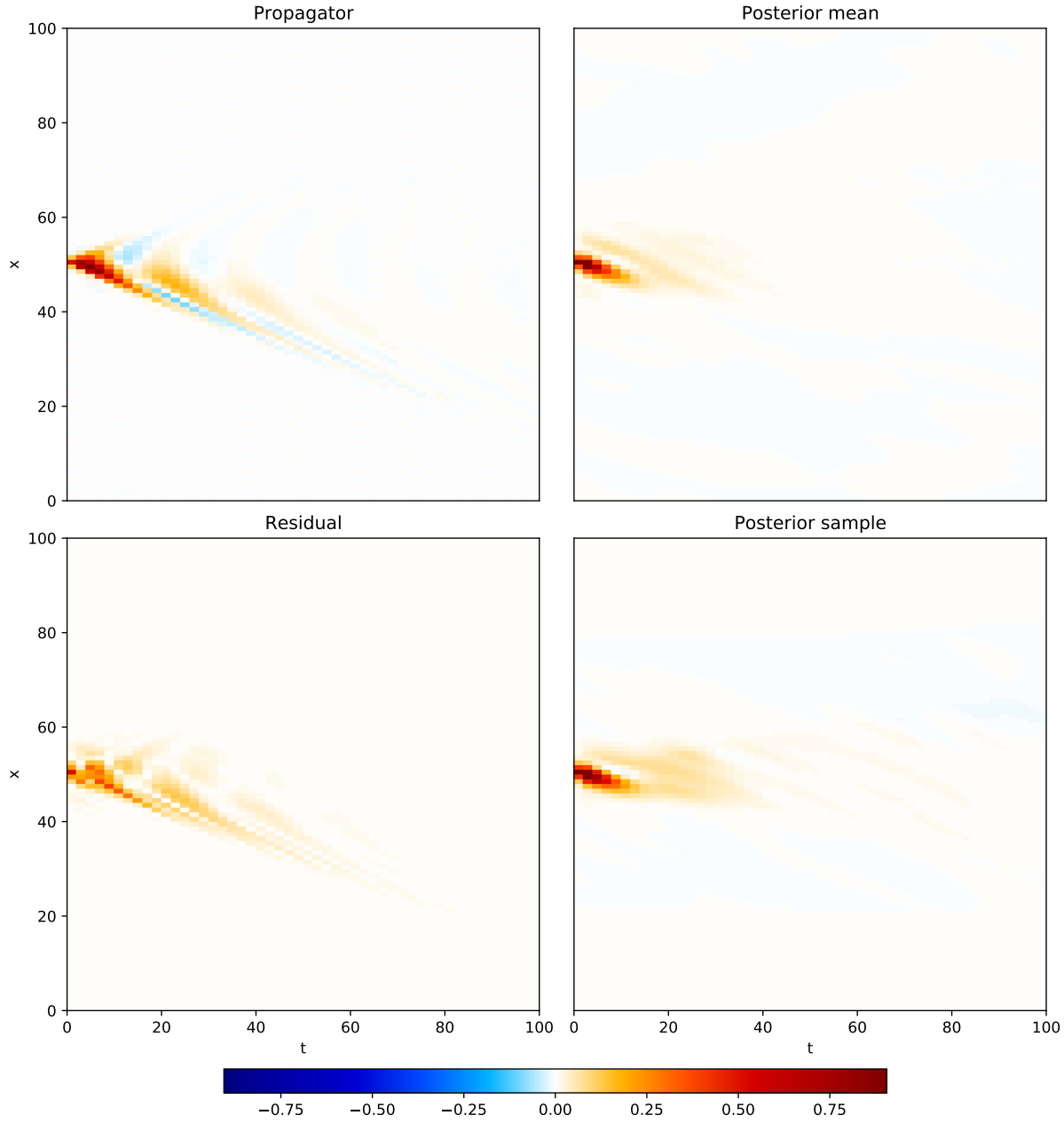

Figure 2.7: **Top:** True Green's function (left) and corresponding posterior mean (right). **Bottom:** Residual of the true Green's function and the reconstruction (left) and a posterior sample for the Green's function (right).

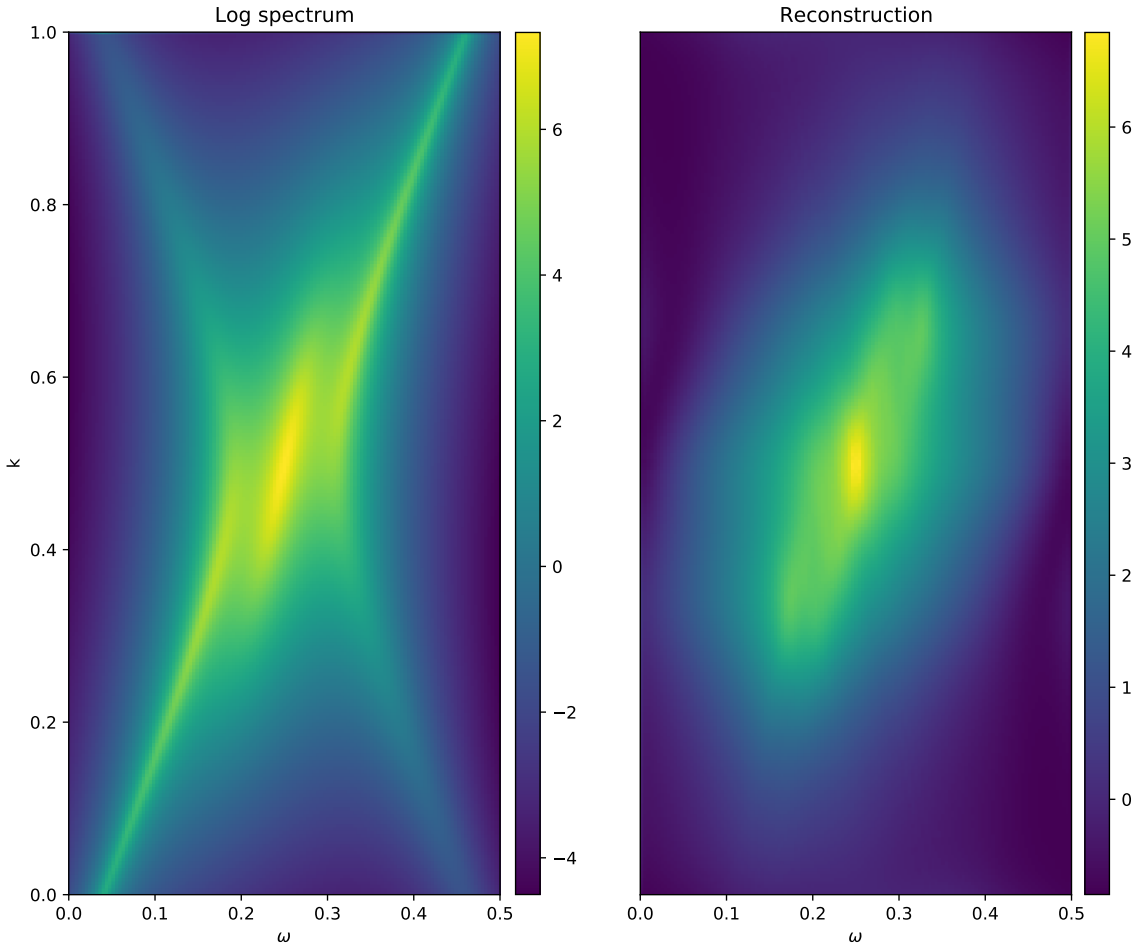

Figure 2.8: Natural logarithmic spectrum of the true Green's function (left) as well as the corresponding posterior mean (right).

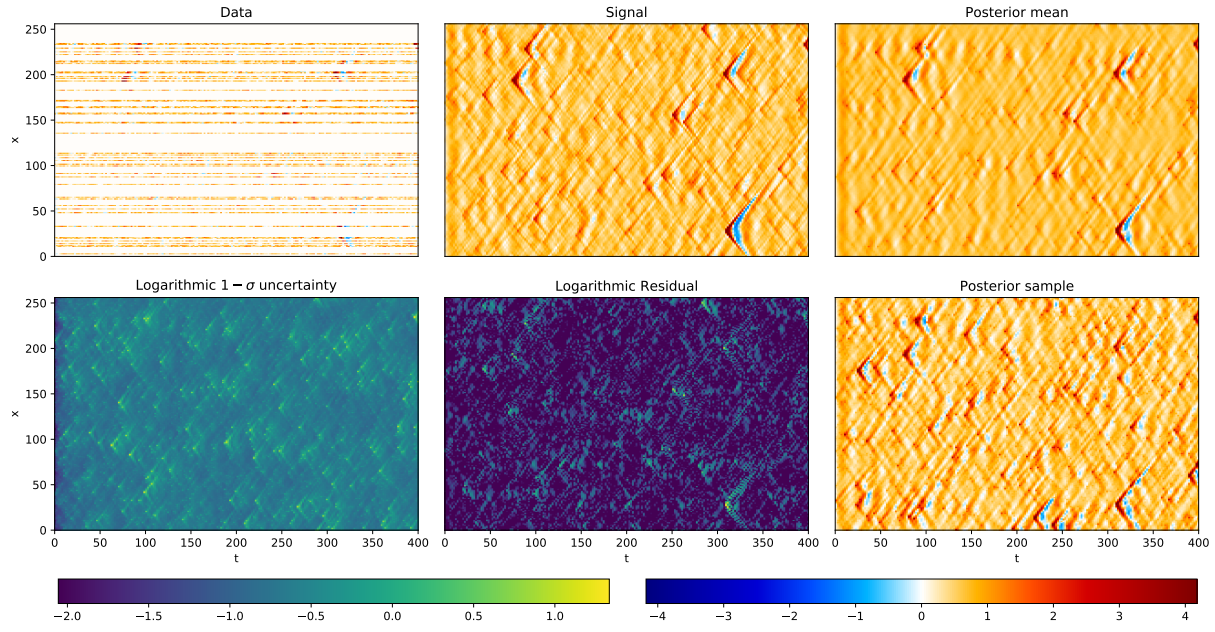

Figure 2.9: **Top:** The left panel shows the sparse and noisy measurement data, drawn from the synthetic signal (middle panel) and the corresponding reconstruction (right panel). **Bottom:** Natural logarithmic one-sigma posterior uncertainty (left panel), natural logarithmic residual of the true signal and the corresponding posterior mean (middle panel), as well as an approximate posterior sample (right panel).

2.9. The unknown excitations are inverse gamma distributed to model strong but sparse excitations. We infer those from measurements of the system at multiple locations together with the dynamics encoding function f. The hyper-parameters of the prior for f (see Eq. (2.31)) are set to $(\sigma, \gamma) = (2.9, 1.7)$. The system used to generate the data exhibits damped traveling waves described by

$$\mathcal{L} = \partial_t^2 - c^2\partial_x^2 + m^2 + \alpha\partial_t - \beta\partial_x \ , \tag{2.50}$$

with $(c, m, \alpha, \beta) = (1.2, 0.04, -0.013, 0.005)$. From an information theoretical point of view this setup is very similar to the previous one since we only changed the measurement response to describe measurements of the temporal evolution at several locations, as well as the prior for excitations to be an inverse gamma prior. Consequently also the inference can be performed in the same way as before.

The setup as well as the reconstruction of the field evolving in space-time are shown in figure 2.9. We see that the reconstruction recovers many sources and the corresponding

propagation. The algorithm uses information from the response of strong excitations to reconstruct the Green's function of the system. Due to the assumed homogeneity in space-time, this information helps to improve the overall reconstruction in other regions. The quality of the reconstruction of single excitations additionally depends on the surrounding measurement scenario.

In figures 2.10 and 2.11 we depict the dynamic propagator as well as the spectrum and reconstructions. We can validate that the Green's function was indeed reconstructed correctly, within uncertainties. We conclude that the task of source detection is possible even in cases where the underlying dynamics is unknown, as long as the assumptions of spatio-temporal homogeneity, causality, and locality hold.

2.6 Conclusion

In this work we considered the problem of reconstructing a random field s, defined in space and time, together with its correlation structure S from noisy and incomplete data d about s. We have shown that this Bayesian hierarchical inference problem can be reformulated to a (theoretically) equivalent problem by means of a generative process, where we aim to infer an excitation field ξ as well as the dynamic response G. Ultimately the eigen-spectrum of G was encoded in the dynamics encoding field f and the propagation-speed encoding parameter $\mathbf{C}$. Together with the excitations ξ they denote the quantity of interest a of the inference problem. We proposed a Gaussian process prior for f which gives rise to a non-parametric description of the dynamic response G. This gives rise to a non-linear generative process for s. As the proposed method is also applicable for non-Gaussian prior distributions for ξ it can also model a variety of other, physically motivated, prior distributions for s. This flexibility is discussed for the example of an inverse gamma distribution for the excitations.

To demonstrate its applicability, the proposed method is applied to several synthetic data examples. These include a one dimensional example where the excitations were known and only G had to be inferred, a 1+1 dimensional example with unknown Gaussian distributed excitations as well as an example with inverse gamma distributed ξ and a measurement response that is sparse in the spatial domain.

As we restricted the prior assumptions for G to the physically motivated concepts of space-time homogeneity, locality, and causality, the method appears to be applicable in a wide range of problems. One particular strength is the non-parametric formulation of the Green's function. This becomes important in scenarios where physical models cannot provide a simple parametric description of evolution so far, to describe the Green's function. In addition, a probabilistic description of excitations is sufficient for inference. Consequently the method is still applicable in cases where external influences cannot be described completely.

All in all, we believe that the proposed method is capable of dealing with current as well as upcoming inference problems involving fields defined over space and time, arising from the context of astrophysical imaging.

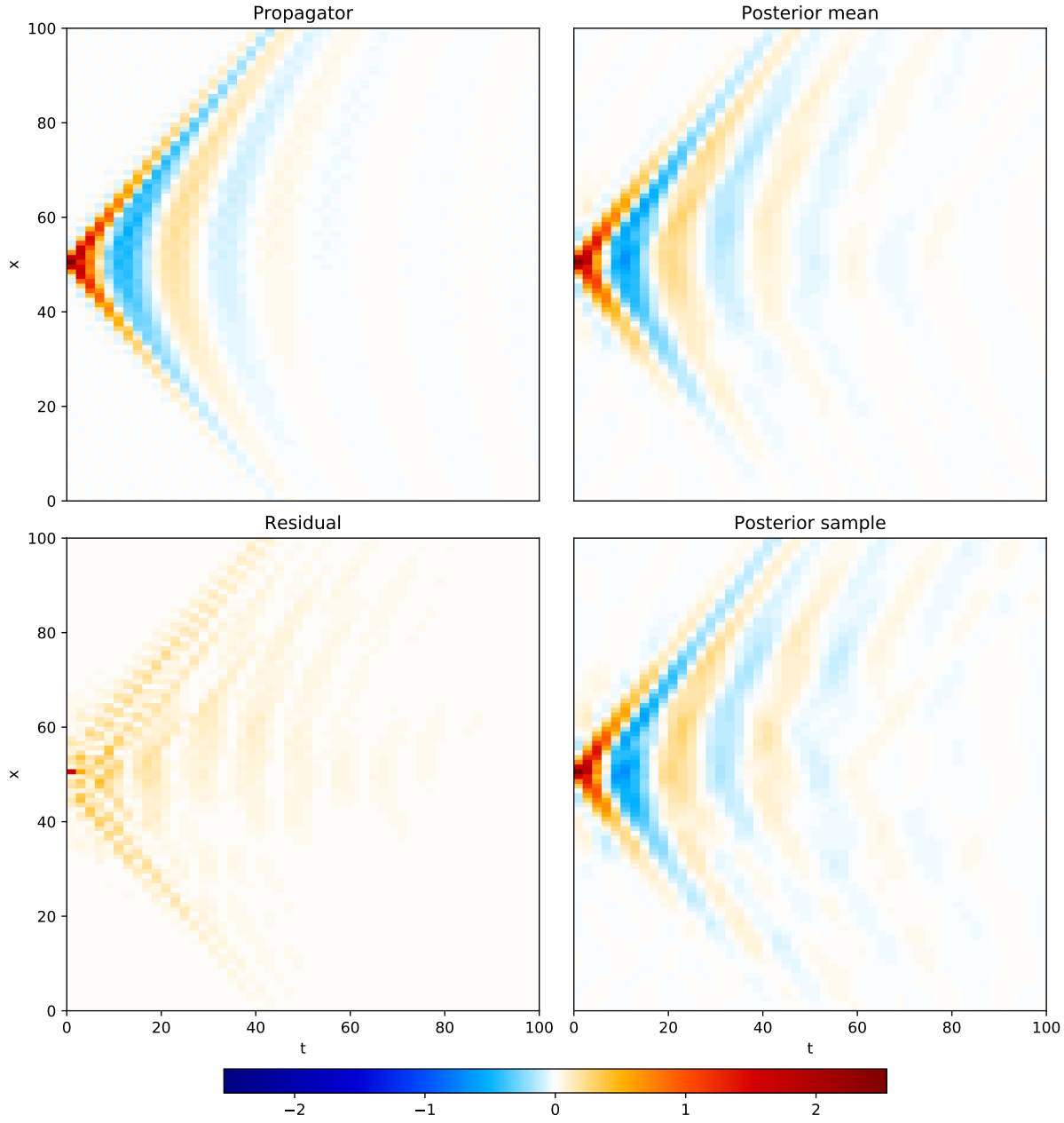

Figure 2.10: **Top:** True Green's function of the process defined in eq. 2.50 (left) and corresponding posterior mean (right). **Bottom:** Residual of the true Green's function and the reconstruction (left) and a posterior sample for the Green's function (right).

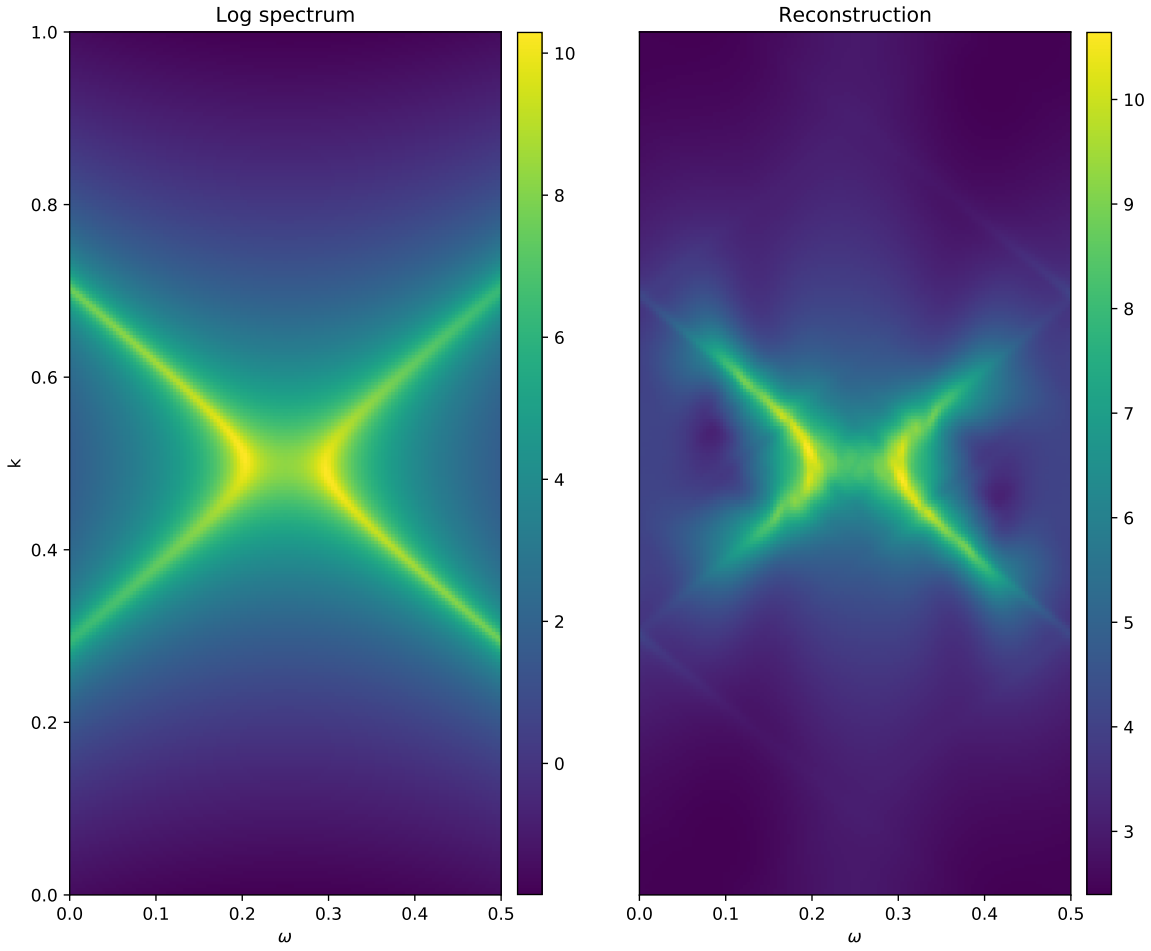

Figure 2.11: Natural logarithmic spectrum of the true Green's function (left) as well as the corresponding posterior mean (right).

Appendix

2.A Light cone prior on a discretized space

As discussed in section 2.3 the concept of causality in space-time results in a restriction of propagation within a light cone. In a continuous description, this restriction is realized via a convolution with a step function of the form

$$\Theta(l^2) = \Theta\left(t^2 - \mathbf{x}^\dagger \mathbf{C}^{-1} \mathbf{x}\right) \ . \tag{2.51}$$

Due to the fact that our calculations are ultimately performed on a finite grid, this definition appears to be somewhat problematic, as it introduces boundary effects along the edges of the step function, when realized on a discretized space. In addition, if we aim to elevate $\mathbf{C}$ to be an unknown parameter of the problem that has to be inferred, gradient based methods are no longer applicable due to the fact that the gradient is zero almost everywhere in space-time (or not defined on the boundary). Therefore we seek to find a way to relax the sharp boundary introduced via the cone, without loosing its useful properties. To do so we borrow an idea from quantum field theory where it turns out that these sharp boundaries are "smeared" out when considering them on a quantum scale. In our case the "quantum" scale can be regarded as the resolution of the discrete representation of space-time, although this analogy is purely artificial.

To achieve this relaxation, consider the following quantity

$$\Delta = \sqrt{-l^2} = \sqrt{-\left(t^2 - \mathbf{x}^\dagger \mathbf{C}^{-1} \mathbf{x}\right)} \ . \tag{2.52}$$

It has two useful properties: For causal (including time-like as well as light-like) points the real part of Δ, $\Re(\Delta)$, is zero as the square root is taken from a negative number. For non-causal (space-like) points the real part becomes positive. Furthermore, for fixed t, $\Re(\Delta)$ is asymptotically linear in x. Therefore, if we consider a Gaussian in $\Re(\Delta)$,

$$\theta(\Delta) \equiv \exp\left(-\frac{1}{2\sigma^2}\Re(\Delta)^2\right) \ , \tag{2.53}$$

we notice that this quantity remains one within the light cone, while it asymptotically falls off like a Gaussian in x, for fixed t. Here σ stands for an optional scaling parameter which controls the width of the Gaussian. In all applications of this paper we replace the function Θ (Eq. (2.51)) with θ and set σ to the size of a few pixels.

Acknowledgements

We would like to thank Jakob Knollmüller, Philipp Arras and Margret Westerkamp for fruitful discussions, Martin Reinecke for his contributions to NIFTy, and two anonymous referees for numerous comments that significantly improved the mathematical and overall presentation of the subject.

Chapter 3

M87* in space, time, and frequency: Interferometric imaging of variable sources

The following chapter contains parts of a manuscript that has been submitted to Nature Astronomy *[8]. It is the result of a collaborative effort with equal contributions made by Philipp Arras, Philipp Haim, Jakob Knollmüller, Reimar Leike, and me. All authors contributed text to this publication. Philipp Arras, Philipp Haim, Jakob Knollmüller, Reimar Leike, and I implemented and tested the instrument response, likelihood, and model. Jakob Knollmüller developed the inference heuristic. Philipp Arras and Jakob Knollmüller performed the hyperparameter study. Philipp Arras and I contributed the amplitude model which features outer products of power spectra. Martin Reinecke provided implementations and numerical optimisation for many of the employed algorithms. Torsten Enßlin coordinated the team and contributed to discussions. The text has been written as a collaborative effort by all of us unless otherwise specified below.*

Abstract

Observing the dynamics of compact astrophysical objects provides insights into their inner workings, thereby probing physics under extreme conditions. The immediate vicinity of an active supermassive black hole with its event horizon, photon ring, accretion disk, and relativistic jets is a perfect place to study general relativity, magnetohydrodynamics, and high energy plasma physics. The recent observations of the black hole shadow of M87* with *Very Long Baseline Interferometry* (VLBI) by the *Event Horizon Telescope* (EHT, [36, 37, 38, 39, 40, 41]) open the possibility to investigate its dynamical processes on time scales of days. In this regime, radio astronomical imaging algorithms are brought to their limits. Compared to regular radio interferometers, VLBI networks typically have fewer antennas and low signal to noise ratios (SNRs). If the source is variable during the observational period, one cannot co-add data on

the sky brightness distribution from different time frames to increase the SNR. Here, we present an imaging algorithm[a] that copes with the data scarcity and the source's temporal evolution, while simultaneously providing uncertainty quantification on all results. Our algorithm views the imaging task as a Bayesian inference problem of a time-varying brightness, exploits the correlation structure between time frames, and reconstructs an entire, $\mathbf{2+1+1}$ dimensional time-variable and spectrally resolved image at once. The degree of correlation in the spatial and the temporal domains is inferred from the data and no form of correlation is excluded a priori. We apply this method to the EHT observation of M87* [35] and validate our approach on synthetic data. The time- and frequency-resolved reconstruction of M87* confirms variable structures on the emission ring on a time scale of days. The reconstruction indicates extended and time-variable emission structures outside the ring itself. However, the data does not provide conclusive evidence for spectral index variations.

[a] `https://gitlab.mpcdf.mpg.de/ift/vlbi_resolve`

3.1 Main part

This section has partly been written by my coauthors. To address the imaging challenge of time-resolved VLBI data, we employ Bayesian inference. In particular, we adopt the formalism of *information field theory* (IFT) [32] for the inference of field-like quantities such as the sky brightness. IFT combines the measurement data and any included prior information into a consistent sky brightness reconstruction and propagates the remaining uncertainties into all final science results. Assuming limited spatial, frequency, and temporal variations, we can work with sparsely sampled data, such as the 2017 EHT observation of M87*.

A related method based on a Gaussian Markov model was proposed by [18] and another approach based on constraining information distances between time frames was proposed by [66]. These methods impose fixed correlations in space or time, whereas our approach adapts flexibly to the demands of the data. We also enforce strict positivity of the brightness and instead of maximizing the posterior probability, we perform a variational approximation, taking uncertainty correlations between all model parameters into account.

Interferometers sparsely probe the Fourier components of the source brightness distribution. The measured Fourier modes, called visibilities, are determined by the orientation and distance of antenna pairs, while the Earth's rotation helps to partly fill in the gaps by moving these projected baselines within the source plane. Since the source is time-variable and we aim at a time-dependent reconstruction, the measurement data have to be subdivided into multiple separate image frames along the temporal axis, leading to an extremely sparse Fourier space coverage in every frame. In the case of the EHT observation of M87*, data were taken during four 8-hour cycles spread throughout seven days. All missing image information needs to be restored by the imaging algorithm, exploiting implicit and explicit assumptions about the source structure.

Physical sources, including M87*, evolve continuously in time. Images of these sources separated by time intervals that are short compared to the evolutionary time scale are thus expected to be strongly correlated. Imposing these expected correlations during the image reconstruction process can inform image degrees of freedom (DOFs) that are not directly constrained by the data.

In radio interferometric imaging, spatial correlations can be enforced by convolving the image with a kernel, either during imaging, as part of the regularisation, or as a post-processing step. In our algorithm, we use a kernel as part of a forward model, where an initially uncorrelated image is convolved with the kernel to generate a proposal for the logarithmic sky brightness distribution, which is later adjusted to fit the data. The specific structure of such a kernel can have substantial impact on the image reconstruction. We infer this kernel in a non-parametric fashion simultaneously with the image. This substantially reduces the risk of biasing the result by choosing an inappropriate kernel, at the cost of introducing redundancies between DOFs of the convolution kernel and those of the pre-convolution image.

Metric Gaussian Variational Inference (MGVI) is a Bayesian inference algorithm that is capable of tracking uncertainty correlations between all involved DOFs, which is crucial for models with redundancies, while having memory requirements that grow only linearly with the number of DOFs [71]. It represents uncertainty correlation matrices implicitly without the need for an explicit storage of their entries and provides uncertainty quantification of the final reconstruction in terms of samples drawn from an approximate Bayesian posterior distribution, with a moderate level of approximation. Compared to methods that provide a best-fit reconstruction, our approach provides a probability distribution, capturing uncertainty.

A limitation of the Gaussian approximation is its uni-modality, as the posterior distribution is multi-modal [106]. Representing multi-modal posteriors in high dimensions is hard if not infeasible. Therefore, our results describe a typical mode of this distribution, taking the probability mass into account.

MGVI is the central inference engine of the Python package *Numerical Information Field Theory* (NIFTy [6])[1], which we use to implement our imaging algorithm, as it permits the flexible implementation of hierarchical Bayesian models. NIFTy turns a forward model into the corresponding backward inference of the model parameters by means of automatic differentiation and MGVI. For time-resolved VLBI imaging, we therefore need to define a data model that encodes all relevant physical knowledge of the measurement process and the brightness distribution of the sky.

This forward model describes in one part the sky brightness, and in another part the measurement process. For the sky brightness, we require strictly positive structures with characteristic correlations in space, time, and frequency. These brightness fluctuations can vary exponentially over linear distances and time intervals, which is represented by a log-normal prior with a Gaussian process kernel. The correlation structure of this process is assumed to be statistically homogeneous and isotropic for space, time, and frequency indi-

[1] `https://gitlab.mpcdf.mpg.de/ift/nifty`

vidually and decoupled for each sub-domain. Consequently the correlations are represented by a direct outer product of rotationally symmetric convolution kernels, or equivalently by a product of one-dimensional, isotropic power spectra in the Fourier domain. We assume the power spectra to be close to power laws with deviations modelled as an integrated Wiener processes on a double logarithmic scale [54]. The DOFs, which finally determine the spatio-temporal correlation kernel, are inferred by MGVI alongside the sky brightness distribution. While the adopted model can only describe homogeneous and isotropic correlations, this symmetry is broken for the sky image itself by the data, which in general enforce heterogeneous and anisotropic structures.

The EHT collaboration has published data averaged down to two frequency bands at 227 GHz and 229 GHz. Therefore, we employ a simplified model for the frequency axis: We reconstruct two separate, but correlated images for these bands, with a priori assumed log-normal deviation on the 1 % level, which amounts to spectral indices of ± 1 within one standard deviation. Our algorithm does not constrain the absolute flux of the two channels. Thus, we can recover the relative spectral index changes throughout the source but not the absolute ones. A detailed description of the sky model is outlined in the methods section.

We further require an accurate model of the instrument response. Just as the prior model is informed by our physical knowledge of the source, the instrument model is informed by our knowledge of the instrument. We consider two sources of measurement noise that cause the observed visibilities to differ from the perfect sky visibilities: additive Gaussian thermal noise and multiplicative, systematic measurement errors. The latter source can be conveniently eliminated by basing the model on derived quantities (closure amplitudes and phases), which are not affected by it. The magnitude of the thermal noise is provided by the EHT collaboration in the data set. Systematic measurement errors are mainly caused by antenna-based effects, e.g. differences in the measurement equipment, atmospheric phase shift, and absorption of the incoming electromagnetic waves. All those effects can be summarized in one complex, possibly time-variable, number per telescope, containing the antenna gain factors and antenna phases.

For VLBI on μas-scale, these effects can be prohibitively large. Fortunately, certain combinations of visibilities are invariant under antenna-based systematic effects, so called closure-phases and -amplitudes [98]. These quantities serve as the data for our reconstruction (for details refer to Methods section).

We apply this method to the EHT data of the super-massive black hole M87*. With a shadow of the size of approximately four light days and reported superluminal proper motions of $6c$ [13], its immediate vicinity is expected to be highly dynamic and subject to change on a time scale of days. The exceptional angular resolution of the EHT allowed for the first time to image the shadow of this super-massive black hole directly and to confirm its variability on horizon scale.

In this letter, we present a time- and frequency-resolved reconstruction of the shadow of M87* over the entire observational cycle of seven days, utilizing correlation in all four dimensions (see figure 3.1). The closure quantities do not contain information on the total flux and the absolute position of the source. Therefore, we normalize our results such that the flux in the entire ring is constant in time and agrees with the results of the EHT

collaboration for the first frame of our reconstruction. To achieve an alignment of the source even in the absence of absolute position information we start the inference with the data of only the first two observation days and continue with all data until convergence.

Figure 3.2 displays the frequency-averaged sample mean image for the first observing day together with its pixel-wise uncertainty. In full agreement with the EHT result, our image shows an emission ring that is brighter on its southern part, most likely due to relativistic beaming effects. Additionally, we obtain two faint extended structures, positioned opposite to each other along the south-western and north-eastern direction. They do not have the shape of typical VLBI-imaging artefacts, i.e. they are not faint copies of the source itself, and similar structures do not appear in any of our validation examples. We conclude that these structures are either of physical origin or due to unmodelled baseline-based calibration artefacts. The presence of additional significant source features, compared to the results by the EHT collaboration, is enabled by the usage of the data of all four observation days at once and thereby partially integrating the information.

Since our reconstruction is based on closure quantities that are not sensitive to absolute flux, the absolute spectral dependency is not constrained. Still, the relative spectral index variations w.r.t. an overall spectrum can be explored (see top row of figure 3.8). The map exhibits a higher relative spectral index in the southern portion of the ring which coincides with its brightest emission spot. However, the uncertainty map indicates that this feature is not significant and similar features falsely appear in the validation (see bottom row of figure 3.8). Therefore, we do not report any significant structures in the spectral behaviour of M87* and continue our analysis with frequency-averaged time frames.

The sky brightness for each day of the observation together with the absolute and relative differences between adjacent days is displayed in figure 3.1. We report mild temporal brightness changes of up to 6 % per day, in particular within the western and southern parts of the ring, validating the observations made by [39]. Figure 3.3 shows the detailed temporal evolution of a selected number of locations and areas. Our method consistently interpolates in between observations. In several locations our reconstruction agrees with the EHT's imaging results, whereas others clearly deviate. Especially at location 7, which corresponds to the extended structure in the south-western direction, the brightness decreases by about 5 % between adjacent days throughout the entire observation. This hints at a real and non-trivial temporal evolution.

Following the analysis of [39], we compute empirical characteristics of the asymmetric ring, i.e. diameter d, width w, orientation angle η, azimuthal brightness asymmetry A, and floor-to-ring contrast ratio f_C. All findings are summarized in table 3.1 and compared to the results of the EHT collaboration [39]: We can confirm the stationary values for diameter d, width w, azimuthal brightness asymmetry A, and floor-to-ring contrast ratio f_C during the seven days and a significant temporal evolution of the orientation angle η. The latter might be caused by flickering of emission spots [85]. We report a slightly larger diameter $d = 45(3)\,\mu$as, which does not significantly deviate from the result published by the EHT Collaboration of $d = 42(3)\,\mu$as [36].

A collection of six validation examples has been assembled to assess accuracy and robustness of our method (figures 3.4, 3.5). Figure 3.6 shows spatial correlation spectra for

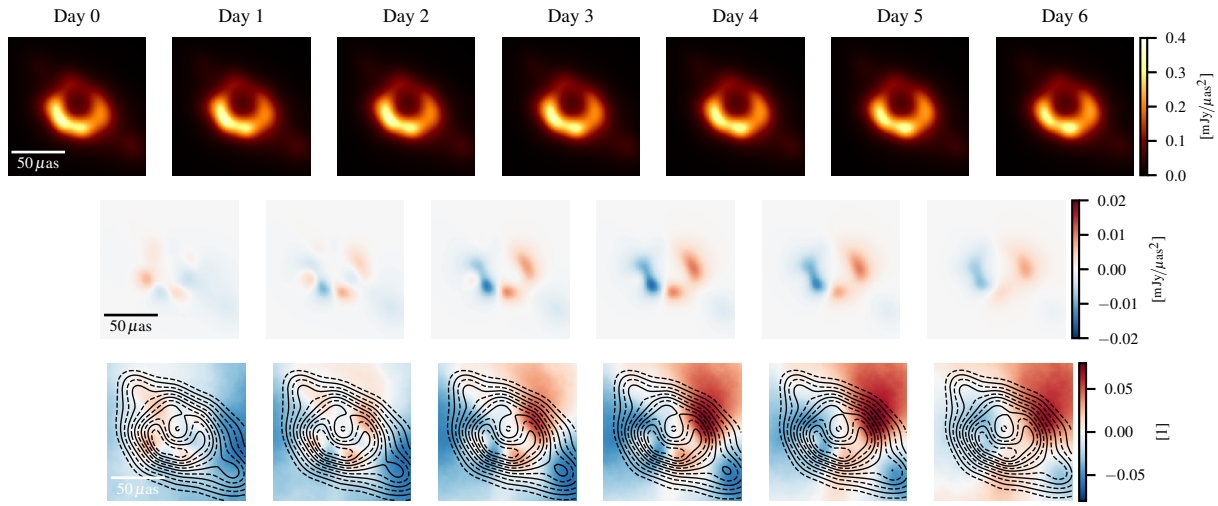

Figure 3.1: A visualisation of the approximate posterior mean. All figures are constrained to half the reconstructed field of view. The first row shows time frames of the image cube, one for each day. The second row visualises the brightness for day $N+1$ minus day N. Red and blue visualises increasing and decreasing brightness over time, respectively. The third row visualises the relative difference in brightness over time. The over-plotted contour lines show brightness in multiplicative steps of $1/\sqrt{2}$ and start at the maximum of the posterior mean of our reconstruction. The solid lines correspond to factors of powers of two from the maximum.

our scientific and validation images. Figure 3.9 displays the results of the imaging methods used by the EHT Collaboration together with our posterior mean and two samples for all observation periods.

In conclusion, we present and validate the first Bayesian imaging method that is capable of simultaneously reconstructing emission over spatial, temporal and spectral dimensions from closure quantities, utilizing correlation and quantifying uncertainties via posterior samples. We provide the first independent confirmation of the overall morphology of the emission ring around M87* and an apparent evolution of its orientation as published by the EHT collaboration. The frequency resolution allows us to obtain a relative spectral index map, together with an uncertainty estimation. For the data set at hand, significant spectral features could not be found. In addition to the emission ring, we resolve significant and potentially dynamic emission structures along the south-western and north-eastern direction. With future observations, our method may help to explore the intricate structure in the spatial, spectral, and temporal domain of M87* and other variable sources. To achieve this, the model can be extended with inference of the prior spectral correlation structure.

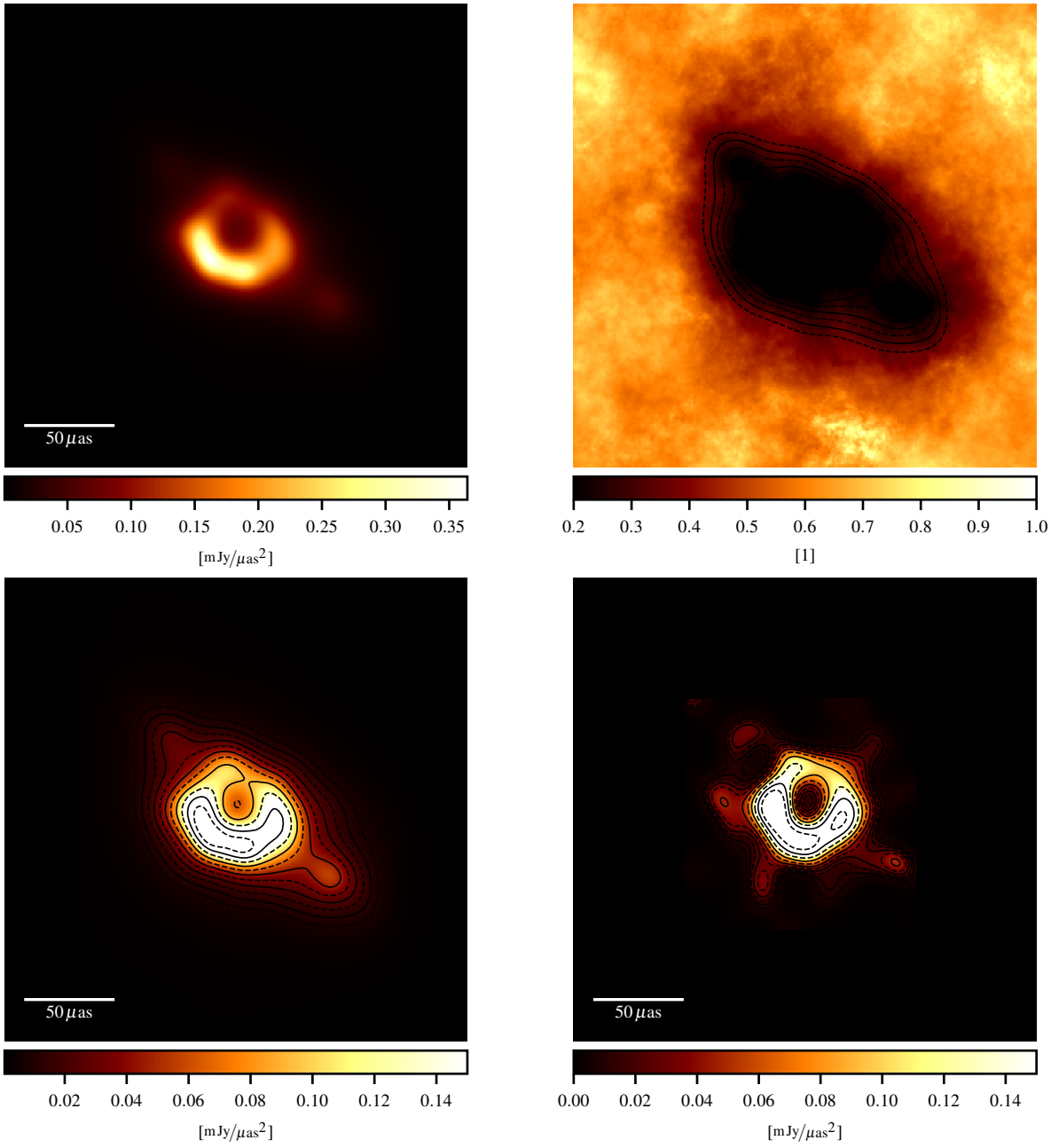

Figure 3.2: The top row shows the reconstructed mean and relative error for the first observing day. Note that the small-scale structure in regions with high uncertainty in the error map is an artefact of the limited number of samples. The bottom left shows a saturated plot of the approximate posterior mean, revealing the emission zones outside the ring. The bottom right shows the result of the EHT-imaging pipeline in comparison, saturated to the same scale and with overplotted contour lines. The over-plotted contour lines show brightness in multiplicative steps of $1/\sqrt{2}$ and start at the maximum of the posterior mean of our reconstruction. The solid lines correspond to factors of powers of two from the maximum.

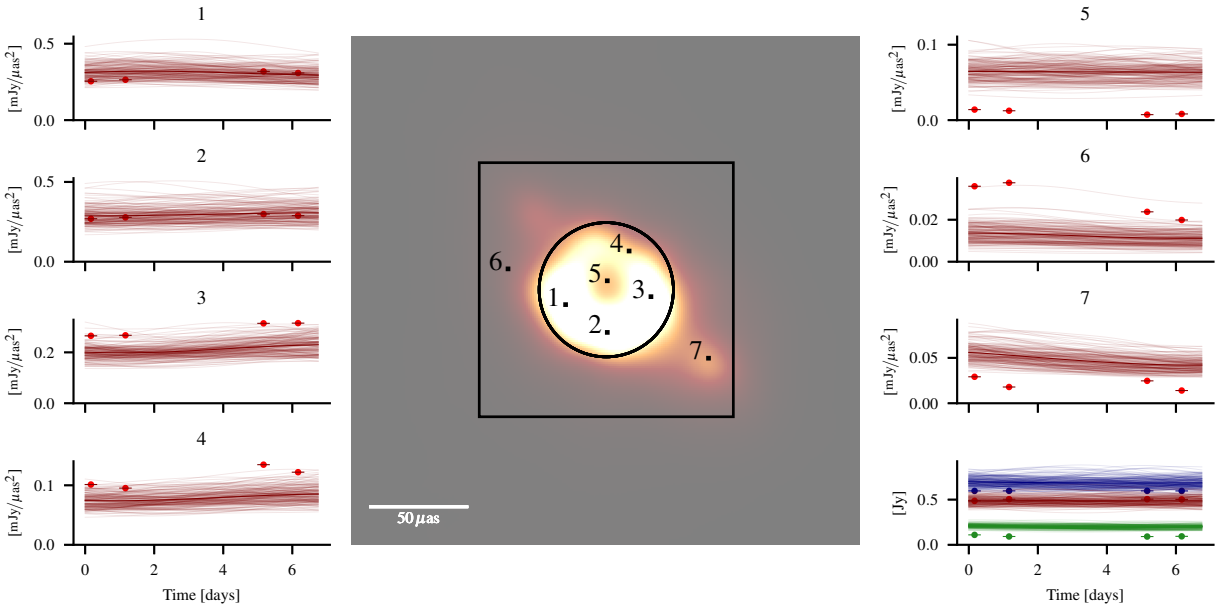

Figure 3.3: The time evolution of the brightness and flux for approximate posterior samples and their ensemble mean at specific sky locations and areas as indicated in the central panel. The peripheral panels show brightness and flux values of samples (thin lines) and their mean (thick lines). Of those, the bottom right one displays the flux inside (red) and outside the circle (green), as well as the sum of the two (blue). For comparability, only brightness within the field of view of the EHT collaboration image, indicated by the black box in the central plot, is integrated. The remaining panels give the local brightness for the different locations labelled by numbers in the central panel. The single-day results from EHT-imaging are indicated as points.

