Metric Gaussian Variational Inference

Jakob Knollmüller

Munich 2020

Metric Gaussian Variational Inference

Jakob Knollmüller

Dissertation
an der Fakultät für Physik
der Ludwig–Maximilians–Universität
München

vorgelegt von
Jakob Knollmüller
aus Deggendorf

München, den 26. August 2020

Erstgutachter: PD Dr. Torsten Enßlin
Zweitgutachter: Prof. Dr. Thomas Kuhr
Tag der mündlichen Prüfung: 26. Oktober 2020

Contents

List of Figures

List of Tables

Zusammenfassung

Ein Hauptergebnis dieser Dissertation ist die Entwicklung von Metric Gaussian Variational Inference (MGVI), einer Methode zur approximativen Inferenz in extrem hohen Dimensionen und für komplexe probabilistische Modelle. Dazu ist zunächst eine hinreichend flexible Approximation erforderlich, um die tatsächliche Posterior-Verteilung genau genug zu erfassen. Desweiteren skaliert die Anzahl der dafür benötigten Parameter unvorteilhaft mit der Anzahl der Modellparameter. Um beispielsweise die Korrelation zwischen allen Modellparametern explizit auszudrücken, ist ihre quadratische Anzahl von Korrelationskoeffizienten erforderlich. Bei Szenarien mit Millionen von Modellparametern ist dies nicht machbar.

MGVI überwindet diese Einschränkung durch das Ersetzen der explizite Kovarianz mit einer impliziten Approximation, die nicht gespeichert werden muss und auf die über Stichproben zugegriffen wird. Dieses Verfahren skaliert linear mit der Problemgröße und erlaubt es, die vollen Korrelationen auch bei extrem großen Problemen zu berücksichtigen. Aus diesem Grund ist es auch auf wesentlich komplexere Setups anwendbar.

MGVI ermöglichte eine Reihe von ehrgeizigen Signalrekonstruktionen, von mir und anderen, welche vorgestellt werden sollen. Dabei handelt es sich um eine zeit- und frequenzaufgelöste Rekonstruktion des Schattens um das Schwarze Loch M87* mit Daten der Event-Horizon-Telescope Kollaboration, eine dreidimensionale tomographische Rekonstruktion von interstellarem Staub innerhalb von 300pc um die Sonne mit Daten des Gaia Satelliten, medizinische Bildgebungsalgorithmen für Computertomographie, einer Faraday-Rotationskarte des gesamten Himmels mithilfe der Kombination mehrerer Datenquellen und schließlich gleichzeitiger Kalibration und Bildgebung mit einem Radiointerferometer.

Das zweite Hauptergebnis dieser Arbeit ist ein Ansatz zur Kombination mehrerer, trainierter neuronaler Netze um Schlüsse in komplexen Fragestellungen zu ziehen. Deep learning erlaubt es, abstrakte Konzepte zu erfassen, indem man sie aus großen Mengen von Trainingsdaten extrahiert, anstatt sie mathematisch explizit zu formulieren. Hier wird ein generatives neuronales Netz als Prior-Verteilung verwendet, während gleichzeitig bestimmte Eigenschaften über Klassifikations- und Regressionsnetze gefordert werden. Die Schlussfolgerung wird dann in Bezug auf die latenten Variablen des Generators durchgeführt, was mit Hilfe von MGVI oder anderen Verfahren erfolgt. Dies ermöglicht es, neue Fragen flexibel durch Bayes'sches Schließen zu beantworten, ohne dass ein neuronales Netz neu trainiert werden muss. Dieser neuartige Ansatz des Bayes'schen Schliessens mit neuronalen Netzen kann auch mit konventionellen Messdaten kombiniert werden.

Abstract

One main result of this dissertation is the development of Metric Gaussian Variational Inference (MGVI), a method to perform approximate inference in extremely high dimensions and for complex probabilistic models. The problem with high-dimensional and complex models is twofold. Fist, to capture the true posterior distribution accurately, a sufficiently rich approximation for it is required. Second, the number of parameters to express this richness scales dramatically with the number of model parameters. For example, explicitly expressing the correlation between all model parameters requires their squared number of correlation coefficients. In settings with millions of model parameter, this is unfeasible.

MGVI overcomes this limitation by replacing the explicit covariance with an implicit approximation, which does not have to be stored and is accessed via samples. This procedure scales linearly with the problem size and allows to account for the full correlations in even extremely large problems. This makes it also applicable to significantly more complex setups.

MGVI enabled a series of ambitious signal reconstructions by me and others, which will be showcased. This involves a time- and frequency-resolved reconstruction of the shadow around the black hole M87* using data provided by the Event Horizon Telescope Collaboration, a three-dimensional tomographic reconstruction of interstellar dust within 300pc around the sun from Gaia starlight-absorption and parallax data, novel medical imaging methods for computed tomography, an all-sky Faraday rotation map, combining distinct data sources, and simultaneous calibration and imaging with a radio-interferometer.

The second main result is an an approach to use several, independently trained and deep neural networks to reason on complex tasks. Deep learning allows to capture abstract concepts by extracting them from large amounts of training data, which alleviates the necessity of an explicit mathematical formulation. Here a generative neural network is used as a prior distribution and certain properties are imposed via classification and regression networks. The inference is then performed in terms of the latent variables of the generator, which is done using MGVI and other methods. This allows to flexibly answer novel questions without having to re-train any neural network and to come up with novel answers through Bayesian reasoning. This novel approach of Bayesian reasoning with neural networks can also be combined with conventional measurement data.

1 Introduction

The success of modern physics is based on the concept that there is no authority beyond the empirical reality and every physical theory is challenged by the experiment. Strictly following this principle has led to a precise and deep understanding of nature on most scales and environments. To further our understanding, ever more complex and sensitive experiments are conducted, demanding a careful and detailed description of all involved quantities. Higher resolutions in space, time, and energy require more storage and computations. The measurements themselves are noisy, incomplete, or even sparse. Possibly weak signals are buried in other, dominant contributions. Information of multiple datasets and sources needs to be combined to recover the quantity of interest.

All these complications bring traditional data analysis approaches to their limits. A strictly probabilistic formulation of the problem within the Bayesian framework provides a path to describe such problems in all their aspects. Prior knowledge is combined with a detailed description of the entire experimental setup. From this a solution to the problem can be obtained by following Bayes theorem, including uncertainties and correlations on all quantities. In practice, this is typically not feasible analytically and approximate solutions either fail to capture the complexity of the result or already severely struggle with moderately sized problems due to their computational scaling behaviour.

This dissertation introduces Metric Gaussian Variational Inference (MGVI) as a method to approximately solve large-scale and complex Bayesian inference problems. MGVI approximates the true posterior with a Gaussian distribution. Here the issue usually is that the number of parameters, required to specify the correlation structure fully, scales quadratically with the number of model parameters. To avoid this issue, often a diagonal covariance is assumed for larger problems. This is a severe simplification, which is ignorant to any correlations within the true posterior distribution. MGVI does not require to explicitly parametrize the covariance. Instead it uses an expression based on the inverse Fisher information metric evaluated at the mean as an approximation. This quantity is typically only a lower bound the true uncertainty, but it contains correlations between all parameters. We can represent this quantity as an implicit operator, which circumvents the quadratic scaling. It is represented in terms of a series of sparse operations, which allows us to apply the metric to vectors, using efficient computer routines for linear algebra. This allows us to draw samples from our approximate distribution, which are vital for the approximation. MGVI performs a series of consecutive approximations to the posterior, iterating between updating the covariance for a given mean and updating the mean given a covariance. This procedure scales linearly in computational time and memory with respect to the model parameters, while still taking correlations between all parameters into account. MGVI allows to holistically approach extremely large and complex problems, exceeding millions of parameters.

I will showcase five distinct applications of MGVI within different parts of astrophysics and medical imaging. These are only partially based on my work, but it illustrates the wide applicability of the method. MGVI allowed to approach the problems in unprecedented scale and/or complexity. The first example demonstrates a space-time- and frequency-resolved reconstruction around the supermassive black hole M87* using the data provided by the Event Horizon Collaboration. In the second example a three-dimensional reconstruction of the interstellar dust from starlight absorption data and accurate parallaxes, as provided by the Gaia satellite, with exceptional spatial resolution is shown. Following a similar measurement principle, example three is about medical imaging using CT data and an elaborate model that is aware of distinct segments with their own correlation structure. We continue with an improved Faraday rotation map of our galaxy, which fuses data of distinct sources to better constrain the overall morphology, using knowledge on the underlying physics and accounting for possible systematics. In the last example we discuss the simultaneous imaging and calibration of a radio-interferometer.

The second main result of this thesis is an approach to perform reasoning on systems that are far too complex to fully comprehend as human, evading a rigorous mathematical description in terms of first principles. The recent advancements in deep learning allow to capture abstract concepts within trained neural networks to solve certain problems. For them it is enough to have enormous amounts of examples as training data available to master their envisioned task. So far, those networks are not capable to perform tasks outside their initial scope and come up with novel concepts and relations. I managed to combine several independently trained neural networks to solve certain problems by performing Bayesian reasoning, which hopefully is a step towards more flexible and intelligent machines in the future as well as other open questions, such as continuous learning, uncertainty quantification, or reasoning in general. This work was partially enabled by the capability of MGVI to efficiently perform approximate inference in high-dimensional settings.

To achieve this, two popular kinds of trained neural networks were used. Classification and regression networks by now are omnipresent in technological applications, as they can efficiently check for certain traits in the input. Deep generative networks can generate realistic examples according to a distribution underlying the training set. This requires an implicit understanding on the topic at hand, including high-order correlations and non-linear features. The generative model transforms a random input vector, following a simple source distribution, into the system sample. The classification networks can then check whether a certain property is present in this generated sample. Bayesian reasoning allows now to invert this relation to generate samples, following certain, predefined properties. Several constraints can be demanded simultaneously, only requiring networks trained on the individual sub-tasks. Instead of laboriously re-training a conditional generative network, this approach allows to flexibly assemble an analogous Bayesian inference problem from a library of prepared networks.

Developing MGVI was not the original goal of this thesis. It emerged from working on several large-scale and complex Bayesian inference problems. A number of conceptual and numerical steps were necessary to come up with this method. To sketch this journey and to document the progress during the research for this thesis, I want to include two of my earlier papers. These are not the main result of this work, but

provide valuable insight towards MGVI. The first one discusses the problem of separating a number of auto-correlated components from several noisy and incomplete measurements. This problem is omnipresent in astrophysics, as we can only observe the entire Universe at once and isolating individual processes is vital to explore the underlying physics. Improving such techniques was the original goal of my thesis. In hindsight we would formulate the algorithm and model differently, but the main ideas of performing a variational inference using samples from the approximation, as well as how the problem is stated remains the same. Now we would use a standardized model together with MGVI to solve for all quantities simultaneously and taking cross-correlations into account.

A discussion of the conceptual and numerical advantages of standardized models is given in the second additional paper. The standardization is a coordinate transformation of a hierarchical, probabilistic model of continuous distributions to new model parameters that are a-priori independent and follow a simple distribution. All the complexity is then stored in the potentially non-linear coordinate transformations. In this form, a large variety of models have conceptually the same structure, which is the reason for the wide applicability of MGVI.

I also worked on a number of other projects, resulting in further papers, which are publicly available but not published or no longer are pursued to be published in a peer-reviewed journal. These papers are omitted from this thesis for the sake of brevity. They either became obsolete before publication due to further developments or resulted in a dead end, not meriting to further pursue publication. For completeness I want to mention them at this point.

I implemented the ideas discussed in Enßlin and Knollmüller [36] to explore ways to improve the numerical performance of the simultaneous reconstruction of a auto-correlated signal and its correlation. In Knollmüller et al. [77] we explored the same problem in more general settings, including different likelihoods and arbitrary, monotonous modifications to the signals. The starblade algorithm [78, 79] approaches the problem of separating an auto-correlated component from a point-like component in an astrophysical context. Here, no noise is assumed and it was meant as an intermediate step in larger reconstructions to speed up the overall convergence. One issue in scenarios with good data and complex models often is that the prior is weak compared to the likelihood. Therefore in reconstructions, the main goal of the algorithm is to first satisfy the likelihood, irrespective on how plausible the separation into the individual components is. Sorting this out is slow and laborious, as the components always have to satisfy the likelihood. The idea behind starblade was to keep the likelihood constant and only optimize with respect to the prior only on the resulting the sub-manifold, and later continue with the full problem. This approach might be interesting to further pursue in the future, but currently it does not generalize well to other problems.

I also contributed to the python package NIFTy [10, 11, 128]. It allows to build large-scale and complex inference algorithms and provides an implementation of MGVI to solve them efficiently. Most of the examples in this thesis are implemented with the help of NIFTy.

Here, I want to briefly outline the structure of this thesis. After a general introduction to the topic of this theses in the first chapter, I give a brief brief introduction of the key concepts of probability theory and Bayesian inference in the second chapter.

The third chapter outlines conceptually the ultimate challenge of reconstructing the Universe simultaneously in space, time, and frequency, using all available information from all instruments and how this can be broken down by a modular imaging framework. The methods presented in this work could be a crucial step towards such a framework.The fourth chapter contains the paper for MGVI as the first main result of this thesis. In the fifth chapter I showcase several examples from astrophysics and medical imaging from me and others that utilize MGVI. The sixth chapter is about Bayesian reasoning with several independently trained, deep neural networks to approach novel problems, the second main result of this thesis. Chapter seven and eight contain important steps towards MGVI, where the first discusses how to separate and reconstruct auto-correlated components from several noisy and incomplete measurements and the second elaborates on the numerical and conceptual advantages of a standardized probabilistic model. I conclude in chapter nine.

2 Probabilistic Reasoning

This introduction follows loosely the first chapter of the script of the lecture on Model-Based Data Analysis Parameter Inference and Model Testing [25], as well as the book Probability Theory, the Logic of Science [63].

2.1 What is Probability?

Everywhere around us, we experience uncertainty and randomness. Is this a fundamental feature of nature or does it emerge from our ignorance about it? For example, the toss of a coin can be fully described as a classical mechanical system. Given the initial conditions, the behaviour of the system can be (in principle) precisely predicted for all times. There is absolutely no randomness involved. Nevertheless, the coin toss is the primary example of a random system. In practice we do not precisely know the initial conditions of a system, but we still want to describe it. Probabilities allow us to do this. A defining feature of the system is that the coin always lands on one of two sides. We can assign a certain probability for either of the outcomes. This is a highly simplified model of the physical system of the coin toss, but captures its essence. The price of this simplification is uncertainty. It expresses our lack of knowledge on the system, or equivalently, our state of belief.

In the same way, we cannot know what the weather will be after a certain amount of time. There is no intrinsic randomness associated with the temporal evolution of the atmosphere, following the Navier-Stokes equations. The randomness emerges from our incomplete knowledge of the entire system at any given point in time. The chaotic nature leads to exponentially diverging trajectories and the system is unstable under infinitesimal perturbation. These perturbations grow with time, as does our uncertainty about the actual weather.

Embracing the ignorance on a system has lead to one of the most successful theories in physics, namely statistical mechanics. It allows to derive the precise properties of macroscopic systems without having to deal with the overwhelming complexity of the microstates.

Finally, quantum mechanics seems to exhibit true randomness. According to the Born rule, the probability of finding a system in a certain state, when measured, is proportional to the amplitude of the wave equation. But can we understand this in a similar fashion to the previous cases? The Schrödinger equation itself is fully deterministic. Evolving the wave function of the entire Universe will therefore not yield anything unexpected. What is different in the case of quantum mechanics is a fundamental limit on how much we can know about the Universe due to the Heisenberg uncertainty principle. The problem is that the observer himself is part of this Universe. From his subjective perspective from the inside, he cannot know the full picture and his best guess is the probabilistic description, originating from the lack of information.

Because of our limited perspective regarding the world around us, understanding probability is a necessity to understanding nature.

2.2 Probability Theory

There are two main paths towards probability theory. The first one is from a measure theoretical perspective by formulating the rules of probability as axioms by Kolmogorov [81] and the second one is the extension of Aristotelian logic towards uncertainty following Cox Theorem [30]. We will not show equivalence between both approaches, but briefly state the underlying assumptions.

2.2.1 The Laws of Probability

The following laws correspond to the Kolmogorov axioms [81]. Let E be a set of elementary events E_i and $\mathscr{F}$ a σ-algebra on E. Let $(E, \mathscr{F}, P)$ be a measure space with $E \in \mathscr{F}$ and probability measure P.

First axiom The probability of every event $A \in \mathscr{F}$ is non-negative and real.

$$P(A) \in \mathbb{R} \text{ and } P(A) \geq 0 \tag{2.1}$$

Second axiom The probability of the full set E is one.

$$P(E) = 1 \tag{2.2}$$

Third axiom For disjunct sets $A, B \in \mathscr{F}$, $AB = \emptyset$, the probability of the joint sets equals the sum of the individual probabilities.

$$P(A, B) = P(A) + P(B) \tag{2.3}$$

Cox's Theorem [30]

The laws of probabilities, as stated above, can also be derived from the set of the following three assumptions, according to Cox's Theorem. The argumentation is based on plausibilities, which are more general than probabilities, but choosing a certain configuration in the end recovers the usual probabilities. The propositions in the stated form are taken from Arnborg and Sjödin [9].

Divisibility and comparability The plausibility of a statement is a real number and is dependent on information we have related to the statement.

Common sense Plausibilities should vary sensibly with the assessment of plausibilities in the model.

Consistency If the plausibility of a statement can be derived in two ways, the results must be equal.

Note that the common sense assumption includes Aristotelian logic in the limiting case of certainty

2.2.2 Properties of Probabilities

From the laws of probability, a number of useful properties can be directly derived. We will make use of these throughout this thesis and want to state them here briefly.

The Conditional Probability of A given B is defined as follows:

$$P(A|B) \equiv \frac{P(A,B)}{P(B)} \tag{2.4}$$

The Product Rule immediately follows from the definition of the conditional probability:

$$P(A,B) = P(A|B)P(B) \tag{2.5}$$

It states that the joint distribution over two events is the product of the conditional probability of the first event given the second, multiplied by the unconditional probability of the second.

Independence of two events A and B is defined via a factorization of the joint distribution into unconditional probabilities.

$$P(A,B) = P(A)P(B) \tag{2.6}$$

The Sum Rule states how the joint probability depends on the unconditioned probabilities and the probability of their junction.

$$P(A,B) = P(A) + P(B) - P(AB) \tag{2.7}$$

Marginalization describes how the unconditional probability of one quantity can be obtained by a probability-weighted sum of the conditional probabilities. This allows us to remove dependence on certain quantities.

$$P(A) = \sum_i P(A|B_i)P(B_i) \tag{2.8}$$

Bayes Theorem describes how to relate knowledge on one quantity to another one through inverting the conditional relation. It immediately follows from the product rule and reads:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}. \tag{2.9}$$

It is the center piece of probabilistic reasoning and will be the foundation of any further considerations. It states how to combine prior knowledge $P(B)$ on the quantity of interest, B with knowledge on A via the likelihood $P(A|B)$ and the evidence $P(A)$ to obtain the posterior probability $P(B|A)$ as an updated state of knowledge on the system B.

Probability Density So far we considered probability distributions $P(A)$ over sets of elementary event, which corresponds to discrete scenarios. We can extend this towards continuous quantities by introducing probability densities. An event A corresponds then to the continuous random variable a falling into intervals associated to A. The probability $P(A)$ is then given by the integral of the probability density over the intervals.

$$P(A) = \int_{I_A} da\, \mathcal{P}(a) \tag{2.10}$$

The above relations straightforwardly extend towards probability densities by replacing sums with integrals and taking differential volumes into account.

Expectation Values of quantities under probability distributions allow to investigate the behaviour of the system at hand. These are probability-weighted averages of functions that depend on the random variables.

$$\langle f(E_i)\rangle_{P(E_i)} \equiv \sum_{E_i \in E} f(E_i) P(E_i). \tag{2.11}$$

In the continuous case this becomes

$$\langle f(x)\rangle_{\mathcal{P}(x)} \equiv \int_{x \in X} dx\, f(x)\, \mathcal{P}(x). \tag{2.12}$$

Here, X is the domain on which the density is defined.

Commonly used expectation values are the mean $\bar{x}$, in which case the function $f(x) = x$ is the identity, and the variance $f(x) = (x - \bar{x})^2$. These are the first and second moment of the probability distribution and allow to express the mean expectation of the random variable, as well as the typical variation around the mean.

2.3 Bayesian Reasoning

This probabilistic framework allows us to confront our theories with hard, real-world data. From this, we can infer properties of the system at hand and reason about its inner workings. The approach is always the same. We have a model that captures our expectations of the system in general, implemented within the prior distribution $\mathcal{P}(x)$. Here, x summarizes all our model parameters. The relation between data y and the state of our system, as encoded in x is expressed in a likelihood $\mathcal{P}(y|x)$, which contains all details of the measurement. The product of these two distribution has to be normalized by the evidence $\mathcal{P}(d)$ to obtain the posterior distribution $\mathcal{P}(x|y)$, which expresses our updated state of knowledge.

2.3.1 The Coin Toss

We want to demonstrate this approach by applying it to the simple toy model of throwing a coin multiple times. From the series of binary outcomes, either heads or tails, we want to learn the underlying probability μ of either result.

Likelihood

For a single trial the likelihood of a certain result $y \in 0, 1$ in this binary setting is given by the Bernoulli distribution:

$$\mathcal{P}(y|\mu) = \mu^y (1-\mu)^{1-y} \tag{2.13}$$

We assume one fixed rate that does not change from one throw to the next and the results are independent from each other. This means we do not have to care about the order of the results. The only relevant quantity is how many k of the N trials show heads. The likelihood of m, our data, is then given by the Binomial distribution.

$$\mathcal{P}(k|\mu) = \binom{N}{k} \mu^k (1-\mu)^{N-k} \tag{2.14}$$

This is the first ingredient for this problem, relating the observation on k to the underlying rate μ. The next step is to formulate our prior knowledge $\mathcal{P}(\mu)$ on the rate. At this point we have many options and the freedom to implement anything we know about the rate. One property we have to respect is that the rate is bounded within the unit interval. We will discuss three different scenarios, but before this we briefly consider the last missing piece for the inference, the evidence $\mathcal{P}(k)$. Given a prior model $\mathcal{P}(\mu)$, the evidence is obtained via marginalization:

$$\mathcal{P}(k) = \int d\mu \, \mathcal{P}(k|\mu)\mathcal{P}(\mu) \tag{2.15}$$

Finally, the posterior distribution encodes to what degree certain rates are compatible with our prior assumptions and the observed data.

$$\mathcal{P}(\mu|k) = \frac{\mathcal{P}(k|\mu)\mathcal{P}(\mu)}{\mathcal{P}(k)} \tag{2.16}$$

Absolute Certainty

We start with discussing the limiting case of absolute certainty. For whatever reason we precisely know the rate μ^*, say, we are dealing with a fair coin (or at least we belief to know this). In this case, we can express the prior in terms of a delta distribution $\mathcal{P}(\mu) = \delta(\mu - \mu^*)$, which vanishes for all values except $\mu = \mu^*$, is normalized, and becomes infinity if the argument vanishes. In this case, we obtain the posterior:

$$\mathcal{P}(\mu|k) = \frac{\mathcal{P}(k|\mu)\delta(\mu-\mu^*)}{\int d\mu \, \mathcal{P}(k|\mu)\delta(\mu-\mu^*)} \tag{2.17}$$

$$= \frac{\mathcal{P}(k|\mu)\delta(\mu-\mu^*)}{\mathcal{P}(k|\mu^*)} = \begin{cases} 0 & \text{if } \mu \neq \mu^* \\ \infty & \text{if } \mu = \mu^* \end{cases} \tag{2.18}$$

$$= \delta(\mu - \mu^*) \tag{2.19}$$

$$= \mathcal{P}(\mu) \tag{2.20}$$

This tells us that in the case we are absolutely convinced about something, there is no observation that can change our mind. We do not gain any new information from measurements. This shows that certainty is the end of reason and we should be careful about the things we think to know for certain.

Uniform Prior

In this second and more realistic case, we do not want to express preference for any rate. A priori, this corresponds to equal probability for any rate, represented by the uniform distribution $\mathcal{P}(\mu) = 1$ over the unit interval. This leads to:

$$\mathcal{P}(\mu|k) = \frac{\mathcal{P}(k|\mu)1}{\int_0^1 d\mu\, \mathcal{P}(k|\mu)1} \tag{2.21}$$

This requires to solve the following integral in the denominator, which is given by the Beta function $B(\alpha, \beta)$:

$$\mathcal{P}(k) = \int_0^1 d\mu \binom{N}{k} \mu^k (1-\mu)^{N-k} \tag{2.22}$$

$$\equiv \binom{N}{k} B(k+1, N-k+1) \tag{2.23}$$

This gives us the posterior distribution:

$$\mathcal{P}(\mu|k) = \frac{\mu^k (1-\mu)^{N-k}}{B(k+1, N-k+1)} \tag{2.24}$$

This is a Beta distribution, which is bound between zero and one. The general form of the Beta distribution is given by

$$\mathcal{B}(\mu|\alpha, \beta) = \frac{\mu^\alpha (1-\mu)^\beta}{B(\alpha, \beta)} \ . \tag{2.25}$$

Interesting properties are its mean and standard deviation, which allows us to summarize what rate to expect and how certain we are about it. These quantities are given by:

$$\langle \mu \rangle_{\mathcal{P}(\mu|k)} = \bar{\mu} = \frac{k+1}{N+2} \tag{2.26}$$

$$\left\langle (\mu - \bar{\mu})^2 \right\rangle_{\mathcal{P}(\mu|k)} = \frac{(k+1)(N-k+1)}{(N+2)^2(N+3)} \tag{2.27}$$

Interesting is the limiting behaviour towards infinitely many tosses. In this case, the additive constants become irrelevant and the rate approaches the fraction of the number of a certain outcome to the number of total trials. The variance vanishes, as the denominator grows faster than the numerator. In this limit we approach a delta distribution, the case we discussed before. Being certain about something is analogous to having seen an infinite amount of trials before. It is therefore not surprising at all that any further tosses do not have any effect.

Another interesting feature of this system is that the uniform prior in the beginning actually is also a Beta distribution with a certain choice of its parameters, i.e. $\alpha = \beta = 1$. The posterior distribution belongs therefore to the same class of distributions as the prior, making it a so-called conjugate prior. For various likelihood distributions, such a prior is available. These make calculating the posterior distribution especially simple, as only the parameters have to be updated according to some rule. This

allows, for example, to perform continual learning, in which the full information is not immediately availble, but arrives continuously. The intermediate posterior can be used as prior in the next step, allowing predictions the entire time.

Such a conjugate prior, however, is somewhat limiting in terms of knowledge that can be expressed.

Dependence on Further Quantities

So far, we only considered a single rate parameter and the only available information were the outcomes of the tosses. In real-world applications, we tend to collect vast amount of additional information that may or may not impact the result of our experiment. For example, we could have multiple persons throwing the coin, different temperatures, different sides facing top before the throw, launching velocities, angular velocities, et cetera. Let us summarize these additional informations in a vector x for every toss. All these quantities somehow impact the outcome of the toss. We want to learn how. For this, we have to build a model. The likelihood itself is still a binary process with an underlying rate. But this rate is now a function of the additional available information $\mu = f(x)$. However, we do not precisely know this function. Instead, we select a parametric family of functions $f_\theta(x)$ and try to learn the parametrization θ. The outcome of f has to be within zero and one, which can be achieved by applying a sigmoid function σ in a last step. A simple example is logistic regression, where each entry in x_i gets its own coefficient θ_i and everything is added up $\mu = \sigma(x^T\theta)$. Possible extensions are to also consider the interactions between quantities by introducing additional parameters for products in $x_i x_j$. Also, a more sophisticated model, which captures a notion of the underlying physics, can be used. A popular approach nowadays are neural networks that consist of a series of local non-linear functions and linear transformations. These are extremely flexible and allow, given enough data, to learn everything. The inference is then performed in terms of the parameters of the function.

$$\mathcal{P}(\theta|y) = \frac{\mathcal{P}(y|\theta)\mathcal{P}(\theta)}{\mathcal{P}(y)} \tag{2.28}$$

This implicitly also depends on x, but this can be regarded as part of the model. The problem with this inference task is the normalization of the posterior, i.e. the evidence.

$$\mathcal{P}(y) = \int d\theta\, \mathcal{P}(y|\theta)\mathcal{P}(\theta) \tag{2.29}$$

Up to special cases, this integral is usually intractable and numerical integration is only feasible for low-dimensional problems. One way around this is to only approximately solve for the posterior distribution. Popular approaches are point estimates by finding the most likely parameter configuration, approximating the posterior distribution with a simpler one, or sampling techniques that draw samples from the true posterior distribution. The first one is relatively fast to obtain, but tends to perform poorly in complex models. The sampling techniques do converge towards the true posterior, but they tend to be computationally extremely expensive. This thesis will present a method how to approximate the posterior with another distribution using variational inference.

2.3.2 Variational Inference

Here, we want to briefly outline the key concept of variational inference. Consider our true posterior distribution $\mathcal{P}(\theta|y)$ and another probability distribution $\mathcal{Q}_\eta(\theta)$ parametrized in terms of variational parameters η. A way to measure the similarity between two distributions is to compare their information content, or overlap. This quantity is given by the Kullback-Leibler divergence (KL) [86].

$$\mathcal{D}_{\mathrm{KL}}(\mathcal{Q}_\eta(\theta)||\mathcal{P}(\theta|y)) \equiv \int d\theta\ \mathcal{Q}_\eta(\theta) \ln \frac{\mathcal{Q}_\eta(\theta)}{\mathcal{P}(\theta|y)} \tag{2.30}$$

$$= \langle \ln \mathcal{Q}_\eta(\theta) \rangle_{\mathcal{Q}_\eta(\theta)} - \langle \ln \mathcal{P}(\theta|y) \rangle_{\mathcal{Q}_\eta(\theta)} \tag{2.31}$$

Note that this quantity is asymmetric in terms of its arguments. The order stated above corresponds to the variational KL. Exchanging the posterior and the approximation corresponds to moment matching, but requires to calculate expectations under the posterior. If we could do this, there would not be a need to approximate. We are therefore considering the variational KL, which only requires expectation values under the approximation. This is only a local approximation of the posterior and tends to under-estimate uncertainties. Mathematically, it is the expectation value of the logarithmic ratio of both distributions over the variational distribution.

Minimizing the KL-divergence with respect to the variational parameters η provides then an approximation to the true posterior.

$$\frac{\partial \mathcal{D}_{\mathrm{KL}}}{\partial \eta} \stackrel{!}{=} 0 \tag{2.32}$$

The KL-divergence still contains the true posterior distribution, including the evidence. However, to achieve the goal above, this term is not required. Using Bayes theorem, we can expand the KL-divergence.

$$\mathcal{D}_{\mathrm{KL}}(\mathcal{Q}_\eta(\theta)||\mathcal{P}(\theta|y)) = \langle \ln \mathcal{Q}_\eta(\theta) \rangle_{\mathcal{Q}_\eta} - \langle \ln \mathcal{P}(y|\theta) \rangle_{\mathcal{Q}_\eta} - \langle \ln \mathcal{P}(\theta) \rangle_{\mathcal{Q}_\eta} + \langle \ln \mathcal{P}(y) \rangle_{\mathcal{Q}_\eta} \tag{2.33}$$

$$= \langle \ln \mathcal{Q}_\eta(\theta) \rangle_{\mathcal{Q}_\eta} - \langle \ln \mathcal{P}(y|\theta) \rangle_{\mathcal{Q}_\eta} - \langle \ln \mathcal{P}(\theta) \rangle_{\mathcal{Q}_\eta} + \ln \mathcal{P}(y) \tag{2.34}$$

The last term is constant in θ and therefore also in η. We can simply drop this term and optimize the remaining terms of the divergence. This makes it possible to approximate the posterior distribution without the necessity of having the evidence available.

The choice of $\mathcal{Q}_\eta(\theta)$ highly impacts how well the true posterior distribution can be captured. This is always a trade-off between accuracy and available resources. In principle, it is possible to perfectly recover the true posterior, e.g. by using renormalizing flows [112], but this involves an enormous amount of variational parameters. Especially in the case of high-dimensional inference problems, computational feasibility is paramount and one reverts to mean-field approximation, which assume independence between all parameters. This is a severe simplification and obtained uncertainties might be unreliable. In one of the next sections of this thesis we will discuss a particular choice of the approximate distribution that captures correlations between all model parameters, while scaling linearly with the problem size. Taking these correlations into account allows to approach problems with millions of parameters and complex relations, as we will demonstrate throughout the remaining thesis.

3 A Picture of The Universe

The Universe around us contains a large variety of rich phenomena. We not only see how it looks today, but also its origins in the furthest distances, as well as its evolution. From the perspective of Earth, everything appears projected onto the celestial sphere. An ever growing fleet of telescopes, on Earth and on satellites, has revealed the physical origins of many of those phenomena. By now, almost every part of the electromagnetic spectrum is covered by some instrument and novel insights require the fusion of data from multiple sources. The emerging field of multi-messenger astronomy utilizes information complementary to the electromagnetic spectrum in form of gravitational waves and neutrinos to study the processes at their violent origin. The Bayesian framework provide a path how to combine all this information consistently.

All these telescopes are linked through looking at the same physical object, the Universe. They might observe it at different locations, frequencies, or times, but they all probe a part of some consistent physical reality. At least in principle, it has to be possible to combine the information from all the different instruments into one coherent picture of the Universe. Lets imagine for a moment that would be possible to do. This imagined complete Universe picture would require at least five dimensions (three spatial, one temporal, and one spectral) and incredible resolution in every direction due to the highly sensitive instruments. Despite its size and complexity, it would only be a picture of the Universe and not a full description of the Universe itself. In the same sense that the picture of a landscape only shows us what we see and not the highly complex geological and biological processes that lead to its emergence. The picture, however, might show us clues about the underlying processes and principles that shaped our Universe and we can study it to further our understanding of the world.

For all practical purposes, this holistic approach of reconstructing a picture of the entire Universe from all available data of all telescopes is far too ambitious to be solved at once. Enormous datasets, complex measurements, high resolutions in five dimensions, high-dimensional posterior inference, et cetera, will provide challenges for generations of scientists to come.

Instead of trying to solve the entire problem at once, we want to build a modular system, consisting of building blocks that can be assembled to solve certain sub-problems. For example to perform imaging for individual instruments, separating astrophysical components, analysing the spectra or temporal evolution of sources, etc. To solve these individual tasks, certain computational blocks for the description of instruments and astrophysical components are required. Adding them successfully to the framework allows to gradually grow the over-all capabilities. For example, being able to perform a temporal and a spectral analysis of a source enables a spectral-temporal analysis that might reveal interesting aspects of its dynamics that otherwise would have been missed in spatial-only and temporal-only analyses. Having descriptions of

two instrument from the individual reconstructions allows us to fuse their data into one single, improved reconstruction.

Such a system needs to have three ingredients. First of all, a probabilistic description of every measurement process involved would be needed. This would contain models of the used instruments in a computer, digital twins of the real devices. If exposed to a simulated sky, such digital twins will provide virtual data, which the instrument could have produced in case this sky would have been reality.

The second ingredient would be a probabilistic description of plausible sky configurations. The sky meaning the sky on any subset of the three spatial, the temporal and spectral dimension.

This description can implement typical properties of an astrophysical sky, for example the positivity of flux. We have to be careful not be too confident about the properties we impose on the sky, as this might blind us for novel and unexpected features. Here, our strategy is to start with a phenomenological description in the beginning and then refine the model to include more and more knowledge on physics, for example the spectral shape of certain emission processes as a function of underlying physical quantities, such as temperatures, densities, compositions of the emission regions.

The first two ingredients would specify the probabilistic inference problem. The last ingredient would be a way to solve it. The inference problem is possibly high-dimensional and involves complex models with non-trivial interactions between the different parts. This thesis tries to provide such a third ingredient by proposing a probabilistic inference method that is well scaling with the large number of degrees of freedom the envisaged imaging problems involve. The method proposed in the next chapter tries to approach such tasks by approximating the posterior distribution with a special Gaussian.

The problem of recovering signals from noisy and incomplete data is not unique to astrophysics. Similar challenges are posed in other fields that rely on imaging and signal reconstruction. Examples are medical-, geo-, or bio-imaging applications. These require their domain-specific modules, but are conceptually inter-operable with the rest of the framework. We demonstrate the reconstruction of Computed Tomography data in Sec. 5.3.

In the following, I want to illustrate a number of such building blocks that can be re-used throughout many different imaging tasks in astrophysics.

3.1 Astrophysical Signals

Typically, our telescopes probe a certain part of the electromagnetic spectrum and therefore sense some kind of photon-emitting density in space, time, and frequency. This emissivity contains the imprint of many physical processes taking place in our Universe. The first step is to develop a simple, phenomenological description of this density. We always have two ways to improve the fidelity of the reconstructed picture. Either we collect more data, or we use more sophisticated models of the signals. A simplistic model, such as developed in the following, already allows us to combine multiple instruments and thereby increase the amount of and scientific yield from the available data. Another advantage of phenomenological models is that they are based

on generic concepts that re-appear throughout the spatial, spectral, and temporal domain, for example correlations and positivity. More physical models require deeper insights into the concrete scenario and they have a higher risk to miss unexpected features. Including such physical refinements would be the second step, which in return would allow us to develop an even deeper understanding than the phenomenological approach will offer. With such a physics based data interpretation we would no longer only image the emissivity itself, but start to learn about the underlying physical quantities directly from raw data. For example, if we assume a thermal spectrum, we can assign temperatures that translate to emissivity. If this is the only complexity we would allow for, we were blind to any non-thermal effects. A conceptually simpler, phenomenological model that only relies on spectral correlation can reproduce the thermal spectrum, as well as additional features. This, however, does not tell us directly a temperature and requires an additional interpretation step to obtain such from it.

We can coarsely classify the appearance of the phenomena and objects in the Universe in two categories: Point-like sources and extended structures. Point-like sources are everything that is bright enough to be visible to us, but too far away to be spatially resolved, for example distant stars, Active Galactic Nuclei (AGN) [106], pulsars [94], etc. Extended structures can be spatially resolved, either due to our proximity or their large scale. In this class, we have for example various components in our galactic environment [5], the distribution of galaxies in the cosmic web [20], or the cosmic microwave background as a remnant of the Big Bang [127].

Point sources are often the dominant category throughout the electromagnetic spectrum and several of them show interesting structures in the temporal direction. However, those appear to us as simple objects in the spatial domain, as they are only seen as bright spots at a certain location, following a characteristic brightness distribution that originates from their location in the Universe. In contrast to that, the extended structures are far more intricate due to the complex, hierarchical structure formation processes from the largest scales to our galactic neighbourhood. One fundamental feature of such extended emission structure is their spatial correlation. Knowing the emission at one location contains information about the emission in its vicinity. This correlation can be described in terms of characteristic length-scales. Another phenomenological observation is the exponential brightness variation on linear spatial scales. Mathematically, one can use non-linear Gaussian processes to implement these features [37]. Many objects exhibit structures that go far beyond simple descriptions via only two-point correlations. Examples are galaxies with spiral arms and jets, filamentary structures in gas and dust, characteristic shapes of supernova remnants, etc. More elaborate models can capture such higher-order correlations, but quickly become complex and not practical for large reconstructions. One way around this might be presented in Chapter 6, where I present a way to utilize deep learning techniques to represent complex systems. Especially interesting is also to include differential equations into the model to re-create parts of the formation of their structures [62]. One huge issue is the superposition of all these objects and phenomena on the celestial sphere.

So far, we talked about some characteristic spatial features of the constituents of our Universe. It appears in a completely different light when we consider the spectral direction. In every direction, we observe the superposition of all emission and absorp-

tion processes along the line of sight. An overview of important emission processes throughout the electromagnetic spectrum can be found in Rybicki and Lightman [120]. In analogy to the spatial direction, we can again coarsely characterize the individual emission processes in two categories. First, continuous spectra that are smooth and exhibit correlation along the spectral direction. Examples of such are thermal spectra, synchrotron emission of a population of relativistic electrons in radio, or Inverse Compton scattering processes in X- and γ-ray. Often, these follow power-laws with cut-offs related to the environment. The second category are line-emissions that originate from quantum-mechanical transitions in atoms and molecules and are narrowly concentrated at a certain photon energy. These are the fingerprints of the chemical composition of an objects, tracing relative abundances of elements or molecules. In the laboratory we can precisely measure the properties of those transitions. Shifts of these lines in the spectrum provides insight in the relative motion to us due to the Doppler effect. The 21cm hydrogen-line, a hyperfine transition, is important to trace the rotation of distant galaxies and it could possibly illuminate the time between cosmic recombination and reionization in the young Universe [107]. Phenomenologically, an arbitrary mixture of such emission processes will consist of a smooth contribution, as well as these characteristic lines. In fact, we can use the same building-blocks as for the spatial direction, this time only in 1D, to model such mixtures. In contrast to the spatial morphology of sources, we do have a good understanding on many of these processes as we can re-create them here on Earth and precisely measure their properties. This knowledge about spectral characteristics of the physical emission processes allows us to disentangle the observed astrophysical spectra into their constituents for a better physical understanding. Especially in the spatio-spectral imaging scenario, we can use this knowledge to separate the sky into its different astrophysical constituents. This allows to isolate these constituents and to study their properties in isolation.

Absorption also takes place in this spectral dimension and it shares many similarities to emission, i.e. there are processes with continuous absorption coefficients, as well as line-absorption, also due to atoms and molecules. In contrast to the emission, absorption requires an illumination from another source to become evident. Stacking multiple sources and absorbers along a line of sight might yield complex spectra. The phenomenological model can capture such, but their interpretation is up to further analysis.

The final direction to be discussed is the temporal domain. Many processes take place on timescales far longer than the observational period, or even the human lifespan. To us, these events appear to be static. This is due to the enormous spatial scales and comparably low speeds. Our Galaxy spans about 100000ly in diameter and most objects move significantly below the speed of light. Fast temporal variations in brightness therefore require processes confined to small volumes, compared to typical speeds in the system. This is the case for a number of point-sources. For example, AGNs are driven by the accretion of matter onto a supermassive black hole in the center of galaxies. These objects only extend up to several light-days, but in the jet, matter approaches the speed of light. This allows us to observe variations on the scale of days or below. Similarly, pulsars are rapidly spinning neutron stars that emit beams of radiation along its magnetic poles, providing a highly periodic signal. This appears to us as a rapidly varying point source, as the neuron stars themselves only exhibit

diameters up to tens of kilometers. Flares of various kinds might show complex temporal evolutions or quasi-periodic oscillations related to the physical processes and environments. A continuous evolution exhibits temporal correlations, which allows us to describe this process again in terms of its characteristic correlation lengths. Periodicity is also a form of strong temporal correlation associated to characteristic frequencies. In the Fourier spectrum of this signal, these are characterized in terms of sharp features at certain locations, analogous to point-sources in the spatial domain or spectral lines in the spectral domain.

In the emerging field of multi-messenger astronomy, photons are complemented the with additional carriers of information, e.g. neutrinos [28] or gravitational waves [2, 3]. They do not appear in the electromagnetic spectrum and they span their own additional direction. Combining all these messengers can no longer be done only in terms of a picture, but require physical models that associate those with one another.

3.2 Astrophysical Data

Astrophysical instruments come in many forms and sizes, coping with the technical challenges to observe certain ranges of the electromagnetic spectrum, or other messengers. On an abstract level, they all probe the sky emissivity in a certain range in time, direction, and energy. Mathematically, the instrument performs operations to the sky, i.e. various transformations and selections, which in the end relates the sky to the data. For example, a lens distorts light passing through. This process follows the rules of linear optics. As light from the sky travels towards the detectors in the telescopes, several such operations might occur. We will therefore describe the instruments as a series of operations in the order they affect the travelling light. We want to give a brief overview of common measurement principles and their challenges.

Modern optical telescopes collect light over large areas through advanced mirror systems, measure the intensity in photosensitive semiconductors, and store the data digitally. Adaptive optics allows to correct for atmospheric distortions and enables large telescopes on the surface of Earth [114]. Similar principles can be applied to the UV and infrared spectrum, but for these wavelengths the atmosphere becomes intransparent. Mounting these telescopes on satellites circumvents this issue, but makes them significantly more expensive. This is also necessary for even higher wavelengths in the X- and γ-ray range. Measuring these ionizing radiations relies usually on detecting individual photons and following their trajectory through the detector, from which incoming directions and energies can be derived. The the highest-energy γ-rays can be measured by observing Earths atmosphere for particle showers. Similarly, neutrino telescopes observe large volumes of water or ice for interaction to later on reconstruct the event.

The previous techniques relied on the particle nature of photons, but for lower energies and larger wavelengths the wave properties become more important. Interferometry is a highly sensitive technology that makes use of the phase information, not only the amplitude of the electromagnetic field. Consider two antennas that record the electromagnetic field. We can super-impose the feeds with a certain phase that corresponds to a location in the sky. Everything from that direction interferes constructively and many other directions cancel out each other. Mathematically, the

antenna pair probes one location in the Fourier-transformed sky brightness. The farther they are apart (and the shorter the wavelength), the higher their resolution is. Large antenna arrays allow many combinations between pairs of antennas, filling in the Fourier plane and thereby providing more information on the sky. As Earth revolves around itself, the position of an array relative to the source changes, filling in even more of the missing information. For a exhaustive discussion see Richard Thompson et al. [113]. Interferometry is mainly used for the radio spectrum, but technical advancements allow its application also in the visible range, starting to yield extraordinary results and could provide unimaginable spatial resolutions in the future [4]. Sec. 5.1 and Sec. 5.5 showcase applications using radio-interferometers.

In some cases, we can measure distances to objects, which allows us to learn something about the three dimensional structure of our surroundings. Examples are parallaxes to nearby objects, various standard candles, or redshift. This data can then be used to reconstruct three-dimensional spatial maps of either the sources themselves, or quantities along the lines of sight through absorption processes. In Sec. 5.2 a three-dimensional reconstruction of dust using starlight absorptions and parallaxes is shown.

One challenge throughout all measurement principles and instruments is calibration. For any kind of analysis, we have to relate the observed data to the processes on the sky, but we do not have perfect knowledge on the internal state of the instrument. This can be due to defects in some of its components. If they are static, we have a chance to learn them and correct for them in the observation. Many properties, however, change with time or have to be newly determined for every observation. Examples are thermal contractions, telescope pointings, Space- and/or Earth-weather, degradations, etc. Calibration is the process of determining all these effects. Conceptually, this is just another inference problem. Given a model of the instrument, we want to learn its internal state from noisy and potentially incomplete data. Similarly to our sky-models, these quantities may vary in spatial directions, for example defects in the optics, temporal directions, e.g. through the atmosphere, or energy directions (some dispersion). The associated technical processes are often extremely complex and usually not highly relevant on their own. Here, phenomenological models are often sufficient for calibration. We can make use of the same building blocks as for the sky model to characterize them, for example time-continuous variations, temporarily localized interference, constant backgrounds, etc. Often for this, it is helpful to observe a known source as calibrator and adjust the respective calibration parameters, but this requires valuable observation time. To reduce this liability, several methods for self-calibration have been proposed. These try to simultaneously perform the imaging and calibration. In the Bayesian framework, this simply requires to extend the sky model by an additional instrument model, which introduces all the calibration parameters to the overall inference problem. This is analogous to adding additional sky component. The image, together with the calibration solution are then provided in terms of the posterior distribution of the overall problem. An example with radio-interferometry data is shown in Sec. 5.5.

This becomes especially interesting when we start to combine multiple instruments. In combination, they provide more information on the sky, which, in turn, allows to obtain a better calibration for either instrument, further enhancing the reconstructed sky image. This follows the Big Data paradigm that more data is not just more data,

but better data. Combining more and more instruments in this fashion allows us to tap into the full potential of already available data.

3.3 A Universial Bayesian Imaging Kit

How do we build a system that can do all these things? The dry answer is modularity and pre-defined interfaces. This allows to build a large library of modules, as well as a straightforward path to extend the framework. Every telescope of a certain class will follow a common model. For example, most space-based telescopes have some pointing, exposure, sensitivity, points-spread, etc. All of those correspond to certain mathematical operations, our fundamental building blocks. Composing them in a certain way implements this instrument, a higher-level building-block. Other telescopes might have different operations or orders. Implementing a new instrument therefore requires the arrangement of existing building blocks, or the creation of new ones. The latter can be added to the pool of already existing modules to extend the framework. The same holds for the sky models, where the phenomenological part strongly relies on correlation modules or localized features of some kind. Every building block requires the tedious implementation of all technicalities and underlying concepts, but afterwards they are available to all other methods.

This is the underlying idea of the Universal Bayesian Imaging Kit (UBIK). It allows to build complex and tailor-made models from pre-built modules to approach signal reconstruction and imaging challenges in space, time, and energy, in astrophysics and elsewhere. At least in principle, it allows the joint analysis of all available data with the high-fidelity and state-of-the-art Bayesian inference methods. The first building blocks are available and several of the mentioned combinations of instruments and signal dimensions have been successfully demonstrated and a collection of them will be presented in Chapter 6, which make use of the inference engine as presented in the next chapter.

So far, only a handful of digital instrument or instrument type twins are available, as well as signal models. For the future, it is intended to extend this significantly in every direction to increase the overall capabilities of UBIK. With increasing complexity and dimensionality of the models, we encounter new problems. The parallelization and data-handling becomes an issue that also has to be addressed. Full parallelization on diverse computing infrastructures, including novel GPGPU and tensor capabilities, requires to split problems along multiple axes. With the complexity, also the number of hyper-parameters increases and setting them becomes harder due to unintuitive interactions. In the end, we have to solve high-dimensional and non-linear optimization problems, which are hard on their own. Our envisioned signal-dimensions with space, time, and energy in high resolution becomes an issue due to the curse of dimensionality. Along certain axes, but also in large parts of an image, not much is happening. Examples are static large-scale features, or small objects embedded in nothingness. We only need resolution at locations where something interesting is happening. This requires the development of sparse and adaptive signal representations, potentially on non-equidistant grids.

These are just a few challenges that need to be addressed in going forward with UBIK. All of them are independent features and can be broken down into sub-

problems, which are relevant problems on their own. Solving them one after another allows to approach larger and more complex inference problem, allowing us to get ever closer to the original vision of reconstructing a picture of the Universe.

4 Metric Gaussian Variational Inference

This chapter is used as a publication currently submitted to the Journal of Machine Learning Research [76]. My contribution includes the development, implementation and testing of the idea and all examples. I also wrote the contents. Torsten Enßlin was involved in all discussions and provided valuable feedback on the entire manuscript. All authors read, commented, and approved the final manuscript.

4.1 Abstract

Solving Bayesian inference problems approximately with variational approaches can provide fast and accurate results. Capturing correlation within the approximation requires an explicit parametrization. This intrinsically limits this approach to either moderately dimensional problems, or requiring the strongly simplifying mean-field approach. We propose Metric Gaussian Variational Inference (MGVI) as a method that goes beyond mean-field. Here correlations between all model parameters are taken into account, while still scaling linearly in computational time and memory. With this method we achieve higher accuracy and in many cases a significant speedup compared to traditional methods. MGVI is an iterative method that performs a series of Gaussian approximations to the posterior. We alternate between approximating the covariance with the inverse Fisher information metric evaluated at an intermediate mean estimate and optimizing the KL-divergence for the given covariance with respect to the mean. This procedure is iterated until the uncertainty estimate is self-consistent with the mean parameter. We achieve linear scaling by avoiding to store the covariance explicitly at any time. Instead we draw samples from the approximating distribution relying on an implicit representation and numerical schemes to approximately solve linear equations. Those samples are used to approximate the KL-divergence and its gradient. The usage of natural gradient descent allows for rapid convergence. Formulating the Bayesian model in standardized coordinates makes MGVI applicable to any inference problem with continuous parameters. We demonstrate the high accuracy of MGVI by comparing it to HMC and its fast convergence relative to other established methods in a number of examples. We investigate real-data applications, as well as synthetic examples of varying size and complexity and up to a million model parameters.

