Diss. ETH no. 29627

RANDOMISING IN AND OVER SEISMOLOGY

A thesis submitted to attain the degree of

DOCTOR OF SCIENCES

(Dr. sc. ETH Zurich)

presented by

Lars Gebraad

M.Sc. in Applied Geophysics,
TU Delft, ETH Zurich, RWTH Aachen

born on 29.09.1995

citizen of The Netherlands

accepted on the recommendation of

Prof. Dr. A. Fichtner
Dr. A. Zunino
Prof. Dr. N. Linde

2024

I seek a thousand answers – I find but one or two

I maintain no discomfiture – my path again renewed

Summary

This thesis presents a collection of works centered on computational statistics in Bayesian seismology. Bayesian seismology interprets inverse problems in seismology as questions of inference, striving not to produce a single answer to an inverse problem, but to ascribe a probability to all possible solutions. To avoid evaluating every potential solution, or scenario, algorithms from computational statistics are necessary. However, the selection of appropriate algorithms is non-trivial, often demanding a deep understanding of the inverse problem at hand and knowledge of the potential algorithms available. This work focuses on the use of a specific algorithm, Hamiltonian Monte Carlo (HMC), and related variants. It's application to Bayesian seismology is studied from various perspectives.

Firstly, a general case study for appraising a computationally demanding inverse problem in seismology with HMC is presented. It is demonstrated that the use of the HMC algorithm enables successful consideration of Full-Waveform Inversion within a Bayesian inference framework, unlocking inference on parameters such as density, which have traditionally been poorly resolved.

This is followed by an effort to quantify the performance of algorithms on a given class of inverse problems. The collection of No-Free-Lunch algorithms precludes any single algorithm from being universally efficient, guiding the investigation into whether HMC and related algorithms might be optimal for a reduced set of relevant problems. While this is confirmed, the attempt is restricted by the curse of dimensionality, confining the analysis to inverse problems of limited dimensionality.

The expertise gained on these appraisal algorithms is subsequently distilled into an accessible and well-documented collection of open-source codes called HMCLab. This collection includes numerous didactic materials aimed at showcasing HMC and its variants to the general geophysicist. It covers various inverse problems and their Bayesian treatment, along with instructions on implementing inverse problems posed by the user.

Next, two approaches to writing efficient wavefield simulation codes are proposed. The first, an open-source package named psvWave, is a C++ written and Python accessible software designed to simulate 2D wavefields in parallel. The second approach demonstrates how to leverage modern unified chips using the Metal Shading Language to accelerate existing C++. Its ease of use is demonstrated on the psvWave package. Efficient wavefield modeling is integral to Bayesian seismology, as reducing computational costs can enable more extensive evaluations of wavefield-based inverse problems.

The thesis concludes with a report on multiple seismological field campaigns that are extensively documented using aerial and ground-based photogrammetry. In the three field campaigns, Structure-from-Motion methods were innovatively used to digitise the field sites. It is shown that these methods are accessible with limited resources and consumer electronics. The digitisation employing remotely operated drones enables safe surveying of hazardous fields and the ability to rapidly create meshes of structures and topography for wavefield simulations, while ground-based imagery offers a low-cost, low-risk alternative.

Zusammenfassung

Diese Dissertation stellt eine Sammlung von Arbeiten vor, die sich mit computergestützten Statistik in der Bayesschen Seismologie befassen. In der Bayesschen Seismologie werden inverse Probleme in der Seismologie als Fragen der Inferenz interpretiert, wobei nicht versucht wird, eine einzige Antwort auf ein inverses Problem zu finden, sondern allen möglichen Lösungen eine Wahrscheinlichkeit zuzuschreiben. Um nicht jede mögliche Lösung oder jedes mögliche Szenario auswerten zu müssen, werden Algorithmen aus der computergestützten Statistik benötigt. Die Auswahl geeigneter Algorithmen ist jedoch nicht trivial und erfordert oft ein gründliches Verständnis des vorliegenden inversen Problems und die Kenntnis der verfügbaren potenziellen Algorithmen. Diese Arbeit konzentriert sich auf die Verwendung eines bestimmten Algorithmus, Hamiltonian Monte Carlo (HMC), und verwandter Varianten. Seine Anwendung auf die Bayessche Seismologie wird aus verschiedenen Perspektiven untersucht.