	$d\,(\mu\text{as})$	$w\,(\mu\text{as})$	$\eta\,(^\circ)$	A	f_C
DIFMAP					
April 5	37.2 ± 2.4	28.2 ± 2.9	163.8 ± 6.5	0.21 ± 0.03	0.5
April 6	40.1 ± 7.4	28.6 ± 3.0	162.1 ± 9.7	0.24 ± 0.08	0.4
April 10	40.2 ± 1.7	27.5 ± 3.1	175.8 ± 9.8	0.20 ± 0.04	0.4
April 11	40.7 ± 2.6	29.0 ± 3.0	173.3 ± 4.8	0.23 ± 0.04	0.5
EHT-IMAGING					
April 5	39.3 ± 1.6	16.2 ± 2.0	148.3 ± 4.8	0.25 ± 0.02	0.08
April 6	39.6 ± 1.8	16.2 ± 1.7	151.1 ± 8.6	0.25 ± 0.02	0.06
April 10	40.7 ± 1.6	15.7 ± 2.0	171.2 ± 6.9	0.23 ± 0.03	0.04
April 11	41.0 ± 1.4	15.5 ± 1.8	168.0 ± 6.9	0.20 ± 0.02	0.04
SMILI					
April 5	40.5 ± 1.9	16.1 ± 2.1	154.2 ± 7.1	0.27 ± 0.03	7×10^{-5}
April 6	40.9 ± 2.4	16.1 ± 2.1	151.7 ± 8.2	0.25 ± 0.02	2×10^{-4}
April 10	42.0 ± 1.8	15.7 ± 2.4	170.6 ± 5.5	0.21 ± 0.03	4×10^{-6}
April 11	42.3 ± 1.6	15.6 ± 2.2	167.6 ± 2.8	0.22 ± 0.03	6×10^{-6}
OUR METHOD (UNCERTAINTY AS PER [39, TABLE 7])					
April 5	44.4 ± 3.4	23.2 ± 5.2	164.9 ± 9.5	0.26 ± 0.04	0.365
April 6	44.4 ± 2.9	23.3 ± 5.4	161.7 ± 5.6	0.24 ± 0.04	0.374
April 10	44.8 ± 2.8	23.0 ± 5.0	176.7 ± 9.8	0.22 ± 0.03	0.374
April 11	44.6 ± 2.8	22.8 ± 4.8	180.1 ± 10.4	0.22 ± 0.03	0.372
OUR METHOD (SAMPLE UNCERTAINTY)					
April 5	44.1 ± 1.2	23.1 ± 2.4	163.9 ± 5.0	0.25 ± 0.03	0.377 ± 0.081
April 6	44.0 ± 1.2	22.9 ± 2.4	161.9 ± 6.0	0.24 ± 0.03	0.385 ± 0.085
April 10	44.6 ± 1.2	22.9 ± 2.5	176.2 ± 6.5	0.22 ± 0.03	0.383 ± 0.089
April 11	44.6 ± 1.2	23.0 ± 2.6	179.8 ± 6.2	0.22 ± 0.03	0.383 ± 0.090

Table 3.1: A comparison of diameter d, width w, orientation angle η, asymmetry A and floor-to-ring contrast ratio f_C as defined by [39, Table 7] and computed for images published by the EHT collaboration (first three sections of table) as well as for our reconstruction (last two sections). Section four provides the result of the estimators and their standard deviations as defined by [39] applied to our posterior mean. Section five provides means and standard deviations based on processing our posterior samples individually through the estimators and by computing mean and standard deviations from these results.

3.2 Methods

This section has partly been written by my coauthors. The reconstruction algorithm relies on Bayesian statistics. Thus, it consists out of two essential components: the likelihood and the prior.

The likelihood is a probabilistic description of the measurement process including details on the measurement device. We choose to describe the measurement in terms of closure quantities that are invariant under antenna-based calibration effects.

The prior model captures all assumptions on the sky brightness distribution. Here we assume positivity at all times, correlation along the temporal, spatial, and spectral direction, as well as the possibility of variations on exponential scale. This is implemented with the help of a Gaussian process prior of the logarithmic brightness distribution with unknown kernel. Below, a non-parametric kernel model is derived that assumes a stochastic process along each dimension individually.

3.2.1 Likelihood

This section has partly been written by Philipp Arras and Reimar Leike. The likelihood of the measured visibilities given the sky brightness distribution s is computed independently for each time frame. The visibilities for all measured data points are assumed to follow the measurement equation in the flat sky approximation:

$$R(s)_{AB} \coloneqq \int e^{-2\pi i(u_{AB}x+v_{AB}y)} s(x,y)\,dx\,dy \tag{3.1}$$

$$=: e^{\rho_{AB}} e^{i\phi_{AB}}. \tag{3.2}$$

Here AB runs through all ordered pairs of antennas A and B for all non-flagged baselines, u_{AB} and v_{AB} are the coordinates of the measured Fourier points, $s(x,y)$ is the sky brightness distribution as a function of sky angles x and y, and R is called measurement response. The visibilities $R(s)_{AB}$ are complex numbers and we represent them in polar coordinates as phases $\phi_{AB}(s) \in \mathbb{R}$ and logarithmic amplitudes $\rho_{AB}(s) \in \mathbb{R}$, i.e. $R(s)_{AB} = \exp(\rho_{AB}(s) + i\,\phi_{AB}(s))$. We assume the thermal noise of the phase and logarithmic amplitude to be independently Gaussian distributed with covariance

$$N = \operatorname{diag}\left(\frac{\sigma^2}{|d|^2}\right) \ , \tag{3.3}$$

where d is the reported visibility data and σ is the reported thermal noise level. The operation $\operatorname{diag}(x)$ denotes a diagonal matrix with x on its diagonal. This is approximately valid for a signal-to-noise ratio larger than 5 [14], which is true for most of our data.

To avoid antenna based systematic effects, we compute closure quantities from these visibilities [14]. Closure phases are obtained by combining a triplet of complex phases of visibilities via:

$$(\phi_{\mathrm{cl}})_{ABC} \coloneqq \phi_{AB} + \phi_{BC} + \phi_{CA}. \tag{3.4}$$

Closure amplitudes are formed by combining the logarithmic absolute value of four visibilities:

$$(\rho_{\mathrm{cl}})_{ABCD} \coloneqq \rho_{AB} - \rho_{BC} + \rho_{CD} - \rho_{DA}. \tag{3.5}$$

These closure quantities are invariant under antenna based visibility transformations of the form

$$R(s)_{AB} \to c_A c_B^* R(s)_{AB} \tag{3.6}$$

for all antennas and multiplicative calibration errors c_A and c_B, where $*$ denotes the complex conjugate.

Note that forming the closure phases is a linear operation on the complex phase, while forming the closure amplitudes is linear in the logarithmic absolute value. We can thus represent these operations using matrices:

$$\rho_{\mathrm{cl}} = L\rho, \quad \phi_{\mathrm{cl}} = M\phi. \tag{3.7}$$

The closure matrices L and M are sparse and contain in every row ± 1 for visibilities associated with the closure, and zero elsewhere.

The noise covariances N_ρ and N_ϕ of the closure quantities are related to N via:

$$N_\rho = \left\langle Ln(Ln)^\dagger \right\rangle_{\mathcal{N}(n|0,N)} = LNL^\dagger \quad \text{and} \quad N_\phi = MNM^\dagger, \tag{3.8}$$

where $\mathcal{N}(n|0,N)$ denotes a Gaussian distribution over n with mean 0 and covariance N. The mixing introduced by applying L and M leads to non-diagonal noise covariance matrices of the closure quantities.

For a given antenna setup (of five or more antennas), more closure quantities can be constructed than visibilities are available, and therefore they provide a redundant description of the data. For the logarithmic amplitudes ρ, we first construct all possible closure quantities and then map to a non-redundant set using the eigen-decomposition of N_ρ. Specifically, we construct a unitary transformation U_ρ where each column of the matrix is an eigenvector corresponding to a non-zero eigenvalue of N_ρ. This transformation provides a map from the space of all possible closure amplitudes to the space of maximal non-redundant sets, with the additional property that the transformed noise covariance becomes diagonal. Specifically

$$U_\rho N_\rho U_\rho^\dagger = \Lambda_\rho \ , \tag{3.9}$$

where Λ_ρ denotes a diagonal matrix with the non-zero eigenvalues of N_ρ on its diagonal. We can combine L and U_ρ to form an operation that maps from the logarithmic amplitudes of visibilities ρ directly to the space of non-redundant closure amplitudes ϱ via

$$\varrho = U_\rho \rho_{\mathrm{cl}} = U_\rho L \rho \ , \tag{3.10}$$

and use it to compute the observed, non-redundant closure amplitude ϱ_d from the published visibility data $d = \exp(\rho_d + i\,\phi_d)$.

The resulting likelihood for closure amplitudes reads

$$\mathcal{P}(\varrho_d|\varrho, L, N) = \mathcal{N}(\varrho_d|\varrho, \Lambda_\rho) \ . \tag{3.11}$$

Closure phases are constructed differently to avoid problems induced by phase wraps. Adding or subtracting 2π from a phase does not change the result, and we need to preserve this symmetry in our algorithm. We thus can only add integer multiples of phases such as (3.4) and this prohibits using a direct matrix decomposition to find a maximal non-redundant closure set.

We build the closure sets to be used in the imaging with the help of a greedy algorithm that processes closure phases in the order of decreasing signal-to-noise ratio, as defined by the inverse of the diagonal of N_ϕ (3.8). The algorithm collects closure sets into M until $\mathrm{rank}(M) = \dim(\phi)$ ensuring that ϕ_{cl} consists of a maximal non-redundant set. In principle, all maximal non-redundant closure sets are equivalent as long as one takes the non-diagonal noise covariance into account. The concrete choice might have a minor impact for our approximation of the closure phase likelihood.

Within our closure set, we can decompose the noise covariance N_ϕ into a unitary matrix U_ϕ and its eigenvalues Λ_ϕ. Instead of working with the phases ϕ_{cl} directly, we use their positions on the complex unit circle $e^{i\phi_{\mathrm{cl}}}$ to define

$$\varphi = U_\phi \ e^{i\phi_{\mathrm{cl}}} = U_\phi \ e^{iM\phi} \ . \tag{3.12}$$

This mitigates the problem of phase wraps at the price of approximating the corresponding covariance. This approximation yields errors below the 1 % level if the signal-to-noise ratio is larger than 10. Most of the data points are above that threshold, and the error decreases quadratically with increasing signal-to-noise ratio. Since data with the lowest standard deviation are also the most informative, we believe the impact of the approximation on the reconstruction to be negligible.

Given the closure phases on the unit circle φ, the corresponding phase likelihood can be written as

$$\mathcal{P}(\varphi_d|\varphi, L, N) = \mathcal{N}(\varphi_d|\varphi, \Lambda_\phi) \ , \tag{3.13}$$

where $\varphi_d = U_\phi \ e^{iM\phi_d}$. Note that (3.13) is a Gaussian distribution on complex numbers with the probability density function as

$$\mathcal{N}(x|y, X) = |4\pi X|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(x-y)^\dagger X^{-1}(x-y)\right) \ , \tag{3.14}$$

and Hermitian covariance X. Complex and real Gaussian distributions only differ in their normalization constant. We do not distinguish between them explicitly, as the normalization is irrelevant for our variational approach.

3.2.2 Modelling the sky brightness

This section has partly been written by Philipp Arras. The sky brightness distribution $s_{xt\nu}$ is defined within a fixed field of view $\Omega_x \subset \mathbb{R}^2$, a time interval $\Omega_t = [0, \bar{t}]$, and frequency

range $\Omega_\nu \subset \mathbb{R}$, which renders it to be a field defined in space, time, and frequency. We assume s to be a priori log-normal distributed:

$$s_{xt\nu} \coloneqq e^{\tau_{xt\nu}} \text{ with } x \in \Omega_x,\ t \in \Omega_t, \text{ and } \nu \in \Omega_\nu \text{ with } \mathcal{P}(\tau|T) \coloneqq \mathcal{N}(\tau|0,T). \tag{3.15}$$

The a priori correlation structure of the logarithmic sky brightness τ is encoded within the covariance T. Choosing a log-normal model allows the sky brightness to vary exponentially on linear spatial, temporal, and frequency scales and ensures the positivity of the reconstructed intensity, similarly to [22, 21].

We perform a basis transformation to a standardised Gaussian distribution $\mathcal{P}(\xi_s) = \mathcal{N}(\xi_s|0,\mathbb{1})$, which allows us to separate the correlation structure from its realization [70]. The new coordinates ξ_s have the same dimension as the original parameters, but are a priori independent:

$$s = e^{A\xi_s} \quad \text{with} \quad AA^\dagger \coloneqq T. \tag{3.16}$$

This defines a generative model which turns standard normal distributed DOFs ξ_s into random variables s that are distributed according to (3.15). Although the information encoded in a distribution is invariant under coordinate transformations, MGVI depends on the choice of coordinates. Therefore, reformulating the entire inference problem in terms of standardised generative models is important to ensure that the prior information is fully captured by an approximation via MGVI. We visualize our generative model in figure (3.7).

3.2.3 Correlations in space, time, and frequency

We do not know the correlation structure of the logarithmic sky brightness a priori, so we include it as part of the model, which has to be inferred from the data. The different dimensions of the sky brightness are governed by completely distinct physical phenomena, which should be reflected in the model.

Setting up such correlations involves a number of intricate technicalities. The main idea is to model the correlations in space, time, and frequency independently using the same underlying model and combine them via outer products. Doing this naively results in degenerate and highly un-intuitive model parameters. The model we introduce in the following avoids these issues, but unfortunately requires a certain complexity.

For now we consider the correlation structure along the different sub-domains individually. A priori we do not want to single out any specific location or direction for the logarithmic sky brightness, which corresponds to statistical homogeneity and isotropy. According to the Wiener-Khinchin theorem, such correlation structures $T^{(i)}$ with $i \in \{\Omega_x, \Omega_t, \Omega_\nu\}$ are diagonal in the Fourier domain and can be expressed in terms of a power spectrum $p_{T^{(i)}}(|k|)$:

$$T^{(i)}_{kk'} = \left(F^{(i)} T^{(i)} \left(F^{(i)}\right)^\dagger\right)_{kk'} = (2\pi)^{D^{(i)}}\, \delta\left(k-k'\right)\ p_{T^{(i)}}(|k|), \tag{3.17}$$

where $F^{(i)}$ and k denote the Fourier transformation and Fourier coordinates associated to the space i, $D^{(i)}$ is the dimension of i, δ denotes the Kronecker delta, and $|k|$ is the

Euclidean norm of the vector k. Here † denotes the adjoint of the operator. We choose our Fourier convention such that no factors of 2π enter the transformation $F^{(i)}$, and thus its inverse has a factor of $1/(2\pi)^{D^{(i)}}$. As we build the model in terms of standardised coordinates ξ_s, we work with the square root of the correlation matrix

$$A^{(i)}_{kk'} = (2\pi)^{D^{(i)}}\,\delta\,(k-k')\;\sqrt{p_{T^{(i)}}(|k|)} =: (2\pi)^{D}\,\delta\,(k-k')\,p^{(i)}(|k|) \tag{3.18}$$

that converts those into the logarithmic brightness $\tau = A\,\xi_s$.

The amplitude spectrum $p^{(i)}(|k|)$ depends on the characteristic length scales of the underlying physical processes, which we do not know precisely. Our next task is to develop a flexible model for this spectrum that expresses our uncertainty and is compatible with a wide range of possible systems. We model the amplitude spectrum in terms of its logarithm:

$$p^{(i)}(|k|) \propto e^{\gamma^{(i)}(|k|)}. \tag{3.19}$$

We do not want to impose any functional basis for this logarithmic amplitude spectrum $\gamma^{(i)}(|k|)$, so we describe it non-parametrically using an integrated Wiener process in logarithmic $l = \log|k|$ coordinates. This corresponds to a smooth, i.e. differentiable, function, with exponential scale dependence [33]. In the logarithmic coordinates l, the zero-mode $|k| = 0$ is infinitely far away from all other modes. Later on we deal with it separately and continue with all remaining modes for now.

The integrated Wiener process in logarithmic coordinates $\gamma^i(l)$ reads:

$$\gamma^{(i)}(l) = m^{(i)} l + \eta^{(i)} \int_{l_0}^{l}\int_{l_0}^{l'} \xi^{(i)}_W(l'')\,dl'\,dl'', \tag{3.20}$$

where l_0 is the logarithm of the first mode greater than zero. Without loss of generality, we set the initial offset to zero. Later on we explicitly parameterise it in terms of a more intuitive quantity. The parameter $m^{(i)}$ is the slope of the amplitude on double-logarithmic scale. It is a highly influential quantity, as it controls the overall smoothness of the logarithmic sky brightness distribution. Specifically, after exponentiation, the spectrum is given as a power law with multiplicative deviations, and the exponent of this power law is given by the slope. Therefore, a spectrum with slope zero indicates the absence of any spatial correlation in the image, whereas a slope of -1 indicates continuous, and -2 differentiable brightness distributions along the respective axis [87]. The parameter $\eta^{(i)}$ describes how much the amplitude spectrum deviates from the power law. These deviations follow the smooth integrated Wiener process and can capture characteristic length scales of the logarithmic brightness distribution. Their precise shape is encoded in the realization $\xi^{(i)}_W \sim \mathcal{N}(\xi^{(i)}_W|0,\mathbb{1})$, which are also parameters of our model and follow a priori the standard Gaussian distribution. We do not want to fix the slope and deviations and therefore impose Gaussian and log-normal priors for $j \in \{m,\eta\}$ respectively, with preference for a certain value $\mu^{(i)}_j$ and expected deviations $\sigma^{(i)}_j$ thereof:

$$m^{(i)} = \mu^{(i)}_m + \sigma^{(i)}_m \xi^{(i)}_m, \quad \eta^{(i)} = e^{\mu^{(i)}_\eta + \sigma^{(i)}_\eta \xi^{(i)}_\eta} \quad \text{with} \quad \xi^{(i)}_j \sim \mathcal{N}(\xi^{(i)}_j|0,\mathbb{1}). \tag{3.21}$$

The amplitude spectrum defines the expected variation $\widetilde{U}^{(i)}$ of the log-brightness around its offset via

$$\widetilde{U}^{(i)} := \int_{k\neq 0} p_{T^{(i)}}(|k|)\, dk = \int_{k\neq 0} e^{2\,\gamma^{(i)}(|k|)}\, dk. \tag{3.22}$$

The relation between $\gamma^{(i)}$ and $\widetilde{U}^{(i)}$ is un-intuitive, but it is critical to constrain the expected variation to reasonable values as it has a severe impact on a priori plausible brightness distributions. Therefore we replace the variance amplitude (i.e. the square root of $\widetilde{U}^{(i)}$) with a new parameter $a^{(i)}$:

$$p^{(i)}(|k|) = a^{(i)}\, \frac{e^{\gamma^{(i)}(|k|)}}{\sqrt{\widetilde{U}^{(i)}}}, \quad \forall k \neq 0. \tag{3.23}$$

Note that this step implicitly determines the offset of the Wiener processes in terms of $a^{(i)}$. We elevate $a^{(i)}$ to be a free model parameter and impose a log-normal model analogous to $\eta^{(i)}$ with hyperparameters $\mu_a^{(i)}$ and $\sigma_a^{(i)}$.

Next, we combine correlation structures in independent sub-domains. For every one of those, i.e. in our case space, time, and frequency, we use an instance of the model described above. We have not yet specified how to deal with the amplitude of the zero-modes $p^{(i)}(0)$, and their treatment emerges from the combination of the sub-domains. The overall correlation structure including all sub-domains is given by the outer product of the sub-spaces:

$$A = \bigotimes_{i\in\{x,t,\nu\}} A^{(i)}. \tag{3.24}$$

This product introduces a degeneracy: $\alpha(A^{(i)} \otimes A^{(j)}) = (\alpha A^{(i)}) \otimes A^{(j)} = A^{(i)} \otimes (\alpha A^{(j)})$ for all $\alpha \in \mathbb{R}^+$. With every additional sub-domain we add one additional degenerate degree of freedom. We can use this freedom to constrain the zero-mode of the amplitude spectrum, and thus remove the degeneracy up to a global factor. For this we normalize the amplitudes in real-space:

$$\widetilde{A}^{(i)} := \left(\frac{1}{V^{(i)}} \int_{\Omega^{(i)}} \left(F^{(i)}\right)^{-1} p^{(i)}\, \mathrm{d}\Omega^{(i)}\right)^{-1} A^{(i)} = \frac{V^{(i)}}{p^{(i)}(0)} A^{(i)}. \tag{3.25}$$

The zero-mode of the normalised amplitude $\widetilde{A}^{(i)}$ can be fixed to the total volume $V^{(i)}$ of the space $\Omega^{(i)}$. Consequently, the overall correlation structure is expressed as

$$A = \alpha \bigotimes_{i\in\{x,t,\nu\}} \widetilde{A}^{(i)}. \tag{3.26}$$

The remaining multiplicative factor α globally sets the scale in all sub-domains and has to be inferred from the data. Additionally, we put a log-normal prior with logarithmic mean μ_α and standard deviation σ_α hyperparameters and a corresponding standard Gaussian parameter ξ_α on this quantity.

This was the last ingredient for the correlation structure along multiple independent sub-domains and serves as a generative prior to infer the correlation structure in a space-time-frequency imaging problem. For the specific application to the EHT observations, however, only data averaged down to two narrow frequency channels is available. Therefore, as we do not expect to be able to infer a sensible frequency correlation structure using only two channels, we simplify (3.26) to explicitly parameterize the frequency correlations as

$$A = \begin{pmatrix} 1 & \epsilon \\ 1 & -\epsilon \end{pmatrix} \left(\alpha \bigotimes_{i \in \{x,t\}} \widetilde{A}^{(i)} \right) , \tag{3.27}$$

where ϵ is a hyperparameter that steers the a priori correlation between the frequency channels.

We briefly summarise all the required hyperparameters and how the generative model for the correlation structure is built. We start with the correlations in the individual sub-domains which we describe in terms of their amplitude spectra $A^{(i)}(\xi^{(i)})$. Four distinct standardised model parameters are inferred from the data, $\xi^{(i)} \coloneqq (\xi^{(i)}_m, \xi^{(i)}_\eta, \xi^{(i)}_W, \xi^{(i)}_a)$. The first describes the slope of the linear contribution to the integrated Wiener process. The second is related to the strength of the smooth deviations from this linear part. The third parameter describes the actual form of these deviations. Finally, the last one describes the real-space fluctuations of the associated field.

The hyperparameters are μ^i_j and σ^i_j for $j \in \{m, \eta, a\}$ specifying the expected mean and standard deviation of the slope $m^{(i)}$ and expected mean and standard deviation for $\ln(\eta), \ln(a)$, which are therefore enforced to be positive. In addition to these, we have to determine the global scale parameter $\alpha(\xi_\alpha)$, for which we also specify the logarithmic mean μ_α and standard deviation σ_α. We determine the values for the hyperparameters of the logarithmic quantities through an additional moment matching step by explicitly specifying the mean and standard deviation of the log-normal distribution. The generative model for the correlation structure is therefore:

$$A(\xi_A) = \begin{pmatrix} 1 & \epsilon \\ 1 & -\epsilon \end{pmatrix} \left(\alpha(\xi_\alpha) \bigotimes_{i \in \{x,t\}} \widetilde{A}^{(i)}(\xi^{(i)}) \right) \quad \text{with} \quad \xi_A = \left(\xi_\alpha, \xi^{(x)}, \xi^{(t)} \right) . \tag{3.28}$$

Combining this with the generative model for the sky brightness itself we end up with the full model:

$$s(\xi) = e^{F^{-1}(A(\xi_A)\ \xi_s)} \quad \text{with} \quad F^{-1} = \bigotimes_{i \in \{x,t\}} \left(F^{(i)} \right)^{-1} . \tag{3.29}$$

Our model is now standardized and all its parameters $\xi = (\xi_A, \xi_s)$ follow a multivariate standard Gaussian distribution. The Bayesian inference problem is fully characterised by the negative logarithm (or information) of the joint probability distribution of data and parameters. Combining the closure likelihoods with the described sky brightness model

therefore yields:

$$\begin{aligned}-\log\Big(\mathcal{P}(\varrho_d,\varphi_d,\xi)\Big) =& \frac{1}{2}\Big(\varrho_d-\varrho(s(\xi))\Big)^\dagger\Lambda_\rho^{-1}\Big(\varrho_d-\varrho(s(\xi))\Big)\\ &+\frac{1}{2}\Big(\varphi_d-\varphi(s(\xi))\Big)^\dagger\Lambda_\phi^{-1}\Big(\varphi_d-\varphi(s(\xi))\Big)\\ &+\frac{1}{2}\xi^\dagger\xi+H_0,\end{aligned} \tag{3.30}$$

where H_0 is a constant that is independent of the latent variables ξ.

Implementation Details *This section has partly been written by my coauthors.* We implement the generative model in NIFTy [6], which also provides an implementation of MGVI utilising auto-differentiation. We represent the spatial domain with 256×256 pixels, each with a length of $1\,\mu$as. In the time domain we choose a resolution of 6 hours for the entire observation period of 7 days, thus obtaining 28 time frames. The implementation of the generative model utilizes the Fast Fourier Transform and thus defines the resulting signal on a periodic domain. To avoid artefacts in the time domain, we add another 28 frames to the end of the observed interval, resulting in a temporal domain twice that size.

For the frequency domain, only two channels are available, and we do not expect them to differ much from each other. Instead of inferring the correlation along this direction, as we do for the spatial and temporal axis, we assume a correlation between the two channels on the 99 % level a priori, i.e. we set $\epsilon = 0.01$.

This adds another factor of 2 of required pixels to the reconstruction. For future reconstructions with deeper frequency sampling we can extend the model and treat this domain equivalently to the space and time domains. Overall we have to constrain $256\times256\times56\times2+$ power spectrum DOFs $\approx 7.4\times10^6$ pixel values with the data.

The Gaussian approximation to the closure likelihoods is only valid in high signal-to-noise regimes [14]. We increase the signal-to-noise ratio by means of an averaging procedure, which subdivides each individual scan into equally sized bins with a length of approximately 2 min. To validate that this averaging is justified we compare the empirical standard deviation of averaged data values with the corresponding thermal noise standard deviation and find their ratio to be 1.48 on average, consistent with the expected $\sqrt{2}$ for complex valued data.

The intra-site baselines of ALMA–APEX and SMT–JCMT probe the sky at scales larger than our field of view. To avoid contamination from external sources, we flag these intra-site baselines and exclude closure quantities that involve the respective pair.

Acknowledgements

We thank Landman Bester and Iniyan Natarajan for discussions regarding VLBI imaging, the Schneefernerhaus for their hospitality, and the four anonymous referees for numerous comments that significantly improved the manuscript. Philipp Arras acknowledges the

financial support by the German Federal Ministry of Education and Research (BMBF) under grant 05A17PB1 (Verbundprojekt D-MeerKAT). Jakob Knollmüller acknowledges the financial support by the Excellence Cluster ORIGINS, which is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy - EXC-2094-390783311.

Data Availability

The data this work is based on have been published by the Event Horizon Collaboration [38, 39] and is available at [35]. We provide a set of 160 antithetic sample pairs of the sky brightness from the approximate posterior distribution, which can be used to propagate uncertainty to any derived quantity. The samples are available at [1].

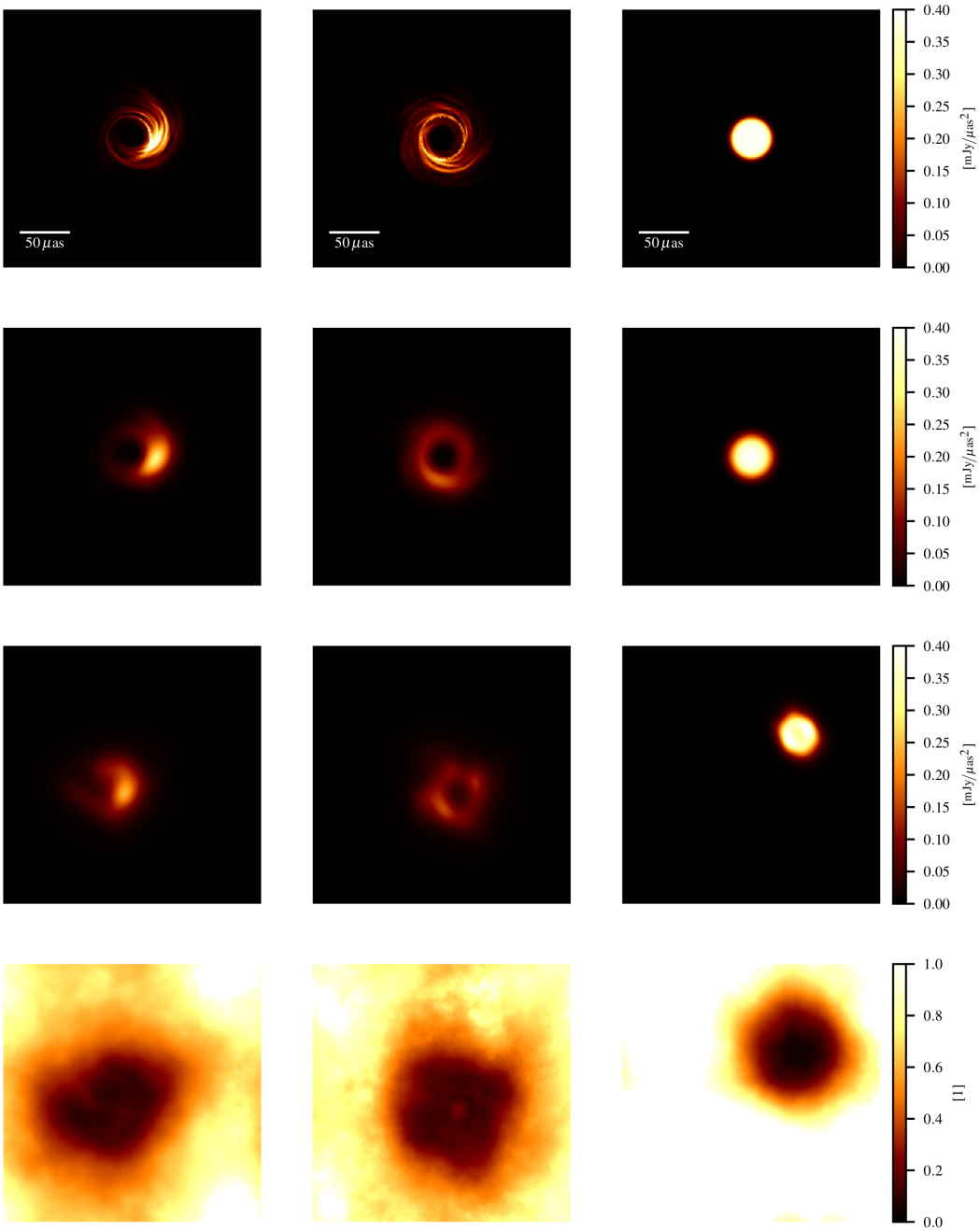

Figure 3.4: Our validation for static sources, showing two scenarios from the EHT imaging challenge and a uniform disk. The rows depict the ground truth, the smoothed ground truth, the approximate posterior mean, and the relative standard deviation for our three static validation examples. The plots in the first three rows are normalized to their respective maximum, are not clipped, and the minimum of the colour bar is zero. In the last row the colour bar is clipped to the interval $[0, 1]$.

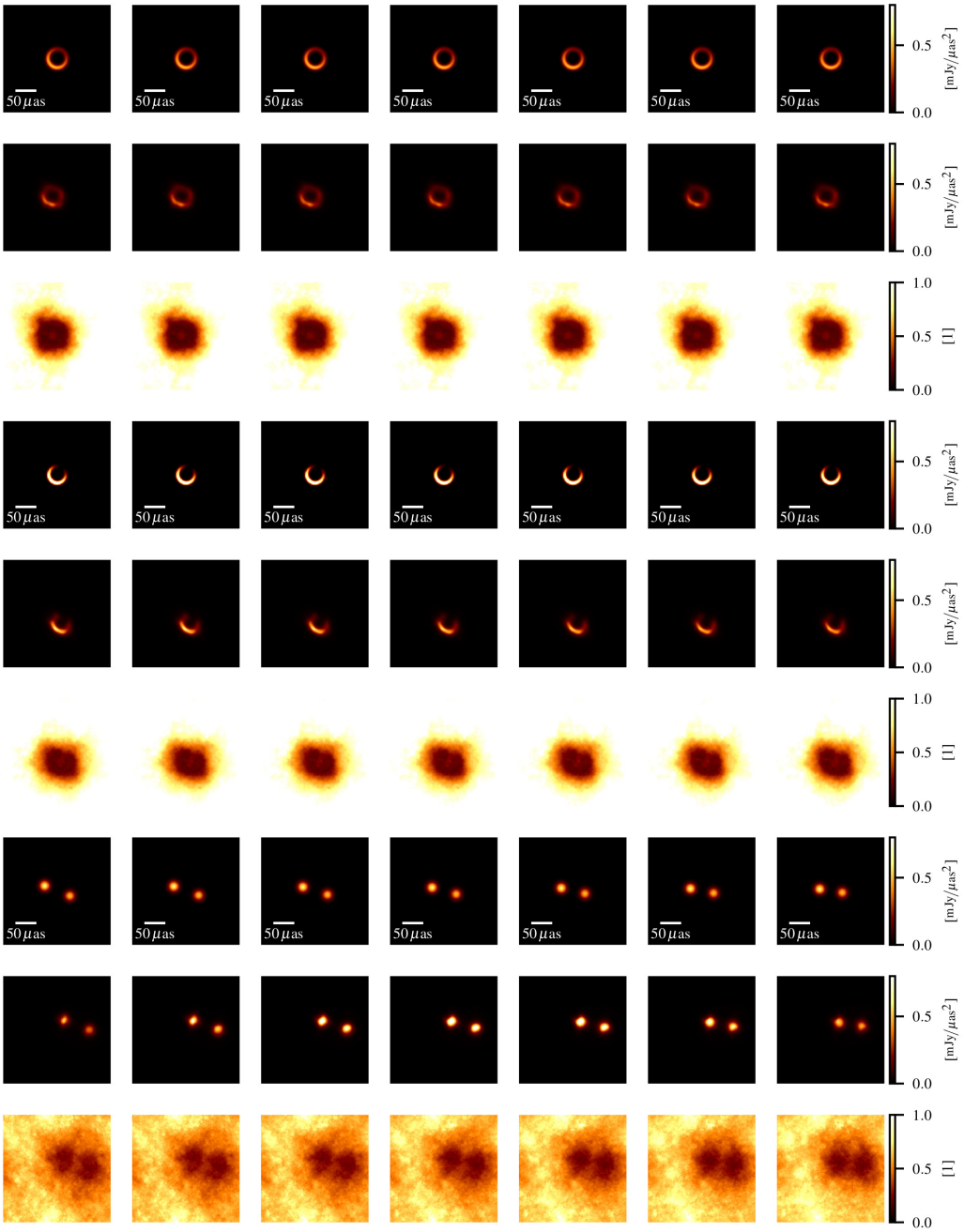

Figure 3.5: Our validation on synthetic observations of time-variable sources. In the figure, time goes from left to right showing slices through the image cube for the first time bin of each day. Different source models are shown from top to bottom: `eht-crescent`, `slim-crescent`, and `double-sources`. For each source the ground truth, the approximate posterior mean of the reconstruction, and the relative posterior standard deviation, clipped to the interval $[0, 1]$, are displayed (from top to bottom). The central three columns show moments in time in which no data is available since data was taken only during the first and last two days of the week-long observation period.

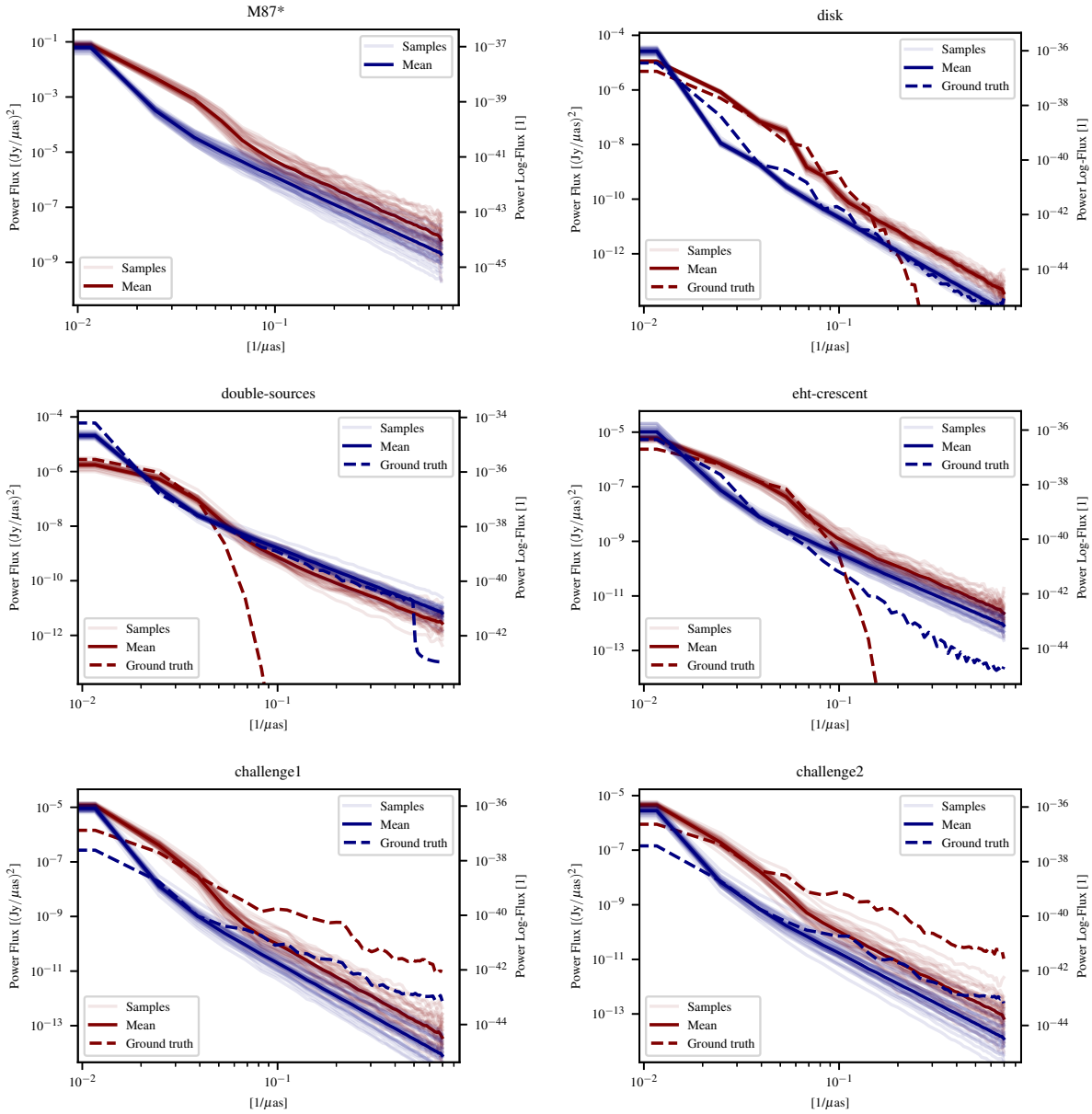

Figure 3.6: The spatial correlation power spectra of our reconstruction for the EHT-observation of M87* (top left panel) and five of our validation data sets. The red curves show the power spectra of the reconstructed brightness. The blue curves show the power spectra of the logarithmic brightness. For the three validation sets, the corresponding power spectra of the ground truth are plotted as a dashed line.

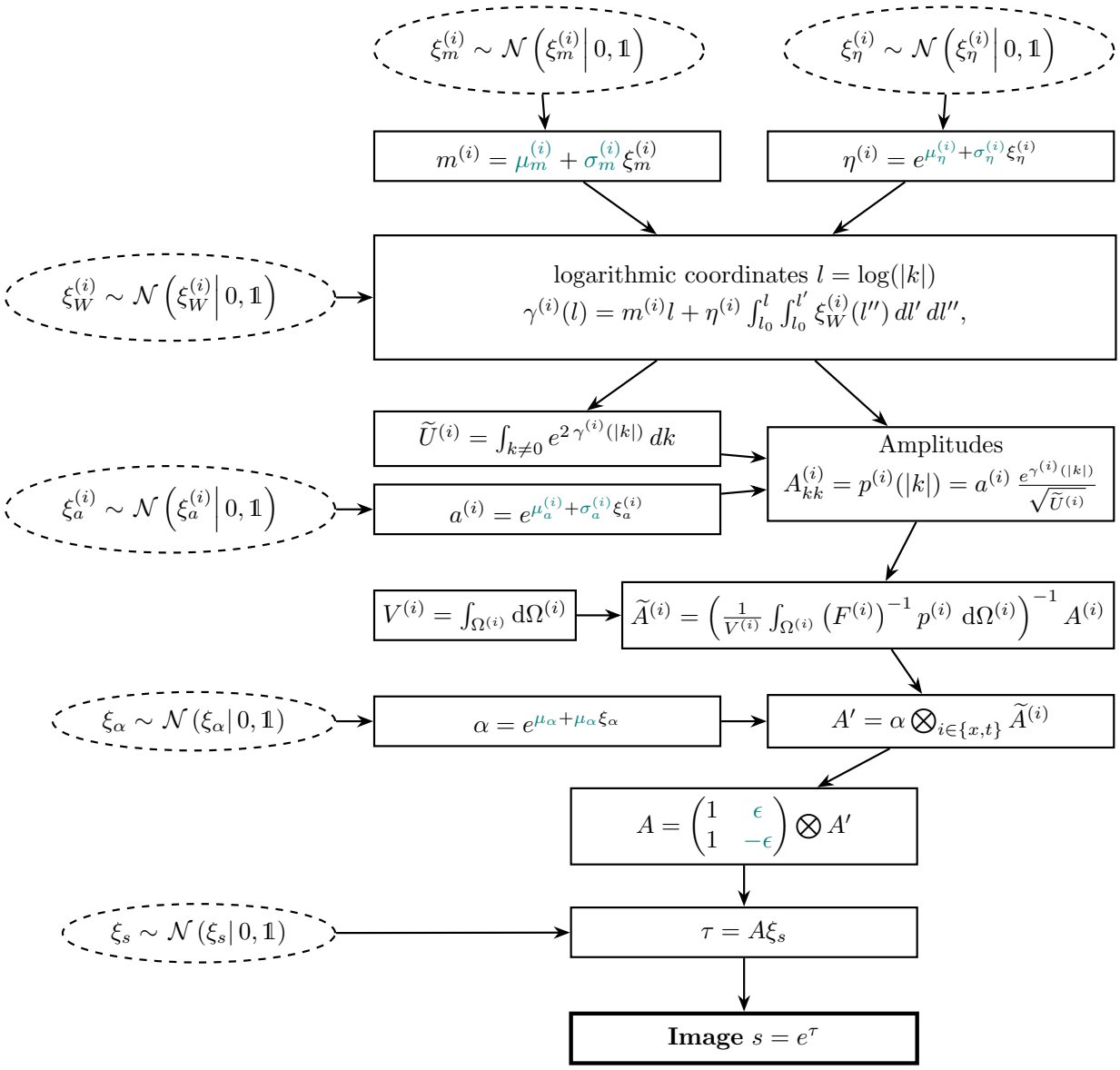

Figure 3.7: A visualization of the hierarchical model that was used as prior on the four-dimensional (frequency, time and space) image s, as described in the methods section. The round dashed nodes represent the inferred latent parameters, which are independent normal distributed a priori. The solid rectangular nodes represent computation steps. Arrows denote dependencies. All hyperparameters are marked in teal. The upper half of the diagram describes our non-parametric model of the power spectra in temporal and spatial domains. The lower half specifies how the four dimensional image is obtained from additional latent parameters and the power spectra.