4.2 Introduction

Performing Bayesian inference in large and complex models is challenging. Analytic posteriors are not available for non-conjugate models and only approximate solutions

are possible. Depending on the requirements and resources, a large variety of approaches is available. MCMC sampling techniques recover the true posterior exactly in the limit of infinite samples, but are computationally expensive. An efficient variant is Hamiltonian Monte Carlo (HMC) [34], which explores the posterior distribution following the Hamilton equations. The choice of the parameter coordinate system is also relevant, as it is a way to decouple the different quantities. To increase sampling efficiency, Betancourt and Girolami [15] proposes to choose a standardized coordinate system, in which the deep hierarchical structure of the problem is resolved and flattened down. Here the reparametrization trick [69] is applied to the model parameters directly.

A completely different approach to solve the inference problem is calculating the Maximum Posterior estimate (MAP). To obtain it, one only has to maximize the posterior probability, which is far easier than sampling the entire posterior density. This makes the MAP approach still applicable in extremely high parameter dimensions. The problem with it is that it does not provide any uncertainty quantification on its own. It is also sensitive to any multi-modal feature or degenerate direction in the posterior distribution. This results in over-fitting the data realization or delivering implausible parameter configurations. One way to fix the shortcoming of the missing uncertainty later on is the Laplace approximation (for details see Bishop [16]). Here the true posterior is approximated with a Gaussian distribution centered around the MAP estimate. The inverse Hessian of the potential landscape is adapted as covariance estimate. Sometimes also the Fisher information metric is used making it a Fisher-Laplace approximation [53, 66]. This requires, however, that MAP provides a reasonable result in the first place, which in complex models often is not the case.

It is therefore better to take the uncertainty already into account when approximating the posterior distribution. A way to do this is Variational Inference. For a comprehensive review on this topic see Blei et al. [18]. Here a family of parametric probability distributions is selected and the variational parameters are optimized by minimizing the Kullback-Leibler (KL) divergence [86] between the approximate distribution and the true posterior distribution. The KL-divergence measures the average information discrepancy between the two distributions. For large problems, the mean-field approximation is commonly used, which scales linearly with the problem size [22, 80]. The approximate distribution factorizes over all individual parameters, ignoring any posterior correlation. Often Gaussian distributions are chosen as the parametric family, which provide an uncertainty associated to the mean position, making it Gaussian Variational Inference [89, 102]. By explicitly parametrizing the covariance, it allows to express correlations between model parameters. Here the problem is the quadratic scaling of the variational parameters with the dimension of the posterior distribution, limiting full-covariance Gaussian variational inference only to moderately sized problems.

In special cases an exact covariance can be parametrized in terms of a quantity that only scales linearly with the model parameters [102]. The associated optimization problem is harder than in the explicit parametrization, but efficient solvers are investigated [67]. We want to approach problems with a more general structure, where the linear scaling is not necessarily available.

The choice of the coordinate systems of the parameters also matter when approximating the posterior distribution. Combining the previously mentioned standard-

ization with Gaussian Variational Inference, one obtains Automatic Differentiation Variational Inference (ADVI) [85]. The standardization extends one common variational approach to any posterior over continuous model parameters, making it extremely flexible. For high dimensional posteriors, however, one is again restricted to the mean-field approach. To avoid the heavy computational load associated with a full-covariance approach, Linear Response (LR-)ADVI has been proposed [49] to first perform mean-field ADVI, and then to construct an uncertainty estimate around the obtained mean utilizing the inverse Hessian of the KL-divergence as an uncertainty estimate instead of the obtained mean-field variance. This covariance estimate measures the sensitivity of the approximation with respect to small variations in the variational parameters, containing cross-correlation between all quantities. It follows the logic of the Laplace approximation by first obtaining a comparably inexpensive estimate, and then fixing certain shortcomings later on. Here again one relies on a simpler method to find a good-enough solution. The uncertainty is then not self-consistent with the mean estimate. A problem of this covariance estimate is again the scaling behavior. The sparsity of the matrix depends on the number of global parameters, which are collectively informed by multiple likelihoods. This is a problem for e.g. Gaussian process regression, where one data point informs all latent parameters in the standardized formulation.

Here we want to propose Metric Gaussian Variational Inference (MGVI) to perform approximate Bayesian inference to extremely high-dimensional and complex posterior distributions. Instead of trying to fix the correlations between all parameters in the end, we take them into account during the optimization to obtain self-consistent mean and uncertainty estimates. We make use of standardized model parameters, as they permit a uniform treatment of many problems and thereby effectively widen the applicability of the method. MGVI does not directly optimize the KL-divergence for a parametric family, instead it performs a number of subsequent Gaussian approximations to the posterior distribution. It iterates between updating the covariance with a term based on the inverse Fisher information metric evaluated at the mean estimate and updating the mean estimate by minimizing the KL-divergence for this given covariance. This procedure is iterated until the mean estimate is consistent with the uncertainty estimate. The covariance estimate is equivalent to the one used for the Fisher-Laplace approximation, as the inverse Hessian of the posterior information is not a valid covariance at every location due to violated positive definiteness. In comparison to the Hessian of the KL-divergence used as covariance estimate in LR-ADVI, our covariance estimate will also be sparse in terms of global parameters, enabling for example large-scale Gaussian process regressions as part of the model. We achieve linear scaling with the posterior dimension by completely avoiding explicitly constructing the covariance at any time. Instead we draw samples from the approximate Gaussian distribution using implicit operators and numerical solutions to large sets of linear equations. All correlations are then stored implicitly within the sample realizations, which are then used to estimate the KL-divergence and its gradient. For minimizing the KL-divergence we rely on efficient Natural Gradient descent [7, 97]. In order to apply MGVI, a number of conditions have to be fulfilled by the underlying model. First, all parameters have to be continuous, and not discrete. Second, the Fisher information metric of the likelihood requires an accessible eigenbasis, which is e.g. the case for independently sampled data. Third, the true posterior has

to be sufficiently Gaussian, and fourth, the standardizing transformation is locally well-approximated by a linear function and higher order terms can be neglected.

In the numerical experiments we apply MGVI to a wide range of different Bayesian inference problems. We validate the method by comparing results to HMC sampling in a synthetic Poisson log-normal Gaussian process regression and a hierarchical logistic regression problem with US presidential election polling data. We demonstrate the scaling of MGVI by approximating a posterior with more than a million parameters in a binary Gaussian process classification problem with simultaneous kernel learning. In this example we also explore the impact of meta-parameter choices for the method. We also apply MGVI to a non-negative matrix factorization problem with a Gamma-Poisson model on the Frey face data set. Throughout the experiments, MGVI has the highest accuracy in most of the used metrics and is always closest to the HMC estimates. It behaves similarly to full-covariance ADVI, as it captures cross-correlation between all parameters, but is in many cases roughly one order of magnitude faster than even mean-field ADVI, as MGVI relies on natural gradient descent and has only half the number of variational parameters.

4.3 Variational Inference

4.3.1 Bayesian Inference

Bayesian inference in general describes how the knowledge on one quantity of a system affects the knowledge on some other quantity of interest, following Bayes theorem:

$$\mathcal{P}(\theta|d) = \frac{\mathcal{P}(d|\theta)\mathcal{P}(\theta)}{\mathcal{P}(d)} \ . \tag{4.1}$$

The posterior distribution $\mathcal{P}(\theta|d)$ of the unknown quantity θ given some known data d is equal to the likelihood $\mathcal{P}(d|\theta)$ of observing the data given a certain configuration of θ multiplied by the prior distribution $\mathcal{P}(\theta)$. This whole expression is normalized by the evidence $\mathcal{P}(d)$.

Prior knowledge on the system is encoded in the prior distribution. The likelihood describes how the observed data is related to the parameters of the model. The main difficulty arises in the calculation of the evidence to obtain a properly normalized posterior distribution.

Often this normalization is analytically intractable, especially in non-conjugate models, which are more flexible to encode knowledge on the system. In such cases one has to approximate the true posterior distribution, for example via Maximum Posterior (MAP), variational inference, or MCMC based sampling techniques.

Instead of working with probability distributions, it is equivalent to discuss the problem in terms of information $\mathcal{H}$, defined as the negative logarithm of a probability distribution $\mathcal{P}$, i.e. $\mathcal{H}(\dots) \equiv -\ln(\mathcal{P}(\dots))$. Bayes theorem in this perspective reads:

$$\begin{aligned} \mathcal{H}(\theta|d) &\equiv -\ln(\mathcal{P}(\theta|d)) && (4.2)\\ &= \mathcal{H}(d|\theta) + \mathcal{H}(\theta) - \mathcal{H}(d) && (4.3)\\ &\widehat{=} \mathcal{H}(d|\theta) + \mathcal{H}(\theta) \ . && (4.4) \end{aligned}$$

In terms of information, the normalization is an additive constant, independent of the quantity of interest. Leaving these terms out is indicated here by the $\widehat{=}$ sign.

4.3.2 Kullback-Leibler Divergence

Variational inference allows to approximate posterior distributions to complex problems within reasonable timescales [18]. One chooses a parametric family of distributions $\mathcal{Q}_\eta(\theta)$ with the variational parameters η and minimizes the average information discrepancy between the true posterior and the approximation, measured by the Kullback-Leibler divergence [86], with respect to these parameters. The KL-divergence is defined as:

$$\mathcal{D}_{\mathrm{KL}}(\mathcal{Q}_\eta(\theta)||\mathcal{P}(\theta|d)) = \int d\theta\ \mathcal{Q}_\eta(\theta) \ln \frac{\mathcal{Q}_\eta(\theta)}{\mathcal{P}(\theta|d)} \tag{4.5}$$

$$\equiv \langle \mathcal{H}(\theta|d) \rangle_{\mathcal{Q}_\eta(\theta)} - \langle \mathcal{H}_\eta(\theta) \rangle_{\mathcal{Q}_\eta(\theta)} \tag{4.6}$$

$$\widehat{=} \langle \mathcal{H}(d,\theta) \rangle_{\mathcal{Q}_\eta(\theta)} - \langle \mathcal{H}_\eta(\theta) \rangle_{\mathcal{Q}_\eta(\theta)}\ . \tag{4.7}$$

The first term is the cross-entropy between the distributions and the second is the Shannon-entropy of the approximation, where $\mathcal{H}_\eta(\theta)$ is the negative logarithm of the approximating distribution. Expectation values are expressed by $\langle \dots \rangle_{\mathcal{P}(\dots)}$, noting the respective distribution as index. In order to minimize the KL-divergence, the normalization of the posterior is irrelevant, as it does not depend on the variational parameters and can be dropped. The expression in the last line is equivalent to the negative Evidence Lower Bound (ELBO) [16]. The parameter solution of minimal KL-divergence provides the variational approximation of the original problem.

For complex models or approximations we cannot calculate the expectation values analytically, but the KL-divergence can be estimated via samples from the approximation. Together with the reparametrization trick [69], the gradients on the variational parameters can be estimated as well. This way we can minimize the KL-divergence in a stochastic optimization procedure even in high dimensions and analytically intractable expectation values.

When approximating the true posterior with another distribution, certain aspects will be lost. Whether a variational approximation is useful or not depends on the problem-specific requirements and available resources. We want to approach problems with an enormous amount of model parameters and reasonable complexity, in which more accurate methods are unfeasible and variational inference can still provides answers.

4.4 Gaussian Variational Inference

Gaussian Variational Inference [102] describes variational inference with parametrized Gaussians as the approximating family. The Gaussian distribution exhibits a number of convenient properties, while still providing uncertainty and correlation between

parameters. In this case the approximate distribution is

$$\mathcal{Q}_\eta(\theta) = \mathcal{G}(\theta|\bar{\theta}, \Theta) \tag{4.8}$$

$$= \frac{1}{|2\pi\Theta|^{\frac{1}{2}}} e^{-\frac{1}{2}(\theta-\bar{\theta})^\dagger \Theta^{-1}(\theta-\bar{\theta})} \ , \tag{4.9}$$

with variational parameters $\eta = (\bar{\theta}, \Theta)$ and corresponding KL-divergence

$$\mathcal{D}_{\mathrm{KL}}\left(\mathcal{G}(\theta|\bar{\theta},\Theta)||\mathcal{P}(\theta|d)\right) \mathrel{\widehat{=}} \left\langle \mathcal{H}(d,\theta)\right\rangle_{\mathcal{G}(\theta|\bar{\theta},\Theta)} - \left\langle \mathcal{H}_{\bar{\theta},\Theta}(\theta)\right\rangle_{\mathcal{G}(\theta|\bar{\theta},\Theta)} \ . \tag{4.10}$$

In order to perform the variational inference of the parameters, the expression above is minimized with respect to the variational mean $\bar{\theta}$ and covariance Θ parameters. The second term in this equation is the Shannon entropy of the approximate Gaussian with the analytic form

$$\left\langle \mathcal{H}_{\bar{\theta},\Theta}(\theta)\right\rangle_{\mathcal{G}(\theta|\bar{\theta},\Theta)} \mathrel{\widehat{=}} \frac{1}{2}\ln|2\pi e\Theta| \ . \tag{4.11}$$

Here $|\ldots|$ expresses a determinant and e is Eulers' number. Note that this expression is independent of the variational mean parameter $\bar{\theta}$. To efficiently optimize the KL-divergence we require gradient information with respect to the variational parameters. Derivatives with respect to the mean and covariance are simply the expected gradient and curvature over the Gaussian distribution, respectively [102].

$$\frac{\partial}{\partial\bar{\theta}}\mathcal{D}_{\mathrm{KL}} = \left\langle \frac{\partial}{\partial\theta}\mathcal{H}(d,\theta)\right\rangle_{\mathcal{G}(\theta|\bar{\theta},\Theta)} \quad \text{, and} \tag{4.12}$$

$$\frac{\partial}{\partial\Theta}\mathcal{D}_{\mathrm{KL}} = \frac{1}{2}\left\langle \frac{\partial^2}{\partial\theta\partial\theta^\dagger}\mathcal{H}(d,\theta)\right\rangle_{\mathcal{G}(\theta|\bar{\theta},\Theta)} - \frac{1}{2}\Theta^{-1} \ . \tag{4.13}$$

For the mean parameter only the cross-entropy term is relevant and if we were to optimize only with respect to this parameter, we avoid the necessity of calculating determinants of possibly large matrices. Setting the derivative with respect to the covariance to zero, we obtain the following implicit relation:

$$\Theta^{-1} = \left\langle \frac{\partial^2}{\partial\theta\partial\theta^\dagger}\mathcal{H}(d,\theta)\right\rangle_{\mathcal{G}(\theta|\bar{\theta},\Theta)} \tag{4.14}$$

$$= \left\langle \frac{\partial\mathcal{H}(d,\theta)}{\partial\theta}\frac{\partial\mathcal{H}(d,\theta)}{\partial\theta^\dagger}\right\rangle_{\mathcal{G}(\theta|\bar{\theta},\Theta)} - \left\langle \frac{1}{\mathcal{P}(d,\theta)}\frac{\partial^2\mathcal{P}(d,\theta)}{\partial\theta\partial\theta^\dagger}\right\rangle_{\mathcal{G}(\theta|\bar{\theta},\Theta)} \ . \tag{4.15}$$

This relation serves as starting point for Metric Gaussian Variational Inference. We will set up an iterative fixed-point scheme where we start with some initial mean value $\bar{\theta}$, and adapt an implicit solution for the covariance, similarly to the expression above. For this Gaussian distribution we can then optimize the KL-divergence only with respect to the mean parameter, keeping the covariance fixed. Once it is optimized, we update the covariance to the implicit solution for the new mean parameter. This procedure is then iterated until convergence. Unfortunately the right side of the above equation is not necessarily compatible with a covariance, as in general it is not

strictly positive definite. The first term, containing the outer product of first derivatives certainly is. Problematic is the second term, which involves second derivatives of the probability distribution. It might contain negative eigenvalues, harming the overall positive definiteness of the covariance of the Gaussian in this approximation. For this reasons we cannot use this expression. It is also a dense matrix for global parameters, which are collectively informed by common likelihoods. We will instead use a similar expression as covariance based on the inverse Fisher information metric as approximation, which overcomes these limitations.

Often the covariance is parametrized explicitly in terms of another matrix A via $\Theta = AA^\dagger$ to ensure positive definiteness. The problem with an explicit parametrization of the variational covariance is the quadratic scaling in the model parameters. It allows only for moderately sized problems. To overcome this limitation, usually a diagonal covariance is assumed, which is a mean-field approach. A diagonal covariance approximation cannot capture correlations between posterior parameters, severely limiting the expressiveness of the result.

We cannot calculate the KL-divergence for arbitrary problems analytically, but it is always possible to approximate the expectation value through sample averages. Therefore, we optimize a stochastic estimate of the KL-divergence with the corresponding stochastic gradient.

$$\langle \mathcal{H}(d,\theta)\rangle_{\mathcal{G}(\theta|\bar{\theta},\Theta)} \approx \frac{1}{N}\sum_{i=1}^{N} \mathcal{H}(d,\theta^i_*) = \frac{1}{N}\sum_{i=1}^{N} \mathcal{H}(d,\bar{\theta}+\Delta\theta^i_*) \tag{4.16}$$

$$\theta^i_* \sim \mathcal{G}(\theta|\bar{\theta},\Theta)) \quad \text{or} \quad \Delta\theta^i_* \sim \mathcal{G}(\theta|0,\Theta)\ . \tag{4.17}$$

We indicate sample realizations with the lower $*$-index, and note Δ for zero-centered Gaussian samples. Splitting the sample in a mean contribution and Gaussian residual $\theta^i_* = \bar{\theta} + \Delta\theta^i_*$ allows us to adapt the samples to an updated mean, which is the reparametrization trick in its simplest form [69]. In the end we will be following an implicit optimization scheme, as briefly discussed above. For this it is therefore sufficient to obtain residual samples $\Delta\theta^i_*$ to learn only the mean $\bar{\theta}$ of the approximate Gaussian for a given covariance.

4.5 Standardization

Deep hierarchical Bayesian models are used to describe sophisticated models and complex dependencies and they strongly vary throughout different applications. To remove large parts of the problem-specific complexity from the variational inference, we prefer to work in standardized parameter coordinates, following Automatic Differentiation Variational Inference (ADVI) [85]. In hierarchical models, certain parameters might be restricted to only a certain parameter range. Performing the variational approximation with a Gaussian in these original coordinates might not be possible due to the infinite support of the Gaussian distribution. In the standard coordinates all parameters follow a priori a standard Gaussian distribution, removing this complication. This transformation opens the door to apply the here proposed algorithm to any problem with continuous parameters. It might not be necessary to standardize problems with infinite support on all parameters, and there the method should also

work in the original coordinates. We do not want to treat this special case separately and choose the more unified standard parametrization. In the hierarchical formulation the interdependence between the different quantities might be strong, resulting in a numerically stiff problem. The hierarchical structure is resolved by applying the reparametrization trick [69] to the model parameters, leading to a flat model. In the context of HMC sampling, these standard coordinates are also used to explore the posterior more efficiently [15]. These numerical and conceptual advantages also apply to variational inference, especially if the true distribution is well approximated with a Gaussian [75].

Conceptually one takes a likelihood $\mathcal{P}(d|\theta)$ together with a hierarchical prior $\mathcal{P}(\theta) = \mathcal{P}(\theta_1|\theta_2 \dots \theta_N) \dots \mathcal{P}(\theta_{N-1}|\theta_N)\mathcal{P}(\theta_N)$ and performs coordinate transformation to uniform parameters using the multivariate distributional transform $\mathcal{F}^{-1}_{\mathcal{P}(\theta)}(\dots)$ [119]. This uses the inverse conditional cumulative density functions, following the logic of inverse transform sampling [32].

$$u \sim \mathcal{U}(u) \tag{4.18}$$
$$\theta = \mathcal{F}^{-1}_{\mathcal{P}(\theta)}(u) \tag{4.19}$$
$$\Rightarrow \theta \sim \mathcal{P}(\theta) \ . \tag{4.20}$$

We draw samples from the prior distribution by drawing samples u from the uniform distribution $\mathcal{U}(u)$, and processing them through $\mathcal{F}^{-1}_{\mathcal{P}(\theta)}(\dots)$. The sample u has finite support on the unit interval and performing a Gaussian approximation in these coordinates is not sensible. A second transformation to standard Gaussian coordinates enables this. The transformation is given by the cumulative density function of the Gaussian $\mathcal{F}_{\mathcal{G}(\xi|0,\mathbb{1})}$.

$$\xi \sim \mathcal{G}(\xi|0,\mathbb{1}) \tag{4.21}$$
$$u = \mathcal{F}_{\mathcal{G}(\xi|0,\mathbb{1})}(\xi) \tag{4.22}$$
$$\Rightarrow u \sim \mathcal{U}(u) \ . \tag{4.23}$$

The resulting ξ parameters are a priori independent and the entire complexity is encoded in the composition of the two transformations $\theta = \mathcal{F}^{-1}_{\mathcal{P}(\theta)} \circ \mathcal{F}_{\mathcal{G}(\xi,\mathbb{1})}(\xi) \equiv f(\xi)$. The probability distribution and its information in these coordinates are

$$\mathcal{P}(d,\xi) = \mathcal{P}\left(d|f(\xi)\right)\mathcal{G}(\xi|0,\mathbb{1}) \tag{4.24}$$
$$\mathcal{H}(d,\xi) \mathrel{\widehat{=}} \mathcal{H}\left(d|f(\xi)\right) + \frac{1}{2}\xi^\dagger \mathbb{1} \xi \ . \tag{4.25}$$

For the rest of the paper we will indicate standardized parameters with ξ, whereas general parameters are θ. The Gaussian approximation in standard coordinates is denoted as $\mathcal{G}(\xi|\bar{\xi},\Xi)$. This standardization allows us to obtain an uncertainty estimate of a certain structure, which enables us to draw samples from the approximate distribution.

4.5.1 Gaussian Variational Inference in Standard Coordinates

It is often stated that in the case of Gaussian prior distributions for all N parameters, Gaussian variational inference only requires $N + M$ variational parameters to

express the full mean and covariance [102], with M being the number of independent likelihood contributions. With standardization we can express any continuous probability distribution in terms of a standard Gaussian prior and a corresponding transformation. We want to emphasize that this statement does not hold for arbitrary transformations. To be precise, it only holds for a linear mixture of latent variables, followed by a point-wise non-linear function. Consider M independent likelihoods with data d_i, parameters θ_i and their relation to the latent Gaussian parameters $\theta_i = f_i(\xi)$. According to Eq. 4.15, the covariance must satisfy the following relation:

$$\Xi^{-1} = \mathbb{1} + \left\langle \sum_{i=1}^{M} \frac{\partial \mathcal{H}(d_i|\theta_i)}{\partial \xi} \frac{\partial \mathcal{H}(d_i|\theta_i)}{\partial \xi^\dagger} \right\rangle_{\mathcal{G}(\xi|\bar{\xi},\Theta)} - \left\langle \sum_{i=1}^{M} \frac{1}{\mathcal{P}(d_i|\theta_i)} \frac{\partial^2 \mathcal{H}(d_i|\theta_i)}{\partial \xi \partial \xi^\dagger} \right\rangle_{\mathcal{G}(\xi|\bar{\xi},\Theta)} \tag{4.26}$$

$$\begin{aligned} = \mathbb{1} &+ \left\langle \sum_{i=1}^{M} \frac{\partial f_i(\xi)}{\partial \xi} \frac{\partial \mathcal{H}(d_i|\theta_i)}{\partial \theta_i} \frac{\partial \mathcal{H}(d_i|\theta_i)}{\partial \theta_i^\dagger} \frac{\partial f_i(\xi)^\dagger}{\partial \xi^\dagger} \right\rangle_{\mathcal{G}(\xi|\bar{\xi},\Theta)} \\ &- \left\langle \sum_{i=1}^{M} \frac{1}{\mathcal{P}(d_i|\theta_i)} \frac{\partial \mathcal{H}(d_i|\theta_i)}{\partial \theta_i} \frac{\partial^2 f_i(\xi)}{\partial \xi \partial \xi^\dagger} \right\rangle_{\mathcal{G}(\xi|\bar{\xi},\Theta)} \\ &- \left\langle \sum_{i=1}^{M} \frac{1}{\mathcal{P}(d_i|\theta_i)} \frac{\partial f_i(\xi)}{\partial \xi} \frac{\partial^2 \mathcal{H}(d_i|\theta_i)}{\partial \theta_i \partial \theta_i^\dagger} \frac{\partial f_i(\xi)}{\partial \xi^\dagger} \right\rangle_{\mathcal{G}(\xi|\bar{\xi},\Theta)} . \end{aligned} \tag{4.27}$$

It is proposed to parametrize this covariance in the following form:

$$\Xi^{-1} = \mathbb{1} + R^\dagger \Lambda R \; . \tag{4.28}$$

with Λ being a diagonal matrix of dimension M, containing the variational parameters for the covariance. This, however, is only be exact if the standardization has the following form:

$$\theta = f(\xi) = g(R\xi) \; . \tag{4.29}$$

Here R is an arbitrary, matrix and g an arbitrary, point-wise, non-linear function. The first and second derivatives of this function with respect to the parameters reads:

$$\frac{\partial f}{\partial \xi} = \frac{\partial g(R\xi)}{\partial \xi} = g'(R\xi) R \tag{4.30}$$

$$\frac{\partial^2 f}{\partial \xi \partial (\xi)^\dagger} = \frac{\partial^2 g(R\xi)}{\partial \xi \partial \xi^\dagger} = R^\dagger g''(R\xi) R \; . \tag{4.31}$$

The parameter-dependent parts $g'(R\xi)$ and $g''(R\xi)$ are diagonal matrices of dimension M, and the matrix R maps from the N-dimensional parameter space to the M-dimensional space.

We insert these derivatives into the expectation values in Eq. 4.27 and pull out the linear R terms out of the integrals, resulting in an expression of the form

$$\Xi^{-1} = \mathbb{1} + R^\dagger \left\langle X_1(\xi) - X_2(\xi) - X_2(\xi) \right\rangle_{\mathcal{G}(\xi|\bar{\xi},\Xi)} R \; . \tag{4.32}$$

Such a term can be exactly approximated by a parametrization of the form Eq. 4.28, as X_1, X_2 and X_3 are diagonal matrices depending on the parameter.

For more general standardization functions $f(\xi)$, containing a number of consecutive linear and and point-wise non-linear transformations, this is not possible, as only the outermost matrix can be pulled out of the expectation value. So in the general, non-linear case, the number of required variational parameters to express the covariance fully does scale quadratically with the number of model parameters.

4.6 Approximating the Covariance

We want to explore the properties of extremely high dimensional posterior distributions through an efficient approximation. The associated volume in such high dimensional spaces is enormous and in it the posterior might exhibit a rich structure. Capturing the posterior structure within the approximation requires a global perspective on it, involving large numbers of parameters to be learned. Already capturing correlations between all model parameters explicitly requires a memory that scales quadratically with the posterior dimension.

In order to avoid such unfavorable scaling, we have to explore the posterior only from a more local perspective, where we only rely on quantities scaling linearly with dimensions. One example for such an approach is the MAP approach. It, however, is susceptible to implausible results, and to getting stuck in local minima and at improbable parameter configurations of elongated valleys along degenerate directions in complex models. The reason for this is that MAP can be regarded as an approximation to the posterior with a delta distribution, which is highly sensitive to local structures in the information landscape.

To avoid this, we have to account for uncertainty all along the way. We want to do this by using a Gaussian distribution to approximate the posterior, which, in addition to a location, also has a scale. This scale is extremely helpful in maneuvering through the landscape outlined by the posterior, as the Gaussian simply cannot fit into all the small local features and degenerate directions a delta distribution is sensitive to. Only structures of the posterior comparable to its own size or larger couple to the Gaussian.

For this, we have to extract an estimate of the posterior uncertainty from a local perspective. The first thing that comes to mind is the Laplace approximation, which uses the inverse Hessian at the location of the MAP solution as a covariance. It explores locally the curvature of the negative log-posterior and associates strongly curved directions with low uncertainty and vice versa. This Laplace approximation is widely used to extract uncertainties from point estimates, but it fundamentally requires the MAP approach to provide reasonable results in the first place.

For our purpose, we cannot use the inverse Hessian as it is not necessarily a valid covariance outside a mode. A covariance matrix exhibits strictly positive eigenvalues, but the Hessian measures curvature, which has vanishing or negative eigenvalues in plateaus and concave directions, respectively, which both are often encountered in high dimensional and complex models. This is the same reason we cannot use the expression given in Eq. 4.15, the implicit solution to the covariance in Gaussian variational inference. Here one could drop the problematic term, which for approximately

Gaussian posteriors will be small anyway and use

$$\Theta^{-1} \approx \left\langle \frac{\partial \mathcal{H}(d,\theta)}{\partial \theta} \frac{\partial \mathcal{H}(d,\theta)}{\partial \theta^\dagger} \right\rangle_{\mathcal{G}(\theta|\bar{\theta},\Theta)} . \tag{4.33}$$

This is precisely the term used in LR-ADVI [49] to approximate the covariance around the mean-field ADVI mean estimate. It is certainly positive definite and somewhat close to the true covariance, but we cannot efficiently represent it in high dimensions without severe limitations on the used models. It is a dense matrix for global parameters, which are collectively informed by the same data points. One example where the overall problem does not factorize into independent sub-problems is Gaussian process regression in the standardized coordinates. We have therefore no access to its eigenbasis without storing and decomposing the dense (sub-)matrices explicitly, something we cannot afford in large problems. We need access to the eigenbasis to generate samples from the approximation, used for estimating the KL-divergence and its gradient. This term itself is an Gaussian expectation value and can be approximated via samples. However, such a sub-sampled matrix is only invertible if the samples constitute a full basis, requiring at least as many samples as parameter dimensions. This, again, is equivalent to storing an entire matrix directly, and therefore not practical. A covariance of this form, however, will serve as the inspiration for the approximation we will be using.

4.6.1 Fisher Information Metric as Covariance

To approach truly large inference problems we require three fundamental properties from the covariance approximation. First, it has to be strictly positive definite, a defining feature of any covariance. Second, it has to resemble the true covariance as closely as possible, at least in limiting cases. Third, the structure of the approximation allows to draw samples from the approximate Gaussian, without the necessity of ever constructing the explicit covariance. All these properties are fulfilled by the covariance proposed in this section based on the inverse Fisher information metric I^{-1}. Inside the mode it is considered to be inferior to the Laplace approximation [66], but it is a valid covariance outside. Nevertheless, sometimes it is used to describe the uncertainty around the MAP location [53]. It measures the sensitivity of the posterior with respect to small parameter variations and it consists out of two parts $I = I_d + I_\theta$. First, the Fisher information metric of the likelihood:

$$I_d(\theta) \equiv \left\langle \frac{\partial \mathcal{H}(d|\theta)}{\partial \theta} \frac{\partial \mathcal{H}(d|\theta)}{\partial \theta^\dagger} \right\rangle_{\mathcal{P}(d|\theta)} . \tag{4.34}$$

In a frequentist setting, the inverse of this metric gives the Cramér-Rao bound [31, 109], a lower bound to the uncertainty of an estimator $\widehat{\theta}$:

$$I_d(\theta)^{-1} \leq \left\langle \left(\theta - \widehat{\theta}\right) \left(\theta - \widehat{\theta}\right)^\dagger \right\rangle_{\mathcal{P}(d|\theta)} . \tag{4.35}$$

The $\leq$ indicates that the right minus the left side of the equation exhibits a positive semi-definite matrix.

The second part is the information metric of the prior distribution, given by:

$$I_\theta = \left\langle \frac{\partial \mathcal{H}(\theta)}{\partial \theta} \frac{\partial \mathcal{H}(\theta)}{\partial \theta^\dagger} \right\rangle_{\mathcal{P}(\theta)} . \tag{4.36}$$

This quantity is a lower bound to the prior variance of an estimator (see Schützenberger [122] for vanishing likelihood), i.e.:

$$I_\theta^{-1} \leq \left\langle \left(\theta - \widehat{\theta}\right) \left(\theta - \widehat{\theta}\right)^\dagger \right\rangle_{\mathcal{P}(\theta)} . \tag{4.37}$$

We now have two bounds on the variance of the estimator, originating from information provided by prior and likelihood. To get to the posterior, we have to add up those two information sources. In the spirit of Gaussian error propagation, we constrain the posterior uncertainty by adding up the corresponding Fisher metrics. So the posterior covariance, compared to the prior one, will be at least reduced by the inverse Fisher metric of the likelihood. We cannot evaluate the resulting term at the location of the ground truth θ, as it is not available. We instead assume the estimator $\widehat{\theta}$ to be sufficiently close to provide a good approximation, which assumes sufficient local Gaussianity in the posterior, an assumption we will rely on later anyway. In this case, the inverse sum of the two metrics, evaluated at the estimator, should tend to be a lower bound to the true posterior variance. This is not a precise inequality and how it behaves in certain conditions will have to be explored in the future or case by case. We expect it to hold for not too extreme models, and otherwise at least to be sufficiently close. Thus, we state

$$I(\widehat{\theta}) \equiv I_d(\widehat{\theta}) + I_\theta \qquad \text{and} \tag{4.38}$$

$$I(\widehat{\theta})^{-1} \lessapprox \left\langle \left(\theta - \widehat{\theta}\right) \left(\theta - \widehat{\theta}\right)^\dagger \right\rangle_{\mathcal{P}(\theta|d)} . \tag{4.39}$$

By construction, the inverse Fisher information metric has only positive eigenvalues and we do not necessarily have to be in a mode, making it valid to use as covariance at every location, compared to the inverse Hessian with its potentially negative eigenvalues.

From now on, we identify the estimator with the estimate $\widehat{\theta}$, and interpret the variance of an estimator as uncertainty around the estimate. The inverse Fisher metric is then a lower bound to this uncertainty. Those two quantities constitute a Gaussian distribution $\mathcal{G}(\theta|\bar{\theta}, I(\widehat{\theta})^{-1})$ with mean $\bar{\theta} \leftarrow \widehat{\theta}$.

Here it is important to distinguish between the estimate $\widehat{\theta}$ and the mean of the Gaussian $\bar{\theta}$, which only initially coincide. In an iterative scheme we will use the location of the estimate $\widehat{\theta}$ to estimate the local uncertainty. While keeping this quantity fixed, we optimize for the mean parameter $\bar{\theta}$ via variational inference, such that the resulting Gaussian better matches the true posterior distribution. At this location we update the parameter estimate $\widehat{\theta} \leftarrow \bar{\theta}$. This way, we resolve the explicit dependence of the uncertainty estimate on the mean parameter and alleviate the necessity of calculating the Shannon entropy terms in the KL-divergence, containing the determinant of the possibly large covariance.

The surrounding landscape of this new estimate will have changed, compared to the previous location, and so will the inverse of the local metric. We set this as new covariance and repeat the procedure. Once the location and uncertainty are self-consistent with the posterior, we have converged to our final approximation. Instead of minimizing the KL-divergence within the family of a parametric distribution, we iteratively solve the locally Gaussian approximation problem, to narrow in towards the posterior mode. This bares a similarity to second order optimization, where always the locally quadratic problem is solved to iteratively find optima.

4.6.2 Standardized Metric

The information metric as an abstract mathematical object is invariant under coordinate transformation. In the previously discussed standard coordinates, the metric has an especially simple structure. A priori we deal with independent, standard Gaussian parameters $\xi \sim \mathcal{G}(\xi|0,\mathbb{1})$, without any hierarchical structure. Here the prior information metric is simply the covariance of the Gaussian, the identity operator:

$$I_\xi = \mathbb{1} \ . \tag{4.40}$$

The standard parameters are related to the original parametrization of the system via the possibly complex nonlinear transformation $\theta = f(\xi)$. The likelihood metric in the standard coordinates therefore is simply the push-forward from the likelihood metric in the original parametrization.

$$\begin{aligned} I_d(\xi) &= \left\langle \frac{\partial \mathcal{H}(d|\xi)}{\partial \xi} \frac{\partial \mathcal{H}(d|\xi)}{\partial \xi^\dagger} \right\rangle_{\mathcal{P}(d|\xi)} && (4.41)\\ &= \left(\frac{\partial f(\xi)}{\partial \xi}\right)^\dagger \left\langle \frac{\partial \mathcal{H}(d|\theta)}{\partial \theta} \frac{\partial \mathcal{H}(d|\theta)}{\partial \theta^\dagger} \right\rangle_{\mathcal{P}(d|\theta)} \frac{\partial f(\xi)}{\partial \xi} && (4.42)\\ &= J(\xi)^\dagger I_d(f(\xi)) J(\xi) \ . && (4.43) \end{aligned}$$

Here $J(\xi) = \frac{\partial f(\xi)}{\partial \xi}$ is the Jacobian of the transformation with respect to the new coordinates.

The Cramér-Rao bound in the standardized coordinates acquires additional curvature terms X [14]:

$$\left\langle \left(\xi - \widehat{\xi}\right)\left(\xi - \widehat{\xi}\right)^\dagger \right\rangle_{\mathcal{P}(d|\xi)} \geq I_d(\xi)^{-1} + X \ . \tag{4.44}$$

We will neglect those additional X terms, restricting ourselves to only parameter models with a sufficiently linear standardization transformation, at least locally. Therefore, we do not expect the method to perform well in the case of models with extreme X terms, which should be hard in general. Extensions of MGVI that treat this term better are left for future research.

For the uncertainty approximation we evaluate this expression at the current parameter estimate. The overall metric in standardized coordinates will therefore always have the following structure:

$$I(\widehat{\xi}) \equiv I_d(\widehat{\xi}) + I_\xi \tag{4.45}$$
$$= J(\widehat{\xi})^\dagger I_d(f(\widehat{\xi}))J(\widehat{\xi}) + \mathbb{1} \ . \tag{4.46}$$
$$\tag{4.47}$$

It only consists of three parts. First, the prior metric, which is the identity operator in the space of standard parameters. Second, the Fisher information metric of the likelihood, which is available for a large number of commonly used likelihoods. And third, the Jacobian of the standardization transformation. This transformation has to be implemented anyway, as it is equivalent to the model implementation. Its Jacobian can then be obtained by auto-differentiation, or consistently applying the chain rule. As long as the likelihood metric in the original coordinates does not require it, none of these quantities have to be stored in the form of a dense matrix. For the common case of independent data points, the likelihood factorizes and thus allowing for an implicit metric. We will elaborate on the concept of implicit operators in the dedicated Section 4.7. Nevertheless, using the inverse of the metric as approximate covariance is a non-diagonal approximation that captures correlations between all involved parameters.

4.6.3 Validity of the Covariance Approximation

The validity of the covariance approximation will depend on the properties of the system at hand. Here we will discuss three limiting cases in which the inverse Fisher metric is an accurate representation of the true uncertainty. The first scenario is asymptotic normality of the posterior distribution under the Bayesian Central Limit Theorem [48]. For a large amount of independently drawn data, the prior information will become irrelevant, according to the Bernstein-von Mises Theorem [134]. The posterior will approach the Gaussian distribution:

$$\mathcal{P}(\xi|d) \approx \mathcal{G}\left(\xi|\widehat{\xi}, I_d(\widehat{\xi})^{-1}\right) \ . \tag{4.48}$$

Here $\widehat{\xi}$ is a Bayesian estimator of ξ. The resulting covariance is equivalent to the term given in Eq. 4.43. Our covariance approximation contains additional to this term also the prior metric. In this highly informed setup, the likelihood will be by far the most dominant term, and the additive $\mathbb{1}$ prior metric becomes irrelevant, obeying the Bernstein-von Mises Theorem. So, in the Bayesian central limit, our approximate covariance coincides with the true posterior uncertainty. In this scenario, however, a MAP estimate, which essentially is the Maximum Likelihood estimate, will also provide reasonable results. Nevertheless, this behavior in the regime of large amounts of data is reassuring.

The opposite case is a vanishing likelihood. If data is scarce, we do not gain much information compared to the prior distribution. In truly large inference problems we easily encounter situations where we have to constrain millions of parameters with only thousands of data points. If we want to approach such problems, it is vital to be accurate in this limit, and the key to this is the standardized parametrization. In the trivial case of no likelihood at all, the posterior is equivalent to the standardized

Gaussian prior and the likelihood contribution to the metric vanishes. Here our approximate covariance will again be equivalent to the true uncertainty.

The inverse Fisher information metric is therefore a good approximation for large, as well as small amounts of data. The remaining question is, how well our approximation interpolates between those two limiting scenarios. So, the prior uncertainty is exact and we combine it with a lower bound of the uncertainty originating from the likelihood by adding the Fisher information metrics. In the limiting case of a Gaussian likelihood and linear standardization we actually obtain the true posterior covariance and our approximation will be also exact. In general cases the fidelity of the uncertainty estimate will depend on how well the inverse likelihood metric describes the uncertainty originating from the data. This is essentially a statement on how tight the Cramér-Rao bound is to the true uncertainty. In the worst case scenario, the inverse likelihood metric vanishes, and our approximation approaches a delta distribution, ultimately resulting in a MAP estimate. If it does not vanish, our approximation will be a better representation of the posterior.

The Cramér-Rao bound can be attained if, and only if, the likelihood is a member of the exponential family [136], which includes a large number of commonly used likelihoods. In such cases, we expect the inverse Fisher metric to well represent the uncertainty and our approximation to be valid. Also in cases where the likelihood is close to a member of the exponential family, the covariance should be reasonable.

In the context of high-dimensional and complex problems, several of these scenarios might be realized within the same model. Certain data might constrain a number of parameters extremely well, whereas other are only weakly informed. The former parameters might be in the regime of the central limit, the later could still be prior-dominated, and others will fall in between. As long as the model is not too extreme, our proposed covariance can capture the uncertainty and correlations in all these regimes simultaneously.

4.7 Implicit Operators

The information metric as a matrix has a dimension of the number of parameters squared. Storing it explicitly on a computer is already unfeasible for relatively small problems. In imaging for example, millions of pixel parameters are not uncommon and we will demonstrate MGVI for such an example at the end. The metric is built out of a collection of linear transformations, projections, and diagonal operators that all can be expressed efficiently by sparse matrices represented by computer routines. The metric itself is therefore expressible as an implicit operator, described by the composition of these simple operators. By construction, the metric is linear and positive definite, and therefore invertible. The inverse of the metric correlates all parameters with each other, usually resulting in a dense matrix expression, which will serve as approximate posterior covariance. This object is of interest during the inference, as well as for posterior analysis. As mentioned before, we cannot afford to store the posterior covariance at any moment explicitly. We have to extract all relevant information on correlations from the implicit metric only. This requires to apply the metric, as well as its inverse to vectors.

The implicit representation allows us to apply the metric $I = \Theta^{-1}$ to some vector

x efficiently.

$$b = \Theta^{-1} x \ . \tag{4.49}$$

More problematic is the application of the covariance Θ, the inverse metric, to some vector b.

$$x = \Theta b \ . \tag{4.50}$$

This matrix inversion can be done by solving Eq. 4.49 numerically for x, equivalent to solving a set of linear equations. The metric is certainly positive definite and hermitian, allowing the use of the Conjugate Gradient algorithm [126] for this inversion. This algorithm makes extensive use of the positive definiteness of the problem, leading to rapid convergence, compared to more general solvers. The resulting vector x then approximately satisfies Eq. 4.50.

4.7.1 Drawing Samples from the Approximation

The numerical inversion is the key to drawing samples from a Gaussian distribution with only an accessible precision matrix. We need those samples to estimate the KL-divergence and in the end they can be used to propagate uncertainty to any quantity of interest, based on the posterior. Those samples can be drawn by following the scheme outlined in Papandreou and Yuille [105], as our approximate covariance conveniently follows the structure of a conditional Gaussian distribution. This procedure scales linearly in time and memory with the dimensionality of the posterior. The main idea is to draw a sample from a Gaussian distribution with the inverse covariance, and then obtain a sample from the actual Gaussian by applying the covariance via numerical inversion.

In general, we can draw samples from a zero-centered Gaussian by drawing independent, white noise η_* in the eigenbasis of the covariance $\Theta = Q\Lambda Q^\dagger$ with eigenvectors Q and eigenvalues on the diagonal of Λ, weighting it with the square-root of the eigenvalues, and transforming it into the original space:

$$\Delta\theta_* = Q\sqrt{\Lambda}\eta_* \ , \text{ therefore} \tag{4.51}$$

$$\Delta\theta_* \sim \mathcal{G}(\theta|0,\Theta) \ . \tag{4.52}$$

Unfortunately, we do not have direct access to Θ, as it is only implicitly given trough its inverse Θ^{-1}, but we can draw samples according to $\Delta\phi_* \sim \mathcal{G}(\phi|0,\Theta^{-1})$, via $\Delta\phi_* = Q\sqrt{\Lambda^{-1}}\eta$. Numerically, we can approximately apply Θ to this sample from the Gaussian with inverse covariance, which yields

$$\Delta\theta_* \equiv \Theta\Delta\phi_* \tag{4.53}$$

$$= Q\Lambda Q^\dagger Q\sqrt{\Lambda^{-1}}\eta_* \tag{4.54}$$

$$= Q\sqrt{\Lambda}\eta_* \ . \tag{4.55}$$

Therefore, $\Delta\theta_*$ is then a sample from $\mathcal{G}(\theta|0,\Theta)$. Note that Q is unitary, therefore its adjoint is the inverse, $Q^\dagger Q = \mathbb{1}$. A set of such samples $\{\Delta\theta_*\}_N$ serves now as a

representation of the intractable, dense covariance.

$$\Theta = \langle \theta\theta^\dagger \rangle_{\mathcal{G}(\theta|0,\Theta)} \approx \frac{1}{N} \sum_{i=1}^{N} \Delta\theta_*^i \Delta\theta_*^{i\dagger} \ . \tag{4.56}$$

In the standardized parametrization, the approximate covariance always has the identical structure:

$$\Xi(\widehat{\xi}) = \left(J(\widehat{\xi})^\dagger I_d(f(\widehat{\xi})) J(\widehat{\xi}) + \mathbb{1} \right)^{-1} \ . \tag{4.57}$$

To draw samples according to the covariance Ξ, we start by drawing from the constituents of Ξ^{-1}:

$$n_* \sim \mathcal{G}\left(n|0, I_d\left(f(\widehat{\xi})\right)\right) \tag{4.58}$$

$$\eta_* \sim \mathcal{G}(\eta|0, \mathbb{1}) \tag{4.59}$$

$$\Delta\phi_* = J(\widehat{\xi})^\dagger n_* + \eta_* \ . \tag{4.60}$$

This requires the likelihood Fisher metric to be accessible in the eigenbasis, which is the case for example with independently sampled data points. Now $\Delta\phi_* \sim \mathcal{G}(\phi|0, \Xi(\widehat{\xi})^{-1})$ is distributed according to the inverse covariance. Using Eq. 4.53, we numerically apply the covariance itself to this sample via conjugate gradient, following Eq. 4.49.

$$\Delta\xi_* = \Xi(\widehat{\xi})\Delta\phi_* \quad \text{, and therefore} \tag{4.61}$$

$$\Delta\xi_* \sim \mathcal{G}(\xi|0, \Xi(\widehat{\xi})) \ . \tag{4.62}$$

These samples are drawn from a zero-mean Gaussian with the correct covariance. Our overall approximation will not be zero-centered, but this corresponds only to a shift by the mean vector $\bar{\xi}$.

$$\bar{\xi} + \Delta\xi_* \sim \mathcal{G}(\xi|\bar{\xi}, \Xi(\widehat{\xi})) \ . \tag{4.63}$$

This is essentially the reparametrization trick [69], which allows us to stochastically approximate the KL-divergence, while still providing gradients to the variational parameters, in our case only $\bar{\xi}$.

Using this procedure, we can draw a set of independent samples from the approximate posterior distribution, which allows us to statistically estimate the Kullback-Leibler divergence. Drawing these samples can be relatively costly, as every sample requires the numerical inversion of the inverse covariance, but drawing several samples is completely independent from each other and it can be done in parallel. Overall we might want to use as little samples as possible to reduce the numerical effort.

Another important point is how accurately the numerical inversion is performed. Of course, a higher accuracy results in better samples, but also requires more computations. The effect of un-converged samples depends mainly of the starting position of the conjugate gradient. Roughly speaking, the conjugate gradient method updates first the most informative directions. These correspond to the smallest eigenvalues of the covariance.

Starting at a sample from the standard Gaussian prior, after n iterations of the conjugate gradient at least the n most informative directions are updated towards the

posterior uncertainty, whereas the remaining directions still have the prior variance. Overall, un-converged samples will have the correct variance for the best informed directions and the remaining directions over-estimate the actual variance, encoded in the approximate covariance. This behavior safeguards us from a number of pitfalls that can be observed in MAP estimators by underestimating, or ignoring uncertainty variance. We will explore the impact of this accuracy on the method in one of the numerical examples in the end.

4.7.2 Antithetic Sampling

We will perform a stochastic estimate the KL-divergence and its gradient. This estimate is subject to sampling noise, which is reduced by increasing the number of samples. This increase will significantly impact the performance of the method. An additional way to reduce the variance of the estimates is antithetic sampling [84]. Here anti-correlated samples are used to obtain better estimates. Because we use a Gaussian approximation to the posterior, generating an additional, totally anti-correlated sample is trivial, as $\bar{\xi} - \Delta\xi^i_*$ is an equally valid sample as $\bar{\xi} + \Delta\xi^i_*$. Consider some monotonic function $g(\xi)$. The antithetic estimator $\widehat{g}^{(a)}$ of the function value is the average over the anti-correlated samples.

$$\widehat{g}^{(a)} = \frac{1}{N}\sum_{i=1}^{N/2}\left(g^i_- + g^i_+\right) \quad \text{, and} \tag{4.64}$$

$$\mathrm{Var}\left(\widehat{g}^{(a)}\right) = \frac{\mathrm{Var}\left(g(\xi)\right)}{N}\left(1 + \varrho_{g_+,g_-}\right) \ . \tag{4.65}$$

Here we indicate $g(\bar{\xi} \pm \Delta\xi^i_*) = g^i_\pm$ and ϱ_{g_+,g_-} is the correlation between the antithetic pairs. The smaller this correlation is, the better the estimate will be. For the parameter mean, i.e. $g(\xi) = \xi$, this variance will completely vanish. For non-linear functions and transformations, the anti-correlation in the samples could be reduced. In the worst case, the pairs are fully correlated, and we fall back to the $N/2$ independent samples in terms of the resulting variance, only wasting computations. However, only artificially constructed systems seem to be capable of showing such behavior.

Empirically we found that adding antithetic samples is extremely helpful in stabilizing the algorithm by counterbalancing extreme fluctuations in certain parameters. We will show in the numerical examples that even as little as one single pair of antithetic samples can be sufficient to obtain reasonable results, at least in the early stages of the procedure. This speeds up the overall convergence of the method, reducing the time to draw samples, as well as reducing the overall number of required samples due to lower variance of the estimates.