Zunächst wird ein allgemeiner Fall zur Bewertung eines rechenintensiven inversen Problems in der Seismologie mit HMC vorgestellt. Es wird gezeigt, dass die Verwendung des HMC-Algorithmus es ermöglicht, die Full-Waveform-Inversion erfolgreich im Rahmen einer Bayesschen Inferenz zu betrachten und die Inferenz von Parametern wie der Dichte zu ermöglichen, die traditionell schlecht gelöst wurden.

Anschließend wird versucht, die Leistung der Algorithmen für eine bestimmte Klasse von inversen Problemen zu quantifizieren. Die Sammlung von No-Free-Lunch-Algorithmen schließt aus, dass ein einzelner Algorithmus universell effizient ist, was zu der Untersuchung führt, ob HMC und verwandte Algorithmen für eine reduzierte Menge relevanter Probleme optimal sein könnten. Obwohl dies bestätigt wird, wird der Versuch durch den Fluch der Dimensionalität eingeschränkt, der die Analyse auf inverse Probleme mit begrenzter Dimensionalität beschränkt.

Das aus diesen Algorithmen gewonnene Fachwissen wurde dann in eine zugängliche und gut dokumentierte Sammlung von Open-Source-Code, genannt HMCLab, destilliert. Diese Sammlung enthält zahlreiche didaktische Materialien, die darauf abzielen, HMC und seine Varianten dem allgemeinen Geophysiker vorzustellen. Es werden verschiedene inverse Probleme und ihre Bayessche Behandlung behandelt, zusammen mit Anleitungen zur Implementierung von durch den Benutzer gestellten inversen Problemen.

Anschließend werden zwei Ansätze für die Erstellung effizienter Wellenfeldsimulationscodes vorgeschlagen. Der erste, ein Open-Source-Paket namens psvWave, ist eine in C++ geschriebene und in Python zugängliche Software zur parallelen Simulation von

2D-Wellenfeldern. Der zweite Ansatz zeigt, wie moderne Unified Chips, die die Metal Shading Language verwenden, zur Beschleunigung von bestehendem C++ verwendet werden können. Die Benutzerfreundlichkeit wird anhand des psvWave-Pakets demonstriert. Eine effiziente Wellenfeldmodellierung ist ein integraler Bestandteil der Bayesschen Seismologie, da eine Reduzierung der Rechenkosten umfangreichere Auswertungen von wellenfeldbasierten inversen Problemen ermöglichen kann.

Die Dissertation schließt mit einem Bericht über mehrere seismologische Feldkampagnen ab, die mit Hilfe von Luft- und Bodenphotogrammetrie ausführlich dokumentiert wurden. In den drei Feldkampagnen wurden innovative Structure-from-Motion-Methoden zur Digitalisierung von Feldstandorten eingesetzt. Es wurde gezeigt, dass diese Methoden mit begrenzten Ressourcen und Unterhaltungselektronik zugänglich sind. Die Digitalisierung mit ferngesteuerten Drohnen ermöglicht die sichere Untersuchung von gefährlichen Feldern sowie die schnelle Vernetzung von Strukturen und Topografie für Wellenfeldsimulationen, während bodengestützte Bilder eine kostengünstige und risikoarme Alternative darstellen.