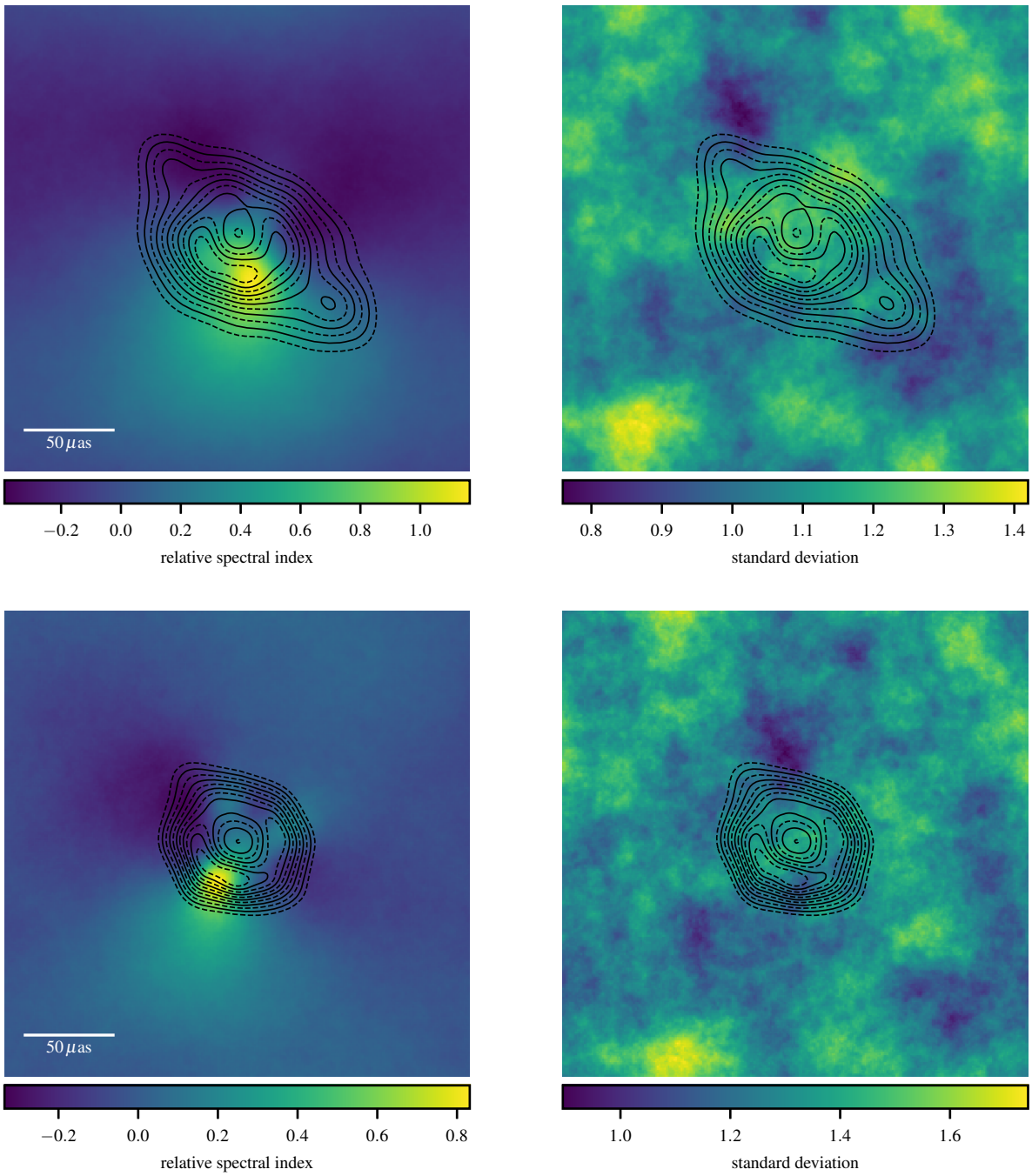

Figure 3.8: The relative spectral index and the pixel-wise uncertainty, as calculated from the 227 GHz and 229 GHz channels for M87* (top) and the `eht-crescent` example (bottom).

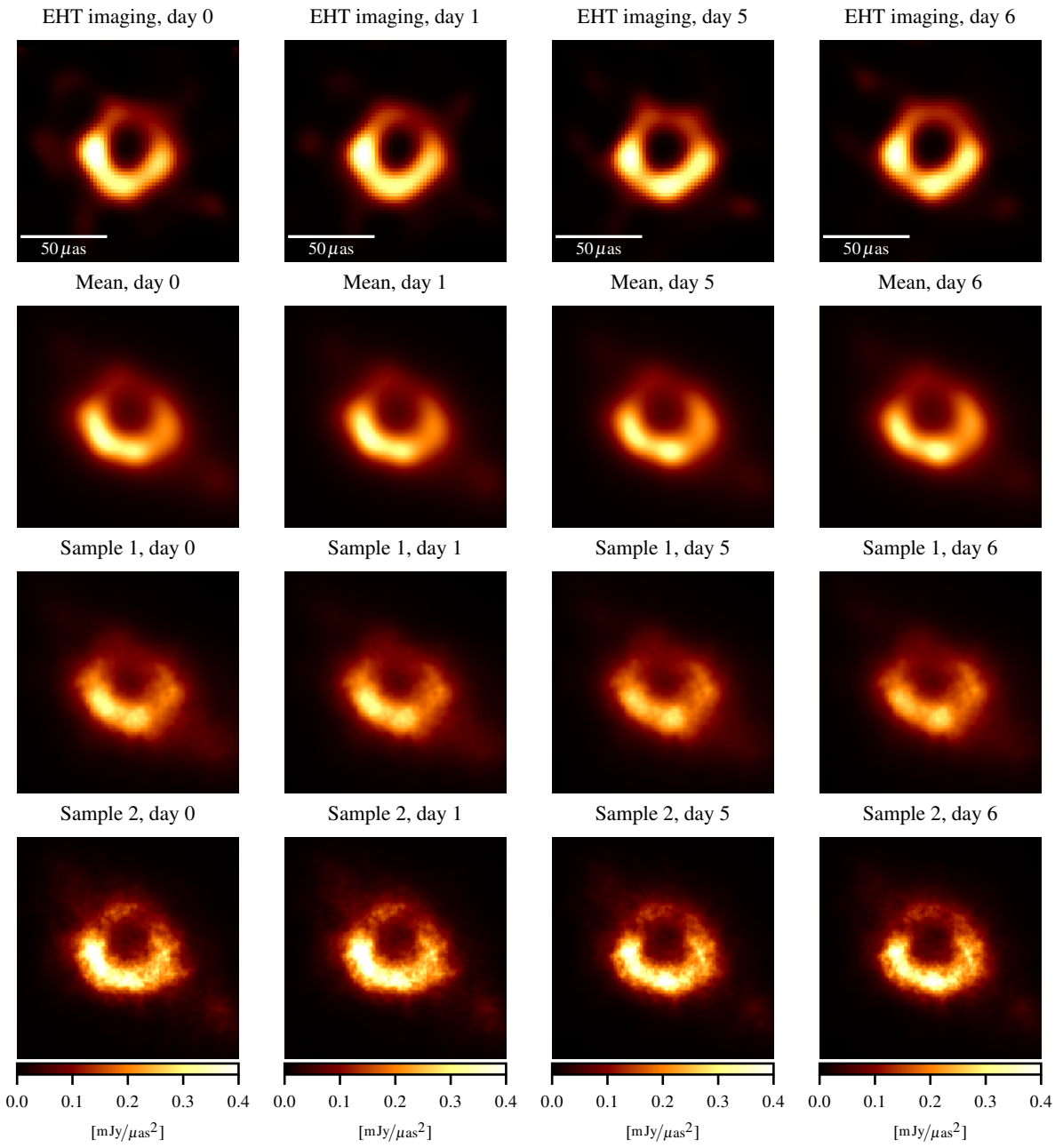

Figure 3.9: Comparison of our imaging result to that of the EHT-imaging pipeline. All panels have the same colour bar. The columns label the four days for which observational data exist. The first row shows snapshot images from the EHT-imaging pipeline for each of the 4 days. The second row shows our mean reconstruction for the same time frame. The third and fourth row each show one posterior sample from our imaging pipeline.

Chapter 4

Probabilistic simulation of partial differential equations

The following chapter has been submitted to Physical Review E *with me as the first author and has been made publicly available on* Arxiv *[45]. All authors read, commented, and approved the final manuscript.*

Abstract

Computer simulations of differential equations require a time discretization, which inhibits to identify the exact solution with certainty. Probabilistic simulations take this into account via uncertainty quantification. The construction of a probabilistic simulation scheme can be regarded as Bayesian filtering by means of probabilistic numerics. Gaussian prior based filters, specifically Gauss-Markov priors, have successfully been applied to simulation of ordinary differential equations (ODEs) and give rise to filtering problems that can be solved efficiently. This work extends this approach to partial differential equations (PDEs) subject to periodic boundary conditions and utilizes continuous Gaussian processes in space and time to arrive at a Bayesian filtering problem structurally similar to the ODE setting. The usage of a process that is Markov in time and statistically homogeneous in space leads to a probabilistic spectral simulation method that allows for an efficient realization. Furthermore, the Bayesian perspective allows the incorporation of methods developed within the context of information field theory such as the estimation of the power spectrum associated with the prior distribution, to be jointly estimated along with the solution of the PDE.

4.1 Introduction

Numerical simulation of partial differential equations (PDEs) has been studied extensively for a long time as PDEs arise naturally in many scientific fields. Recently, fully probabilistic approaches to simulation have been proposed [90, 92], many of them within the

context of probabilistic numerics (PN) [23, 69]. Many probabilistic numerical methods aim to disentangle traditional numerical algorithms into the prior assumptions as well as the (artificial) observations that appear within the algorithm [100]. This provides an uncertainty quantification within the context of Bayesian reasoning [58] and often has led to new variants of the algorithms by replacing prior assumptions [68].

In this work we aim to discuss probabilistic numerical simulation within the context of information field theory (IFT) [31], that is information theory for quantities that are defined over continuous spaces (i.E. fields). Previous works towards an information field theoretical consideration of PDE simulation has been established by means of information field dynamics (IFD) [30, 77]. IFD aims to construct a simulation step that is optimal in the information theoretical sense, that is minimal loss of information about the system between subsequent simulation steps. In this work, however, we follow a line of argument more closely related to PN rather then IFD. We discuss the relations to IFD in further detail once we established the main properties of the probabilistic solver. Nevertheless, the usage of IFT allows for an application of non-parametric estimation of power spectra [33] to the task of PDE simulation. This enables us to construct more sophisticated filters that adapt to the correlation structure of the simulated process.

We notice that our approach has considerable structural similarities to a recent reformulation of probabilistic simulation of ordinary differential equations (ODEs) by means of nonlinear Bayesian filtering [107], however here applied to PDEs.

4.1.1 Introduction to IFT and notation

In IFT we consider fields s^x that are defined over some continuous domain $\Omega \subset \mathbb{R}^d$ where d denotes the dimensionality of the space and x may label a location in a coordinate system on Ω. We aim to provide probabilistic reasoning for fields, and therefore we need to define probability distributions for fields. To this end we equip the function space $L^2\{\Omega\}$ with a scalar product defined as

$$a^\dagger b \equiv \int_\Omega a_x^* \; b^x \; \mathrm{d}x \; , \tag{4.1}$$

where $*$ denotes complex conjugation. Consequently, applications of linear operators $O : L^2\{\Omega\} \to L^2\{\Omega\}$ are denoted as

$$b^x = (Oa)^x = O^x_{\;x'} \; a^{x'} = \int_\Omega O^x_{\;x'} \; a^{x'} \; \mathrm{d}x' \; , \tag{4.2}$$

where we also introduced the continuous version of the Einstein sum convention. This allows us to define a Gaussian distribution with mean m and covariance D for a field s via

$$\begin{aligned} P(s) &= \mathcal{G}(s-m, D) \\ &\equiv \frac{1}{|2\pi D|^{\frac{1}{2}}} \; e^{-\frac{1}{2}(s-m)^\dagger D^{-1}(s-m)} \; , \end{aligned} \tag{4.3}$$

where $|\bullet|$ denotes the functional determinant. (For further details see e.g. [32]). In order to perform inference we additionally need to define a mapping $R : L^2\{\Omega\} \to \mathbb{R}^N$ (often

referred to as response, or design-matrix) that maps a field s to some discrete measurement data $d \in \mathbb{R}^N$. Similar to Eq. (4.2) we write

$$d^i = (Rs)^i = R^i{}_x s^x = \int_\Omega R^i{}_x \; a^x \; \mathrm{d}x \; . \tag{4.4}$$

If we aim to apply the adjoint of R (denoted as $R^\dagger$), however, we get that

$$b^x = (R^\dagger d)^x = (R^\dagger)^x_i d^i \equiv \sum_{i=1}^{N} R^x_i \; d^i \; , \tag{4.5}$$

as we define the scalar product in discrete space as a sum.

4.2 Probabilistic Simulation within IFT

To summarize some key results of probabilistic simulation required for PDE simulation, we start with a brief discussion of ODE simulation and show its relation to Bayesian filtering. For an extensive overview of PN methods for ODE simulation please refer to [101, 107].

4.2.1 Probabilistic ODE simulation

Consider an ODE of the form

$$\dot{s}^t \equiv \frac{\partial s^t}{\partial t} = f\left(s^t\right) \quad \text{with initial condition} \quad s^{t_0} = s^0 \; , \tag{4.6}$$

where $s^t \in \mathbb{R}^M$ denotes the state of the system at time t and $f : \mathbb{R}^M \to \mathbb{R}^M$ is a (non-linear) map.

A Bayesian approach to simulation can the be formulated as: Given some prior knowledge on the field s given as $P(s|s^0)$ we aim to constrain this prior via artificial observations such that it solves Eq. (4.6). The resulting posterior distribution is thus informed via the information in the observations, as well as the prior assumptions. To realize the ODE constraints, we may define a continuous data-set d^t as

$$d^t = \dot{s}^t - f\left(s^t\right) \; , \tag{4.7}$$

and require that $d^t = 0 \quad \forall t$. In general, however, this gives rise to an infinite set of non-tractable constraints and given only finite computational resources, leads to non-computable posterior distributions. Therefore, in the spirit of PN, we require this constraint to be satisfied only at a discrete set of moments in time $T \equiv \{t_i\}_{i \in \{0,\dots,N-1\}}$ via

$$d = R\left(\dot{s} - f\left(s\right)\right) \quad \text{with} \quad R^i_t = \delta\left(t_i - t\right) \; , \tag{4.8}$$

and then require $d^i = 0 \quad \forall i \in \{0, ..., N-1\}$. Note that the choice of R has an impact on the resulting simulation scheme as it introduces a measure and consequently a PN method

for simulation is only fully specified given a prior distribution of the continuous process, as well as a measurement operation. The specific choice of R considered in this work has the desirable property that

$$Rf\left(s\right) = f\left(Rs\right) \ . \tag{4.9}$$

As it will turn put, this property allows us to set up a simulation scheme that only requires to construct the distribution of Rs and $R\dot{s}$ from the prior.

To do so, consider the special case of a Gaussian prior for s of the form of Eq. (4.3). Furthermore let

$$x = \begin{pmatrix} \bar{s} \\ \dot{\bar{s}} \end{pmatrix} \equiv \begin{pmatrix} Rs \\ R\dot{s} \end{pmatrix} = \begin{pmatrix} Rs \\ R\partial_t s \end{pmatrix} \ , \tag{4.10}$$

where ∂_t denotes the derivative w.r.t. t. As Gaussian distributions are closed under affine transformations, we get that x is also Gaussian distributed with mean

$$m_x = \begin{pmatrix} Rm \\ R\dot{m} \end{pmatrix} \ , \tag{4.11}$$

and covariance X

$$X = \begin{pmatrix} RDR^\dagger & RD\partial_t^\dagger R^\dagger \\ R\partial_t DR^\dagger & R\partial_t D\partial_t^\dagger R^\dagger \end{pmatrix} \ , \tag{4.12}$$

where $\partial_t^\dagger$ denotes taking the derivative to the left (i.E. the second index of D in this case). We can use these results to construct the posterior distribution of s given $d = 0$. Let $\underline{T} \equiv [t_0, \infty) \setminus T$ and let $\underline{s}$ be all s^t with $t \in \underline{T}$, we get that

$$\begin{aligned} P\left(s|d=0, s^0\right) &= \int \mathrm{d}\dot{\bar{s}} \ P\left(\underline{s}, \bar{s}, \dot{\bar{s}}|d=0\right) && (4.13) \\ &\propto \int \mathrm{d}\dot{\bar{s}} \ P\left(d=0|\underline{s}, \bar{s}, \dot{\bar{s}}\right) \ P\left(\underline{s}|\bar{s}, \dot{\bar{s}}\right) \ P\left(\bar{s}, \dot{\bar{s}}|s^0\right) \\ &= \int \mathrm{d}\dot{\bar{s}} \ \delta\left(\dot{\bar{s}} - f\left(\bar{s}\right)\right) \ P\left(\underline{s}|\bar{s}, \dot{\bar{s}}\right) \ P\left(\bar{s}, \dot{\bar{s}}|s^0\right) \\ &= P\left(\underline{s} \middle| x = \begin{pmatrix} \bar{s} \\ \dot{\bar{s}} = f\left(\bar{s}\right) \end{pmatrix}\right) \ P\left(x = \begin{pmatrix} \bar{s} \\ \dot{\bar{s}} = f\left(\bar{s}\right) \end{pmatrix} \middle| s^0\right) \ . && (4.14) \end{aligned}$$

First, we notice that the posterior for all $\underline{s}$ remains a Gaussian distribution irrespective of f and is equal to the conditional distribution of s given the values and the first derivatives at all T. Furthermore we may write

$$P\left(x = \begin{pmatrix} \bar{s} \\ \dot{\bar{s}} = f\left(\bar{s}\right) \end{pmatrix} \middle| s^0\right) = P\left(\dot{\bar{s}} = f\left(\bar{s}\right)|\bar{s}, s^0\right) P\left(\bar{s}|s^0\right) \ , \tag{4.15}$$

which ultimately renders the task of simulation a non-linear Bayesian regression problem in $\bar{s}$ [107].

Gauss-Markov processes

For general Gaussian priors, i.E. for general D (see Eq. (4.3)), this approach scales with N^2 (N^3 in case of unknown hyper-parameters in D) as we need to compute conditional distributions for all T. Therefore, as proposed by e.g. [101], one can achieve linear scaling in N via usage of Gauss-Markov processes. In this work we restrict ourselves to the simple case of an integrated Wiener process (IWP), however a generalization to higher order Gauss-Markov process priors is possible as provided by [101]. The IWP may be defined as

$$\ddot{s}^t = \sigma \xi^t \quad \text{with} \quad \xi \sim \mathcal{G}\left(\xi, \mathbb{1}\right) \ , \tag{4.16}$$

and yields the conditional distribution for s^t and $\dot{s}^t$ given their values at a previous time step:

$$\begin{aligned} &P\left(\begin{pmatrix} s^{t_i} \\ \dot{s}^{t_i} \end{pmatrix} \middle| \begin{pmatrix} s^{t_{i-1}} \\ \dot{s}^{t_{i-1}} \end{pmatrix}\right) \\ &= \mathcal{G}\left(\begin{pmatrix} s^{t_i} \\ \dot{s}^{t_i} \end{pmatrix} - \begin{pmatrix} 1 & \Delta_i \\ 0 & 1 \end{pmatrix} \begin{pmatrix} s^{t_{i-1}} \\ \dot{s}^{t_{i-1}} \end{pmatrix}, \sigma^2 \begin{pmatrix} \Delta_i^3/3 & \Delta_i^2/2 \\ \Delta_i^2/2 & \Delta_i \end{pmatrix}\right) \ , \end{aligned} \tag{4.17}$$

where $\Delta_i = t_i - t_{i-1}$.

Using the IWP prior, the posterior Eq. (4.15) reads

$$\begin{aligned} P\left(\bar{s}|d, s^0\right) &\propto \prod_{i=1}^{N-1} P\left(\begin{pmatrix} s^{t_i} \\ f\left(s^{t_i}\right) \end{pmatrix} \middle| \begin{pmatrix} s^{t_{i-1}} \\ f\left(s^{t_{i-1}}\right) \end{pmatrix}\right) \\ &= \prod_{i=1}^{N-1} \left[P\left(\dot{s}^{t_i} = f\left(s^{t_i}\right) | s^{t_i}, s^{t_{i-1}}, \dot{s}^{t_{i-1}} = f\left(s^{t_{i-1}}\right)\right)\right. \\ &\quad \left. P\left(s^{t_i} | s^{t_{i-1}}, \dot{s}^{t_{i-1}} = f\left(s^{t_{i-1}}\right)\right)\right] \ . \end{aligned} \tag{4.18}$$

In words, the observations constructed via R only affect the posterior locally and therefore the Markov property of the prior remains present in the posterior. As a consequence the Bayesian filtering problem defined in Eq. (4.15) decomposes into a set of $N-1$ subsequent filtering problems, one for each s^{t_i}.

4.2.2 PDEs with periodic boundary conditions

To construct a probabilistic method for PDEs consider a generic PDE in $1+1$ dimensions for a scalar field s of the form

$$\dot{s}^{tx} = f\left(s^{tx}, \left(s^{(1)}\right)^{tx}, \left(s^{(2)}\right)^{tx}, ...\right) \ , \tag{4.19}$$

with $f : \mathbb{R} \otimes \mathbb{R} \otimes ... \to \mathbb{R}$, and $s^{(c)}$ denotes the cth spatial derivative of s. We restrict the discussion to scalar fields in $1+1$ dimensions but note that an extension to higher dimensions and vector fields is possible. Furthermore we only consider PDEs that are

compatible with periodic boundary conditions in the spatial domain[1], and, without loss of generality, require the size of the spatial domain to be equal to one.

For a probabilistic solver, we require a prior distribution for s. We remain in the setting of a Gauss-Markov prior and additionally assume independence of space and time prior correlations. I.e. we assume that

$$\left\langle s^{tx} s^{t'x'} \right\rangle_{P(s)} = C^{tt'} \; S^{xx'} = C^{tt'} S\left(|x - x'|\right) \;\; , \tag{4.20}$$

where we include the additional assumption that the spatial correlation structure is a priori statistical homogeneous and isotropic. We set C such that s follows an IWP in time. Furthermore, we define s in terms of its Fourier series

$$s^{tx} = \sum_{k=-\infty}^{\infty} \tilde{s}^{tk} e^{2\pi i k x} \;\; , \tag{4.21}$$

and use the fact that the Fourier modes $\tilde{s}$ of a statistically homogeneous process become statistically independent in Fourier space. The prior assumptions additionally imply that the time evolution of each Fourier mode $\tilde{s}^k$ follows an IWP of the form

$$\ddot{\tilde{s}}^{tk} = \sigma^k \; \xi^{tk} \quad \text{with} \quad \xi \sim \mathcal{G}\left(\xi, \mathbb{1}\right) \;\; , \tag{4.22}$$

with σ such that $|\sigma|^2$ equals the Fourier spectrum associated with the spatial covariance S.

Discrete Measurements

In analogy to the ODE discussion we have to define a discrete set of measurements in order to arrive at a computable posterior distribution. We may use a measurement operator of the form

$$R^{ij}_{tx} = M^i_t \; B^j_x = \delta(t_i - t) \; \delta(x_j - x) \;\; , \tag{4.23}$$

i.E. each measurement singles out a specific location in space-time. We notice that arbitrary (e.g. random) space-time locations again renders the simulation to scale with N^2 (N^3). To minimize this computational burden more sophisticated methods of choosing design points in space-time have been proposed. E.g. [23] aims to choose design points such that the posterior uncertainty is minimized, i.E. by minimizing the trace or the determinant of the posterior covariance w.r.t. the locations of the design points. For many PDEs, however, it is important to satisfy the equation at many points simultaneously in order to arrive at a good numerical approximation. Therefore, in this work, we make use of the specific prior structure to arrive at an almost linear scaling of the proposed method.

To this end we notice that due to the Markov property of the IWP, the distribution at a later time, given all Fourier modes in the past, only depends on the latest Fourier modes.

[1] Other boundary conditions can be enforced by modification of the dynamical equations in the here presented approach, and possibly by a zero padding area between those in the periodic domain. We leave this to future research.

In analogy to Eq. (4.17), for each Fourier mode k we get an independent Markov process of the form

$$\begin{aligned} &P\left(\begin{pmatrix}\tilde{s}^{ik}\\ \dot{\tilde{s}}^{ik}\end{pmatrix}\middle|\begin{pmatrix}\tilde{s}^{(i-1)k}\\ \dot{\tilde{s}}^{(i-1)k}\end{pmatrix}\right) \\ &= \mathcal{G}\left(\begin{pmatrix}\tilde{s}^{ik}\\ \dot{\tilde{s}}^{ik}\end{pmatrix} - \begin{pmatrix}1 & \Delta_i\\ 0 & 1\end{pmatrix}\begin{pmatrix}\tilde{s}^{(i-1)k}\\ \dot{\tilde{s}}^{(i-1)k}\end{pmatrix}, \left|\sigma^k\right|^2 \begin{pmatrix}\Delta_i^3/3 & \Delta_i^2/2\\ \Delta_i^2/2 & \Delta_i\end{pmatrix}\right), \end{aligned} \tag{4.24}$$

with $\tilde{s}^{ik} = (M\tilde{s})^{ik}$.

However, the process only remains Markov if we keep all (infinitely many) modes in memory. If we additionally require the spatial locations to be on the same regular grid with K points, i.E. $x_j = {}^{j}/_{K}$, we notice that we can construct a discrete Markov process since

$$\begin{aligned} &e^{2\pi i(k+nK)x_j} = e^{2\pi ik}e^{2\pi inKj/K} = e^{2\pi ik} \\ &\forall j \in \{0,1,...,K-1\}\ ,\ n \in \mathbb{Z}\ . \end{aligned} \tag{4.25}$$

I.e. each Fourier mode k shifted by multiples of K coincides with the mode k for each location on the grid. Consequently we can represent the field values on the grid using only K modes as

$$\begin{aligned} \left(\bar{s}^{(c)}\right)^{tj} \equiv \left(Bs^{(c)}\right)^{tj} &= \sum_{k=-K/2+1}^{K/2} \left(\tilde{\bar{s}}^{(c)}\right)^{tk} e^{2\pi ikx_j} \\ &\equiv \mathcal{F}_k^j \left(\tilde{\bar{s}}^{(c)}\right)^{tk}\ , \end{aligned} \tag{4.26}$$

where we defined the discrete Fourier transformation $\mathcal{F}$. The finite Fourier modes $\tilde{\bar{s}}$ are defined in terms of $\tilde{s}$ as

$$\left(\tilde{\bar{s}}^{(c)}\right)^{tk} = \sum_{n=-\infty}^{\infty} \left(2\pi i\left(k+nK\right)\right)^c \tilde{s}^{t(k+nK)}\ . \tag{4.27}$$

Each discrete Fourier mode can be expressed in terms of an infinite sum of Gaussian random variables and thus itself is Gaussian. Note that for each spatial derivative c, however, the terms within the sum are different and therefore the summation results in a vector $\tilde{\bar{\mathbf{s}}} = \left(\tilde{\bar{s}}^{(0)}, \tilde{\bar{s}}^{(1)}, ...\right)$ of correlated Gaussian random variables, one for each spatial derivative involved in the PDE. The reason for this is that even though the field and its derivatives can be represented on the same grid, taking the derivative does not commute with the discretization operation B.

The infinite Fourier modes $\tilde{s}^k$ are solutions of the IWP process defined in Eq. (4.24), and therefore we may use an analogous derivation for the discrete representation of the

time derivatives $\dot{\tilde{\mathbf{s}}} = \left(\dot{\tilde{s}}^{(0)}, \dot{\tilde{s}}^{(1)}, ...\right)$ to arrive at a discrete Markov prior of the form

$$\begin{aligned} & P\left(\begin{pmatrix}\tilde{\mathbf{s}}^{ik}\\ \dot{\tilde{\mathbf{s}}}^{ik}\end{pmatrix}\middle|\begin{pmatrix}\tilde{\mathbf{s}}^{(i-1)k}\\ \dot{\tilde{\mathbf{s}}}^{(i-1)k}\end{pmatrix}\right) \\ & = \mathcal{G}\left(\begin{pmatrix}\tilde{\mathbf{s}}^{ik}\\ \dot{\tilde{\mathbf{s}}}^{ik}\end{pmatrix} - \begin{pmatrix}1 & \Delta_i\\ 0 & 1\end{pmatrix}\begin{pmatrix}\tilde{\mathbf{s}}^{(i-1)k}\\ \dot{\tilde{\mathbf{s}}}^{(i-1)k}\end{pmatrix}, \begin{pmatrix}\Delta_i^3/3 & \Delta_i^2/2\\ \Delta_i^2/2 & \Delta_i\end{pmatrix}\otimes \mathbf{D}^k\right) \\ & \forall k \in [-K/2+1, K/2] \ , \end{aligned} \tag{4.28}$$

where $\tilde{\tilde{\mathbf{s}}}^{ik} = M_t^i \tilde{\tilde{\mathbf{s}}}^{tk}$ and $\otimes$ denotes the tensor product. The discrete Fourier mode covariance $\mathbf{D}^k$ takes the form

$$\begin{aligned} \left(\mathbf{D}^k\right)^{cd} &\equiv \left\langle \left(\tilde{\tilde{s}}^{(c)}\right)^k \left(\tilde{\tilde{s}}^{(d)}\right)^k \right\rangle \\ &= (-1)^d \sum_{n=-\infty}^{\infty} (2\pi i(k+nK))^{c+d} \left|\sigma^{k+nK}\right|^2 \ . \end{aligned} \tag{4.29}$$

The Markov property of the IWP remains in the discrete representation of the field since we defined the space and time correlations to be independent a priori. See Appendix 4.A for a derivation of $\mathbf{D}^k$.

The discrete Fourier transformation defined in Eq. (4.26) is invertible, and therefore we can construct the measurement equation associated with the PDE (Eq. (4.19)) in terms of the Fourier modes as

$$d^{ik} = \dot{\tilde{\tilde{s}}}^{ik} - \left(\mathcal{F}^{-1} f\left(\mathcal{F}\tilde{\tilde{s}}^{(0)}, \mathcal{F}\tilde{\tilde{s}}^{(1)}, ...\right)\right)^{ik} \equiv \left(\dot{\tilde{\tilde{s}}} - g\left(\tilde{\tilde{\mathbf{s}}}\right)\right)^{ik} \ . \tag{4.30}$$

Posterior distribution

In direct analogy to the ODE setting, we can combine the observational data d with the prior to construct a posterior distribution. Let $\mathbf{u} = \left(\tilde{\tilde{s}}^{(0)}, \tilde{\tilde{s}}^{(1)}, ...\right)$ be the discretized Fourier space field values and their higher order spatial derivatives and $\mathbf{v} \equiv \left(\dot{\tilde{\tilde{s}}}^{(1)}, ...\right)$ be the time derivative of the spatial derivatives in $\mathbf{u}$, we get that

$$\begin{aligned} & P\left(\mathbf{u}, \mathbf{v} | d=0, \mathbf{u}^0, \mathbf{v}^0\right) \\ & \propto \prod_{i=1}^{N-1} P\left(\begin{pmatrix}\mathbf{u}^i\\ \dot{\tilde{\tilde{s}}}^i = g\left(\mathbf{u}^i\right)\\ \mathbf{v}^i\end{pmatrix}\middle|\begin{pmatrix}\mathbf{u}^{i-1}\\ \dot{\tilde{\tilde{s}}}^{i-1} = g\left(\mathbf{u}^{i-1}\right)\\ \mathbf{v}^{i-1}\end{pmatrix}\right) \\ & = \prod_{i=1}^{N-1}\left[P\left(\mathbf{v}^i\middle|\begin{pmatrix}\mathbf{u}^i\\ \dot{\tilde{\tilde{s}}}^i = g\left(\mathbf{u}^i\right)\end{pmatrix}, \begin{pmatrix}\mathbf{u}^{i-1}\\ \dot{\tilde{\tilde{s}}}^{i-1} = g\left(\mathbf{u}^{i-1}\right)\\ \mathbf{v}^{i-1}\end{pmatrix}\right)\right. \\ & \left. P\left(\begin{pmatrix}\mathbf{u}^i\\ \dot{\tilde{\tilde{s}}}^i = g\left(\mathbf{u}^i\right)\end{pmatrix}\middle|\begin{pmatrix}\mathbf{u}^{i-1}\\ \dot{\tilde{\tilde{s}}}^{i-1} = g\left(\mathbf{u}^{i-1}\right)\\ \mathbf{v}^{i-1}\end{pmatrix}\right)\right] \ . \end{aligned} \tag{4.31}$$

Here, the involved conditional distributions can be directly constructed from Eq. (4.28). We notice that the distribution of $\mathbf{v}^i$ remains Gaussian and we can directly sample it once we solved the simulation step for $\mathbf{u}^i$ by constructing the conditional distribution of $\mathbf{v}^i$ from Eq. (4.28). The distribution of $\mathbf{u}^i$ may again be rewritten in terms of a non-linear filter as

$$\begin{aligned} P\left(\begin{pmatrix}\mathbf{u}^i\\ \dot{\tilde{s}}^i = g\left(\mathbf{u}^i\right)\end{pmatrix}\middle|\begin{pmatrix}\mathbf{u}^{i-1}\\ \dot{\tilde{s}}^{i-1} = g\left(\mathbf{u}^{i-1}\right)\\ \mathbf{v}^{i-1}\end{pmatrix}\right) = \\ = P\left(\dot{\tilde{s}}^i = g\left(\mathbf{u}^i\right)\middle|\mathbf{u}^i, \begin{pmatrix}\mathbf{u}^{i-1}\\ \dot{\tilde{s}}^{i-1} = g\left(\mathbf{u}^{i-1}\right)\\ \mathbf{v}^{i-1}\end{pmatrix}\right) \times \\ P\left(\mathbf{u}^i\middle|\begin{pmatrix}\mathbf{u}^{i-1}\\ \dot{\tilde{s}}^{i-1} = g\left(\mathbf{u}^{i-1}\right)\\ \mathbf{v}^{i-1}\end{pmatrix}\right) \end{aligned} \tag{4.32}$$

Eq. (4.31) and (4.32) describe the central results of our work. Under the given prior assumptions and measurement setting the posterior becomes a Markov process in time in the finite state vector $\mathbf{u}^i$. Furthermore, each time step is presented as a non-linear Bayesian filtering problem, where the second probability on the r.h.s. in Eq. (4.32) is a Gaussian prior distribution in $\mathbf{u}^i$ that acts as a predictive step to construct the next step from the previous one. The first distribution may be regarded as a (in general non-linear) likelihood which acts as a regularization by comparing the time derivative $\dot{\tilde{s}}^i$ constructed via the PDE from $\mathbf{u}^i$, to the conditional distribution of $\dot{\tilde{s}}^i$ that arises from the previous step and the prior process. See Algorithm 1 for a pseudo-code description of the resulting algorithm.

4.2.3 Posterior properties

It is noteworthy that, in contrast to the ODE setting, even though we use an IWP prior in time, it is in general not sufficient to only store the field values on the grid. We also have to keep the involved spatial derivatives $\mathbf{u}$ and, maybe even more surprising, the spatial derivatives of the first time derivative $\mathbf{v}$ in memory, in order to be fully consistent with the continuous prior process. In fact, as the spatial derivatives of the first time derivative do not enter the PDE, we may analytically integrate over these quantities, but the resulting process would loose the Markov property, which we believe is in general not desirable. However, as we have seen, once we have solved the inference problem for $\mathbf{u}$ we can directly sample $\mathbf{v}$ as the conditional distribution remains Gaussian.

On the other hand, given a fixed step size, the spatial resolution, and a spectrum $|\sigma|^2$, we may rewrite the posterior distribution in terms of the generative process associated with the predictive prior of $\mathbf{u}$. This reads

$$\mathbf{u}^i = \mathbf{u}^{i-1} + \Delta_i \begin{pmatrix} g\left(\mathbf{u}^{i-1}\right)\\ \mathbf{v}^{i-1}\end{pmatrix} + \sqrt{\Delta_i^3/3}\, \mathbf{U}\Lambda\mathbf{r}^i \;, \tag{4.33}$$

with $\mathbf{r}^i \sim \mathcal{G}(\mathbf{r}^i, \mathbb{1})$, and where $\mathbf{U}\Lambda\mathbf{U}^\dagger$ denotes the eigen-decomposition of the prior covariance $\mathbf{D}$ with $\mathbf{U}$ being a unitary matrix and Λ a real diagonal matrix. Note that due to the homogeneity of the prior this covariance takes a block diagonal form in k and therefore we only need to decompose a set of K independent $(o+1)$-dimensional matrices where o is the highest spatial derivative involved in the PDE. We notice that for fast decaying spectra (resulting in a strong spatial smoothness) the eigenvalues Λ also decrease very fast. That means that we can define a precision level prior to the simulation up to which we want to keep track of discretization contributions, and set all eigenvalues below this threshold and all associated components in $\mathbf{r}^i$ to zero. This may reduce the burden of storing additional quantities on the grid.

4.2.4 Power spectrum estimation

So far we only considered the case of a given prior power spectrum σ. In practical applications, however, it is often unclear prior to the simulation which spatial correlation structure one should choose given the initial state and the PDE. A strongly decaying spectrum enforcing too much smoothness might result in a poor performance of the simulation algorithm as small scale structures are missing while a very flat spectrum might over-represent these scales and consequently leads to very high uncertainties.

Since we have formulated the simulation problem by means of Bayesian inference, it is straightforward to elevate the power spectrum to an unknown quantity that has to be inferred along with the solution. To this end we may write

$$P(\mathbf{u}, \mathbf{v}, \sigma | d, \mathbf{u}^0, \mathbf{v}^0) \propto P(\mathbf{u}, \mathbf{v} | d, \mathbf{u}^0, \mathbf{v}^0, \sigma)\ P(\sigma)\ , \tag{4.34}$$

where $P(\mathbf{u}, \mathbf{v} | d, \mathbf{u}^0, \mathbf{v}^0, \sigma)$ is defined via Eq. (4.31).

A more difficult question is how to construct a useful prior distribution for σ as in order to construct the distribution of $\mathbf{u}$ and $\mathbf{v}$ we have to compute the infinite sums associated with $\mathbf{D}$ (see Eq. (4.29)). In this work we follow an approach originally developed for power spectrum estimation within the context of Bayesian imaging [9]: First consider the spectrum on a double logarithmic scale as

$$\sigma(|k|) = e^{\tau(l)} \quad \text{with} \quad l = \log(|k|)\ . \tag{4.35}$$

This provides a useful scale for power spectra as power laws appear as straight lines on this scale. As power-law shaped spectra are reasonable for many physical processes, we aim to construct a prior that, in absence of further information, follows a power law. Furthermore we require that deviations from this power-law are smooth (i.E. differentiable) on log-log-scale. To this end we assume that τ solves an IWP process in the log-coordinates l of the form

$$\frac{\partial^2 \tau}{\partial l^2} = \sigma_\tau \xi \quad \text{with} \quad \xi \sim \mathcal{G}(\xi, \mathbb{1})\ , \tag{4.36}$$

where σ_τ is a positive scaling factor. Finally, we realize this process on a regular grid in l with L pixels, up to a maximal value $l_{\max}$, and approximate all intermediate values of τ

via bi-linear interpolation in l. This allows us to approximately compute the covariance $\mathbf{D}$ by summing up all contributions to the sum up to $n_{\max}$ with $l_{\max} = \log(n_{\max}K)$ where K is the number of pixels of the spatial grid. The bi-linear interpolation additionally allows to approximately compute the sum directly from the values of τ on the logarithmic grid l without the need to realize a high resolved version of σ on linear scale. Furthermore, as we define a regular grid on logarithmic scale in $|k|$ we can easily extend the spectrum to extremely large values of $|k|$ (large $n_{\max}$), far below the smallest resolved scales of the simulation. For a detailed discussion of these prior properties see e.g. [9] and [8].

We notice that a time invariant spectrum constructed this way renders the full posterior to be non-Markov since all steps depend on the same spectrum. We can restore the Markov property by introducing a different spectrum for each time step τ^i. Specifically, we assume the spectrum to be piecewise constant for the length of the time step, but different for each step. Furthermore, to increase stability, we may assume that the power spectra of subsequent steps are correlated, which is a reasonable assumption since we do not expect the statistical properties to vary arbitrarily strong between two subsequent time steps. A simple way to introduce such correlations is by assuming that τ follows a discrete time Wiener process, that is

$$\tau^i = \tau^{i-1} + \Delta_i \tilde{\tau}^i \ . \tag{4.37}$$

Specifically the current log-spectrum τ^i can be constructed from the previous one τ^{i-1} and a random component $\tilde{\tau}^i$. We let $\tilde{\tau}^i$ be distributed according to an IWP in the log-Fourier coordinates l, as defined via Eq. (4.36). This renders the full time-Fourier process for τ to be a discrete Wiener Process in time and an IWP in the log-Fourier coordinates l.

4.2.5 Composed algorithm

The full algorithm using power spectrum estimation may be denoted as:

Given the previous state $X^{(i-1)} = \left(\mathbf{u}(i-1), \mathbf{v}(i-1), \tau^{(i-1)}\right)$, use the posterior distribution constructed from Eq. (4.32) and Eq. (4.34) to compute an estimate (or sample) for $\mathbf{u}^i$ and τ^i via e.g. a joint Maximum a Posteriori (MAP) estimate, a Variational approximation, or Monte Carlo based sampling. Use this estimate (sample) in the distribution of $\mathbf{v}^i$ (see Eq. (4.31)) to sample $\mathbf{v}^i$ conditional to $\mathbf{u}^i$, τ^i and the previous state $X^{(i-1)}$. Given the new full state X^i we may repeat the procedure to compose a new time-step. For a pseudo code representation see Algorithm 2

Initial conditions

We notice that initial conditions s^0, evaluated on the grid, do not fully determine the initial state X^0 that is needed to start the simulation as X^0 also consists of the spatial derivatives of the continous field, evaluated on the grid, and the initial power spectrum τ^0. However, there are multiple ways to estimate an initial state X^0 given s^0. For example we may estimate the large scale (scales that are resolved by the simulation grid) power spectrum from the initial conditions directly and accompany this estimate with a consistent initial

Algorithm 1: PDE simulation with fixed spectrum

Input: $\mathbf{u}^0$, $\mathbf{v}^0$, σ, PDE
for $i = 1$ *to* N **do**
  Given $(\mathbf{u}^{i-1}, \mathbf{v}^{i-1})$ and σ, solve Bayesian filtering problem (Eq. (4.32)) to get an estimate (sample) for $\mathbf{u}^i$
  Given $\mathbf{u}^i$ use Eq. (4.31) to sample $\mathbf{v}^i$
end
Output: $\{(\mathbf{u}^i, \mathbf{v}^i)\}_{i\in\{1,\dots,N\}}$

Algorithm 2: PDE simulation with variable spectrum

Input: $\mathbf{X}^0$, PDE
for $i = 1$ *to* N **do**
  Given $\mathbf{X}^{i-1}$, solve the joint Bayesian filtering problem of Eqs. (4.32) and (4.34) to get an estimate (sample) for $\mathbf{u}^i$ and τ^i
  Given $\mathbf{u}^i$ and τ^i use Eq. (4.31) to sample $\mathbf{v}^i$
  $\mathbf{X}^i \leftarrow (\mathbf{u}^i, \mathbf{v}^i, \tau^i)$
end
Output: $\{\mathbf{X}^i\}_{i\in\{1,\dots,N\}}$

guess for the small scale spectrum. Given this spectrum, it is straightforward to estimate the spatial derivatives needed for X^0, given the spectrum and s^0 via Gaussian regression. We may even perform a probabilistic estimate and sample from the corresponding distribution to construct X^0 in order to propagate the uncertainty that arises from insufficient knowledge of the initial state into the simulation.

In this work, however, we want to study the performance of the simulation algorithm itself, and therefore assume that the initial state X^0 is fully given, i.E. we start with an initial condition that allows us to compute the spatial derivatives analytically.

4.3 Applications

In the following we present the application of the proposed methods to two systems, the diffusion equation as well as the viscous Burgers equation. All applications are conducted on the same regular grid in space, with 128 pixels and periodic boundary conditions. The power spectra are realized on a logarithmic regular grid with 500 pixels and a maximal value $l_{\max}$ corresponding to an effective Fourier space 100 times the resolution of the simulation grid. This large effective Fourier space ensures that, at any point in the given examples, the spectra are numerically zero outside this region.