4.8 Metric Gaussian Variational Inference

At this point we want to summarize the key concepts of Metric Gaussian Variational Inference. MGVI performs a series of approximations of a complex posterior with Gaussian distributions. The covariance of the approximating Gaussian is extracted from the local properties of the true posterior, describing the vicinity around the

current mean estimate and it consists of the inverse Fisher information metric of likelihood and prior. Given this covariance, the approximate distribution is shifted to better represent the true posterior by minimizing the KL-divergence between truth and approximation with respect to the mean parameter. Given this covariance, the posterior is now optimally approximated by the Gaussian. However, at this new location, the vicinity around the mean might have changed, and we possibly represent the uncertainty better by again adapting the local properties of the true posterior. We iterate this procedure until the mean estimate is self-consistent with the uncertainty estimate.

MGVI cannot capture multi-modal structure in the posterior, as a Gaussian distribution is used to describe it. It also breaks down for severely non-linear models where second order terms of the transformation cannot be neglected. In the limits of small amounts of data, Gaussian posteriors, and the Bayesian central limit, MGVI will provide excellent results. In large-scale problems, certain parameters might be constrained extremely well, whereas others are almost uninformed by the data. Here MGVI can capture both limits simultaneously.

The standardization procedure of the hierarchical model might be optional for models with parameters of infinite support, but more complex models often contain parameters restricted to certain ranges. In this case the posterior cannot be approximated with a Gaussian. Standardization allows to approach these problems as well, as outlined for ADVI [85]. Here we will not treat the special case where MGVI is used in hierarchical coordinates, and we will only discuss the more unified, and structurally simpler case in standard coordinates. This results in non-Gaussian solutions for the original parametrization, as the approximation transforms according to the standardization transformation to the original parameters.

In this parametrization the information of the joint distribution of data and standardized parameters ξ with standardization transformation f always reads

$$\mathcal{H}(d, \xi) = \mathcal{H}(d|f(\xi)) + \frac{1}{2}\xi^\dagger \mathbb{1} \xi \ , \tag{4.66}$$

as outlined in Eq. 4.25. We want to variationally approximate the posterior corresponding to this model with a Gaussian distribution of the form (Eq. 4.8)

$$\widetilde{\mathcal{P}}(\xi|\bar{\xi}, \Xi) = \mathcal{G}(\xi|\bar{\xi}, \Xi) \ . \tag{4.67}$$

For an initial parameter estimate $\widehat{\xi}$, we construct the initial mean value $\bar{\xi} = \widehat{\xi}$ and the uncertainty estimate from the local Fisher information metric:

$$\Xi = \Xi(\widehat{\xi}) = \left(J(\widehat{\xi})^\dagger I_d(f(\widehat{\xi})) J(\widehat{\xi}) + \mathbb{1} \right)^{-1} \ . \tag{4.68}$$

Here $I_d(f(\widehat{\xi}))$ is the Fisher metric of the likelihood and $J(\widehat{\xi})$ the Jacobian of the standardizing transformation evaluated at the latent parameter estimate $\widehat{\xi}$ and $\mathbb{1}$ the prior metric, the identity operator in standard coordinates. This is a non-diagonal full-rank, positive definite matrix that correlates all parameters with another. We cannot store it explicitly at any time, but its inverse, the precision matrix can be well represented by a collection of sparse, implicit operations. In order to work with the covariance, we do have to rely on numerical operator inversion, as outlined in Sec. 4.7.

Given this covariance, we want to match the Gaussian distribution as closely as possible to the true posterior distribution by minimizing the KL-divergence with respect to $\bar{\xi}$, while keeping the covariance fixed:

$$\mathcal{D}_{\mathrm{KL}}\left(\mathcal{G}(\xi|\bar{\xi},\Xi(\widehat{\xi}))||\mathcal{P}(\xi|d)\right) \mathrel{\widehat{=}} \langle \mathcal{H}(d,\xi)\rangle_{\mathcal{G}(\xi|\bar{\xi},\Xi(\widehat{\xi}))} \tag{4.69}$$

$$\approx \frac{1}{N}\sum_{i=1}^{N} \mathcal{H}(d,\bar{\xi}+\Delta\xi_*^i), \text{ with} \tag{4.70}$$

$$\xi_* \sim \mathcal{G}(\xi|0,\Xi(\widehat{\xi})) \ . \tag{4.71}$$

When minimizing only with respect to the mean of a Gaussian, the Shannon entropy term is irrelevant for the KL-divergence, which therefore simplifies to the cross-entropy. We approximate the expectation value with a set of samples drawn from our approximation following the implicit sampling scheme described in Sec. 4.7.1. To minimize the stochastic estimate of the KL-divergence with respect to $\bar{\xi}$, we calculate the gradient, as well using these samples:

$$\frac{\partial \mathcal{D}_{\mathrm{KL}}}{\partial \bar{\xi}} = \left\langle \frac{\partial \mathcal{H}(d,\xi)}{\partial \xi} \right\rangle_{\mathcal{G}(\xi|\bar{\xi},\Xi(\widehat{\xi}))} \tag{4.72}$$

$$\approx \frac{1}{N}\sum_{i=1}^{N} \frac{\partial \mathcal{H}}{\partial \xi}\left(d,\bar{\xi}+\Delta\xi_*^i\right) \ . \tag{4.73}$$

To efficiently optimize the stochastic estimate of the KL-divergence, we rely on a (relaxed) natural gradient descent [6, 97]. We do have the Fisher information metric of the problem available anyway, so we use it to weight the gradient with the local inverse metric, followed by a line-search along this direction to account for non-quadratic features in the landscape. We repeat this procedure until the KL-divergence is minimized. The Fisher information metric of this stochastic estimate of the loss function is the average of the individual metrics evaluated at the sample location. As the samples collectively move through the landscapes, coupled by the mean, we re-evaluate the averaged metric at intermediate steps towards the minimum.

$$\left\langle \Xi^{-1}\right\rangle(\bar{\xi}) \equiv \frac{1}{N}\sum_{i=1}^{N} \Xi^{-1}(\bar{\xi}+\Delta\xi_*^i) \ . \tag{4.74}$$

The sum of implicit operators is still an implicit operator, and we can approximately apply the inverse of it to the gradient. The result is roughly the natural gradient, which we use as descent direction:

$$\Delta_{\bar{\xi}} = \left\langle \Xi^{-1}\right\rangle^{-1} \frac{\partial \mathcal{D}_{\mathrm{KL}}}{\partial \bar{\xi}} \ . \tag{4.75}$$

Now that we optimized the KL-divergence for the fixed covariance, we obtained a new parameter estimate in form of the mean of the variational Gaussian. We continue to repeat this procedure until the mean is self-consistent with the uncertainty estimate and it no longer changes. In Al. 1 we present a sketch of the MGVI algorithm.

Algorithm 1: Metric Gaussian Variational Inference

Input: Data d, Likelihood $\mathcal{P}(d|\theta)$, Fisher metric $I_d(\theta)$, Standardization $\theta = f(\xi)$

Initialize global iteration counter $i = 0$
Initialize $\widehat{\xi}^{(0)} = 0$ or small perturbation
while $\widehat{\xi}$ *not converged* **do**
  Construct covariance approximation $\Xi(\widehat{\xi}^{(i)})$ (Eq. 4.57)
  for N *samples* **do**
    Draw sample $n_* \sim \mathcal{G}(n|0, I_d(f(\widehat{\xi}^{(i)})))$ (Eq. 4.58)
    Draw sample $\eta_* \sim \mathcal{G}(\eta|0, \mathbb{1})$ (Eq. 4.59)
    Calculate $\Delta\phi_* = J(\widehat{\xi}^{(i)})^\dagger n_* + \eta_*$ (Eq. 4.60)
    Solve $\Delta\xi_* = \Xi(\widehat{\xi}^{(i)})\Delta\phi_*$ implicitly via numerical inversion (Eq. 4.61)
    Store $\Delta\xi_*$ (and $-\Delta\xi_*$) in the set of samples $\{\Delta\xi_*\}_N^{(i)}$
  end
  Set $\bar{\xi}^{(0)} \leftarrow \widehat{\xi}^{(i)}$
  Initialize local iteration counter $j = 0$
  while $\mathcal{D}_{\mathrm{KL}}$ *not minimized* **do**
    Estimate $\frac{\partial \mathcal{D}_{\mathrm{KL}}^{(i)}}{\partial \bar{\xi}}(\bar{\xi}^{(j)})$ with samples $\{\Delta\xi_*\}_N^{(i)}$ (Eq. 4.73)
    Construct Fisher information metric $\left\langle \Xi^{(i)-1} \right\rangle (\bar{\xi}^{(j)})$ (Eq. 4.74)
    Solve for natural gradient $\Delta_{\bar{\xi}}^{(j)} = \left\langle \Xi^{(i)-1} \right\rangle^{-1} \frac{\partial \mathcal{D}_{\mathrm{KL}}^{(i)}}{\partial \bar{\xi}}$ implicitly (Eq. 4.50)
    Find step-length η via line search of $\mathcal{D}_{\mathrm{KL}}^{(i)}\left(\bar{\xi}^{(j)} - \eta\Delta_{\bar{\xi}}^{(j)}\right)$ (Eq. 4.70)
    Update $\bar{\xi}^{(j+1)} \leftarrow \bar{\xi}^{(j)} - \eta\Delta_{\bar{\xi}}^{(j)}$
    Increment local iteration counter j
  end
  Update $\widehat{\xi}^{(i+1)} \leftarrow \bar{\xi}^{(j)}$
  Increment global iteration counter i
end
return $\widehat{\xi} \leftarrow \widehat{\xi}^{(i)}$
return $\{\Delta\xi_*\}_N \leftarrow \{\Delta\xi_*\}_N^{(i-1)}$

Initializing the parameter estimate with zero can be problematic due to vanishing gradients and numerical artifacts, which is resolved by using Gaussian noise with small variance instead. The convergence of $\widehat{\xi}$ can be determined by observing the changes between iterations. Because we use samples to determine all relevant quantities for the optimization, we are always subject to sampling errors. The more samples we use, the more accurate our solutions will be, so we can only convergence within the intrinsic sampling noise, given a number of samples. For more samples, we can achieve deeper convergence, and in practice we will increase the number samples throughout the algorithm.

Other meta-parameters of the algorithm are the accuracy of the numerical inversion to draw samples, i.e. the number of performed conjugate gradient steps, and how well we optimize the KL-divergence for a given parameter estimate. We will illustrate and discuss the impact of certain choices in the second numerical example. To use antithetical sampling for better stochastic estimates, one simply also includes $-\Delta\xi_*$ to the set of samples $\{\Delta\xi_*\}_N^{(i)}$.

In Al. 1 we use an approximate Relaxed Newton scheme to optimize the KL-divergence in the inner while-loop, but in principle any optimization scheme could be used. Especially the Newton-CG algorithm also performs well. In any case, we recommend to make use of the Fisher information for the optimization, as we have all ingredients available anyway and it can provide enormous speed-ups in high-dimensional problems.

In the end, MGVI provides a parameter estimate $\widehat{\xi}$ and a set of samples $\{\Delta\xi_*\}_N$, which together are samples from the approximate Gaussian distribution. This parameter estimate is self-consistent with the uncertainty estimate provided by the used approximation. The samples can then be used to propagate the uncertainty to any quantity of interest.

4.9 Numerical Examples

We will demonstrate MGVI in several examples, showcasing a diverse spectrum of applications, of both, synthetic- and real-data applications. We compare our approach to MAP estimates, HMC, mean-, and full-covariance ADVI.

In the first example we discuss the problem of inferring the rate of a Poisson distribution described as a log-Gaussian process. This process exhibits a squared exponential kernel of known amplitude and width.

The second example demonstrates the well behaved scaling of MGVI with the problem size, as well as its viability in the context of complex models with conceptually distinct parameters. Here we discuss the problem of binary Gaussian process classification in two dimensions with non-parametric kernel estimation. The data consists of binary values with associated location. The likelihood is the Bernoulli distribution and its rate is linked through a sigmoid function to a Gaussian process with unknown kernel. The size of the posterior exceeds one million model parameters. The computation and storage of a dense covariance as used by ADVI with a full covariance is computationally unfeasible as it would require to maintain 10^{12} entries. This problem size and complexity prohibits validation with the other methods and we compare the result of MGVI only to mean-field ADVI, as well as the underlying truth. Addition-

ally, we showcase and discuss the impact of several important meta-parameter choices on a smaller scale version of this problem.

The third example solves a non-negative matrix factorization problem on the Frey Face data-set assuming a Gamma-Poisson model.

In the last example we explore a hierarchical logistic regression problem involving polling data of the 1988 US presidential election with several regressors. We use a simplified model to again validate MGVI against HMC, as well as all the other methods, and a more complex model to discuss the convergence behavior of MGVI.

For an even larger numerical example with real data we refer to Leike and Enßlin [91], where a three dimensional dust map in our galactic vicinity is reconstructed in a resolution of 256^3 voxels from dust absorption measures and star locations obtained by the Gaia satellite. This problem involved a truncated Gaussian likelihood with log-normal prior and unknown kernel, analogous to the model used in the second example. The reconstruction was conducted using the here described MGVI procedure.

MGVI is further used by Arras et al. [12] to jointly calibrate a radio interferometer data set and perform its imaging. This allows to use the stationarity of the science target to obtain better calibration solutions, which in turn lead to better image reconstructions.

Another application of MGVI with multiple components and data fusion in spherical geometries is outlined in Hutschenreuter and Enßlin [58], where the Galactic Faraday depth sky is reconstructed from a rotation measure catalogue and free-free emission data.

Finally, Frank et al. [44] formulate locality and causality priors to learn the dynamics of a field from noisy and incomplete observation. Again, the inference of this field together with its dynamics is done via MGVI.

Performance metrics: Comparing different methods against each other is not straightforward. The preference of one method over another depends on many different factors, e.g. required accuracy, uncertainty quantification, or available resources. Any performance metric will only tell something about a certain aspect of the methods and we do not have a universal scale available to strictly determine the superiority of one method over another.

Ideally we want to validate against the true posterior distribution, but usually we do not have it available. MCMC methods allow to draw samples from the true posterior distribution, requiring large computational resources. Where it is feasible, we will use HMC as a reference, comparing the other methods against, but this restricts us to relatively small inference problems.

To explore the high-dimensional settings, for which MGVI was developed, we have to use a different approach. In real-world applications we do not know the true parameters underlying the data, but in a simulation we do. Performing the inference on such simulated data sets, we can always use the ground truth as reference scale and explore how well a method performs.

A simple metric in such a setting is the root mean squared error (RMS) of the reconstruction. For some scalar estimator $\widehat{x}$, for example the model evaluated at the

mean, it reads:

$$\mathrm{RMS} = \sqrt{\frac{1}{N}\sum_i \left(x^{(i)}_{\mathrm{true}} - \widehat{x}^{(i)}\right)^2} \,. \tag{4.76}$$

In many applications the goal is to get as closely as possible to some underlying truth, and therefore a lower RMS error should correspond to a better result.

Another quality criterion of a method is the capability to accurately estimate its uncertainty associated with the prediction. Large deviations to the ground truth are acceptable in cases a large variance is expected. For this we weight the absolute residual with the predicted standard deviation, corresponding to the average significance of the residual in terms of standard deviations:

$$\mathrm{AS} = \frac{1}{N}\sum_i \frac{1}{\widehat{\sigma}^{(i)}} |x^{(i)}_{\mathrm{true}} - \widehat{x}^{(i)}| \,. \tag{4.77}$$

In the case of a Gaussian posterior this quantity should be close to 1, expressing how significant on average the ground truth is, given in units of standard deviations. The posterior distributions we investigate will not be Gaussian, especially due to the non-linear transformations involved, but it should still provide an insight into the behavior of methods relative to each other, as long as the posterior resembles remotely a Gaussian and the non-linearity is not too extreme.

In large real-data applications sampling is unfeasible and the ground truth is unknown. In such cases we split the total data in a small sub-set d' for cross-validation of the approximation by evaluating how likely these reference data appear and use only the remaining data d for the inference. For this, we can calculate the predictive likelihood:

$$\mathcal{P}(d'|d) = \int d\theta\, \mathcal{P}(d'|\theta)\mathcal{Q}_\eta(\theta) \approx \frac{1}{N}\sum_{\theta_i} \mathcal{P}(d'|\theta_i) \,. \tag{4.78}$$

Here θ_i are samples drawn from the approximate distribution $\mathcal{Q}_\eta(\theta)$ fitted to the posterior given the remaining data d. It measures how predictive the obtained distribution is for the reference data. Generally a large value tells us that we can well extrapolate towards unobserved regions, which usually is desired. Nevertheless, this performance metric punishes uncertainty in a prediction. To see this, consider the maximum likelihood solution on the reference data. It is a point estimate and maximizes, by definition, the predictive likelihood. Now consider an uncertainty around this point, for example in form of a Laplace distribution. Every sample drawn from this presumably better approximation will have a lower predictive likelihood, and will therefore appear worse in this metric. We will encounter such a scenario in our examples and to make the comparison more fair, we will also state the predictive likelihood evaluated only at the latent mean parameter, corresponding to a best guess.

4.9.1 Poisson Log-Normal

Setup

In this example we discuss the inference of the rate λ of a Poisson likelihood providing count data d, where the logarithmic rate is modeled as a Gaussian process with

squared exponential kernel of known amplitude and width. The count data for our experiment is displayed in Fig. 4.1. The Poisson likelihood reads:

$$\mathcal{P}(d|\lambda) = \prod_i \mathcal{P}(d_i|\lambda_i) \quad \text{, with} \tag{4.79}$$

$$\mathcal{P}(d_i|\lambda_i) = \frac{\lambda_i^{d_i} e^{-\lambda_i}}{d_i!} \ . \tag{4.80}$$

Its Fisher information metric with respect to this rate parameter is:

$$I_d(\lambda) = \widetilde{\lambda}^{-1} \ . \tag{4.81}$$

This is a diagonal matrix, indicated by the tilde, in the data space with the inverse of the rate λ on its diagonal. A tilde over a vector raises it to diagonal matrix, i.e. $\widetilde{a}_{ij} = \delta_{ij} a_i$. The rate is expressed in terms of the exponential of a Gaussian process $\lambda = Re^s$ with prior distribution $\mathcal{P}(s) = \mathcal{G}(s|0, S)$ and some linear response operator R. Assuming a stationary, or homogeneous and isotropic kernel, the kernel can be expressed in terms of a spectral density in the harmonic domain, i.e. $S = \mathbb{F}^{-1}\widetilde{\mathbb{P}p}\,\mathbb{F}$ where $\mathbb{F}$ indicates the Fourier transformation, $\mathbb{P}^\dagger$ is the projection of the one-dimensional spectral density onto the Fourier space of the signal coordinates, also one-dimensional in this example, but in general multi-dimensional. Here $p(k) = \sqrt{2\pi}\sigma^2 l\, e^{-2\pi l^2 k^2}$ represents the squared exponential, or Gaussian, correlation kernel in Fourier space (in one dimension). The parameter l is a characteristic length-scale, σ^2 a variance parameter and k is the harmonic coordinate. This defines the mathematical setup of this first example.

The next step is to standardize. As the prior is already Gaussian, we simply have to identify $S = AA^\dagger$ with $A = \mathbb{F}^{-1}\widetilde{\mathbb{P}p^{\frac{1}{2}}}$ and rewrite $s = A\xi$. With this reparametrization we express the information of the problem for a given spectrum as

$$\mathcal{H}(d, \xi) \mathrel{\widehat{=}} -d^\dagger \ln R e^{A\xi} + 1^\dagger R e^{A\xi} + \frac{1}{2}\xi^\dagger \mathbb{1} \xi \ . \tag{4.82}$$

The $1^\dagger$ indicates a scalar product with the one vector, corresponding to the integration over the space. The overall standardization reads

$$\lambda = f(\xi) \tag{4.83}$$

$$= Re^{A\xi} \ . \tag{4.84}$$

This function allows us to build the local approximation to the covariance. Here the parameter dependence is still relatively simple and the Jacobian can be calculated by hand.

$$\Xi^{-1} = J(\widehat{\xi})^\dagger \widetilde{f(\widehat{\xi})^{-1}} J(\widehat{\xi}) + \mathbb{1} \tag{4.85}$$

$$= A^\dagger \widetilde{e^{A\widehat{\xi}}}^\dagger R^\dagger \widetilde{\frac{1}{Re^{A\widehat{\xi}}}} \widetilde{Re^{A\widehat{\xi}}} A + \mathbb{1} \ . \tag{4.86}$$

We see that the metric is composed out of a collection of operators that can be simply implemented and combined.

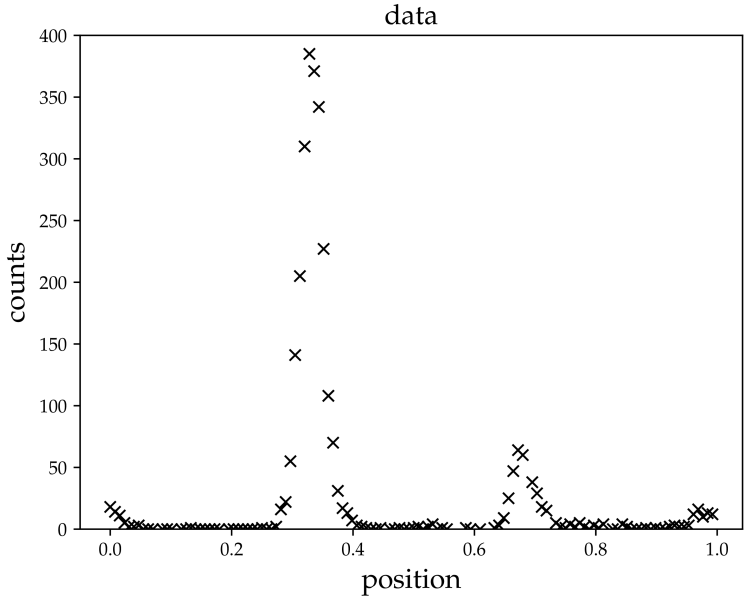


Figure 4.1: A Poisson realization drawn according to a log-normal process with squared exponential kernel on linear scale.

Now, we approximate the posterior probability implied by the model as described by Eq. 4.82 using MGVI. We start the optimization with one single pair of antithetic samples. Initially we perform three natural gradient steps and use 25 conjugate gradient iterations to draw the sample. After twenty global iterations we start to increase the number of samples and natural gradient steps by one until the thirtieth iteration and steadily increase the sampling accuracy by a total factor of four. Initially we do not want to waste computations for unnecessary accuracy and this purely heuristic scheme is derived from the meta-parameter discussion of the next example.

The problem, as well as MGVI, mean-field (mf-) and full-covariance (fc-) ADVI [85], HMC [34], and the Laplace approximation are implemented within Python using the NIFTy5[1] package [10, 124, 128].

The posterior samples obtained from HMC will serve as the reference in the validation of our approach. We run five HMC chains in the standard coordinates and obtain a minimal effective sample size of $\mathrm{ESS}_{\mathrm{min}} = 108$ and average $\mathrm{ESS}_{\mathrm{mean}} = 220$. The average Gelman-Rubin test statistic for the five chains is $\widehat{R}_{\mathrm{mean}} = 1.016$ with maximum $\widehat{R}_{\mathrm{max}} = 1.052$.

For ADVI we perform both, a fully parametrized covariance, as well as a mean-field approximation, estimating only a diagonal covariance. Using the full covariance limits the possible problem size and we will stick to 128 parameters to describe the Gaussian process, as well as 128 equidistant data points. For the optimization procedure we follow the stochastic gradient descent scheme proposed by Kucukelbir et al. [85].

The data is is drawn according to the model and the realization is shown in Fig. 4.1. We monitor the performance and convergence by withholding 10% of the data points and calculating the predictive likelihood of the intermediate result.

[1]NIFTy documentation: `http://ift.pages.mpcdf.de/NIFTy/`
NIFTy code: `https://gitlab.mpcdf.mpg.de/ift/NIFTy`

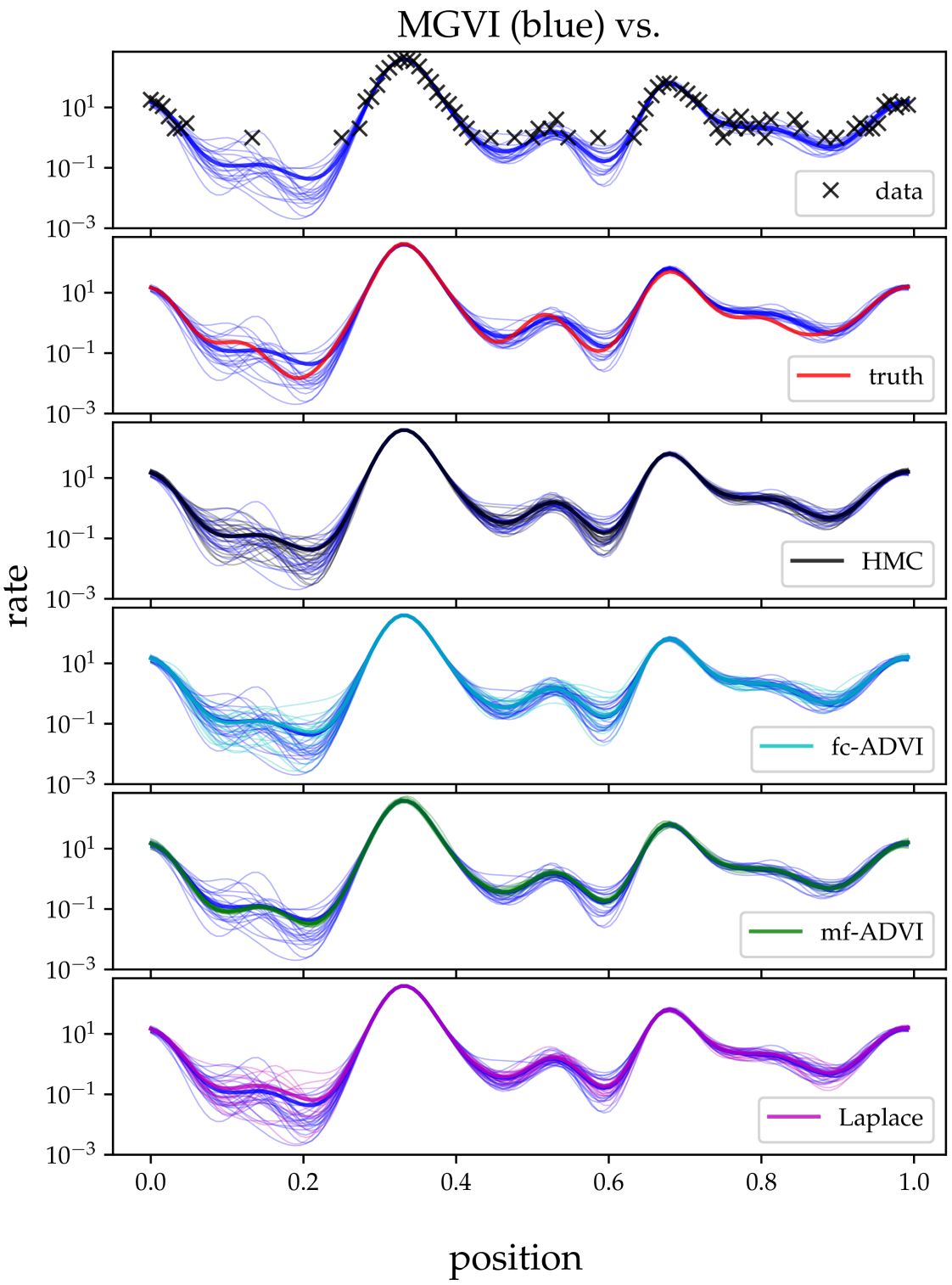


Figure 4.2: Reconstructed rates and posterior samples provided by MGVI in comparison to those from various other methods.

Results

All methods recover the underlying rate quite well. The obtained rates $\lambda = Re^{A\xi}$ are shown in Fig. 4.2 for MGVI and all the other methods. The uncertainty of the different estimates are indicated by a set of posterior samples drawn around their corresponding mean rates for all methods. Visually, all methods, except mean-field ADVI, provide similar results. The later severely underestimates the true posterior variance in areas of large uncertainty and overestimates it in regions well-determined by the data.

We also note that overall the relative uncertainty is higher in regions of low counts and smaller in regions of high counts. This is expected from a Poisson likelihood, as its variance σ_d^2 is equal to its rate λ and therefore the relative uncertainty increases with decreasing rate, $\sigma_d/\lambda = 1/\sqrt{\lambda}$.

The two-point correlation matrix of the rate constructed from samples is shown in Fig 4.3. Here again, all correlations do look similar, except mean-field ADVI. The correlation is diagonal dominant and spatially structured. Strong short-range correlations are wrapped by a band of anti-correlation, decaying towards zero for large distances of the points. The periodic boundary condition of this setup is showing up in the top right and bottom left corners. This pattern originates from the squared exponential kernel and is modified by the data. High-signal regions appear here to be more narrow, and the correlation is farther extending in low-rate regions. Here the mf-ADVI correlation structure is agnostic to spatial structure, compromising between high-data and uninformed regions due to the limited expressibility of a mean-field approximation.

We provide a snapshot of all methods against HMC in Fig. 4.3, scattering the values for two locations against each other in three distinct scenarios. This provides a visual impression on how well correlations, as well as the marginal probabilities, are captured by the approximations. MGVI, fc-ADVI and the Laplace approximation match closely to the HMC samples, but mf-ADVI strongly underestimates the variance and, as observed in the correlation matrix, does not express much of the posterior covariance structure.

To validate this impression we plot the mean and standard deviation of the log-rate at every location obtained by HMC against the four other methods, as shown in Fig. 4.5. Towards large mean rates, all methods agree quite well, as they are well determined by the data. For small rates, the methods differ slightly, but systematically. Here the Laplace approximation, as well as mf-ADVI tends to overestimate the rate, whereas fc-ADVI and MGVI underestimate it, but in that agree well. Thus, it seems that MGVI behaves similarly to fc-ADVI here.

The standard deviations are structurally more interesting. MGVI and the Laplace approximation agree well with the HMC results and fc-ADVI appears slightly shifted towards overestimated variances, which might be a remnant of insufficient convergence within the assigned computational budget. Finally, mf-ADVI is agnostic to parameter-specific uncertainty and only on average correct, over- and underestimating the standard deviation roughly the same amount of time, as already seen in the correlation matrix and the scatter plots.

Collapsing these plots down to the RMS error relative to HMC gives Tab. 4.1. Here MGVI is the best method in terms of reproducing the mean, as well as the standard de-

Table 4.1: The RMS errors of mean and standard deviation of the pixel-wise log-rate with respect to HMC

RMS HMC against	MGVI	fc-ADVI	mf-ADVI	Laplace
mean	0.041	0.076	0.16	0.17
standard deviation	0.023	0.080	0.42	0.032

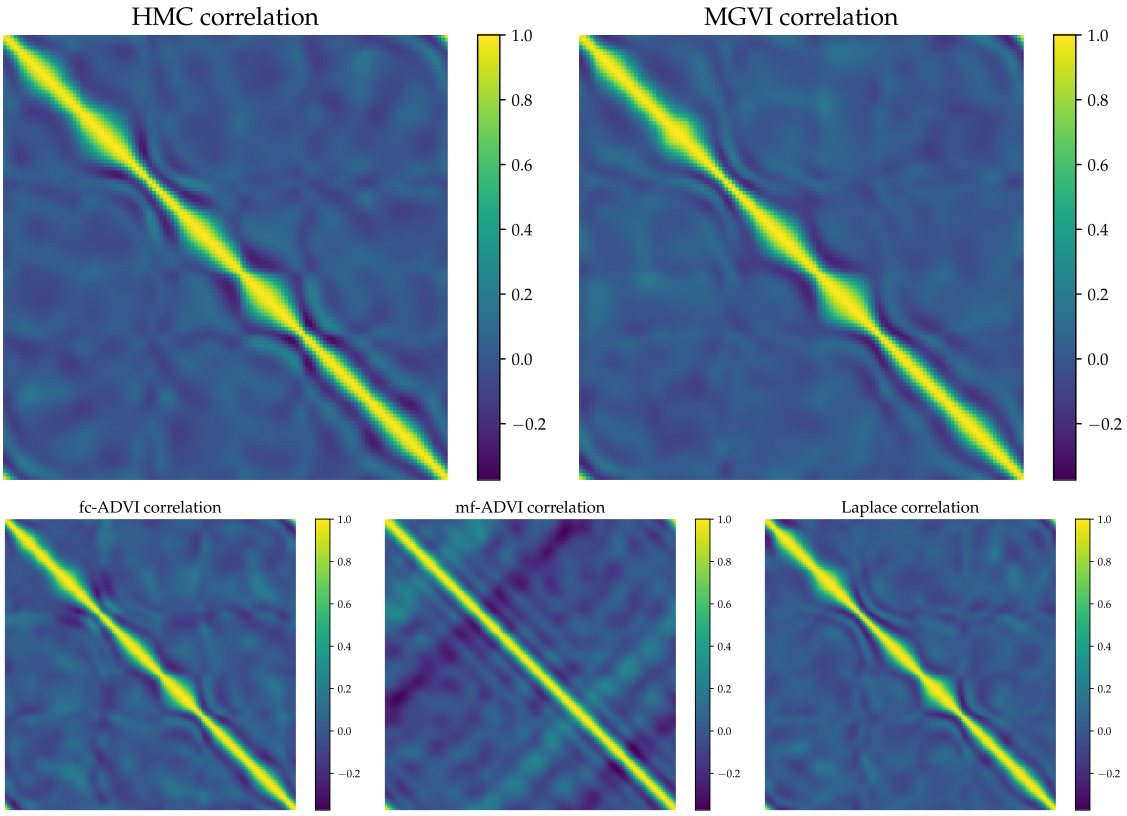


Figure 4.3: The sampled correlation structures for the different methods.

viation obtained via HMC. Regarding the mean, fc-ADVI gives a similar, but slightly worse result, but mf-ADVI and the Laplace approximation exhibit significantly larger errors. Surprisingly, the latter exhibits one of the best standard deviations, better than fc-ADVI, which still might improve for even longer optimization. The deviations for mf-ADVI are, as expected, just off, dramatically.

Overall, MGVI seems to be slightly better, but on par in terms of accuracy with the other methods in this example. Its true strength only becomes evident by considering the required computational time to obtain these results, as well as the in principle linear scaling behavior in terms of memory and computations. In the following we will discuss the temporal evolution of several quantities during the optimization for MGVI and both ADVI variants.

Convergence behavior

The first quantity we monitor during the optimization is the predictive likelihood on unobserved data. For this purpose we withheld 10% of the data points to track how

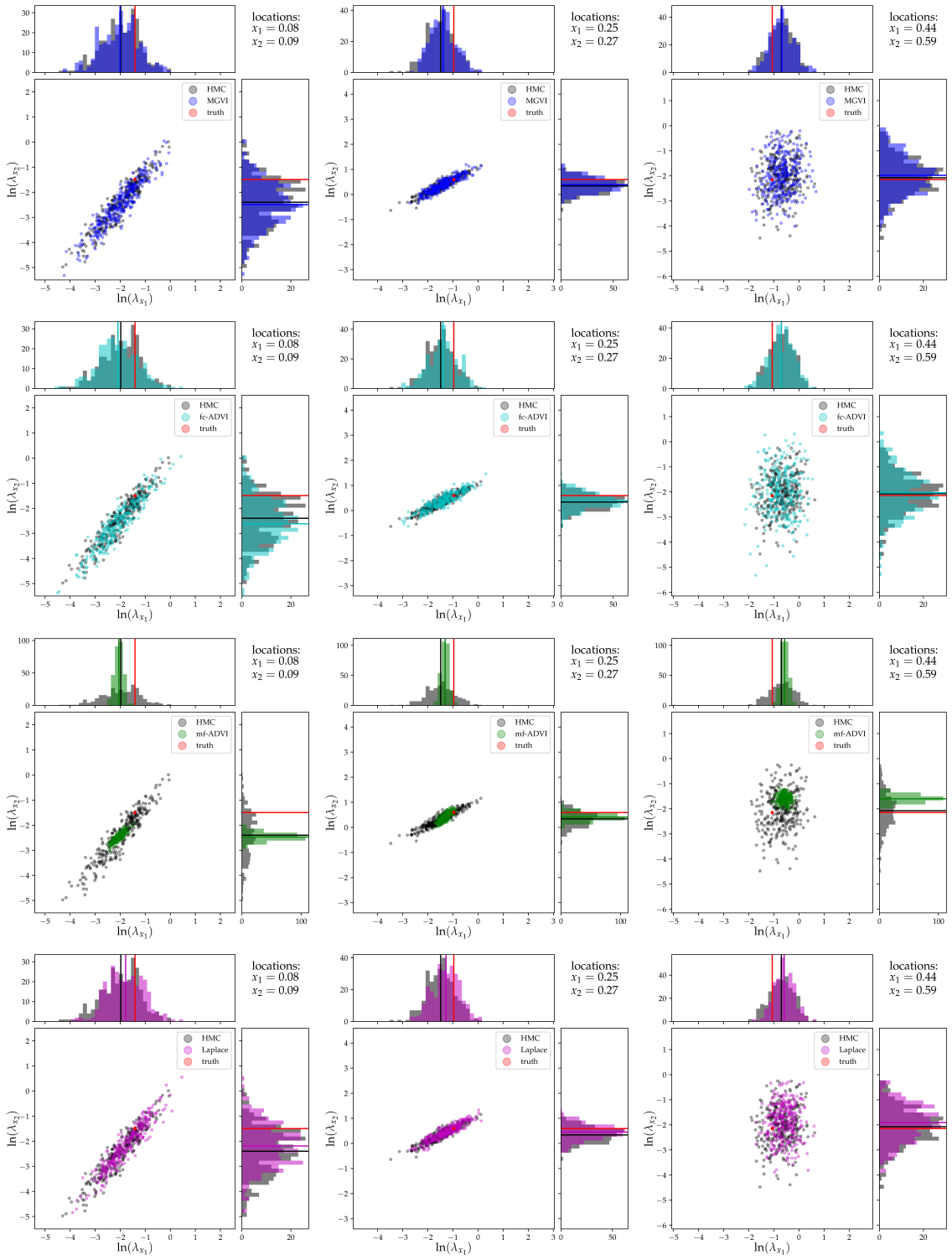


Figure 4.4: Scatter-plots of the logarithmic posterior rates at two close-by locations in a low-count region. The posterior samples from MGVI are compared to those of all other methods that provide posterior samples. The true rates are indicated as well.

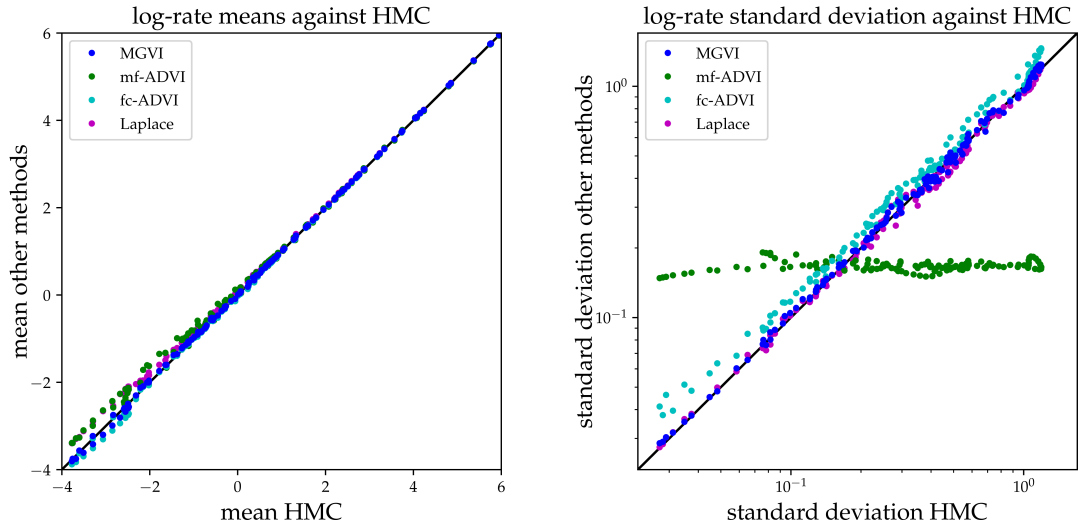


Figure 4.5: The parameter means and standard deviations of the Poisson log-normal problem from the different methods plotted against the HMC results.

well those are explained by the current state of some method. In this example we do have access to the underlying true rate, which allows us to monitor the RMS error to this ground truth, as well as how well the remaining residual is captured by the predicted uncertainty in terms of significance, which, for a Gaussian would be one sigma. For the definition of these quantities see Sec. 4.9. All results are shown in Fig. 4.6. For the predictive likelihood, by far the fastest method is a MAP estimate. Using second order natural gradient descent, this method converges within less than 0.06 seconds. MGVI is significantly slower, but also rapidly converges in terms of the predictive likelihood. After only 0.1 seconds and drawing new samples twice it no longer changes significantly for the remaining time. After roughly 2 second, the next method to converge is mf-ADVI. Ten times longer is required by HMC, which takes roughly 20 seconds to complete its burn-in. This is consistent with the one order of magnitude speedup reported in Kucukelbir et al. [85]. By far the slowest method is fc-ADVI, requiring roughly 1000 seconds (or 16.6 minutes) to achieve comparable predictivity. In this example MGVI is slower, but comparable to a MAP estimate, more than one order of magnitude faster than mf-ADVI, two orders of magnitude faster than HMC and four orders of magnitude than fc-ADVI, which on this problem scale is barely feasible.

The RMS error to the true rate can be used as another indicator how fast the methods converges. In the case of MGVI, not much happens after the first global iteration, requiring 0.1 seconds. The RMS error of both ADVI methods steadily drop down to the final level, mf-ADVI being initially faster, but fc-ADVI catches up before final convergence.

Interesting is the behavior of the average significance, characterizing how well the deviations from the true rate are explained by a Gaussian approximation using the samples provided by the methods. It also allow us to evaluate how well the covariance of each method has converged. For MGVI this seems to be the case after 10 seconds. This coincides with the increase of samples used to estimate the KL-divergence. With

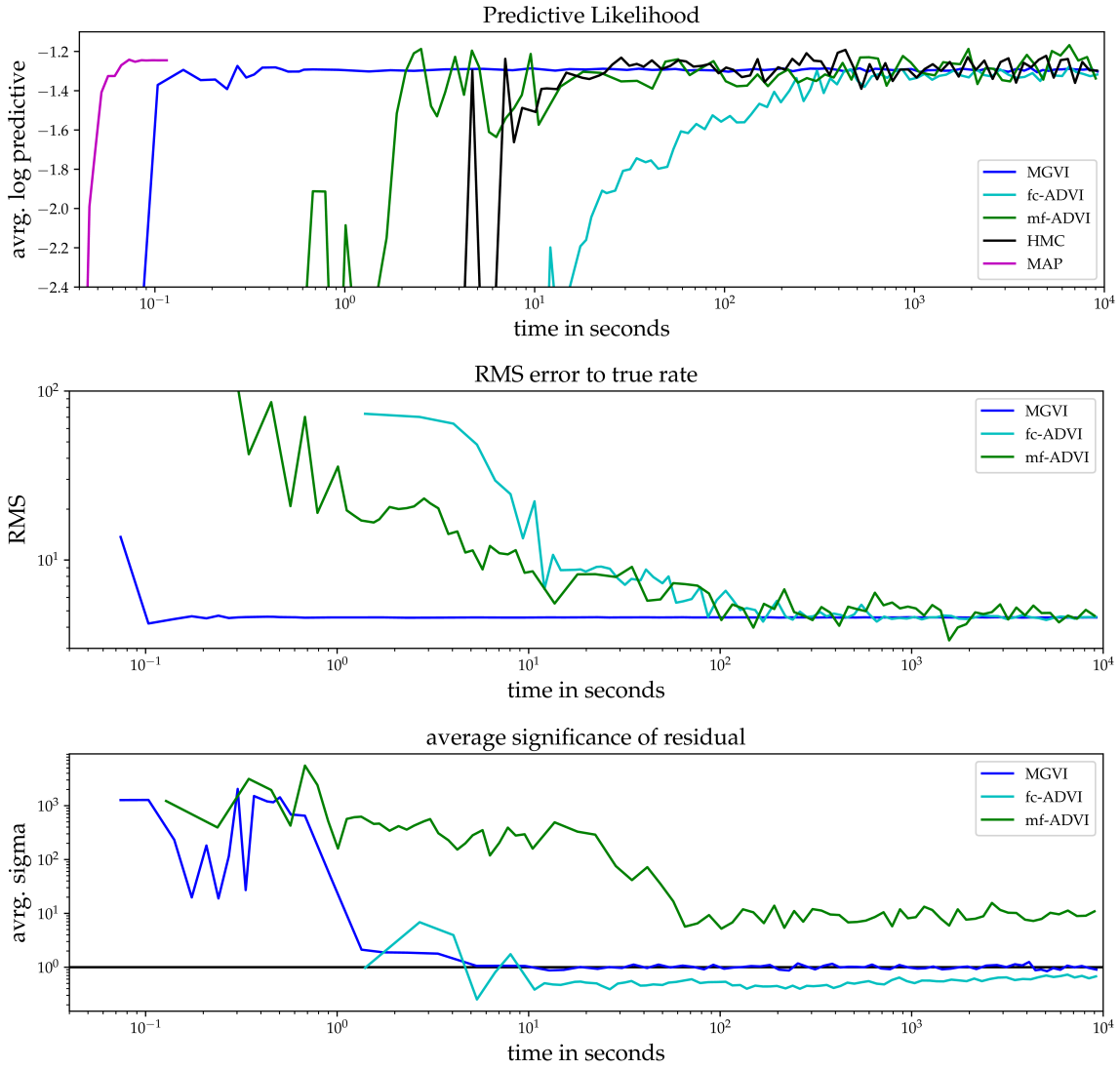


Figure 4.6: The performance metrics for the Poisson log-normal problem for all methods. The curves are smoothed by a moving average after the first ten points and equidistantly sampled on a logarithmic scale.

more samples the uncertainties are probed better and the average significance of the residuals are spot on the one sigma level, hinting at a quite Gaussian posterior. Although mf-ADVI converges quickly in terms of the predictive likelihood, here we observe drifts within the first 100 seconds. This is even more extreme in the case of fc-ADVI, which only drifts gradually and probably did not fully converge, still slightly overestimating the variance. This is consistent with the shift relative to the HMC standard deviations.

Overall, MGVI is fast because it has intrinsically fewer parameters and (quasi-) second order optimization can be used. Also the observation that means of Gaussian approximations converge fast and a covariances slowly might also contribute to the rapid convergence behavior of MGVI. Only the mean has to be optimized for a given covariance, and once it converged a new, plausible covariance is adapted, without having to laboriously optimize for it.

4.9.2 Binary Gaussian Process Classification with Non-Parametric Kernel

In the second example we apply MGVI to a much higher dimensional problem and more complex context, making it unfeasible for a fully parametrized covariance. Binary Gaussian process classification is used to attribute regions to certain classes and identify boundaries between them. A comprehensive overview can be found in Kuss and Rasmussen [88] and Nickisch and Rasmussen [101]. In addition to the typical formulation, we also infer the underlying kernel non-parametrically. With this extension a Laplace approximation will not provide reasonable results as parameters are degenerate. In this example we compare MGVI to mf-ADVI, which is still capable of coping with such extremely high dimensional problems. We consider binary data in two spatial dimensions, measured only at certain locations. The likelihood is a Bernoulli distribution and its rate parameter is described by a sigmoid function applied to an underlying Gaussian process. The kernel of this process is unknown and will be modeled non-parametrically as well. We assume a stationary, isotropic kernel and model it by two spectral components. The first component follows a power law that is modified by the second component, a log-Gaussian process with a smooth kernel. Overall the spectral density is parametrized by two power-law parameters, an amplitude and the spectral index, and the Gaussian process parameters for the component modifying this power-law. This model is inspired by systems with underlying processes favoring certain length-scales and is well-suited for imaging applications.

Setup

The likelihood in this example is the Bernoulli distribution that reads

$$\mathcal{P}(d|\mu) = \prod_i \mathcal{P}(d_i|\mu_i) \quad \text{, with} \tag{4.87}$$

$$\mathcal{P}(d_i|\mu_i) = \mu_i^d (1-\mu_i)^{1-d_i} \ . \tag{4.88}$$

for some rate parameter on the unit interval $\mu \in (0,1)$ and binary outcome $d \in \{0,1\}$. The Fisher information metric for this likelihood is

$$I_d(\mu) = \widetilde{\mu(1-\mu)}^{-1}. \tag{4.89}$$

The rate μ is linked to a Gaussian process $s \sim \mathcal{G}(s|0,S)$ by a sigmoid function and a linear response:

$$\mu = R\sigma(s) \tag{4.90}$$

$$= R\,\frac{1}{2}(1+\tanh(s))\,. \tag{4.91}$$

The kernel S of this process is assumed to be stationary and isotropic and can be expressed as $S = \mathbb{F}^{-1}\widetilde{\mathbb{P}p}\,\mathbb{F}$ with spectral density p. This quantity itself is to be learned and it is modeled according to

$$p(k) = e^{a\,\ln k + b + \tau_k}\,. \tag{4.92}$$

The first two terms in the exponent model a power-law kernel, which is linear on double-logarithmic scale, with power a and amplitude b, both of which get a Gaussian prior with assumed mean ($\bar{a}$ and $\bar{b}$) and variance (σ_a and σ_b). The last term in the exponent follows an integrated Wiener process on logarithmic spatial scale, ergo a differentiable function, according to the known kernel $T = AA^\dagger$. The integrated Wiener process follows a power-law kernel with power four. We treat it analogously to the other correlation kernel S. This prior is standardized by performing a Fourier transformation on logarithmic coordinates and multiplication with the square root of the power-law spectrum. The graphical structure of this described model is shown in Fig. 4.7.

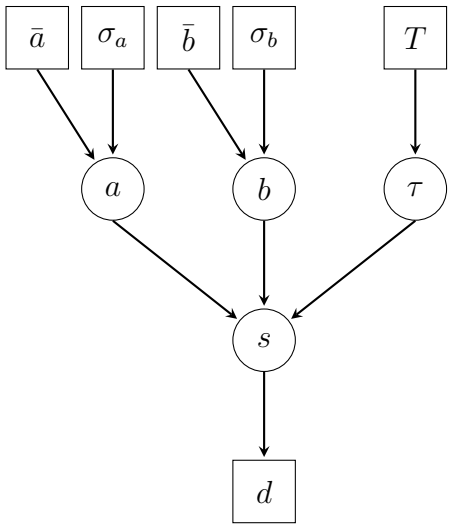


Figure 4.7: The graphical structure of the binary Gaussian process classification with non-parametric kernel.