Samenvatting

Dit proefschrift presenteert een verzameling van werken gericht op computationele statistiek in Bayesiaanse seismologie. Bayesiaanse seismologie interpreteert inverse problemen in de seismologie als inferentievragen, waarbij niet gestreefd wordt naar een enkel antwoord op een invers probleem, maar naar het toekennen van een waarschijnlijkheid aan alle mogelijke oplossingen. Om te voorkomen dat elke mogelijke oplossing of scenario moet worden geëvalueerd, zijn algoritmen uit de computationele statistiek nodig. De selectie van geschikte algoritmen is echter niet triviaal en vereist vaak een grondig begrip van het inverse probleem in kwestie en kennis van de mogelijke beschikbare algoritmen. Dit werk richt zich op het gebruik van een specifiek algoritme, Hamiltonian Monte Carlo (HMC), en verwante varianten. De toepassing ervan op Bayesiaanse seismologie wordt vanuit verschillende perspectieven bestudeerd.

Ten eerste wordt een algemene casus gepresenteerd voor het beoordelen van een rekenintensief invers probleem in de seismologie met HMC. Aangetoond wordt dat het gebruik van het HMC-algoritme het mogelijk maakt om Full-Waveform Inversion met succes te beschouwen binnen een Bayesiaans inferentiekader, waardoor de inferentie van parameters zoals dichtheid, die traditioneel slecht werden opgelost, wordt ontsloten.

Dit wordt gevolgd door een poging om de prestaties van algoritmen te kwantificeren op een bepaalde klasse van inverse problemen. De verzameling No-Free-Lunch algoritmen sluit uit dat een enkel algoritme universeel efficiënt is, waardoor het onderzoek wordt geleid naar de vraag of HMC en verwante algoritmen optimaal zouden kunnen zijn voor een beperkte verzameling relevante problemen. Hoewel dit wordt bevestigd, wordt de poging beperkt door de vloek van de dimensionaliteit, waardoor de analyse beperkt blijft tot inverse problemen met een beperkte dimensionaliteit.

De expertise die is opgedaan met deze algoritmen is vervolgens gedistilleerd in een toegankelijke en goed gedocumenteerde verzameling van open-source codes genaamd HMCLab. Deze collectie bevat veel didactisch materiaal om HMC en zijn varianten uit een te zetten voor de algemene geofysicus. Het behandelt verschillende inverse problemen en hun Bayesiaanse behandeling, samen met instructies voor het implementeren van inverse problemen die door de gebruiker worden gesteld.

Vervolgens worden twee benaderingen voor het schrijven van efficiënte golfveldsimulatiecodes voorgesteld. De eerste, een open-source pakket genaamd psvWave, is C++ geschreven en Python toegankelijke software ontworpen om 2D golfvelden parallel te simuleren. De tweede benadering laat zien hoe moderne geunificeerde chips met behulp

van de Metal Shading Language gebruikt kunnen worden om bestaande C++ te versnellen. Het gebruiksgemak wordt gedemonstreerd op het psvWave pakket. Efficiënte golfveldsimulatie is een integraal onderdeel van Bayesiaanse seismologie, omdat het verlagen van de rekenkosten uitgebreidere evaluaties van inverse problemen op basis van golfvelden mogelijk maakt.

Het proefschrift sluit af met een verslag van meerdere seismologische veldcampagnes die uitgebreid zijn gedocumenteerd met behulp van fotogrammetrie vanuit de lucht en vanaf de grond. In de drie veldcampagnes werden Structure-from-Motion methoden innovatief gebruikt om de veldlocaties te digitaliseren. Er wordt aangetoond dat deze methoden toegankelijk zijn met beperkte middelen en consumentenelektronica. De digitalisering met behulp van op afstand bediende drones maakt veilig onderzoek van gevaarlijke velden mogelijk, evenals de mogelijkheid om snel mazen van structuren en topografie te maken voor golfveldsimulaties, terwijl beelden vanaf de grond een goedkoop alternatief met weinig risico's bieden.