4.3.1 Diffusion equation

To emphasize the influence of the spectrum on the simulation we start with the simple case of a diffusion equation, that is

$$\dot{s} = f(s) = \nu \, s^{(2)} \ , \ \nu > 0 \ , \tag{4.38}$$

and choose a Gaussian profile as the initial state. In Figure 4.1 we depict the MAP estimate of the first step for a step size of $\Delta_1 = 0.04$, and for $\nu = 0.01$. We show two different modes of the simulation scheme: the case of a given generic power spectrum of the form $|\sigma^k|^2 \propto |k|^{-6}$ as well as the case where we optimize for the spectrum together with the solution. As a comparison, we also compute the solution given by the trapezodial rule, where in this case the spatial derivatives are computed via discrete Fourier derivatives, i.E. $\left(s^{(2)}\right)^k = (2\pi i k)^2 \left(s^{(0)}\right)^k$. This method may serve as a standard comparison as it also requires the differential equation to be satisfied for the current as well as the future state simultaneously and therefore is an implicit method of second order, such as the two approaches proposed in this work are. We see in Figure 4.1 that compared to the standard method, both approaches are closer to the ground truth, with the optimized spectrum being slightly closer.

Furthermore, in Figure 4.2, we compare the ground truth to the posterior mean of the simulation and also depict the posterior uncertainty of the problem. We approximate the posterior distribution via the empirical Bayes approach, that is, we use the Maximum a posterior (MAP) estimate of the logarithmic power spectrum τ^* and compute the conditional posterior distribution of the solution s, given τ^*. This conditional posterior is analytically computable since the linear dynamics together with a Gaussian prior distribution results in a Gaussian posterior for s, given τ^*. We see that the posterior mean is in agreement with the ground truth within posterior uncertainties. Furthermore, on the right hand side of Figure 4.2, we depict the residual between the ground truth and the reconstruction as a function of the step size for various locations. Again, the deviation agrees with the uncertainties and furthermore we notice that due to the fact that the prior is stationary, and the diffusion equation is linear and stationary, the posterior distribution also remains a stationary process in space and therefore the posterior uncertainty is the same for every location.

Finally, in Figure 4.3 we depict the time evolution of the simulation together with the ground truth and the estimated power spectra for every time step. As a comparison, we also depict the time evolution for a simulation setting where we used the power spectra computed from subsequent steps of the ground truth, and solved the simulation problem conditional to these spectra.

We see that as time progresses, the initially sharp spatial distribution tends to decay and smooth out over the spatial domain. Consequently, the reconstructed power spectra show less power on small scales as time progresses and only large scale power remains. Furthermore, the overall magnitude of the power spectrum decreases, which indicates that the uncertainty (and therefore the local error) of later time steps become smaller. This adaptive control of the spectrum leads to a better quantification of the local error and

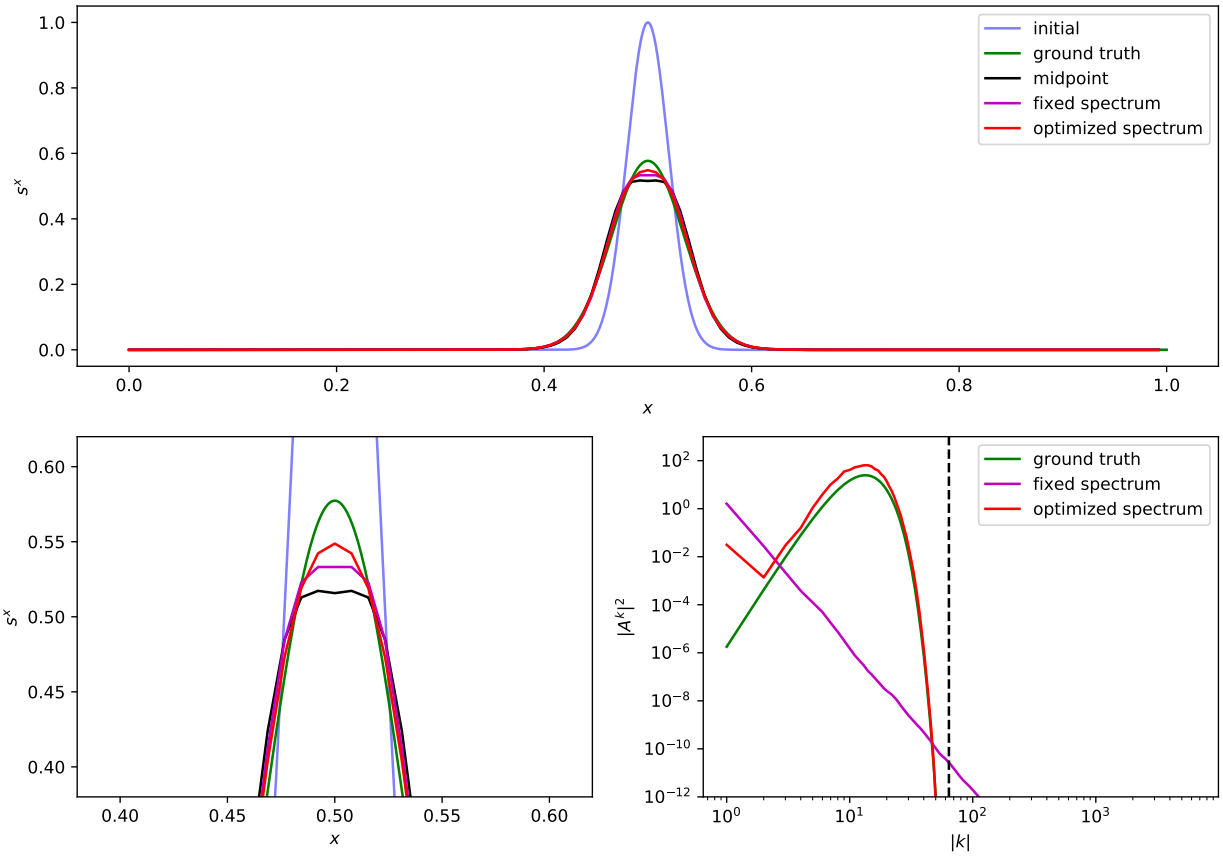

Figure 4.1: **Top:** First time step of the simulation of the diffusion equation with an initial Gaussian profile (blue). The green line corresponds to the ground truth, the black line to the midpoint rule, the purple line to the posterior mean of the reconstruction using a fixed power spectrum $\propto |k|^{-6}$, and the red line corresponds to the MAP estimate of the simulation with an adaptive power spectrum. **Bottom left:** Detailed version of the simulation step zoomed into the central region. **Bottom right:** Power spectra of the simulation on a double-logarithmic scale. Purple: Spectrum of the simulation step with a fixed spectrum. Red: MAP estimate of the optimized spectrum. Green: Ground truth of the spectrum. Here ground truth refers to the spectrum that was reconstructed using the true time evolution as a realization of the corresponding Gaussian prior distribution. The black dashed line indicates the largest harmonic mode corresponding to the resolution of the simulation.

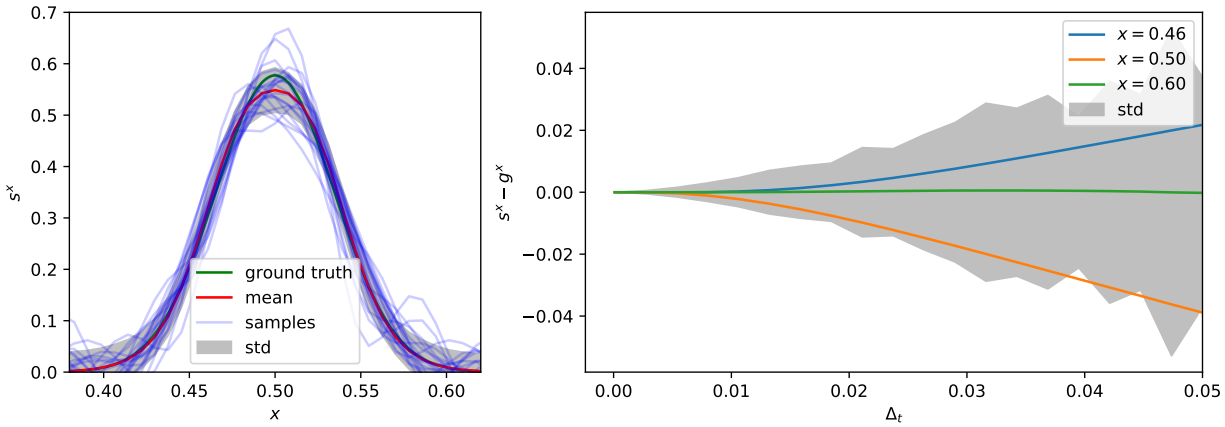

Figure 4.2: **Left:** Ground truth (green), posterior mean (red), posterior samples (light blue), and posterior standard deviation (gray) of the first time step of the diffusion equation. The posterior samples as well as the standard deviation were conducted by means of the empirical Bayes' approach. Specifically, the posterior distribution conditional to the MAP estimate of the optimized spectrum is used. **Right:** Colored lines: Residual difference between the ground truth and the posterior mean at multiple locations of the spatial domain as a function of step size Δ_t. The corresponding posterior standard deviation (valid for any location) is given as the gray contour.

therefore also leads to a more sophisticated control of the global error of the system. We notice, however, that the inferred power spectra of intermediate steps are substantially different from the power spectra of the ground truth. First, on the largest scales the reconstructed power spectra has more power compared to the ground truth. This is a common issue that appears when jointly inferring a field with its power spectrum, as for these modes inference is very degenerate and consequently mostly dominated by the prior assumptions. A more suitable prior in terms of more restrictive hyper-parameters might improve this behaveiour. The second difference becomes apparent for small scale modes where there is too much power around $|k| \in [20, 30]$. We believe that this effect is rooted in the large step size of the given simulation setting: The first steps of the ground truth show a rapid decay of these modes which cannot fully be captured by the simulation step and thus power remains on these scales that gets picked up by the power spectra estimate. However, as time progresses, the power of these scales eventually decay due to the diffusive dynamics of the process.

4.3.2 Burger's equation

As a second example, we study the performance of the proposed approach in the context of the (viscous) Burgers equation. Specifically

$$\dot{s} + \; s \; s^{(1)} = \nu \; s^{(2)} \; . \tag{4.39}$$

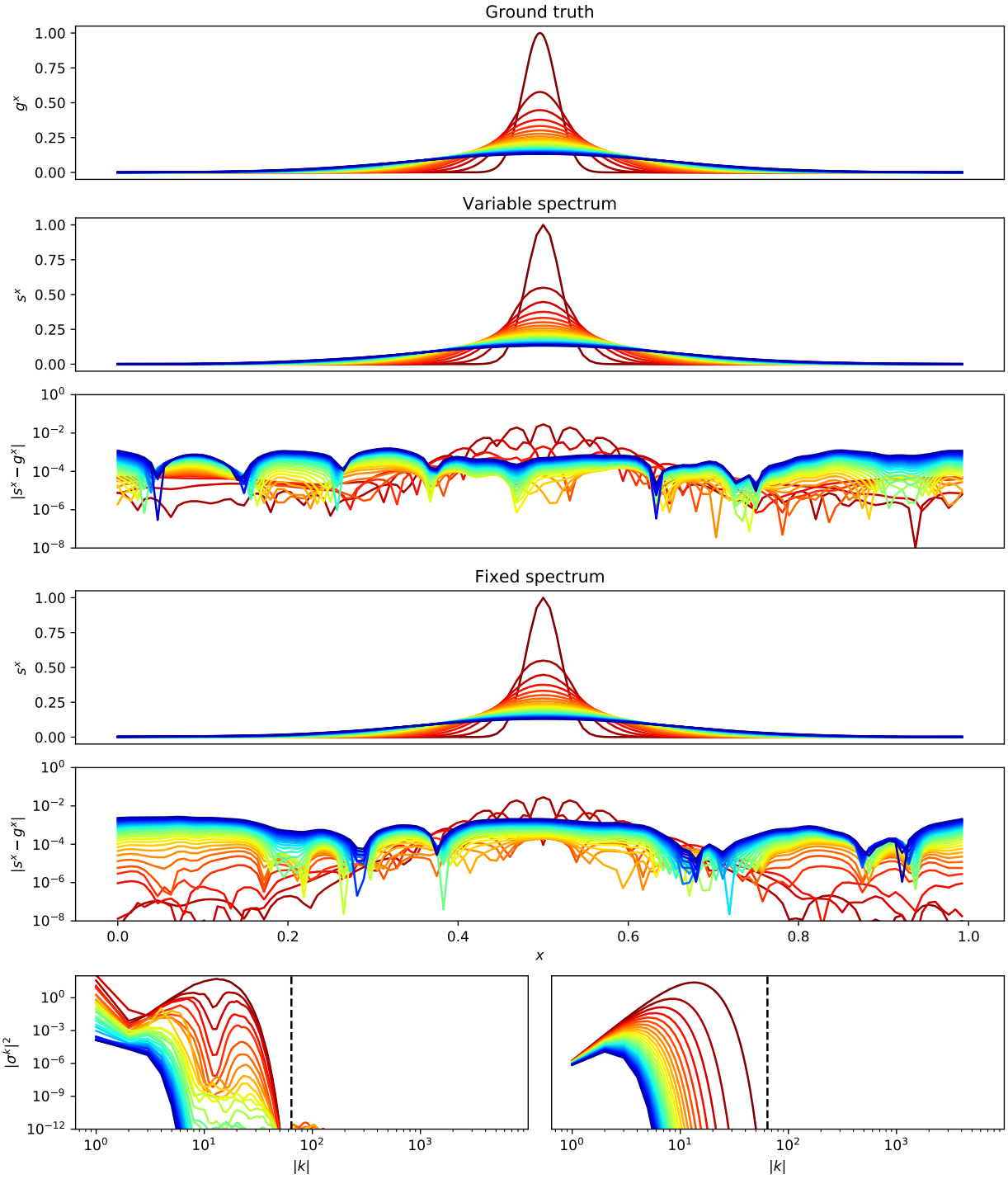

Figure 4.3: Color coded time evolution of the diffusion equation. Red indicates early times and blue indicates the latest time-steps. **Top to bottom:** Ground truth, reconstruction using a variable spectrum (i.E. joint optimization for solution and spectrum), residual norm between ground truth and reconstruction, reconstruction using the fixed spectrum derived from the ground truth, and corresponding residual norm. **Bottom left:** Reconstructed power spectra for each time-step of the joint optimization case. **Bottom Right:** Power spectra computed from the ground truth.

We again start with a Gaussian profile as the initial state and set $\Delta_i = 3 \times 10^{-3}$ and $\nu = 4 \times 10^{-3}$.

The Burger's equation is known to develop strong shock waves for small viscosity ν, which means that in contrast to the diffusion equation, small scale structures become more relevant as time progresses. Indeed we find that if we compute the power spectra of subsequent time steps from the ground truth (see bottom right of Figure 4.4) we see how the spectrum gains power on small scales, while the large scale power remains almost unchanged. In addition we also notice that after a few time steps there is non-negligible power on scales that are smaller then the smallest resolved scales of the simulation grid.

It turns out that, when applying the adaptive simulation to this setup (see Figure 4.4), it is only possible to consistently infer the power spectra along with the solution for scales that are also resolved by the simulation grid. As we only require the differential equation to be satisfied on the grid, there is no direct information about smaller scales that enter the reconstruction and therefore the power spectrum estimation, and ultimately also the simulation itself breaks down as the shock forms. This leads us to the conclusion that using only the feedback of small scales to the large scales provides insufficient information to properly infer the small scale statistics. Without further prior information, we believe that the only way to properly access these scales is via resolving them on a grid with high enough resolution.

However, we notice that it is possible to circumvent the need of realizing the process on a high resolution grid, via the usage of appropriate prior information. To this end consider the middle panels of Figure 4.4, where we used the power spectra estimated from the ground truth to construct a simulation scheme with fixed spectrum on the same resolution as the adaptive one (i.E. a spatial discretization of 128 pixels). It turns out that in contrast to the adaptive scheme, the simulation remains stable and is in agreement with the ground truth long after the adaptive scheme diverged. This result highlights the second key mechanism of a probabilistic treatment of PDE simulation: even though the spatial resolution appears to be insufficient to fully resolve the state, the consistent treatment of discretization via the introduction of spatial derivatives as additional random variables allows for a simulation that remains in agreement with the ground truth. As the correct power spectra are given in this setup, they provide small scale structures consistent with the given PDE and in turn allow for a correct feedback of the small (unresolved) scales to larger (resolved) scales.

4.4 Comparison to IFD

In this work, as well as in IFD there exists the concept of a measurement operator R that specifies the evaluated values of the field. In IFD the resulting measurements are the quantities that are ultimately stored on a computer for a given time-step, meaning that if R singles out a finite set of spatial locations, as used in this work, the corresponding field values are stored. In contrast, in this work not only the field values but also the spatial derivatives involved in the PDE are stored. However, we note that one can alter the measurement operator of IFD to measure not only the values but also the spatial

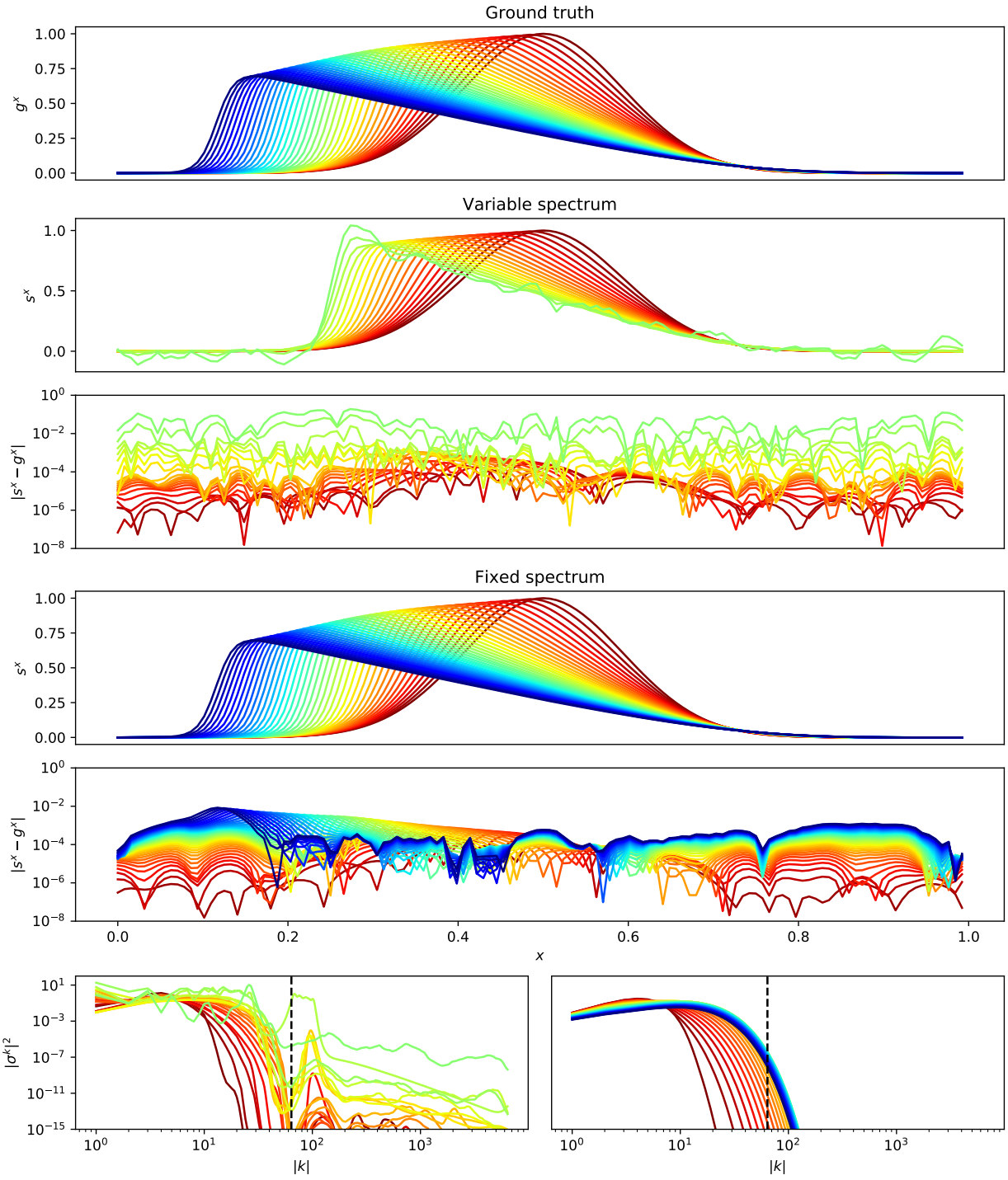

Figure 4.4: Same composition as Figure 4.3, but for the time evolution of the Burgers equation.

derivatives, to result at the same set of quantities that have to be stored. The important difference is that while in IFD this is a choice made by the user, in this work it is a result of the method in order to arrive at a computable distribution that is consistent with the continuous prior process.

Furthermore, in this work R also defines the set of space-time locations at which the process has to fulfill the PDE. This is fundamentally different from IFD as IFD aims to fulfill the PDE at every location. As a consequence there is no need for a prior time correlation in IFD as, in case of a Gaussian prior, the only quantity necessary to translate between the finite state and the distribution of the field is a prior spatial correlation structure. However, we note that for most non-linear applications, the exact time evolution that is required for IFD is not available and thus an approximation has to be made there, which is not captured in a probabilistic fashion. Consequently uncertainties arising from approximated time evolution are not captured within IFD, while the approach in this work takes into account these uncertainties and aims to fill the time gap via the assumed prior time correlation structure. However, requiring the PDE to be satisfied only at a discrete set of locations is also problematic as we have seen, in particular when we aim to infer the prior correlation structure (i.E. the power spectra) on scales that are not resolved by R.

4.5 Conclusion

In this work we derived a fully probabilistic framework for simulation of PDEs subject to periodic boundary conditions. The proposed method makes use of continuous space-time Markov process priors that are stationary in space, and incorporates artificial observational constraints that require the PDE to be satisfied on a regular grid. The Markov property allows for a formulation of the posterior such that the distribution of the current state is only conditional on the state at the previous time-step. The state of the system, however, not only consists of the field values realized on the grid, but also consists of the values of all spatial derivatives involved in the PDE. Only if these random variables are kept track of, the discrete Markov realization is consistent with the continuous process. Furthermore, the usage of prior distributions that are stationary in space, together with sampling on a regular and periodic grid with K pixels, allows for an efficient $K \log(K)$ scaling of a single step of the algorithm via incorporation of Fast Fourier Transforms.

The Bayesian analysis of the problem allows for inference of hyper parameters, such as the spatial correlation structure, i.E. the prior power spectrum, alongside with the solution of the simulation. To this end we incorporate a non-parametric method of power spectra estimation, originally developed for Bayesian imaging by means of information field theory. The resulting joint estimation of spectrum and realization of the process leads to a simulation scheme that is closer to the ground truth compared to a method with a fixed, generic spectrum, and also allows for a more sophisticated error analysis in terms of the posterior uncertainty. We notice, however, that without further prior information about the small scale statistics, the inference of the power spectrum is only valid up to scales that are resolved by the simulation grid. As we have seen in the application to the

Burgers equation, once scales below the grid resolution become relevant for the solution, the estimation of the spectrum becomes inaccurate, and as a consequence the simulation starts to diverge from the true solution. If an accurate estimation of the small scale spectra are available, however, we notice that it is possible to use these spectra for a low-resolution simulation that remains consistent with the high-resolution setting.

Finally we may conclude that the approach for probabilistic PDE simulation provides novel insights into the interplay between prior assumptions entering a simulation algorithm and the involved PDE. However, additional work, in particular concerning small (unresolved) scale statistics, has to be done in order to improve the performance and stability of the proposed approach.

On the other hand, in addition to Bayesian uncertainty quantification, a fully probabilistic approach to simulation enables several novel key properties compared to traditional numerical simulation. For example, as the analysis gives rise to a posterior probability distribution that may be separated into a generative prior and a likelihood, it is straightforward to incorporate the simulation into a larger inference framework, in order to estimate for example parameters of the PDE or initial conditions, from observational data.

In addition, modern day machine learning techniques can be used to speed up the simulation algorithm. In particular neural networks have already successfully been applied to simulation using training data composed via traditional numerical simulation as an input (see e.g. [103]). On the other hand, to circumvent the need of generating training data, which might be very expensive, [91] has demonstrated that is possible to train a neural network to approximate the solution directly by minimizing the squared norm of the deviations of the PDE from zero at a discrete set of space-time locations using only the initial state and the PDE as an input. However, in [91], it has also been demonstrated that training a network to reproduce the internal stages of a high-order Runge-Kutta scheme rather than solely minimizing the squared norm associated with the PDE, appears to be more efficient due to the additional prior assumptions incorporated in the Runge-Kutta scheme. As Runge-Kutta type methods have a probabilistic interpretation in terms of a Gaussian process prior [100], these results indicate that on one hand, neural networks are capable of approximating simulation steps, and on the other hand that a probabilistic posterior distribution for simulation, as derived in this work, may provide a more sophisticated measure for neural-network training. Specifically the posterior distribution is informed about both, the differential equation being satisfied, and a notion of continuity (and differentiability) in space and time in terms of the prior assumptions.

All in all, we believe that the probabilistic approach to simulation, in particular in terms of probabilistic numerics, is capable to provide further insights into numerical simulation, and to generalize existing algorithms. However, further work has to be done in order to arrive at a class of simulation algorithms that are capable of tackling broader classes of physically relevant PDEs.

Appendix

4.A Discrete prior

Consider a Gaussian random field s^{tx} with $x \in [0,1]$ on a periodic domain and $t \in [t_0, \infty)$. Furthermore s has statistically homogeneous and isotropic statistics in space and follows an IWP in time. Specifically:

$$s^{tx} = \sum_{k=-\infty}^{\infty} \tilde{s}^{tk} e^{2\pi i k x} \tag{4.40}$$

$$\ddot{\tilde{s}}^{tk} = \sigma^k \, \xi^{tk} \quad \text{with} \quad \xi \sim \mathcal{G}(\xi, \mathbb{1}) \; . \tag{4.41}$$

$$\tag{4.42}$$

If we define a discretization operation of the form

$$R^{ij}{}_{tx} = M^i_{\ t} \; B^j_{\ x} = \delta(t_i - t) \; \delta(x_j - x) \; , \tag{4.43}$$

with $x_j = {}^j\!/_K$ for $j \in \{0, 1, ..., K-1\}$, it follows from Eqs. (4.41) and (4.42) that all Fourier modes are independent and follow IWP processes of the form:

$$\begin{aligned} &P\left(\begin{pmatrix}\tilde{s}^{ik}\\ \dot{\tilde{s}}^{ik}\end{pmatrix}\middle|\begin{pmatrix}\tilde{s}^{(i-1)k}\\ \dot{\tilde{s}}^{(i-1)k}\end{pmatrix}\right) \\ &= \mathcal{G}\left(\begin{pmatrix}\tilde{s}^{ik}\\ \dot{\tilde{s}}^{ik}\end{pmatrix} - \begin{pmatrix}1 & \Delta_i\\ 0 & 1\end{pmatrix}\begin{pmatrix}\tilde{s}^{(i-1)k}\\ \dot{\tilde{s}}^{(i-1)k}\end{pmatrix}, \left|\sigma^k\right|^2 \begin{pmatrix}{}^{\Delta_i^3}\!/_3 & {}^{\Delta_i^2}\!/_2\\ {}^{\Delta_i^2}\!/_2 & \Delta_i\end{pmatrix}\right), \end{aligned} \tag{4.44}$$

with $\Delta_i = t_i - t_{i-1}$ and $\tilde{s}^{ik} = (M\tilde{s})^{ik}$.

As x_j is sampled on a regular grid, from Eq. (4.25) we get that

$$\left(\tilde{\tilde{s}}^{(c)}\right)^{ik} = \sum_{n=-\infty}^{\infty} \left(2\pi i \left(k + nK\right)\right)^c \tilde{s}^{i(k+nK)} \; , \tag{4.45}$$

$$\left(\dot{\tilde{\tilde{s}}}^{(c)}\right)^{ik} = \sum_{n=-\infty}^{\infty} \left(2\pi i \left(k + nK\right)\right)^c \dot{\tilde{s}}^{i(k+nK)} \; , \tag{4.46}$$

with $k \in [-{}^K\!/_2 + 1, {}^K\!/_2]$.

Proposition: The random vectors $\tilde{\tilde{\mathbf{s}}} = \left(\tilde{\tilde{s}}^{(0)}, \tilde{\tilde{s}}^{(1)}, ...\right)$ and $\dot{\tilde{\tilde{\mathbf{s}}}} = \left(\dot{\tilde{\tilde{s}}}^{(0)}, \dot{\tilde{\tilde{s}}}^{(1)}, ...\right)$ are Gaussian distributed according to Eq. (4.28).

As the involved discretization operation is a linear operation, it is sufficient to show that the mean and covariance take the proposed form, since $\tilde{s}$ and $\dot{\tilde{s}}$ are itself Gaussian

distributed. For the mean we get that

$$\begin{aligned}
\left\langle \left(\tilde{\bar{s}}^{(c)}\right)^{ik} \right\rangle &= \sum_{n=-\infty}^{\infty} \left(2\pi i \left(k+nK\right)\right)^c \left\langle \tilde{s}^{i(k+nK)} \right\rangle \\
&= \sum_{n=-\infty}^{\infty} \left(2\pi i \left(k+nK\right)\right)^c \left(\tilde{s}^{(i-1)(k+nK)} + \Delta_i \dot{\tilde{s}}^{(i-1)(k+nK)}\right) \\
&= \underbrace{\sum_{n=\infty}^{\infty} \left(2\pi i \left(k+nK\right)\right)^c \tilde{s}^{(i-1)(k+nK)}}_{=\left(\tilde{\bar{s}}^{(c)}\right)^{(i-1)k}} \\
&+ \Delta_i \underbrace{\sum_{n=-\infty}^{\infty} \left(2\pi i \left(k+nK\right)\right)^c \dot{\tilde{s}}^{(i-1)(k+nK)}}_{=\left(\dot{\tilde{\bar{s}}}^{(c)}\right)^{(i-1)k}} \\
&= \left(\tilde{\bar{s}}^{(c)}\right)^{(i-1)k} + \Delta_i \left(\dot{\tilde{\bar{s}}}^{(c)}\right)^{(i-1)k} ,
\end{aligned} \tag{4.47}$$

and similarly

$$\begin{aligned}
\left\langle \left(\dot{\tilde{\bar{s}}}^{(c)}\right)^{ik} \right\rangle &= \sum_{n=-\infty}^{\infty} \left(2\pi i \left(k+nK\right)\right)^c \left\langle \dot{\tilde{s}}^{i(k+nK)} \right\rangle \\
&= \sum_{n=-\infty}^{\infty} \left(2\pi i \left(k+nK\right)\right)^c \dot{\tilde{s}}^{(i-1)(k+nK)} \\
&= \left(\dot{\tilde{\bar{s}}}^{(c)}\right)^{(i-1)k} .
\end{aligned} \tag{4.48}$$

For the equal time covariance we get

$$\begin{aligned}
&\left\langle \left(\left(\dot{\tilde{\bar{s}}}^{(c)}\right)^{ik} - \left\langle \left(\dot{\tilde{\bar{s}}}^{(c)}\right)^{ik} \right\rangle \right) \left(\left(\dot{\tilde{\bar{s}}}^{(d)}\right)^{iq} - \left\langle \left(\dot{\tilde{\bar{s}}}^{(d)}\right)^{iq} \right\rangle \right)^* \right\rangle \\
&= \sum_{n,m=-\infty}^{\infty} \left(2\pi i \left(k+nK\right)\right)^c \left(-2\pi i \left(q+mK\right)\right)^d \times \\
&\underbrace{\left\langle \left(\dot{\tilde{s}}^{i(k+nK)} - \left\langle \dot{\tilde{s}}^{i(k+nK)} \right\rangle \right) \left(\dot{\tilde{s}}^{i(q+mK)} - \left\langle \dot{\tilde{s}}^{i(q+mK)} \right\rangle \right)^* \right\rangle}_{\delta_{nm}\delta_{kq}\left|\sigma^{k+nK}\right|^2 \Delta_i} \\
&= \delta_{kq}\ \Delta_i\ (-1)^d \sum_{n=-\infty}^{\infty} \left(2\pi i \left(k+nK\right)\right)^{c+d} \left|\sigma^{k+nK}\right|^2 \\
&= \delta_{kq}\ \Delta_i \left(\mathbf{D}^k\right)^{cd} ,
\end{aligned} \tag{4.49}$$

where we recover the definition of $\mathbf{D}^k$ (Eq. (4.29)). An analogous computation of the covariance of $\tilde{\bar{s}}^{(c)}$ yields the same result with Δ_i being replaced by ${}^{\Delta_i^3}/_3$. Similarly the cross correlation between $\tilde{\bar{s}}^{(c)}$ and its time derivative also results in the same covariance with a pre-factor of ${}^{\Delta_i^2}/_2$.

Acknowledgements

We would like to thank Reimar Leike and Philipp Arras for fruitful discussions and constructive feedback throughout the development process.

Chapter 5

Geometric variational inference

The following chapter has first been published in Entropy *with me as the first author [47]. All authors read, commented, and approved the final manuscript.*

Abstract

Efficiently accessing the information contained in non-linear and high dimensional probability distributions remains a core challenge in modern statistics. Traditionally, estimators that go beyond point estimates are either categorized as Variational Inference (VI) or Markov-Chain Monte-Carlo (MCMC) techniques. While MCMC methods that utilize the geometric properties of continuous probability distributions to increase their efficiency have been proposed, VI methods rarely use the geometry. This work aims to fill this gap and proposes geometric Variational Inference (geoVI), a method based on Riemannian geometry and the Fisher information metric. It is used to construct a coordinate transformation that relates the Riemannian manifold associated with the metric to Euclidean space. The distribution, expressed in the coordinate system induced by the transformation, takes a particularly simple form that allows for an accurate variational approximation by a normal distribution. Furthermore, the algorithmic structure allows for an efficient implementation of geoVI which is demonstrated on multiple examples, ranging from low-dimensional illustrative ones to non-linear, hierarchical Bayesian inverse problems in thousands of dimensions.

5.1 Introduction

In modern statistical inference and machine learning it is of utmost importance to access the information contained in complex and high dimensional probability distributions. In particular in Bayesian inference, it remains one of the key challenges to approximate samples from the posterior distribution, or the distribution itself, in a computationally fast and accurate way. Traditionally, there have been two distinct approaches towards this problem: the direct construction of posterior samples based on Markov Chain Monte-Carlo (MCMC)

methods [51, 19, 81], and the attempt to approximate the probability distribution with a different one, chosen from a family of simpler distributions, known as variational inference (VI) [16, 60, 96, 75] or variational Bayes' (VB) methods [70, 44, 104]. While MCMC methods are attractive due to their theoretical guarantees to reproduce the true distribution in the limit, they tend to be more expensive compared to variational alternatives. On the other hand, the family of distributions used in VI is typically chosen ad-hoc. While VI aims to provide an appropriate approximation within the chosen family, the entire family may be a poor approximation to the true distribution.

In recent years, MCMC methods have been improved by incorporating geometric information of the posterior, especially by means of Riemannian manifold Hamilton Monte-Carlo (RMHMC) [53], a particular hybrid Monte-Carlo (HMC) [29, 12] technique that constructs a Hamiltonian system on a Riemannian manifold with a metric tensor related to the Fisher information metric of the likelihood distribution and the curvature of the prior. For VI methods, however, the geometric structure of the true distribution has rarely been utilized to motivate and enhance the family of distributions used during optimization. One of the few examples being [99] where the Fisher metric has been used to reformulate the task of VI by means of α-divergencies in the mean-field setting.

In addition, a powerful variational approximation technique for the family of normal distributions utilizing infinitesimal geometric properties of the posterior is Metric Gaussian Variational Inference (MGVI) [71]. In MGVI the family is parameterized in terms of the mean m, and the covariance matrix is set to the inverse of the metric tensor evaluated at m. This choice ensures that the true distribution and the approximation obtain the same geometric properties infinitesimally, i.E. at the location of the mean m. In this work we extend the geometric correspondence used by MGVI to be valid not only at m, but also in a local neighborhood of m. We achieve this extension by means of an invertible coordinate transformation from the coordinate system used within MGVI, in which the curvature of the prior is the identity, to a new coordinate system in which the metric of the posterior becomes (approximately) the Euclidean metric. We use a normal distribution in these coordinates as the approximation to the true distribution and thereby establish a non-Gaussian posterior in the MGVI coordinate system. The resulting algorithm, called geometric Variational Inference (geoVI) can be computed efficiently and is inherently similar to the implementation of MGVI. This is not by mere coincidence: To linear order, geoVI reproduces MGVI. In this sense, the geoVI algorithm is a non-linear generalization of MGVI that captures the geometric properties encoded in the posterior metric not only infinitesimally, but also in a local neighborhood of this point. We include an implementation of the proposed geoVI algorithm into the software package Numerical Information Field Theory (NIFTy [6]), a versatile library for signal inference algorithms.

5.1.1 Mathematical setup

Throughout this work, we consider the joint distribution $P(d, s)$ of observational data $d \in \Omega$ and the unknown, to be inferred signal s. This distribution is factorized into the likelihood of observing the data, given the signal $P(d|s)$, and the prior distribution $P(s)$. In general,

only a subset of the signal, denoted as s', may be directly constrained by the likelihood, such that $P(d|s) = P(d|s')$, and therefore there may be additional hidden variables in s, that are unobserved by the data, but part of the prior model. Thus the prior distribution $P(s)$ may posses a hierarchical structure that summarizes our knowledge about the system prior to the measurement, and s represents everything in the system that is of interest to us, but about which our knowledge is uncertain a priori. We do not put any constraints on the functional form of $P(s)$, and assume that the signal s solely consists of continuous real valued variables, i.E. $s \in X \subset \mathbb{R}^M$. This enables us to regard s as coordinates of the space on which $P(s)$ is defined and to use geometric concepts such as coordinate transformations to represent probability distributions in different coordinate systems. Probability densities transform in a probability mass preserving fashion. Specifically let $f : \mathbb{R}^M \to X$ be an invertible function, and let $s = f(\xi)$. Then the distributions $P(s)$ and $P(\xi)$ relate via

$$\int P(s)\ \mathrm{d}s = \int P(\xi)\ \mathrm{d}\xi\ . \tag{5.1}$$

This allows us to express $P(s)$ by means of the pushforward of $P(\xi)$ by f. We denote the pushforward as

$$P(s) = (f \star P(\xi))\,(s) = \int \delta(s - f(\xi))\ P(\xi)\ \mathrm{d}\xi = \left(P(\xi)\ \left|\left| \frac{\mathrm{d}f}{\mathrm{d}\xi} \right|\right|^{-1} \right)\Bigg|_{\xi = f^{-1}(s)} . \tag{5.2}$$

Under mild regularity conditions on the prior distribution, there always exists an f that relates the complex hierarchical form of $P(s)$ to a simple distribution $P(\xi)$ [17]. We choose f such that $P(\xi)$ takes the form of a normal distribution with zero mean and unit covariance and call such a distribution a *standard distribution*:

$$P(\xi) = \mathcal{N}(\xi; 0, \mathbb{1})\ , \tag{5.3}$$

where $\mathcal{N}(\xi; m, D)$ denotes a multivariate normal distribution in the random variables ξ with mean m and covariance D.

We may express the likelihood in terms of ξ as

$$P(d|\xi) \equiv P(d|s' = f'(\xi))\ , \tag{5.4}$$

where f' is the part of f that maps onto the observed quantities s'. In general, f' is a non-invertible function and is commonly referred to as *generative model* or *generative process*, as it encodes all the information necessary to transform a standard distribution into the observed quantities, subject to our prior beliefs. Using equation (5.4) we get by means of Bayes' theorem, that the posterior takes the form

$$P(\xi|d) = \frac{P(\xi, d)}{P(d)} = \frac{P(d|\xi)\ \mathcal{N}(\xi; 0, \mathbb{1})}{P(d)}\ . \tag{5.5}$$

Using the push-forward of the posterior, we can recover the posterior statistics of s via

$$P(s|d) = (f \star P(\xi|d))(s) \ , \tag{5.6}$$

which means that we can fully recover the posterior properties of s, which typically has a physical interpretation as opposed to ξ. In particular equation (5.6) implies that if we are able to draw samples from $P(\xi|d)$ we can simply generate posterior samples for s since $s = f(\xi)$.

5.2 Geometric properties of posterior distributions

In order to access the information contained in the posterior distribution $P(\xi|d)$, in this work, we wish to exploit the geometric properties of the posterior, in particular with the help of Riemannian geometry. Specifically, we define a Riemannian manifold using a metric tensor, related to the Fisher Information metric of the likelihood and a metric for the prior, and establish a (local) isometry of this manifold to Euclidean space. The associated coordinate transformation gives rise to a coordinate system in which, hopefully, the posterior takes a simplified form despite the fact that probabilities do not transform in the same way as metric spaces do. As we will see, in cases where the isometry is global, and in addition the transformation is (almost) volume-preserving, the complexity of the posterior distribution can be absorbed (almost) entirely into this transformation.

To begin our discussion, we need to define an appropriate metric for posterior distributions. To this end, consider the negative logarithm of the posterior, sometimes also referred to as information Hamiltonian, which takes the form

$$\mathcal{H}(\xi|d) \equiv -\log(P(\xi|d)) = \mathcal{H}(d|\xi) + \mathcal{H}(\xi) - \mathcal{H}(d) \ . \tag{5.7}$$

A common choice to extract geometric information from this Hamiltonian is the Hessian $\mathcal{C}$ of $\mathcal{H}$. Specifically

$$\mathcal{C}(\xi) \equiv \frac{\partial^2 \mathcal{H}(\xi|d)}{\partial\xi\partial\xi'} = \frac{\partial^2 \mathcal{H}(d|\xi)}{\partial\xi\partial\xi'} + \mathbb{1} \equiv \mathcal{C}_{d|\xi}(\xi) + \mathbb{1} \ , \tag{5.8}$$

where the identity matrix arises from the curvature of the prior (information Hamiltonian). While $\mathcal{C}$ provides information about the local geometry, it turns out to be unsuited for our approach to construct a coordinate transformation, as it is not guaranteed to be positive definite for all ξ. An alternative, positive definite, measure for the curvature can be obtained by replacing the Hessian of the likelihood with its Fisher information metric [42], defined as

$$\mathcal{M}_{d|\xi}(\xi) = \left\langle \frac{\partial\mathcal{H}}{\partial\xi}\frac{\partial\mathcal{H}}{\partial\xi'} \right\rangle_{P(d|\xi)} = \left\langle \frac{\partial^2\mathcal{H}(d|\xi)}{\partial\xi\partial\xi'} \right\rangle_{P(d|\xi)} = \left\langle \mathcal{C}_{d|\xi}(\xi) \right\rangle_{P(d|\xi)} \ . \tag{5.9}$$

The Fisher information metric can be understood as a Riemannian metric defined over the statistical manifold associated with the likelihood [93], and is a core element in the

field of information geometry [4] as it provides a distance measure between probability distributions [20]. Replacing $\mathcal{C}_{d|\xi}$ with $\mathcal{M}_{d|\xi}$ in equation (5.8) we find

$$\mathcal{M}(\xi) \equiv \mathcal{M}_{d|\xi}(\xi) + \mathbb{1} = \left\langle \mathcal{C}_{d|\xi}(\xi) \right\rangle_{P(d|\xi)} + \mathbb{1} = \langle \mathcal{C}(\xi) \rangle_{P(d|\xi)} \ , \tag{5.10}$$

which, from now on, we refer to as the metric $\mathcal{M}$. As the Fisher metric of the likelihood is a symmetric, positive-semidefinite matrix, we get that $\mathcal{M}$ is a symmetric, positive-definite matrix for all ξ. It is noteworthy that upon insertion, we find that the metric $\mathcal{M}$ is defined as the expectation value of the Hessian of the posterior Hamiltonian $\mathcal{C}$ w.r.t. the likelihood $P(d|\xi)$. Therefore, in some way, we may regard $\mathcal{M}$ as the measure for the curvature in case the observed data d is unknown, and the only information given is the structure of the model itself, as encoded in $P(d|\xi)$. This connection is only of qualitative nature, but it highlights a key limitation of $\mathcal{M}$ when used as the defining property of the posterior geometry. From a Bayesian perspective, only the data d that is actually observed is of relevance as the posterior is conditioned on d. Therefore a curvature measure that arises from marginalization over the data must be sub-optimal compared to a measure conditional to the data, as it ignores the local information that we gain from observing d. Nevertheless, we find that in many practical applications $\mathcal{M}$ encodes enough relevant information about the posterior geometry that it provides us with a valuable metric to construct a coordinate transformation. It is noteworthy that attempts have been provided to resolve this issue via a more direct approach to recover a positive definite matrix from the Hessian of the posterior while retaining the local information of the data. E.g. in [11], the SoftAbs non-linearity is applied to the Hessian and the resulting positive definite matrix is used as a curvature measure. In our practical applications, however, we are particularly interested in solving very high dimensional problems, and applying a matrix non-linearity is currently too expensive to give rise to a scalable algorithm for our purposes. Therefore we rely on the metric $\mathcal{M}$ as a measure for the curvature of the posterior, and leave possible extensions to future research.