Reparametrizing the model parameters yields the following relation to the original

rate μ:

$$\mu = f(\xi) \tag{4.93}$$

$$= R\sigma \left(\mathbb{F}^{-1} \left(\mathbb{P} e^{(\bar{a}+\sigma_a \xi_a)\, \widetilde{\ln k + \bar{b}} + \sigma_b \xi_b + A\, \xi_\tau} \right) \xi_s \right) \ . \tag{4.94}$$

This reparametrized model has therefore a highly non-linear likelihood in terms of its parameters, where $\xi = (\xi_s, \xi_a, \xi_b, \xi_\tau)^\dagger$. This expression has to be read from the right to the left, which is the direction of the generative model. A series of linear and point-wise non-linear operations are performed on the latent model parameters to generate μ. $\mathbb{F}$ and $\mathbb{P}$ are again the Fourier transformation and the isotropic projection of a 1D spectrum to a 2D Fourier space. Here again, the tilde indicate that the quantity below is raised to a diagonal operator. Obtaining this function is tedious but straightforward and can be done automatically, given the hierarchical structure of the model. We spare the reader the expressions of the Jacobian of the function with respect to its parameters $J(\xi) = \frac{\partial f(\xi)}{\partial \xi}$ as this should be implemented using auto-differentiation, which NIFTy5 [10] provides to us. Structurally, the problem is now identical to the previous one with information and approximate covariance:

$$\mathcal{H}(d, \xi) \mathrel{\widehat{=}} - d^\dagger \ln f(\xi) - (1-d)^\dagger \ln\left(1 - f(\xi)\right) + \frac{1}{2} \xi^\dagger \mathbb{1} \xi \tag{4.95}$$

$$\Xi(\widehat{\xi}) = \left(J(\widehat{\xi})^\dagger \Big(\widetilde{f(\widehat{\xi})(1 - f(\widehat{\xi}))} \Big)^{-1} J(\widehat{\xi}) + \mathbb{1} \right)^{-1} \ . \tag{4.96}$$

Regarding the numerical setup, we consider 2^{19} binary data points on a two dimensional plane organized in a checkerboard. We use 1024×1024 parameters to describe the Gaussian process underlying this rate. The spectral density is parametrized by additional two parameters for the power-law and 64 for the non-parametric part ξ_τ, resulting in overall more than a million parameters, which is completely out of reach for explicit covariance parametrization. For simplicity periodic boundaries were assumed. Due to the large parameter dimension, we initially choose to use 100 conjugate gradient iterations, and increase it towards 400 at the end. Otherwise we use the setup from the previous example.

Results

The synthetic data, the true underlying rate, the mf-ADVI, as well as MGVI results are shown in Fig. 4.8. The data is only sampled at certain locations and due to the binary output appears noisy. It exhibits spatial characteristics, predominately showing one class over the other in certain regions. The true rate, from which the data was drawn, shows rich features on all scales. The largest of them can also be seen in the data directly, but small-scale features are washed out due to the Bernoulli noise. The MAP solution to this problem (not shown) does not provide a plausible posterior estimate and completely over-fits the data.

The mean rate recovered by MGVI matches up to a certain scale exceptionally well to the true rate. Even in unobserved areas the structures are recovered correctly to

some extent (as can be seen e.g. in the top right and bottom left corners). Small scales cannot be recovered as the data does not contain much information on them. This is also reflected in the standard deviation at each location. The highest uncertainty is, as expected, in the not observed areas, reproducing the checkerboard pattern. The standard deviation is also modulated by the rate itself. The more a certain region is attributed to one class, the lower its uncertainty. The uncertainty is especially high at the boundaries between the classes.

Also mf-ADVI recovers the underlying rate quite well, with maybe slightly less sharp features, but certainly comparable to the MGVI result. The main difference lies in the uncertainty estimate, which completely lacks the spatial features attributed to the incomplete checkerboard sampling of the data. Nevertheless, it shows the error associated with the nonlinear error propagation from the Gaussian process to the rate. This is similar to the behavior observed in the previous example. Compared to the standard deviation from MGVI, the uncertainty seems to be larger in areas with observed data and significantly smaller in unobserved regions.

The recovered spectral density is shown in Fig. 4.9. At the largest scales, and therefore the smallest modes, the true correlation structure is correctly recovered within the error by MGVI, indicated by a set of samples. Even most large-scale spectral features are identified correctly by the algorithm. At a some point towards smaller scales the uncertainty increases significantly. This is also the point where the recovered spectrum diverges from the true one. This might indicate incomplete convergence, however, those highly uncertain parameters are the last ones to converge anyway and are affected the most by the stochastic estimation of the KL-divergence. Even on those scales the trues spectrum is not completely out of the error bound and seems consistent with the recovered spectrum. The mean spectrum obtained by mf-ADVI is similar to MGVI, but is slightly shifted down for the most part, but the error estimates are again not spatially modulated, not reflecting the deviations from the truth.

Regarding the result, MGVI seems to be better at describing the true posterior distribution with slightly more accurate means but superior uncertainties. Also important is how fast MGVI achieved this result, compared to mf-ADVI. For this we track again, as in the previous example, the predictive likelihood of all the unobserved data, filling in the checkerboard. Additionally we track the RMS error of the mean to the ground truth, as well as the average significance of the residual.

The convergence behavior of MGVI and mf-ADVI are also shown in Fig. 4.9 as well. Here MGVI shows the first accurate results after roughly 200 seconds in terms of RMS error and predictive likelihood. Shortly after this, the predictive likelihood of the mean and samples significantly diverge. This coincides with the increase from a single pair of antithetic samples to a gradually larger number, shown by the steep drop in the average significance. We suspect that the system found a self-consistent solution with that single sample and it got strongly disturbed by the presence of more samples. Finally the system recovers and converges to a similar predictivity and consistent error. In this example, mf-ADVI is significantly slower to achieve comparable predictivity and RMS error by roughly one order of magnitude, but those levels are achieved.

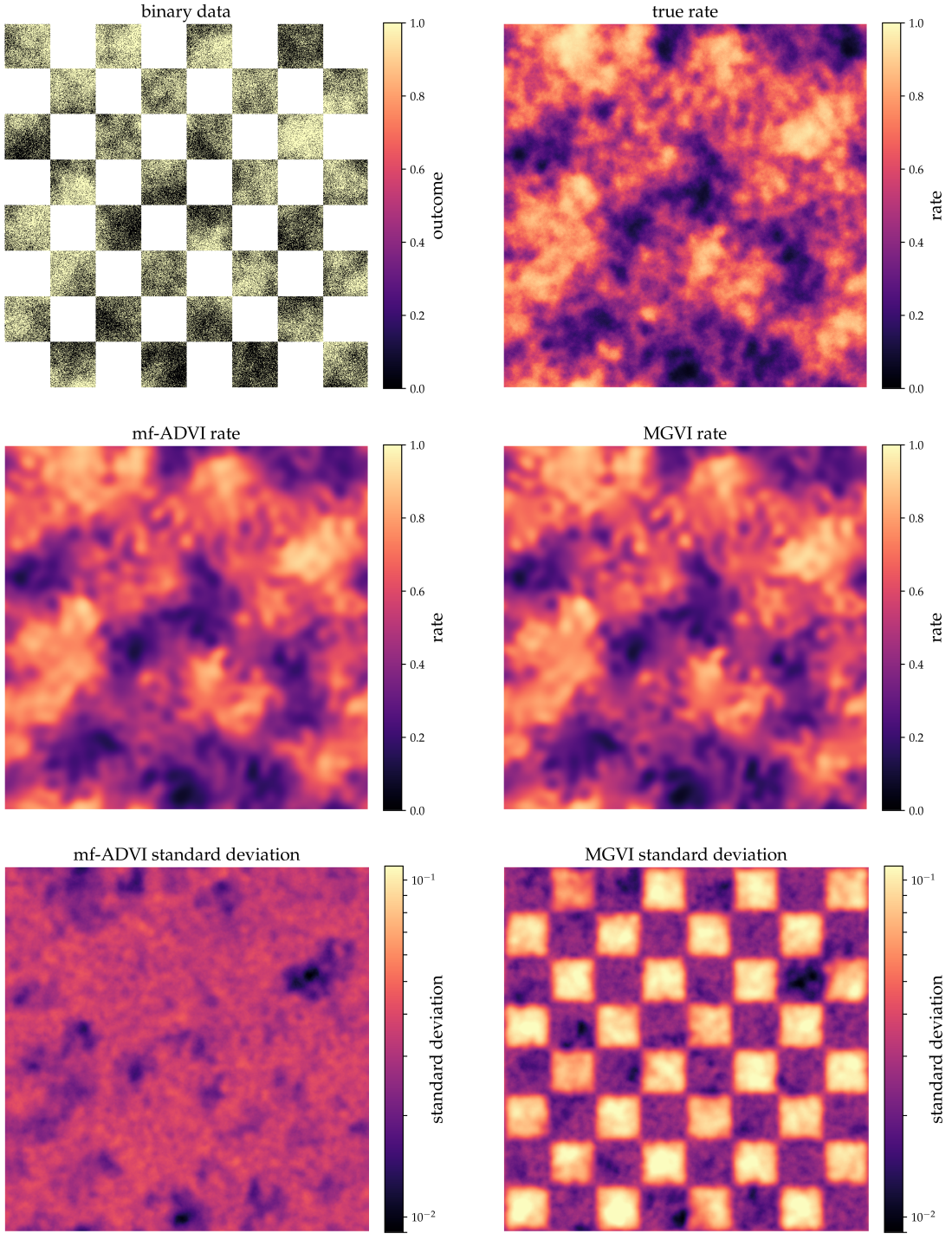


Figure 4.8: The data and true rate, as well as the mf-ADVI and MGVI means and standard deviations. Note that the small-scale noise in the data can lead to a color blend that does not seem to be part of the used color scheme.

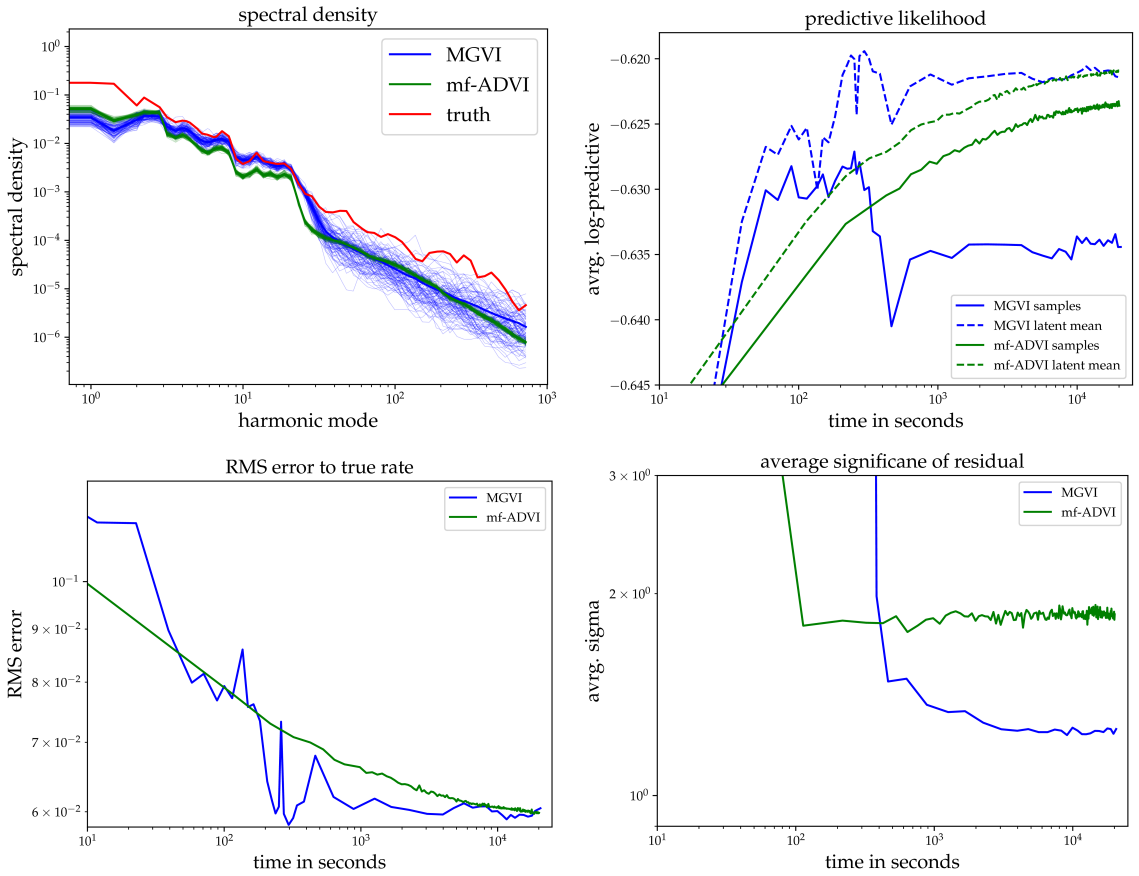


Figure 4.9: Recovered spectral density (top left), and the performance metrics for the binary Gaussian process classification problem, comparing MGVI to mf-ADVI.

Discussion of meta parameters

MGVI requires a number of meta parameters that will affect the performance and accuracy of the method and we cannot provide a universally applicable recipe on how to set them. Here we want to showcase how the choice of a parameter tends to impact the method, but only in an isolated case. We also restrict ourselves to a discussion on individual parameters, not their interactions, as the possible combinations are overwhelming. To illustrate their impact we use the identical setup as in the last example, just a factor of 64 smaller with 128×128 Gaussian process parameters and half that data points. We again track the predictive likelihood, RMS error to the ground truth, and average significance of the residual for different meta parameter settings. Here we discuss the sampling accuracy, the number of natural gradient steps for a set of samples, the number of overall samples and the effect of using antithetic samples. We only vary one parameter per example, keeping all other parameters at some reasonable value. The default sampling accuracy are 30 conjugate gradient steps, for a set of samples we make 10 natural gradient steps and use 10 independent samples without antithetic counterparts. The results are shown in Fig. 4.10.

The first meta parameter is the sampling accuracy, describing how many conjugate gradient steps are used to draw an approximate sample according to the covariance. Here one starts with a prior sample and every conjugate gradient iteration removes variance along the eigendirections corresponding to the consecutive largest eigenvalues of the metric. How many steps are required will strongly depend on the problem at hand and especially on the eigenspectrum of the metric. If it drops fast, only a few iterations are sufficient, otherwise more are required. The predictive likelihood, as well as the RMS error in the top row of Fig. 4.10 show that too few iterations will affect the result, but increasing the number rapidly converges towards a common plateau. Already 9 iterations seem to be sufficient in this example. Extremely interesting is the average significance in this case. Regardless of the sampling accuracy, the result will have a consistent error estimate, absorbing insufficient convergence into uncertainties and avoiding a misleading result.

The next meta parameter is the number of natural gradient steps for a given set of samples. This number essentially controls how well an intermediate approximation converges before new samples are drawn at the obtained location. Newly drawn samples will usually not match the true posterior as well as the old samples, for which the KL was optimized, as they probe other directions and it takes some optimization steps to catch up, during which the problem itself is not yet further optimized. Taking too few steps will not lead to good results, as only the sampling stochasticity is chased. This can be observed in the plots. Overall, a deeper convergence for an intermediate approximation reduces the variance of the results and converges better overall. One danger is the over-fitting of the sample realization, not collecting the progress in the mean parameter. This can mainly occur for a small number of samples, as the variances are not probed well.

The number of samples to estimate the KL divergence critically impacts the performance of MGVI. In terms of required computations, everything scales linearly with the number of used samples, so using as few as possible is desired. A single sample is certainly insufficient, as it does not define a variance and the result will be a MAP estimate shifted by the sampled residual, if the KL is fully optimized. The behavior

of MGVI for different sample numbers are shown in the third row of Fig. 4.10. Clearly two and four samples are not enough to converge towards a reasonable solution. For more than eight samples it converges and it seems that more samples allow for deeper convergence and reduced stochastic behavior. It is worth noting that stochasticity is sometimes an advantage, as one can escape local minima, making it more reliable to finding good solutions.

Finally, the last row shows the impact of using antithetic pairs of samples, using not only mean plus the residual as sample, but also minus the residual. This way the mean of the samples and the mean parameter always coincide, stabilizing the gradient estimate significantly, while requiring to draw only half the number of samples. This way MGVI already converges towards reasonable results using only one single sample together with its antithetic counterpart, as shown in the plots. It is still relatively noisy, but compared to using two independent samples, as shown in the row above, far more robust. More samples again reduce the stochasticity even further.

Overall, a higher accuracy, more samples, or more steps are always favorable for higher accuracy of the approximation, but they come at the price of computational effort, so the hard task is to counterbalance those two contradicting goals. Especially towards the beginning of the procedure, high precision might not be needed, as the landscape changes rapidly anyway, but towards the end, as everything starts to settle down and converge, it might be worth to invest into higher accuracy. In our experience we find it useful to gradually tune up the parameters, especially the number of samples to be fast and inaccurate in the beginning and then converge by adding samples, but how to optimally steer MGVI in general is unclear.

4.9.3 Non-Negative Matrix Factorization

In Non-Negative Matrix Factorization models, the data is described as a positive mixture of positive components, or factors. The goal is to find a lower-dimensional description of the data, which can be used to predict unobserved values. The data d should be described by a data matrix D, which is the product of a mixture matrix M and a component matrix C, which are to be learned:

$$D = MC. \tag{4.97}$$

We choose a Gamma-Poisson model, assuming a Poisson likelihood and Gamma-priors on all entries of the matrices, enforcing positivity on all quantities. The problem is standardized by reparametrization defined by the inverse CDF of the Gamma distribution and CDF of the standard Gaussian. Both functions do not have an analytic expression, but can be approximated numerically.

$$M = \mathcal{F}^{-1}_{\mathrm{Gamma}(M|\alpha_M,\beta_M)} \circ \mathcal{F}_{\mathcal{G}(\xi_M,\mathbb{1})}(\xi_M) \equiv f_M(\xi_M) \tag{4.98}$$

$$C = \mathcal{F}^{-1}_{\mathrm{Gamma}(C|\alpha_C,\beta_C)} \circ \mathcal{F}_{\mathcal{G}(\xi_C,\mathbb{1})}(\xi_C) \equiv f_C(\xi_C)\ . \tag{4.99}$$

These equations are to be read element-wise for every matrix entry. The standardized problem information then reads

$$\mathcal{H}(d,\xi) = d^\dagger \ln\left(f_M(\xi_M) f_C(\xi_C)\right) + 1^\dagger \left(f_M(\xi_M) f_C(\xi_C)\right) + \frac{1}{2}\xi^\dagger \mathbb{1} \xi\ . \tag{4.100}$$

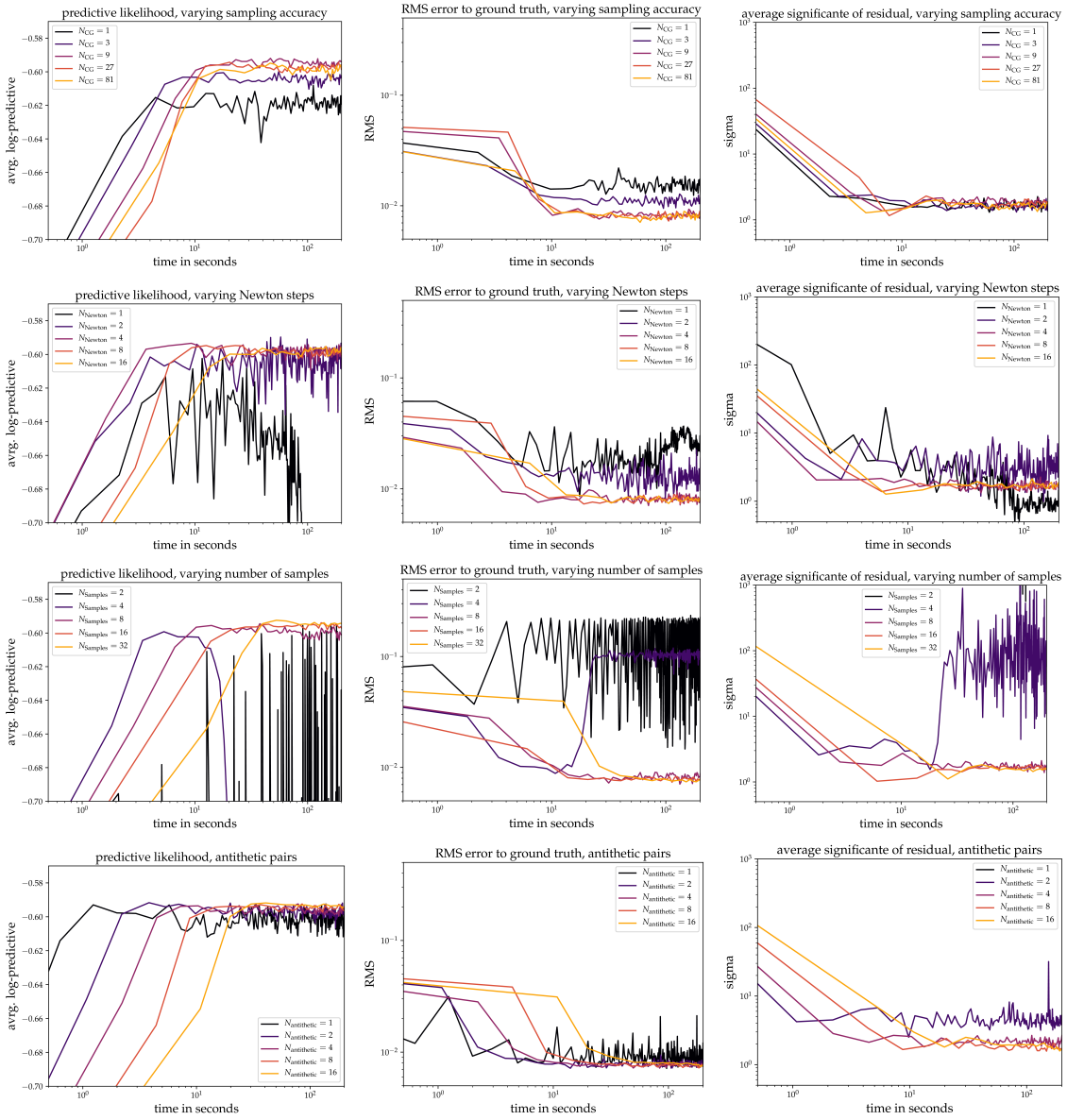


Figure 4.10: The results of the meta-parameter exploration. The different performance metrics are shown from left to right, the different meta-parameters from top to bottom.

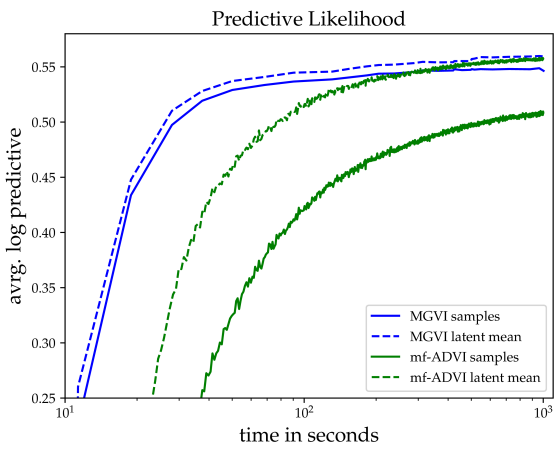
Here ξ is the concatenation of ξ_C and ξ_M. As in the first example, we have again a Poisson likelihood and its metric is given by Eq. 4.81. We apply this model to the Frey face data set, consisting of 1965 images of a sequence of facial expressions in a resolution of 28x20 pixels, assuming ten components. All parameters of the Gamma distribution are chosen to be 1, and we randomly mask 10% of pixels to calculate the predictivity of different methods per elapsed time. In addition to that, the bottom part of one frame is fully covered by the mask, and we will show how well it is recovered. Overall the model has 25160 free parameters to be learned and we compare the performance of MGVI to mean-field ADVI. In this example we do have relatively good data, so it is not as relevant to frequently refresh the samples to explore the uncertainty and we can afford to optimize deeper in each global iteration to achieve overall faster convergence. Therefore we initially perform 10 natural gradient steps together with one pair of antithetic samples, compared to the three steps in the previous example, but otherwise we also increase the number of samples starting after twenty iterations. The initial sampling accuracy are 50 iterations, increasing it to 200 towards the end.

The predictive likelihoods during the optimization for both methods, the results on the half masked frame, as well as the recovered components are shown in Fig. 4.11. The predictivity of the MGVI samples and mean converge rapidly towards the same value, indicating low uncertainty. After 40 seconds MGVI seems relatively converged as the slope strongly decreases. The predictivity of the mf-ADVI mean achieves comparable levels to MGVI after 200 seconds, but the predictive likelihood of the samples is significantly lower. As the discrepancy between predictivity of the mean and the samples are a proxy to the variance of the distribution, it seem that mf-ADVI severely struggles to compress towards the posterior mode, crippling down the overall convergence. MGVI does not have this problem, as the covariance adapts to the environment of the mean, and it can therefore contract towards the posterior mode far more rapidly. Regarding the half masked frame, the mean for both methods matches very closely. Interesting is the pixel-wise standard deviation, shown below the mean. For MGVI it is clearly structured and aligns with regions of facial variability, for example around the mouth and the eyebrows. The variance is especially high in the masked region, whereas mf-ADVI shows less pronounced features, and even lower variance within the masked half.

4.9.4 Hierarchical Logistic Regression

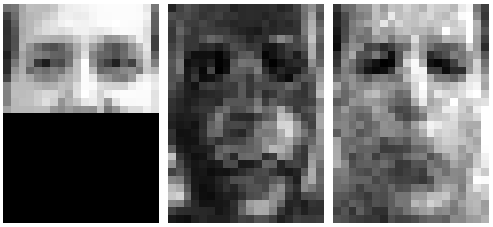
In this example we discuss two hierarchical logistic regression problems involving polling data from the US 1988 presidential election, using the models discussed in Gelman and Hill [45] and we will follow the analysis of Kucukelbir et al. [85]. The data set involves 13544 points on age, gender, ethnicity, education, region, state, and polling behavior. We consider two logistic regression models of different complexity to predict the polling behavior. The smaller model contains only information on state, gender and ethnicity, allowing full posterior sampling with HMC as reference. The larger model utilizes the full data set and it allows us insight into the convergence behavior of MGVI.

The likelihood is given by a Bernoulli distribution, as stated in Eq. 4.87 and corresponding metric is given in Eq. 4.89. The data is the polling result and it is modeled



(a) Predictive likelihood of MGVI compared to mf-ADVI in seconds. Dashed lines are the predictivity of the mean, solid lines of the samples.

(b) ground truth. (c) MGVI mean. (d) ADVI mean.



(e) masked data. (f) MGVI uncertainty. (g) ADVI uncertainty.

(h) MGVI components. (i) ADVI components.

Figure 4.11: The predictivity of both methods (top left), means and standard deviations for a certain frame together with ground truth and data (top right), and recovered components (bottom).

by a rate μ, depending via a logit link on regression coefficients and the design matrix containing X.

$$\mu = \sigma(X^\dagger \beta) \ . \tag{4.101}$$

Here σ is again a sigmoid function.

A Simple Model

For the simple model the rate is described by only a subset of all categories

$$\mu = \sigma \left(\beta_0 + x_{\text{gender}} \beta_{\text{gender}} + x_{\text{ethnicity}} \beta_{\text{ethnicity}} + \beta_{\text{state}}[x_{\text{state}}] \right) \quad , \tag{4.102}$$

with binary data on gender and ethnicity, and multi-class labels on the state. Additionally we set a standard Gaussian prior on β_0, $\beta_{\text{ethnicity}}$ and β_{gender}. To make it a hierarchical problem, all β_{state} coefficients follow also independent Gaussian priors, but with a priori unknown standard deviation σ_{state}, shared among them. We give it a uniform prior on the unit interval. Compared to the model described in Gelman and Hill [45], we choose more restrictive priors for convergence reasons, especially for HMC, but also MGVI. We will elaborate on this later when discussing the full model.

For our analysis we compare MGVI, fc- and mf-ADVI, a Laplace approximation, as well as HMC. Our initial sampling accuracy are only 25 conjugate gradient steps due to the relatively low number of problem parameters, and we increase it to 100 towards the end and otherwise we use the setup from the first two examples. Regarding convergence, we ran MGVI and the ADVI methods for a total of 1000 seconds each, although all, except fc-ADVI, converged within seconds, as did the MAP estimate. After a burn-in and parameter tuning phase we sampled with five chains for several hours, ending with mean Gelman-Rubin test statistic $\widehat{R}_{\text{mean}} = 1.002$ over all parameters and maximum $\widehat{R}_{\text{max}} = 1.009$. The smallest effective sample size was $\text{ESS}_{\text{min}} = 500$, which is the number of samples we use for our analysis. Fig. 4.12 shows scatter plots of different model parameters against each other, comparing HMC to the other methods. MGVI (blue) performs remarkably well, as it is almost indistinguishable from HMC in all cases, matching in mean, variance, and correlation. As expected, fc-ADVI (cyan) also captures the true posterior distribution quite well, but only at extremely high computational cost. As in the previous examples, mf-ADVI (green) does not capture any correlations, but also tends to under-estimate the uncertainty. The recovered mean clearly differs form the sampled posterior mean, and for the more nonlinear σ_{state} parameter, a systematic shift is observed. The Laplace approximation works decently for some parameters, for others only the variance is off, and for other directions straightforwardly fails, as it can be observed to happen most severely in the last panel.

Fig. 4.13 shows the means and standard deviations of all model parameters and methods against the HMC results. Again, in both plots MGVI seems to be superior to mf-ADVI, as well as the Laplace approximation, which is also supported by the RMS errors, as shown in Tab. 4.2. Here MGVI has significantly smaller mean errors compared to all other methods. Compared to fc-ADVI, the error is only a third, and to mf-ADVI one seventh. In the standard deviations the difference is not as severe,

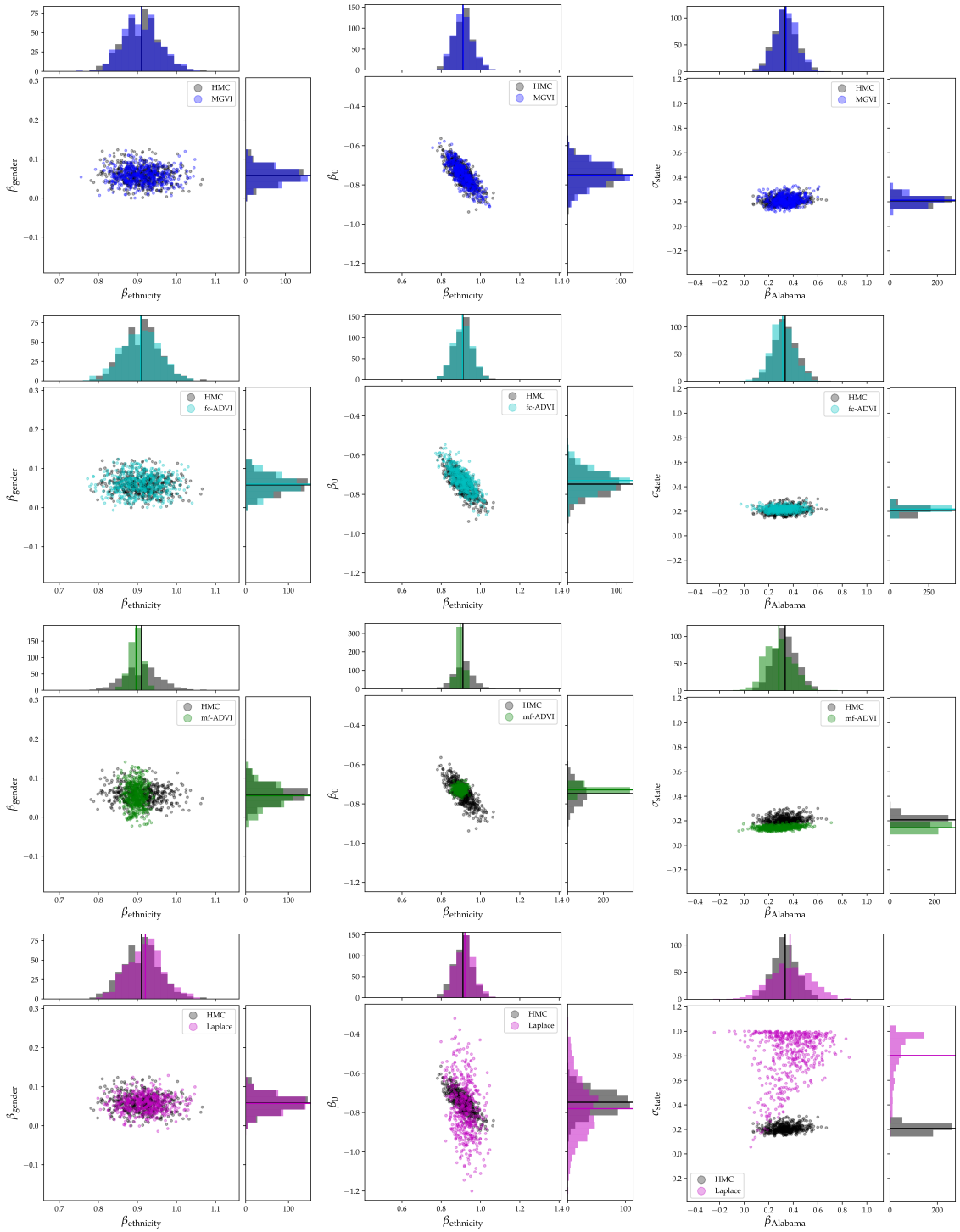


Figure 4.12: Scatter plots for certain parameter combinations for all the different methods in comparison to HMC in the logistic regression example. The parameter pairs vary from left to right, the methods from top to bottom.

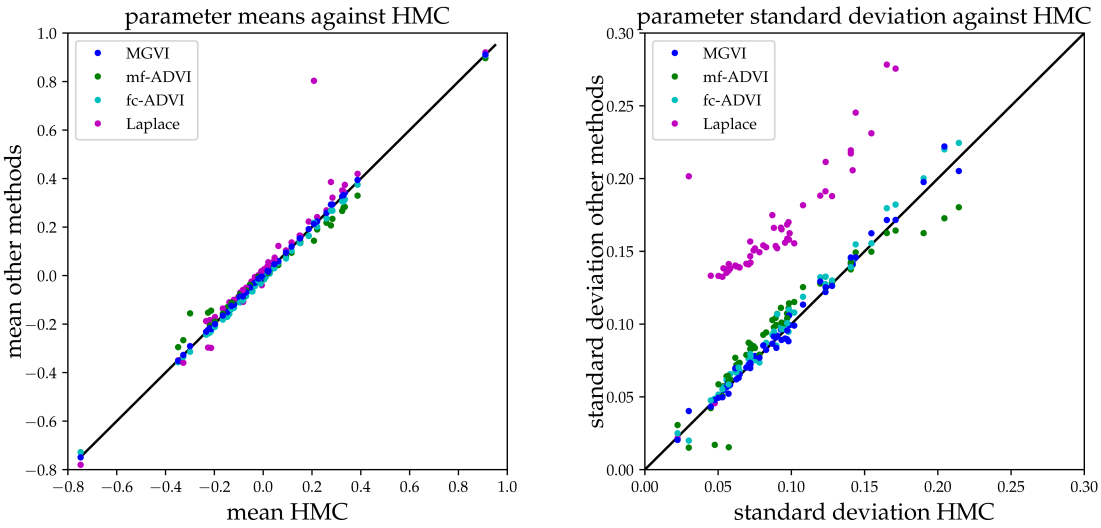


Figure 4.13: Mean (left) and standard deviation estimates (right) for all parameters and methods plotted against the HMC estimates.

Table 4.2: The RMS error of parameter means and standard deviations relative to HMC.

RMS HMC against	MGVI	fc-ADVI	mf-ADVI	Laplace
mean	0.0047	0.015	0.035	0.13
standard deviation	0.0051	0.0067	0.014	0.14

but also there MGVI is the closest to HMC. In the means, the Laplace approximation is only completely off once, namely for the hierarchical σ_{state} parameter. For the standard deviations, Laplace is rarely correct, most uncertainties appear far too large. Several points are outside the plot, with deviations up to 0.9. Overall, MGVI seems to be the best among the tested methods for this problem in terms of accuracy.

The Full Model

The full model, as described in detail in [45], additionally takes further regressors into account, such as the multi-class variables of age, education and region, as well as combinations of categories, and previous election results. In addition, now the coefficients of all categories follow a Gaussian prior with a priori unknown standard deviation. As in the simple problem, a uniform, hierarchical prior with some upper limit is imposed on those. In the original model the interval $[0, 100]$ is proposed, corresponding to largely uninformative prior distributions. From a Bayesian inference perspective, the posterior is extremely far away from the prior distribution, containing much more information. This is a problem for every method starting somewhat close to the prior distribution. For HMC it is hard to find the posterior mode to explore, making sampling in such scenarios inefficient and laborious. MGVI also experiences something similar, which can be seen in Fig. 4.14. In this case it is significantly slower in the beginning, compared to mf-ADVI. Everywhere, except close to the posterior

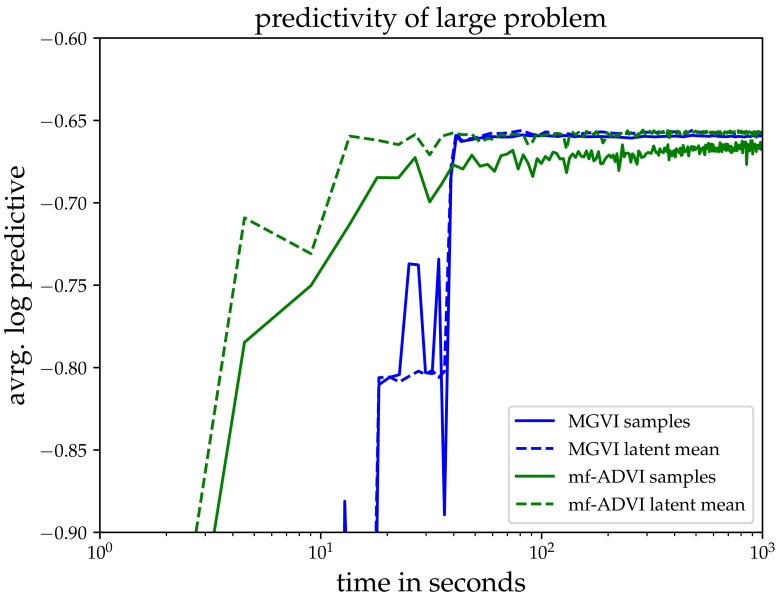


Figure 4.14: The predictive likelihood of MGVI and mf-ADVI in the large logistic regression example.

mode, the metric demands a large variance of the Gaussian and the stochastic nature of the optimization may result in a new location still far away from the posterior mode with practically unchanged metric. Therefore, only by chance the mode is found, and once it is, MGVI will quickly contract its variance and converges. For MGVI we found that in this case the stochasticity introduced by only a single pair of antithetic samples in combination with deep convergence for this given sample yields best results. This way we can escape local minima and flat energy landscapes. The initial position of mf-ADVI is close to a delta distribution and will therefore initially mimic the behavior of a MAP estimate, which is much better suited for such a scenario with unconstraining priors and strong likelihood. The mean converges relatively fast, but the sample average of the predictive likelihood is slow, as seen in the other examples as well.

To overcome the limitations of MGVI in problems with weak priors and strong likelihoods, one could come up with heuristic schemes to artificially reduce the sample variance in the beginning, also imitating MAP, or possibly even starting with mf-ADVI and later on switching to MGVI, keeping the mean estimate.

4.10 Conclusion

We proposed Metric Gaussian Variational Inference (MGVI) as a general method to perform approximate Bayesian inference for high-dimensional and complex posterior distributions. MGVI scales linearly in terms of memory and computations with the problem size, making it applicable in scenarios with millions of parameters. MGVI iterates between approximating the covariance with the inverse Fisher information metric at the current mean estimate and optimizing the KL-divergence to the true

posterior for the current covariance estimate to update the mean. Drawing samples from the approximate distribution via implicit sampling avoids storing the covariance explicitly at any point in time, leading to the linear scaling. The samples are used for an stochastic estimate of the KL-divergence and its gradient. The variance of these estimates can be reduced via antithetic sampling and the optimization is performed via natural gradient descent. The algorithm has converged once the mean estimate is self-consistent with the covariance. The result is a set of samples from the approximate posterior distribution that implicitly represent correlations between all parameters, going beyond a mean-field approximation while circumnavigating the quadratic scaling of an explicit covariance.

In our numerical experiments we demonstrate the accuracy of MGVI by comparing it to HMC samples, outperforming the Laplace approximation, mean-field, and even full-covariance ADVI. In addition to this, in most examples MGVI is significantly faster than the ADVI methods, as well as HMC. Applying MGVI in a diverse set of different contexts illustrates the versatility of the method. In the logistic regression example we have shown that MGVI is intrinsically different to an Laplace approximation, as it finds better solutions in complex models, mimicking the behavior of full-covariance ADVI. MGVI is also suited for large-scale image reconstruction problems with millions of parameters and complex models and it provides accurate uncertainty quantification, which can be used to propagate errors to any derived science result.

For the future it is left to explore the limits of MGVI, both, numerically and theoretically. For the problem dimensionality we do not see any conceptual limitation, except the linear scaling. More problematic is model complexity, and especially degenerate parameter directions. Those lead to numerical stiffness, and therefore slow convergence. Solving certain sub-problems individually, or using tempering methods might result in overall faster convergence. Better justified heuristics for the meta-parameter choices will have to be developed. It is also unclear how to deal with truly large data sets, i.e. too large to work with at once, as the Fisher information metric requires it. How sub-sampling the likelihood and mini-batching the data affects MGVI is to be explored. On the theoretical side, the properties of the proposed covariance approximation and the convergence behavior of the method have to be explored more rigorously.

Finally, we hope that MGVI will open the door to even larger and more complex Bayesian inference problems in the future.

Acknowledgments

We acknowledge Philipp Arras, Philipp Frank, Maksim Greiner, Sebastian Hutschenreuter, Reimar Leike, Daniel Pumpe, Martin Reinecke and Theo Steininger for fruitful discussions.

5 Applications of MGVI

5.1 The Variable Shadow of M87* [13]

The Event Horizon Telescope (EHT) Collaboration published the first image of the immediate vicinity of a black hole [38, 39, 40, 41, 42, 43], revealing its shadow surrounded by accreting plasma. This image was reconstructed from Very Long Baseline Interferometry (VLBI) Radio data. To achieve the required resolution, the signal from multiple radio telescopes was correlated, effectively building an interferometer the size of Earth. The target was the super-massive black hole M87* in the center of the nearby M87 galaxy. Due to its enormous size, this target exhibits a comparable angular extension the black hole in the center of the Milky Way, Sagittarius A*, but as it is roughly a thousand times larger, is far less variable. It only varies on the time-scales of days, instead of minutes, which allows to integrate more data into one image.

The main challenge for better reconstructions is the variability. The temporal evolution of the source has to be resolved, especially in the case of the strongly varying Sagittarius A*, for which no reconstruction or data has been published yet. Introducing an additional time direction dramatically increases the number of required parameters, as an additional dimension is introduced. Also, the data becomes strongly diluted, reducing the available information for every pixel.

As we have shown in Arras et al. [13], the key to overcome the severe data-sparsity in spatial and temporal directions are correlations together with a rigorous probabilistic treatment of the reconstruction problem to account for uncertainty. To demonstrate the capabilities of the method, we obtained the first time-resolved reconstruction of the immediate vicinity of a black hole, using the data of M87*, provided by the EHT collaboration. The reconstruction spans the entire observational cycle of seven days, containing four 8-hour observations.

The following part will briefly discuss the approach the results from this publication. All the details can be found in the paper. I contributed to the team effort through work on the implementation and the testing of the likelihood and the reconstruction algorithm.

A radio-interferometer in general consists of multiple antennas. The measured radio-waves are correlated with each other according to a certain time-delay, which determines the pointing of the telescope. These correlations correspond to a point-measurement of the Fourier-transformed sky-brightness. Imaging is then required to turn the sparsely-probed information on the Fourier plane back into a full image of the sky. VLBI utilizes the same measurement principles, but the antennas belong to a large variety of different instruments, placed thousands of kilometers apart. This makes the calibration, in contrast to conventional radio-interferometers, a lot more challenging. All antenna-based effects can be removed by combining several ampli-

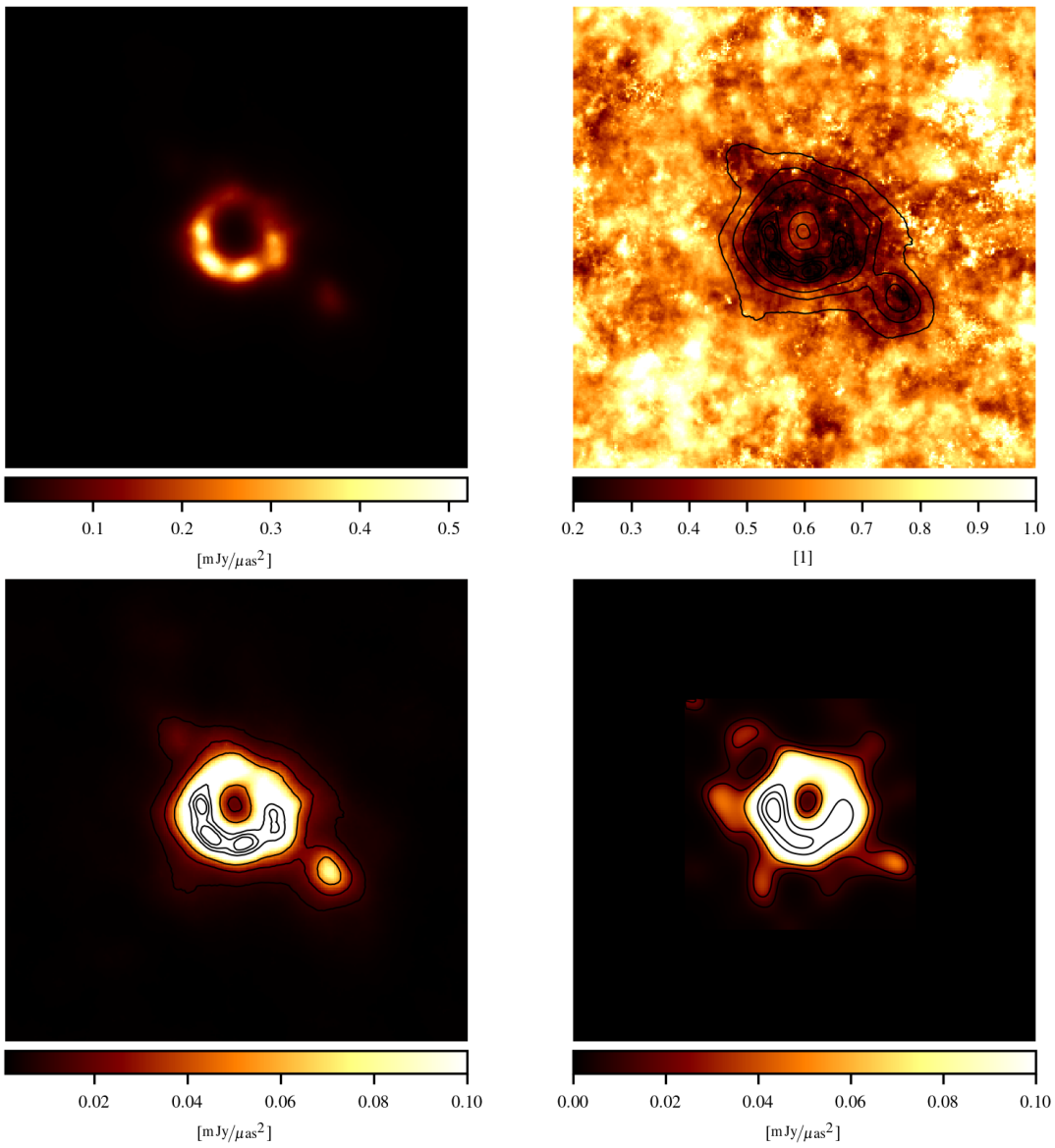


Figure 5.1: The first row shows the reconstructed mean and relative error, averaged over the entire observational period of 7 days. The first panel of the bottom row is a saturated plot of the time averaged posterior mean, revealing the emission zones outside the ring. The overplotted contour lines show flux at 0.009, 0.023, 0.051, 0.188, 0.282, 0.376 Jy/μas^2. The last panel on the bottom row shows the result of the EHT-imaging pipeline in comparison, also saturated and with overplotted contour lines. This figure and caption text is taken from Arras et al. [13].

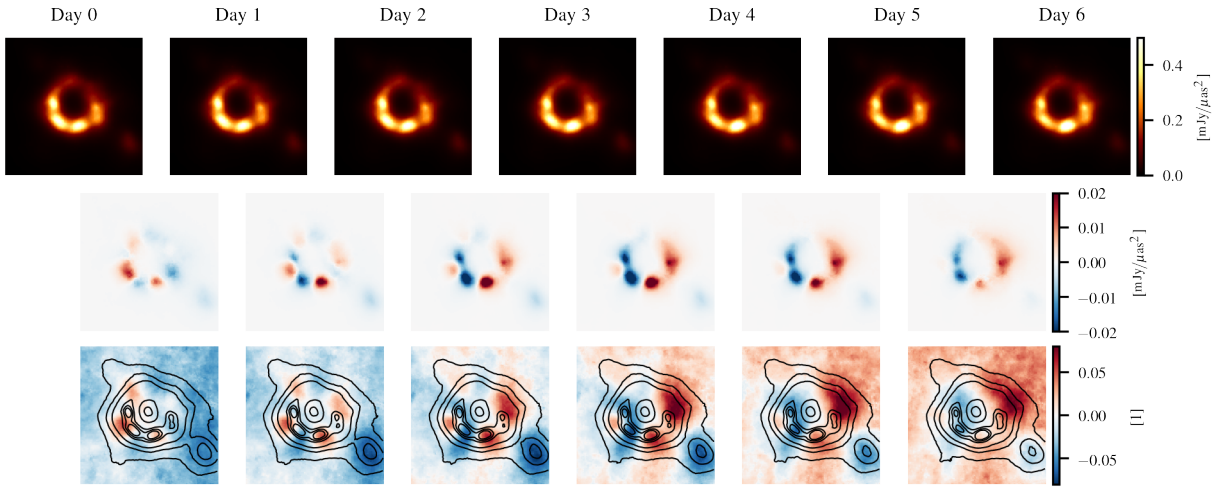


Figure 5.2: Visualisation of the posterior mean. All figures are constrained to half the reconstructed field of view. The first row shows time frames of the image cube, one for each day. The second row visualises the absolute difference between adjacent days. Blue and red visualises increasing and decreasing flux over time, respectively. The third row visualises the relative difference in flux over time. The overplotted contour lines show the flux at 0.009, 0.023, 0.051, 0.188, 0.282, 0.376 Jy/μas^2 of our posterior mean. This figure and caption text is taken from Arras et al. [13].

tudes and phases in a certain way to form closure quantities [17, 115], which are then used to perform the imaging. The problem with these is, that also the information of absolute positions and total flux is lost.

These closure quantities are the basis of the image reconstruction. In order to recover the original signal from this highly incomplete and noisy measurement, it is unavoidable to impose certain assumptions. One fundamental assumption is the positivity of the emitting density. In case of the black hole shadow, we expect an extended emission structure, which exhibits spatial correlation. We also assume a strong correlation in the temporal direction due to the physical size of the object in the order of light days. We also expect exponential brightness variations on linear spatial scales. All these assumptions can be implemented using a log-normal model together with a Gaussian process correlation structure. This correlation is described as an outer product of a spatial and a temporal correlation structure. A priori, we assume independence between those two directions.

The correlation structure reflects assumptions on physical processes. We are not absolutely sure what to expect in the immediate vicinity of a black hole, or at least we want to allow for multiple scenarios. We therefore do not impose the spatial and temporal correlation structure directly, but also learn it from the data. For this, we assume a priori statistical homogeneity and isotropy in the spatial and temporal direction separately. This way, the correlation structure is diagonal in the Fourier domain, according to the Wiener-Khinchin theorem [68, 135], and can be expressed in terms of power-spectra. These are also modelled in terms of a log-normal model

with an integrated Wiener process prior on double-logarithmic scale.