Contents

Part I

The preliminaries

Chapter 1

Introduction

Science should be a narrative. Perhaps not when we consider academic publications in highly specialised journals, but when we're communicating our scientific story to the wider scientific community and beyond, in communication, outreach and of course talking to your non-academic peers. Narratives make the scientific facts palatable and digestible and help to explain our motivations. However, storytelling is not the only goal. Maintaining the process of sharing technical information with the peers in your field is a cornerstone to performing science. Academic, sterile publications have their place and help to ensure the utility of our works for other scientists.

Therefore, I've structured this work to embed my key academic publications within a broader narrative. This approach illustrates not only my personal motivations behind these works but also the developments and drivers from the seismological field that enabled and guided this research. I will also discuss the projects I've been involved in, where I believe my experience with algorithmic solutions made a significant contribution.

So, why did I dive into randomisation in Seismology? I have a fondness for algorithms, simulations, technology, and collaboration. I aimed to explore what I thought were promising avenues of possibility, perhaps making me the most opportunistic PhD student to submit a thesis to my professor. At the same time, I tried to apply my methods to every inverse problem I came across. In the end, my research did not follow the original research plan outlined in December 2019, but the journey nevertheless has become very insightful. In its totality it becomes a chaotic, random story that seems to reflect my own attention span.

This narrative draws on multiple threads from academic history which have enabled the present work. I will attempt to keep this history brief, because if you're reading this, there's a high chance you've written a similar introduction. The following sections will discuss Bayesian inference, computational statistics and seismic tomography.

1.1 Bayesian inference

Our ability to understand the world around us is fundamentally rooted in inference - the process of updating our beliefs based on observations and experiments. This is a critical procedure, as it provides a basis on which to quantify the scientific method. Updating our beliefs within an inference framework allows scientists to refine or reject them further.

The cornerstone of statistical inference is a work from 1763. In this year, the posthumously published seminal work of Thomas Bayes [Bayes, 1763] introduced the mathematical relationship that allows one to combine pieces of information in the form of probability distributions, forming the basis of quantitative belief updating. This work influenced the nascent field significantly enough to have the statistical analysis of beliefs propagated through models termed Bayesian inference.

An early adopter of Bayesian inference was Pierre-Simon Laplace, who independently developed an analogue to Bayes' Theorem in 1774 [Laplace, 1774] and used it in various studies. His principle VI in Laplace [1814], loosely translated from the original text, reads as follows:

> Each of the causes to which an observed event can be attributed is indicated with greater plausibility the more probable it is that, assuming this cause exists, the event will occur. The probability of the existence of any of these causes is therefore a fraction, with the numerator being the probability of the event resulting from that cause, and the denominator being the sum of similar probabilities related to all causes. If these various causes, considered a priori, are unequally probable, instead of the probability of the event resulting from each cause, the product of this probability and that of the cause itself must be used. This is the fundamental principle of this branch of the analysis of hazards, which consists of tracing events back to their causes.

This wording may be heavy compared to contemporary descriptions, but the key elements of Bayes' Theorem are recognisable. In modern notation, we simply write

$$p(A\,|\,B) = \frac{p(B\,|\,A)\,p(A)}{p(B)}. \tag{1.1}$$

In this notation, the distribution $p(\cdot)$ quantifies the probability of an event. In the context of inference, an event is a very broad concept. It can be the event of observing a specific outcome of an experiment, or the event that nature itself is in a specific state. In Laplace's interpretation, Bayes' Theorem attempts to link two events: a cause A, and an observation B. The distributions $p(A)$ and $p(B)$ are the prior beliefs of how likely events A and B are.

It wasn't until the twentieth century that major developments in the field of Bayesian inference led to the style of parameter estimation currently practised. Harold Jeffreys introduced the concept of a non-informative expectation for an event [Jeffreys, 1939], as a means to remain as unbiased as possible when evaluating hypotheses. Edwin Thompson Jaynes argued that in the case of competing distributions, the maximum entropy distribution [Jaynes, 1957] should be used. This is, in essence, the least committal distribution and can be considered a quantitative variant of Occam's Razor.