5.2.1 Coordinate transformation

Our goal is to construct a coordinate system y and an associated transformation g, that maps from ξ to y, in which the posterior metric $\mathcal{M}$ takes the form of the identity matrix $\mathbb{1}$. The motivation is that if $\mathcal{M}$ captures the geometric properties of the posterior, a coordinate system in which this metric becomes trivial should also be a coordinate system in which the posterior distribution takes a particularly simple form. For an illustrative example see figure 5.1. To do so, we require the Fisher metric of the likelihood $\mathcal{M}_{d|\xi}$ to be the pullback of the Euclidean metric. Specifically we propose a function $x(\xi)$ such that

$$\mathcal{M}_{d|\xi} \overset{!}{=} \left(\frac{\partial x}{\partial \xi}\right)^T \frac{\partial x}{\partial \xi} \ , \tag{5.11}$$

where T denotes the adjoint of a matrix. As outlined in Appendix 5.A, for many practically relevant likelihoods such a decomposition is possible by means of an inexpensive to evaluate

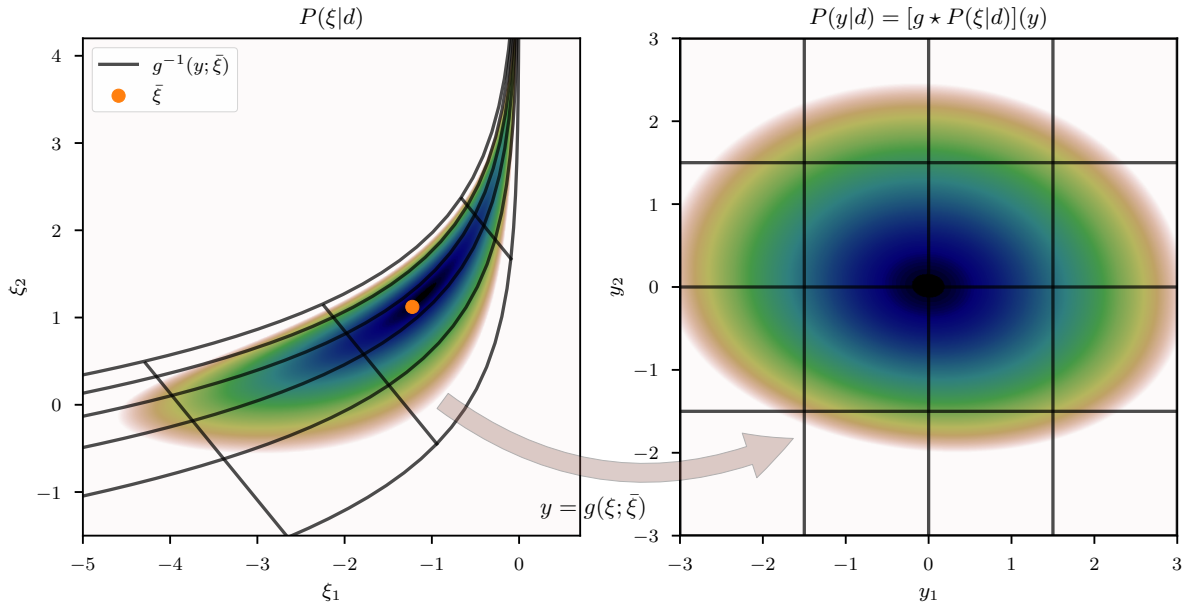

Figure 5.1: Non-linear posterior distribution $P(\xi|d)$ in the standard coordinate system of the prior distribution ξ (left) and the transformed distribution $P(y|d)$ (right) in the coordinate system y where the posterior metric becomes (approximately) the identity matrix. $P(y|d)$ is obtained from $P(\xi|d)$ via the push-forward through the transformation g which relates the two coordinate systems. The functional form of g is derived in section 5.2.1 and depends on an expansion point $\bar{\xi}$ (orange dot in the left image), and g is set up such that $\bar{\xi}$ coincides with the origin in y. To visualize the transformation, the coordinate lines of y (black mesh grid on the right) are transformed back into ξ-coordinates using the inverse coordinate transformation g^{-1} and are displayed as a black mesh in the original space on the left. In addition, note that while the transformed posterior $P(y|d)$ arguably takes a simpler form compared to $P(\xi|d)$, it does not become trivial (e.g. identical to a standard distribution) as there remain small asymmetries in the posterior density. There are multiple reasons for these deviations which are discussed in more detail in section 5.2.2 once we established how the transformation g is constructed.

function x.[1] Given x, we can rewrite the posterior metric $\mathcal{M}$ as

$$\mathcal{M} = \left(\frac{\partial x}{\partial \xi}\right)^T \frac{\partial x}{\partial \xi} + \mathbb{1} \ . \tag{5.12}$$

In order to relate this metric to Euclidean space, we aim to find the isometry g that relates the Riemannian manifold associated with the metric $\mathcal{M}$ to Euclidean space. Specifically we seek to find an invertible function g satisfying

$$\mathcal{M}(\xi) = \left(\frac{\partial x}{\partial \xi}\right)^T \frac{\partial x}{\partial \xi} + \mathbb{1} \stackrel{!}{=} \left(\frac{\partial g}{\partial \xi}\right)^T \frac{\partial g}{\partial \xi} \ . \tag{5.13}$$

In general, i.E. for a general function $x(\xi)$, however, this decomposition does not exist globally. Nevertheless, there exists a transformation $g(\xi;\bar{\xi})$ based on an approximative Taylor series around an expansion point $\bar{\xi}$, that results in a metric $\tilde{\mathcal{M}}(\xi)$ such that

$$\mathcal{M}(\xi) \approx \tilde{\mathcal{M}}(\xi) \equiv \left(\frac{\partial g(\xi;\bar{\xi})}{\partial \xi}\right)^T \frac{\partial g(\xi;\bar{\xi})}{\partial \xi} \ , \tag{5.14}$$

in the vicinity of $\bar{\xi}$. This transformation g can be obtained up to an integration constant by Taylor expanding equation (5.14) around $\bar{\xi}$ and solving for the Taylor coefficients of g in increasing order. We express g in terms of its Taylor series using the Einstein sum convention as

$$g(\xi;\bar{\xi})^i = \bar{g}^i + \bar{g}^i{}_{,j} \left(\xi - \bar{\xi}\right)^j + \bar{g}^i{}_{,jk} \left(\xi - \bar{\xi}\right)^j \left(\xi - \bar{\xi}\right)^k + ... \ , \tag{5.15}$$

where repeated indices get summed over, $a_{,i}$ denotes the partial derivative of a w.r.t. the ith component of ξ, and $\bar{s}$ denotes a (tensor) field $s(\xi)$, evaluated at the expansion point $\bar{\xi}$. We begin to expand equation (5.14) around $\bar{\xi}$ and obtain for the zeroth order

$$\bar{\mathcal{M}}_{ij} \equiv \bar{x}^\alpha{}_{,i} \bar{x}^\alpha{}_{,j} + \delta_{ij} \stackrel{!}{=} \bar{g}^\alpha{}_{,i} \ \bar{g}^\alpha{}_{,j} \ . \tag{5.16}$$

Expanding equation (5.14) to first order yields

$$\bar{x}^\alpha{}_{,ik} \bar{x}^\alpha{}_{,j} + \bar{x}^\alpha{}_{,i} \bar{x}^\alpha{}_{,jk} \stackrel{!}{=} \bar{g}^\alpha{}_{,ik} \bar{g}^\alpha{}_{,j} + \bar{g}^\alpha{}_{,i} \bar{g}^\alpha{}_{,jk} \ , \tag{5.17}$$

and therefore

$$\bar{g}^i{}_{,jk} = \bar{\mathcal{M}}^{i\gamma} \bar{g}^\beta{}_{,\gamma} \bar{x}^\alpha{}_{,\beta} \bar{x}^\alpha{}_{,jk} \ , \tag{5.18}$$

where $\bar{\mathcal{M}}^{ij} = \left(\bar{\mathcal{M}}^{-1}\right)_{ij}$ denotes the components of the inverse of $\bar{\mathcal{M}}$.

[1] Here with "inexpensive" we mean that applying the function $x(\xi)$ has a similar computational cost compared to applying the likelihood function $P(d|\xi)$ to a specific ξ.

Thus, to first order in the metric (meaning to second order in the transformation) the expansion remains solvable for a general x. Proceeding with the second order, however, we get that

$$\bar{x}^{\alpha}{}_{,ikl}\bar{x}^{\alpha}{}_{,j} + \bar{x}^{\alpha}{}_{,ik}\bar{x}^{\alpha}{}_{,jl} + \bar{x}^{\alpha}{}_{,il}\bar{x}^{\alpha}{}_{,jk} + \bar{x}^{\alpha}{}_{,i}\bar{x}^{\alpha}{}_{,jkl} \stackrel{!}{=} \bar{g}^{\alpha}{}_{,ikl}\bar{g}^{\alpha}{}_{,j} + \bar{g}^{\alpha}{}_{,ik}\bar{g}^{\alpha}{}_{,jl} + \bar{g}^{\alpha}{}_{,il}\bar{g}^{\alpha}{}_{,jk} + \bar{g}^{\alpha}{}_{,i}\bar{g}^{\alpha}{}_{,jkl} \ , \quad (5.19)$$

which does not exhibit a general solution for $\bar{g}^{i}{}_{,jkl}$ in higher dimensions due to the fact that the third derivative has to be invariant under arbitrary permutation of the latter three indices jkl. However, in analogy to equation (5.18), we may set

$$\bar{g}^{i}{}_{,jkl} = \bar{\mathcal{M}}^{i\gamma}\bar{g}^{\beta}{}_{,\gamma}\bar{x}^{\alpha}{}_{,\beta}\bar{x}^{\alpha}{}_{,jkl} \ , \quad (5.20)$$

which cancels the first and the last term of equation (5.19), and study the remaining error which takes the form

$$\begin{aligned}
\epsilon_{ijkl} &= \bar{x}^{\alpha}{}_{,ik}\bar{x}^{\alpha}{}_{,jl} + \bar{x}^{\alpha}{}_{,il}\bar{x}^{\alpha}{}_{,jk} - \bar{g}^{\alpha}{}_{,ik}\bar{g}^{\alpha}{}_{,jl} - \bar{g}^{\alpha}{}_{,il}\bar{g}^{\alpha}{}_{,jk} \\
&= \bar{x}^{\alpha}{}_{,ik}\bar{x}^{\alpha}{}_{,jl} + \bar{x}^{\alpha}{}_{,il}\bar{x}^{\alpha}{}_{,jk} - \bar{x}^{\alpha}{}_{,ik}\bar{x}^{\gamma}{}_{,\alpha}\bar{\mathcal{M}}^{\gamma\delta}\bar{x}^{\delta}{}_{,\beta}\bar{x}^{\beta}{}_{,jl} - \bar{x}^{\alpha}{}_{,il}\bar{x}^{\gamma}{}_{,\alpha}\bar{\mathcal{M}}^{\gamma\delta}\bar{x}^{\delta}{}_{,\beta}\bar{x}^{\beta}{}_{,jk} \\
&= \bar{x}^{\alpha}{}_{,ik}\left(\delta_{\alpha\beta} - \bar{x}^{\gamma}{}_{,\alpha}\bar{\mathcal{M}}^{\gamma\delta}\bar{x}^{\delta}{}_{,\beta}\right)\bar{x}^{\beta}{}_{,jl} + \bar{x}^{\alpha}{}_{,il}\left(\delta_{\alpha\beta} - \bar{x}^{\gamma}{}_{,\alpha}\bar{\mathcal{M}}^{\gamma\delta}\bar{x}^{\delta}{}_{,\beta}\right)\bar{x}^{\beta}{}_{,jk} \ . \quad (5.21)
\end{aligned}$$

Let $X^{i}{}_{j} \equiv \bar{x}^{i}{}_{,j}$, the expression in the parentheses takes the form

$$\mathbb{1} - X\bar{\mathcal{M}}^{-1}X^{T} = \mathbb{1} - X\left(\mathbb{1} + X^{T}X\right)^{-1}X^{T} = \left(\mathbb{1} + XX^{T}\right)^{-1} \equiv M \ , \quad (5.22)$$

and thus equation (5.21) reduces to

$$\epsilon_{ijkl} = \bar{x}^{\alpha}{}_{,ik}M_{\alpha\beta}\bar{x}^{\beta}{}_{,jl} + \bar{x}^{\alpha}{}_{,il}M_{\alpha\beta}\bar{x}^{\beta}{}_{,jk} \ . \quad (5.23)$$

The impact of this error contribution can be qualitatively studied using the spectrum $\lambda(M)$ of the matrix M. This spectrum may exhibit two extreme cases, a so-called likelihood dominated regime, where the spectrum $\lambda(XX^{T}) \gg 1$, and a prior dominated regime where $\lambda(XX^{T}) \ll 1$. In the likelihood dominated regime, we get that $\lambda(M) \ll 1$ and thus the contribution of the error is small, whereas in the prior dominated regime $\lambda(M) \approx 1$ which yields an $\mathcal{O}(1)$ error. However, in the prior dominated regime, the entire metric $\mathcal{M}$ is close to the identity as we are in the standard coordinate system of the prior ξ and therefore higher order derivatives of x are small. As a consequence, the error is of the order $\mathcal{O}(1)$ only in regimes where the third (and higher) order of the expansion is negligible compared to the first and second order. An exception occurs when the expansion point is close to a saddle point of x, i.E. in cases where the first derivative of x becomes small (and therefore the metric is close to the identity), but higher order derivatives of x may be large. For the moment, we proceed under the assumption that the change of x, as a function of ξ, is sufficiently monotonic throughout the expansion regime. We discuss the implications of violating this assumption in section 5.5.3.

If we proceed to higher order expansions of equation (5.14), we notice that a repetitive picture emerges: The leading order derivative tensor $\bar{g}^i{}_{,j\ldots}$ that appears in the expansion may be set in direct analogy to equations (5.18) and (5.20) as

$$\bar{g}^i{}_{,j\ldots} = \bar{\mathcal{M}}^{i\gamma}\bar{g}^\beta{}_{,\gamma}\bar{x}^\alpha{}_{,\beta}\bar{x}^\alpha{}_{,j\ldots} \,, \tag{5.24}$$

where ... denotes the higher order derivatives. The remaining error contributions at each order take a similar form as in equation (5.23), where the matrix M reappears in between all possible combinations of the remaining derivatives of x that appear using the Leibniz rule. Note that for increasing order, the number of terms that contribute to the error also increases. Specifically for the nth order expansion of equation (5.14) we get $m = \sum_{k=1}^{n-1} \binom{n}{k}$ contributions to the error. Therefore, even if each individual contribution by means of M is small, the expansion error eventually becomes large once high order expansions become relevant. Therefore the proposed approximation only remains locally valid around $\bar{\xi}$.

Nevertheless, we may proceed to combine the derivative tensors of g determined above in order to get the Jacobian of the transformation g as

$$\begin{aligned} g^i{}_{,j}(\xi) &\equiv \bar{g}^i{}_{,j} + \bar{g}^i{}_{,jk}\left(\xi-\bar{\xi}\right)^k + \frac{1}{2}\bar{g}^i{}_{,jkl}\left(\xi-\bar{\xi}\right)^k\left(\xi-\bar{\xi}\right)^l + \ldots \\ &= \bar{g}^i{}_{,j} + \bar{\mathcal{M}}^{i\alpha}\bar{g}^\beta{}_{,\alpha}\bar{x}^\gamma{}_{,\beta}\left(\bar{x}^\gamma{}_{,jk}(\xi-\bar{\xi})^k + \frac{1}{2}\bar{x}^\gamma{}_{,jkl}(\xi-\bar{\xi})^k(\xi-\bar{\xi})^l + \ldots\right) \,, \end{aligned} \tag{5.25}$$

or equivalently

$$\bar{g}^\alpha{}_{,i}g^\alpha{}_{,j}(\xi) = \delta_{ij} + \bar{x}^\alpha{}_{,i}\left(\bar{x}^\alpha{}_{,j} + \bar{x}^\alpha{}_{,jk}(\xi-\bar{\xi})^k + \frac{1}{2}\bar{x}^\alpha{}_{,jkl}(\xi-\bar{\xi})^k(\xi-\bar{\xi})^l + \ldots\right) \,. \tag{5.26}$$

From the zeroth order, equation (5.16), we get that $\bar{g}^i{}_{,j} = \left(\sqrt{\bar{\mathcal{M}}}\right)^i_j$ up to a unitary transformation, and we can sum up the Taylor series in x of equation (5.26) to arrive at an index free representation of the Jacobian as

$$\frac{\partial g}{\partial \xi} = \sqrt{\bar{\mathcal{M}}}^{-1}\left(\mathbb{1} + \left(\frac{\partial x}{\partial \xi}\bigg|_{\bar{\xi}}\right)^T\frac{\partial x}{\partial \xi}\right) \,. \tag{5.27}$$

Upon integration, this yields a transformation

$$g(\xi) - g(\bar{\xi}) = \sqrt{\bar{\mathcal{M}}}^{-1}\left(\xi - \bar{\xi} + \left(\frac{\partial x}{\partial \xi}\bigg|_{\bar{\xi}}\right)^T\left(x(\xi) - x(\bar{\xi})\right)\right) \,. \tag{5.28}$$

The resulting transformation takes an intuitive form: The approximation to the distance between a point $g(\xi)$ and the transformed expansion point $g(\bar{\xi})$ consists of the distance w.r.t. the prior measure $\left(\xi - \bar{\xi}\right)$ and the distance w.r.t. the likelihood measure

$\left(x(\xi) - x(\bar{\xi})\right)$, back-projected into the prior domain using the local transformation at $\bar{\xi}$. Finally, the metric at $\bar{\xi}$ is used as a measure for the local curvature. Equation (5.28) is only defined up to an integration constant, and therefore, without loss of generality, we may set $g(\bar{\xi}) = 0$ to obtain the final approximative coordinate transformation as

$$g(\xi;\bar{\xi}) = \sqrt{\bar{\mathcal{M}}}^{-1}\left(\xi - \bar{\xi} + \left(\frac{\partial x}{\partial \xi}\Big|_{\bar{\xi}}\right)^T \left(x(\xi) - x(\bar{\xi})\right)\right) \equiv \sqrt{\bar{\mathcal{M}}}^{-1}\,\tilde{g}(\xi;\bar{\xi}) \ . \tag{5.29}$$

5.2.2 Basic properties

In order to study a few basic properties of this transformation, for simplicity, we first consider a posterior distribution with a metric that allows for an exact isometry g_{iso} to Euclidean space. Specifically let g_{iso} be a coordinate transformation satisfying equation (5.13). The posterior distribution in coordinates $y = g_{\text{iso}}(\xi)$ is given via the push-forward of $P(\xi|d)$ through g_{iso} as

$$P(y|d) \propto \left(g_{\text{iso}} \star P(\xi|d)\right)(y) = \left(P(\xi|d)\ \left|\left|\frac{\partial g_{\text{iso}}}{\partial \xi}\right|\right|^{-1}\right)\Bigg|_{\xi = g_{\text{iso}}^{-1}(y)} = \frac{P(\xi|d)}{\sqrt{|\mathcal{M}(\xi)|}}\Bigg|_{\xi = g_{\text{iso}}^{-1}(y)} \ , \tag{5.30}$$

and the information Hamiltonian takes the form

$$\mathcal{H}(y|d) = \left(\mathcal{H}(\xi|d) + \frac{1}{2}\log\left(|\mathcal{M}(\xi)|\right)\right)\Big|_{\xi = g_{iso}^{-1}(y)} + \mathcal{H}_0 \equiv \tilde{\mathcal{H}}(\xi = g_{iso}^{-1}(y)) + \mathcal{H}_0 \ , \tag{5.31}$$

where $\mathcal{H}_0$ denotes y independent contributions. We may study the curvature of the posterior in coordinates y given as:

$$\mathcal{C}(y) = \frac{\partial \xi}{\partial y}\left(\frac{\partial^2 \tilde{\mathcal{H}}(\xi)}{\partial \xi \partial \xi'}\right)\left(\frac{\partial \xi'}{\partial y'}\right)^T + \frac{\partial \tilde{\mathcal{H}}(\xi)}{\partial \xi}\frac{\partial^2 \xi}{\partial y \partial y'} \quad \text{with} \quad \xi = g_{\text{iso}}^{-1}(y) \quad \text{and} \quad \xi' = g_{\text{iso}}^{-1}(y') \ , \tag{5.32}$$

which we can use to construct a metric $\mathcal{M}(y)$ in analogy to equation (5.10) by taking the expectation value of the curvature w.r.t. the likelihood. This yields

$$\begin{aligned}\mathcal{M}(y) &= \frac{\partial \xi}{\partial y}\mathcal{M}(\xi)\left(\frac{\partial \xi'}{\partial y'}\right)^T + \frac{1}{2}\frac{\partial \xi}{\partial y}\left(\frac{\partial^2 \log\left(|\mathcal{M}(\xi)|\right)}{\partial \xi \partial \xi'}\right)\left(\frac{\partial \xi'}{\partial y'}\right)^T + \left\langle \frac{\partial \tilde{\mathcal{H}}(\xi)}{\partial \xi}\right\rangle_{P(d|\xi)}\frac{\partial^2 \xi}{\partial y \partial y'} \\ &\equiv \frac{\partial \xi}{\partial y}\mathcal{M}(\xi)\left(\frac{\partial \xi'}{\partial y'}\right)^T + \mathcal{R}(y) = \mathbb{1} + \mathcal{R}(y) \ .\end{aligned} \tag{5.33}$$

The first terms yields the identity, as it is the defining property of g_{iso}. Furthermore, in case we are able to say that $\mathcal{R}(y)$ is small compared to the identity, we notice that the quantity $\mathcal{M}(\xi)$ (equation (5.10)), that we referred to as the posterior metric, approximately transforms like a proper metric under g_{iso}. In this case we find that the isometry g_{iso} between the Riemannian manifold associated with $\mathcal{M}(\xi)$ and the Euclidean space is also a transformation that removes the complex geometry of the posterior. To further study $\mathcal{R}(y)$, we consider its two contributions separately, where for the first part, the log-determinant (or logarithmic volume), we get that it becomes small compared to the identity if

$$\mathcal{M}(\xi) \overset{!}{\gg} \frac{1}{2}\left(\frac{\partial^2 \log\left(|\mathcal{M}(\xi)|\right)}{\partial\xi\partial\xi'}\right) . \tag{5.34}$$

Therefore, the curvature of the log-determinant of the metric has to be much smaller then the metric itself. To study the second term of $\mathcal{R}(y)$, we may again split the discussion into a prior and a likelihood dominated regime, depending on the ξ at which we evaluate the expression. In a prior dominated regime we get that

$$\frac{\partial^2 \xi}{\partial y \partial y'} \approx 0 , \tag{5.35}$$

as the metric is close to the identity in this regime (and therefore $\xi \approx y$). In a likelihood dominated regime we get that $\tilde{\mathcal{H}} \approx \mathcal{H}(d|\xi)$ and therefore

$$\left\langle \frac{\partial\tilde{\mathcal{H}}(\xi)}{\partial\xi} \right\rangle_{P(d|\xi)} \approx \left\langle \frac{\partial\mathcal{H}(d|\xi)}{\partial\xi} \right\rangle_{P(d|\xi)} = -\left\langle \frac{1}{P(d|\xi)} \frac{\partial P(d|\xi)}{\partial\xi} \right\rangle_{P(d|\xi)} = 0 . \tag{5.36}$$

So at least in a prior dominated regime, as well as a likelihood dominated regime, the posterior Hamiltonian transforms in an analogous way as the manifold, under the transformation g_{iso}, if equation (5.34) also holds true.

For a practical application, however, in all but the simplest cases the isometry g_{iso} is not accessible, or might not even exist. Therefore, in general, we have to use the approximation $g(\xi;\bar{\xi})$, as defined in equation (5.29), instead. We may express the transformation of the metric using $g(\xi;\bar{\xi})$, and find that

$$\begin{aligned}\mathcal{M}(y) = \sqrt{\bar{\mathcal{M}}} \left(\mathbb{1} + \left(\frac{\partial x}{\partial\xi}\right)^T \frac{\partial x}{\partial\xi}\bigg|_{\bar{\xi}}\right)^{-1} \left(\mathbb{1} + \left(\frac{\partial x}{\partial\xi}\right)^T \frac{\partial x'}{\partial\xi'}\right) \left(\mathbb{1} + \left(\frac{\partial x}{\partial\xi}\bigg|_{\bar{\xi}}\right)^T \frac{\partial x'}{\partial\xi'}\right)^{-1} \sqrt{\bar{\mathcal{M}}} \\ + \tilde{\mathcal{R}}(y) ,\end{aligned} \tag{5.37}$$

now with $\xi = g^{-1}\left(y;\bar{\xi}\right)$ and analogous for ξ'. $\tilde{\mathcal{R}}$ is defined by replacing g_{iso} with g for the entire expression of $\mathcal{R}$. We notice that this transformation does not yield the identity, except when evaluated at the expansion point $\xi = \bar{\xi}$. Therefore, in addition to the error $\tilde{\mathcal{R}}$ there is a deviation from the identity related to the expansion error as one moves away from $\bar{\xi}$.

At this point we would like to emphasize that the posterior Hamiltonian $\mathcal{H}$ and the Riemannian manifold constructed from $\mathcal{M}$ are only loosely connected due to the errors described by $\tilde{\mathcal{R}}$ and the additional expansion error. They are arguably small in many cases and in the vicinity of $\bar{\xi}$, but we do not want to claim that this correspondence is valid in general (see section 5.5.3). Nevertheless, we find that in many cases this correspondence works well in practice. Some illustrative examples are given in section 11.

5.3 Posterior approximation

Utilizing the derived coordinate transformation for posterior approximation is mainly based on the idea that in the transformed coordinate system, the posterior takes a simpler form. In particular we aim to remove parts (if not most) of the complex geometry of the posterior, such that a simple probability distribution, e.g. a Gaussian distribution, yields a good approximation.

5.3.1 Direct approximation

Assuming that all the errors discussed in the previous section are small enough, we may attempt to directly approximate the posterior distribution via a unit Gaussian in the coordinates y as in this case the transformed metric $\mathcal{M}(y)$ is close to the identity. As the coordinate transformation g, defined via equation (5.29), is only known up to an integration constant by construction, the posterior approximation is achieved by a shifted unit Gaussian in y. This shift needs to be determined, which we can do by maximizing the transformed posterior distribution

$$P(y|d) \propto \left(P(\xi, d) \left| \left| \frac{\partial g}{\partial \xi} \right| \right|^{-1} \right) \Bigg|_{\xi = g^{-1}(y;\bar{\xi})} , \tag{5.38}$$

w.r.t. y. Here $g^{-1}(y;\bar{\xi})$ denotes the inverse of $g(\xi;\bar{\xi})$ w.r.t. its first argument. Equivalently we can minimize the information Hamiltonian $\mathcal{H}(y|d)$, defined as

$$\mathcal{H}(y|d) \equiv -\log\left(P(y|d)\right) = \left(\mathcal{H}(\xi, d) + \frac{1}{2} \log\left(\left| \tilde{\mathcal{M}} \right| \right) \right) \Bigg|_{\xi = g^{-1}(y;\bar{\xi})} \equiv \tilde{\mathcal{H}}(\xi = g^{-1}(y;\bar{\xi})) \, . \tag{5.39}$$

Minimizing $\mathcal{H}(y|d)$ yields the maximum a posterior solution y^* which, in case the posterior is close to a unit Gaussian in the coordinates y, can be identified with the shift in y. As g is an invertible function, we may instead minimize $\tilde{\mathcal{H}}$ w.r.t. ξ and apply g to the result in order to obtain y^*. Specifically

$$y^* \equiv \underset{y}{\operatorname{argmin}} \left(\mathcal{H}(y|d) \right) = g\left(\underset{\xi}{\operatorname{argmin}} \left(\tilde{\mathcal{H}}(\xi) \right) \right) \, . \tag{5.40}$$

Therefore we can circumvent the inversion of g at any point during optimization. Now suppose that we use any gradient based optimization scheme to minimize for ξ, starting from some initial position ξ^0. If we set the expansion point $\bar{\xi}$, used to construct g, to be equal to ξ^0, we notice that

$$\tilde{\mathcal{M}}(\bar{\xi}) = \mathcal{M}(\bar{\xi}) \tag{5.41}$$

$$\left.\frac{\partial \tilde{\mathcal{M}}}{\partial \xi}\right|_{\xi=\bar{\xi}} = \left.\frac{\partial \mathcal{M}}{\partial \xi}\right|_{\xi=\bar{\xi}} , \tag{5.42}$$

as the expansion of the metric is valid to first order by construction. Therefore if we set the expansion point $\bar{\xi}$ to the current estimate of ξ after every step, we can replace the approximated metric $\tilde{\mathcal{M}}$ with the true metric $\mathcal{M}$ and arrive at an optimization objective of the form

$$\bar{\xi} = \underset{\xi}{\operatorname{argmin}} \left(\mathcal{H}(\xi, d) + \frac{1}{2} \log \left(|\mathcal{M}(\xi)| \right) \right) . \tag{5.43}$$

Note that $g(\bar{\xi}; \bar{\xi}) = 0$ by construction, and therefore $y^* = 0$, as there is a degeneracy between a shift in y and a change of the expansion point $\bar{\xi}$. Once the optimal expansion point $\bar{\xi}$ is found, we directly retrieve a generative process to sample from our approximation to the posterior distribution. Specifically

$$P(y|d) \approx \mathcal{N}(y; 0, \mathbb{1}) \tag{5.44}$$

$$\xi = g^{-1}(y; \bar{\xi}) , \tag{5.45}$$

where g^{-1} is only implicitly defined using equation (5.29) and therefore its inverse application has to be approximated numerically in general.

Numerical approximation to sampling

Recall that

$$y = g(\xi; \bar{\xi}) = \sqrt{\bar{\mathcal{M}}}^{-1} \tilde{g}(\xi; \bar{\xi}) . \tag{5.46}$$

To generate a posterior sample for ξ we have to draw a random realization for y from a unit Gaussian, and then solve equation (5.46) for ξ. To avoid the matrix square root of $\bar{\mathcal{M}}$, we may instead define

$$z \equiv \sqrt{\bar{\mathcal{M}}}\, y = \tilde{g}(\xi; \bar{\xi}) \quad \text{with} \quad P(z) = \mathcal{N}(z; 0, \bar{\mathcal{M}}) . \tag{5.47}$$

Sampling from $P(z)$ is much more convenient then constructing the matrix square root, since

$$\bar{\mathcal{M}} = \mathbb{1} + \left(\left(\frac{\partial x}{\partial \xi} \right)^T \frac{\partial x}{\partial \xi} \right) \Bigg|_{\bar{\xi}} , \tag{5.48}$$

and therefore a random realization may be generated using

$$z = \eta_1 + \left(\frac{\partial x}{\partial \xi} \Big|_{\bar{\xi}} \right)^T \eta_2 \quad \text{with} \quad \eta_i \sim \mathcal{N}(\eta_i; 0, \mathbb{1}) \ , \ i \in \{1, 2\} \ . \tag{5.49}$$

Finally, a posterior sample ξ is retrieved by inversion of equation (5.47). We numerically approximate the inversion by minimizing the squared difference between z and $\tilde{g}(\xi)$. Specifically,

$$\xi = \underset{\xi}{\operatorname{argmin}} \left(\frac{1}{2} \left(z - \tilde{g}(\xi) \right)^T \left(z - \tilde{g}(\xi) \right) \right) \ . \tag{5.50}$$

Note that if g is invertible then also $\tilde{g}$ is invertible as $\bar{\mathcal{M}}$ is a symmetric positive definite matrix. Therefore the quadratic form of equation (5.50) has a unique global optimum at zero which corresponds to the inverse of equation (5.47).

In practice, this optimum is typically only reached approximately. For an efficient numerical approximation, throughout this work, we employ a second order quasi-Newton method, named NewtonCG [86], as implemented in the NIFTy framework. Within the NewtonCG algorithm, we utilize the metric $\tilde{\mathcal{M}}(\xi)$ as a positive-definite approximation to the curvature of the quadratic form in equation (5.50). Furthermore, its inverse application, required for the second order optimization step of NewtonCG, is approximated with the conjugate gradient (CG) [59] method, which requires the metric to be only implicitly available via matrix-vector products. In addition, in practice we find that the initial position ξ^0 of the minimization procedure can be set to be equal to the prior realization η_1 used to construct z (equation (5.49)) in order to improve convergence as $\xi = \eta_1$ is the solution of equation (5.47) for all degrees of freedom unconstrained by the likelihood. Alternatively, for weakly non-linear problems, initializing ξ^0 as the solution of the linearized problem

$$z = \tilde{g}(\xi; \bar{\xi})\Big|_{\xi=\bar{\xi}} + \frac{\partial \tilde{g}}{\partial \xi}\Big|_{\xi=\bar{\xi}} \left(\xi - \bar{\xi} \right) = \left(\mathbb{1} + \left(\frac{\partial x}{\partial \xi} \right)^T \frac{\partial x}{\partial \xi} \right) \Bigg|_{\xi=\bar{\xi}} \left(\xi - \bar{\xi} \right) \ , \tag{5.51}$$

can significantly improve the convergence. The full realization of the sampling procedure is summarized in algorithm 3.

Properties

We may qualitatively study some basic properties of the coordinate transformation and the associated approximation using illustrative one and two dimensional examples. To

Algorithm 3: Approximate posterior samples using inverse transformation

Function `drawSample(`*Location* $\bar{\xi}$*, Transformation* $x(\xi)$*, Jacobian* $\frac{\partial x}{\partial \xi}$`):`
- $A \leftarrow \left.\frac{\partial x}{\partial \xi}\right|_{\xi=\bar{\xi}}$
- $\eta_1 \sim \mathcal{N}(\eta_1; 0, \mathbb{1})$
- $\eta_2 \sim \mathcal{N}(\eta_2; 0, \mathbb{1})$
- $z \leftarrow \eta_1 + A^T \eta_2$
- $\xi^0 \leftarrow \eta_1$ or $\xi^0 \leftarrow$ `Solve(`$z = \left(\mathbb{1} + A^T A\right)\left(\xi^0 - \bar{\xi}\right)$`)` for ξ^0 (see Eq. (5.51))
- **Function** `Energy(`ξ`):`
  - $\tilde{g} \leftarrow \xi - \bar{\xi} + A^T \left(x(\xi) - x(\bar{\xi})\right)$
  - **return** $\frac{1}{2}\left(z - \tilde{g}\right)^T \left(z - \tilde{g}\right)$
- $\xi^* \leftarrow$ `NewtonCG(Energy,` ξ^0`)`
- **return** ξ^*

this end, consider a one dimensional log-normal prior model with zero mean and standard deviation σ_p of the form

$$s(\xi) = e^{\sigma_p \xi} \quad \text{with} \quad P(\xi) = \mathcal{N}(\xi; 0, 1) \ , \tag{5.52}$$

from which we obtain a measurement d subject to independent, additive Gaussian noise with standard deviation σ_n such that the likelihood takes the form

$$P(d|\xi) = \mathcal{N}(d; s(\xi), \sigma_n^2) \ . \tag{5.53}$$

The posterior distribution is given as

$$P(\xi|d) \propto P(d|\xi)\ P(\xi) = \mathcal{N}(d; s(\xi), \sigma_n^2)\ \mathcal{N}(\xi; 0, 1) \ , \tag{5.54}$$

and its metric takes the form

$$\mathcal{M}(\xi) = \left(\frac{1}{\sigma_n}\frac{\partial s(\xi)}{\partial \xi}\right)^2 + 1 = \left(\frac{\sigma_p}{\sigma_n}\right)^2 e^{2\sigma_p \xi} + 1 \ . \tag{5.55}$$

In this one dimensional example we can construct the exact transformation g_{iso} that maps from ξ to the transformed coordinates y, by integrating the square root of equation (5.55) over ξ. The resulting transformation can be seen in the central panel of figure 5.2, for an example with $\sigma_p = 3$ and $\sigma_n = 0.3$ and measured data $d = 0.5$. In addition, we depict the approximated transformation $g(\xi; \bar{\xi})$ for multiple expansion points $\bar{\xi} \in \{-1, -0.6, -0.2\}$. We see that the function approximation quality depends on the choice of the expansion point $\bar{\xi}$ as the approximation error is smallest in the vicinity of $\bar{\xi}$. In order to transform the posterior distribution P (Eq. (5.54)) into the new coordinated system, not all parts of the transformation are equally relevant and therefore different expansion points result in more/less complex transformed distributions (see top panel of figure 5.2). Finally, if

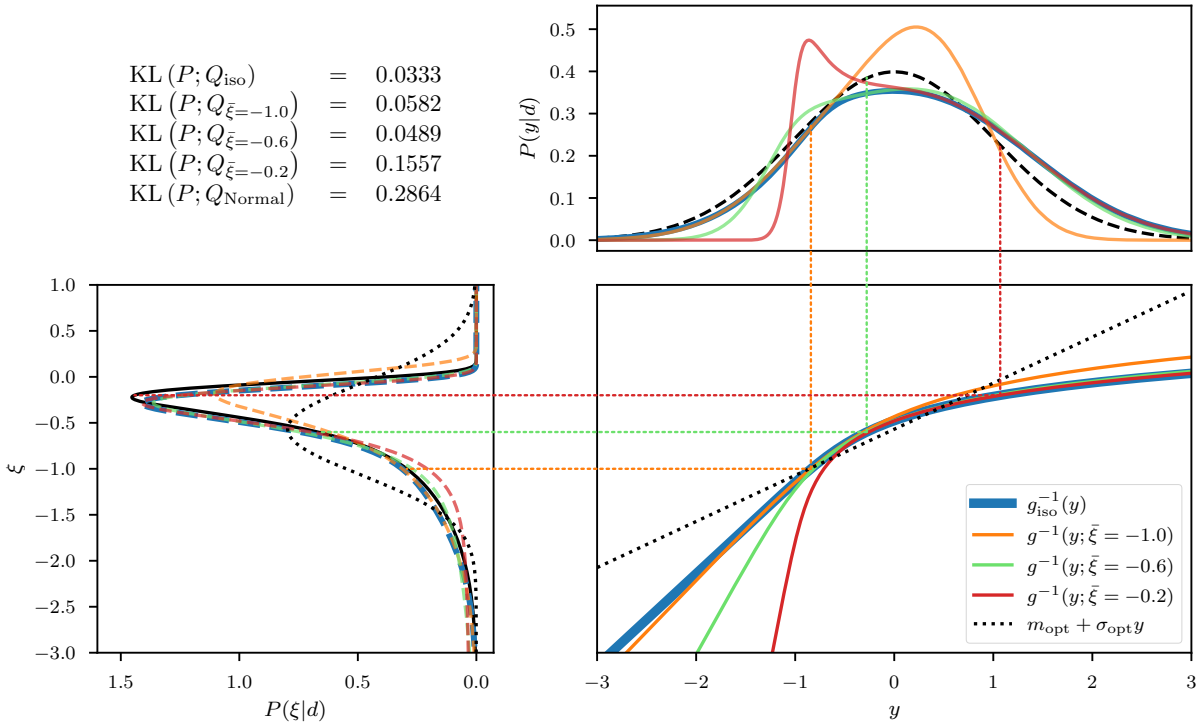

Figure 5.2: Illustration of the coordinate transformation for the one-dimensional log-normal model (equation (5.54)). The true posterior $P(\xi|d)$, displayed as the black solid line in the left panel, is transformed into the coordinate system y using the optimal transformation g_{iso} (blue), as well as three approximations g thereof with expansion points $\bar{\xi} \in \{-1, -0.6, -0.2\}$ (orange, green, red). The resulting distributions $P(y|d)$ are displayed in the top panel of the figure as solid lines, color coded according the used transformation g (or g_{iso} in case of blue). The black, dashed line in the top panel displays a standard distribution in y. The location of the expansion point $\bar{\xi}$, and its associated point in y, is highlighted via the color coded, dotted lines. Finally, the direct approximations to the posterior associated with the transformations, meaning the push-forwards of the standard distribution in y using the inverse of the various transformations g^{-1}, are displayed in the left panel as dashed lines, color coded according to their used transformation. As a comparison, the "optimal linear approximation" (black dotted line in the central panel), which corresponds to the optimal approximation of the posterior with a normal distribution in ξ (black dotted line in left panel), is displayed as a comparison. To numerically quantify the information distance between the true distribution P and its approximations $Q_\bullet$, the Kullback-Leibler (KL) divergences between P and $Q_\bullet$ are displayed in the top left of the image. The numerical values of the KL are given in nats (meaning the KL is evaluated in the basis of the natural logarithm).

we use a standard distribution in the transformed coordinates y and transform it back using the inverse transformations $g^{-1}(y;\bar{\xi})$, we find that the approximation quality of the resulting distributions $Q_{\bar{\xi}}$ depends on $\bar{\xi}$. The distributions are illustrated in the left panel of figure 5.2 together with the Kullback-Leibler divergence KL between the true posterior distribution P and the approximations $Q_{\bar{\xi}}$. We also illustrate the "geometrically optimal" approximation using a standard distribution in y and the optimal transformation g_{iso} and find that while the approximation error becomes minimal in this case, it remains non-zero. Considering the discussion in section 5.2.2, this result is to be expected due to the error contribution from the change in volume associated with the transformation g. As a comparison we also depict the optimal linear approximation of P, that is a normal distribution in the coordinates ξ with optimally chosen mean and standard deviation. We see that even the worst expansion point $\bar{\xi} = -0.2$, that is far away from the optimum, still yields a better approximation of the posterior.

As a second example we consider the task of inferring the mean m and variance v of a single, real valued Gaussian random variable d. In terms of $s = (m, v)$, the likelihood takes the form

$$P(d|s) = \mathcal{N}(d; m, v) \ . \tag{5.56}$$

Furthermore we assume a prior model for s by means of a generative model of the form

$$m = \xi_1 \quad \text{and} \quad v = \exp\left[3(\xi_2 + 2\xi_1)\right] \ , \tag{5.57}$$

where ξ_1 and ξ_2 follow standard distributions a priori. This artificial model results in a linear prior correlation between the mean and the log-variance and thus introduces a non-linear coupling between m and v. The resulting two dimensional posterior distribution $P(\xi_1, \xi_2)$ can be seen in the left panel of figure 5.3, together with the two marginals $P(\xi_1)$ and $P(\xi_2)$ for a given measurement $d = 0$. We approximate this posterior distribution following the direct approach described in section 5.3.1, where the expansion point $\bar{\xi}$ is obtained from minimizing the sum of the posterior Hamiltonian and the log-determinant of the metric (see Eq. (5.43)). The resulting approximative distribution Q_{D} is shown in the right panel of figure 5.3, where the location of $\bar{\xi}$ is indicated as a blue cross. In comparison to the true distribution, we see that both, the joint distribution as well as the marginals are in a good agreement qualitatively, which is also supported quantitatively by a small difference of the KL between P and Q_{D} (see figure 5.3). The difference between P and Q_{D} appears to increase in regions further away from the expansion point, which is to be expected due to the local nature of the approximation. However, non-linear features such as the sharp peak at the "bottom" of P (figure 5.3), are also present in Q_{D}, although slightly less prominent. This demonstrates that relevant non-linear structure can, to some degree, be captured by the coordinate transformation g derived from the metric $\mathcal{M}$ of the posterior.