This prior model, together with the closure likelihoods for phase and amplitudes, constitutes a Bayesian inference problem, which we approximately solve using MGVI. We set a spatial resolution of $1\mu\text{as}^2$ and a field of view of $256 \times 256\mu\text{as}^2$. The temporal resolution are 6 hours over 7 days. In addition to this, we have two highly-correlated frequency channels and double the time period to circumvent the periodic boundary conditions of the Fast Fourier Transformation. In total, this results in 7.3 million model parameters, which are constrained by roughly 1000 data points. In scenarios with such extreme data sparsity, it is mandatory to properly propagate uncertainties and to take correlations into account. MGVI enables such reconstructions.

With this, we obtain the first time- and frequency-resolved reconstruction of a black hole shadow. The time-averaged result is shown in Fig. 5.1, together with the uncertainty and a reconstruction from the EHT-imaging pipeline. We clearly recover a circular structure with a crescent-like brightness distribution. The bottom is brighter, whereas the top part fainter, which is probably due to relativistic beaming. The overall result is higher resolved, but fully consistent with previous reconstructions, as published by the EHT collaboration [41]. In addition to the circular structure, we recover a significant emission region directed towards the bottom right corner.

The temporal evolution throughout the observational period is illustrated in Fig. 5.2. Here, the day-averaged images together with their absolute and relative changes are shown. We clearly observe variability on the time-scale of days. On a large scale, we observe a dimming on the left side and an increase in brightness on the right. Certain spots on the ring become brighter or dimmer, most notably on the bottom and slightly to its left. The structure outside the ring gets fainter the entire time.

Overall we introduced a method to perform time-resolved reconstructions of extended sources from VLBI data and demonstrated it's applicability to real-world problems by obtaining the first time- and frequency-resolved reconstruction of a black hole shadow. This method was enabled by the capability of MGVI to efficiently solve such high-dimensional problems. In the future, this method could possibly be used to approach sources with much stronger variability, for example Sagittarius A*.

5.2 Resolving Nearby Dust Clouds [92]

Determining the tree-dimensional structure around us in the Universe is challenging. Due to the enormous distances, most things appear projected onto the celestial sphere. One way to obtain distance information on nearby objects are parallaxes. Changing angles towards the source during the revolvement of Earth around the Sun can be used for triangulation. Larger distances require higher precision because of smaller angles. The Gaia satellite [108] obtained extremely accurate parallaxes for hundreds of millions of stars in our Galaxy, providing an astonishing insight in the three-dimensional structure of our galactic environment. One of its constituents is interstellar dust. It absorbs starlight, preferentially higher frequencies, and thereby effectively reddens the light spectrum. This therefore traces the amount of dust along the line of sight between Earth and the star. Combining this absorption with the distances from parallaxes provides information of the three dimensional distribution of dust. Estimates for both are provided by Gaia [23].

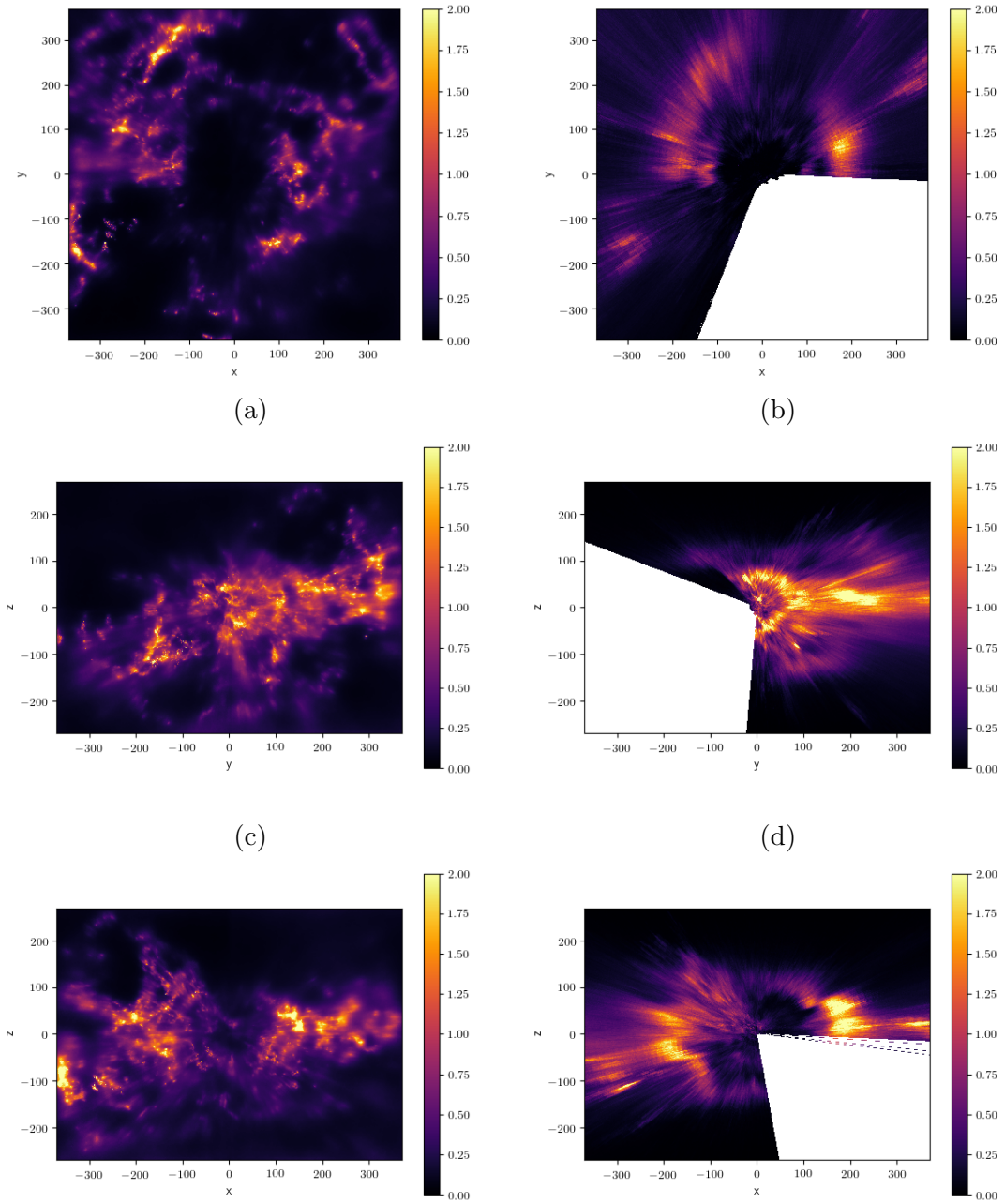


Figure 5.3: Column density comparison of this current reconstruction (Leike et al. [92], left column) and that of Green et al. [51] (right column). The rows show integrated dust extinction for sightlines parallel to the z- x- and y-axis respectively. Note that for Green et al. [51] we show the integrated extinction only if more than 50% of the projected voxels exist in the reconstruction. This figure and caption text is taken from Leike et al. [92].

Leike and Enßlin [91] used these data to reconstruct a three-dimensional map of the dust within a 600pc cube centered around the Sun in unprecedented resolution with more than 17 million voxels. In a improved reconstruction, Leike et al. [92] used a more accurate likelihood together with better extinction estimates obtained from cross-matching. In addition to that, the volume was split into octants around the sun and reconstructed independently. This allowed for an overall increase in volume to $800\text{pc} \times 800\text{pc} \times 600\text{pc}$ and to a voxel-resolution of 2pc. Each of the octant contains 50 million parameters, resulting in a total of 400 million parameters in the reconstruction.

I was not directly involved in this work, but the used tomographic algorithm is based on MGVI, which enabled reconstructions on such a large scale.

Such tomographic reconstructions are challenging for a number of reasons. They need to entangle many interrelated unknowns. All line of sights in the volume start at a star with known distance and end in the center. The most information is therefore available towards the center, where the density of lines is highest and it drops off towards the edges of the reconstructed volume. The main issue with three-dimensional reconstructions is the enormous number of required parameters to achieve high resolutions. A Gaussian likelihood was constructed from the extinction estimates. The prior for the dust density follows a log-normal distribution with a Gaussian process kernel. This way, positivity is enforced, spatial correlations are encoded and exponential density-variations on linear spatial scales are allowed. The correlation structure depends strongly on the physical processes shaping the morphology of the dust. Instead of imposing a certain correlation, it was also recovered directly from the data. For this, statistical homogeneity and isotropy of the logarithmic density was assumed, allowing the correlation to be expressed in terms of a power-spectrum. It was assumed to follow a log-normal distribution with an integrated Wiener process kernel on logarithmic Fourier modes. This model is analogous to the one used for the black hole reconstruction, except that only one single correlation structure for all spatial directions is used.

The results, together with the result from a similar method are shown in Fig. 5.3. It shows the projection of the dust density along the three spatial axes. The reconstruction clearly shows the local bubble, an absence of dust in the immediate vicinity of Earth, as well as fine filaments throughout the volume. Compared to the other methods, the spatial resolution is significantly higher. As MGVI is used to obtain the result, cross-correlations between all parameters have been taken into account and uncertainties on all quantities are available, allowing to propagate errors to any quantity of interest. In the future, it is envisioned to significantly increase the reconstructed volume, as well as the spatial resolution.

5.3 Computed Tomography with Segment-Aware Priors

The fundamental problem of astrophysical and medical imaging is the same. The goal is to reconstruct some underlying truth from noisy and incomplete data. In one case the target is some phenomenon in our Universe, in the other the constituents of some patients body. Medical imaging has become an integral part of modern di-

agnostics and medical research. Especially three-dimensional reconstructions provide deep insights into the causes for certain problems and several techniques are commonly used, for example Computed Tomography (CT), Magnetic Resonance Imaging (MRI), and Positron Emission Tomography (PET). All these techniques are strongly related to astrophysical approaches. A CT relies on line of sight absorption if X-ray emission, fundamentally identical to the dust reconstruction in the previous example. Reconstructing MRI data poses the same problem as an radio-interferometric data, as the Fourier-plane is sparsely probed. For PET scans the decay numbers and locations of radio-nuclides are recorded, analogously to X- and Gamma-ray telescopes. All of these methods produce large amounts of data, probing the three-dimensional structure of the patient.

One goal for the development of reconstruction algorithms is to reduce the number required data. This has two advantages. First, it reduces the required time in the extremely expensive instruments, providing a larger number of patients access to such advanced technology. Second, CT and PET rely on sources for ionizing radiation to penetrate the patients body. This constitutes a health risk and minimizing the exposure is paramount. To achieve this, it is required to use the available data to its fullest extent and the Bayesian framework provides the path to this goal. Using prior knowledge on the constituents of the human body can significantly reduce the amount of required data. Using a probabilistic approach, uncertainties associated to the final reconstruction are provided. Quantifying uncertainty is essential to derive reliable conclusions from the images, on which important decisions on the treatments of patients are based. The work presented in the following is part of the master thesis of Philipp Haim, who developed a probabilistic model to capture many relevant aspects of a human body and used it to reconstruct images from X-ray CT-data. I co-supervised this project and was involved in discussions.

The human body is segmented into several distinct component with certain properties, for example different kinds of tissues and bones. They exhibit characteristic densities, morphology, and are usually auto-correlated. Within the body, every location can only be inhabited by one single component, due to spatial restrictions. Fundamentally, physical densities are measured, which are strictly positive.

For every component, two Gaussian random fields with a-priori unknown correlation structures were used. One expresses the morphology, capturing the characteristic density and lengths-scales. The other indicates how much the component is present at a certain location. Because these indicator fields sum up to one over all components, they exclude each other. One of the components is corresponds to an X-ray transparent material, i.e. air, with density zero to describe the environment around the patient. The number of expected components has to be specified beforehand, but their location and morphology is automatically identified and determined by the reconstruction algorithm. A graphical representation of this model is shown in Fig. 5.5. To recover the correlation structures of all involved fields, the same model as in the other examples was used.

The associated inference problem is highly non-linear and requires several fields the size of the final reconstructed image. MGVI was capable of dealing with the scale and complexity, providing reasonable reconstructions of the image itself, but also for the individual components. Results for the reconstruction of a two-dimensional slice from a X-ray CT-scan of a chest is shown in Fig. 5.4. Here, three components for the body

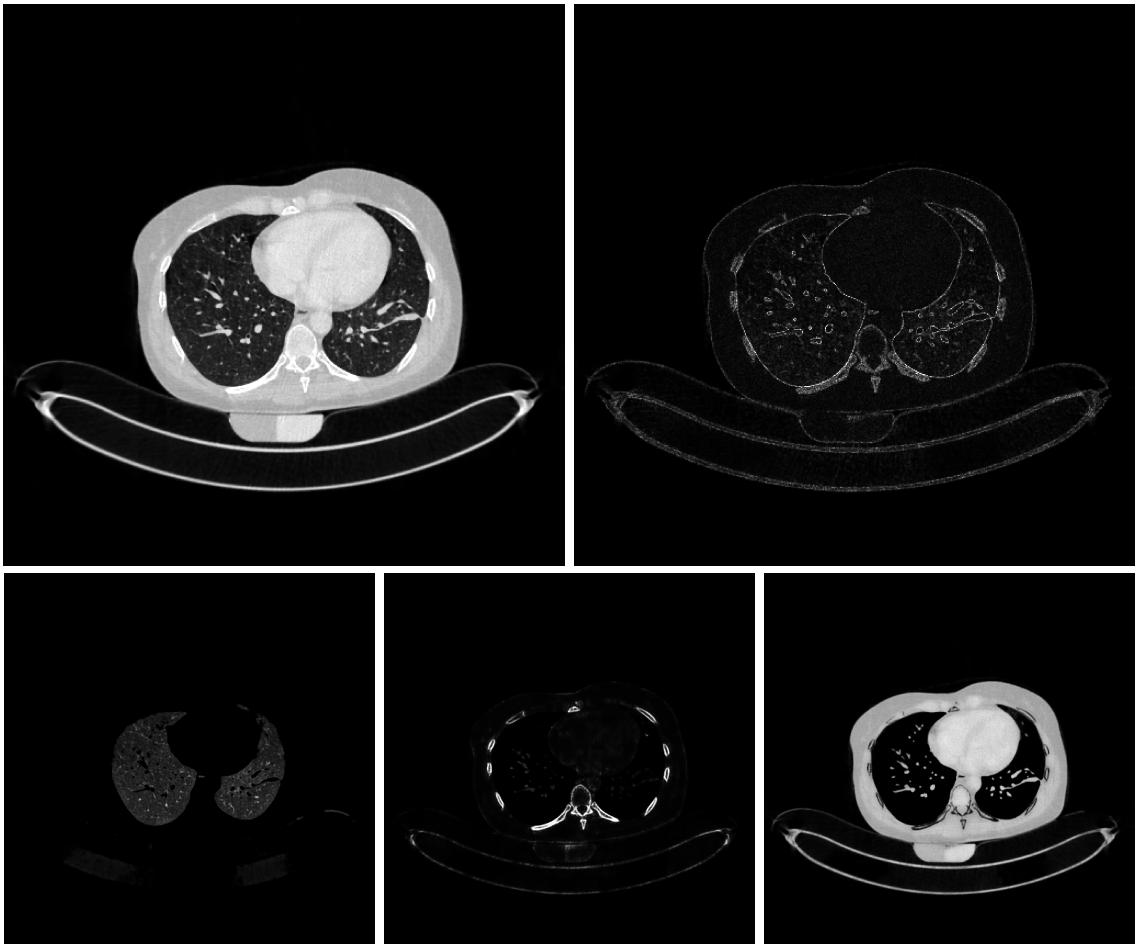


Figure 5.4: The reconstructed image (top left) and its point-wise standard-deviation (top right). The bottom row shows the identified segments. The figures were provided by Philipp Haim.

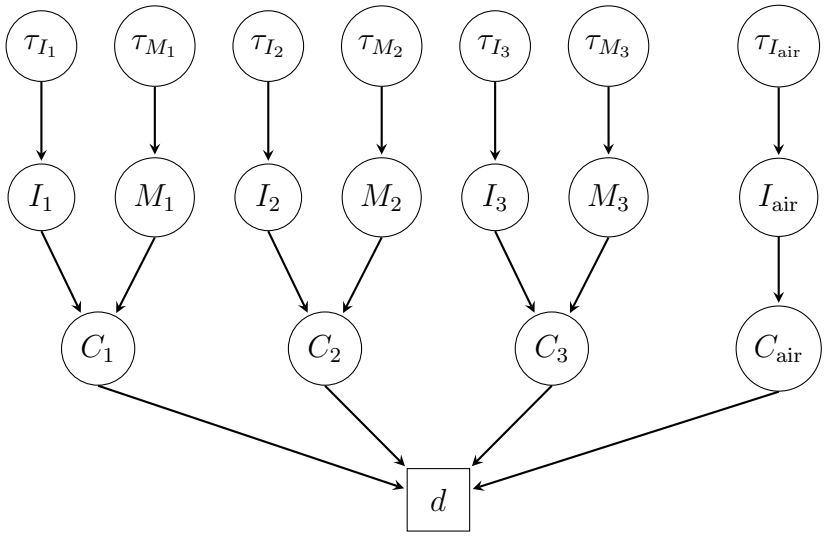


Figure 5.5: The graphical structure of the segment-aware prior model used for the reconstruction of the CT data. Every component C_i consists of an indicator field I_i and morphology field M_i. All these fields have their own, a-priori unknown, correlation structure τ_i. One component is X-ray transparent, omitting a dedicated morphology field. This figure was provided by Philipp Haim.

and one for air were assumed. The overall reconstruction provides a high-resolution image of the patients interior. The three identified components clearly correspond to characteristic constituents of the interior. The first segment shows lung-tissue, which is morphologically distinct to everything else, exhibiting low average density and filamentary structure. The second segment corresponds to bones, showing high characteristic density and larger scale correlations. The final segment contains the combination of several other types of tissue, which share common densities and length-scales. The pixel-wise uncertainty of the full image indicates higher values at the boundary between the distinct segments.

In the future, it is envisioned to extend the reconstruction to three-dimensional reconstructions, utilizing the additional correlations. To additionally reduce the amount of required data, better descriptions of the measurement, including energy-dependent characteristics of the radiation and its absorption, and accounting for count-statistics are pursued.

5.4 The Galactic Faraday Depth Sky Revisited [58]

The polarization of the radiation received from the Universe provides a valuable insight into the magnetic structure of the sources. One example are high-energy free electrons spiralling in magnetic fields, emitting synchrotron radiation in radio frequencies. One issue with the measurement of polarized sources is Faraday rotation. Magnetic fields along the propagation direction of the wave rotate the polarization angle. In our Galaxy, the interstellar medium is filled with magnetized plasma, distorting the polarization in a characteristic manner. The amount of rotation is directly

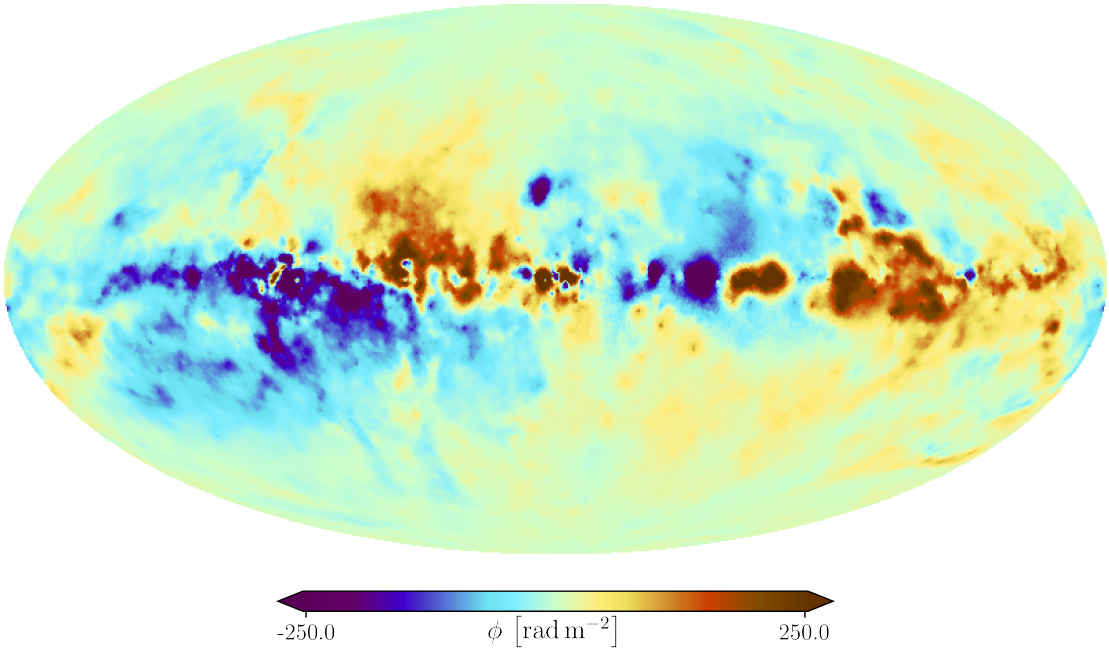


Figure 5.6: The recovered Faraday rotation map using the described method. This figure is taken from Hutschenreuter and Enßlin [58].

dependent on the amount of magnetic fields and the number of thermal electrons along the line of sight, as well as the squared wavelength. Longer wave-lengths are stronger affected and the characteristic wavelength dependence can be used to measure the line of sight integrated combination of thermal electron density and line of sight magnetic field component.

To study polarized, astrophysical sources it is vital to account for the Faraday rotation in our Galaxy. Extra-galactic, polarized point-sources, e.g. pulsars, can be used to estimate the Faraday rotation occurring at individual locations on the sphere. These measurements are highly incomplete with respect to the full sky and have a galactic and extra-galactic Faraday contribution. The extra-galactic component is spatially uncorrelated, but has unknown magnitude for every source. The galactic Faraday rotation does exhibit spatial correlation, as it follows the distribution of magnetic fields and thermal electrons around us. These electrons are also emitting free-free radiation in the microwave regime, as observed by the Planck satellite. This emission is Bremstrahlung due to the Coulomb interaction between the electrons and ions. Hutschenreuter and Enßlin [58] used rotation measures from extra-galactic point-sources together the free-free emission maps provided by the Planck satellite to obtain improved maps of the Galactic Faraday rotation. I was not directly involved in this work, but it beautifully showcases how MGVI can be used to approach large-scale and complex inference problems with multiple distinct data sets in a holistic manner. Several quantities are inferred simultaneously to model the interesting physics, as well as to account for numerous systematics. The Faraday map is the product of an amplitude component and magnetic field component. The amplitude contains the

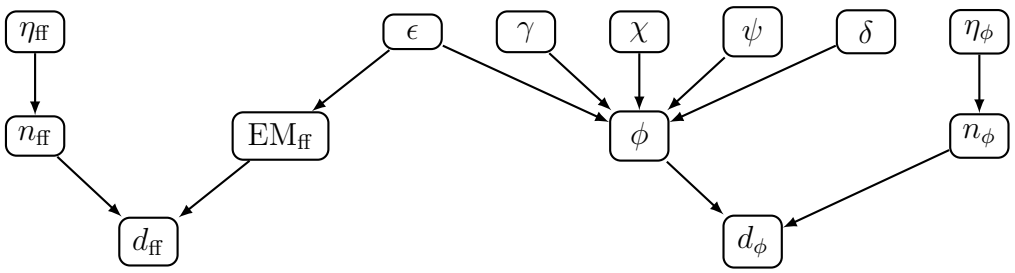


Figure 5.7: The graphical structure of the Faraday model. The figure is taken from Hutschenreuter and Enßlin [58]. Here, ϵ represents the electron density, which enters both maps, the free-free, as well as the Faraday map. γ, ψ, and δ represent systematic effects and χ determines the direction of the magnetic field. All combined make up the Faraday map ϕ. The free-free map EM_{ff} is only modelled in terms of ϵ. The noise covariances n_{ff} and n_ϕ are dependent on th parameters η_{ff} and η_ϕ and are associated to the free-free data d_{ff} and rotation measures d_ϕ. All quantities in the top row correspond to the model parameters. The middle row are the derived quantities and the bottom row contains the data.

electron density, which also imprints into the free-free map. Three additional fields are used to model certain systematic deviations between between the Faraday map and the free-free map. Spatial correlations for all these quantities are assumed, which are a priori unknown. The same approach as in the previous examples is used to learn them simultaneously from the data. In addition to this, the noise covariance on the input data is unreliable, due to either Faraday rotation occurring in the host galaxy of the sources, or component separation artefacts in the free-free map. To account for this, the noise covariance on both data sets is also modelled in terms of additional parameters for each data point. Here, the inverse gamma distribution is used as a prior on the entries of the diagonal noise covariance. This allows for potentially high uncertainty associated with individual data-points, accounting for systematic errors. Overall, the model contains five fields on the sphere, corresponding to magnetic field strengths and directions, electron densities and systematics. All these have unknown correlation structures. In addition to this, a noise covariance for all data points is reconstructed. The graphical structure of the probabilistic model is shown in Fig. 5.7. MGVI is then used to solve for all quantities simultaneously.

The recovered Faraday map using this method is shown in Fig. 5.6. It provides a significant improvement in regions that are only sparsely sampled by the original Faraday data. This map generally agrees with previous maps, but overall resolves smaller structures. Especially, the disk region in the center is improved. An additional result of this reconstruction is a modified free-free emission map (not shown), which only contains contributions that are consistent with the rotation measures, removing certain artefacts.

In the future, it is straightforward to extend this model to contain additional data sets or use models with further physical knowledge.

5.5 Unified Radio Interferometric Calibration and Imaging with Joint Uncertainty Quantification [12]

To conclude this chapter, I want to mention one last application of MGVI. Radio-interferometers are extremely sensitive devices that allow to probe the non-thermal Universe. Modern interferometers allow to measure with ever more accuracy and resolution. All these instruments require careful calibration to unfold their full potential. One important problem is the time-variable screen between the antennas and the source due to atmospheric or ionospheric effects. It affects the amplitudes and phases of the received electromagnetic waves, which leads to apparent shifts and changes of brightness of the source. Therefore, a calibrator source with known location and brightness is usually also observed. During the measurement, the telescope switches between the science target and the calibrator in regular intervals. For the analysis, the calibrator data is used to find solutions for the amplitude and phase calibrations, which are then used for the imaging of the science target.

Instead of splitting this procedure into two separate tasks, Arras et al. [12] simultaneously performs the antenna-based calibration and imaging. To model diffuse emission, a log-normal model with a Gaussian process prior and a priori unknown correlation structure is assumed, analogously to the other examples. To make use of the temporal correlation of the calibration solutions, a Gaussian process model for the phases and a log-normal model with Gaussian process prior for the amplitudes in the temporal direction was assumed.

By combining the calibration and the imaging, the advantages are manifold First, the science target itself is typically static on timescales of the measurement. This helps to further constrain the calibration solutions in between calibrator observations. Better calibrations in turn lead to better images, possibly reducing the required time devoted for calibration, or obtain improved results. MGVI implicitly also takes the cross-correlation between all quantities into account, which allows to propagate calibration uncertainty into the final result. A reconstructed image of the supernova remnant SN1006 observed with the VLA telescope together with its uncertainty and exemplary phase and amplitude calibrations are shown in Fig. 5.8.

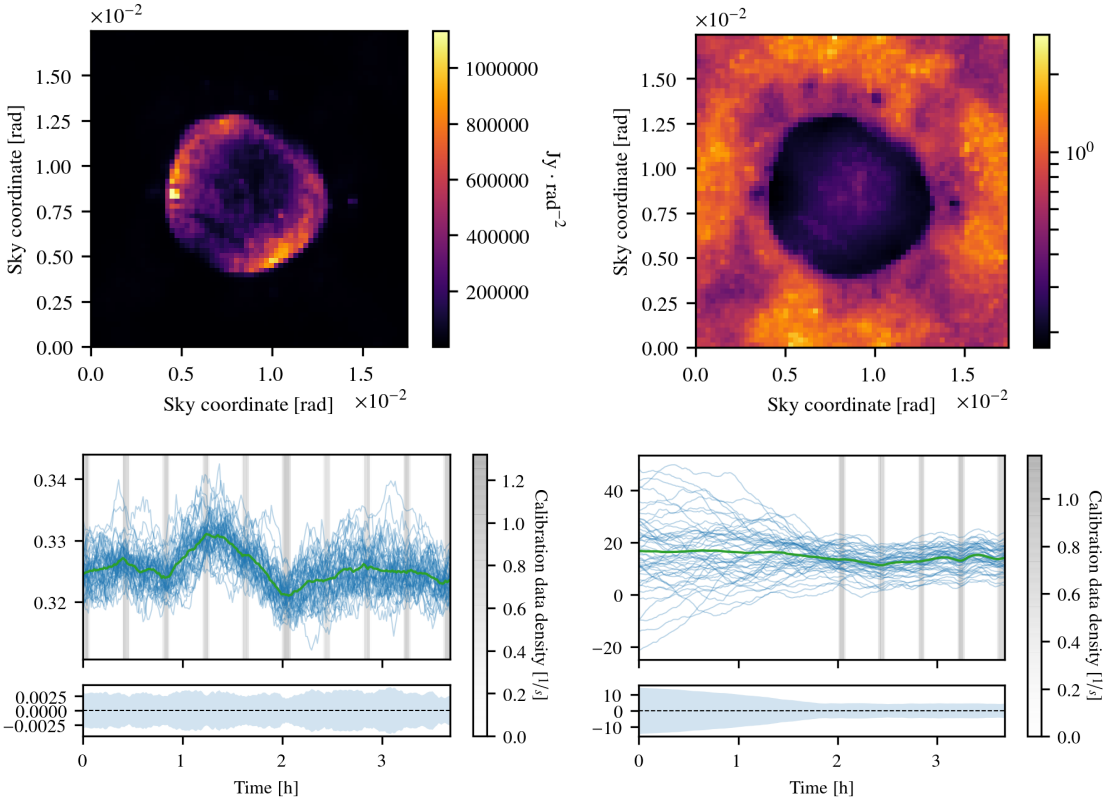


Figure 5.8: The results for a simultaneous calibration and reconstruction of SN1006. The figures are taken from Arras et al. [12]. The top left shows the obtained image, top right contains the corresponding relative uncertainty. The bottom shows the temporal evolution of the amplitude (left) and phase (right) calibration throughout the observation.

6 Bayesian Reasoning with Deep-Learned Knowledge

This chapter is used as a publication currently submitted to the Conference on Neural Information Processing Systems (NeurIPS) [73]. My contribution includes the development, implementation and testing of the idea and all examples. I also wrote the contents. Torsten Enßlin was involved in all discussions and provided valuable feedback on the entire manuscript. All authors read, commented, and approved the final manuscript.

6.1 Abstract

We use independently trained neural networks to represent abstract concepts and combine them through Bayesian reasoning to approach tasks outside their initial scope. Prior knowledge is provided by deep generative models and classification or regression networks are used to express knowledge on complex features of the system. The task at hand is then formulated as a Bayesian inference problem, which we approximately solve through variational or sampling techniques. We demonstrate how this leads to an alternative way to obtain conditional generative models. By imposing multiple constraints at once, we formulate riddles and solve them through reasoning. We also demonstrate how additional information on features can be combined with conventional noisy measurements to reconstruct high-resolution images of human faces.

6.2 Introduction

Reasoning is the act of combining knowledge from multiple sources to come up with the solution to a problem. If the available information is incomplete and uncertain, the knowledge is represented in terms of probabilities. Bayesian inference provides the framework to approach this task, as it is the generalization of Aristotelian logic into the realm of uncertainty [30].

The questions that can be answered this way are limited in one's capability to formulate the problem mathematically. Elaborate models require a deep understanding of a system, but allow deep insights in its inner workings. Nevertheless, many interesting systems are far too complex and their features too abstract to directly write down a mathematical model. Here deep learning provide a way to capture the complexity by training neural networks on examples. The trained network is then a surrogate model of the concept it was trained on.

In this paper we want to use several networks trained on different tasks to jointly answer novel questions on complex systems by performing Bayesian reasoning. We

achieve this by using deep generative models, as obtained from Variational Auto-Encoders (VAEs) [69] or Generative Adversarial Nets (GANs) [50], as prior distributions to describe these systems. Deep classification or regression networks allow us to express knowledge on abstract properties, which we can use to constrain them further. The answer to a question is then given by the posterior distribution over the latent variables of the generator subject to the posed constraint. We will approximate this posterior distribution either variationally [16], or explore it using Hamiltonian Monte Carlo [34].

This allows us to rely on already trained networks and flexibly use them in different contexts. We do not have to train an entirely new network for every task and only have to appropriately arrange available building blocks.

Our approach could be used to develop a wide range of novel methods. In this place we can only provide a proof of concept of the fundamental idea and illuminate certain facets. We do not yet understand all aspects and implications, rendering our algorithmic choices ad-hoc. In our examples we first discuss a generator subject to one constraint, second, how to combine multiple constraints to solve a non-trivial riddle through reasoning, and third, how to combine the approach with conventional measurement data in a large-scale inference problem using current network architectures.

6.3 Related Work

Deep generative models as description of complex systems are especially helpful in various imaging related problems, such as de-noising, in-painting or super-resolution by recovering the latent variables of the generator [21, 93]. Even untrained generators can provide good models for such tasks [132]. The mentioned methods rely on point-estimates. Using the deep generative models in a Bayesian context, the posterior distribution allows to account for complex uncertainty structures [19, 139]. In essence we follow these works and extend them by adding constraints on abstract system features, which are expressed through trained classification or regression networks.

Many of the tasks above are also directly approached by deep learning through end-to-end solutions, for example super-resolution [33]. Similarly, our approach provides an alternative path towards conditional generative models. Usually these are obtained by providing labels or more general constraints [98, 111, 133, 140] during a training phase. We instead use unconstrained generators and flexibly add further constraints through independently trained networks in a modular fashion, allowing to combine information from multiple sources. Similar methods also manipulate the latent variables of unconstrained generators to change samples into a desired direction [141]. Also more implicit information in form of the preference of one sample over the other can guide the generation [55]. Our posterior distribution is in principle also a manipulation of the latent variable distribution, such that the constraints are fulfilled.

Interestingly we can regard our approach as a form of continual variational learning [100]. When approximating our posterior variationally, we conceptually add an additional set of layers to the original generator, which then satisfy the posed constraint. Repeating this for multiple constraints, we can partially train generators through abstract knowledge on the subject, instead of data only.

6.4 Deep-Learned Knowledge

6.4.1 Deep Generative Priors

Knowledge on any system can be encoded within a probability distribution $x \sim \mathcal{P}(x)$. An equivalent representation of this knowledge is a generative model $x = G(\xi)$. These transform latent variables with simple distributions $\xi \sim \mathcal{P}(\xi)$ to system realizations x. Deep generative models, such as GANs and VAEs, allow to represent complex systems that evade an explicit mathematical formulation. They acquire their understanding through large amounts of examples within a training set.

Conceptually, the use of generative models corresponds to the reparametrization trick [69] applied to the model parameters. The Bayesian inference problem in terms of the thereby introduced latent model variables is

$$\mathcal{P}(\xi|y) = \frac{\mathcal{P}\left(y|G(\xi)\right)\mathcal{P}(\xi)}{\mathcal{P}(y)} \ . \tag{6.1}$$

This way, we have a generative Bayesian model with a simple distribution as prior. This equation tells us how any kind of new information through y restricts the latent parameters ξ. Without loss of generality we will be using a standard Gaussian distribution over the latent variable $\mathcal{P}(\xi) = \mathcal{N}(\xi|0, \mathbb{1})$, as this provides a convenient parametrization for inference [15, 85]. If the original generator was not trained on Gaussian latent samples, we can convert them to such through an additional transformation.

6.4.2 Constraints through Neural Networks

Knowledge on a system x is often represented in terms of a constraint $y = F(x)$ involving some feature extracting function F. In our case the functions F will be neural networks trained to extract certain features. Given a desired feature value y as data, a candidate system x is judged for its adherence to the feature via a likelihood probability, i.e. $\mathcal{P}(y|F(x))$. For continuous quantities a convenient choice is the Gaussian distribution

$$\mathcal{P}(y|F(x)) = \mathcal{N}(y|F(x), N) \ . \tag{6.2}$$

Here $F(x)$ serves as the mean of the Gaussian and N is the covariance. The more certain we are about the estimate, the narrower we can center the distribution around the mean.

In the case of discrete categories, the function F might provide classification probabilities $p_i(x)$ for the feature of x being in class i. A more appropriate choice for a likelihood could then be the categorical distribution

$$\mathcal{P}(y|F(x)) = \mathcal{C}(y|F(x)) = p_y(x) \ . \tag{6.3}$$

This distribution describes the outcome of one draw. It does not directly encode how much we trust the estimate of the network. A simple way to introduce this is to raise this distribution to a certain power α.

$$\mathcal{P}(y|F(x)) \propto \mathcal{C}^\alpha(y|F(x)) = p_y^\alpha(x) \ . \tag{6.4}$$

Positive integer values for α are equivalent to multiple consecutive draws with the corresponding outcome, resembling the multinomial distribution. This has roughly the same effect as the narrower variance of the Gaussian in the continuous case and we will use this parameter to encode our certainty in a categorical feature.

6.5 Bayesian Reasoning with Deep-Learned Knowledge

We now use a deep generative model $x = G(\xi)$, which encodes our prior knowledge on the system and feed its output into the classification or regression network $F(x)$ to check whether the abstract property is fulfilled. This concatenation $F \circ G(\xi)$ relates the latent variable to the data in the likelihood. The prior itself is the source distribution of the latent variables ξ. Bayes theorem allows us to combine the associated probability distributions to obtain the posterior distribution over latent variables that are compatible with the constraint,

$$\mathcal{P}(\xi|y) = \frac{\mathcal{P}\left(y|F \circ G\left(\xi\right)\right)\mathcal{N}(\xi|0,\mathbb{1})}{\mathcal{P}(y)} \ . \tag{6.5}$$

This poses a non-conjugate Bayesian inference problem in terms of the latent variables of the generator. Arbitrarily many constraints, either abstract through networks or conventional measurements, can be considered by including additional likelihoods.

6.5.1 Approximate Inference

Due to the nonlinear structure in F and G, the evidence $\mathcal{P}(y)$ will not be available analytically, so we have to rely on approximations to the posterior distribution. The associated approximation problem might be challenging due to the high dimensionality of the posterior and its hardly comprehensible shape due to the posed constraints. Choosing the right method for an application will highly depend on the requirements, but this does not change how the problem is approached. In general, we want to capture the true posterior distribution as closely as possible. Sampling techniques, such as Hamiltonian Monte Carlo (HMC) [34] allow us to draw samples from this true posterior, but require large amounts of computational resources. This method allows us to verify the fundamental validity of our approach and we will use it in one of our smaller examples.

Variational inference [18] can be much faster than HMC and does not only provides samples, but an entire probability distribution. This distribution can be used as a prior in a future problem to perform continual learning. The true posterior $\mathcal{P}(\xi|y)$ is approximated with another distribution $\mathcal{Q}_\varphi(\xi)$ within a parametrized family by minimizing their Kullback-Leibler divergence [86] with respect to the variational parameters φ, which is equivalent to maximizing the Evidence Lower Bound (ELBO) [16]. The reparametrization trick [69] allows to express the approximation in terms of a deterministic function $\xi = H_\varphi(\zeta)$ and a transformed random variable ζ that follows a simple source distribution. We stochastically estimate the ELBO and its gradient through samples from the approximation [56]. The deterministic reparametrization is a generative model for latent variables ξ that are compatible with the posed constraint.

Figure 6.1: The posterior mean (leftmost columns) and samples (other columns) obtained from HMC (left) and a Gaussian mean-field approximation (right).

Therefore, the concatenation of the unconstrained generator with this reparametrization, i.e. $G \circ H_\varphi(\zeta)$, gives a conditional generator with the same mathematical structure as the original one. Thus, the variational inference provides an additional set of network layers on top of the original input layer, which are responsible to satisfy the additional constraints.

The accuracy of variational inference depends on the capability of the approximate distribution to capture the true posterior. Flexible approaches, such as Normalizing Flows [112] allow in principle for arbitrary accuracy, but at high computational costs. For this reason, here we will only consider simple Gaussian approximations, which are significantly faster. To avoid an explicit parametrization of the full covariance, we will use a mean-field approach [102]. Interestingly, due to the standard Gaussian prior, this is equivalent to Automatic Differentiation Variational Inference [85]. As additional method we will be using Metric Gaussian Variational Inference (MGVI) for larger problems [76], which also captures correlations between all quantities implicitly.

6.6 Demonstrations

6.6.1 Conditional Generators

In the first example [72] we want to illustrate how our approach provides an alternative way to obtain a conditional generative model. As likelihood we use the categorical distribution, containing the trained classification network $F(x)$ attached to the output of the generator $x = G(\xi)$. The prior over latent variables ξ is the source distribution of the generator, i.e. the standard Gaussian. The posterior is then proportional to the product of prior and likelihood with a certain choice for α to control its strength.

$$\mathcal{P}(\xi|y) \propto \mathcal{C}^\alpha(y|F \circ G(\xi))\, \mathcal{N}(\xi|0, \mathbb{1}) \tag{6.6}$$

Here we constrain a generator of hand-written digits to a certain label. As generative model we use a Wasserstein-GAN [8, 52] with three hidden layers, convolutional

architecture, and 128 latent variables trained on the MNIST dataset [90]. The digit classification is performed by a deep three-layer convolutional neural network [83] trained on cross-entropy and achieving 98% test accuracy. We strongly enforce the constraint by setting $\alpha = 100$. All networks are implemented in tensorflow [1] and the inference problem is solved in NIFTy [10]. In the next two examples we used a single core of an Intel Xeon CPU E5-2650 with 2.3GHz. We did not fully optimize for run-time.

In this example we explore the true posterior distribution via HMC sampling. For every digit we use eight chains to draw $80,000$ samples in total, after disregarding an initial burn-in and tuning phase. We are aiming for an acceptance rate of 0.6 and adapt a diagonal mass matrix. For every sample we perform 10 leapfrog integration steps and all chains are initialized at a prior sample. One chain requires 36 minutes. To ensure convergence, we calculate the Gelman-Rubin test statistic $\widehat{R}$ for all latent variables [46]. Throughout all digits, the largest value we encounter is $\widehat{R} = 1.03$ for the constraint that the digit is a zero. For all other cases, the largest value is about another order of magnitude closer to unity. The mean value over all variables and cases is $\widehat{R} = 1.001$, which indicates well-converged chains. The smallest effective sample size, which accounts for auto-correlation is $N_{\text{eff}} = 4169$ for one parameter of the digit four and on average $N_{\text{eff}} = 5474$. This small difference indicates that we explore almost all directions equally well. This could be due to the Gaussian prior and only a small amount of additional information provided by the likelihood, making it a well-conditioned posterior landscape. We judge the quality of the samples by checking whether they satisfy the posed constraint according to the classifier. Surprisingly, in all cases, except for the digit 4, all samples are classified correctly. For this exception still 98% of the samples fulfill the posed constraint. For these results we used every tenth sample, i.e. 8000 for each digit.

We also perform a variational mean-field approximation with a Gaussian. For this we follow the same procedure as described in the next example, optimizing for 600 seconds, starting with 10 samples and increase them every 120s by another 10. For our analysis we use 300 samples from the resulting distribution. Compared to the HMC samples, we have slightly more errors with an average accuracy of 99% and only the digits zero, one, and five are exclusively classified correctly. The mean for every digit, as well as representative samples for both methods are shown in Fig. 6.1. As HMC explores the full posterior, we obtain morphologically diverse samples. These also expose the shortcomings of the original generator. The variational approximation provides more distinguished, but highly uniform samples, which is to be expected due to under-estimation of true variance by the mean-field approach.

6.6.2 Solving Riddles

Here we want to solve a riddle by enforcing multiple constraints simultaneously. We know a priori that we are looking for three single-digit numbers. We want them to fulfill the five constraints outlined in Tab. 6.1. The only viable solution is the combination 134. In the model the three digits are generated through three instances of the same generator used in the previous example, i.e. $x = G(\xi)$, resulting in a total of 384 latent variables in ξ. For each of the five constraints we assemble a function $F_i(x)$ that checks whether it is fulfilled or not. For the constraints I,IV and V those

Table 6.1: The riddle discussed in Sec. 6.6.2.

	There are three numbers:
I.	The first number is odd.
II.	The second is two larger than the first.
III.	The first plus the second equal the third.
IV.	The third number is not a seven.
V.	It also contains no closed circle.

correspond to three independently trained convolutional neural networks applied to the respective digit. For the fourth constraint we re-use the digit classification network from the previous example. The other two constraints use the same architecture, but are trained on the respective task. The remaining constraints II and III involve multiple numbers simultaneously. Both require again the classification probabilities of the digits to calculate how likely they are satisfied. For every digit this is a 10 dimensional vector. The mathematical logics are directly implemented into the model, represented by a 2-tensor A and a 3-tensor B with ones at locations corresponding to valid expressions and zeros elsewhere. Contracting these tensors with the classification probabilities provides the overall probability the constraint is fulfilled or not. This way we included explicit domain knowledge into the reasoning system. The graphical structure of the inference problem is outlined in Fig. 6.2.

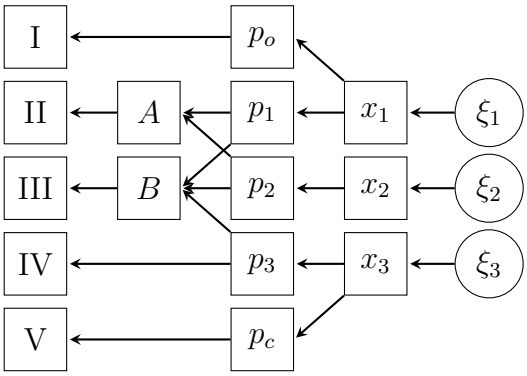


Figure 6.2: The graphical structure of the riddle solver. Here p_i indicates the classification probability given the respective network. (o for odd, c for circle and 1,2,3 for the digits)

The solution is given in terms of the posterior distribution, which is proportional to the product of all likelihood terms and the prior.

$$\mathcal{P}(\xi|y_I,\ldots,y_V) \propto \prod_{i\in I\ldots V} \mathcal{C}^{\alpha}\left(y_i|F_i(G(\xi))\right) \times \mathcal{N}(\xi|0,\mathbb{1}) \tag{6.7}$$

Here we perform a Gaussian mean-field approximation using variational inference. The challenge in this case is the multi-modality of the posterior distribution. There are many combinations of numbers that partially fulfill the constraints and therefore constitute a mode in the posterior. An annealing [116] strategy partially mitigates

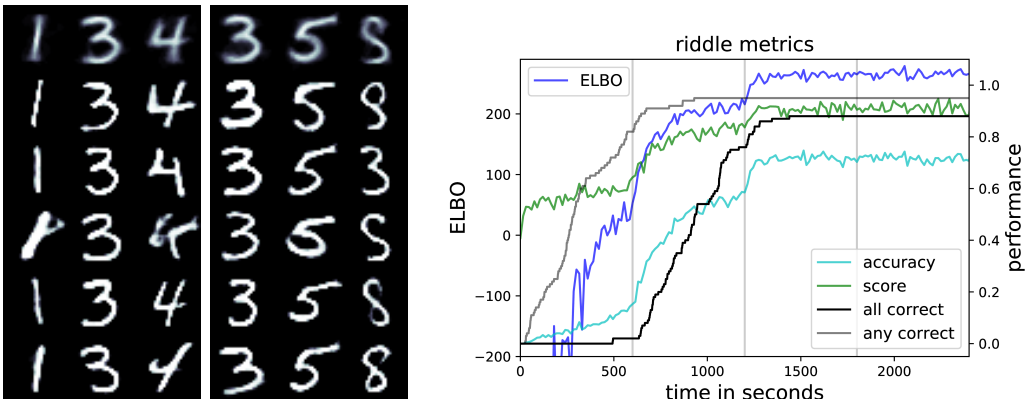


Figure 6.3: Left: Two examples for provided solutions for the riddle. The left one shows a correct solution, whereas the other one is only almost correct. The top rows show the ensemble means. Right: The ELBO and various other metrics of the riddle during the optimization. Vertical lines indicate the increases in α.

this issue. Initially we choose a small α to allow the optimization to explore the posterior landscape and later on increase it to learn the structure of the resulting mode.

Regarding the optimization scheme, we had issues with the convergence of common momentum-based stochastic optimizers. We found that non-stochastic optimizers work reasonably well in combination with good estimates of the gradients. Antithetic sampling [84] allows to reduce the stochasticity of the estimates on gradient and loss. Every sample drawn from the approximation is accompanied by a totally anti-correlated partner, obtained by mirroring the sample at the center of the Gaussian. We then employ a line-search along the gradient direction to update the variational parameters. Five sample pairs are used to estimate the ELBO, its gradient and the other quantities. The overall run-time is 2400 seconds. We choose $\alpha \in \{0.5, 1, 3, 10\}$ and increase it after every 600 seconds. To illustrate the behavior quantitatively we repeat the approximation for 100 different random seeds.

In the end, 88% of the runs end up with the correct solution. To track the progress during the optimization we follow the evolution of five quantities. First, the ELBO for the final $\alpha = 10$ as the overall optimization goal. Second, the average conditional categorical likelihood over samples and constraints as a score to quantify how well the conditions are met. Third, our accuracy in terms of the average conditional categorical likelihood that the samples show the correct solution. These quantities are between zero and one. The fourth and fifth show cumulatively the fraction of runs that were able to achieve any or exclusively correct samples up to that time. All these quantities, averaged over the hundred runs, are shown in the right panel of Fig. 6.3. The ELBO increases in three steps, plateauing before the increases in α. There seems to be no large difference for the posterior whether α is 3 or 10. In the first quarter the constraints are only weakly enforced and the approximation can explore the posterior distribution. For most runs the first occurrence of a correct sample falls in this section and afterwards this line flattens. The capability to explore is in contradiction to well approximate the local mode, so only in a small number of

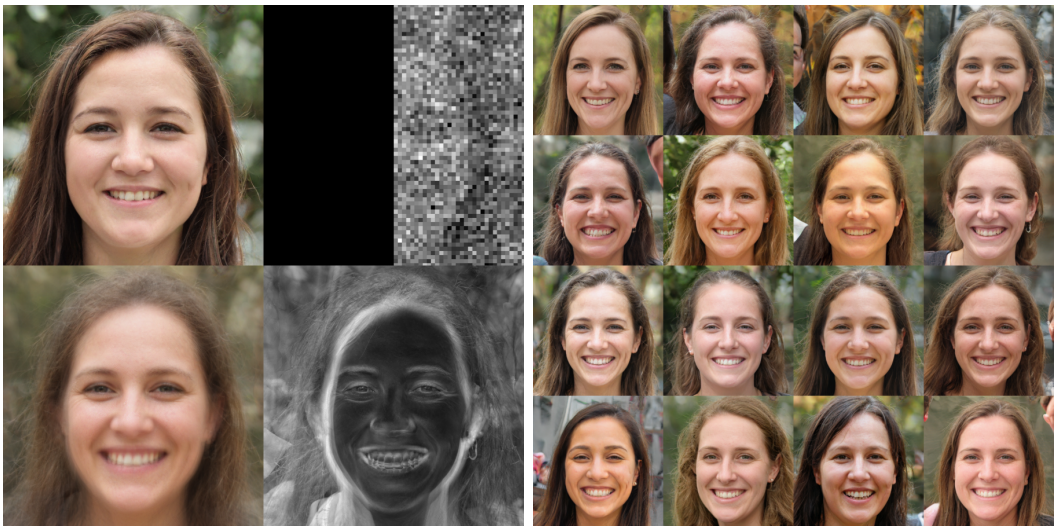


Figure 6.4: Setup and results (left panel) for the face reconstruction problem with ground truth (top left), masked and corrupted data (top right), the mean reconstructed image (bottom left), the pixel-wise standard deviation (bottom right), and samples (right panel).

cases exclusively correct samples are achieved in this phase. At the end, the score is significantly above the accuracy, so although all constraints are quite satisfied, the solution itself is not necessarily correct.