The final building block of our modern interpretation of Bayesian inference is Albert Tarantola's work on inverse problem theory. Tarantola's insights into the probabilistic nature of inverse problems have shaped the way contemporary inference is conducted. His early work showed us that to solve inverse problems means to [Tarantola and Valette, 1982], and his interpretation and notation were further solidified in his extensive textbook Tarantola [2005].

The concept of Bayesian inference is more than just a powerful tool in statistical analysis. As a method of testing hypotheses, it has found widespread usage across the natural sciences. Moreover, the mechanism of Bayesian inference itself has also been used to model how the human brain learns, wherein a person updates their beliefs in accordance with the rules of probability [e.g., Jaynes, 1988, Angela, 2007, Fletcher and Frith, 2009, Clark, 2013]. For all its applicability, it might be worth regarding Bayesian inference not as a method, but a philosophy.

1.2 Computational statistics

A gap of roughly two centuries separates the inception of Bayesian inference from the advent of the modern computer. However, it is the computer that finally allows Bayesian inference to live up to its potential. Pure mathematics alone turns out to be inadequate for solving arbitrary inference problems, and ways to create numerical approximations of distributions are needed.

In 1948, the first programmable multi-purpose computer was soon to be leveraged by titans of twentieth-century particle physics. At Los Alamos National Laboratory, the site of the recently declassified Manhattan project, the work of Nicholas Metropolis, Stanislaw Ulam, and John von Neumann birthed the field of computational statistics. Their work led to the development of the Metropolis algorithm [Metropolis et al., 1953], later refined by Wilfred Keith Hastings into the Metropolis-Hastings algorithm [Hastings, 1970]. This method and its many derivatives have become ubiquitous in modern computational sciences. This early application of computational power to statistical problems continues to influence scientific research, including the present work.

In 1987, Simon Duane and colleagues developed the Hybrid Monte Carlo algorithm Duane et al. [1987], driven by the challenges in lattice field theory. Lattice field theory is a computational method in theoretical physics where spacetime is discretised and fields are defined at these lattice points, enabling the simulation of quantum mechanics through the sampling of possible field states. As the number of particles and the complexity of fields increased, generating new states became prohibitively expensive. Their approach merged the randomness of the Metropolis-Hastings algorithm with the deterministic behavior of simulating the equations of motion [Duane, 1985, shows how this was used to compute intractable integrals as well], into a hybrid algorithm. Over the years, the Hybrid Monte Carlo algorithm has become known as the Hamiltonian Monte Carlo (HMC) algorithm, as the motion of the particles can be solved by Hamilton's Equations. Some of the notable works on HMC include Neal's extensive review [Neal, 2011], the extension of the dynamics to curved space to integrate the distribution's second-order information [Girolami and Calderhead, 2011], and the No U-Turn sampler [Hoffmann and Gelman, 2014].

Recent advances in computational statistics have seen the development of parallelised MCMC [Geyer, 1991], transdimensional sampling [Green, 1995], sequential Monte Carlo [Liu and Chen, 1998], and variational inference methods [e.g. Liu and Wang, 2019]. As the range of computational methods has expanded dramatically, it could be argued that the most significant recent development is the emergence of the "No-Free-Lunch" theorems [Wolpert and Macready, 1997]. These theorems state that no single method can serve as a silver bullet. Therefore, assessing the attributes of a specific inference problem and subsequently choosing the appropriate appraisal method remains, for the time being, a skill exclusive to initiated experts.

1.3 Seismic tomography

One of the key competences of Earth scientists is to unravel the state of the Earth by analysing observations. Akin to the the process Laplace described in his principle VI, the observations of events are linked to causes through inference —- a crucial task, given that many processes of the Earth occur beyond our direct observation. In seismology, this process is further refined to the analysis of seismic waves.

Seismic waves, both those that occur naturally and anthropogenically, carry a wealth of information. Over the last century and a half, the careful study of the wavefield from various sources has enabled seismologists to distinguish multiple modes of vibration and travel paths, leading to the distinction of various seismic phases in observations. These different phases carry knowledge about both the originating processes and the subsurface they travelled through. Seismic tomography primarily aims to decode this subsurface imprint on seismic phases, attempting to construct approximations of the Earth and ultimately perform inferences about the state of the Earth based on seismic observations.