Although these low-dimensional, illustrative examples appear promising, there remains one central issue left to be addressed before the approach can be applied to high-dimensional

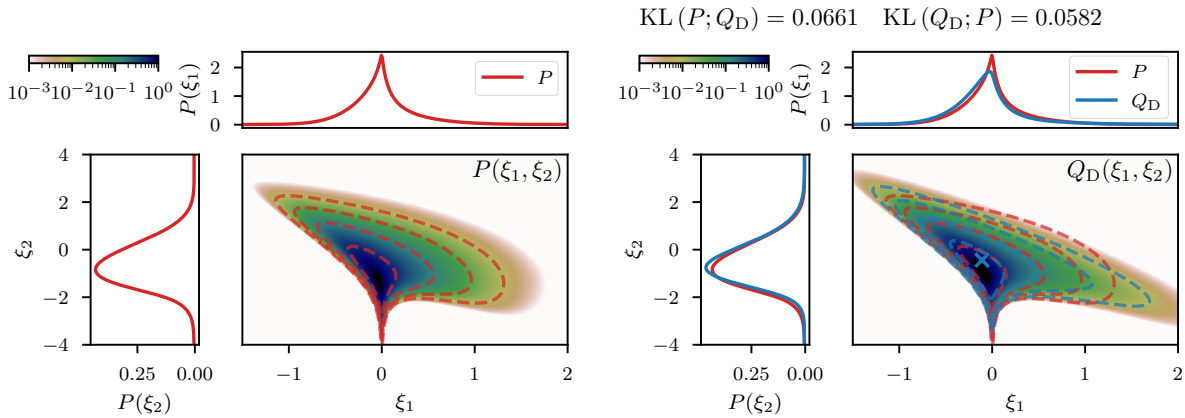

Figure 5.3: Left: posterior distribution P in the standard coordinates $\xi_{1/2}$ for the inference of the mean and variance of a normal distribution (equations (5.56) and (5.57)). The central panel shows the two dimensional density and the red dashed lines are logarithmically spaced contours. The top and left sub-panels display the marginal posterior distributions for ξ_1 and ξ_2, respectively. Right: Approximation Q_D to the posterior distribution using the direct method (section 5.3.1). As a comparison, the contours (red dashed) and the marginal distributions (red solid) of the true posterior distribution P are displayed in addition to the approximation. The blue cross in the central panel denotes the location of the expansion point used to construct Q_D. Above the panel we display the optimal ($\mathrm{KL}(P;Q_\mathrm{D})$) and variational ($\mathrm{KL}(Q_\mathrm{D};P)$) Kullback-Leibler divergences between P and Q_D.

problems. In particular, the direct approach possesses a substantial additional computational burden compared to e.g. a maximum a posteriori (MAP) estimate in ξ which is obtained by minimizing the posterior Hamiltonian $\mathcal{H}$. For the direct approach, the optimization objective Eq. (5.43) consists not only of $\mathcal{H}$, but also of the log-determinant of the metric $\mathcal{M}$. In all but the simplest examples this term cannot be computed directly but has to be approximated numerically as in high dimensions an explicit representation of the matrix becomes infeasible and $\mathcal{M}$ is only implicitly accessible through matrix vector products (MVPs). There are a variety of stochastic log-determinant (more specifically trace-log) estimators based on combining Hutchinsons' trace-estimation [61] with approximations to the matrix logarithm using e.g. Chebychev polynomials [57], Krylov subspace methods [108], or moment constrained estimation based on Maximum Entropy [43]. While all these methods provide a significant improvement in performance compared to directly computing the determinant, they nevertheless typically require many MVPs in order to yield an accurate estimate. For large and complex problems, evaluating an MVP of $\mathcal{M}$ is dominated by applying the Jacobian of x, more precisely of the generative process $s'(\xi)$, and its adjoint to a vector. Similarly, evaluating the gradient of $\mathcal{H}$ is also dominated by an MVP that invokes applying the adjoint Jacobian of $s'(\xi)$. Therefore the computational overhead compared to a MAP estimate in ξ is, roughly, multiplicative in the number of MVPs. For large, non-linear problems, this quickly becomes infeasible as nonlinear opti-

mization typically requires many steps to reach a sensible approximation to the optimum.

Nevertheless there remain some important exceptions, in which a fast and scalable algorithm emerges. In particular recall that

$$\log\left(|\mathcal{M}|\right) = \mathrm{tr}\left(\log\left(\mathbb{1} + \left(\frac{\partial x}{\partial \xi}\right)^T \frac{\partial x}{\partial \xi}\right)\right) = \mathrm{tr}\left(\log\left(\mathbb{1} + \frac{\partial x}{\partial \xi}\left(\frac{\partial x}{\partial \xi}\right)^T\right)\right) \ , \tag{5.58}$$

where the last equality arises from applying the matrix determinant lemma. Therefore in cases where the dimensionality of the so-called data-space (i.E. the target space of x) is much smaller then the dimensionality of the signal space (the domain of ξ), the latter representation of the metric is of much smaller dimension. Thus in cases where either the signal- or the data-space is small, or in weakly non-linear cases (i.E. if $\mathcal{M}$ is close to the identity), the log-determinant may be approximated efficiently enough to give rise to a fast and scalable algorithm. For the (arguably most interesting) class of problems where neither of these assumptions is valid, however, the direct approach to obtain the optimal expansion point becomes too expensive for practical purposes as none of the log-determinant estimators scale linearly with the size of the problem in general.

5.3.2 Geometric Variational inference (geoVI)

As we shall see, it is possible to circumvent the need to compute the log-determinant of the metric at any point, if we employ a specific variant of a variational approximation to obtain the optimal expansion point. To this end, we start with a variational approximation to the posterior P, assuming that the approximative distribution $\tilde{Q}$ is given as the unit Gaussian in y transformed via g. To this end let

$$\tilde{Q}(\xi|\bar{\xi}) = \mathcal{N}\left(g(\xi;\bar{\xi});0,\mathbb{1}\right)\left|\left|\frac{\partial g(\xi;\bar{\xi})}{\partial \xi}\right|\right| \ , \tag{5.59}$$

denote the approximation to the posterior conditional to the expansion point $\bar{\xi}$. The variationally optimal $\bar{\xi}$ can be found by optimization of the forward Kullback-Leibler divergence between $\tilde{Q}$ and P, as given via

$$\begin{aligned}
\mathrm{KL}\left(\tilde{Q}|P\right) &\equiv \int \log\left(\frac{\tilde{Q}(\xi|\bar{\xi})}{P(\xi|d)}\right)\tilde{Q}(\xi|\bar{\xi})\mathrm{d}\xi \\
&= \langle \mathcal{H}(\xi|d)\rangle_{\tilde{Q}(\xi|\bar{\xi})} - \left\langle \mathcal{H}_{\tilde{Q}}(\xi|\bar{\xi})\right\rangle_{\tilde{Q}(\xi|\bar{\xi})} \\
&= \langle \mathcal{H}(\xi|d)\rangle_{\tilde{Q}(\xi|\bar{\xi})} + \frac{1}{2}\left\langle \log\left(\left|\tilde{\mathcal{M}}(\xi)\right|\right)\right\rangle_{\tilde{Q}(\xi|\bar{\xi})} + \mathrm{KL}_0 \ ,
\end{aligned} \tag{5.60}$$

where KL_0 denotes contributions independent of $\bar{\xi}$, and $\mathcal{H}(\xi|d)$ and $\mathcal{H}_{\tilde{Q}}$ denote the Hamiltonians of the posterior and the approximation, respectively. We notice that in this form, a minimization of the KL w.r.t. $\bar{\xi}$ does not circumvent a computation of the log-determinant

of the metric. Within the KL, this term arises from the entropy of the approximation $\tilde{Q}$, and can be understood as a measure of the volume associated with the distribution. In order to avoid this term, our idea is to propose an alternative family of distributions $Q_m(\xi|\bar{\xi})$, defined as a shifted version of $\tilde{Q}$. Specifically we let $\xi \to m + \xi - \bar{\xi}$ such that the distribution may be written as

$$Q_m(\xi|\bar{\xi}) = \tilde{Q}(\xi|\bar{\xi})\Big|_{\xi=\xi+\bar{\xi}-m} \equiv Q(r|\bar{\xi})\Big|_{r=\xi-m} \quad \text{with} \quad r = \xi - \bar{\xi}\ , \tag{5.61}$$

where we also introduced the residual r, which measures the deviations from $\bar{\xi}$, and the associated distribution $Q(r|\bar{\xi})$. In words, $Q_m(\xi|\bar{\xi})$ is the distribution using the residual statistics r, around an expansion point $\bar{\xi}$, but shifted to m. One can easily verify that the entropy related to Q_m becomes independent of m, as shifts are volume-preserving transformations. Therefore we may use some fixed expansion point $\bar{\xi}$, and find the optimal shift m using the KL which now may be written as

$$\begin{aligned} \mathrm{KL}\left(Q_m|P\right) &= \langle \mathcal{H}(\xi = m + r, d)\rangle_{Q(r|\bar{\xi})} + \frac{1}{2}\left\langle \log\left(\left|\tilde{\mathcal{M}}(\xi)\right|\right)\Big|_{\xi=\bar{\xi}+r}\right\rangle_{Q(r|\bar{\xi})} + \mathrm{KL}_0 \\ \widehat{\mathrm{KL}} &= \langle \mathcal{H}(\xi = m + r, d)\rangle_{Q(r|\bar{\xi})}\ , \end{aligned} \tag{5.62}$$

where $\widehat{\mathrm{KL}}$ denotes the KL up to m independent contributions. After optimization for m, we can update to a new expansion point, and use it to define a new family of distributions Q_m which are a more appropriate class of approximations. In general, the expectation value in $\widehat{\mathrm{KL}}$ cannot be computed analytically, but it can be approximated using a set of N samples $\{r_i^*\}_{i\in\{1,\dots,N\}}$, drawn from $Q(r|\bar{\xi})$, which yields

$$\widehat{\mathrm{KL}}\left(Q_m|P\right) \approx \frac{1}{N}\sum_{i=1}^{N} \mathcal{H}(\xi = m + r_i^*, d) \quad \text{with} \quad r_i^* \sim Q(r|\bar{\xi})\ . \tag{5.63}$$

Sampling from $Q(r|\bar{\xi})$ is defined as in section 5.3.1, where the sampling procedure for $\tilde{Q}(\xi|\bar{\xi})$ is described, with the addition that a sample r^* is is obtained from a sample for ξ^* as $r^* = \xi^* - \bar{\xi}$.

Optimizing $\widehat{\mathrm{KL}}$ w.r.t. m yields the variational optimum for the distribution $Q_m(\xi|\bar{\xi})$, given a fixed, predetermined expansion point $\bar{\xi}$. In order to move the expansion point $\bar{\xi}$ towards the optimal point, its location is updated subsequently and the KL is re-estimated using novel samples from $Q(r|\bar{\xi})$ with an updated $\bar{\xi}$. Specifically, we initialize the optimization algorithm at some position m^0, set $\bar{\xi} = m^0$ to obtain a set of samples $\{r_i^*\}^{(0)}_{i\in\{1,\dots,N\}}$, and use this set to approximate the KL. This approximation is then used to obtain an optimal shift m^1. Given this optimal shift, a new expansion point $\bar{\xi} = m^1$ is defined and used to obtain a novel set of samples $\{r_i^*\}^{(1)}_{i\in\{1,\dots,N\}}$ which defines a new estimate for the KL. This estimate is furthermore used to obtain a novel optimal m, and so on. An illustrative view of this procedure is given in figure 5.4. Finally, the entire procedure of optimizing the KL for m and re-estimation of the KL via a novel expansion point is repeated until the algorithm converges to an optimal point $m^* = \bar{\xi}^*$. To optimize $\widehat{\mathrm{KL}}$ for m, we again

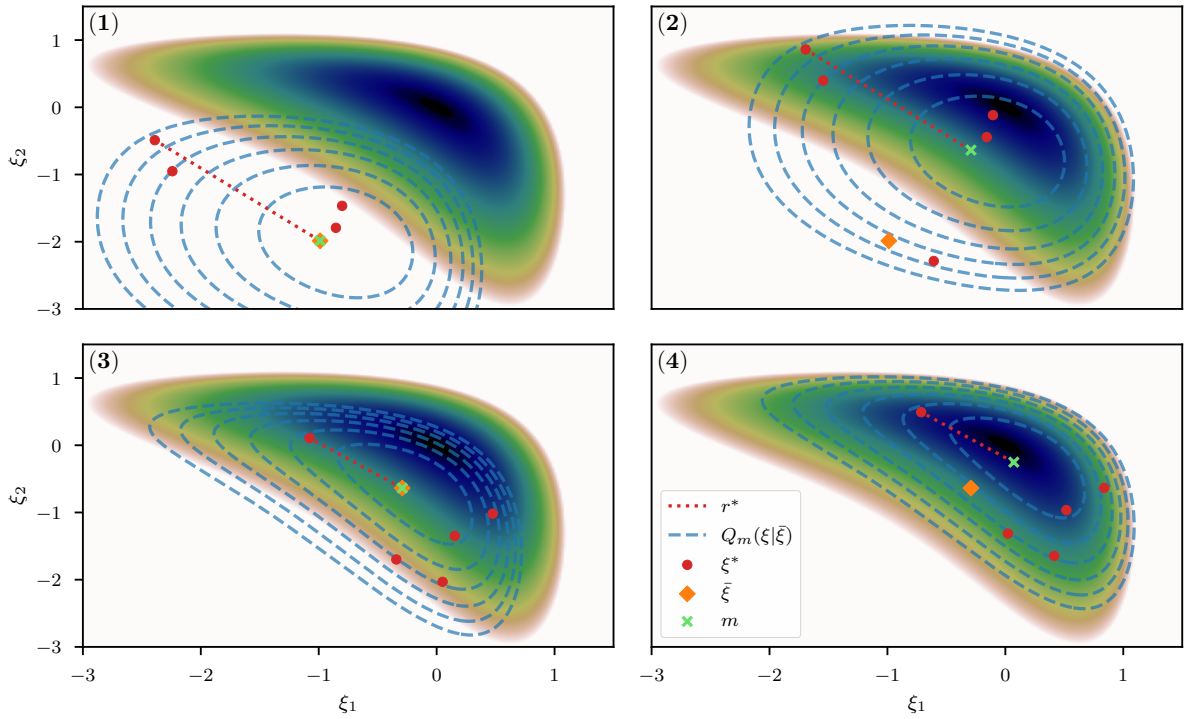

Figure 5.4: (1) – (4): Visualization of the geoVI steps. (1): A randomly initialized shift m (green cross) is used to set the initial expansion point $\bar{\xi}$ (orange dot) which in turn defines the initial approximation $Q_m(\xi|\bar{\xi}$ (blue dashed contours) used to generate a set of samples ξ^* (red dots). (2): The KL (equation (5.63)), estimated from the samples, is used to optimize for m, which results in a shift of $Q_m(\xi|\bar{\xi}$ away from the expansion point $\bar{\xi}$. The residual statistics r^* derived from the geometry around $\bar{\xi}$, however, remains unchanged during this shift and therefore, at the new location m, becomes a bad representation of the local geometry. Thus, in (3), the expansion point is set to the current estimate of m, which yields an update to the approximation $Q_m(\xi|\bar{\xi}$. Finally, we generate samples from this update and use them to optimize the re-estimated KL for m which again results in a shift as seen in (4). Within the full geoVI algorithm this procedure is iterated until convergence.

employ the NewtonCG algorithm, and use the average of the metric $\mathcal{M}$ as a proxy for the curvature of $\widehat{\mathrm{KL}}$ to perform the optimization step. Specifically we use

$$\widehat{\mathcal{M}}(m) = \frac{1}{N}\sum_{i=1}^{N}\mathcal{M}(\xi = m + r_i^*) \ , \tag{5.64}$$

as the metric of $\widehat{\mathrm{KL}}$. We call this algorithm the *geometric Variational Inference* (geoVI) method. A pseudo-code summary of geoVI is given in algorithm 4.

Algorithm 4: Geometric Variational Inference (geoVI)

Input: Likelihood $\mathcal{H}(d|\xi)$, Transformation $x(\xi)$, Jacobian $\frac{\partial x}{\partial \xi}$

Function `Energy(`ξ`):`
  return $\mathcal{H}(d|\xi) + \frac{1}{2}\xi^T\xi$
$\mathsf{m} \sim \mathcal{N}(m, 0, \mathbb{1})$
while m *not converged* **do**
  $\bar{\xi} \leftarrow \mathsf{m}$
  samples $\leftarrow$ *empty list*
  for $i = 1$ *to* N **do**
    $\xi^* \leftarrow$ `drawSample(`$\bar{\xi}$, x, $\frac{\partial x}{\partial \xi}$`)` (see Algorithm 3)
    $r^* \leftarrow \xi^* - \bar{\xi}$
    Insert r^* into samples
  end
  Function `geoKL(`ξ`):`
    kl $\leftarrow 0$
    for r^* *in* samples **do**
      kl $\leftarrow$ kl + `Energy(`$\xi + r^*$`)`
    end
    return $\frac{1}{N}$ kl
  $\mathsf{m}^* \leftarrow$ `NewtonCG(geoKL,` m`)`
  $\mathsf{m} \leftarrow \mathsf{m}^*$
end
posteriorSamples $\leftarrow$ *empty list*
for r^* *in* samples **do**
  $\xi^* \leftarrow \mathsf{m} + r^*$
  Insert ξ^* into posteriorSamples
end

Output: posteriorSamples

Numerical sampling within geoVI

It is noteworthy that, as described in section 5.3.1, an implementation of the proposed sampling procedure for the residual r, and as a result also of the geoVI method itself, inevitably

relies on numerical approximations to realize a sample for r. To better understand the impact of such approximations, we have to consider its impact on the distribution $Q(r|\bar{\xi})$. To this end, we denote with f the function that, given the expansion point $\bar{\xi}$, turns two standard distributed random vectors η_1 and η_2 into a random realization of r. Specifically

$$r = f(\eta_1, \eta_2; \bar{\xi}) \quad \text{with} \quad \eta_{1/2} \sim \mathcal{N}\left(\eta_{1/2}; 0, \mathbb{1}\right) \ , \tag{5.65}$$

where the functional form of f is defined by combination of equation (5.49) and (5.50). using f we may write the geoVI distribution Q as

$$Q(r|\bar{\xi}) = \int\int \delta\left(r - f(\eta_1, \eta_2; \bar{\xi})\right) \ \mathcal{N}\left(\eta_1; 0, \mathbb{1}\right) \ \mathcal{N}\left(\eta_2; 0, \mathbb{1}\right) \ \mathrm{d}\eta_1 \mathrm{d}\eta_2 \ . \tag{5.66}$$

Any numerical algorithm used to approximate the sampling, irrespective of its exact form, may be described by replacing the function f, leading to exact sampling from Q, with some approximation $\widehat{f}$ which leads to an approximation of the distribution for r, which we denote as $\widehat{Q}(r|\bar{\xi})$. Therefore, in a way, the geoVI result using a numerical approximation for sampling can be understood as the variational optimum chosen from the family of distributions $\widehat{Q}$, rather than Q. Therefore, even for a non-zero approximation error in $\widehat{f}$, the result remains a valid optimum of a variational approximation, it is simply the family of distributions used for approximation that has changed. This finding is of great relevance in practice, as there is typically a trade off between numerical accuracy of the generated samples and computational efforts. Thus we may achieve faster convergence at a cost of accuracy in the approximation, but without completely detaching from the theoretical optimum, so long as $\widehat{f}$ remains sufficiently close to f. Nevertheless, as motivated in the introduction, it is important for the chosen family to contain distributions close to the true posterior, and therefore it remains important that the family $\widehat{Q}$ remains close to the family of Q as only for Q the geometric correspondence to the posterior has been established. A detailed study to further quantify this result, is left to future work.

MGVI as a first order approximation

We can compare the geoVI algorithm to the aforementioned variational approximation technique called Metric Gaussian variational inference (MGVI), and notice some key similarities. In particular the optimization heuristics with repeated alternation between sampling of r^* and optimization for m is entirely equivalent. The difference occurs in the distribution $Q(r|\bar{\xi})$ used for approximation. In MGVI, Q is assumed to be a Gaussian distribution in r, as opposed to the Gaussian distribution in the transformed space y used in geoVI. Specifically

$$Q_{\mathrm{MGVI}}(r|\bar{\xi}) \equiv \mathcal{N}(r; 0, \bar{\mathcal{M}}^{-1}) \ , \tag{5.67}$$

where the inverse of the posterior metric $\mathcal{M}$, evaluated at the expansion point $\bar{\xi}$, is used as the covariance. As it turns out, the distribution Q_{MGVI} arises naturally as a first order

approximation to the coordinate transformation used in the geoVI approach. Specifically if we consider the geoVI distribution of r given in terms of a generative process

$$r = g^{-1}(y;\bar{\xi}) - \bar{\xi} \quad \text{with} \quad y \sim \mathcal{N}(y;0,\mathbb{1}) \;, \tag{5.68}$$

and expand it around $y = 0$ to first order, we get that

$$\begin{aligned} r &= g^{-1}(0,\bar{\xi}) - \left(\left.\frac{\partial g(\xi,\bar{\xi})}{\partial \xi}\right|_{\xi=\bar{\xi}}\right)^{-1} y + \mathcal{O}\left(y^2\right) - \bar{\xi} \\ &= \bar{\xi} - \bar{\xi} - \sqrt{\bar{\mathcal{M}}}\left(\mathbb{1} + \left(\left(\frac{\partial x}{\partial \xi}\right)^T \frac{\partial x}{\partial \xi}\right)\Bigg|_{\xi=\bar{\xi}}\right)^{-1} y + \mathcal{O}\left(y^2\right) \\ &= -\left(\sqrt{\bar{\mathcal{M}}}\right)^{-1} y + \mathcal{O}\left(y^2\right) \;. \end{aligned} \tag{5.69}$$

Therefore, to first order in y, we get that

$$Q(r|\bar{\xi}) = \int \delta\left(r + \left(\sqrt{\bar{\mathcal{M}}}\right)^{-1} y\right) \mathcal{N}(y;0,\mathbb{1}) \,\mathrm{d}y = \mathcal{N}(r;0,\bar{\mathcal{M}}^{-1}) = Q_{\mathrm{MGVI}}(r|\bar{\xi}) \;. \tag{5.70}$$

This correspondence shows that geoVI is a generalization of MGVI in non-linear cases. This is a welcome result, as numerous practical applications [62, 110, 7] have shown that already MGVI provides a sensible approximation to the posterior distribution. On the other hand, it provides further insight in which cases the MGVI approximation remains valid, and when it reaches its limitations. In particular if

$$\mathcal{M}(m+r) \approx \bar{\mathcal{M}} \;, \quad \forall r = g^{-1}(y,\bar{\xi}) - \bar{\xi} \quad \text{with} \quad y \sim \mathcal{N}(y;0,\mathbb{1}) \;, \tag{5.71}$$

we get that the first order approximation of equation (5.69) yields a close approximation of the inverse and geoVI reduces to the MGVI algorithm. In contrast, geoVI with its non-linear inversion requires the log determinant of the metric $\mathcal{M}$ to be approximately constant throughout the sampling regime. This is a much less restrictive requirement then equation (5.71), as the variation of eigenvalues of $\mathcal{M}$ is considered on a logarithmic scale whereas it is considered on linear scale in equation (5.71). Furthermore, the log-determinant is invariant under unitary transformations which means that local rotations of the metric, and therefore changes in orientation as we move along the manifold, can be captured by the non-linear approach, whereas equation (5.71) does not hold any more if the orientation varies as a function of r. Therefore we expect the proposed approach to be applicable in a more general context, while still retaining the MGVI properties, as it reproduces MGVI in the linear limit.

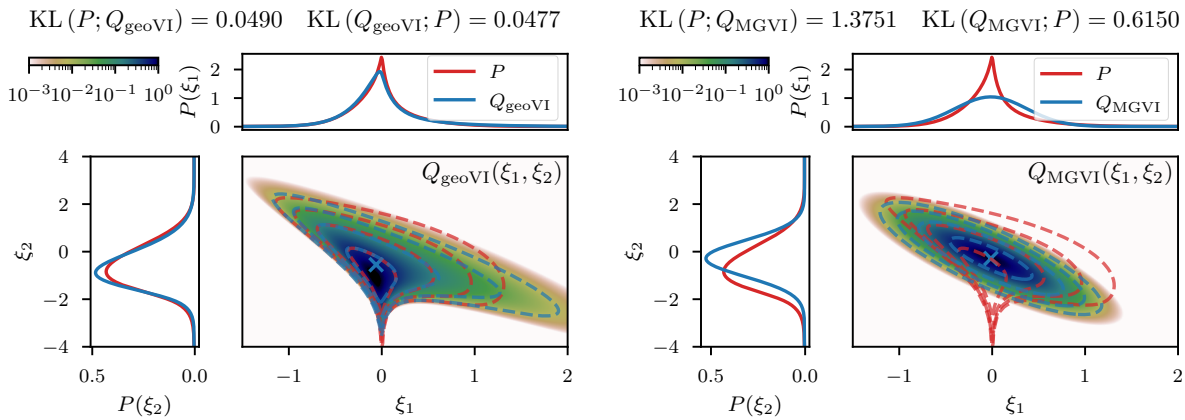

Figure 5.5: The geoVI and MGVI approximations of the two-dimensional example described in section 11. We display the same quantities as for the direct approximation shown in figure 5.3.

5.3.3 Examples

We can visually compare the geoVI and the MGVI algorithm using the two-dimensional example previously mentioned in section 11. In analogy to figure 5.3 we depict the approximation to the posterior density together with its two marginals in figure 5.5. We see that geoVI yields a similar result compared to the direct approach here, while it provides a significant improvement compared to the approximation capacity of MGVI.

To conclude the illustrative examples, we consider a single observation of the product of a normal and a log-normal distributed quantity subject to independent, additive Gaussian noise. The full model consists of a likelihood and a prior of the form

$$P(d|\xi_1, \xi_2) = \mathcal{N}(d; \xi_1 e^{\xi_2}, \sigma_n^2) \quad \text{with} \quad \xi_{1/2} \sim \mathcal{N}(\xi_{1/2}; 0, 1) \ . \tag{5.72}$$

This example should serve as an illustration of the challenges that arise when attempting a separation of non-linearly coupled quantities from a single observation. Such separation problems reappear in section 5.4 in much more intertwined and high dimensional examples, but much of the structural challenges can already be seen in this simple two-dimensional problem. Figure 5.6 displays the results of the direct approach as well as the geoVI and MGVI methods for a measurement setting of $d = -0.3$ and $\sigma_n = 0.1$. As a comparison, we also depict the results from performing a variational approximation using a normal distribution with a diagonal covariance, also known as a mean-field approximation (MFVI), as well as an approximation with a normal distribution using a full-rank matrix as its covariance (FCVI). Both, the diagonal as well as the full-rank covariance are considered parameters of the distribution, and have to be optimized for in addition to the mean of the normal distribution. An efficient implementation thereof is described in [75]. We notice that both, the direct and the geoVI approach manage to approximate the true posterior distribution well, although the KL values indicate that the approximation by geoVI is

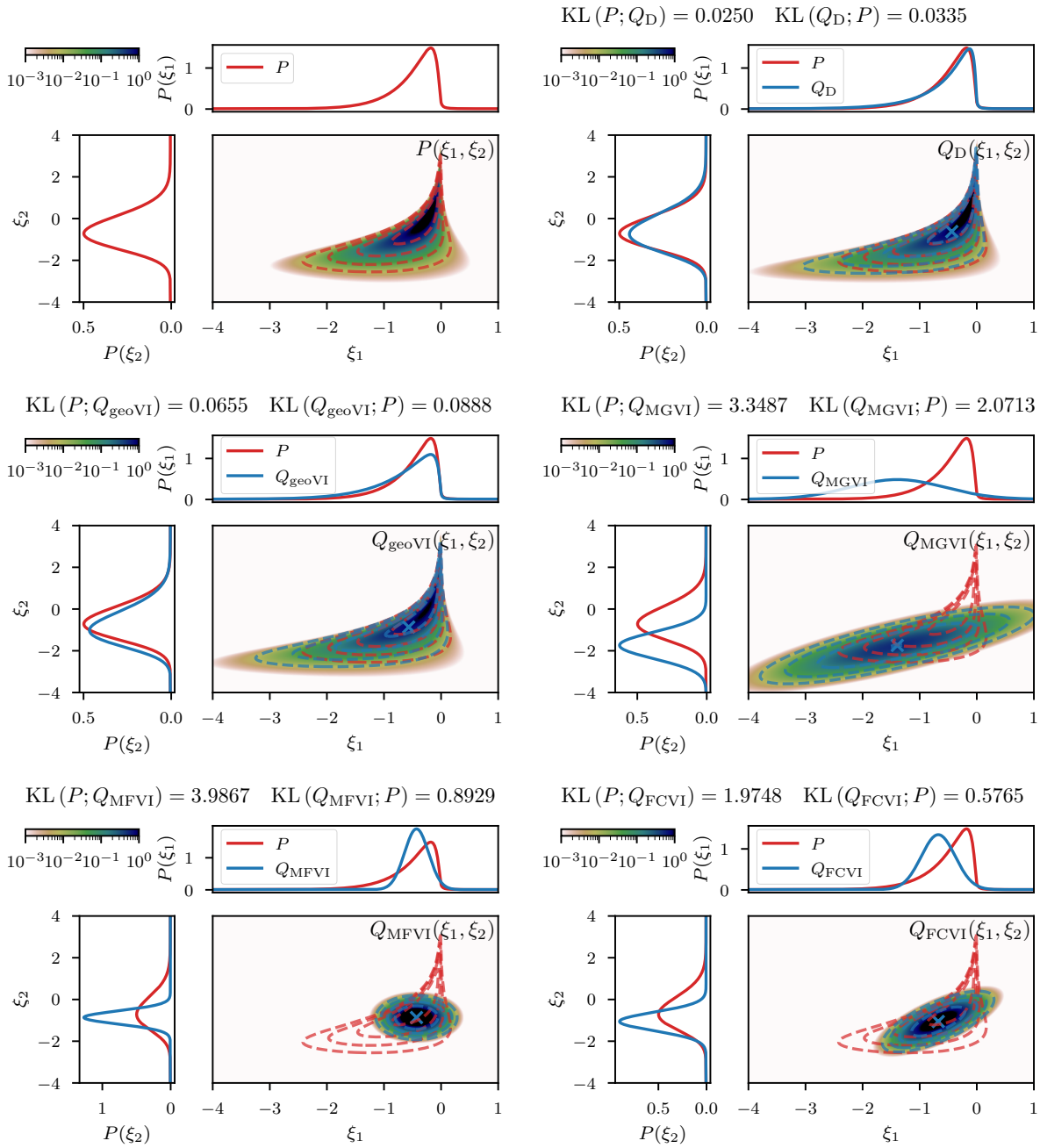

Figure 5.6: Same setup as in figures 5.3 and 5.5 but for a Gaussian measurement of the product of a normal distributed quantity ξ_1 and a log-normal distributed one ξ_2 as described in the second example of section 5.3.3. From top to bottom and from left to right: ground truth P, direct approximation Q_{D}, geoVI approximation Q_{geoVI}, MGVI approximation Q_{MGVI}, mean-field approximation Q_{MFVI}, and the normal approximation with a full-rank covariance Q_{FCVI}.

worse by ≈ 0.016 nats compared to the direct approach. Here the passive update of the expansion point used in this approach reaches its limitations as in cases where the posterior distribution becomes increasingly narrow towards the optimal expansion point, the static sample statistics of r can get stuck during optimization and increasingly repeated resampling becomes necessary as one moves closer to the optimum. Nevertheless, the geoVI approximation remains a good approximation to the true distribution, especially when compared to the approaches using a normal distribution such as MGVI, MFVI, and FCVI.

5.4 Applications

To investigate the performance of the geoVI algorithm in high dimensional imaging problems, we apply it to two mock data examples and compare it to the results using MGVI. In the first example, which serves as an illustration, the geoVI results are additionally compared to the results obtained from applying a Hamiltonian Monte-Carlo (HMC) sampler [29] to the mock example (see section 5.5.2 for further information on HMC). The second example is an illustration of a typical problem encountered in astrophysical imaging. Both examples consist of hierarchical Bayesian models with multiple layers which are represented as a generative process. One particularly important process for the class of problems at hand are statistically homogeneous and isotropic Gaussian processes with unknown power spectral density, for which a flexible generative model has been presented in [8]. This process is at the core of a variety of astrophysical imaging applications [79, 8, 9, 63], and therefore an accurate posterior approximation of problems involving this model is crucial. To better understand the inference challenges that arise in problems using this particular model, we briefly summarize some of its key properties.

5.4.1 Gaussian processes with unknown power spectra

Consider a zero mean, square integrable random process s_x defined on a L-dimensional space subject to periodic boundary conditions which, for simplicity, we assume to have size one. Specifically let $x \in \Lambda = [0, 1]^L$ and thus $s \in \mathcal{L}^2(\Lambda)$. A Gaussian process

$$P(s) = \mathcal{N}(s; 0, S) \ , \tag{5.73}$$

with mean zero and covariance function S_{xy} is said to be statistically homogeneous and isotropic, if S is a function of the Euclidean distance between two points i.E.

$$S_{xy} = S(|x - y|) \ . \tag{5.74}$$

Furthermore, as implied by the Wiener Wiener-Khinchin theorem [112], the linear operator associated with S becomes diagonal in the Fourier space associated with Λ, and therefore s may be represented in terms of a Fourier series with coefficients $\tilde{s}_k$, where k labels

the Fourier coefficients. These coefficients are independent, zero mean Gaussian random variables with variance

$$\left\langle |\tilde{s}_k|^2 \right\rangle_{P(s)} \equiv P_s(|k|) \ , \tag{5.75}$$

which is also known as the power spectrum P_s of s. As P_s encodes the correlation structure of s, its functional form is crucial to determine the prior statistical properties of s. In [8] a flexible, non-parametric prior process for the power spectrum has been proposed by means of a Gauss-Markov process on log-log-scale. This process models the spectrum as a straight line on log-log-scale (resulting in a power law in $|k|$ on linear scale) with possible continuous deviations thereof. These deviations are itself defined as a Gauss-Markov process (specifically an integrated Wiener process) and their respective variance is, among others, an additional scalar parameter steering the properties of this prior process that are also considered to be random variables that have to be inferred. These parameters are summarized in Table 5.1. A more formal derivation of this model in terms of a generative process relating standard distributed random variables ξ_p to a random realization $P_s(\xi_p)$ of this prior model, is given in appendix 5.B.

Name	Description	Prior distribution
offset std.	Prior standard deviation of the overall offset of s from zero	Log-normal
fluctuations	Prior amplitude of the variation of s around its offset	Log-normal
slope	Exponent of the power law related to P_s	Normal
flexibility	Amplitude of deviations from the power-law on log-log-scale	Log-normal
asperity	Smoothness of the deviations as a function of $\log(\|k\|)$	Log-normal

Table 5.1: Table of additional parameters

In order to use this prior within a larger inference model, the underlying space has to be discretized such that the solution of the resulting discrete problem remains consistent with the continuum. We achieve this discretization by means of a truncated Fourier series for s such that s may be written as

$$s = \mathcal{F}^\dagger \left(\sqrt{P_k(\xi_p)}\xi \right) \quad \text{with} \quad P(\xi) = \mathcal{N}(\xi; 0, \mathbb{1}) \ , \tag{5.76}$$

where $\mathcal{F}$ denotes a discrete Fourier transformation (DFT) and $\mathcal{F}^\dagger$ its back-transformation. If we additionally evaluate s on a regular grid on Λ, we can replace the DFT with a fast Fourier transformation (FFT) which is numerically more efficient. For a detailed description on how the spatial discretization is constructed please refer to [82, 46]. In this

work, however, we are primarily interested in evaluating the approximation quality of the proposed algorithm geoVI, and therefore, from now on, we regard all inference problems involving this random process to be high, but finite, dimensional Bayesian inference problems and ignore the fact that it was constructed from a corresponding continuous, infinite dimensional, inference problem.

5.4.2 Log-normal process with noise estimation

As a first example we consider a log-normal process e^s, defined over a one-dimensional space, with s being a priori distributed according to the aforementioned Gaussian process prior with unknown power spectrum. The observed data d (see top panel of figure 5.7) consists of a partially masked realization of this process subject to additive Gaussian noise with standard deviation σ_n. In addition to s and its power spectrum P_s, we also assume σ_n to be unknown prior to the observation and place a log normal prior on it. Therefore the corresponding likelihood takes the form

$$P(d|s, \sigma_n) = \mathcal{N}(d; Re^s, \sigma_n^2) \ . \tag{5.77}$$

We apply the geoVI algorithm (figure 5.7),the MGVI algorithm (figure 5.8), and an HMC sampler (figure 5.9) to this problem and construct a set of 3000 approximate posterior samples for all methods. The HMC results serve as the true reference here, as the true posterior distribution is too high dimensional to be accessible directly and HMC is known to reproduce the true posterior in the limit of infinite samples. Considering solely the reconstruction of e^s, we see that both methods, geoVI and MGVI, agree with the true signal largely within their corresponding uncertainties. Overall we find that the geoVI solution is slightly closer to the ground truth compared to MGVI and the posterior uncertainty is smaller for geoVI in most regions, with the exception of the unobserved region, where it is larger compared to MGVI (see residual plot of figures 5.7 and 5.8). In this region MGVI appears to slightly underestimate the posterior uncertainty. In addition, in the bottom panels of figures 5.7 and 5.8, we depict the posterior distribution of the noise standard deviation σ_n as well as the posterior mean of the power spectrum P_s, together with corresponding posterior samples. Here the difference between geoVI and MGVI becomes evident more visibly, which, to some degree, is to be expected due to the more non-linear coupling of σ_n and P_s to the data compared to e^s. Indeed we find that the posterior distribution of σ_n recovered using MGVI is overestimating the noise level of the reconstruction. The geoVI algorithm is also slightly overestimating σ_n, however we find that for geoVI the posterior yields $\sigma_n^{\text{geoVI}} = 0.220 \pm 0.026$ which places the true value of $\sigma_n = 0.2$ approximately 0.8-sigma away from the posterior mean. In contrast, for MGVI, we get that $\sigma_n^{\text{MGVI}} = 0.233 \pm 0.011$ for with the ground truth is almost a 3-sigma event. Considering the HMC results (bottom panel of figure 5.9), the geoVI results appear to be closer to the HMC distribution compared to MGVI, although the HMC distribution for σ_n is broader and even closer to the true value then geoVI. In addition we find that the overall reconstruction quality of the power spectrum P_s is significantly increased when moving

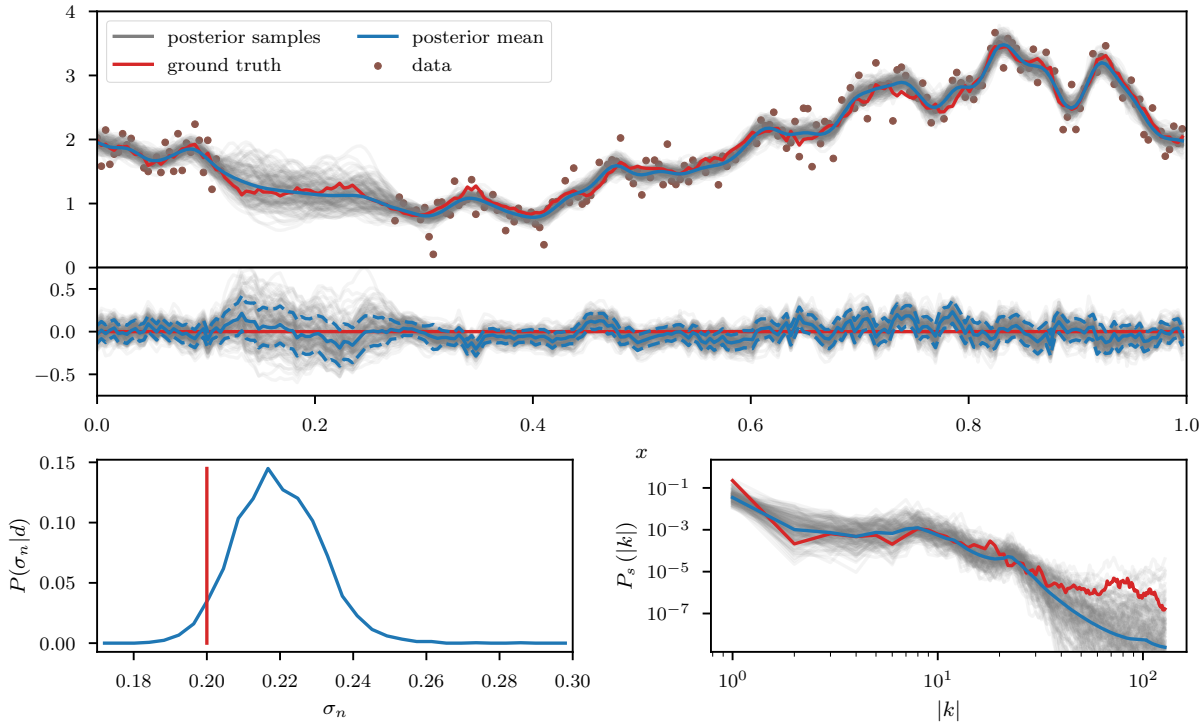

Figure 5.7: Posterior approximation using the geoVI algorithm for the log-normal process described in section 5.4.2. Top: The ground truth realization of the log-normal process e^s (red line) and the corresponding data (brown dots) used for reconstruction. The blue line is the posterior mean, and the gray lines are a subset of the posterior samples obtained from the geoVI approximation. Below we depict the residual between the ground truth and reconstruction, including the residuals for the posterior samples. The blue dashed line corresponds to the one-sigma uncertainty of the reconstruction. Bottom left: Approximation to the marginal posterior distribution (blue) of the noise standard deviation σ_n. The red vertical line indicates the true value of $\sigma_n = 0.2$ used to construct the data. Bottom right: Power spectrum P_s of the logarithmic quantity s. Red displays the ground truth, blue the posterior mean, and the gray lines are posterior samples of the power spectrum.

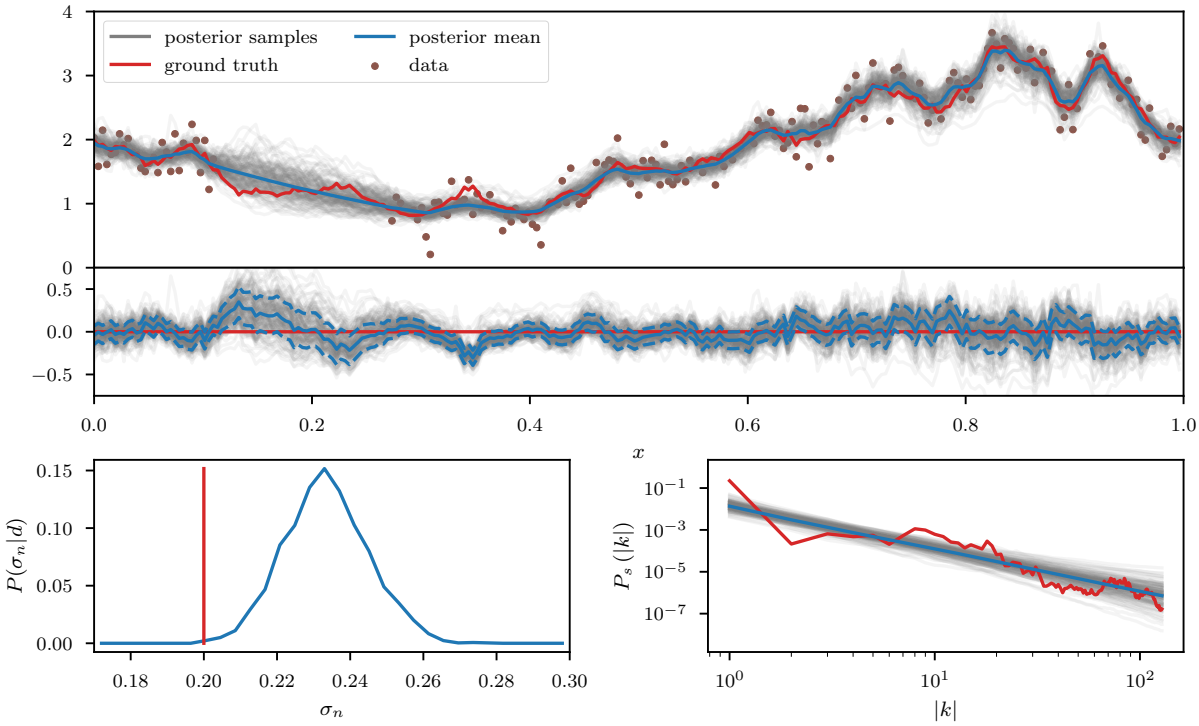

Figure 5.8: Same setup as in figure 5.7, but for the approximation using the MGVI algorithm.

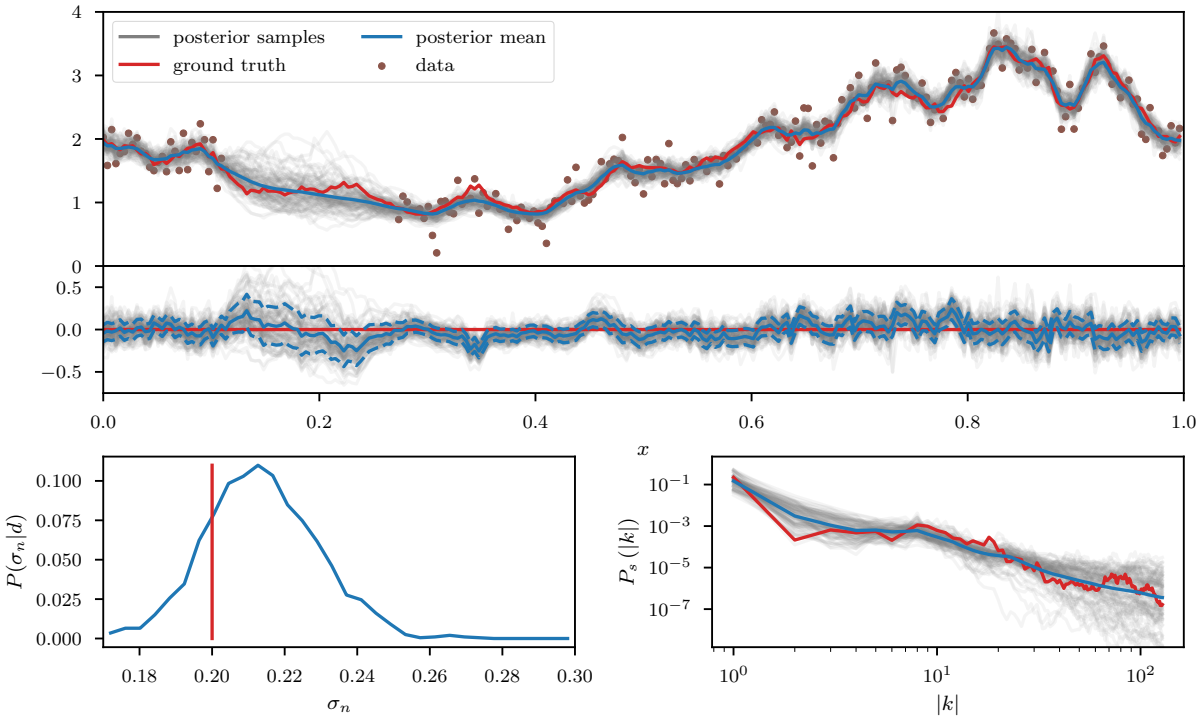

Figure 5.9: Same setup as in figure 5.7, but for the approximation using the HMC sampling.

from MGVI to geoVI. While MGVI manages to recover the overall slope of the power-law, it fails to reconstruct the devations from this power-law as well as the overall statistical properties of P_s as encoded in the parameters of table 5.1. In contrast, the geoVI algorithm is able to pick up some of these features and recovers posterior statistical properties of the power spectrum similar to the ground truth. In addition the posterior uncertainty appears to be on a reasonable scale, as opposed to the MGVI reconstruction which significantly underestimates the posterior uncertainty. The structures on the smallest scales (largest values for $|k|$), however, appear to be underestimated by the geoVI mean, although the posterior uncertainty increases significantly in these regimes. In comparison to HMC we find that the results are in agreement for the large scales, although the geoVI uncertainties appear to be slightly larger. On small scales, the methods deviate stronger, and the under-estimation of the spectrum seen by geoVI is absent in the HMC results.