How this is possible can be seen in the two examples shown in the left panel of Fig. 6.3. Here samples and their mean from two different runs are shown. The first one correctly identifies the solution, whereas the second one ends up in an almost correct mode. In this, all constraints are satisfied, except that the last digit has no closed circle. Qualitatively all runs that do not show the correct solution end up in this mode. Note that the last digit tends to avoid closing the circle to comply with all constraints. This is possible because two distinct networks are used to check for the constraints and these samples correspond to a fringe case in which both are satisfied. To avoid this behavior, additional information can be added to the likelihood. In fact, our third constraint, i.e. the last digit not being a seven, is mathematically not necessary. Its purpose is to remove a similar local minimum to make it easier to find the correct solution. Based on this, a heuristic could be developed to iteratively add insights on wrong solutions to the problem to come up with the correct answer.

6.6.3 Reconstructing Faces

In the last example we reconstruct the image of a face from degraded, noisy, and incomplete data, making use of the additional information of age and gender. We compare it to a reconstruction using only the image data. As generative model of face images $x = G(\xi)$, the stylegan network trained on the Flickr-Faces-HQ data set is used [65]. It generates photo-realistic images of faces in a resolution of 1024×1024 pixels from 512 latent parameters. The generative model contains 23 million trained weights. Age and gender estimates are obtained via two networks, $F_a(x)$ and $F_g(x)$, with the same ResNet-50 architecture [54], trained on the IMDB-WIKI dataset [117, 118]. Each

of them contains 10 million weights. A ground truth is drawn from the generator and degraded in several ways. The three color channels are added up to generate a gray-scale image. Then, the resolution is reduced to 64×64 pixels via coarse-graining, followed by masking the left part. These steps are summarized in the linear degradation operator $F_I(x)$. Finally, Gaussian white noise with unit variance is added, providing image data y_I and noise covariance $N_I = \mathbb{1}$. We obtain age and gender estimates of the ground truth by applying the corresponding classifier. To account for different input and output shapes, a re-scaling from 1024×1024 to 224×224 pixels links both types of networks.

All likelihood terms contain the generator applied to the respective operator or network. The data from the degraded image enters via a Gaussian likelihood. For the age prediction we calculate the weighted average provided by the classification probabilities, resulting in a continuous estimate. On this we also impose a Gaussian likelihood, centered around the true age y_a and assuming a standard deviation of one year, i.e. $N_a = 1$. The gender is enforced via a categorical likelihood with data y_g in favor of the respective category and we use an $\alpha = 10$. The posterior is then proportional to

$$\mathcal{P}(\xi|y_I, y_a, y_g) \propto \mathcal{N}(y_I|F_I(G(\xi)), N_I)\, \mathcal{N}(y_a|F_a(G(\xi)), N_a)\, \mathcal{C}^\alpha(y_g|F_g(G(\xi)))\, \mathcal{N}(\xi|0, \mathbb{1}) \,. \tag{6.8}$$

According to the networks, the ground truth is 33 years old and female. The posterior distribution is approximated using MGVI [76] instead of the previous mean-field approach to achieve faster convergence. We perform 15 iterations with five pairs of antithetic samples and 30 natural gradient steps. In the last iteration we increase to 20 samples to use in further analysis. MGVI is an iterative procedure, not directly optimizing an ELBO. We determine convergence through vanishing changes between iterations. The used hardware in this example is a Intel Xeon CPU E5-2680 with 2.4GHz together with a NVIDIA GeForce GTX 1080ti. The run time for the full problem is about 20 hours.

The setup and mean result with variance are shown in the Fig. 6.4, together with a set of representative samples. The mean of sample images bares striking similarity to the ground truth, including facial expression, overall position, and the outdoor setting in the background. The pixel-wise standard deviation shows high certainty in the central parts of the face, whereas smaller-scale features such as hairstyle and background-details are washed out. The samples appear homogeneous in style, age, and gender. In comparison to that, samples for using only the image data are shown in the left panel of Fig. 6.5. These illustrate the remaining information in the degraded image. It seems there is an outdoor setting and that the most likely female person smiles. The age seems not well-constrained, as the visual spread is far larger than that of the previous samples. This is plausible, as age is mostly associated to small-scale features, which are removed through the degradation. Including the additional information of age and gender allows to reduce the variance in these directions.

In the right panel of Fig. 6.5 the evolution of the perceived age, as well as the RMS error of the image to the ground truth during the optimization is shown. We use here iterations instead of time, as evaluating the age and gender networks is comparably slow. Without the additional information, the mean age is roughly correct throughout the optimization, but the variance is large. Given an age, the samples

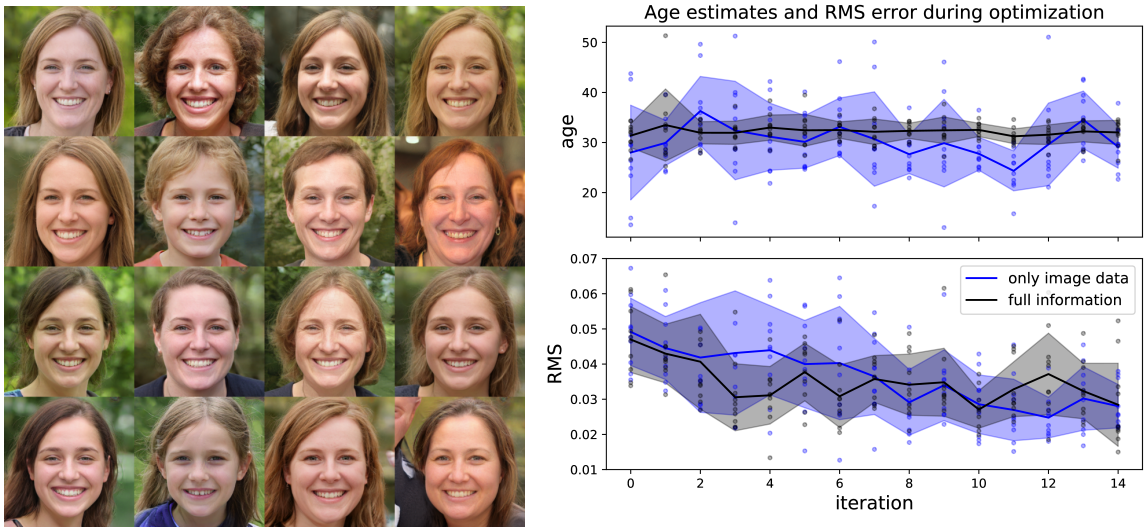


Figure 6.5: Face samples only informed by the image data without age and gender (left panel). Age estimates and RMS of the samples during the optimization for both cases (right panel).

are strongly concentrated around it. The RMS error to the ground truth decreases in both cases similarly, but this is unsurprising, as image data is available to both. Standard deviations are large due to the high noise level.

6.7 Conclusion

We demonstrated how to impose non-trivial constraints represented by deep neural networks to complex systems described through deep generative models. This provides an alternative path for conditional generators. Combining several constraints, we can build a collection of neural networks that jointly solve tasks through reasoning, quantifying their uncertainty. The approach is applicable to state-of-the-art architectures and high-dimensional posterior distributions. Knowledge on high-level concepts can be included to support reconstructions with conventional measurement data.

One could envision large-scale reasoning systems that can flexibly answer a large variety of complex questions by assembling appropriate modules from a library of trained networks. Such systems can also incorporate newly arriving information and support continual learning. The optimal configuration, architectures, and inference schemes for the reasoning need to be identified by future research.

Furthermore, our approach poses the question whether reasoning in human/natural intelligence works via similar processes, the on the fly connection of generative and discriminative networks. As backwards deductions are computational expensive, repeatedly occurring inference tasks are better written into forward models. This entirely depends on knowledge that is already stored within networks and does not require further external input. Does this happen when we dream?

Broader Impact

The approach presented in this paper addresses the fundamental problem of combining complex information from different sources to come up with novel insights through reasoning. This might have consequences we not yet envision. The approach presented could allow to build systems that flexibly derive conclusions for changing tasks. Therefore, it could be a step towards more generic artificial intelligence, with all the benefits and dangers this implies.

Besides the possibility of intentionally malicious use, unintended biases enter through the trained networks and will influence the reasoning. Also the reasoning itself can be wrong due to insufficient exploration of the posterior distribution.

The intrinsic duality of this technology requires more than the usual caution from all sides.

7 Noisy ICA of Auto-Correlated Components

This chapter is additionally a journal publication in Physical Review E [74]. My contribution includes the development, implementation and testing of the idea and all examples. I also wrote the contents. Torsten Enßlin was involved in all discussions and provided valuable feedback on the entire manuscript. All authors read, commented, and approved the final manuscript.

7.1 Abstract

We present a new method for the separation of superimposed, independent, auto-correlated components from noisy multi-channel measurement. The presented method simultaneously reconstructs and separates the components, taking all channels into account and thereby increases the effective signal-to-noise ratio considerably, allowing separations even in the high noise regime. Characteristics of the measurement instruments can be included, allowing for application in complex measurement situations. Independent posterior samples can be provided, permitting error estimates on all desired quantities. Using the concept of information field theory, the algorithm is not restricted to any dimensionality of the underlying space or discretization scheme thereof.

7.2 Introduction

The separation of independent sources in multi-channel measurements is a fundamental challenge in a large variety of different contexts in the fields of science and technology. Large interest in such methods comes from bio-medicine, namely neural-science to investigate brain activities [96], but also in the analysis of financial time series [71] or for the separation of astrophysical components in our universe [27], to name a few.

Mainly two distinct approaches to component separation exist, namely Principle Component Analysis (PCA) and Independent Component Analysis (ICA).

PCA performs a linear transformation of the data to obtain mutually uncorrelated, orthogonal directions, which one calls the principle components. For different principle components s_1 and s_2 their covariance vanishes if averaged over the data set:

$$\langle s_1 s_2 \rangle - \langle s_1 \rangle \langle s_2 \rangle = 0 \tag{7.1}$$

PCA is very useful in situations the data can be described by orthogonal processes. However, this does not imply independence, therefore higher order correlations may

not vanish [60]. The number of principle components one obtains depends on the dimension of the involved data spaces. Some of these components are due to processes generating the data, others might just be due to noise. Drawing a line between those classes of components requires careful consideration of the context the data was obtained in.

ICA speaks of independent components s_1 and s_2 if and only if their probability distributions factorize

$$\mathcal{P}(s_1, s_2) = \mathcal{P}(s_1)\mathcal{P}(s_2) \tag{7.2}$$

ICA Algorithms try to estimate independent components by maximizing some measure of independence. Several such measures are used, such as kurtosis, negentropy, or mutual information, to name a view. These all rely on the non-Gaussian statistics of the components. A mixture of Gaussian components is still Gaussian and does not have the non-orthogonal, relevant directions used in traditional ICA. Therefore, it us often assumed that non-Gaussianity is a prerequisite of ICA. However, the exploitation of auto-correlations in a temporal or spatial domain breaks this Gaussian symmetry and allows identification of the components [131].

An example for a PCA method which can be used in a rather similar setting to the one discussed in this work is the multivariate singular spectrum analysis (MSSA)[47]. It can also be used in noisy multi-channel measurement situations, taking auto-correlations into account. This is done by extending the original channel measurements with a number of time-delayed versions of the vectors. Then one calculates the correlation matrix of all possible channels and time delays. Diagonalizing this matrix leads to the orthogonal principle components, incorporating temporal correlations via the time delays. The most relevant principle components can then be used to describe main features of the data, allowing to analyze dynamical properties of the underlying system.

We, however, want to identify the truly independent components in the data, as characterized by Eq. 7.2. For this goal a PCA method based on the weaker criteria of Eq. 7.1 is a suboptimal approach.

On the side of Independent Component Analysis we have a large variety of widely used algorithms, the more popular ones include FastICA [95] and JADE [26], which rely on the above mentioned independence measures in noise free environments. The often inherent temporal or spatial correlation of the individual components is also not used. An algorithm which uses them is AMUSE [131] which exploits time structure in a noise-free scenario.

A problem for ICA methods is often the presence of measurement noise. The noise prohibits a unique recovery of the individual components and demands for a probabilistic describtion of the problem. Several approaches have been made to solve this problem by using maximum likelihood methods [59] or Gaussian mixtures [99]. In essence this method will follow a similar path.

The general advice in the literature so far, however, is to first de-noise the measurement and then to treat the results as noiseless, processing then with suitbale ICA methods [61]. This approach severely suffers in the high noise regime as it is limited to the signal-to-noise ratio of the individual measurements.

The method we want to present combines the concept of auto-correlation with noisy measurements and thereby overcomes this restriction by reconstructing and separat-

ing the components simultaneously, combining the information across all measured channels and thereby vastly increases the effective signal-to-noise ratio while taking spatial or temporal correlations of the individual components into account. Using this method we can improve the result by adding additional channels and satisfying results are obtained even in high noise environments.

We achieve this by following the Bayesian framework to consistently include auto-correlations to a posterior estimate on the components. The posterior, however, is not accessible analytically and the maximum posterior estimate is insufficient for this problem. We will therefore present an approximation to the true posterior which is capable of capturing its essential features. We will use the Kullback-Leibler divergence to optimally estimate the model parameters in an information theoretical context.

Furthermore we will formulate the components as physical fields without the requirement of specifying any discretization. This allows us to use the language of information field theory (IFT) [37] to develop an abstract algorithm free of any limitations to be used in a specific grid or on a specific number of dimension.

IFT is information theory for fields, generalizing the concept of probability distributions of functions over continuous spaces. In this framework we can formulate a prior distribution encoding the auto-correlation of the components.

Fist we describe the generic problem of noisy independent component analysis. In the next section we formulate auto-correlations in continuous spaces and how to include them in our model. Ways how to approximate the model in a feasible fashion are discussed in Sec. 7.5. In order to infer all parameters we have to draw samples from the approximate posterior. We describe a procedure how to obtain such samples. After briefly stating the full algorithm, we discuss its convergence and demonstrate its performance on two numerical examples showcasing different measurement situations.

7.3 Noisy ICA

The noisy ICA [29] describes the situation of multiple measurements of the same components in different mixtures in the presence of noise. Each individual measurement i at some position or time x results in data $d_{i,x}$ which has a noise contribution $n_{i,x}$, as well as a linear combination M_{ij} of all components $s_{j,x}$ which also have some spatial or temporal structure. The data equation for this is process is given by

$$d_{i,x} = M_{ij} s_{j,x} + n_{i,x} \ . \tag{7.3}$$

Here we use the summation convention over multiple indices. The mixture M_{ij} acts on all positions equally and therefore does not depend on a position index. We can simplify the notation of the equation above by simply dropping the position index, interpreting those quantities as vectors. What remains are the measurement and component indices.

$$d_i = M_{ij} s_j + n_i \tag{7.4}$$

We can go even further by introducing the multi-measurement vector d as a vector of vectors, containing the individual measurements, and noise n, as well as the multi-component vector s, consisting of all components. Then one can use the usual matrix

multiplication with the mixture M to end up with the index-free formulation of this equation as

$$d = Ms + n\,. \tag{7.5}$$

We want to modify this expression in two ways. The first one is to describe the components not as vectors, but as fields. On the one hand true components usually should resemble some physical reality, which is not limited to any discretization and therefore best described by a continuous field, therefore

$$s_{j,x} \to s_j(x)\ . \tag{7.6}$$

On the other hand our data d can never be a continuous field with infinite resolution, as this would correspond to an infinite amount of information. It is therefore necessary to introduce a description of the measurement process, where some kind of instrument probes the physical reality in form of the mixed continuous components. In general this instrument is a linear operator with a continuous physical domain and discrete target, the data space. Including this response operator R, the data equation becomes

$$d_{i,X} = \int dx\, R_i(X,x) M_{ij} s_j(x) + n_{i,X}\ . \tag{7.7}$$

The capital letter X represents the discrete positions of the data, whereas x is the continuous position. We can again drop all indices and state the equation above in operator notation

$$d = RMs + n\ . \tag{7.8}$$

We have now decoupled the domains of the data from the components. We can also not represent components with an infinite resolution, once we want to do numerical calculations we have to somehow specify a discretization, but introducing the response operator allows us to choose representations completely independent from the data and the measurement process. The response operator also allows us to consider any linear measurements, using different instruments for the individual channels. One can easily include masking operations, convolutions, transformations or any other linear instrument specific characteristics in a consistent way.

We will now derive the likelihood of this data model. The noise n will be assumed to be Gaussian with known covariance N and vanishing mean in the data domain. We describe it as

$$\mathcal{P}(n) = \mathcal{G}(n,N) = \frac{1}{|2\pi N|^{\frac{1}{2}}} e^{-\frac{1}{2} n^\dagger N^{-1} n}\ . \tag{7.9}$$

The expression $n^\dagger$ is the complex conjugated, transposed noise vector. This leads to a scalar in the exponent via matrix multiplication. Using this data equation, we can derive the likelihood of the data d, given components s, mixture M and noise realization n. This is a delta distribution as the data is fully determined by the given quantities.

$$\mathcal{P}(d|s,M,n) = \delta(d - (RMs + n)) \tag{7.10}$$

However, the realization of the noise is not of interest and we will marginalize it out using the Gaussian noise model given in Eq. 7.9.

$$\mathcal{P}(d|s,M) = \int \mathcal{D}n\, \delta(d-(RMs+n))\mathcal{G}(n,N) \tag{7.11}$$

$$= \mathcal{G}(d-RMs,N) \tag{7.12}$$

Taking the negative logarithm provides us with the information Hamiltonian[1] of the likelihood, also called negative log-likelihood.

$$\begin{aligned}\mathcal{H}(d|s,M) \equiv & -\ln\left[\mathcal{P}(d|s,M)\right] \\ &= \frac{1}{2}(d-RMs)^\dagger N^{-1}(d-RMs) \\ &\quad + \frac{1}{2}\ln|2\pi N| \end{aligned} \tag{7.17}$$

7.4 Auto-Correlation

The components we want to separate exhibit auto-correlation and we want to exploit this essential property. A component $s_i(x)$ has some value at each location in its continuous domain. We define the scalar product for two fields as $j^\dagger s \equiv \int dx j^*(x)s(x)$, where $j^*(x)$ expresses the complex conjugate of the field j at position x. We can express the two-point auto-correlation as

$$S_i(x,x') \equiv \langle s_i(x)s_i^*(x')\rangle_{\mathcal{P}(s_i)}\ , \tag{7.18}$$

which is a linear operator encoding the internal correlation of the component s_i. Assuming statistical homogeneity, the correlation between two locations $S_i(x,x')$ only depends on their position relative to each other.

$$S_i(x,x') = S_i(x-x') \tag{7.19}$$

Furthermore, we can now apply the Wiener-Khinchin theorem [68] and identify the eigenbasis of the correlation with the associated harmonic domain, which for flat

[1] The information Hamiltonian emerges from the analogy (or equivalence) of information theory to statistical physics:

$$\mathcal{P}(s|d) = \frac{\mathcal{P}(s,d)}{\mathcal{P}(d)} \equiv \frac{e^{-\mathcal{H}(s,d)}}{\mathcal{Z}(d)}, \text{ with information Hamiltonian} \tag{7.13}$$

$$\mathcal{H}(s,d) = -\ln\mathcal{P}(s,d) \text{and partition function} \tag{7.14}$$

$$\mathcal{Z}(d) = \int Ds\, e^{-\mathcal{H}(s,d)} \tag{7.15}$$

$\mathcal{H}(s,d)$ therefore contains all available information on the signal s and is often a more practical object perform do calculations with than the equivalent probability distributions, as information Hamiltonians are additive:

$$\mathcal{H}(s,d) = \mathcal{H}(d|s) + \mathcal{H}(s) \tag{7.16}$$

spaces corresponds to the Fourier basis. This is convenient for the implementation of the algorithm because it allows us to apply the correlation operator in Fourier space, where it is just a diagonal operation and efficient implementations for the Fourier transformation of the components are available as well. This approach stays feasible even for high resolutions of the components, as the representation of the covariance scales roughly linearly in Fourier space, but quadratically in position space.

For components with correlations in more than one dimension, it might also be advantageous to assume statistical isotropy. With this, the correlation only depends on the absolute value of the distance between two points. We can then express the correlation structure by a one-dimensional power spectrum.

In this paper we assume the correlation structure of a component to be known. In principle it could also be inferred from the data with critical filtering [103]. The idea of critical filtering is to parametrize the power spectrum and additionally infer its parameters. This allows us to separate auto-correlated components without knowing the correlation structure beforehand. Critical filtering has been successfully applied in multiple applications [24, 64, 125] and can be included straightforwardly in this model. In order to keep the model simple we choose not to discuss this case in detail here.

We use the known correlation structure S_i to construct a prior distribution of the components s_i, informing the algorithm about the auto-correlation. The least informative prior with this property will be a Gaussian prior with vanishing mean and covariance S_i

$$\mathcal{P}(s_i) = \mathcal{G}(s_i, S_i) \,. \tag{7.20}$$

Conceptually this is a Gaussian distribution over the continuous field s_i. In any numerical application we have to represent the field in a discretized way and this distribution becomes a regular multivariate Gaussian distribution again. Assuming independence of the individual components, the prior distributions factorize and we write

$$\mathcal{P}(s) = \prod_i \mathcal{G}(s_i, S_i) \tag{7.21}$$

$$\equiv \mathcal{G}(s, S) \,. \tag{7.22}$$

The product of Gaussian distributions can be written in the compact form of a combined Gaussian over the multi-component vector s with block-diagonal correlation structure expressing the independence of the different components of each other, i.e. $\langle s_i(x) s_j(x') \rangle_{\mathcal{P}(s)} = 0$ for $i \neq j$. The prior independence actually implements the underlying assumption of any ICA method as stated in Eq. 7.2.

It is worth emphasizing that this way of formulating the correlation structure allows us to apply the resulting algorithm regardless of the dimension. At the end we will demonstrate one dimensional cases for illustration purposes, but without any changes the algorithm generalizes to two, three and n-dimensional situations. Even correlations on curved spaces such as on a sphere can be considered by replacing the Fourier basis with the corresponding harmonic basis.

The information Hamiltonian of this prior distribution is given by

$$\mathcal{H}(s) = \frac{1}{2} s^\dagger S^{-1} s + \frac{1}{2} \ln |2\pi S| \,. \tag{7.23}$$

We now constructed a likelihood from our data model and a prior distribution over the components, encoding their auto-correlation. Using Bayes theorem we can derive the posterior distribution over the components s and their mixture M via

$$\mathcal{P}(s, M|d) = \frac{\mathcal{P}(d|s, M)\mathcal{P}(s, M)}{\mathcal{P}(d)} \,. \tag{7.24}$$

We did not discuss any prior distribution over the mixture M as we do not want to restrict it in any way. Any problem-specific insights about the mixture should be expressed right here. The prior distributions can in our case be written as

$$\mathcal{P}(s, M) \propto \mathcal{P}(s) = \mathcal{G}(s, S) \tag{7.25}$$

and thereby implicitly assuming a flat and independent priors on the entries of M. The evaluation of $\mathcal{P}(d)$ is not feasible as it involves the integration over both, the mixture and signal of the joint probability distribution. Therefore we have to think of approximative approaches. First we state the posterior information Hamiltonian $\mathcal{H}(s, M|d) = -\ln(\mathcal{P}(s, M|d))$ without any component or mixture independent terms.

$$\begin{aligned}\mathcal{H}(s, M|d) =& \frac{1}{2} s^\dagger M^\dagger R^\dagger N^{-1} R M s - s^\dagger M^\dagger R^\dagger N^{-1} d \\ &+ \frac{1}{2} s^\dagger S^{-1} s + \text{const}(d).\end{aligned} \tag{7.26}$$

7.5 Approximating the Posterior

A typical approach to a problem like this is to take the most likely posterior value as an estimate of the parameters. This is achieved by minimizing the information Hamiltonian above. It can be interpreted as an approximation of the posterior distribution with delta distributions peaked at the most informative position in the sense of minimal Kullback-Leibler (KL) divergence [86] between true and approximated posterior. For this the latter can be written as

$$\widetilde{\mathcal{P}}_{\text{MAP}}(s, M|d) = \delta(s - s_{\text{MAP}})\delta(M - M_{\text{MAP}}) \,. \tag{7.27}$$

This approximation turned out to be insufficient for a meaningful separation of the components as we will illustrate in Sect. 7.9. Iterating the minimization with respect to the components and the mixture we do not obtain satisfying results. The maximum posterior estimate is known to over-fit noise features. This has severe consequences in this component separation as it relies on iterative minimization of the Hamiltonian with respect to one of the parameters. In each step we over-fit which affects the consecutive minimization. In this way we accumulate errors in the parameters, leading to unrecognizable, strongly correlated components. During the minimization the MAP algorithm approaches reasonable component separations but it does not converge to those and continues to accumulate errors, converging somewhere else. This behavior can be seen in Figure 7.1, showing the deviation of current estimates to the true components for the MAP case, as well as the algorithm discussed in the following.

Our strategy to solve this problem is to choose a richer model to approximate the posterior distribution which is capable to capture uncertainty features and reduce

over-fitting. Instead of using a delta-distribution to describe the posterior components, we choose a variational approach using a Gaussian distribution whose parameters have to be estimated. For the posterior mixture we stay with the initial description of the point estimate as this turns out to be sufficient for many applications. We therefore approximate the true posterior with a distribution of the form

$$\widetilde{\mathcal{P}}(s, M|d) = \mathcal{G}(s - m, D)\delta(M - M_*) \,. \tag{7.28}$$

In this approximation we describe our posterior knowledge about the mixture M with the point-estimate M_* and about the components s by a Gaussian distribution with mean m and covariance D . We will use m as the estimate of our posterior components and the covariance D describes the uncertainty structure of this estimate. Compared to the prior covariance S, the posterior covariance does not have to be diagonal in the harmonic domain, as the likelihood typically breaks the homogeneity.

The main problem of this approximation is the point-estimate of the posterior mixture M. With this we assume perfect knowledge about the mixture with absolute certainty. This is certainly not justified due to the probabilistic nature of the problem, but this is true for every point-estimate in any context. This approximation also affects the posterior covariance of the components D which will contain the mixture. As we assume no uncertainty in it, we will not consider any errors in the mixture and therefore underestimate the true uncertainty of the components. In the low noise regime this effect is negligible, it will become larger for low signal-to-noise ratio, as we will see in the numerical examples. This model, however, seems to perform reasonably well in relatively high noise regimes, but one has to take the error estimates with caution and keep in mind that those will be underestimated. One could easily think of a more complex model performing even better and more accurate in the high noise case. For example also approximating the mixture with a Gaussian distribution or using one large Gaussian distribution, also accounting for cross-correlations between components and the mixture. Those models come with the cost of dramatically increased analytical and numerical complexity. As no analytic form of the posterior is available, the best solution possible can be obtained from sampling the posterior, which can can become computationally extremely expensive, as the dimensionality of the problem scales with the resolution of the components. We choose the approximation given in Eq. 7.28 as it should capture the relevant quantities while being as simple as possible.

In order to estimate the parameters of the distribution in Eq. 7.28, we have to minimize its KL divergence to the initial posterior. The divergence is defined as

$$\mathrm{KL}\left[\widetilde{\mathcal{P}}(s, M|d)||\mathcal{P}(s, M|d)\right] \equiv \tag{7.29}$$

$$\equiv \int \mathcal{D}s\, \mathcal{D}M\, \widetilde{\mathcal{P}}(s, M|d) \ln\left[\frac{\mathcal{P}(s, M|d)}{\widetilde{\mathcal{P}}(s, M|d)}\right] \tag{7.30}$$

$$= \langle \mathcal{H}(s, M_*|d)\rangle_{\mathcal{G}(s-m,D)} - \langle \ln\left[\mathcal{G}(s - m, D)\right]\rangle_{\mathcal{G}(s-m,D)} \,. \tag{7.31}$$

The integration over the mixture just replaces every M by M_*. In order to keep the expressions shorter we will drop from now on the star and will use in all further

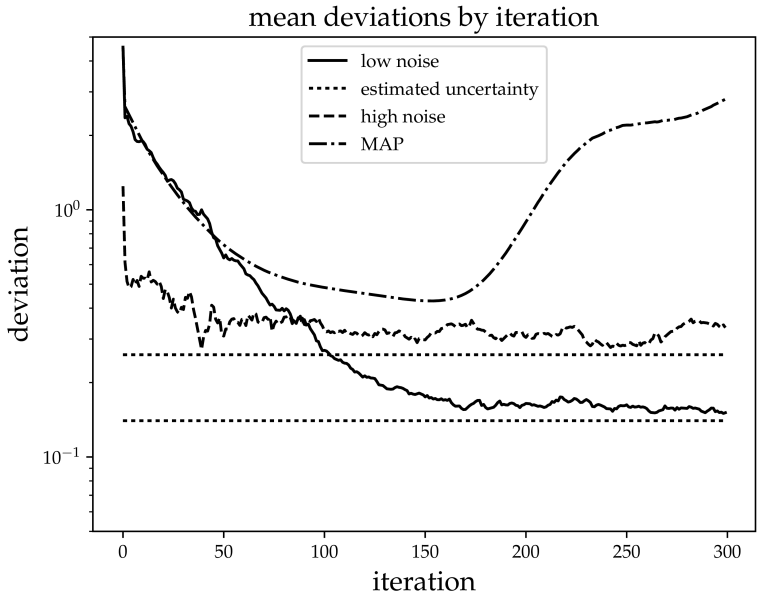


Figure 7.1: The mean deviation of the current estimates from the true components during the minimization for all three example scenarios compared to the estimated uncertainty of the final result.

calculations just the symbol M. We now have to calculate Gaussian expectation values of the total information Hamiltonian. We can perform this calculation with the cyclical property of the trace operation and the identity

$$\langle ss^\dagger \rangle_{\mathcal{G}(s-m,D)} = mm^\dagger + D \,. \tag{7.32}$$

The second expectation value in the KL-divergence corresponds to the entropy of the Gaussian distribution. The analytic expression then reads

$$\begin{aligned} \mathrm{KL} = &\frac{1}{2} m^\dagger M^\dagger R^\dagger N^{-1} R M m + \frac{1}{2} \mathrm{Tr}\left[M^\dagger R^\dagger N^{-1} R M D\right] \\ &- m^\dagger M^\dagger R^\dagger N^{-1} d \\ &+ \frac{1}{2} m^\dagger S^{-1} m + \frac{1}{2} \mathrm{Tr}\left[S^{-1} D\right] \\ &+ \mathrm{Tr}\left[1 + \ln(2\pi D)\right] \,. \end{aligned} \tag{7.33}$$

We have to minimize this expression with respect to all parameters of our approximate distribution, namely m, D and M.

We will start with the posterior component mean m. Comparing the terms of the Hamiltonian in Eq. 7.26 containing s with the ones in the KL containing m we find their analogous structure. Given some mixture M the minimum will be identical. We can solve for the posterior mean by setting the derivative of the KL-divergence with

respect to it to zero:

$$\frac{\delta \mathrm{KL}}{\delta m^\dagger} \stackrel{!}{=} 0 \tag{7.34}$$

$$= -M^\dagger R^\dagger N^{-1} d + M^\dagger R^\dagger N^{-1} R M m + S^{-1} m \tag{7.35}$$

$$\Rightarrow m = \left(M^\dagger R^\dagger N^{-1} R M + S^{-1}\right)^{-1} M^\dagger R^\dagger N^{-1} d \tag{7.36}$$

The structure of the solution looks familiar. In fact it is the Wiener filter solution [135] for a known mixture. We can also solve for the posterior covariance:

$$\frac{\delta \mathrm{KL}}{\delta D} \stackrel{!}{=} 0 \tag{7.37}$$

$$\Rightarrow D = \left(M^\dagger R^\dagger N^{-1} R M + S^{-1}\right)^{-1} \tag{7.38}$$

This also turns out to be the Wiener covariance for known mixture. We can then define the information source j and write the approximate posterior mean in terms of the Wiener filter formula:

$$j \equiv M^\dagger R^\dagger N^{-1} d \tag{7.39}$$

$$m = Dj \tag{7.40}$$

If we know the mixture M a Gaussian with mean m and covariance D would be the exact posterior of the components given the data.

Now we also have to calculate the derivative of the KL-divergence with respect to the entries of the mixture matrix M_{ij} while keeping m and D fixed. In this calculation the trace term

$$\frac{1}{2}\mathrm{Tr}\left[M^\dagger R^\dagger N^{-1} R M D\right] \tag{7.41}$$

in the divergence does not vanish as it contains the mixture M and will gives rise to the required uncertainty corrections, regularizing the mixture and therefore making the algorithm converge. Unfortunately this term is numerically challenging. The trace of an operator can be extracted via operator probing [57, 123]. This involves multiple numerical operator inversions using conjugate gradient method which is computationally expensive.

We will choose another approach which avoids the trace expressions by taking them implicitly into account. For this purpose we still have to solve multiple linear systems, but we found the new approach to be numerically more stable and more general as it can also be applied in cases we do not have explicit expressions.

In order to obtain the analytic expression of the KL-divergence we calculated the expectation value of the information Hamiltonian with respect to the approximate posterior Gaussian distribution which gave rise to the trace terms in the first place. To avoid them we will consider the KL-divergence before performing the averaging over the approximating Gaussian and keep it that way during the derivation of the

gradient. To estimate the resulting expressions we approximate the averaging by replacing it with an average over samples drawn from the distribution $\mathcal{G}(s-m,D)$. All relevant terms in the KL-divergence concerning M read:

$$\begin{aligned}\mathrm{KL}_M = \frac{1}{2}\langle s^\dagger M^\dagger R^\dagger N^{-1} R M s\rangle_{\mathcal{G}(s-m,D)} \\ -\langle d^\dagger N^{-1} R M s\rangle_{\mathcal{G}(s-m,D)}\end{aligned} \tag{7.42}$$

For the minimization with respect to the mixture M we assume the posterior mean m and covariance D to be fixed, so we can calculate the derivative of the expression above with respect to the mixture ignoring the expectation value:

$$\begin{aligned}\frac{\delta \mathrm{KL_M}(s,M|d)}{\delta M_{ij}} = \langle s^\dagger M^\dagger R^\dagger N^{-1} R \mathbb{1}_{ij} s\rangle_{\mathcal{G}(s-m,D)} \\ -\langle d^\dagger N^{-1} R \mathbb{1}_{ij} s\rangle_{\mathcal{G}(s-m,D)}\end{aligned} \tag{7.43}$$

The operator $\mathbb{1}_{ij}$ with $(\mathbb{1}_{ij})_{i'j'} = \delta_{ii'}\delta_{jj'}$ singles out the position of the entry ij. It is of the same shape as the mixture matrix with all entries zero, except for the one at the position ij. Comparing this term to the derivative of the information Hamiltonian

$$\begin{aligned}\frac{\delta \mathrm{H}(s,M|d)}{\delta M_{ij}} = s^\dagger M^\dagger R^\dagger N^{-1} R \mathbb{1}_{ij} s \\ - d^\dagger N^{-1} R \mathbb{1}_{ij} s\end{aligned} \tag{7.44}$$

which is used in the maximum posterior approximation, the main difference to our method becomes apparent. In the maximum posterior approach only a point estimate for the components s is used. Our approach replaces the minimization of the Hamiltonian with the minimization of the mean Hamiltonian under the approximated Gaussian, taking the uncertainty structure of the components into account.

Setting the our mixture gradient to zero allows us to solve for the mixture in a Wiener-filter-like fashion

$$M = \left\langle s^\dagger \mathbb{1}^\dagger R^\dagger N^{-1} R \mathbb{1} s\right\rangle^{-1} \left\langle d^\dagger N^{-1} R \mathbb{1} s\right\rangle \tag{7.45}$$

The first part serves as a Wiener covariance and the second term corresponds to an information source for M.

At some point we have to evaluate all those expectation values numerically to minimize the divergence with respect to the mixture M.

The terms we want to calculate are expectation values of the Gaussian distribution $\mathcal{G}(s-m,D)$, but they will introduce impractical trace terms. Instead we want to approximate it with a set of L samples $\{s^*\}$ distributed according to $\mathcal{G}(s-m,D)$ using the sampling distribution.

$$\mathcal{G}(s-m,D) \approx \frac{1}{L}\sum_{l=0}^{L} \delta(s-s_l^*) \tag{7.46}$$

Using this distribution, the expectation values are replaced with the average over the set of samples. In the next section we will discuss how to obtain those samples from the distribution.

7.6 Approximate Posterior Sampling

Drawing samples from the approximate posterior distribution for the components is challenging, as we do not have direct access to its eigenbasis in which the correlation structure is diagonal. If we had, we could draw independent Gaussian samples with mean zero and variance one in each dimension, weight them with the square root of the eigenvalue to adjust to the correct variance and apply the transformation to position space given by the eigenvectors. At this point the sample has the correct correlation structure and has only be adjusted to the correct mean by adding it.

The main task is therefore to get samples with the correct correlation structure. In the case of our approximate posterior

$$\widetilde{\mathcal{P}}(s|d) = \mathcal{G}(s-m, D), \tag{7.47}$$

we have to find residuals $(s-m)$ which satisfy

$$\langle (s-m)(s-m)^\dagger \rangle_{\mathcal{G}(s-m,D)} = D \ . \tag{7.48}$$

Obviously we do not have access to the true components s. What we do have is a prior belief about them. Its correlation is diagonal in the Fourier domain for each component and we can easily generate a samples from it using the description above.

$$s' \curvearrowleft \mathcal{G}(s, S) \tag{7.49}$$

Those components s' have nothing to do with the true components, except their correlation structure. We now want to find an m' which satisfies

$$\langle (s'-m')(s'-m')^\dagger \rangle_{\mathcal{G}(s'-m',D)} = D \ . \tag{7.50}$$

The posterior covariance is a Wiener Filter covariance described by

$$D^{-1} = M^\dagger R^\dagger N^{-1} R M + S^{-1} \tag{7.51}$$

with given mixture M, instrument response R, noise covariance N and prior signal covariance S. We can reconstruct the quantity m' from the data we would have obtained if s' were the real components. We therefore have to simulate the measurement process of the arbitrary sample s' using our linear data equation

$$d' = RMs' + n' \ . \tag{7.52}$$

We can draw an noise realization n' from the prior noise distribution $\mathcal{G}(n, N)$ which is diagonal in data space. On this mock data we simply perform a Wiener Filter reconstruction

$$m' = Dj' \text{ , with} \tag{7.53}$$

$$j' \equiv \ M^\dagger R^\dagger N^{-1} d' \ . \tag{7.54}$$

This is the numerical costly part of the sampling procedure as it involves a conjugate gradient to solve the system.

However, once we obtained the reconstruction m' of the mock signal s' we can calculate the residual, which follows exactly the correlation structure encoded in D. The components s' are now a sample drawn from the distribution $\mathcal{G}(s'-m', D)$. What we actually want is a sample s^* from $\mathcal{G}(s^* - m, D)$, originating from our true data. The residuals of both distributions have the same statistical properties, thus we can therefore set them equal and solve for s^*.

$$s' - m' \stackrel{!}{=} s^* - m \tag{7.55}$$

$$s^* = s' - m' + m \tag{7.56}$$

Those components s^* now exactly behave according to $\mathcal{G}(s^* - m, D)$ with mean m and covariance D. We can use samples drawn by this procedure to calculate the expectation values we need in the minimization process for the mixture.

Furthermore, we can use the samples to easily estimate arbitrary posterior properties, such as the uncertainty of our component estimate.

Let us briefly summarize this approach for approximate posterior sampling. We start with a sample s' drawn independently from the component prior, use those to set up a mock observation, which provides us with mock data d'. We Wiener filter this data to get the posterior mean m'. The only thing we are interested in from this calculations is the residual $s' - m'$, as it allows us to construct a sample s^* from the mean m of the distribution we are actually interested in. The more samples we draw this way the better the sampling distribution approximates the true distribution. However, we want to use as few samples as possible as their calculation is computationally expensive, not only during the sampling procedure, but also their usage in all further calculations, such as gradient estimations. During the alternating minimization with respect to the mean components m and the mixture M, we have to permanently recalculate the samples as the mean and mixture is constantly changing. We found that it is practical to start with few samples and to increase their number during the inference. Note that the KL divergence is not fully calculated and also only estimated through the samples, therefore this estimate inherits stochastic variations.

7.7 The Algorithm

Now we have all the tools to set up an iterative scheme to minimize the KL divergence in order to infer the parameters of our approximation.

In order to use this algorithm we need knowledge on the characteristic noise behavior encoded in the correlation structure N, as well as on the statistical properties of the individual components described by the prior covariance S. In addition we have to specify the number of components we want to infer.

We will start with a random guess for the mixture M and use it to estimate our initial mean components m and covariance D under the assumption the initial guess of the mixture is correct using the Wiener Filter.

$$D^{-1} = M^\dagger R^\dagger N^{-1} R M + S^{-1} \tag{7.57}$$

$$j = M^\dagger R^\dagger N^{-1} d \tag{7.58}$$

$$m = Dj \tag{7.59}$$

We have now the first estimate of the approximate posterior distribution $\mathcal{G}(s - m, D)$. In order to estimate a new mixture we can draw a set of independent samples $\{s^*\}$ from this distribution using the procedure described in the previous section.

$$\{s^*\} \curvearrowleft \mathcal{G}(s - m, D) \tag{7.60}$$

We use those to replace the Gaussian expectation values with averages over the sampling distribution, which allows us to solve for a new estimate for the mixture, using the Wiener-Filter-like formula:

$$M = \left\langle s^\dagger \mathbb{1}^\dagger R^\dagger N^{-1} R \mathbb{1} s \right\rangle^{-1}_{\{s^*\}} \left\langle d^\dagger N^{-1} R \mathbb{1} s \right\rangle_{\{s^*\}} \tag{7.61}$$

Now we have to take care of the multiplicative degeneracy between between the components and their mixture vector. We therefore normalize each column of the mixture to an L_2-norm of $||M_j^\dagger|| = 1$, multiplying each component mean accordingly by the normalization factor to keep the product Mm unchanged.

If the power-spectrum of the components are unknown, we perform here a critical filter step [103], which we choose not to discuss at this point.

This way we obtain a new estimate for the mixture, which allows us to estimate new component means and covariances, which allows us to draw new samples, which we can use for a new mixture, and so on, until the algorithm converges. We will discuss its converges in the next section.

However, after the algorithm has converged we can use the samples to calculate any posterior quantity of interest involving the components and estimate its uncertainty. One example would be the spatial uncertainty of the component reconstruction by evaluating

$$D_{xx} = \left\langle (s_x - m_x)^2 \right\rangle_{\{s^*\}} . \tag{7.62}$$

7.8 On its Convergence

Each estimate of a new parameter on its own will reduce the remaining KL divergence between our approximated posterior and the true posterior, at least stochastically. The stochasticity is due to the noise introduced by the sampling and can be reduced by using more samples, for the price of high computational cost. Let us briefly discuss the symmetries, structure and minima of the Kullback-Leibler divergence as it is stated in Eq. 7.33. We start with two likelihood contributions

$$\mathrm{KL} \mathrel{\widehat{=}} \frac{1}{2} m^\dagger M^\dagger R^\dagger N^{-1} R M m - m^\dagger M^\dagger R^\dagger N^{-1} d. \tag{7.63}$$

Here, we have a unique minimum for the mixed components Mm due to the quadratic structure in the case $R^\dagger N^{-1} R$ is a full rank operator, otherwise its null space is unconstrained.

In addition to this, individual mixtures M and component means m, the terms above exhibit two symmetries, as we can multiply the mixtures for each components with arbitrary factors while dividing the corresponding components by according factors. This introduces a submanifold of minimal energy.

Finally, we can just interchange the components while also swapping the entries of the mixing matrix. Depending on the number of components, we get additional $c!$ times as many minima, with c being the number of components.

The only other terms concerning the mixture and components are their other likelihood contribution and the component prior term

$$\mathrm{KL} \mathrel{\widehat{=}} \frac{1}{2}\mathrm{Tr}\left[M^\dagger R^\dagger N^{-1} R M D\right] + \frac{1}{2} m^\dagger S^{-1} m \ , \tag{7.64}$$

which are both quadratic terms in m and M, respectively, with positive sign and therefore do not introduce additional minima, but eliminate some degeneracy. First of all, these terms constrain the null space degeneracy to one single point. The multiplicative degeneracy between M and m is also broken as both quadratic terms regularize like L_2 norms. What remains is the degeneracy of the multiplication of M and m with -1 for each component, allowing for 2^c possibilities. Thus, instead of entire submanifolds of optimal solutions we end up with a total of $2^c c!$ minima of the KL divergence with respect to M and m.

In the case that the prior covariances for the individual components in S are not identical, the interchange symmetry is broken and all those minima do not have the same divergence anymore. Therefore, using a gradient descent method we do not necessarily end up in a global minimum. This can be solved by discrete optimization steps, trying all possible permutations of the mixture and components and picking the one with smallest KL divergence.

This problem also vanishes if one also infers the prior correlation structure as the prior then adapts to the chosen permutation, leading to a global minimum for sure.

We have seen that in the case of the same prior correlation structures for all components all minima of the divergence are global minima and we therefore will converge to an optimal solution irrespective of the starting position.

The speed of convergence, however, is hard to estimate as we rely on the iteration of consecutive minimizations of our parameters. Each individual minimization converges rather quickly, depending on the condition numbers of the matrices involved, as we invert them by the conjugate gradient method. The total convergence rate should depend on the correlation between the component means m and mixtures M in the KL divergence. The less they are correlated, the faster the individual parameters should reach a minimum. Strong correlations, however, do not allow for large steps, therefore being slower.

Practically the computational effort highly depends on the choice of various quantities. The algorithm is divided into two distinct minimizations for the mean components m and mixture M of different dimensionality, which is the main source of computational cost. The dimensionality of the component part scales linearly with the number of components and their resolution, at least in the one-dimensional case. For higher dimensional components the resolution scales accordingly. The costly part in this minimization is the numerical inversion of an implicit operator in order to solve a Wiener Filter problem. The minimization with respect to the mixture is rather cheap, having the dimension of number of components times number of data channels. Drawing one posterior samples, however requires a Wiener Filter of the complexity of the first part. We therefore want to keep the number of samples as low

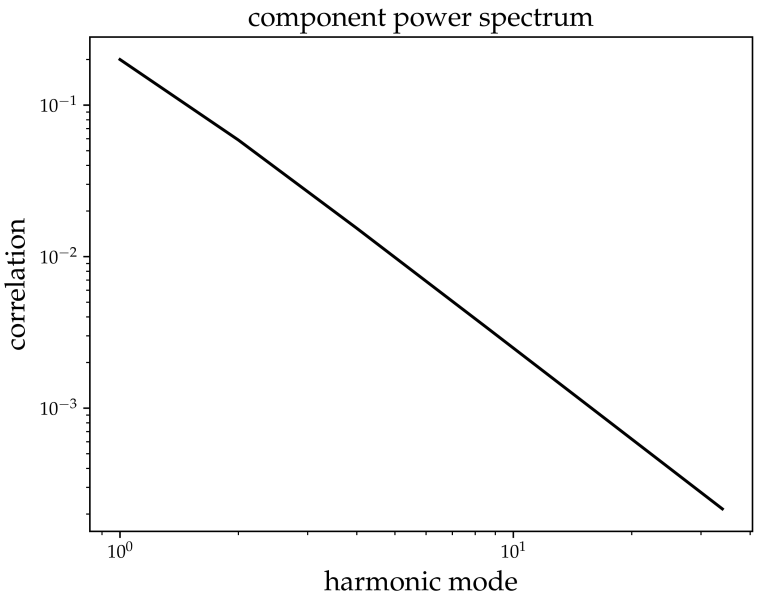


Figure 7.2: The correlation structure of both components in Fourier space in double logarithmic representation.

as possible, at least at the beginning of the inference. We can increase the number of samples towards the end, reducing the statistical sampling noise.

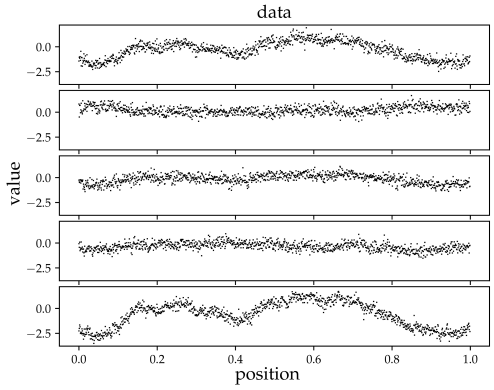
The entire algorithm consists of a large number of consecutive minimizations. The accuracy to which each of them is performed greatly effects the overall performance. We want to avoid unnecessary accuracy wherever possible, as all parameters are changing constantly and for the mixture the KL divergence is only a statistical estimate with uncertainties itself. Therefore we would waste computation if we aim for high accuracy initially. Towards the end, as the number of samples increases, one might also increase the accuracy. How to optimally steer this is rather difficult and currently requires case by case optimization.

7.9 Numerical Examples

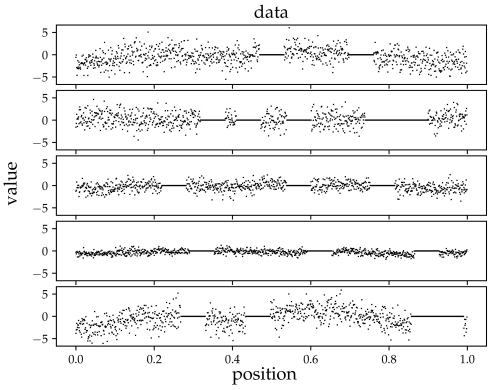
We implemented the algorithm as outlined above in Python using the package NIFTy (Numerical Information Field Theory) [124], allowing a coordinate free implementation. For our two numerical examples we will use synthetic data generated according to the model. The first one will describe a rather simple case with moderate, but present noise.

In the second example we will challenge the algorithm with a more realistic measurement. We will model randomly failing measurement sensors by masking areas of the data set. In addition each sensor will exhibit a individual noise covariance of significantly increased strength. For the comparison we will use the same component realizations and mixture as used before.

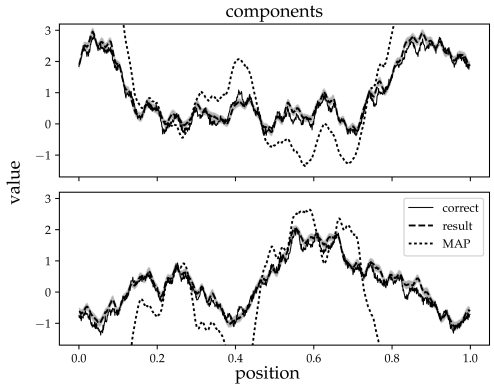
In our examples, we measure five different mixtures of two independent compo-



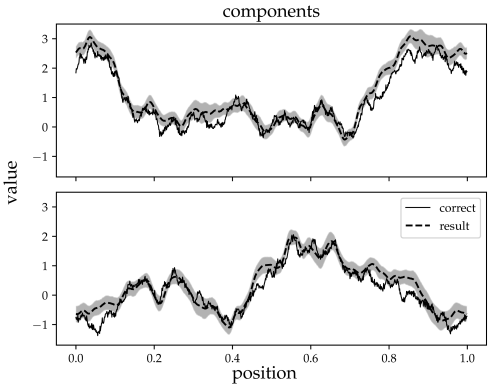
(a) Data of the first scenario in five channels from noisy measurements of two linearly mixed components.