Somewhere around the middle of the 19th century the word seismology was invented, and the field came with it. It marked the start of scientists using seismic phases to discover key properties of the inner structure of the Earth. These very early studies often used the time-of-flight between setting off a source and recording it at a receiver [Mallet and Mallet, 1858], to infer the wave speed of the subsurface.

Building on the basic concept of a seismic phase's time-of-flight, ray tomography [Dziewoński et al., 1977, Aki et al., 1977, Spakman et al., 1993, Hilst et al., 1997, Grand et al., 1997, Gorbatov and Kennett, 2003] was one of the first practical approaches used to probe the Earth's deep interior. Advances in both computational resources and computational seismology subsequently facilitated the development of volumetric sensitivities in seismic phase analysis [finite frequency tomography, see Yomogida, 1992, Dahlen et al., 2000, Friederich, 2003, Yoshizawa and Kennett, 2004, Sigloch et al., 2008] and the comprehensive treatment of wave propagation physics [Full-Waveform Inversion (FWI), see Bamberger et al., 1977, 1982, Lailly, 1983, Tarantola, 1984, Gauthier et al., 1986].

Although the term 'FWI' might imply that this is the final iteration of tomography, practical constraints have so far limited the fitting of a true full waveform. The computational cost associated with simulating the entire bandwidth of data has thus far constrained seismologists to consider relatively low frequencies at any scale. Despite the computational limitations of FWI, its applications have ranged from seismic exploration [Sirgue et al., 2010, Prieux et al., 2013, Warner et al., 2013], in regional tomographies [Chen et al., 2007, Fichtner et al., 2009, Tape et al., 2010, Krischer et al., 2018], to the entire Earth [French and Romanowicz, 2014, Bozdağ et al., 2016, Fichtner et al., 2018a, Thrastarson et al., 2022]. Current efforts are focusing on reducing the computational cost for FWI and other simulation-based methods [van Herwaarden et al., 2020].

1.4 Motivation

If I had to distill my research into a single subject, it would be the study of the HMC algorithm; its workings, its applicability to seismology and other natural sciences, as well as its extendability. I've come to learn that it is a useful algorithm, that can be made efficient for a wide class of problems we encounter in our research performed at Seismology and Wave Physics. Like any algorithm, it is not, and can not be, a silver bullet. This is demonstrated by other works that perform Bayesian FWI, such as Thurin et al. [2019], Huang et al. [2020], Guo et al. [2020], Zhang and Curtis [2021] and Zhang et al. [2023a]. However, our continued work with the algorithm does allow for the practical usage of it, naturally disproportionally so compared to algorithms we do not research actively.

The primary aim of my doctoral research was to explore the applicability of HMC to the FWI problem, ranging from the synthetic scale to global waveform tomography. Despite seismologists probing the Earth with seismic waves for many years, the field has yet to construct satisfactory estimations of density and attenuation. These parameters are heavily regularised in deterministic FWI across the scales, and as a result the derived models are not widely trusted. The development of Bayesian methods for assessing the information that data carries about these parameters allows for a more careful inference about them.

However, as this work will demonstrate, genuine understanding cannot be achieved without exploration. In the spirit of Bayesian thinking, I found value in traversing the landscape of Bayesian seismology. While the development of global probabilistic models did not materialise as initially hoped, owing to unforeseen limitations in HMC scaling, or perhaps fortuitous diversions, my research evolved into a series of studies centered around Bayesian inference, significant in their own right. The broadened understanding of HMC and other gradient-based inference algorithms is crucial for the future of the field, as many of the inverse problems that seismologists encounter are strongly non-linear, ill-posed, and of high dimensionality. Studying the FWI problem using HMC provides key insights that also have implications for the deterministic approach to this inverse problem. The analysis of the problem is further accelerated by developing more performant approaches to simulating the wave equation. Lastly, by creating an accessible software package for carrying out Bayesian seismology, this work provides a springboard for more seismologists to incorporate Bayesian methods into their research. The end result is a thesis on computational statistics in seismology, with a flourish of small projects in other branches of seismology.