To further study the posterior distribution of the various scalar parameters that enter the power spectrum model (see Table 5.1), as well as the noise standard deviation σ_n, we depict the reconstructed marginal posterior distributions for all pairs of inferred scalar parameters. Figures 5.10, 5.11, and 5.12 show the posterior distributions recovered using geoVI, MGVI, and HMC, respectively. All parameters are displayed in their corresponding standard coordinate system, i.E. they all follow a zero-mean unit variance normal distribution prior to the measurement. From an inference perspective, some of these parameters are very challenging to reconstruct, as their coupling to the observed data is highly non-linear and influenced by many other parameters of the model. In turn, their values are highly influential to the statistical properties of more interpretable variables such as the observed signal e^s and its spectrum P_s. We see that despite these challenges the geoVI posterior appears to give reasonable results, that are largely in agreement with the ground truth, within uncertainties. Thus the algorithm appears to be able to pick up parts of the non-linear structure of the posterior, which is validated when compared to the MGVI algorithm, as for these parameters the MGVI reconstruction (figure 5.11) does not yield reliable results any more. This is to be expected in case of significant non-linearity as MGVI is the first order approximation of geoVI. In comparison to HMC (figure 5.12), however, we find that there remain some differences in the recovered posterior distributions. The HMC results regarding the "fluctuations" and "noise std." parameters are more centered on the ground truth and in particular the posterior distribution of the "slope" parameter is significantly different and more constrained, compared to the geoVI results. These differences indicate that there remain some limitations to the recovered geoVI results in the regime of highly non-linear parameters of the model which we may associate to the theoretical limitations discussed in section 5.2.2.

5.4.3 Separation of diffuse emission from point sources

In a second inference problem we consider the imaging task of separating diffuse, spatially extended and correlated emission e^s from, bright, but uncorrelated point sources p in an image. This problem is often encountered within certain astrophysical imaging problems [10] where the goal is to recover the emission of spatially extended structures such as gas

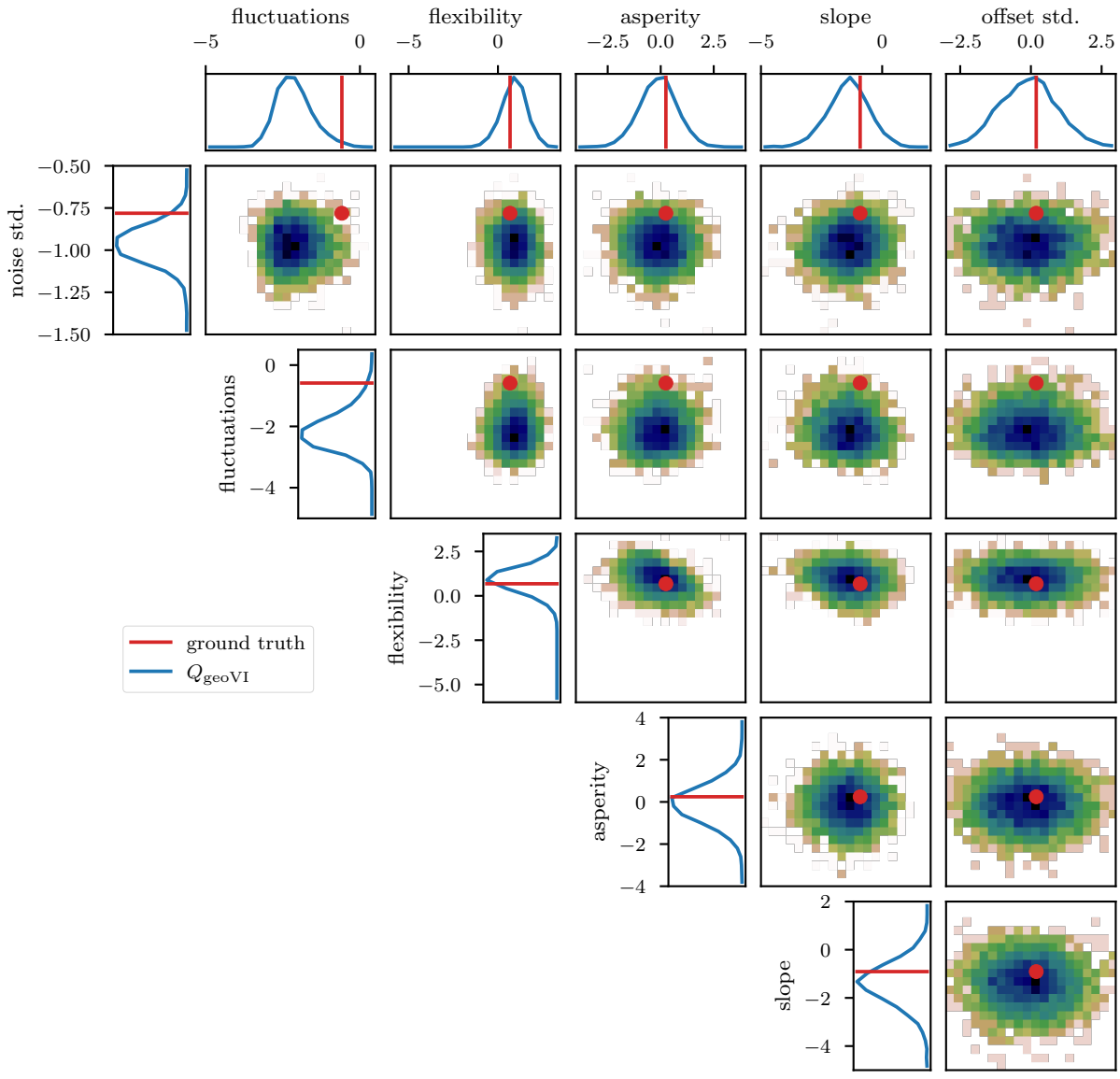

Figure 5.10: Posterior distributions of the scalar parameters that enter the forward model of the power spectrum (Table 5.1), and the noise standard deviation. All parameters, including the noise parameter, are given in their corresponding prior standard coordinate system, i.E. have a normal distribution with zero mean and variance one as a prior distribution. Each square panel corresponds to the joint posterior of the parameter in the respective row and column. In addition, for each row and each column the one-dimensional marginal posteriors of the corresponding parameter are displayed as blue lines. The red lines in the 1-D, and the red dots in the 2-D plots denote the values of the ground truth used to realize the ground truth values of the spectrum P_s, the signal e^s, and finally the observed data d.

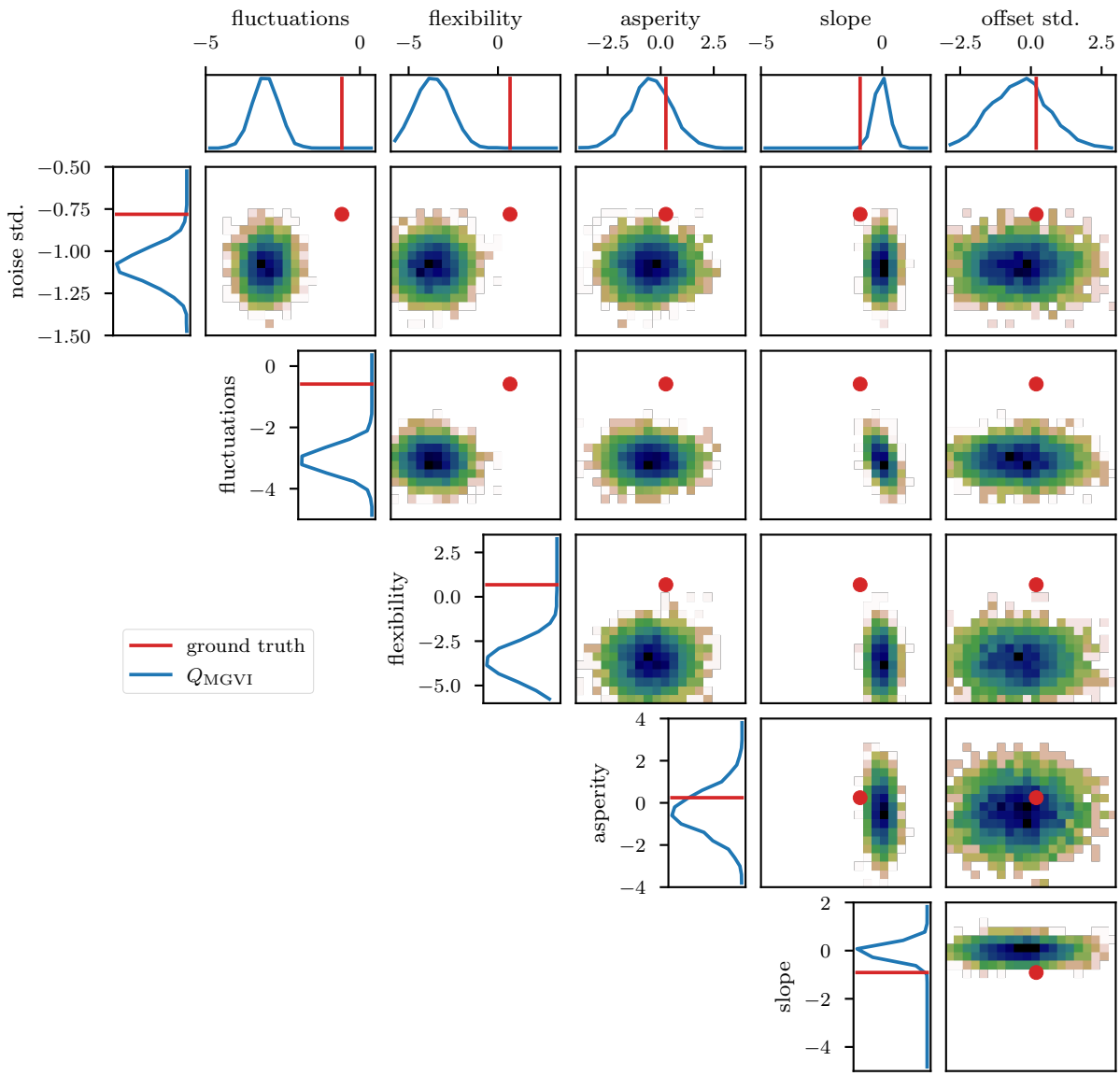

Figure 5.11: Same setup as in figure 5.10, but for the approximation using the MGVI algorithm.

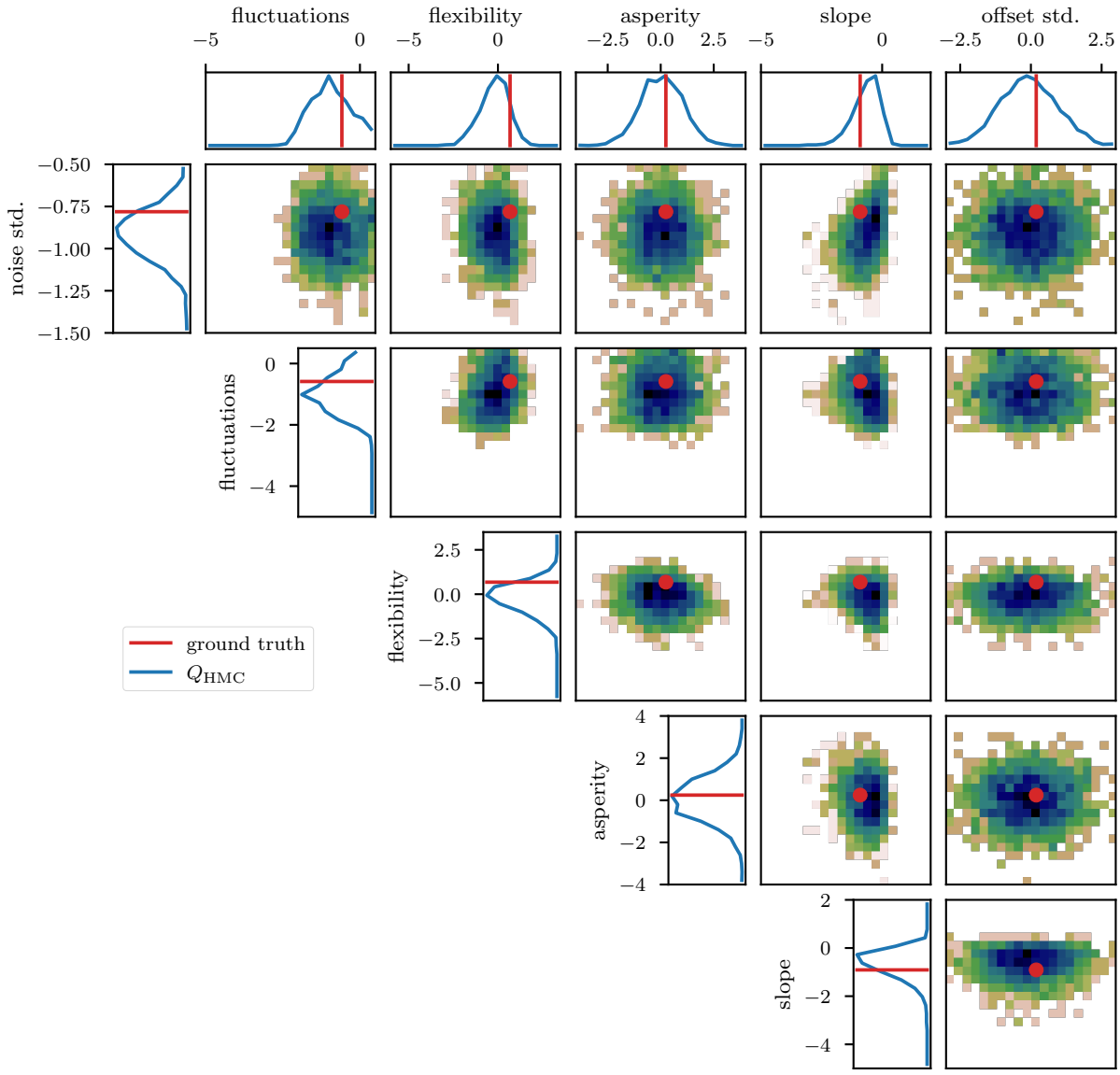

Figure 5.12: Same setup as in figure 5.10, but for the approximation using HMC sampling.

or galactic dust. This emission usually gets superimposed by the bright emission of stars (point sources) in the image plane, and only their joint emission can be observed. In this example we assume that the emission is observed through a detection device that convolves the incoming emission with a spherical symmetric point spread function R and ultimately measures photon counts on a pixelated grid. Specifically we may define a Poisson process with count rate

$$\lambda = R(p + e^s) \quad \text{with} \quad P(d|\lambda) = \prod_i \frac{(\lambda_i)^{d_i}\, e^{-\lambda_i}}{(k_i)!} \;, \tag{5.78}$$

where i labels the pixels of the detector. We assume the diffuse emission to follow a statistically homogeneous and isotropic log-normal distribution with unknown prior power spectrum. Thus s is again distributed according to the prior process previously given in section 5.4.1. The point sources follow an inverse-gamma distribution at every point (x, y) of the image plane, given as

$$P(p_{xy}) = \frac{q^\alpha}{\Gamma(\alpha)} \, (p_{xy})^{-\alpha-1} \exp\left(-\frac{q}{p_{xy}}\right) \;, \tag{5.79}$$

where in the particular example we used $(\alpha, q) = (2, 3)$. A visualization of the problem setup with the various stages of the observation process is given in figure 5.13.

We employ the geoVI and MGVI algorithms to infer all, a priori standard distributed, degrees of freedom of the model and recover the power spectrum P_s for the diffuse emission together with its hyper parameters, the realized emission e^s and the point sources p, from the Poisson count data d. The reconstructed two dimensional images of p and e^s are displayed in figure 5.14 together with the recovered count rate λ, and compared to their respective ground truth. We find that in this example there is barely a visible difference between the reconstructed diffuse emission of MGVI and geoVI. Both reconstructions are in good agreement with the ground truth. For the point sources, we find that the brightest sources are well recovered by both algorithms, while geoVI manages to infer a few more of the fainter sources as opposed to MGVI. Nevertheless, for both algorithms, the posterior mean does not recover very faint sources present in the true source distribution. This can also be seen in figure 5.15, where we depict the per-pixel flux values for all locations in the image against their reconstructed values, for both, the diffuse emission and the point sources. We find that the MGVI and the geoVI are, on average, in very good agreement. It is noteworthy that the deviations between the true and the reconstructed flux values increases towards smaller values, which is to be expected due to the larger impact of the Poisson noise. For the spatially independent point sources, there appears to be a transition regime around a flux of ≈ 50, below which point sources become barely detectable. All in all, for the diffuse emission as well as the point sources, both reconstruction methods apparently yield similar results, consistent with the ground truth. In addition, in figure 5.16, we depict the inferred power spectra. We find that the overall shape is reconstructed well by both algorithms, but smaller, more detailed features can only be recovered using geoVI. In addition we find that the statistical properties of the spectrum, as indicated

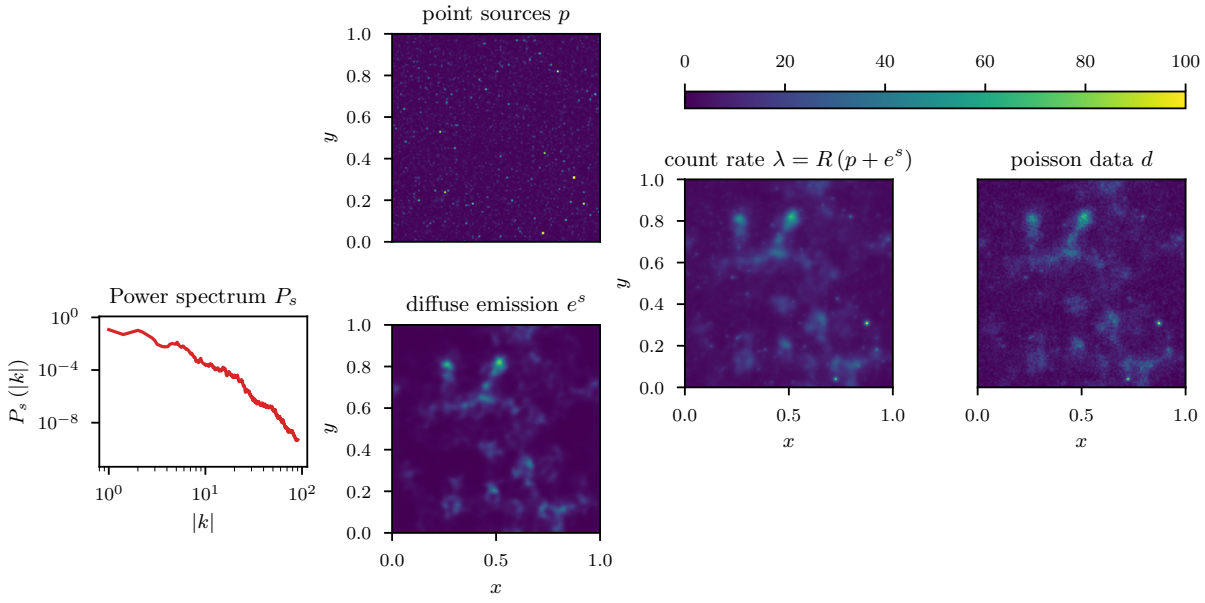

Figure 5.13: Graphical setup of the separation problem discussed in section 5.4.3. Random realization of the power spectrum P_s (left) which is used to generate the log-signal s, which, after exponentiation, models the diffuse emission on the sky e^s. The point sources p (top panel in the middle), which are a realization of the position-independent inverse-gamma process, get combined with the diffuse emission and the result is convolved with a spherical symmetric point spread function R to yield the per-pixel count rate λ which is ultimately used as the rate in a Poisson distribution used to realize the count data d.

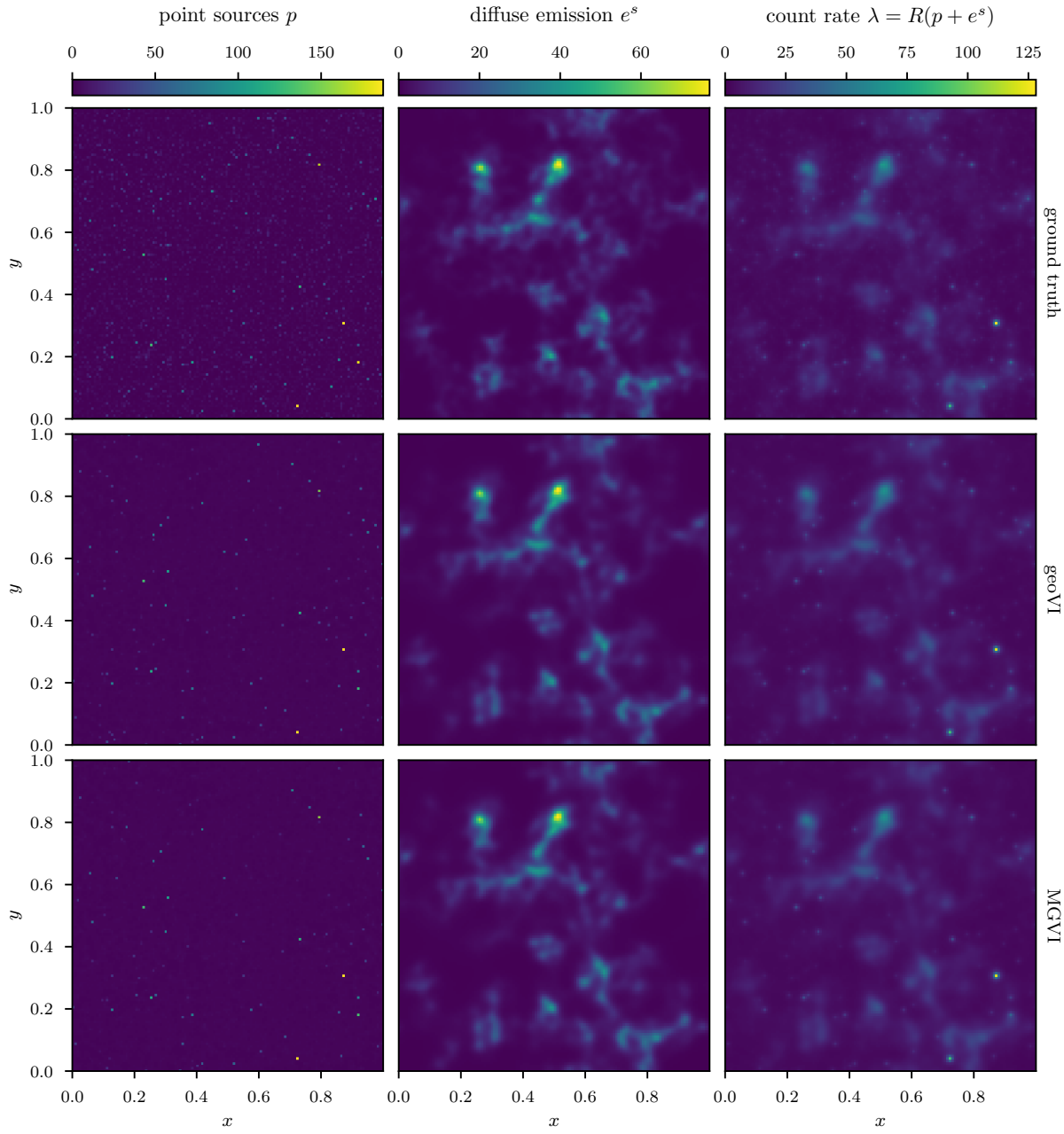

Figure 5.14: Comparison of the ground truth (top row) to the geoVI (middle row) and the MGVI (bottom row) algorithms. The middle and bottom rows show the posterior means for (from left to right) the point sources p, the diffuse emission e^s, and the count rate λ.

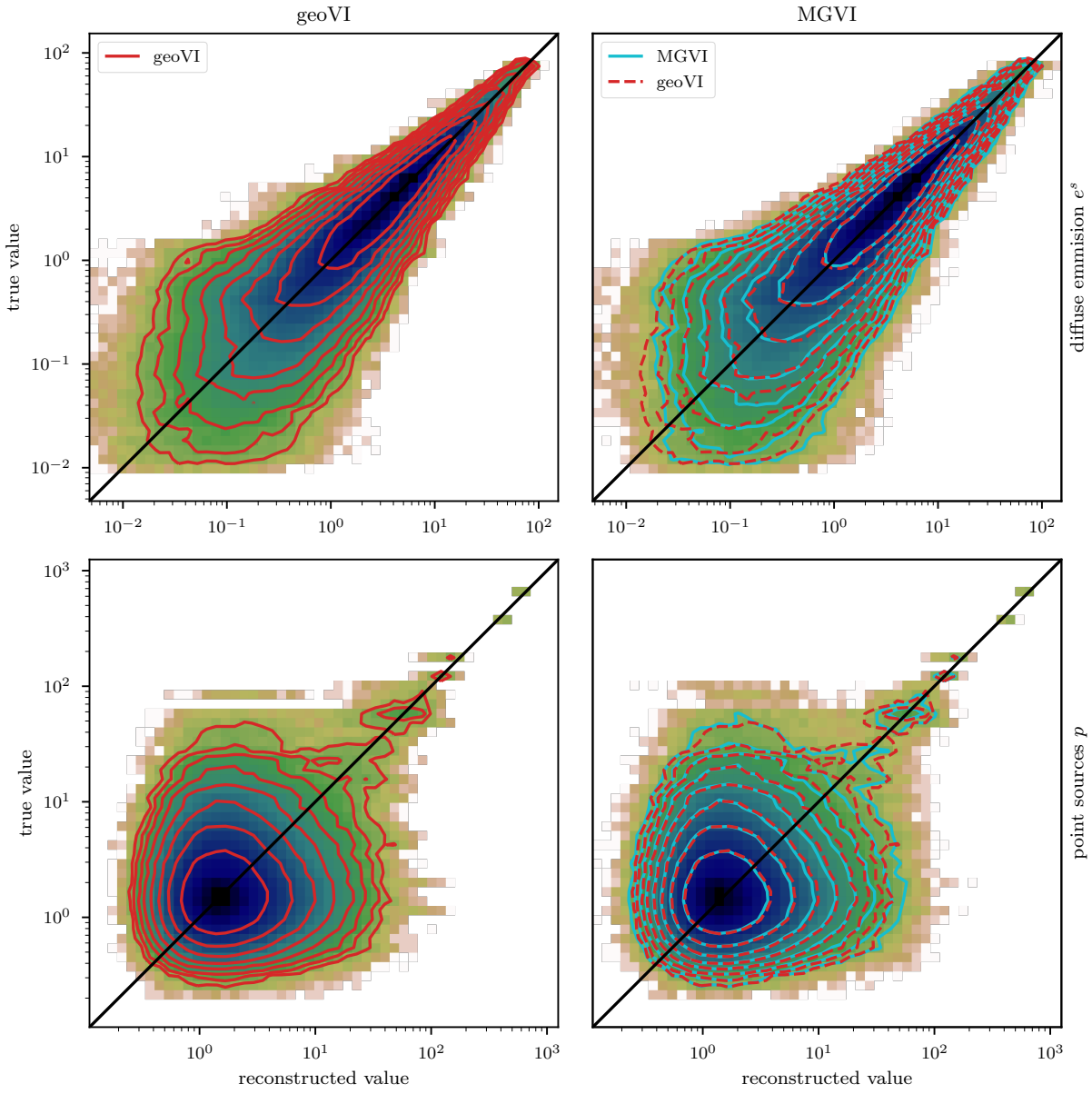

Figure 5.15: Comparison of the per-pixel flux between the ground truth (y-axis) and the reconstruction (x-axis) for the diffuse emission e^s (top row), and the point sources p (bottom row). The left column shows the geoVI result where the density of pixels is color-coded ranging from blue, where the density is highest, to green towards lower densities. The red lines indicate contours of equal density. The right column displays the same for the MGVI reconstruction, with the corresponding density contours now displayed in light blue. The red dashed contours are the density contours of the geoVI case, shown for comparison.

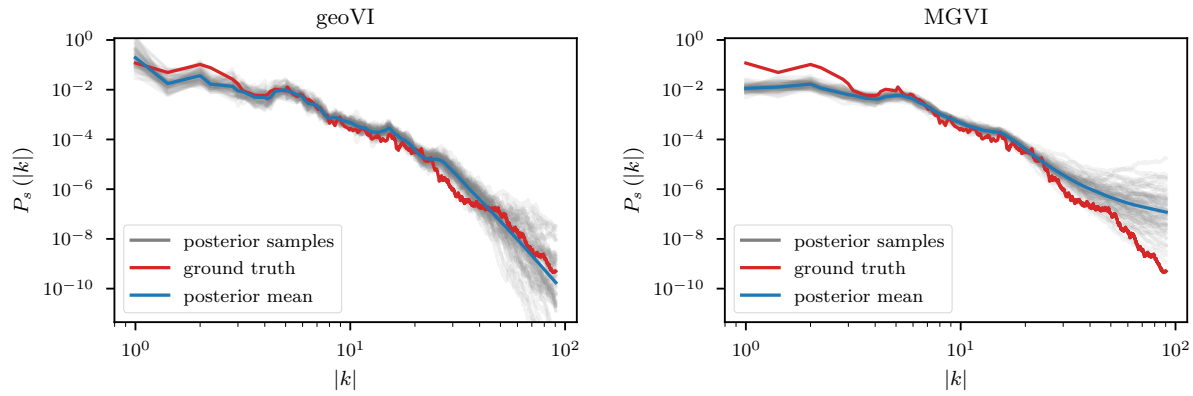

Figure 5.16: Power spectrum P_s of the logarithm of the diffuse emission s. The red line is the ground truth, the blue line the posterior mean, and the gray lines a subset of posterior samples for the geoVI (left) and MGVI (right) approximations.

by e.g. the roughness of the spectrum as a function of the Fourier modes $|k|$, are well reconstructed by geoVI and in agreement with the true spectrum, whereas the MGVI reconstruction, including the posterior samples, appear to be systematically too smooth compared to the ground truth. As discussed in the previous example in more detail, the parameters that enter the model to determine these properties of the spectrum are highly non-linearly coupled and influenced by the observed data and therefore the linear approximation as used in MGVI becomes, at some point, invalid.

5.5 Further properties and challenges

Aside from the apparent capacity to approximate non-linear and high-dimensional posterior distributions, there are some further properties that can be derived from geoVI and the associated coordinate transformation. In the following, we show how to obtain a lower bound to the evidence using the geoVI results. Furthermore, we outline a way to utilize the coordinate transformation in the context of Hamilton Monte-Carlo (HMC) sampling. Finally, some limitations remain to the approximation capacity of geoVI in its current form, which are discussed in section 5.5.3

5.5.1 Evidence lower bound (ELBO)

With the results of the variational approximation at hand, we can provide an Evidence lower bound (ELBO). To this end consider the Hamiltonian $\mathcal{H}(\xi|d)$ of the posterior which takes the form

$$\mathcal{H}(\xi|d) = \mathcal{H}(\xi, d) - \mathcal{H}(d) = \mathcal{H}(d|\xi) + \frac{1}{2}\xi^T\xi + \frac{1}{2}\log\left(|2\pi\mathbb{1}|\right) - \mathcal{H}(d)\ , \tag{5.80}$$

and the Hamiltonian of the approximation $\mathcal{H}_Q$ as a function of r, given as

$$\mathcal{H}_Q(r|\bar{\xi}) = \frac{1}{2} g(\bar{\xi}+r;\bar{\xi})^T g(\bar{\xi}+r;\bar{\xi}) + \frac{1}{2}\log\left(|2\pi\mathbb{1}|\right) - \frac{1}{2}\log\left(\left|\tilde{\mathcal{M}}(\bar{\xi}+r)\right|\right) \ . \tag{5.81}$$

Using these Hamiltonians, we may write the variational approximation as

$$\begin{aligned} \mathrm{KL}(Q;P) &= \left\langle \mathcal{H}(\xi = \bar{\xi}+r|d)\right\rangle_{Q(r|\bar{\xi})} - \left\langle \mathcal{H}_Q(r|\bar{\xi})\right\rangle_{Q(r|\bar{\xi})} \\ &= \left\langle \mathcal{H}(\xi = \bar{\xi}+r,d)\right\rangle_{Q(r|\bar{\xi})} - \mathcal{H}(d) - \left\langle \mathcal{H}_Q(r|\bar{\xi})\right\rangle_{Q(r|\bar{\xi})} \ . \end{aligned} \tag{5.82}$$

As $\mathcal{H}(d) = -\log(P(d))$, we can derive a lower bound for the logarithmic evidence $P(d)$ using the KL as

$$\log(P(d)) \geq \log(P(d)) - \mathrm{KL}(Q;P) \ , \tag{5.83}$$

where the lower bound becomes maximal if the KL becomes minimal. Thus we may use our final expansion point $\bar{\xi}$ obtained from minimizing the KL together with the Hamiltonians (equations (5.80) and (5.81)) to arrive at

$$\begin{aligned} &\log(P(d)) - \mathrm{KL}(Q;P) = \\ &= \frac{1}{2}\mathrm{tr}(\mathbb{1}) - \left\langle \mathcal{H}(d|\xi = \bar{\xi}+r) + \frac{1}{2}\left(\bar{\xi}+r\right)^T\left(\bar{\xi}+r\right) + \frac{1}{2}\log\left(\left|\tilde{\mathcal{M}}(\bar{\xi}+r)\right|\right)\right\rangle_{Q(r|\bar{\xi})} \\ &\approx \frac{1}{2}\mathrm{tr}(\mathbb{1}) - \frac{1}{N}\sum_{i=1}^{N}\left(\mathcal{H}(d|\xi = \bar{\xi}+r_i^*) + \frac{1}{2}\left(\bar{\xi}+r_i^*\right)^T\left(\bar{\xi}+r_i^*\right) + \frac{1}{2}\log\left(\left|\tilde{\mathcal{M}}(\bar{\xi}+r_i^*)\right|\right)\right) \ , \end{aligned} \tag{5.84}$$

where $\{r_i^*\}_{i\in\{1,\dots,N\}}$ are a set of samples drawn from the approximation $Q(r|\bar{\xi})$. Under the assumption that the log determinant of $\tilde{\mathcal{M}}$ is approximately constant throughout the typical set reached by Q, we may replace its sample average with the value obtained at $\bar{\xi}$ to arrive at

$$\begin{aligned} \log(P(d)) - \mathrm{KL}(Q;P) \approx{}& \frac{1}{2}\mathrm{tr}(\mathbb{1}) - \frac{1}{2}\log\left(\left|\bar{\mathcal{M}}\right|\right) \\ &- \frac{1}{N}\sum_{i=1}^{N}\left(\mathcal{H}(d|\xi = \bar{\xi}+r_i^*) + \frac{1}{2}\left(\bar{\xi}+r_i^*\right)^T\left(\bar{\xi}+r_i^*\right)\right) \ , \end{aligned} \tag{5.85}$$

where we also replaced the metric of the expansion $\tilde{\mathcal{M}}$ with the metric of the posterior $\mathcal{M}$ as they are identical when evaluated at the expansion point $\bar{\xi}$ (see equation (5.41)). The assumption that the log determinant does not vary significantly within the typical set is also a requirement for the approximation Q to be a close match for the posterior P and in

turn it is a necessary condition for the ELBO to be a tight lower bound to the evidence as only in this case the KL may become small. Therefore equation (5.85) is a justified simplification in case the entire variational approximation itself is justified. Nevertheless it may be useful to compute the log determinant also for the posterior samples if feasible, as it provides a valuable consistency check for the method itself.

5.5.2 RMHMC with metric approximation

As initially discussed in the introduction, aside from Variational inference methods there exist Markov chain Monte-Carlo (MCMC) methods that utilize the geometry of posterior to increase sampling efficiency. A recently introduced Hybrid Monte-Carlo (HMC) method called Riemannian manifold HMC (or RMHMC) utilizes the same posterior metric as discussed in this work in order to define a Riemannian manifold on which the HMC integration is performed. As one of the key results presented here yields an approximate isometry for this manifold, we like to study the impact of the proposed coordinate transformation on RMHMC. To do so, recall that in HMC the random variable $\xi \in \mathbb{R}^M$, which is distributed according to a posterior distribution $P(\xi)$, is accompanied by another random variable $p \in \mathbb{R}^M$, called momentum, and their joint distribution $P(\xi, p)$ is factorized by means of the posterior $P(\xi)$, and the conditional distribution $P(p|\xi)$. The main idea of HMC is to regard the joint Hamiltonian $\mathcal{H}(\xi, p) = -\log(P(\xi, p))$ as an artificial Hamiltonian system that can be used to construct a new posterior sample from a previous one by following trajectories of the Hamiltonian dynamics. In particular suppose that we are given some random realization ξ^0 of $P(\xi)$, we may use the conditional distribution $P(p|\xi^0)$ to generate a random realization p^0. Given a pair (ξ^0, p^0), HMC solves the dynamical system associated with the Hamiltonian $\mathcal{H}(\xi, p)$ for some integration time t^*, to obtain a new pair (ξ^*, p^*). As Hamiltonian dynamics is both energy and volume preserving by construction, one can show that if (ξ^0, p^0) is a random realization of $P(\xi, p)$, then also (ξ^*, p^*) is. This procedure may be repeated until a desired number of posterior samples is collected. In practice, the performance of an HMC implementation for a specific distribution $P(\xi)$ strongly depends on the choice of conditional distribution $P(p|\xi)$. To simplify the Hamiltonian trajectories and enable a fast traversion of the posterior, RMHMC has been proposed which utilizes a position dependent metric for the conditional distribution of the momentum which takes the form

$$P(p|\xi) = \mathcal{N}(p; 0, \mathcal{M}(\xi)) \ , \tag{5.86}$$

where $\mathcal{M}(\xi)$ denotes the metric associated with the posterior $P(\xi)$ as introduced in section 5.2 in equation (5.10). The associated Hamiltonian takes the form

$$\mathcal{H}(p, \xi) = \frac{1}{2} p^T \mathcal{M}(\xi)^{-1} p + \frac{1}{2} \log\left(|\mathcal{M}(\xi)|\right) + \mathcal{H}(\xi) \ . \tag{5.87}$$

In direct analogy of the discussion in section 5.2, the motivation of utilizing the metric is that the resulting Hamiltonian system can be understood as being defined on the Riemannian manifold associated with $\mathcal{M}$. Therefore the geometric complexity is absorbed into the shape of the manifold, and the trajectories become particularly simple. In practice,

however, numerical integration of the system related to equation (5.87) is challenging, as in general $\mathcal{H}(p,\xi)$ is non-separable. Here, our coordinate transformation may come in handy, as a Hamiltonian using the approximated metric $\tilde{\mathcal{M}}$ (equation (5.14)) instead of $\mathcal{M}$ becomes separable. Specifically replacing $\mathcal{M}$ in equation (5.87) yields

$$\begin{aligned}\mathcal{H}(p,\xi) &= \frac{1}{2}p^T\tilde{\mathcal{M}}(\xi;\bar{\xi})^{-1}p + \frac{1}{2}\log\left(\left|\tilde{\mathcal{M}}(\xi;\bar{\xi})\right|\right) + \mathcal{H}(\xi) \\ &= \frac{1}{2}p^T\left(\left(\frac{\partial g(\xi;\bar{\xi})}{\partial\xi}\right)^T\frac{\partial g(\xi;\bar{\xi})}{\partial\xi}\right)^{-1}p + \tilde{\mathcal{H}}(\xi;\bar{\xi})\ .\end{aligned} \tag{5.88}$$

This modified system allows for a canonical transformation of the form

$$\begin{pmatrix} y \\ v \end{pmatrix} \leftarrow \begin{pmatrix} g(\xi;\bar{\xi}) \\ \left(\frac{\partial g(\xi;\bar{\xi})}{\partial\xi}\right)^T p \end{pmatrix}\ , \tag{5.89}$$

in which the Hamiltonian (equation 5.88) takes the form

$$\mathcal{H}(v,y) = \frac{1}{2}v^Tv + \tilde{\mathcal{H}}(\xi;\bar{\xi})\Big|_{\xi=g^{-1}(y;\bar{\xi})} \equiv T(v) + V(y)\ , \tag{5.90}$$

and therefore $\mathcal{H}$ is separable in the momenta v and the position y. This separability is an interesting property as it has the potential to simplify the integration step used within RMHMC. However, in its current form, we find that there are multiple issues with this approach that prevent an efficient implementation in practice. For one, the transformation g depends on an expansion point $\bar{\xi}$, which becomes a hyper-parameter of the method that has to be determined (possibly in the warm-up phase). In addition, unlike the direct approach discussed in section 5.3.1, we cannot circumvent the inversion of g, which is only implicitly available in general, as it has to be computed for every integration step related to $\mathcal{H}(v,y)$ (equation (5.90)). Therefore, numerical integration of the system may be simpler, but evaluation of $\mathcal{H}(v,y)$ becomes more expensive. Finally, the approximation of the metric may become invalid as we move far away from the expansion point $\bar{\xi}$, and therefore the applicability compared to an RMHMC implementation using the full metric $\mathcal{M}$ is limited. Nevertheless we find the existence of a separable approximation to the Hamiltonian system very interesting, and think that the (or a similar) transformation g and its associated coordinate system (y,v) might be of relevance in the future development of RMHMC algorithms.

5.5.3 Pathological cases

As discussed in section 5.2.1, one property that can violate our assumptions are non-monotonic changes in the metric. To this end, consider a sigmoid-normal distributed random variable, and a measurement subject to additive, independent noise of the form

$$P(d|\xi) = \mathcal{N}(d;\sigma\left(\sigma_p\xi\right),\sigma_n^2) \quad \text{with} \quad P(\xi) = \mathcal{N}(\xi;0,1)\ , \tag{5.91}$$

where $\sigma(\bullet)$ denotes the sigmoid function. The resulting posterior, its associated coordinate transformation, as well as its geoVI approximation, is displayed in figure 5.17 for a case with $(\sigma_p, \sigma_n, d) = (3, 0.2, 0.2)$. We find that similar to the one-dimensional log-normal example of section 11, the approximation quality depends on the chosen expansion point. However, the changes in approximation quality are much more drastic as in the log-normal example. In particular, due to the sigmoid non-linearity, there exists a turning point in the coordinate transformation g, and if we choose an expansion point close to this point, we see that the approximation to the transformation strongly deviates from the optimal transformation as we move away from this point. As a result, in this case the approximation to the posterior (left panel of figure 5.17) obtains a heavy tail that is neither present in the true posterior nor the approximation using the optimal transformation. Nevertheless there may very well also exist a case where such a heavy tail is present in the optimal approximation to the transformation. Even in the depicted case, where the tail is only present for sub-optimal choices of the expansion point, an optimization algorithm might have to traverse this sub-optimal region to reach the optimum. Thus the heavy tail can lead to extreme samples for some intermediate approximation, and therefore the geoVI algorithm could become unstable.

In a second example we consider a bi-modal posterior distribution, generated from a Gaussian measurement of a polynomial. Specifically we consider a likelihood of the form

$$P(d|\xi) = \mathcal{N}\left(d; \xi^4 + \xi, 1\right) \ , \tag{5.92}$$

with ξ being a priori standard distributed. As can be seen in figure 5.18, this scenario leads to a bi-modal posterior distribution with two well separated, asymmetric modes. We find that the geometrically optimal transformation g_{iso} also leads to a bi-modal distribution in the transformed coordinates, however the local asymmetry and curvature of each mode has approximately been removed. Thus while an approximation of the posterior by means of a single unit Gaussian distribution is apparently not possible, each mode may be approximated individually, at least in case the modes are well separated. If we consider the approximation of the coordinate transformation used within geoVI, and choose as an expansion point the optimal point associated with one of the two modes, we get that for the chosen mode the approximation remains valid and the transformation is close to the optimal transformation. However, if we move away from the mode towards the other mode, the approximation quickly deviates from g_{iso} and eventually becomes non-invertible. Therefore only the approximation of one of the modes is possible. Here, care must be taken, as in practical applications the inversion of g is performed numerically and one has to ensure that the inversion does not end up on the second branch of the transformation.