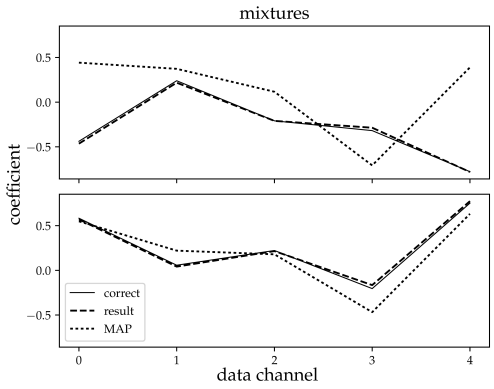
(b) Measurements with failing sensors and varying noise levels in scenario two. Note the changed scales.



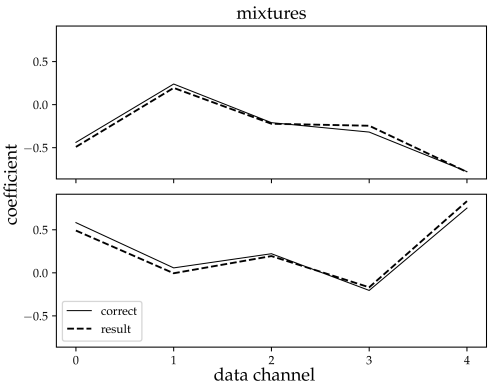
(c) Correct, reconstructed and maximum posterior components with error estimate in scenario one of the data shown in Fig. 7.3a



(d) Reconstruction of the independent components using the noisy data set of scenario two shown in Fig. 7.3b with error estimates.



(e) Correct, reconstructed and maximum posterior mixtures in scenario one.



(f) Correct, reconstructed and maximum posterior mixtures in scenario one.

Figure 7.3: The setups and results for a high- and low-noise scenario.

nents. Each channel consists of 1024 data points probing equally spaced locations of the unit interval over which our periodic components live. In the first example the measurements are corrupted by noticeable noise of zero mean and diagonal covariance of $\sigma_n^2 = 0.1$. The response operator R in this case is just the identity operator $R_{xy} = \delta(x - y)$. The data are illustrated in Figure 7.3a. Both components are generated by drawing a realization from the prior distribution $\mathcal{P}(s)$ with power spectrum

$$P_c(k) = \frac{1}{4k^2 + 1} \,. \tag{7.65}$$

This describes the spatial correlation by a falling power law in Fourier space[2] which is typical for many physical processes.This function is shown in Figure 7.2. By choosing the same power spectrum for both components we can ignore the problem of multimodality of the probability distributions as all minima are equally global minima. The values of the mixture entries are drawn independently from a Gaussian distribution with vanishing mean and unit variance. Afterwards the entries corresponding to one component are normalized to fix the multiplicative degeneracy between mixture and component.

The number of samples s^* used to estimate the mixture was initially one per iteration and was increased to 25 at the end of the reconstruction. We iterated the algorithm 300 times, after which the reconstruction converged. The results of the analysis are shown in Figure 7.3c and 7.3e. The reconstructions are corrected for the degeneracy of the signs and are compared with the true corresponding components and mixtures while keeping the product Mm constant. We can clearly recover the morphological structure of the distinct components with high accuracy. The one sigma uncertainty contours estimated as $\sqrt{D_{xx}}$ quantify the estimated error reasonably well. The structure of the mixture is recovered, only small deviations from the true mixture are present. We can even recover relatively small structures of the components as the algorithm uses the combined information of all channels simultaneously, increasing the effective signal-to-noise ratio, leading to higher resolutions. Denoising each individual channel first and then applying a noise free ICA method cannot reach that resolution as it is limited to the signal to noise ratio of the individual channels.

We also show the result of maximizing the posterior with respect to s and M for this scenario. The initial components are not recovered and the suggested solutions are highly anti-correlated. This demonstrates the necessity of the uncertainty corrections emerging from the presented model, represented by Eq. 7.41 and the averages in Eq. 7.42.

In the second example we used the same setup as before with the same two components and five data channels. We only modified our measurement instrument to resemble typical properties of true sensors. We randomly masked 22% of the total area by sequences of 64 measurement points each. Additionally we assign each sensor an individual noise covariance. The noise level will be significantly higher in this case, ranging from a factor of two up to 25 times the variance compared to the previous example. The data are shown in Figure 7.3b. By eye it is hard to identify any components, only hints of correlated structures can be recognized. We can encode the

[2]We use here the numerical Fourier convention of $f(k) = \int_0^1 dx\, f(x)\, e^{2\pi i k x}$.

failing sensors in the instrument response operator R as masks and the varying noise in the noise operator N and run exactly the same algorithm as before. The result can be seen in Figure 7.3d and 7.3f, again with corrected signs and compared to the true corresponding components. The morphological structure is recovered despite the significantly more hostile conditions. The overall uncertainty is consequently higher than before and therefore small scales are not as well resolved. Due to the masking we observe modulations in the uncertainty structure. In some parts the uncertainty does not fully cover the deviations from the true components. As we do not take the uncertainty structure of the mixture into account we have probably underestimated the error. In the mixture we also observe larger deviations from the correct mixture, but in general we recover it well.

The convergence behavior for all three examples can be seen in Figure 7.1. It shows the mean deviation of the current estimate m_t of the components from the true components s in each iteration step t, corrected for the degeneracy. We calculated it according to

$$\epsilon_t = \sqrt{\frac{(m_t - s)^\dagger (m_t - s)}{l}} \,, \tag{7.66}$$

where l is the number of sites, given by the resolution of the components. We would expect this quantity not to become smaller than the expected deviations originating from the error estimate of the final result, which therefore sets the lower limit. It is shown as the two horizontal lines for the high and low noise case. During the inference the mean deviation declines towards this limit for both cases, but does not reach it.

This indicates that the error estimate of the result underestimates the error slightly, a finding that is not surprising as we do not take uncertainties of the mixture into account. The higher the noise level the more this effect becomes relevant, whereas in the low noise case it is almost negligible. We can also observe the statistical nature of the minimization due to the sampling in the noisy trajectory. Compared to that the maximum posterior minimization follows a smooth line. In this plot we can also nicely see the divergence of using just maximum posterior. It starts approaching the true components, roughly at the same speed as the KL-approach in the same situation, but then slows down and starts to accumulate errors and clearly diverges, while the other method continues converging.

7.10 Summary

We derived a new method which allows for the separation of independent components from noisy measurements exploiting their auto-correlation. This was done by first describing the measurement process as a linear mixture of component fields which are observed by some linear measurement instruments under additive, Gaussian noise. From this model we derived the likelihood. Assuming homogeneity of the auto-correlation of the components we could express their correlation structure as a diagonal operator in the Fourier basis. From this assumption we derived the least informative prior distribution over the components in form of a Gaussian distribution. No prior assumptions about the mixture entries were made, but such could be added easily.

Using the model likelihood and the component prior, we derived an expression for the posterior probability distribution by applying Bayes theorem.

As this expression was not accessible analytically we approximated it by the product of a Gaussian distribution for the components and a delta distribution for the mixture entries. In order to infer the parameters of the approximation we proposed a scheme to minimize the Kullback-Leibler divergence of this distribution to the true posterior. It involved iterative Wiener filtering of the components and the mixture. For estimating the mixture we considered uncertainty corrections originating from the Gaussian approximation for the component maps. These turned out to be essential for obtaining accurate estimates of the mixture matrices. A joint MAP estimate of fields and mixtures tends to provide incorrect results. In order to evaluate the corrections we outlined an approach how to draw independent samples from the approximate Gaussian posterior distribution.

In two numerical examples we demonstrated the applicability of the derived algorithm. The first case involved moderate noise and recovered the true components and mixtures with high accuracy. The estimated error of the components was reliable. The second example models randomly failing sensors and a significantly higher, varying noise level applied to the same components. The morphology of the mixture and components was recovered here as well, the error was slightly underestimated due to the involved point estimate of the mixture. Overall the algorithm delivered satisfying results and can also be applied in complex measurement situations in the high noise regime.

Acknowledgments

We acknowledge helpful discussions and comments on the manuscript by Martin Dupont, Reimar Leike, Sebastian Hutschenreuter, Natalia Porqueres, Daniel Pumpe, and two anonymous referees.

8 Encoding Prior Knowledge in the Structure of the Likelihood

This chapter is used as a publication currently submitted to the Journal of Machine Learning Research [76]. My contribution includes the development, implementation and testing of the idea and all examples. I also wrote the contents. Torsten Enßlin was involved in all discussions and provided valuable feedback on the entire manuscript. All authors read, commented, and approved the final manuscript.

8.1 Abstract

The inference of deep hierarchical models is problematic due to strong dependencies between the hierarchies. We investigate a specific transformation of the model parameters based on the multivariate distributional transform. This transformation is a special form of the reparametrization trick, flattens the hierarchy and leads to a standard Gaussian prior on all resulting parameters. The transformation also transfers all the prior information into the structure of the likelihood, hereby decoupling the transformed parameters a priori from each other. A variational Gaussian approximation in this standardized space will be excellent in situations of relatively uninformative data. Additionally, the curvature of the log-posterior is well-conditioned in directions that are weakly constrained by the data, allowing for fast inference in such a scenario. In an example we perform the transformation explicitly for Gaussian process regression with a priori unknown correlation structure. Deep models are inferred rapidly in highly and slowly in poorly informed situations. The flat model show exactly the opposite performance pattern. A synthesis of both, the deep and the flat perspective, provides their combined advantages and overcomes the individual limitations, leading to a faster inference.

8.2 Introduction

Hierarchical Bayesian models make it possible to express the complex relations in real systems by combining a priori domain knowledge with data. The prior knowledge is updated by the observed data to obtain information about the system at hand. Such models can exhibit deep hierarchies, relating a large number of conceptually distinct parameters in non-trivial fashions. The inference of the posterior parameters can be extremely problematic, especially for large and complex models due to strong dependencies between the quantities and the resulting numerical stiffness. A way to overcome such limitations is to perform coordinate transformations of the parameters to less interdependent ones. In the context of Hamiltonian Monte Carlo (HMC) techniques it was proposed to perform a linear coordinate transformation [87] to decouple

the parameters, which in the discussed case leads to a white, standard Gaussian prior for the new parameters. This way the numerical performance was increased. Another, more general transformation scheme was proposed by Betancourt and Girolami [15] to flatten the deep hierarchical structure, to decouple the parameters by introducing auxiliary parameters or performing a reparametrization of existing ones. The same kind of transformation is also known as reparametrization trick [69] to learn the parameters of an approximate distribution. It is used in Automatic Differentiation Variational Inference [85] to transform the original model into a standardized space, where a variational approximation is conducted. In this paper we discuss this transformation for general hierarchical Bayesian models and numerical implications of performing variational approximations in these transformed parameters in terms of fidelity and convergence.

We derive this standardizing transformation in two steps. First we transform the original parameters of the deep hierarchy model to independent, uniformly distributed parameters, using the multivariate distributional transform [119]. This step already removes the deep hierarchy and introduces a uniform prior. The uniform prior is problematic for many inference schemes as it limits the parameter space to the unit interval but does not provide further gradient information for the parameters. Thus, in a second step we then transform the uniform parameters to a Gaussian parameters with unit variance and vanishing mean. The overall transformation is then a non-linear, deterministic machinery which relates uniform, white Gaussian parameters to the original parameters of the deep model. The prior information of the deep model is stored in the structure of the transformation itself. The Gaussian prior allows us to make quantitative statements about the conditioning of the curvature of the log-posterior in different scenarios, which largely determines the difficulty numerical inference schemes face. The transformation leads to an optimal conditioning in parameter directions that are only poorly constrained by the data. Directions that are highly constrained by the data result in bad conditioning in this transformed space. Although the transformation discussed does not change the statistical model, inference schemes that involve approximations do depend on the choice of the coordinate system.

As an illustrating numerical example we discuss a Gaussian process regression with also unknown correlation structure. To infer the correlation structure as well statistical homogeneity and isotropy is assumed. The correlation structure can then be expressed in terms of a one-dimensional power spectrum that fully specifies it and has to be inferred together with the signal. To be specific about the statistical process generating the signal we assume the signal to be the result of a zero mean Gaussian process with a kernel as implied by the unknown power spectrum. The log-power spectrum is assumed to be generated by a Gaussian process as well, this time with a known smoothness enforcing kernel. In this scenario we can conduct the identical approximation in both coordinate systems, allowing us to investigate the effect of the transformation on the numerics. We compare the convergence behavior of this deep and highly coupled model with the transformed, flat model in situations with different amounts of data. Here we find that the deep model performs well in highly informed cases and struggles in low information scenarios. The flat model behaves the other way round. Alternating between the two perspectives in a numerical scheme overcomes their individual limitations and provides an overall increased performance.

8.2.1 Related works

Inverse transform sampling [32] is used to generate random variables according to an arbitrary distribution from uniform samples. This is done via the inverse of the cumulative density function (CDF), or quantile function. Its multivariate generalization is the multivariate distributional transform [119], which encodes all the internal hierarchical dependencies of the model. We will use it to reparametrize the original model to obtain the flat formulation.

The reparametrization trick allows to infer parameters of complex approximate posterior distributions via variational inference. The problem is to take gradients with respect to the parameters of the approximation as the Kullback-Leibler divergence is only estimated statistically via samples from the approximation [121]. Reparametrizing the coordinates of the problem in a certain way allows to separate the randomness of drawing samples from a deterministic modification by the parameters. This can, for example, be achieved analogously to the inverse transform sampling by using the inverse CDF. The advantage of the reparametrization of the distribution is, that now gradients can be calculated with respect to the parameters of the distribution, which only appear in the deterministic part. It is introduced in the Auto Encoding Variational Bayes (AEVB) algorithm [69] to make the parameters of the approximate distribution part of the network architecture. Samples to approximate the variational bound can then be drawn from some simple distribution, allowing the inference of all parameters.

Automatic Differentiation Variational Inference (ADVI) performs a transformation of the problem to a set of standardized, real-space parameters where then a variational approximation is conducted [85]. We construct this transformation in terms of the multivariate distributional transform and investigate theoretical and numerical properties of performing approximations in this transformed space.

Normalizing flows are used for non-parametric density estimation [129, 130]. To apply those one tries to find a set of transformations of the parameters of a simple distribution, such that the transformed distribution matches the target distribution as closely as possible. Here it is important to keep track of the functional determinants introduced by the transformation, in order to keep the resulting distribution normalized. The learned transformation stores all the complexity of the approximate posterior in a deterministic way, whereas the randomness originates from a simple distribution. Similarly, the transformation captures the complex structure of the deep model and relates it back to a simple prior distribution. Instead of learning a suitable transformation, the standardizing transformation makes use of the structure of the hierarchical model. A method that makes use of both, the reparametrization trick, as well as normalizing flow is the variational inference with normalizing flows [112].

8.3 Basics and notation

8.3.1 Bayesian inference

In Bayesian inference the prior knowledge $\mathcal{P}(\theta)$, expressed as a probability density function (PDF) of some quantity θ should be updated after we obtained data d. This data is related to the quantity of interest by a likelihood $\mathcal{P}(d|\theta)$ of observing the data, given θ. The updated knowledge is the posterior density $\mathcal{P}(\theta|d)$ and is calculated according to Bayes theorem:

$$\mathcal{P}(\theta|d) = \frac{\mathcal{P}(d|\theta)\mathcal{P}(\theta)}{\mathcal{P}(d)} \tag{8.1}$$

In order to do so, one has to normalize the joint density $\mathcal{P}(d,\theta) = \mathcal{P}(d|\theta)\mathcal{P}(\theta)$ with respect to θ, which is done by division with the evidence $\mathcal{P}(d)$.

An alternative way how to describe probabilities is in terms of their information. Information is the negative logarithm of the distribution, $\mathcal{H}(.) = -\ln\mathcal{P}(.)$. Compared to the distributions, which are multiplicative, information is additive. This makes it a more convenient quantity to deal with and they will be adapted later in this work. Bayes theorem in terms of information reads:

$$\mathcal{H}(\theta|d) = -\ln(\mathcal{P}(\theta|d)) \tag{8.2}$$
$$= \mathcal{H}(d|\theta) + \mathcal{H}(\theta) - \mathcal{H}(d) \tag{8.3}$$

The task of Bayesian inference usually comes down to dealing with the normalization term $\mathcal{H}(d)$. In many cases it is not accessible analytically and sampling techniques or approximate inference has to be applied that avoid the calculation of this term. Those only require all terms depending explicitly on the parameters. In order to shorten the notation we will absorb all terms of constant information, therefore parameter independent, into one single information term $\mathcal{H}_0$ and introduce the symbol $\widehat{=}$ that indicates equality up to parameter independent, constant terms.

$$\mathcal{H}(\theta|d) = \mathcal{H}_0 + \mathcal{H}(d|\theta) + \mathcal{H}(\theta) \tag{8.4}$$
$$\widehat{=}\, \mathcal{H}(d|\theta) + \mathcal{H}(\theta) \tag{8.5}$$

8.3.2 Variational Inference

Variational inference [18] is a powerful tool to approximate an intractable posterior $\mathcal{P}(\theta|d)$ distribution with simpler distribution $\widetilde{\mathcal{P}}(\theta|\varphi)$ parametrized by a set of parameters φ. This is done by minimizing the variational Kullback-Leibler divergence (KL) [86] that is defined as:

$$\mathcal{D}_{\mathrm{KL}}\left(\widetilde{\mathcal{P}}(\theta|\varphi)||\mathcal{P}(\theta|d)\right) \equiv \int D\theta\, \widetilde{\mathcal{P}}(\theta|\varphi) \ln\left[\frac{\mathcal{P}(\theta|d)}{\widetilde{\mathcal{P}}(\theta|\varphi)}\right] \tag{8.6}$$
$$\widehat{=}\, \langle\mathcal{H}(d,\theta)\rangle_{\widetilde{\mathcal{P}}(\theta|\varphi)} - \langle\widehat{\mathcal{H}}(\theta|\varphi)\rangle_{\widetilde{\mathcal{P}}(\theta|\varphi)} \tag{8.7}$$

In the second line parameter independent terms are dropped, as they are irrelevant for the minimization. This includes the normalization constant of the posterior, which

typically is the origin of the intractability of the posterior distribution. This term is also equivalent to the negative evidence lower bound (ELBO) [16]. In order to perform the optimization, we do not have to be able to compute the expectation value explicitly, it is sufficient to be capable of drawing samples from the approximate distribution [121].

We will focus on two kinds of variational approximations, point-like, as well as Gaussian approximations. It can be argued whether fitting delta distributions $\widetilde{\mathcal{P}}(\theta|\varphi) = \delta(\theta - \theta^*)$ are truly a variational method, however minimizing the variational KL and maximizing the posterior distribution results in the same procedure. A point estimate on certain parameters might be justified, especially if they are strongly constrained by the problem and they can be inferred significantly faster compared to more sophisticated approaches.

In cases the uncertainty should not be neglected, a variational Gaussian approximation is a powerful tool to infer posterior quantities, as well as their correlations and uncertainties. Gaussian distributions make it simple to sample from them, and derivatives with respect to their parameters can also be calculated easily, as the following expressions hold [102]:

$$\frac{d\mathcal{D}_{\mathrm{KL}}}{d\bar{\theta}} = \left\langle \frac{d\mathcal{H}(d,\theta)}{d\theta} \right\rangle_{\mathcal{G}(\theta-\bar{\theta},\Theta)} \tag{8.8}$$

$$\Theta^{-1} = \left\langle \frac{d^2\mathcal{H}(d,\theta)}{d\theta d\theta^\dagger} \right\rangle_{\mathcal{G}(\theta-\bar{\theta},\Theta)} \tag{8.9}$$

We will make use of these properties in the example.

8.3.3 Transforming Probability Densities

Probability densities are differential quantities and to turn them into probabilities they have to be equipped with the differential of their arguments, which are often not stated explicitly, and be integrated over. Keeping track of the differential is relevant for coordinate transformations, as one has to take the differential volume change into account in form of the Jacobian determinant. Assume some transformation $\theta' = f(\theta)$ of θ to a new set of parameters θ'. The probability distribution $\mathcal{P}(\theta)$ then transforms as follows:

$$\theta' = f(\theta) \tag{8.10}$$

$$\mathcal{P}(\theta)d\theta = \left|\frac{df^{-1}(\theta')}{d\theta'}\right| \mathcal{P}(\theta')d\theta' \tag{8.11}$$

$$= \mathcal{P}'(\theta')d\theta' \ . \tag{8.12}$$

The vertical lines $|\,.\,|$ indicate the absolute value of the determinant of the matrix expression inside. The new distribution $\mathcal{P}'(\theta')$ is the combination of the functional determinant of the transformation and the old distribution of the transformed argument. This way the new distribution is properly normalized.

8.4 Multivariate Distributional Transform and Standard Gaussian Priors

We want to focus on the transformation that transforms a continuous distribution to a uniform distribution on the unit interval. This is achieved by the quantile function, also known as the inverse cumulative density function (CDF). From the uniformly distributed variables we can then transform to white, standard Gaussian coordinates with the CDF of this Gaussian. In the one dimensional case, the quantile transformation reads:

$$\theta_1 = \mathcal{F}^{-1}_{\mathcal{P}(\theta_1)}(u_1) \quad \text{,where} \tag{8.13}$$

$$\mathcal{F}_{\mathcal{P}(\theta_1)}(\theta_1) \equiv \int_{-\infty}^{\theta_1} d\theta_1' \mathcal{P}(\theta_1') \ . \tag{8.14}$$

The first equation is equivalent to inverse transform sampling [32] and the second equation defines the CDF. The derivative of the CDF with respect to its argument is therefore again the original PDF.

In general the prior $\mathcal{P}(\theta)$ with $\theta = (\theta_1 \ldots, \theta_n)^{\mathrm{T}}$ can be expressed in terms of its hierarchical structure

$$\mathcal{P}(\theta) = \mathcal{P}(\theta_n|\theta_1 \ldots \theta_{n-1}) \ldots \mathcal{P}(\theta_2|\theta_1)\mathcal{P}(\theta_1). \tag{8.15}$$

From this separation we can build up iteratively the the transformation to the hierarchical parameters θ from a set of uniformly distributed parameters u.

$$\theta_1 = \mathcal{F}^{-1}_{\mathcal{P}(\theta_1)}(u_1) \tag{8.16}$$

$$\theta_2 = \mathcal{F}^{-1}_{\mathcal{P}(\theta_2|\theta_1)}(u_2) \tag{8.17}$$

$$\vdots$$

$$\theta_n = \mathcal{F}^{-1}_{\mathcal{P}(\theta_n|\theta_1,\theta_2\ldots,\theta_{n-1})}(u_n) \tag{8.18}$$

This is the multivariate distributional transform [119] for u_i being drawn from the uniform distribution $\mathcal{U}(u_i)$ within the interval $[0, 1]$. From this parametrization we can then change to the white, standard Gaussian distribution in a second step. $\mathcal{G}(\xi, \mathbb{1})$ with uniform, diagonal covariance and vanishing mean. In order to be able to find this transformation in practice, it is necessary that one has explicit access to the hierarchical structure of the model and that CDF's either are available or can be approximated efficiently. This limits its applicability to some extent, but even deep hierarchical models are typically constructed by combining simple distributions and transformations.

We will now perform the above stated coordinate transformation to the uniformly distributed parameters within the joint PDF $\mathcal{P}(d, \theta)$ of the data and parameters for a given likelihood $\mathcal{P}(d|\theta)$ and prior distribution $\mathcal{P}(\theta)$ via $\mathcal{P}(d, \theta) = \mathcal{P}(d|\theta)\mathcal{P}(\theta)$.

$$\mathcal{P}(d|\theta)\mathcal{P}(\theta)d\theta = \mathcal{P}(d|\theta)\left[\mathcal{P}(\theta_1)\dots\mathcal{P}(\theta_n|\theta_1\dots\theta_{n-1})\right]\left|\frac{d\theta}{du}\right|du \tag{8.19}$$

$$= \mathcal{P}(d|\theta)\left[\frac{d}{d\theta_1}\mathcal{F}_{\mathcal{P}(\theta_1)}(\theta_1)\dots\frac{d}{d\theta_n}\mathcal{F}_{\mathcal{P}(\theta_n|\theta_1\dots\theta_{n-1})}(\theta_n)\right]\left|\frac{d\theta}{du}\right|du \tag{8.20}$$

$$= \mathcal{P}\left(d|\mathcal{F}^{-1}_{\mathcal{P}(\theta)}(u)\right)\left[\frac{d}{d\theta_1}\mathcal{F}_{\mathcal{P}(\theta_1)}\left(\mathcal{F}^{-1}_{\mathcal{P}(\theta_1)}(u_1)\right)\right.$$
$$\left.\dots\frac{d}{d\theta_n}\mathcal{F}_{\mathcal{P}(\theta_n|\theta_1\dots\theta_{n-1})}\left(\mathcal{F}^{-1}_{\mathcal{P}(\theta_n|\theta_1\dots\theta_{n-1})}(u_n)\right)\right]\left|\frac{d\theta}{du}\right|du \tag{8.21}$$

$$= \mathcal{P}\left(d|\mathcal{F}^{-1}_{\mathcal{P}(\theta)}(u)\right)\left[\frac{du_1}{d\theta_1}\dots\frac{du_n}{d\theta_n}\right]\left|\frac{d\theta}{du}\right|du \tag{8.22}$$

$$= \mathcal{P}\left(d|\mathcal{F}^{-1}_{\mathcal{P}(\theta)}(u)\right)\left|\frac{du}{d\theta}\right|\left|\frac{d\theta}{du}\right|du \tag{8.23}$$

$$= \mathcal{P}\left(d|\mathcal{F}^{-1}_{\mathcal{P}(\theta)}(u)\right)du \tag{8.24}$$

Here, we first expanded the prior probability into a hierarchical structure and substituted to an integral over the uniform parameters. We then expressed those individual prior probabilities in terms of derivatives of the CDF. Inserting the transformations for each parameter we obtained identity operations by construction. What remained is the product of the derivatives of the new parameters with respect to the old ones that exactly canceled the Jacobian determinant of the transformation. The uniform prior is implicitly present in the last expression, as the parameters u are only defined on the unit interval, where the uniform distribution takes a value of one.

For numerical purposes this coordinate space is inconvenient to perform inference due to the compact support of the uniform distribution. To avoid this, we will transform from these uniform parameters into a set of independent Gaussian parameters of unit variance and zero mean. This coordinate transformation is again done via the CDF. It takes the following form:

$$u = \mathcal{F}_{\mathcal{G}(\xi,\mathbb{1})}(\xi) \tag{8.25}$$

$$= \frac{1}{2} + \frac{1}{2}\mathrm{erf}\left(\frac{\xi}{\sqrt{2}}\right) \quad , \tag{8.26}$$

where the error function is defined as

$$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}dt \quad , \tag{8.27}$$

and we adopt the convention that scalar functions are applied to vectors componentwise. Performing the transformation explicitly yields

$$\mathcal{P}\left(d|\mathcal{F}^{-1}_{\mathcal{P}(\theta)}(u)\right)du = \mathcal{P}\left(d|\mathcal{F}^{-1}_{\mathcal{P}(\theta)}(u)\right)\frac{du}{d\xi}d\xi \tag{8.28}$$

$$= \mathcal{P}\left(d|\mathcal{F}^{-1}_{\mathcal{P}(\theta)}\circ\mathcal{F}_{\mathcal{G}(\xi,\mathbb{1})}(\xi)\right)\left[\frac{d}{d\xi}\mathcal{F}_{\mathcal{G}(\xi,\mathbb{1})}(\xi)\right]d\xi \tag{8.29}$$

$$= \mathcal{P}\left(d|\mathcal{F}^{-1}_{\mathcal{P}(\theta)}\circ\mathcal{F}_{\mathcal{G}(\xi,\mathbb{1})}(\xi)\right)\mathcal{G}(\xi,\mathbb{1})d\xi \quad . \tag{8.30}$$

The overall standardizing transformation $\mathcal{C}_{\mathcal{P}(\theta)}(\xi)$ of the deep hierarchical parameters θ to the white, standard Gaussian parameters ξ is therefore the composition, indicated by $\circ$, of the two individual transformations.

$$\theta = \mathcal{F}^{-1}_{\mathcal{P}(\theta)} \circ \mathcal{F}_{\mathcal{G}(\xi,\mathbb{1})}(\xi) \tag{8.31}$$

$$\equiv \mathcal{C}_{\mathcal{P}(\theta)}(\xi) \tag{8.32}$$

Finally, we can rewrite the original probability in terms of the new parameters.

$$\mathcal{P}(d|\theta)\mathcal{P}(\theta)d\theta = \mathcal{P}\left(d|\mathcal{C}_{\mathcal{P}(\theta)}(\xi)\right)\mathcal{G}(\xi,\mathbb{1})d\xi \tag{8.33}$$

Explicit examples for a simple hierarchical model and multivariate Gaussian prior distributions are given in Appendices 8.8 and 8.8, reproducing the reparametrization trick.

8.5 Approximations of the Transformed Distributions

Approximating posterior distributions allows to infer posterior quantities even for large models within reasonable computational effort. In general it matters in which coordinate system the approximation is conducted. We will discuss the impact of this transformation on two popular approximations in the standardized parameters. The first one will be the maximum posterior (MAP) estimate that is obtained by minimizing the information of the posterior with respect to the parameters. The second approach is to perform a variational approximation with a Gaussian distribution in the standardized coordinates.

8.5.1 Maximum Posterior

A maximum posterior approximation is cheap to compute and can provide meaningful results, if the parameters are constrained reasonably well and uncertainties are small. Conceptually, the true posterior distribution is approximated by a delta distribution at a location that has to be inferred. Minimizing the KL divergence in this case is identical to minimizing the information $\mathcal{H}(\theta|d)$ with respect to the parameters θ. Performing the approximation in the standardized coordinate system will not necessarily maximize the posterior in the original parametrization. We can discuss the two limiting cases of uninformative likelihood and extremely constraining data.

In the case of an uninformative likelihood, maximizing the posterior will be close to maximizing the white, standard Gaussian information. The result will be a delta distribution peaked close to the the origin. This location splits the probability mass in any direction in half. Transforming this distribution back into the original coordinates this location corresponds to the median of the prior distribution rather than the (or a) mode that would be the result of MAP in the original space. Especially for heavily skewed or multi-modal prior distributions the results differ substantially.

In the limiting case of highly informative data the true posterior distribution will be narrowly peaked around the true parameter value. In this situation the essential features of the posterior can be captured by an approximation with a delta distribution, neglecting uncertainty. The true posterior then also transforms almost like a delta

distribution, which will also be narrowly peaked around the transformed maximum. In the highly informed case it therefore does not matter much in which coordinates the approximation is conducted.

Any situation in between the extreme cases will exhibit characteristics of both. In general, performing a MAP approximation in the standard coordinates will push towards median prior configurations, unless the data tells otherwise. This is different to the approximation in original coordinates, which favors maximum prior configurations in the absence of additional information. Towards more conclusive data it becomes irrelevant in which coordinates the posterior is maximized.

8.5.2 Variational Gaussian

A natural choice to approximate the true posterior in the transformed coordinates is the Gaussian distribution. This is demonstrated for ADVI [85]. The transformed prior distribution is just a standard Gaussian. If the data updates the prior only slightly, the true posterior will still be close to a Gaussian distribution, which is captured well by the approximation. The strength of variational inference is to take the uncertainty of the problem into account and thereby prevents over-fitting to some extent. This is especially important if the uncertainty is high, which is the case for uninformative data. This is exactly the situation where the variational Gaussian approximation in the standardized space captures the true uncertainty of the actual posterior the best.

A variational Gaussian approximation will also approximate the true posterior well if the likelihood in the transformed coordinates is close to a Gaussian distribution in the parameters, as combining a Gaussian likelihood and prior results in another Gaussian.

For well constrained parameters this approximation will behave similarly to the MAP approximation, as discussed before. If the likelihood introduces multi-modality into the posterior, the Gaussian approximation will choose one mode and approximate the true posterior locally. In these situations one might consider more flexible distributions to approximate the posterior, for example a Gaussian mixture, as demonstrated in Kucukelbir et al. [85].

Overall, the variational Gaussian approximation of the true posterior in the standardized space will be excellent if the data modifies the posterior only slightly.

8.6 Optimization and Conditioning

In order to perform the approximate inference, we do have to optimize a target functional numerically. This problem might be a non-convex optimization and we are not guaranteed to find a global optimum. Here we will discuss the convergence properties of local optimization procedures of the MAP and variational Gaussian approximation in the standardized coordinates. To perform the optimization we do only have access to local information in terms of derivatives of the loss function at the current position in parameter space. Here the negative gradient shows in the direction of the steepest descent and the curvature informs about how the terrain is changing along the different directions. This information is used in a number of Newton and quasi-Newton algorithms to minimize the target functional. The numerical difficulty is encoded in

the condition number of the curvature, the ratio of the absolutes of its largest and smallest eigenvalue. The larger the spread of the eigenvalues, the harder the problem is to solve numerically. For non-convex problems the curvature might exhibit negative eigenvalues, especially far from a minimum. In this case a Newton optimization breaks down, as the step will be performed in the wrong direction. A way to overcome this is to approximate the true curvature with a quadratic approximation to the curvature with the Gauss-Newton method that is always positive definite. For our discussion we will consider only convex curvature. Otherwise the same argumentation holds for an approximate Gauss-Newton curvature.

Conditioning of general models: The general curvature of the information for a deep hierarchical model reads:

$$C = \frac{d^2}{d\theta d\theta^\dagger}\mathcal{H}(d|\theta) + \frac{d^2}{d\theta d\theta^\dagger}\mathcal{H}(\theta) \tag{8.34}$$

The conditioning of this curvature will entirely depend on the concrete model, but a number of general properties influence the conditioning strongly. Large absolute entries in the curvature matrix will contribute to a bad conditioning, wherever they occur. The second derivative with respect to two parameters will be large if the relevant terms themselves are highly informative, as well as if the interaction of the parameters within these terms is strong. Highly informative and strongly coupled terms will therefore contribute to bad conditioning. Not only individual, large entries are problematic, but also a large number of relatively small entries associated with one individual parameter. This case is discussed in Betancourt and Girolami [15] and illustrated there in form of a high dimensional funnel.

Preconditioning: A common technique to reduce the condition number of a linear problem is preconditioning [126]. The idea is to have an approximation of the original problem that has an easily accessible inverse. This approximation is pulled out of the initial matrix, taking already care of the largest and smallest eigenvalues. It remains to solve a better conditioned problem. For example, we want to have the inverse of a matrix that consists of a simple, invertible, and dominant contribution B plus a small modification A. The condition number is therefore dominated by the matrix B. A way to precondition this problem is to pull B out of it:

$$(A+B)^{-1} = B^{-1}(AB^{-1} + \mathbb{1})^{-1} \tag{8.35}$$

Now it remains to numerically invert $(AB^{-1} + \mathbb{1})^{-1}$ instead of the initial problem. Here the smallest eigenvalue is bounded by 1 and the largest eigenvalue relates to the largest eigenvalue of AB^{-1} (plus one), which is smaller than the product of the largest eigenvalues of A and B^{-1}. As A is only a small modification, its largest eigenvalue will also be small, and therefore pulling out the dominant contribution B leads to a better conditioned problem.

Curvature of MAP and Variational Gaussian: The curvature of the MAP approximation, which is used in the Laplace approximation to obtain an uncertainty estimate,

is the second derivative of the information. It reads in the standardized coordinates:

$$C_{\mathcal{H}} = \frac{d^2}{d\xi d\xi^\dagger}\mathcal{H}(d|\xi) + \mathbb{1} \tag{8.36}$$

For a variational Gaussian approximation we can pull derivatives with respect to the mean inside the expectation value and take the derivative with respect to the original parameters [102]:

$$C_{\mathcal{D}_{\mathrm{KL}}} = \left\langle \frac{d^2}{d\xi d\xi^\dagger}\mathcal{H}(d|\xi) + \mathbb{1} \right\rangle_{\mathcal{G}(\xi - m_\xi, D_\xi)} \tag{8.37}$$

The curvature for the mean of the Gaussian approximation is the mean of the information curvature over the approximate distribution. This structure allows us to investigate the overall conditioning of the curvature in terms of the largest and smallest eigenvalues introduced by the likelihood. The structure of both curvatures above correspond to the one of the preconditioned problem in the previous paragraph and we can discuss them in the same way.

Conditioning: The above considerations allow us to make statements on the conditioning of the problem, as the condition number is given by

$$\kappa = \frac{\lambda_{\max} + 1}{\lambda_{\min} + 1}. \tag{8.38}$$

Here $\lambda_{\max}$ and $\lambda_{\min}$ are the largest and smallest eigenvalues of the likelihood information curvature. The magnitude of the eigenvalues relates to the uncertainty in the corresponding eigendirections of the posterior. This property is used for the Laplace approximation. Large eigenvalues indicate low posterior variance, and therefore well constrained parameter directions, and vice versa. Certain directions might not be constrained by the data at all and the posterior uncertainty is the prior one, which is indicated by $\lambda_{\min} = 0$. Because of this, the smallest eigenvalue of the full curvature cannot become smaller than 1.

The overall conditioning of the problem therefore mainly depends on $\lambda_{\max}$. The larger it is, the worse the overall conditioning. The inference of the model will therefore be faster for less informative data. If the smallest eigenvalue $\lambda_{\min}$ is significantly above one, the data constrains all parameters well and the prior will not have much influence on the conditioning.

The bad conditioning for highly informative data can be explained by the structure of this likelihood. All parameters are directly involved into explaining the data. Within the likelihood the parameters are to some extent degenerate, as several might explain the same features. The prior breaks this degeneracy, favoring certain parameter configurations above others. If the likelihood is now extremely strong, the influence of the prior almost vanishes. The first goal of the algorithm will be to minimize the likelihood, irrespective of the prior plausibility. To restore this plausibility, the optimization has to also minimize the prior contributions. This is now only possible by following a trajectory that keeps the likelihood almost constant. This quasi-hard-constraint introduces narrow, high-dimensional valleys into the information function the optimization has to navigate through.

For well constrained parameters we might prefer a MAP approximation. As previously discussed, the resulting estimate should not depend significantly on the coordinates chosen. A way to circumvent the bad conditioning in the highly informed case might be to perform the optimization in the original parameter space with the deep hierarchy.

8.7 Numerical Example

To illustrate the above considerations we present a numerical example. We will explore the inference of a linear Gaussian process s with unknown kernel operator S from incomplete data with additive, Gaussian noise n . This is an important problem, as Gaussian processes [82] are widely used to model continuous functions or auto-correlated quantities [110]. They are also well-suited for image reconstruction, for example in an astrophysical context [37, 64, 70, 104, 125]. The auto-correlation of the process is the result of the properties of the observed system, which might not be known a priori, so it also has to be learned from the data. Parametric [137], as well as non-parametric [138] models have been proposed to infer the correlation structure. The latter work assumes a mixture of Gaussian profiles, which, in principle, can represent any viable spectral density for a sufficiently large number of basis functions. Alternatively the spectrum itself can also be described using a Gaussian process for the logarithmic spectrum, ensuring positive definiteness [35]. This log-normal prior on the spectral density can, for example, enforce spectral smoothness [103]. We will base our discussion on this latter description, but the results should hold for any parametrization of the correlation function.

8.7.1 Gaussian Process with Spectral Smoothness

Gaussian processes are defined over continues spaces, and therefore the involved quantities will be functions and linear operators instead of vectors and matrices. We will additionally consider the presence of a general, linear response function R, which can, for example, select out individual locations where we measure the signal, resulting in our data-points. This operator is necessary to relate the continuous signal s to discrete-valued data. The data is generated according to

$$d = Rs + n. \tag{8.39}$$

This results in a Gaussian likelihood with information of the form:

$$\mathcal{H}(d|s) = \frac{1}{2}(d - Rs)^\dagger N^{-1}(d - Rs) + \frac{1}{2}\ln|2\pi N| \tag{8.40}$$

The information of a Gaussian process prior reads:

$$\mathcal{H}(s|S) = \frac{1}{2}s^\dagger S^{-1}s + \frac{1}{2}\ln|2\pi S| \tag{8.41}$$

The kernel $S(x, x')$ should be homogeneous, or stationary, and it therefore only depends on the relative distance of two points $S(x - x')$ that allows us to express the

correlation as a diagonal operator in the harmonic basis. The additional assumption of isotropy allows for one dimensional kernel functions that only depend on the relative distance $S(|x - x'|)$. The correlation structure is then a diagonal operator in the harmonic basis and fully characterized by its spectral density, according to the Wiener-Khintchin theorem [68, 135]. The isotropy assumption implies that the spectral density only depends on the absolute values of the coordinates in the harmonic space. This can be expressed via an isotropy operator $\mathbb{P}$, that distributes a one dimensional power spectrum into the diagonal of the covariance operator in the harmonic space. The relation between the correlation structure in position space S and a one dimensional power spectrum p therefore reads:

$$S = \mathbb{F}^\dagger \widehat{(\mathbb{P}p)} \mathbb{F} \tag{8.42}$$

The hat over $\widehat{(\mathbb{P}p)}$ indicates the transformation into a diagonal operator and the harmonic transformation is expressed in terms of $\mathbb{F}$. For flat geometries this is the Fourier transformation. In order to enforce the positive definiteness of the correlation structure, p has to be strictly positive, but its values can vary strongly. In many cases one can assume a smooth power spectrum. A suitable choice of a prior to implement these characteristics is a log-normal Gaussian process prior $\mathcal{LN}(p, T)$. The kernel T implements the degree of desired smoothness. As this contains the hard constraint of positivity, we will instead reparametrize the power spectrum in terms of the logarithmic power spectrum $p = e^\tau$, which transforms the hyper-prior to the Gaussian Process prior $\mathcal{G}(\tau, T)$. The total information is obtained by adding up all likelihood, prior, and hyper-prior terms. Disregarding all parameter independent terms, it is:

$$\begin{aligned}\mathcal{H}(d, s, \tau) \mathrel{\widehat{=}} & \frac{1}{2}(d - Rs)^\dagger N^{-1}(d - Rs) \\ & + \frac{1}{2} s^\dagger \mathbb{F}^\dagger \widehat{(\mathbb{P}e^{-\tau})} \mathbb{F} s + \frac{1}{2}\mathrm{Tr}\left(\mathbb{P}\tau\right) + \frac{1}{2}\tau^\dagger T^{-1}\tau\end{aligned} \tag{8.43}$$

We will now apply the standardizing transformation to flatten the hierarchy of the Bayesian model. Due to the assumed statistical homogeneity of the signal, we have access to the eigenbasis of the prior correlation structure. Here $\mathbb{F}$ is the Fourier transformation, which allows us to take the square root of the eigenvalues to standardize the s parameter as outlined in Appendix 8.8 .

$$s = \mathbb{F}\left(\widehat{\mathbb{P}e^{\frac{1}{2}\tau}}\right)\xi \tag{8.44}$$

Performing this substitution introduces the dependency on τ into the likelihood and removes the prior terms depending on τ.

The same procedure can be applied to standardize τ as well, which is also described by a Gaussian process. In order to do so we have to express the smoothness kernel T in terms of its eigenbasis and eigenvalues. Smoothness should be a lack of curvature of τ on a logarithmic scale. We can therefore describe the inverse kernel as $T^{-1} = \frac{1}{\sigma^2}\Delta^\dagger\Delta$. The Δ operator implements the Laplace operator on a logarithmic coordinate system and σ the expected deviations from a smooth spectrum. The larger it is, the more curvature is accepted and vice versa. The Laplace operator is diagonal in

the associated harmonic domain and the diagonal elements contained the squared harmonic coordinate l of this logarithmic space.

$$\Delta = \mathbb{V}^\dagger l^2 \mathbb{V} \tag{8.45}$$

This $\mathbb{V}$ operator is the harmonic transformation in the space of the one dimensional logarithmic power spectrum in logarithmic coordinates. We can now express τ in terms of the standard parameters ζ:

$$\tau = \mathbb{V}\frac{\sigma}{l^2}\zeta \tag{8.46}$$

With both transformation in place, the transformed information of the full problem reads:

$$\begin{aligned}\mathcal{H}(d,\xi,\zeta) \widehat{=}\ & \frac{1}{2}\left(d - R\mathbb{F}\left(\widehat{\mathbb{P}e^{\frac{1}{2}\mathbb{V}\frac{\sigma}{l^2}\zeta}}\right)\xi\right)^\dagger N^{-1}\left(d - R\mathbb{F}\left(\widehat{\mathbb{P}e^{\frac{1}{2}\mathbb{V}\frac{\sigma}{l^2}\zeta}}\right)\xi\right) \\ & + \frac{1}{2}\xi^\dagger \mathbb{1}\xi + \frac{1}{2}\zeta^\dagger \mathbb{1}\zeta\end{aligned} \tag{8.47}$$

Now the entire prior knowledge, such as the concepts of homogeneity, isotropy, spectral and spatial smoothness, and positivity, are absorbed into the likelihood. This now implements the problem of the inference of a Gaussian process with unknown correlation structure in form of a characteristic and generative sequence of linear and nonlinear operations between a priori white parameters. The steps to perform the transformation were technical, but straight forward.

Inference

So far we only formulated the identical problem in two equivalent ways, a deep hierarchical model and a flat hierarchical model with a standard Gaussian white prior, but we still have to perform the inference. In this case the performed coordinate transformations were fully linear. This way we can perform the same approximation in both coordinates systems and the resulting distributions will be transformed versions of each other. We will minimize the variational KL divergence between the true posterior distribution and an approximate distribution. The approximation will be a product of a Gaussian for the signal and a point-estimate for the power spectrum. This illustrates a variational Gaussian approximation, as well as point estimates.

The approximate distribution reads:

$$\mathcal{P}(s,\tau|m,D,\tau^*) = \widetilde{\mathcal{P}}(s,\tau) \equiv \mathcal{G}(s-m,D)\delta(\tau-\tau^*) \tag{8.48}$$

The task is to adjust the parameters m, D and τ^* such that the KL divergence between this distribution and the true posterior is minimized. Up to parameter independent, constant terms the KL divergence reads:

$$\mathcal{D}_{\mathrm{KL}}\left(\mathcal{P}(s,\tau|d)||\widetilde{\mathcal{P}}(s,\tau)\right) \widehat{=} \langle \mathcal{H}(s,\tau|d)\rangle_{\widetilde{\mathcal{P}}(s,\tau)} - \langle \widetilde{\mathcal{H}}(s,\tau)\rangle_{\widetilde{\mathcal{P}}(s,\tau)} \tag{8.49}$$

$$\begin{aligned}\widehat{=}\ & \frac{1}{2}\left\langle (d-Rs)^\dagger N^{-1}(d-Rs) + \frac{1}{2}s^\dagger \mathbb{F}^\dagger(\widehat{\mathbb{P}e^{-\tau^*}})\mathbb{F}s\right\rangle_{\mathcal{G}(s-m,D)} \\ & + \frac{1}{2}\mathrm{Tr}\left(\mathbb{P}\tau^*\right) + \frac{1}{2}\tau^{*\dagger}T^{-1}\tau^* - \mathrm{Tr}\ln\left[2\pi e D\right]\end{aligned} \tag{8.50}$$

The last term corresponds to the entropy of the Gaussian distribution and the delta distribution is already integrated out. Because the information is fully quadratic in s, we can directly solve for the posterior covariance, as well as the mean for a given τ^* [102]:

$$D = \left(R^\dagger N^{-1} R + \mathbb{F}^\dagger(\widehat{\mathbb{P}e^{-\tau^*}})\mathbb{F}\right)^{-1} \tag{8.51}$$

$$j = R^\dagger N^{-1} d \tag{8.52}$$

$$m = Dj \ . \tag{8.53}$$

$\mathcal{P}(s|d,\tau^*) = \mathcal{G}(s-m,D)$ is the Wiener filter solution for a given correlation structure τ^* and it minimizes the KL divergence without the need of a dedicated optimization. The minimization with respect to τ^* requires the evaluation of the KL divergence and its gradient with respect to these parameters. In order to estimate the Gaussian expectation value we will draw samples from the Gaussian distribution for the current value of τ^* and perform the optimization on a stochastic estimate of the KL divergence [121].

The inference procedure now iterates between estimating m and D for a given τ^* and minimizing a stochastic estimate of the KL-divergence with respect to τ^* for the current parameters of m and D.

Analogous terms can be calculated for the standardized coordinates and the inference procedure will be identical.

Implementation and results

Setup: In the concrete example we will consider a two dimensional Gaussian process drawn from the prior distribution. We generate data by measuring the process at random locations and we additionally add white Gaussian noise. We will present three times the identical setup, varying only the amount of data points. We choose a resolution of 128×128 pixel. The first scenario will consist of a measurements of all locations, in the second scenario we randomly sample 10%, or roughly 1600 locations, and in the last case we only take 0.5%, or 83 data points.

The problem is implemented in python using the NIFTy package [124, 128]. We start the optimization of both models with equivalent initial states. All hyperparameters of the involved minimizers are the same as well.

Using the true underlying kernel, we can compute the posterior mean m_{wf} of this simpler problem by evaluating the Wiener filter for the correct spectrum. This better informed estimator will serve as our reference point and we compare both methods in terms of the root mean squared errors (RMS), which are defined as:

$$\epsilon = \sqrt{(m - m_{\mathrm{wf}})^\dagger (m - m_{\mathrm{wf}})} \tag{8.54}$$

By monitoring this quantity we can discuss the convergence behavior of the different methods in the different scenarios.

We do not track the KL divergence as measure for convergence, as it would require the calculation of the entropy term $\mathrm{Tr}\,\ln[2\pi e D]$. The posterior covariance D has $128^2 \times 128^2$ entries and the eigenbasis is not explicitly available. Evaluating the matrix

logarithm scales with the third order in the dimension, requiring 128^6 computations in every step, making this term numerically not accessible. In order to perform the inference itself it is sufficient to have the operator implicitly available.

The overall convergence of the algorithm is determined by the convergence of the power spectrum parameters, as we immediately have the mean and variance of the Gaussian given this spectrum. In order to discuss the convergence behavior it is important to identify which parts of the power spectrum are prior dominated and which are strongly constrained by the likelihood. In order to do so we have to explore how the data contains information on the correlations of different scales. Our knowledge on the true process realization is limited in two ways, namely white Gaussian noise and incomplete coverage.

This noise affects all scales equally, and the relevant quantity to look at is the signal-to-noise ratio. Scales with high power will not be affected much, and are therefore well constrained by the data. If the signal power drops significantly below the noise power, the prior information will be the dominant term.

The random sampling behaves differently. If we randomly select a small number of positions to measure, their average distance is large and we are mainly informed about the largest scales. We are informed about the small scales by having data points close to each other. The more data points we obtain, the more likely it becomes for them to lie close to another one, providing information on the small scale fluctuations.

In our case both effects will be superimposed, but the sparse sampling of data will be the main contribution. With it we will stir between prior and likelihood dominance.

The results can be seen in Fig. 8.1. The three columns correspond to the three different cases. The first row shows the actual data, the second row the posterior mean for the correct kernel. The third row illustrates the RMS error of the current estimate after each algorithmic update during the minimization. The last row shows the progression of the spectra during the inference.

Observations:

1. The flat hierarchy model converges faster in all cases in terms of RMS error.

2. In recovering the true spectrum, both models have complementary strengths and weaknesses. The deep hierarchy is superior in the data dominated case and the flat hierarchy in the prior dominated case.

3. Alternating both methods improves convergence with respect to each and the recovery of the true spectrum.

Convergence: The convergence results can be seen in the third row in Fig 8.1. The flat hierarchy model converges faster in all three cases in terms of the RMS error. In the full data case both methods converge to the identical error, but the flat hierarchy requires one order of magnitude less iterations. In the scattered data cases, the deep model is faster at the beginning, but after a certain amount of steps, the flat model outpaces the other one significantly. In the sparse data case, the deep hierarchy seemingly stops converging after an initial drop off, whereas the flat model reaches a significantly lower error. This illustrates the advantages of the approximation in the

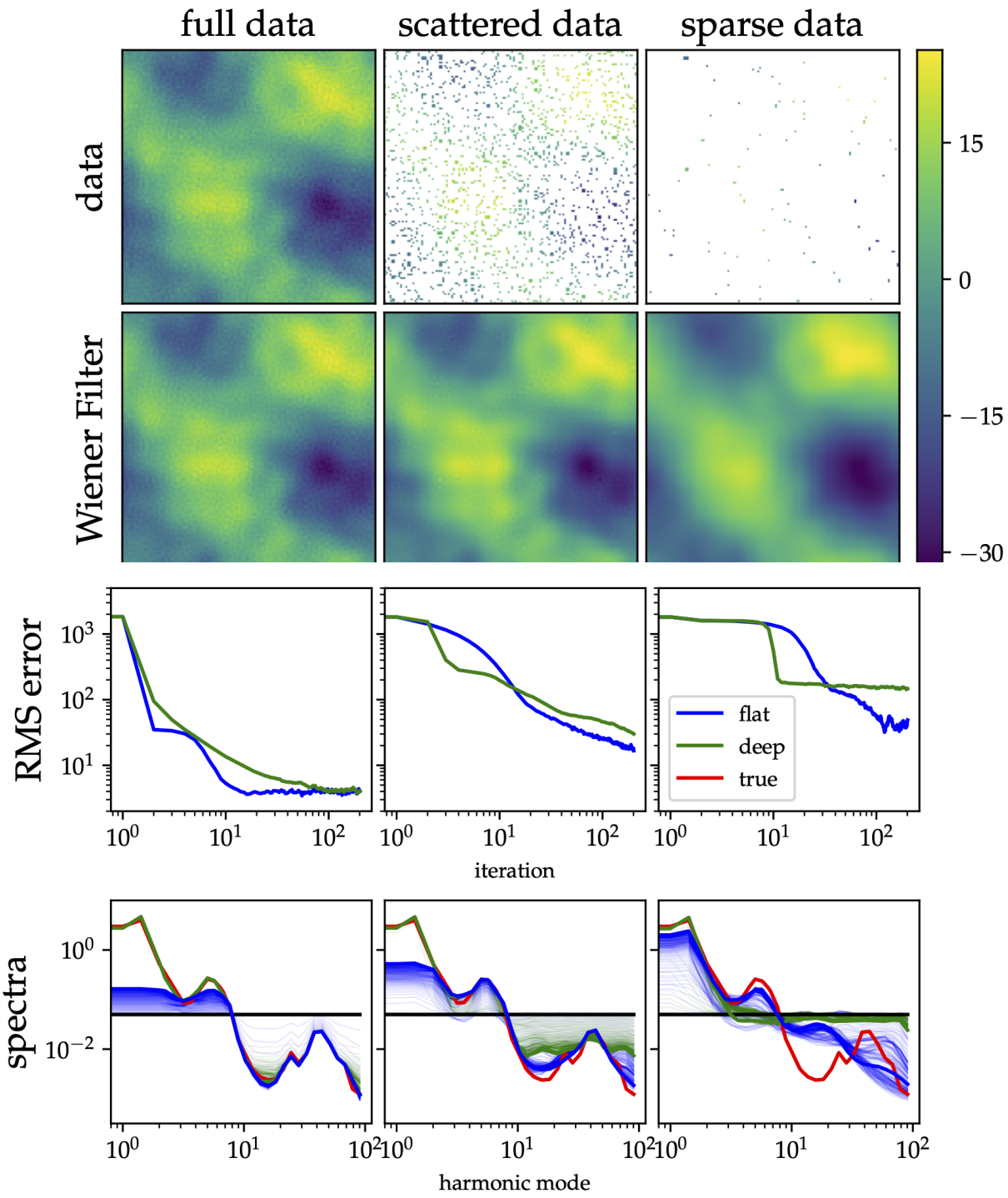


Figure 8.1: The setup and results in the three cases. The column corresponds to the data set with full coverage, the second one to the scattered data set, and the last one to the sparse data set. The first row shows the data in these cases, the second row depicts the Wiener filter reconstruction given the correct correlation structure. The third row illustrates the RMS error of the current reconstruction to the Wiener Filter solution for both coordinates and the last row shows the progression during the optimization.

standardized space for relatively uninformative data. In the scattered data case, the convergence behavior is similar, with slight advantage for the flat model. The problem in this case is that the large scales are well constrained by the data, whereas the small scales should still be prior dominated. Because the deep hierarchy converges fast for well constrained parameters and struggles for the less constrained parameters, the over-all convergence is bottle-necked by the most uninformative parameters. The flat model shows exactly the opposite behavior in the different regimes, and is therefore bottle-necked by the most informed parameters. These behaviors become especially apparent in the convergence of the power spectra.

Spectra: We will now discuss the evolution of the power spectra during the minimization. These are depicted in the last row of Fig. 8.1. Both methods start with spectra at the same horizontal line. From this location the spectra move into the direction of the true correlation structure. We can track fast movement by a low density of lines and slow movement by a high density. Convergence can be identified by small fluctuations around a common spectrum. We also observe the evolution to slow down and almost stopping at some point. This behavior appears as a color gradient in the figure, which progressively gets filled stronger.

The deep hierarchy model is extremely fast in picking up scales, which are well constrained by the data. This can be seen by the density of green lines at the largest scales. In the full data case it immediately jumps close to the true spectrum. For sparser data this is a bit delayed, but still fast. This slowing down cannot only be observed between the cases, but also within the different scales in one setting. The smaller scales are less constraint by the data and for those there are a large number of intermediate lines. In the scattered data case this behavior can be seen very well. There the small scales just slowly drop down from the initial position towards the correct spectrum, bottle-necking the convergence.

In the sparse data case, this slow down behavior is even more dominant. Everything except the largest scales did not move away significantly from the initial position and the minimization is incapable to provide a result within a reasonable amount of time.

The overall behavior can be explained by the deep hierarchical structure of the model. We use the current correlation structure to estimate the mean of the process. For parameters the data is good, this current prior correlation structure does not affect its estimate strongly and if data are sparse, the posterior remains close to the prior estimate. This updated process is then used to adjust the correlation structure within the prior distribution. Therefore well constraint scales are immediately set to reasonable values, whereas the weakly constraint modes are already consistent with the current estimate of the power spectrum. The algorithm gets caught in a loop of self-fulfilling prophecies, which cripples down any progress on scales with little data.

The flat hierarchy model behaves to some extent in an opposite manner. It is fast to recover scales which are weakly constraint by the data, but struggles for well informed scales. The more data it is available, the slower the spectrum approaches the true solution. In situations where the deep hierarchy has problems, this flat model recovers the spectra reasonably. Especially in the sparse data case it is capable on capturing the basic features of the true spectrum. Besides the clearly identified secondary bump, it also estimates the slope of the true spectrum correctly. Spectral features on even

smaller scales could not be identified due to the lack of information. Here the correct spectrum is smoothly interpolated, as we would expect for a Gaussian process in weakly informed scenarios. The flat model is excellent in combining small amounts of data with prior knowledge. Interestingly, its incapability of recovering the correct spectrum in the high data density case does not reflect in the reduction of the RMS error, where the flat model consistently out-competes its competitor.

This behavior can be explained by a multiplicative degeneracy within the likelihood. We re-parametrized the original signal in terms of an amplitude and an excitation:

$$s = A\xi \tag{8.55}$$

This allows to multiply one quantity with some factor, if the other one is divided by it accordingly. In the algorithm we first update the excitation ξ for a given amplitude. In the case of a strong likelihood, the white prior on ξ is almost irrelevant and the excitation will pick up any access power, which is not explained by the initial guess of the amplitude. In the consecutive update of the amplitude itself, this power is hidden inside the excitation and the amplitude cannot pick it up. The likelihood is immediately satisfied and the problem remains in the separation between excitation and amplitude, which is only governed by the priors. The high values in the excitation are incompatible with the prior, which wants to push them down, but it has to act against the extremely strong likelihood. This way, the prior will only slightly push down the values, which releases some power to be picked up by the amplitude. The stronger the likelihood, the slower this process will be and the push from the excitation prior becomes weaker, the progression of the power spectrum will slow down.

Alternating coordinates: Both parametrizations have severe limitations, which is best depicted in the spectra for the scattered data. On large scales, the flat model struggles with internal degeneracy, whereas for small scales the deep model is fighting with self-fulfilling prophecies. Both methods do also have their advantages. The deep model is excellent in the high density data regime, whereas the flat hierarchy allows to recover features even in low information environments.

In this case we can alternate between the coordinates without altering the approximation, due to the linearity of the transformation. In general this is not possible, but as discussed in Sec. 8.5.1 delta approximations for well constrained parameters can also be optimized in transformed coordinates.

Alternating between both methods might allow to overcome the limitations of either one, leading to overall faster convergence and more reasonable spectra. We will restrict our investigation to the scattered data case, where the limitations of both methods were most apparent. The results can be seen in Fig. 8.2. In terms of convergence, alternating the procedures allows to exceed the performance of both methods on their own. The alternating method follows the initial drop of the deep hierarchy model and for the rest it behaves as the flat model, providing the overall lowest error. This can also be seen in the progression of the spectra. The large scales behave like the deep model, rapidly jumping to the correct value, but also match the flat model behavior on the smallest scales.

Overall, the variational approximation in the standardized coordinates demonstrates its numerical superiority in the case of a weak likelihood, or more general,

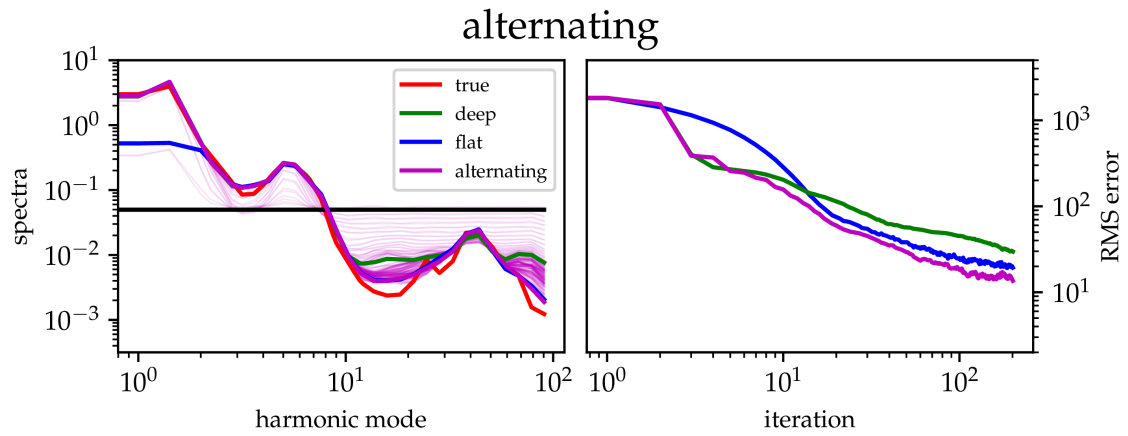


Figure 8.2: The results for alternating between the two parametrizations during the inference. Left is the progression of the power spectra and right the RMS error to the Wiener filter.

in directions the likelihood is relatively uninformative. The overall convergence is limited by the best informed directions, but the other directions converge fast nevertheless. The deep hierarchical parametrization shows the opposite behavior and a MAP estimate for well constrained parameters allows us to alternate between the parametrization, allowing for overall faster convergence.

8.8 Conclusion

We showed how the multivariate distributional transform can be used to transform a general Bayesian model over continuous quantities into a standardized coordinate system in which the prior becomes a standard Gaussian distribution. We discussed the behavior of popular approximations in these new coordinates. A maximum posterior approximation in this space will be oriented towards median prior configurations in the original space, instead of the maximum. For a strong likelihood, the resulting distributions will be transformed versions of each other. A variational Gaussian approximation in the standardized coordinates will be excellent in cases of a weak likelihood, as the shape of the posterior will only be slightly modified compared to the prior, and therefore a Gaussian approximation can capture the true posterior well. This is also the case if the likelihood, containing the standardizing transformation, is still close to a Gaussian.

The simple structure of the model in the standardized coordinates allowed us to investigate the numerical behavior of optimization schemes in terms of the conditioning of the curvature. Its smallest eigenvalue is larger or equal one, and the overall conditioning mainly depends on the largest eigenvalue of the likelihood. This eigenvalue is large if the data contains large amounts of information on certain parameter directions, and small otherwise. This makes the inference of the approximation parameters easier for small amounts, or uninformative data. This makes a Gaussian approximation in the standardized coordinates, as proposed for ADVI [85] a fast and accurate inference procedure especially in cases of sparse data.

We explored numerically the inference of a Gaussian process with unknown correlation structure, which was described by a smooth log-Gaussian process, this time with known, smoothness enforcing kernel. The linearity of the standardizing transformation allowed us to perform the identical approximation in the original, as well as the standardized coordinates. The approximation was a variational Gaussian with full covariance for the Gaussian process, and a point estimate for the power spectrum. We found, as expected, that the flat hierarchical model was superior for less informed parameters, whereas the deep hierarchy outperformed for well-constrained parameters. Both scenarios occur also within the same inference problem, and the convergence speed of either approximation is bottle-necked by its sub-optimal directions. We solved this problem by alternating the optimization between the two coordinate systems, which harnessed the strength of both and lead to an overall faster convergence. We proposed that parameters well constrained by the data should be approximated by delta distributions, which allows their inference in either coordinates. This should improve the inference speed of large, hierarchical Bayesian models.

Acknowledgment

We acknowledge Philipp Arras, Philipp Frank, Maksim Greiner, Sebastian Hutschenreuter, Reimar Leike and Martin Reinecke for fruitful discussions and remarks on the manuscript.

Appendix A: A Simple Transformation Example

We perform this transformation explicitly to a one dimensional example with a hierarchical Bayesian model with two parameters. We consider some likelihood $\mathcal{P}(d|\alpha)$ that depends on a parameter α. The prior distribution on this is a Gaussian with the standard deviation σ and a known mean μ as hyper-parameters. The hyper-prior on this will be an exponential distribution depending on a known constant λ. A similar example is discussed in Betancourt and Girolami [15].

$$\mathcal{P}(\alpha|\sigma) = \mathcal{G}(\alpha - \mu, \sigma^2) \qquad \alpha, \mu \in \mathbb{R} \tag{8.56}$$

$$\mathcal{P}(\sigma) = \lambda e^{-\lambda\sigma} \qquad \sigma, \lambda \in \mathbb{R}^+ \tag{8.57}$$

To calculate the transformations onto white priors we require the CDF of a white Gaussian, which is stated in Eq. 8.26, as well as the inverse CDF's of the prior distributions:

$$\mathcal{F}^{-1}_{\mathcal{P}(\alpha|\sigma)}(u) = \mu + \sqrt{2}\sigma \,\mathrm{erf}^{-1}\,(2u-1) \tag{8.58}$$

$$\mathcal{F}^{-1}_{\mathcal{P}(\sigma)} = -\frac{1}{\lambda}\ln(1-u) \tag{8.59}$$

We can then express our encoding of the prior structure in the likelihood as

$$\alpha = \mathcal{C}_{\mathcal{P}(\alpha|\sigma)}(\xi_1) = \mathcal{F}^{-1}_{\mathcal{P}(\alpha|\sigma)} \circ \mathcal{F}_{\mathcal{G}(\xi,\mathbb{1})}(\xi_1) \tag{8.60}$$

$$= \mu + \sqrt{2}\sigma \,\mathrm{erf}^{-1}\left(2\left(\frac{1}{2} + \frac{1}{2}\mathrm{erf}\left(\frac{\xi_1}{\sqrt{2}}\right)\right) - 1\right) \tag{8.61}$$

$$= \mu + \sigma\xi_1. \tag{8.62}$$

This is just a simple re-scaling and shifting of the white Gaussian ξ parameter within the likelihood. This is exactly the example given for the reparametrization trick [69] and the proposed transformation for hierarchical HMC [15].

Analogously we perform the transformation for the second parameter σ:

$$\sigma = \mathcal{C}_{\mathcal{P}(\sigma)}(\xi_2) \tag{8.63}$$

$$= -\frac{1}{\lambda}\ln\left(\frac{1}{2} - \frac{1}{2}\mathrm{erf}\left(\frac{\xi_2}{\sqrt{2}}\right)\right) \tag{8.64}$$

With this we can substitute σ in the likelihood to obtain full white prior distributions. We can relate back to the original parameter of the likelihood via

$$\alpha = \mu - \frac{1}{\lambda}\ln\left(\frac{1}{2} - \frac{1}{2}\mathrm{erf}\left(\frac{\xi_2}{\sqrt{2}}\right)\right)\xi_1 \tag{8.65}$$

and the posterior can be written as

$$\mathcal{P}(\xi_1, \xi_2|d) = \frac{\mathcal{P}(d|\xi_1, \xi_2)\mathcal{P}(\xi_1)\mathcal{P}(\xi_2)}{\mathcal{P}(d)} \tag{8.66}$$

instead of

$$\mathcal{P}(\alpha, \sigma|d) = \frac{\mathcal{P}(d|\alpha)\mathcal{P}(\alpha|\sigma)\mathcal{P}(\sigma)}{\mathcal{P}(d)}. \tag{8.67}$$

A graphical representation of the of the conditional dependence of the variables within the two models, the initial and the one re-parametrized to have white Gaussian random variables, is depicted in Fig. 8.3. With this transformation we flattened down the hierarchy of the original, deep Bayesian model. Both models contain the identical information. This is now encoded in the nonlinear structure of the parameters within the likelihood. One could continue adding higher levels of prior hierarchies to the problem, but flattening them as well is straight forward, as long as the CDF's are available. As it is pointed out in Betancourt and Girolami [15], the parameters are now independent, conditioned on the data.

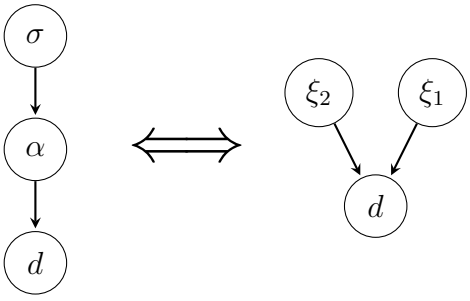


Figure 8.3: The graphical structure of the original model with a deep hierarchy and the flattened structure of the transformed model.

Appendix B: Transforming Multivariate Gaussians

Multivariate Gaussian distributions and their generalizations to infinite dimensions, Gaussian processes [82, 110] are an important class of distributions, especially to express prior knowledge. They take the following form:

$$\mathcal{P}(s) = \mathcal{G}(s, S) = \frac{1}{|2\pi S|^{\frac{1}{2}}} e^{-\frac{1}{2}s^\dagger S^{-1} s} \tag{8.68}$$

Here we do not have to perform the transformation explicitly, as we can find the white, standard Gaussian parametrization through a set of linear transformations. First, we have to express the correlation structure in terms of its eigenbasis.

$$S = \mathbb{F}^\dagger \widetilde{S} \mathbb{F} \tag{8.69}$$

Here the unitary transformation $\mathbb{F}$ is built from the normalized eigenvectors and the diagonal matrix $\widetilde{S}$ consists of the corresponding eigenvalues. The inverse in this basis can be calculated easily by inverting each individual eigenvalue and it reads

$$S^{-1} = \mathbb{F}^\dagger \widetilde{S}^{-1} \mathbb{F}. \tag{8.70}$$

With this we can rewrite the exponent of the Gaussian distribution as

$$\frac{1}{2} s^\dagger S^{-1} S = \frac{1}{2} s^\dagger \mathbb{F}^\dagger \widetilde{S}^{-1} \mathbb{F} s \;\equiv\; \frac{1}{2} \widetilde{s}^\dagger \widetilde{S}^{-1} \widetilde{s}. \tag{8.71}$$

The eigenvalues of S encode the variance of the process along their corresponding eigendirections and it is the squared standard deviation. We can split up this diagonal covariance into two amplitude matrices by taking the matrix square root:

$$\frac{1}{2} \widetilde{s}^\dagger \sqrt{\widetilde{S}^{-1}}^\dagger \sqrt{\widetilde{S}^{-1}} \widetilde{s} = \frac{1}{2} \widetilde{s}^\dagger \sqrt{\widetilde{S}^{-1}}^\dagger \mathbb{1} \sqrt{\widetilde{S}^{-1}} \widetilde{s} \tag{8.72}$$

Finally, we can introduce the new variable $\xi = \sqrt{\widetilde{S}^{-1}} \widetilde{s}$ and the exponential of this Gaussian distribution becomes

$$\frac{1}{2} s^\dagger S^{-1} s = \frac{1}{2} \xi^\dagger \mathbb{1} \xi \tag{8.73}$$

with the full transformation

$$s = \mathbb{F} \sqrt{\widetilde{S}} \xi \equiv A \xi \ . \tag{8.74}$$

This is therefore analogous to the one dimensional case as given in Eq. 8.62. We weight white, random excitations ξ in the eigenbasis of the correlation structure with the corresponding amplitudes contained within A in terms of the standard deviation and transform it back to the space we are interested in. The original correlation structure is expressed in terms of it as $S = AA^\dagger$.

9 Conclusion

This dissertation has led to the development of Metric Gaussian Variational Inference, which allows to approximately solve probabilistic inference problems of enormous scale and complexity. It has so far enabled to solve a number of previously unfeasible problems, providing outstanding scientific results. MGVI performs a series of variational approximations to the posterior distribution of a probabilistic inference problem with a Gaussian distribution. Instead of explicitly parametrizing its full covariance, an implicitly represented expression based on the inverse Fisher information metric is used. Correlations between all model parameters are stored within a set of samples, drawn from the approximation. MGVI alternates between updating the mean of the Gaussian approximation by minimizing the Kullback-Leibler divergence and updating the samples due to changes in the local metric. This procedure scales linearly in time and memory, despite implicitly accounting for a quadratically scaling correlation.

I have outlined a number of applications in which MGVI was used to answer scientifically relevant questions, involving large-scale reconstructions and complex signal models. It allowed to recover a time-resolved reconstruction of the immediate vicinity of the super-massive black hole M87*, despite extremely scarce VLBI data. It was used to obtain a three-dimensional map of interstellar dust, involving tens of million of model parameters together with starlight absorption data and parallaxes. Following the same measurement principle, MGVI also allowed to recover a two-dimensional slice of a patients chest from X-ray CT data, using a segment-aware prior model. Combining two distinct data-sets allowed to produce an improved Faraday rotation map, using an empirical model together with physical insight. In one last example, MGVI helped to simultaneously perform the calibration and imaging of radio-interferometric data, benefiting both procedures to achieve improved results.

This thesis also proposes a way how to include trained, deep neural networks into the Bayesian framework to perform reasoning. This alleviates the necessity of mathematically formulating a model directly, and instead the relevant features are picked up by the network from a large training set. With this, complex questions can be asked by assembling the trained networks, and MGVI allows to rapidly find approximate answers through reasoning to come up with novel ideas. Using this approach allows to combine the strengths of deep learning and Bayesian reasoning to come up with novel ways to approach inference for complex problems. This might be an important step to achieve more intelligent and flexible artificial systems.

MGVI is still new, but has already partly sparked all these results. In the near future, this list will be extended by further examples. The reoccurring theme with all of them is the increase in complexity. More and more aspects of a problem are included and simultaneously solved, as illustrated by the joint calibration and imaging radio-interferometric data.

Every sub-problem addresses a certain aspect of the entire inference problem, which are often reoccurring throughout several tasks. One example is the modelling of dif-

fuse emission on the sky, which was widely used throughout the examples. The descriptions of such sub-problems serve as building blocks for complex and highly specialized inference algorithms, tailored towards a specific problem. This approach highly benefits from a large library of suitable models available and this has to be one goal of future developments. This library, together with MGVI as underlying inference machinery could constitute a powerful reconstruction framework, the Universal Bayesian Imaging Kit. Here, MGVI allows for significantly larger and more complex inference problems, compared to other approaches. The limitations of MGVI, as well as a deeper mathematical understanding will have to be explored in future research.

By allowing to approach large and complex problems in a holistic and flexible way, MGVI is perfectly suited to face the challenges of ever larger and more sensitive experiments.

I want to conclude with a word of caution. The technologies developed in this thesis enable novel ways to extract relevant information from data. As demonstrated in the examples, this can have major benefits to many fields. However, as a physicist, it is my obligation to also consider all consequences of my research. The approaches discussed in this thesis can also be used maliciously to intrude the personal freedoms of unsuspecting targets, potentially within large-scale surveillance systems. Even if the intended use is legitimate, unintended biases can enter through the deep-learned modules and potentially manifest themselves in the conclusions of such systems. Using these technologies therefore demands careful ethical considerations and not to blindly trust the provided results. These always have to be criticised in terms of potential bias.

Bibliography

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL `https://www.tensorflow.org/`. Software available from tensorflow.org.

[2] Benjamin P Abbott, Richard Abbott, TD Abbott, MR Abernathy, Fausto Acernese, Kendall Ackley, Carl Adams, Thomas Adams, Paolo Addesso, RX Adhikari, et al. Observation of gravitational waves from a binary black hole merger. *Physical review letters*, 116(6):061102, 2016.

[3] Benjamin P Abbott, Robert Abbott, TD Abbott, F Acernese, K Ackley, C Adams, T Adams, P Addesso, RX Adhikari, VB Adya, et al. Gravitational waves and gamma-rays from a binary neutron star merger: Gw170817 and grb 170817a. *The Astrophysical Journal Letters*, 848(2):L13, 2017.

[4] Roberto Abuter, Matteo Accardo, A Amorim, N Anugu, Gerardo Avila, N Azouaoui, Myriam Benisty, Jean-Philippe Berger, Nicolas Blind, Henri Bonnet, et al. First light for gravity: Phase referencing optical interferometry for the very large telescope interferometer. *Astronomy & Astrophysics*, 602:A94, 2017.

[5] R Adam, Peter AR Ade, N Aghanim, MIR Alves, M Arnaud, M Ashdown, J Aumont, C Baccigalupi, AJ Banday, RB Barreiro, et al. Planck 2015 results-x. diffuse component separation: Foreground maps. *Astronomy & Astrophysics*, 594:A10, 2016.

[6] Shun-ichi Amari. Neural learning in structured parameter spaces-natural Riemannian gradient. In *Advances in neural information processing systems*, pages 127–133, 1997.

[7] Shun-ichi Amari. *Information geometry and its applications*. Springer, 2016.

[8] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein Generative Adversarial Networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 214–223. PMLR, 2017.

[9] Stefan Arnborg and Gunnar Sjödin. On the foundations of bayesianism. In *AIP Conference Proceedings*, volume 568, pages 61–71. American Institute of Physics, 2001.

[10] Philipp Arras, Mihai Baltac, Torsten A Ensslin, Philipp Frank, Sebastian Hutschenreuter, Jakob Knollmueller, Reimar Leike, Max-Niklas Newrzella, Lukas Platz, Martin Reinecke, et al. Nifty5: Numerical information field theory v5. *Astrophysics Source Code Library*, 2019.

[11] Philipp Arras, Mihai Baltac, Torsten A Ensslin, Philipp Frank, Sebastian Hutschenreuter, Jakob Knollmueller, Reimar Leike, Max-Niklas Newrzella, Lukas Platz, Martin Reinecke, et al. Nifty5: Numerical information field theory v5. *Astrophysics Source Code Library*, 2019.

[12] Philipp Arras, Philipp Frank, Reimar Leike, Rüdiger Westermann, and Torsten Enßlin. Unified radio interferometric calibration and imaging with joint uncertainty quantification. *arXiv preprint arXiv:1903.11169*, 2019.

[13] Philipp Arras, Philipp Frank, Philipp Haim, Jakob Knollmüller, Reimar Leike, Martin Reinecke, and Torsten Enßlin. The variable shadow of m87. *arXiv preprint arXiv:2002.05218*, 2020.

[14] Axel Barrau and Silvere Bonnabel. A note on the intrinsic cramer-rao bound. In *International Conference on Geometric Science of Information*, pages 377–386. Springer, 2013.

[15] Michael Betancourt and Mark Girolami. Hamiltonian monte carlo for hierarchical models. *Current trends in Bayesian methodology with applications*, 79: 30, 2015.

[16] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006. ISBN 0387310738.

[17] Lindy Blackburn, Dominic W Pesce, Michael D Johnson, Maciek Wielgus, Andrew A Chael, Pierre Christian, and Sheperd S Doeleman. Closure statistics in radio interferometric data. *arXiv preprint arXiv:1910.02062*, 2019.

[18] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112 (518):859–877, 2017.

[19] Vanessa Böhm, François Lanusse, and Uroš Seljak. Uncertainty Quantification with Generative Models. In *4th workshop on Bayesian Deep Learning*. NeurIPS, 2019.

[20] J Richard Bond, Lev Kofman, and Dmitry Pogosyan. How filaments of galaxies are woven into the cosmic web. *Nature*, 380(6575):603–606, 1996.

[21] Ashish Bora, Ajil Jalal, Eric Price, and Alexandros G Dimakis. Compressed Sensing using Generative Models. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 537–546. JMLR. org, 2017.

[22] Michael Braun and Jon McAuliffe. Variational inference for large-scale models of discrete choice. *Journal of the American Statistical Association*, 105(489): 324–335, 2010.

[23] AGA Brown, A Vallenari, T Prusti, JHJ de Bruijne, C Babusiaux, CAL Bailer-Jones, Gaia Collaboration, et al. Gaia data release 2. summary of the contents and survey properties. *arXiv preprint arXiv:1804.09365*, 2018.

[24] Daniel Buscombe. Spatially explicit spectral analysis of point clouds and geospatial data. *Computers & Geosciences*, 86:92–108, 2016.

[25] Allen Caldwell. *Lecutre on Model-Based Data Analysis Parameter Inference and Model Testing*. 2019.

[26] Jean-François Cardoso and Antoine Souloumiac. Blind beamforming for non-gaussian signals. In *IEE Proceedings F (Radar and Signal Processing)*, volume 140, pages 362–370. IET, 1993.

[27] Jean-François Cardoso, Jacques Delabrouille, and Guillaume Patanchon. Independent component analysis of the cosmic microwave background. In *4th International Symposium on Independent Component Analysis and Blind Signal Separation (ICA03)*, 2003.

[28] IceCube Collaboration et al. Multimessenger observations of a flaring blazar coincident with high-energy neutrino icecube-170922a. *Science*, 361(6398), 2018.

[29] Pierre Comon. Independent component analysis, a new concept? *Signal processing*, 36(3):287–314, 1994.

[30] Richard T Cox. Probability, frequency and reasonable expectation. *American journal of physics*, 14(1):1–13, 1946.

[31] Harald Cramér. *Mathematical methods of statistics*, volume 9. Princeton university press, 1946.

[32] Luc Devroye. Sample-based non-uniform random variate generation. In *Proceedings of the 18th conference on Winter simulation*, pages 260–265. ACM, 1986.

[33] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image Super-Resolution using Deep Convolutional Networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):295–307, 2015.

[34] Simon Duane, Anthony D Kennedy, Brian J Pendleton, and Duncan Roweth. Hybrid monte carlo. *Physics letters B*, 195(2):216–222, 1987.

[35] Torsten A Enßlin and Mona Frommert. Reconstruction of signals with unknown spectra in information field theory with parameter uncertainty. *Physical Review D*, 83(10):105014, 2011.

[36] Torsten A Enßlin and Jakob Knollmüller. Correlated signal inference by free energy exploration. *arXiv preprint arXiv:1612.08406*, 2016.

[37] Torsten A Enßlin, Mona Frommert, and Francisco S Kitaura. Information field theory for cosmological perturbation reconstruction and nonlinear signal analysis. *Physical Review D*, 80(10):105005, 2009.

[38] The EHT Collaboration et al. First m87 event horizon telescope results. i. the shadow of the supermassive black hole. *ApJL*, 875:1, 2019. URL `https://iopscience.iop.org/article/10.3847/2041-8213/ab0ec7`.

[39] The EHT Collaboration et al. First m87 event horizon telescope results. ii. array and instrumentation. *ApJL*, 875:2, 2019. URL `https://iopscience.iop.org/article/10.3847/2041-8213/ab0c96`.

[40] The EHT Collaboration et al. First m87 event horizon telescope results. iii. data processing and calibration. *ApJL*, 875:3, 2019. URL `https://iopscience.iop.org/article/10.3847/2041-8213/ab0c57`.

[41] The EHT Collaboration et al. First m87 event horizon telescope results. iv. imaging the central supermassive black hole. *ApJL*, 875:4, 2019. URL `https://iopscience.iop.org/article/10.3847/2041-8213/ab0e85`.

[42] The EHT Collaboration et al. First m87 event horizon telescope results. v. physical origin of the asymmetric ring. *ApJL*, 875:5, 2019. URL `https://iopscience.iop.org/article/10.3847/2041-8213/ab0f43`.

[43] The EHT Collaboration et al. First m87 event horizon telescope results. vi. the shadow and mass of the central black hole. *ApJL*, 875:6, 2019. URL `https://iopscience.iop.org/article/10.3847/2041-8213/ab1141`.

[44] Philipp Frank, Reimar Leike, and Torsten A Enßlin. Field dynamics inference for local and causal interactions. *arXiv preprint arXiv:1902.02624*, 2019.

[45] Andrew Gelman and Jennifer Hill. *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press, 2006.

[46] Andrew Gelman, Donald B Rubin, et al. Inference from Iterative Simulation using Multiple Sequences. *Statistical science*, 7(4):457–472, 1992.

[47] Michael Ghil, MR Allen, MD Dettinger, K Ide, D Kondrashov, ME Mann, Andrew W Robertson, A Saunders, Y Tian, F Varadi, et al. Advanced spectral methods for climatic time series. *Reviews of geophysics*, 40(1), 2002.

[48] J. Ghosh and R. Ramamoorthi. Bayesian nonparametrics. *Springer Series in Statistics*, 16, 01 2011.

[49] Ryan Giordano, Tamara Broderick, and Michael I Jordan. Covariances, robustness and variational bayes. *The Journal of Machine Learning Research*, 19(1): 1981–2029, 2018.

[50] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[51] Gregory M. Green, Edward F. Schlafly, Catherine Zucker, Joshua S. Speagle, and Douglas P. Finkbeiner. A 3D Dust Map Based on Gaia, Pan-STARRS 1 and 2MASS. *arXiv e-prints*, art. arXiv:1905.02734, May 2019.

[52] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of Wasserstein GANs. In *Advances in neural information processing systems*, pages 5767–5777, 2017.

[53] Marcelo Hartmann and Jarno Vanhatalo. Laplace approximation and natural gradient for gaussian process regression with heteroscedastic student-t model. *Statistics and Computing*, pages 1–21, 2018.

[54] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[55] Eric Heim. Constrained Generative Adversarial Networks for Interactive Image Generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10753–10761, 2019.

[56] Matthew D Hoffman, David M Blei, Chong Wang, and John Paisley. Stochastic Variational Inference. *The Journal of Machine Learning Research*, 14(1):1303–1347, 2013.

[57] Michael F Hutchinson. A stochastic estimator of the trace of the influence matrix for Laplacian smoothing splines. *Communications in Statistics-Simulation and Computation*, 19(2):433–450, 1990.

[58] Sebastian Hutschenreuter and Torsten A Enßlin. The galactic faraday depth sky revisited. *Astronomy & Astrophysics*, 633:A150, 2020.

[59] Aapo Hyvärinen. Independent component analysis in the presence of gaussian noise by maximizing joint likelihood. *Neurocomputing*, 22(1):49–67, 1998.

[60] Aapo Hyvärinen and Erkki Oja. Independent component analysis: algorithms and applications. *Neural networks*, 13(4):411–430, 2000.

[61] Aapo Hyvärinen, Juha Karhunen, and Erkki Oja. *Independent component analysis*, volume 46. John Wiley & Sons, 2004.

[62] Jens Jasche and Benjamin D Wandelt. Bayesian physical reconstruction of initial conditions from large-scale structure surveys. *Monthly Notices of the Royal Astronomical Society*, 432(2):894–913, 2013.

[63] Edwin T Jaynes. *Probability theory: The logic of science*. Cambridge university press, 2003.

[64] H Junklewitz, MR Bell, M Selig, and TA Enßlin. Resolve: A new algorithm for aperture synthesis imaging of extended emission in radio astronomy. *Astronomy & Astrophysics*, 586:A76, 2016.

[65] Tero Karras, Samuli Laine, and Timo Aila. A Style-Based Generator Architecture for Generative Adversarial Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019.

[66] Robert E Kass and Adrian E Raftery. Bayes factors. *Journal of the american statistical association*, 90(430):773–795, 1995.

[67] Mohammad Emtiyaz Khan, Aleksandr Aravkin, Michael Friedlander, and Matthias Seeger. Fast dual variational inference for non-conjugate latent gaussian models. In *International Conference on Machine Learning*, pages 951–959, 2013.

[68] Aleksandr Khintchin. Korrelationstheorie der stationären stochastischen Prozesse. *Mathematische Annalen*, 109(1):604–615, 1934.

[69] Diederik P Kingma and Max Welling. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations*. ICLR, 2014.

[70] FS Kitaura and TA Enßlin. Bayesian reconstruction of the cosmological large-scale structure: methodology, inverse algorithms and numerical optimization. *Monthly Notices of the Royal Astronomical Society*, 389(2):497–544, 2008.

[71] Kimmo Kiviluoto and Erkki Oja. Independent component analysis for parallel financial time series. In *ICONIP*, volume 2, pages 895–898, 1998.

[72] Jakob Knollmüller. DeepReasoning. `https://gitlab.mpcdf.mpg.de/ift/deepreasoning`, 2020.

[73] Jakob Knollmüller and Torsten Enßlin. Bayesian Reasoning with Deep-Learned Knowledge. *arXiv preprint arXiv:2001.11031*, 2020.

[74] Jakob Knollmüller and Torsten A Enßlin. Noisy independent component analysis of autocorrelated components. *Physical Review E*, 96(4):042114, 2017.

[75] Jakob Knollmüller and Torsten A Enßlin. Encoding prior knowledge in the structure of the likelihood. *arXiv preprint arXiv:1812.04403*, 2018.

[76] Jakob Knollmüller and Torsten A Enßlin. Metric Gaussian Variational Inference. *arXiv preprint arXiv:1901.11033*, 2019.

[77] Jakob Knollmüller, Theo Steininger, and Torsten A Enßlin. Inference of signals with unknown correlation structure from nonlinear measurements. *arXiv preprint arXiv:1711.02955*, 2017.

[78] Jakob Knollmüller, Philipp Frank, and Torsten A Enßlin. Separating diffuse from point-like sources-a bayesian approach. *arXiv preprint arXiv:1804.05591*, 2018.

[79] Jakob Knollmüller, Philipp Frank, and Torsten A Ensslin. Starblade: Star and artefact removal with a bayesian lightweight algorithm from diffuse emission. *Astrophysics Source Code Library*, 2018.

[80] David A Knowles and Tom Minka. Non-conjugate variational message passing for multinomial and binary regression. In *Advances in Neural Information Processing Systems*, pages 1701–1709, 2011.

[81] A Kolomogoroff. *Grundbegriffe der Wahrscheinlichkeitsrechnung*, volume 2. Springer-Verlag, 1933.

[82] Daniel G Krige. A statistical approach to some basic mine valuation problems on the witwatersrand. *Journal of the Southern African Institute of Mining and Metallurgy*, 52(6):119–139, 1951.

[83] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet Classification with Deep Convolutional Neural Networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[84] Dirk P Kroese, Thomas Taimre, and Zdravko I Botev. *Handbook of monte carlo methods*, volume 706. John Wiley & Sons, 2013.

[85] Alp Kucukelbir, Dustin Tran, Rajesh Ranganath, Andrew Gelman, and David M Blei. Automatic differentiation variational inference. *The Journal of Machine Learning Research*, 18(1):430–474, 2017.

[86] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

[87] Malte Kuss and Carl Edward Rasmussen. Assessing approximate inference for binary gaussian process classification. *Journal of machine learning research*, 6 (Oct):1679–1704, 2005.

[88] Malte Kuss and Carl Edward Rasmussen. Assessing approximate inference for binary gaussian process classification. *Journal of machine learning research*, 6 (Oct):1679–1704, 2005.

[89] Miguel Lázaro-Gredilla and Michalis K Titsias. Variational heteroscedastic gaussian process regression. In *ICML*, pages 841–848, 2011.

[90] Yann LeCun, Léon Bottou, Yoshua Bengio, Patrick Haffner, et al. Gradient-Based Learning applied to Document Recognition. *Proceedings of the IEEE*, 86 (11):2278–2324, 1998.

[91] RH Leike and TA Enßlin. Charting nearby dust clouds using gaia data only. *arXiv preprint arXiv:1901.05971*, 2019.

[92] RH Leike, M Glatzle, and TA Enßlin. Resolving nearby dust clouds. *arXiv preprint arXiv:2004.06732*, 2020.

[93] Zachary C Lipton and Subarna Tripathi. Precise Recovery of Latent Vectors from Generative Adversarial Networks. In *Workshop Track*. ICLR, 2017.

[94] Duncan Ross Lorimer and Michael Kramer. Handbook of pulsar astronomy. *hpa*, 2012.

[95] Davide Maino, A Farusi, Carlo Baccigalupi, Francesca Perrotta, AJ Banday, L Bedini, Carlo Burigana, Gianfranco De Zotti, KM Górski, and E Salerno. All-sky astrophysical component separation with Fast Independent Component Analysis (FASTICA). *Monthly Notices of the Royal Astronomical Society*, 334 (1):53–68, 2002.

[96] Scott Makeig, Anthony J Bell, Tzyy-Ping Jung, Terrence J Sejnowski, et al. Independent component analysis of electroencephalographic data. *Advances in neural information processing systems*, pages 145–151, 1996.

[97] James Martens. New insights and perspectives on the natural gradient method. *arXiv preprint arXiv:1412.1193*, 2014.

[98] Mehdi Mirza and Simon Osindero. Conditional Generative Adversarial Nets. *arXiv preprint arXiv:1411.1784*, 2014.

[99] Eric Moulines, J-F Cardoso, and Elisabeth Gassiat. Maximum likelihood for blind separation and deconvolution of noisy signals using mixture models. In *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*, volume 5, pages 3617–3620. IEEE, 1997.

[100] Cuong V Nguyen, Yingzhen Li, Thang D Bui, and Richard E Turner. Variational Continual Learning. In *6th International Conference on Learning Representations*, 2018.

[101] Hannes Nickisch and Carl Edward Rasmussen. Approximations for binary gaussian process classification. *Journal of Machine Learning Research*, 9(Oct):2035–2078, 2008.

[102] Manfred Opper and Cédric Archambeau. The variational gaussian approximation revisited. *Neural computation*, 21(3):786–792, 2009.

[103] Niels Oppermann, Marco Selig, Michael R Bell, and Torsten A Enßlin. Reconstruction of Gaussian and log-normal fields with spectral smoothness. *Physical Review E*, 87(3):032136, 2013.

[104] Niels Oppermann, Henrik Junklewitz, Maksim Greiner, Torsten A Enßlin, Takuya Akahori, Ettore Carretti, Bryan M Gaensler, Ariel Goobar, Lisa Harvey-Smith, Melanie Johnston-Hollitt, et al. Estimating extragalactic faraday rotation. *Astronomy & Astrophysics*, 575:A118, 2015.

[105] George Papandreou and Alan L Yuille. Gaussian sampling by local perturbations. In *Advances in Neural Information Processing Systems*, pages 1858–1866, 2010.

[106] Bradley M Peterson. *An introduction to active galactic nuclei*. Cambridge University Press, 1997.

[107] Jonathan R Pritchard and Abraham Loeb. 21 cm cosmology in the 21st century. *Reports on Progress in Physics*, 75(8):086901, 2012.

[108] Timo Prusti, JHJ De Bruijne, Anthony GA Brown, A Vallenari, C Babusiaux, CAL Bailer-Jones, U Bastian, M Biermann, DW Evans, L Eyer, et al. The gaia mission. *Astronomy & Astrophysics*, 595:A1, 2016.

[109] C Radhakrishna Rao. Information and the accuracy attainable in the estimation of statistical parameters. In *Breakthroughs in statistics*, pages 235–247. Springer, 1992.

[110] Carl Edward Rasmussen. Gaussian processes in machine learning. In *Advanced lectures on machine learning*, pages 63–71. Springer, 2004.

[111] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative Adversarial Text to Image Synthesis. In *Proceedings of the 33th International Conference on Machine Learning-Volume 48*. JMLR. org, 2016.

[112] Danilo Jimenez Rezende and Shakir Mohamed. Variational Inference with Normalizing Flows. In *Proceedings of the 32th International Conference on Machine Learning-Volume 37*. JMLR. org, 2015.

[113] A Richard Thompson, James M Moran, and George W Swenson Jr. *Interferometry and synthesis in radio astronomy*. Springer Nature, 2017.

[114] François Roddier. *Adaptive optics in astronomy*. Cambridge university press, 1999.

[115] AEE Rogers, HF Hinteregger, AR Whitney, CC Counselman, II Shapiro, JJ Wittels, WK Klemperer, WW Warnock, TA Clark, LK Hutton, et al. The structure of radio sources 3c 273b and 3c 84 deduced from the'closure'phases and visibility amplitudes observed with three-element interferometers. *The Astrophysical Journal*, 193:293–301, 1974.

[116] Kenneth Rose, Eitan Gurewitz, and Geoffrey Fox. A Deterministic Annealing Approach to Clustering. *Pattern Recognition Letters*, 11(9):589–594, 1990.

[117] Rasmus Rothe, Radu Timofte, and Luc Van Gool. DEX: Deep EXpectation of apparent age from a single image. In *IEEE International Conference on Computer Vision Workshops (ICCVW)*, December 2015.

[118] Rasmus Rothe, Radu Timofte, and Luc Van Gool. Deep Expectation of Real and Apparent Age from a Single Image without Facial Landmarks. *International Journal of Computer Vision (IJCV)*, July 2016.

[119] Ludger Rüschendorf. On the distributional transform, sklar's theorem, and the empirical copula process. *Journal of Statistical Planning and Inference*, 139(11): 3921–3927, 2009.

[120] George B Rybicki and Alan P Lightman. *Radiative Processes in Astrophysics*. John Wiley & Sons, 2008.

[121] Tim Salimans, David A Knowles, et al. Fixed-form variational posterior approximation through stochastic linear regression. *Bayesian Analysis*, 8(4):837–882, 2013.

[122] MP Schützenberger. A generalization of the fréchet-cramér inequality to the case of bayes estimation. *Bull. Amer. Math. Soc*, 63(142), 1957.

[123] Marco Selig, Niels Oppermann, and Torsten A Enßlin. Improving stochastic estimates with inference methods: Calculating matrix diagonals. *Physical Review E*, 85(2):021134, 2012.

[124] Marco Selig, Michael R Bell, Henrik Junklewitz, Niels Oppermann, Martin Reinecke, Maksim Greiner, Carlos Pachajoa, and Torsten A Enßlin. NIFTY– Information Field Theory-A versatile PYTHON library for signal inference. *Astronomy & Astrophysics*, 554:A26, 2013.

[125] Marco Selig, Valentina Vacca, Niels Oppermann, and Torsten A Enßlin. The denoised, deconvolved, and decomposed fermi γ-ray sky-an application of the d3po algorithm. *Astronomy & Astrophysics*, 581:A126, 2015.

[126] Jonathan Richard Shewchuk et al. An introduction to the conjugate gradient method without the agonizing pain, 1994.

[127] George F Smoot, Charles L Bennett, A Kogut, EL Wright, J Aymon, NW Boggess, ES Cheng, G De Amici, S Gulkis, MG Hauser, et al. Structure in the cobe differential microwave radiometer first-year maps. *The Astrophysical Journal*, 396:L1–L5, 1992.

[128] Theo Steininger, Jait Dixit, Philipp Frank, Maksim Greiner, Sebastian Hutschenreuter, Jakob Knollmüller, Reimar Leike, Natalia Porqueres, Daniel Pumpe, Martin Reinecke, et al. Nifty 3-numerical information field theory-a python framework for multicomponent signal inference on hpc clusters. *arXiv preprint arXiv:1708.01073*, 2017.

[129] EG Tabak and Cristina V Turner. A family of nonparametric density estimation algorithms. *Communications on Pure and Applied Mathematics*, 66(2):145–164, 2013.

[130] Esteban G Tabak, Eric Vanden-Eijnden, et al. Density estimation by dual ascent of the log-likelihood. *Communications in Mathematical Sciences*, 8(1):217–233, 2010.

[131] L Tong, VC Soon, YF Huang, and RALR Liu. Amuse: a new blind identification algorithm. In *Circuits and Systems, 1990., IEEE International Symposium on*, pages 1784–1787. IEEE, 1990.

[132] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep Image Prior. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9446–9454, 2018.

[133] Aaron Van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al. Conditional Image Generation with PixelCNN Decoders. In *Advances in neural information processing systems*, pages 4790–4798, 2016.

[134] A. W. Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

[135] Norbert Wiener. *Extrapolation, interpolation, and smoothing of stationary time series*, volume 2. MIT press Cambridge, 1949.

[136] RA Wijsman et al. On the attainment of the cramér-rao lower bound. *The Annals of Statistics*, 1(3):538–542, 1973.

[137] Christopher KI Williams and Carl Edward Rasmussen. Gaussian processes for regression. In *Advances in neural information processing systems*, pages 514–520, 1996.

[138] Andrew Wilson and Ryan Adams. Gaussian process kernels for pattern discovery and extrapolation. In *International Conference on Machine Learning*, pages 1067–1075, 2013.

[139] Ga Wu, Justin Domke, and Scott Sanner. Conditional Inference in Pre-Trained Variational Autoencoders via Cross-Coding. *arXiv preprint arXiv:1805.07785*, 2018.

[140] Xinchen Yan, Jimei Yang, Kihyuk Sohn, and Honglak Lee. Attribute2image: Conditional Image Generation from Visual Attributes. In *European Conference on Computer Vision*, pages 776–791. Springer, 2016.

[141] Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A Efros. Generative Visual Manipulation on the Natural Image Manifold. In *European Conference on Computer Vision*, pages 597–613. Springer, 2016.

Acknowledgments

I want to start with thanking my PhD supervisor Torsten Enßlin. He fully supported me in every aspect throughout the entire time, providing invaluable advice in any situation. His door was always open and he gave me the freedom to explore my own ideas. Thank you!

I want to thank the generation of PhD students before me for for their help and advice during my master thesis and the early stages of my PhD. I remember long sessions patiently explaining all the intricate details of Fourier transformations, volume factors, binbounds, trace probing, and the pundex. Thank you!
Vanessa Böhm, Sebastian Dorn, Mahsa Ghaempanah, Maxim Greiner, Daniel Pumpe, and Theo Steininger.

I want to thank my fellow PhD students for endless discussions on models, NIFTy, numerics, and everything else during countless coffee breaks and Friday afternoons. We built something amazing. Thank you all and good luck! Philipp Arras, Philipp Frank, Sebastian Hutschenreuter Ivan Kostyuk, Reimar Leike, and Natalia Porqueres.

I want to thank Martin Reinecke for his support in any numerical issue. What you do is magic.

Finally, I want to thank all the master students that joined and left during my time in the group. I want to especially thank the students I had the opportunity to co-supervise:
Gordian Edenhofer, Philipp Haim, Daniel Köglmayr, and Lukas Platz.

I had a great time at the MPA! Again, thank you all!