This summarizes the two main issues that may render a geoVI approximation of a posterior distribution invalid. The challenges and issues related to multi modality appear to be quite fundamental, as in its current form, the geoVI method falls into the category of methods that utilize local information of the posterior which all suffer from the inability to deal with more than a single mode. The problems related to turning points are more specific for geoVI, and its implications need to be further studied in order to generalize

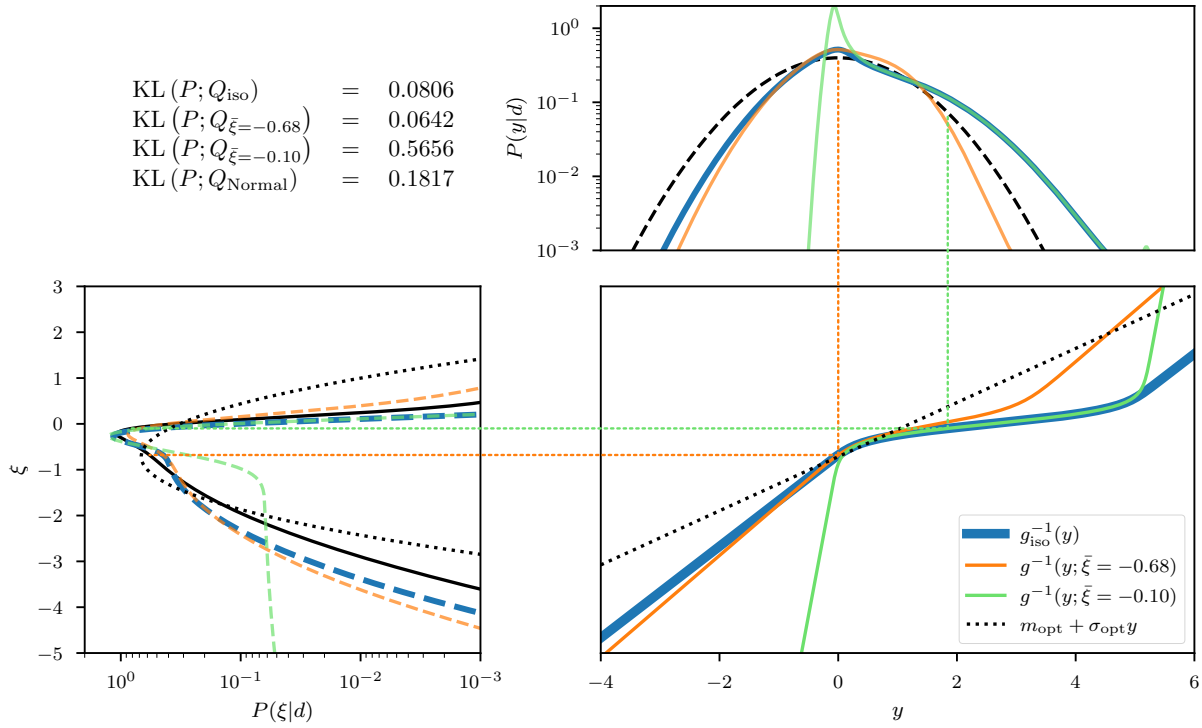

Figure 5.17: Same setup as in figure 5.2, but for the sigmoid-normal distributed case. In addition to the exact isometry g_{iso}, the approximation using the optimal expansion point $\bar{\xi} = -0.68$ and a pathological heavy-tail example using $\bar{\xi} = -0.1$ is displayed.

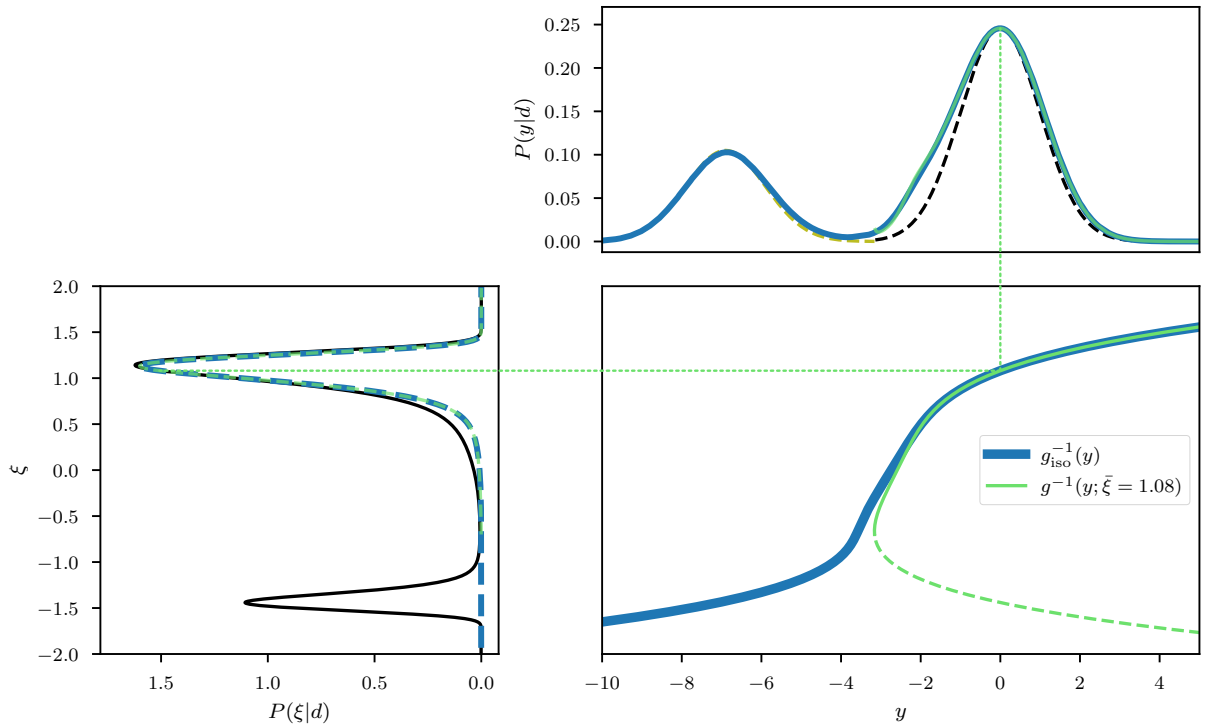

Figure 5.18: Second pathological example, given as a bi-modal posterior distribution. The setup is similar to figures 5.2 and 5.17, where in this example only the (locally) optimal expansion point $\bar{\xi} = 1.08$ is used.

its range of applicability in the future. One promising finding is that this issue appears to be solely related to the local approximation of the transformation with a "bad" expansion point, as the geometrically optimal transformation g_{iso} apparently does not show such behavior. Therefore an extension of the current approximation technique using e.g. multiple expansion points, or identifying and excluding these "bad" expansion points during optimization, may provide a solution to this problem. At the current stage of the development, however, it is unclear how to incorporate such ideas into the algorithm without loss of the functional form of g that allows for the numerically efficient implementation at hand.

5.6 Summary and Outlook

In this work we introduced a coordinate transformation for posterior distributions that yields a coordinate system in which the distributions take a geometrically simple form. In particular we construct a metric as the sum of the Fisher metric of the likelihood and the identity matrix for a standard prior distribution, and construct the transformation that relates this metric to the Euclidean metric. Using this transformed coordinate system, we introduce geometric Variational Inference (geoVI), where we perform a variational ap-

proximation of the transformed posterior distribution with a normal distribution with unit covariance. As the coordinate transformation is only approximately available and utilizes an expansion point around which it is most accurate, the VI task reduces to finding the optimal expansion point such that the variational KL between the true posterior and the approximation becomes minimal. There exists a numerically efficient realization that enables high-dimensional applications of geoVI because even though the transformation is non-volume preserving, geoVI avoids a computation of the related log-determinant of the Jacobian of the transformation at any point. The expansion point used to generate intermediate samples is only passively updated. Furthermore, the application of the constructed coordinate transformation is similar to the cost of computing the gradient of the posterior Hamiltonian. In addition, to generate random realizations, computing the appearing matrix square root of the metric can be entirely avoided, and the inverse transformation is achieved implicitly by second order numerical inversion.

Despite being an approximation method, we find that geoVI is successfully applicable in non-linear, but uni-modal settings, which we demonstrated with multiple examples. We see that non-linear features of the posterior distribution can accurately be captured by the coordinate transformation in low-dimensional examples. This property may translate into high dimensions, as it increases the overall reconstruction quality there when compared to its linearized version MGVI. Nevertheless we also find remaining pathological cases in which further development is necessary to achieve a good approximation quality.

In addition to posterior approximation, geoVI results can be used in order to provide an evidence lower bound (ELBO) which is used for model comparison. Finally we demonstrate the overlap to another posterior sampling technique based on Hamilton Monte-Carlo (HMC), that utilizes the same metric used in geoVI, called Riemannian manifold HMC.

All in all, the geoVI algorithm, and more generally the constructed approximative coordinate transformation, are a fast and accurate way to approximate non-linear and high-dimensional posterior distributions.

Appendix

5.A Likelihood transformations

In order to construct the coordinate transformation $x(\xi)$ introduced in section 5.2.1, we require that the Fisher metric of the likelihood $\mathcal{M}_{d|\xi}$ may be written as the pullback of the Euclidean metric. Recall that the likelihood expressed in coordinates ξ is obtained from the likelihood $P(d|s')$ with $s' = f'(\xi)$ (see equation (5.4)). Therefore we may express $\mathcal{M}_{d|\xi}$ as

$$\mathcal{M}_{d|\xi}(\xi) = \left(\frac{\partial s'}{\partial \xi}\right)^T \mathcal{M}_{d|s'} \frac{\partial s'}{\partial \xi} \ . \tag{5.93}$$

Thus the task reduces to construct a transformation $x(s')$ that recovers $\mathcal{M}_{d|s'}$ from the Euclidean metric if we set the full transformation to be $x(\xi) \equiv x(s' = f'(\xi))$. Specifically we require for $x(s')$

$$\mathcal{M}_{d|s'} \overset{!}{=} \left(\frac{\partial x}{\partial s'}\right)^T \frac{\partial x}{\partial s'} \ . \tag{5.94}$$

Below, in table 5.2, we give a summary of multiple commonly used likelihoods, their respective Fisher metric, and the associated transformation $x(s')$.

Name	$\mathcal{H}(d\|s')$	Metric $\mathcal{M}$	Trafo. $x(s')$
Normal	$\frac{1}{2}(d-s')^T N^{-1}(d-s') + \text{cst.}$	N^{-1}	$\sqrt{N^{-1}}s'$
Poisson	$1^T s' - d^T \log(s') + \text{cst.}$	$1/s'$	$\frac{1}{2}\sqrt{s'}$
Inv. Gamma	$(\alpha+1)^T \log(s') + \beta^T \left(\frac{1}{s'}\right) + \text{cst.}$	$\frac{\alpha+1}{s'^2}$	$\sqrt{\alpha+1}\log(s')$
Student-T	$\frac{\theta+1}{2}\log\left(1+\frac{s'^2}{\theta}\right) + \text{cst.}$	$\frac{\theta+1}{\theta+3}$	$\sqrt{\frac{\theta+1}{\theta+3}}s'$
Bernoulli	$-d^T \log(s') - (1-d)^T \log(1-s') + \text{cst.}$	$\frac{1}{s'(1-s')}$	$-2\tan^{-1}\left(\sqrt{s'}\right)$

Table 5.2: List of common likelihood distributions with their respective Hamiltonian $\mathcal{H}(d|s')$, their Fisher Metric $\mathcal{M}(d|s')$, and the associated coordinate transformation $x(s')$ satisfying equation (5.94).

For some likelihoods, however, such a decomposition is not accessible in a simple form. One example that is being used in this work is a normal distribution with unknown mean m and variance v. The Hamiltonian of a one dimensional example takes the form

$$\mathcal{H}(d|m,v) = \frac{1}{2}\frac{(d-m)^2}{v} + \frac{1}{2}\log(v) + \text{cst.} \ , \tag{5.95}$$

and the corresponding fisher metric for $s' = (m, v)$ is

$$\mathcal{M}_{d|s'} = \begin{pmatrix} \frac{1}{v} & 0 \\ 0 & \frac{1}{2v^2} \end{pmatrix} \ . \tag{5.96}$$

While there is no simple decomposition by means of the Jacobian of some function x, there is an approximation available for which x takes the form

$$x(s') = \begin{pmatrix} \frac{d-m}{\sqrt{v}} \\ \frac{1}{2}\log(v) \end{pmatrix} \quad \text{with} \quad \frac{\partial x}{\partial s'} = \begin{pmatrix} -\frac{1}{\sqrt{v}} & -\frac{d-m}{2v^{3/2}} \\ 0 & \frac{1}{2v} \end{pmatrix} \ . \tag{5.97}$$

We can compute the approximation to the metric and find

$$\left(\frac{\partial x}{\partial s'}\right)^T \frac{\partial x}{\partial s'} = \begin{pmatrix} \frac{1}{v} & \frac{d-m}{2v^2} \\ \frac{d-m}{2v^2} & \frac{1}{4v^2} + \frac{(d-m)^2}{4v^2} \end{pmatrix} \ . \tag{5.98}$$

Note that as opposed to the Fisher metric, this approximation depends on the observed data d. In fact we can recover the Fisher metric from this approximation by taking the expectation value w.r.t. the likelihood. Specifically

$$\left\langle \left(\frac{\partial x}{\partial s'}\right)^T \frac{\partial x}{\partial s'} \right\rangle_{\mathcal{N}(d;m,v)} = \begin{pmatrix} \frac{1}{v} & 0 \\ 0 & \frac{1}{2v^2} \end{pmatrix} = \mathcal{M}_{d|s'} \ , \tag{5.99}$$

and therefore it may be regarded as a local approximation using the observed data. All examples of this work that use a normal distribution where in addition to the mean also the variance is inferred, use this approximation.

5.A.1 Multiple likelihoods

In general, we may encounter measurement situations where multiple likelihoods are involved, e.g. if we aim to constrain s' with multiple data-sets simultaneously. Specifically consider a set of D data-sets $\{d_i\}_{i\in\{1,...,D\}}$, and an associated mutually independent set of likelihoods, such that the joint likelihood takes the form

$$P(d_1, ..., d_D|s') = \prod_{i=1}^{D} P(d_i|s') \ , \tag{5.100}$$

we get that the corresponding Fisher metric takes the form

$$\mathcal{M}_{d_1,...,d_D|s'}(s') = \sum_{i=1}^{D} \mathcal{M}_{d_i|s'} \ . \tag{5.101}$$

If we assume that we have, for every individual metric $\mathcal{M}_{d_i|s'}$, an associated transformation $x_i(s')$ available that satisfies equation (5.94), we see that we can stack them together to form a combined transformation

$$x(s') \equiv (x_1(s'), ..., x_D(s'))^T \ , \tag{5.102}$$

that automatically satisfies (5.94) for the joint metric $\mathcal{M}_{d_1,...,d_D|s'}$.

5.B Correlated Field model

Here, we give a brief description of the generative model for power spectra and resulting Gaussian processes used in section 5.4. For a detailed and extended derivation please refer to [8].

A random realization $s \in \mathcal{L}(\Lambda)$ of a statistically homogeneous and isotropic Gaussian process $P(s)$, defined over an L dimensional domain $\Lambda = [0, 1]^L$, subject to periodic boundary conditions along each dimension, may be represented as a Fourier series via

$$s_x = \left(\mathcal{F}^\dagger A \xi\right)_x \equiv \sum_k e^{-2\pi i k x} A(|k|)\ \xi_k \quad \text{with} \quad \xi_k \sim \mathcal{N}(\xi; 0, 1)\ \forall k \ , \tag{5.103}$$

where $k = (k_1, ..., k_L) \in \mathcal{Z}^L$ is a multi-index labeling the individual Fourier components, and $|k|$ denotes its Euclidean norm. In order to discretize s on a computer, we may truncate this Fourier series, i.E. by replacing the infinite index k with a finite index that truncates at some maximal $k_{\max}$. The operator $\mathcal{F}$ denotes the Fourier transformation and $\mathcal{F}^\dagger$ its corresponding back-transformation (or their discrete versions in case of truncation). The so-called amplitude spectrum A may be identified with the square root of the power-spectrum P_s of the process (specifically P_s being the eigen-spectrum of the linear operator associated with the covariance of the prior probability $P(s)$). Therefore we proceed to construct a model for A rather then P_s as it is more convenient for a generative model. The non-parametric prior model for A is largely built on the assumption that power spectra (and therefore also amplitude spectra) do not vary arbitrarily for similar $|k|$, which in turn allows us to assume that the values of A are, to some degree, correlated. A prominent example of a physically plausible spectrum is a power-law $P_s = |k|^\alpha$ and therefore it turns out to be more convenient to represent A on a log-log-scale, specifically

$$\tau_l \equiv \log\left(A(|k|)\right)\big|_{|k|=e^l} \ , \tag{5.104}$$

since power-laws become straight lines on these scales. As k is a regularly spaced index, the new index $l = \log(|k|)$ is an irregularly spaced index starting from the smallest non-zero mode labeled as l_0 (the origin with $|k| = 0$ is treated separately). To exploit correlations in the prior of τ_l, we define a random process $\tilde{\tau}(l)$ over a continuous domain $O = [l_0, \infty)$ ($O = [l_0, l_{\max} = \log(|k_{\max}|)]$ in the truncated case), and evaluate this process on the irregularly spaced locations on which τ_l is defined. The prior process used for $\tilde{\tau}$ is a Gauss-Markov process given in terms of a linear stochastic differential equation of the form

$$\frac{\partial}{\partial l}\begin{pmatrix}\tilde{\tau}(l)\\ y(l)\end{pmatrix} + \begin{pmatrix}0 & -1\\ 0 & 0\end{pmatrix}\begin{pmatrix}\tilde{\tau}(l)\\ y(l)\end{pmatrix} = \sigma\begin{pmatrix}\epsilon\ \eta_l\\ \xi_l\end{pmatrix} \ , \quad \text{with} \quad {}^{\eta_l}/_{\xi_l} \sim \mathcal{N}({}^{\eta_l}/_{\xi_l}; 0, 1)\ \forall l \in O \ . \tag{5.105}$$

A Markov process can easily be realized on an irregular grid utilizing its transition probability which in this case takes the form

$$P\left(\begin{pmatrix}\tau_l\\ y_l\end{pmatrix}\middle|\begin{pmatrix}\tau_{l_0}\\ y_{l_0}\end{pmatrix}\right) = \mathcal{N}\left(\begin{pmatrix}\tau_l\\ y_l\end{pmatrix}; \begin{pmatrix}1 & \Delta_l\\ 0 & 1\end{pmatrix}\begin{pmatrix}\tau_{l_0}\\ y_{l_0}\end{pmatrix}, \sigma^2\begin{pmatrix}{}^{\Delta_l^3}/_3 + \epsilon^2\Delta_l & {}^{\Delta_l^2}/_2\\ {}^{\Delta_l^2}/_2 & \Delta_l\end{pmatrix}\right) \tag{5.106}$$

with $\Delta_l = l - l_0$. We notice that in absence of stochastic deviations (e.g. if $\sigma = 0$), the solution is a straight line with slope y_{l_0} that determines the exponent of the power-law, and therefore becomes a variable of the model on which we place a Gaussian prior with a negative prior mean (to a priori favor falling power laws). The offset τ_{l_0} becomes, after exponentiation, an overall scaling factor that sets the variance of the stochastic process s. Thus τ_{l_0} (specifically its exponential) is also a variable of the model which we refer to as "fluctuations" in Table 5.1. Similarly, the zero-mode (i.E. $A(|k| = 0)$), which is not included in τ, is set to be a log-normal distributed random variable which we refer to as "offset std.". Finally σ (named flexibility) and ϵ (named asperity) become both log-normal

distributed variables (again see table 5.1) that determine the variance and shape of the deviations of τ from a straight line (i.E. the deviations of A from a power-law).

Acknowledgements

The authors would like to thank Philipp Arras for his detailed feedback on the manuscript, Sebastian Hutschenreuther for his hands on feedback to the early versions of geoVI, Jakob Knollmüller for the development of the MGVI algorithm, and Martin Reinecke for his contributions to NIFTy.

Chapter 6

Conclusion

This thesis studies two of the biggest hurdles faced when attempting to apply Bayesian inference in practice: the construction of an appropriate prior model and the approximation of the resulting posterior distribution. When applied to the inference of physical quantities, the associated physical theories provide a valuable source for prior information. Partially utilizing this physical knowledge to constrain the most relevant aspects of a system, while remaining agnostic regarding all other aspects has been an overarching goal of the methods developed in this thesis. This ensures that the developed models remain uninformative and broad enough to avoid a false bias, i.E. to ensure that the true configuration of a system is well within the set of a priori plausible realizations, while simultaneously being specific enough such that their information together with the observational data allows for a successful reconstruction.

To this end, in chapter 2, a model for the statistically homogeneous correlations of a system defined in space and time is developed. These correlations can be directly related to the average response of the system to random, external excitations which in turn allows to encode the constraint of a causal propagation of this response into a prior model for the correlation structure. Pairing this constraint with the prior assumption of locality, specifically by favoring the most local response consistent with the data over apparent non-local responses, results in a model for the prior correlation structure that appears restrictive enough to enable successful inference in practice. Specifically, an incomplete and noisy observation of a single realization of the system apparently provides sufficient additional information to infer both, the prior correlation structure as well as the specific realization of the system itself, as is demonstrated by multiple numerical examples. In the next chapter an even simpler, less informative prior model is constructed where the prior correlation structure is solely constraint using the statistically homogeneous and also isotropic correlation information of a single (sub-)space. To construct the correlation structure in e.g. space and time, two instances of this model, one for each sub-space, are combined via an outer product. The applicability of this model is demonstrated via a reconstruction of the galactic center of M87 using data provided by the Event Horizon Telescope. Aside from being a valuable scientific result in itself, it provides an impressive validation of the applicability of this prior model as well as the concepts developed in this

work. In addition, since the introduction of those prior models into the software package NIFTy, they have been utilized in numerous applications where they often appear within the context of much more complex models involving e.g. non-linear transformations or a combination of multiple instances of those priors. Their results have not only demonstrated the applicability of these prior models, but also provide evidence for the validity of the underlying idea that, at some level, a statistical description solely based on the correlation structure of a system, paired with a physically motivated prior model for those correlations, appears to be sufficient to allow a successful reconstruction in practice.

Next, in chapter 4, the applicability of the proposed prior models to the task of simulating partial differential equations is discussed. It is demonstrated how a variant of the statistically homogeneous and isotropic prior for correlation structures of the previous chapter can be utilized to refine and improve the spatial discretzation operations involved in numerical PDE simulation. In particular, it is shown that the specific form of these discretization rules is determined by the specification of a prior model for plausible PDE solutions, which in addition also provides an associated uncertainty measure. Furthermore, combining this prior for spatial correlations with a Gauss Markov process in time results in a probabilistic simulation method that retains a similar computational complexity compared to the implicit Euler method, while providing an improvement to its numerical accuracy, as demonstrated using two examples. Furthermore, the probabilistic nature of the proposed method equips it with a natural uncertainty measure to determine its own simulation error, which appears to be accurate so long as the prior correlation structure is appropriately set. Further work needs to be done, however, in order to cast the proposed simulation method into an algorithm that is efficiently applicable to large-scale simulation problems.

The last chapter of this thesis, chapter 5, focuses on the task of approximating posterior probability distributions. In particular, the development and implementation of a variant of approximate inference is discussed, the geoVI algorithm. Its main contribution to variational inference is the construction of a family of approximate distributions that are, by construction, geometrically close to the true distribution. The geometric description of a distribution is given by a manifold with an associated metric tensor that, in some sense, assigns small distances to parameter pairs which are considered to be similar by the distribution and large distances to pairs that are not. While the exact form of this metric tensor remains subject to active research, the empirical evidence provided by the numerical examples suggest that geoVI appears useful to perform accurate approximations of probability distributions in practice. Furthermore, as proven in chapter 5, geoVI reduces to the MGVI algorithm in case all non-linear aspects of the manifold are ignored while providing an improved approximation in case those aspects become significant. Therefore, the numerous successful applications of MGVI[1] provide additional confidence in the applicability of geoVI as the reconstructions are only expected to improve via this extension. In fact, at the time of writing, several ongoing inference projects that utilize NIFTy have switched

[1]Even the results given in chapters 2 - 4 use MGVI simply because geoVI has not been developed at the time the research was carried out.

to the latest version of the package in which geoVI is introduced and have validated the increase in approximation quality also in practice. It will be exciting to see what novel research this increase in accuracy may enable.

To conclude this thesis, it may be reemphasized that, from a purely information theoretical point of view and ignoring all computational constraints, disregarding any form of information is never a desired goal. It does not matter whether this is done by only partially encoding prior knowledge or by approximating the resulting posterior distribution. Both cases are ideally reduced to a bare minimum in practice as they, if done consistently, result in an increase of uncertainty regarding the results and, if done inconsistently, may even result in wrong results. In real astrophysical inference tasks, however, the current state of model complexity and the computational resources available render this hypothetically optimal scenario impossible in almost all cases. Even in the foreseeable future, given the scale of planned future telescopes and the expected development of available computational resources, having a way to successfully perform approximate inference remains highly relevant.

Bibliography

[1] *Time-resolved reconstruction of M87* (Version 1.1)*, 2021.

[2] B. Abbott, R. Abbott, R. Adhikari, P. Ajith, B. Allen, G. Allen, R. Amin, S. Anderson, W. Anderson, M. Arain et al. , *Reports on Progress in Physics* **72** (2009), 076901.

[3] T. Accadia, F. Acernese, M. Alshourbagy, P. Amico, F. Antonucci, S. Aoudia, N. Arnaud, C. Arnault, K. Arun, P. Astone et al. , *Journal of Instrumentation* **7** (2012), P03012.

[4] S. Amari und H. Nagaoka: *Methods of Information Geometry*. Translations of mathematical monographs. American Mathematical Society, 2000.

[5] L. Arnold: *Stochastic Differential Equations: Theory and Applications*. Dover Books on Mathematics. Dover Publications, 2013.

[6] P. Arras, M. Baltac, T.A. Ensslin, P. Frank, S. Hutschenreuter, J. Knollmueller, R. Leike, M.N. Newrzella, L. Platz, M. Reinecke et al. , *Astrophysics Source Code Library* (2019).

[7] P. Arras, H.L. Bester, R.A. Perley, R. Leike, O. Smirnov, R. Westermann und T.A. Enßlin, *Astronomy & Astrophysics* **646** (2021), A84.

[8] P. Arras, P. Frank, P. Haim, J. Knollmüller, R. Leike, M. Reinecke und T. Enßlin, *arXiv e-prints* (2020), arXiv:2002.05218.

[9] P. Arras, P. Frank, R. Leike, R. Westermann und T.A. Enßlin, *A&A* **627** (2019), A134.

[10] E. Bertin und S. Arnouts, *Astronomy and astrophysics supplement series* **117** (1996), 393.

[11] M. Betancourt: *A general metric for Riemannian manifold Hamiltonian Monte Carlo. A general metric for Riemannian manifold Hamiltonian Monte Carlo*, In *International Conference on Geometric Science of Information*. Springer (2013) Seiten 327–334.

[12] M. Betancourt, *arXiv preprint arXiv:1701.02434* (2017).

[13] J. Biretta, W. Sparks und F. Macchetto, *The Astrophysical Journal* **520** (1999), 621.

[14] L. Blackburn, D.W. Pesce, M.D. Johnson, M. Wielgus, A.A. Chael, P. Christian und S.S. Doeleman, *ApJ* **894** (2020), 31.

[15] M.R. Blanton, M.A. Bershady, B. Abolfathi et al. , *The Astrophysical Journal* **154** (2017), 28.

[16] D.M. Blei, A. Kucukelbir und J.D. McAuliffe, *Journal of the American statistical Association* **112** (2017), 859.

[17] V.I. Bogachev, A.V. Kolesnikov und K.V. Medvedev, *Sbornik: Mathematics* **196** (2005), 309.

[18] K.L. Bouman, M.D. Johnson, A.V. Dalca, A.A. Chael, F. Roelofs, S.S. Doeleman und W.T. Freeman, *IEEE Transactions on Computational Imaging* **4** (2018), 512.

[19] S. Brooks, A. Gelman, G. Jones und X.L. Meng: *Handbook of markov chain monte carlo*. CRC press, 2011.

[20] N.N. Cencov: *Statistical decision rules and optimal inference*. Nummer 53. American Mathematical Soc., 2000.

[21] A.A. Chael, M.D. Johnson, K.L. Bouman, L.L. Blackburn, K. Akiyama und R. Narayan, *The Astrophysical Journal* **857** (2018), 23.

[22] A.A. Chael, M.D. Johnson, R. Narayan, S.S. Doeleman, J.F. Wardle und K.L. Bouman, *The Astrophysical Journal* **829** (2016), 11.

[23] J. Cockayne, C. Oates, T. Sullivan und M. Girolami: *Probabilistic numerical methods for PDE-constrained Bayesian inverse problems. Probabilistic numerical methods for PDE-constrained Bayesian inverse problems*, In *AIP Conference Proceedings*, Band 1853. AIP Publishing LLC (2017) Seite 060001.

[24] G. Collaboration, *Astronomy and Astrophysics* **595** (2016), A1.

[25] S.L. Cotter, G.O. Roberts, A.M. Stuart und D. White, *Statistical Science* (2013), 424.

[26] R.T. Cox, *American journal of physics* **14** (1946), 1.

[27] L. Devroye: *Sample-based non-uniform random variate generation. Sample-based non-uniform random variate generation*, In *Proceedings of the 18th conference on Winter simulation*. (1986) Seiten 260–265.

[28] C.R. Doering, *Physics Letters A* **122** (1987), 133.

[29] S. Duane, A.D. Kennedy, B.J. Pendleton und D. Roweth, *Physics letters B* **195** (1987), 216.

[30] T.A. Enßlin, *Phys. Rev. E* **87** (2013), 013308.

[31] T.A. Enßlin, *AIP Conference Proceedings* **1553** (2013), 184.

[32] T.A. Enßlin, *Annalen der Physik* **531** (2019), 1800127.

[33] T.A. Enßlin und M. Frommert, *Phys. Rev. D* **83** (2011), 105014.

[34] T.A. Enßlin, M. Frommert und F.S. Kitaura, *Physical Review D* **80** (2009), 105005.

[35] Event Horizon Telescope Collaboration. *First M87 EHT Results: Calibrated Data*, 2019.

[36] Event Horizon Telescope Collaboration et al. , *Astrophys. J. Lett.* **875** (2019), L1.

[37] Event Horizon Telescope Collaboration et al. , *Astrophys. J. Lett.* **875** (2019), L2.

[38] Event Horizon Telescope Collaboration et al. , *Astrophys. J. Lett.* **875** (2019), L3.

[39] Event Horizon Telescope Collaboration et al. , *Astrophys. J. Lett.* **875** (2019), L4.

[40] Event Horizon Telescope Collaboration et al. , *Astrophys. J. Lett.* **875** (2019), L5.

[41] Event Horizon Telescope Collaboration et al. , *Astrophys. J. Lett.* **875** (2019), L6.

[42] R.A. Fisher: *Theory of statistical estimation. Theory of statistical estimation*, In *Mathematical Proceedings of the Cambridge Philosophical Society*, Band 22. Cambridge University Press (1925) Seiten 700–725.

[43] J. Fitzsimons, D. Granziol, K. Cutajar, M. Osborne, M. Filippone und S. Roberts: *Entropic trace estimates for log determinants. Entropic trace estimates for log determinants*, In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer (2017) Seiten 323–338.

[44] C.W. Fox und S.J. Roberts, *Artificial intelligence review* **38** (2012), 85.

[45] P. Frank und T.A. Enßlin, *arXiv e-prints* (2020), arXiv:2010.06583.

[46] P. Frank, R. Leike und T.A. Enßlin, *Annalen der Physik* **533** (2021), 2000486.

[47] P. Frank, R. Leike und T.A. Enßlin, *Entropy* **23** (2021).

[48] P. Frank, T. Steininger und T.A. Enßlin, *Phys. Rev. E* **96** (2017), 052104.

[49] Frank, Philipp, Jasche, Jens und Enßlin, Torsten A., *A&A* **595** (2016), A75.

[50] M.G. Genton, *Journal of machine learning research* **2** (2001), 299.

[51] C.J. Geyer, *Statistical science* (1992), 473.

[52] J.W. Gibbs: *Elementary Principles in Statistical Mechanics: Developed with Especial Reference to the Rational Foundation of Thermodynamics*. Cambridge Library Collection - Mathematics. Cambridge University Press, 2010.

[53] M. Girolami und B. Calderhead, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **73** (2011), 123.

[54] M. Goldman, *Ann. Math. Statist.* **42** (1971), 2150.

[55] W. Grecksch und P.E. Kloeden, *Bulletin of the Australian mathematical society* **54** (1996), 79.

[56] M. Guardiani, P. Frank, A. Kostić, G. Edenhofer, J. Roth, B. Uhlmann und T. Enßlin. *Non-parametric Bayesian Causal Modeling of the SARS-CoV-2 Viral Load Distribution vs. Patient's Age*, 2021.

[57] I. Han, D. Malioutov und J. Shin: *Large-scale log-determinant computation through stochastic Chebyshev expansions. Large-scale log-determinant computation through stochastic Chebyshev expansions*, In *International Conference on Machine Learning*. PMLR (2015) Seiten 908–917.

[58] P. Hennig, M.A. Osborne und M. Girolami, *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* **471** (2015), 20150142.

[59] M.R. Hestenes, E. Stiefel et al. : *Methods of conjugate gradients for solving linear systems*, Band 49. NBS Washington, DC, 1952.

[60] M.D. Hoffman, D.M. Blei, C. Wang und J. Paisley, *Journal of Machine Learning Research* **14** (2013).

[61] M.F. Hutchinson, *Communications in Statistics-Simulation and Computation* **18** (1989), 1059.

[62] S. Hutschenreuter, C.S. Anderson, S. Betti, G.C. Bower, J.A. Brown, M. Brüggen, E. Carretti, T. Clarke, A. Clegg, A. Costa et al. , *arXiv preprint arXiv:2102.01709* (2021).

[63] Hutschenreuter, Sebastian und Enßlin, Torsten A., *A&A* **633** (2020), A150.

[64] E.T. Jaynes: *Probability theory: The logic of science*. Cambridge university press, 2003.

[65] H. Jeffreys: *The theory of probability*. OUP Oxford, 1998.

[66] M.D. Johnson, K.L. Bouman, L. Blackburn, A.A. Chael, J. Rosen, H. Shiokawa, F. Roelofs, K. Akiyama, V.L. Fish und S.S. Doeleman, *The Astrophysical Journal* **850** (2017), 172.

[67] K. Karhunen: *Über lineare Methoden in der Wahrscheinlichkeitsrechnung.* Annales Academiae Scientiarum Fennicae: Ser. A 1. Sana, 1947.

[68] H. Kersting und P. Hennig: *Active Uncertainty Calibration in Bayesian ODE Solvers. Active Uncertainty Calibration in Bayesian ODE Solvers*, In *Proceedings of the 32nd Conference on Uncertainty in Artificial Intelligence (UAI).* AUAI Press (Juni 2016) Seiten 309–318.

[69] H. Kersting und M. Mahsereci: *A Fourier State Space Model for Bayesian ODE Filters. A Fourier State Space Model for Bayesian ODE Filters*, In *Workshop on Invertible Neural Networks, Normalizing Flows, and Explicit Likelihood Models, ICML.* (2020).

[70] D.P. Kingma und M. Welling: *Auto-Encoding Variational Bayes. Auto-Encoding Variational Bayes*, In *2nd International Conference on Learning Representations.* ICLR (2014).

[71] J. Knollmüller und T.A. Enßlin, *arXiv preprint arXiv:1901.11033* (2019).

[72] J. Knollmüller, P. Frank und T.A. Ensslin. *STARBLADE: STar and Artefact Removal with a Bayesian Lightweight Algorithm from Diffuse Emission*, Mai 2018.

[73] J. Knollmüller, P. Frank und T.A. Enßlin. *Separating diffuse from point-like sources - a Bayesian approach*, 2018.

[74] A.N. Kolmogorov und A.T. Bharucha-Reid: *Foundations of the theory of probability: Second English Edition.* Courier Dover Publications, 2018.

[75] A. Kucukelbir, D. Tran, R. Ranganath, A. Gelman und D.M. Blei, *The Journal of Machine Learning Research* **18** (2017), 430.

[76] S. Kullback und R.A. Leibler, *The annals of mathematical statistics* **22** (1951), 79.

[77] R.H. Leike und T.A. Enßlin, *Phys. Rev. E* **97** (2018), 033314.

[78] Leike, R. H. und Enßlin, T. A., *A&A* **631** (2019), A32.

[79] Leike, R. H., Glatzle, M. und Enßlin, T. A., *A&A* **639** (2020), A138.

[80] M. Loève: *Probability Theory.* Nummer v. 1 in Graduate texts in mathematics. Springer, 1977.

[81] P. Marjoram, J. Molitor, V. Plagnol und S. Tavaré, *Proceedings of the National Academy of Sciences* **100** (2003), 15324.

[82] S.S. Matti Lassas, Eero Saksman, *Inverse Problems & Imaging* **3** (2009), 87.

[83] Milosevic, Sara, Frank, Philipp, Leike, Reimar H., Müller, Ancla und Enßlin, Torsten A., *A&A* **650** (2021), A100.

[84] Müller, Ancla, Hackstein, Moritz, Greiner, Maksim, Frank, Philipp, Bomans, Dominik J., Dettmar, Ralf-Jürgen und Enßlin, Torsten, *A&A* **620** (2018), A64.

[85] K. Nalewajko, M. Sikora und A. Różańska, *A&A* **634** (2020), A38.

[86] J. Nocedal und S.J. Wright: *Numerical Optimization*. second. Auflage. Springer, New York, NY, USA, 2006.

[87] B. Oksendal: *Stochastic differential equations: an introduction with applications*. Springer Science & Business Media, 2013.

[88] N. Oppermann, M. Selig, M.R. Bell und T.A. Enßlin, *Phys. Rev. E* **87** (2013), 032136.

[89] P.E. Protter. In *Stochastic integration and differential equations*. Springer (2005), Seiten 249–361.

[90] M. Raissi, P. Perdikaris und G.E. Karniadakis, *Journal of Computational Physics* **348** (2017), 683 .

[91] M. Raissi, P. Perdikaris und G.E. Karniadakis. *Physics Informed Deep Learning (Part I): Data-driven Solutions of Nonlinear Partial Differential Equations*, 2017.

[92] M. Raissi, P. Perdikaris und G.E. Karniadakis, *SIAM Journal on Scientific Computing* **40** (2018), A172.

[93] C.R. Rao. In *Breakthroughs in statistics*. Springer (1992), Seiten 235–247.

[94] C.E. Rasmussen: *Gaussian processes in machine learning. Gaussian processes in machine learning*, In *Summer School on Machine Learning*. Springer (2003) Seiten 63–71.

[95] N. Reeb, S. Hutschenreuter, P. Zehetner, T. Ensslin, S. Alves, M. André, M. Anghinolfi, G. Anton et al. . *Studying Bioluminescence Flashes with the ANTARES Deep Sea Neutrino Telescope*, 2021.

[96] D. Rezende und S. Mohamed: *Variational inference with normalizing flows. Variational inference with normalizing flows*, In *International Conference on Machine Learning*. PMLR (2015) Seiten 1530–1538.

[97] A.J. Roberts, *ANZIAM Journal* **45** (2003), 1.

[98] A. Rogers, H. Hinteregger, A. Whitney, C. Counselman, I. Shapiro, J. Wittels, W. Klemperer, W. Warnock, T. Clark, L. Hutton et al. , *The Astrophysical Journal* **193** (1974), 293.

[99] A. Saha, K. Bharath und S. Kurtek, *Journal of the American Statistical Association* **115** (2020), 822. PMID: 33041402.

[100] M. Schober, D. Duvenaud und P. Hennig: *Probabilistic ODE Solvers with Runge-Kutta Means. Probabilistic ODE Solvers with Runge-Kutta Means*, In *Advances in Neural Information Processing Systems 27*. Curran Associates, Inc. (2014) Seiten 739–747.

[101] M. Schober, S. Särkkä und P. Hennig, *Statistics and Computing* (2018).

[102] M. Selig, M.R. Bell, H. Junklewitz, N. Oppermann, M. Reinecke, M. Greiner, C. Pachajoa und T.A. Enßlin, *A&A* **554** (2013), A26.

[103] J. Sirignano und K. Spiliopoulos, *Journal of Computational Physics* **375** (2018), 1339 .

[104] V. Šmídl und A. Quinn: *The variational Bayes method in signal processing*. Springer Science & Business Media, 2006.

[105] A.M. Stuart, *Acta Numerica* **19** (2010), 451–559.

[106] H. Sun und K.L. Bouman. *Deep Probabilistic Imaging: Uncertainty Quantification and Multi-modal Solution Characterization for Computational Imaging*, 2020.

[107] F. Tronarp, H. Kersting, S. Särkkä und P. Hennig, *Statistics and Computing* **29** (2019), 1297.

[108] S. Ubaru, J. Chen und Y. Saad, *SIAM Journal on Matrix Analysis and Applications* **38** (2017), 1075.

[109] M.P. van Haarlem, M.W. Wise, A.W. Gunst, G. Heald, J.P. McKean, J.W.T. Hessels, A.G. de Bruyn, R. Nijboer, J. Swinbank und R. Fallows, *A&A* **556** (2013), A2.

[110] C. Welling, P. Frank, T. Enßlin und A. Nelles, *Journal of Cosmology and Astroparticle Physics* **2021** (2021), 071.

[111] M. Westerkamp, I. Ovchinnikov, P. Frank und T. Enßlin. *Dynamical field inference and supersymmetry*, 2021.

[112] N. Wiener: *Extrapolation, interpolation, and smoothing of stationary time series, with engineering applications.* Nummer ix, 163 p. in Stationary time series. Technology Press of the Massachusetts Institute ofTechnology, 1950.

Acknowledgements

I would like to use the opportunity to say thank you to all people who were supporting me throughout the entire time of my PhD.

First and foremost, thank you *Torsten Enßlin* for being the supervisor and mentor that you are. Whatever form of input I was seeking, whether it was a discussion of novel ideas, a fresh view on things when I got stuck, or your incredibly rapid feedback on a manuscript, your door has always been open. I really enjoyed the discussions we had about so many aspects of information theory, its role in science, but also its incarnations in everyday life. Thank you also for what you have built at MPA with the Information Field Theory group. I really enjoyed being a part of it for my entire time at MPA. Even though so many people joined and left the group, and even in the times of a global pandemic, you somehow managed to retain a positive and constructive atmosphere for everyone involved.

Thank you *Reimar Leike*, for being tireless in discussing so many novel, and sometimes really crazy, ideas with me. I will never forget your infamous "five minutes at a blackboard" discussions which, even though they sometimes left me exhausted, have always been an inspiration. My thanks also go to *Philipp Arras*, who has been an amazing office mate throughout my entire PhD. I am very grateful for your expertise regarding so many programming related aspects and working with you has taught me a lot about it.

Special thanks and gratitude goes out to *Martin Reineke* for his tremendous efforts involving almost every aspect of NIFTy as well as most additional numerical aspects of this work. What you do is truly amazing.

In addition, I want to thank all the other PhD students and post-docs that I was fortunate enough to share time with in the group: *Vanessa Böhm, Sebastian Dorn, Vincent Eberle, Gordian Edenhofer, Mahsa Ghaempanah, Maksim Greiner, Johannes Harth-Kizerow, Sebastian Hutschenreuther, Sebastian Kehl, Ivan Kostyuk, Max Newrzella, Jakob Knollmüller, Natalia Porqueres, Daniel Pumpe, Jakob Roth, Theo Steininger, Ann-Kathrin Straub.* It has been a pleasure to work with all of you and I am very thankful for all the time we had together.

My thanks also goes to all the master students I was fortunate enough to co-supervise during their thesis: *Vincent Eberle, Morten Giese, Matteo Guardiani, Sara Milosevic, Margret Westerkamp, and Johannes Zacherl.* I really enjoyed the participation in your projects and it has been amazing to be able to explore so many different aspects of information theory and its applications. I also want to thank all the numerous other students who were an integral part of the group. All of you have contributed to this amazing atmosphere that

I was fortunate enough to experience here.

Finally, I am so grateful for all the emotional support of my family and my friends that I had throughout the entire time. To my parents: Thank you for everything. Without your constant support none of this would have been possible.