1.5 Outline

This thesis is structured into four parts. Part I, titled "The preliminaries", comprises simply this introduction. Part II, "First author works", consists of separate first author works I have produced throughout my doctoral studies. These chapters are included verbatim, following the style of a cumulative thesis. Chapter 2 presents the application of HMC to the FWI problem. Chapter 3 describes the software package HMCLab, developed collaboratively by Andrea Zunino and myself in two distinct programming languages. Chapter 4 attempts to provide a numerical estimate of how cost-effective inference algorithms can be, providing numerical insights into the No-Free-Lunch theorems. Chapter 5 outlines an open-source parallel implementation of 2D elastic FWI, facilitating rapid prototyping for geophysical inversion techniques. Chapter 6 exploits recent hardware advances in consumer electronics to demonstrate how numerical simulations may be accelerated on portable hardware. Lastly, Chapter 7 details a series of field campaigns designed to support different branches of seismology. The details on the publication and peer-review status are provided at the beginning of each chapter. Part III, "Scientific collaborations", highlights productive collaborations with individuals outside of my research group. Chapter 8 details collaborations within seismology, and Chapter 9 outlines those outside of geophysics. As for these works I am not the first author, I will reproduce the relevant abstracts of the works, and highlight my relevant contribution. Finally, Part IV, "Synthesis", synthesises all aspects of my research, offering insights into promising research areas and providing concluding remarks.

Part II

First author works

Chapter 2

Bayesian Full-Waveform Inversion using Hamiltonian Monte Carlo

Chapter published as L. Gebraad, C. Boehm, and A. Fichtner. Bayesian Elastic Full-Waveform Inversion Using Hamiltonian Monte Carlo. *J. Geophys. Res.*, 125(3):e2019JB018428, Feb. 2020. doi: 10.1029/2019JB018428.

Abstract

We present a proof of concept for Bayesian elastic full-waveform inversion in 2D. This is based on (1) Hamiltonian Monte Carlo sampling of the posterior distribution, (2) the computation of misfit derivatives using adjoint techniques, and (3) a mass matrix tuning of the Hamiltonian Monte Carlo algorithm that accounts for the different sensitivities of seismic velocities and density. We apply our method to two synthetic end-member scenarios with different dimension D that are particularly relevant in the context of full-waveform inversion: low-dimensional models ($D < 100$) with potentially large variations in material parameters, and high-dimensional models ($D > 30'000$) describing smaller-scale variations of lower amplitude relative to some background. For both end members, the Hamiltonian Monte Carlo sampling reliably recovers important aspects of the posterior, including means, covariances, skewness, as well as 1D and 2D marginals. Depending on the strength of material variations, the posterior can be significantly non-Gaussian. This suggests to replace local methods for uncertainty quantification based on Gaussian assumptions by proper sampling of the posterior. In addition to P-wave and S-wave velocity, the sampling provides constraints on density structure that are free from subjective regularization artifacts.

2.1 Introduction

2.1.1 Full-waveform inversion

While having been conceptualized already in the late 1970's and early 1980's [Bamberger et al., 1977, 1982, Lailly, 1983, Tarantola, 1984, Gauthier et al., 1986], practical full-waveform inversion (FWI) is a comparatively recent addition to the seismological toolbox. Based on numerical wave propagation through potentially complex Earth models, it is the natural extension of ray tomography [Dziewoński et al., 1977, Aki et al., 1977, Spakman et al., 1993, Hilst et al., 1997, Grand et al., 1997, Gorbatov and Kennett, 2003] and finite-frequency tomography [Yomogida, 1992, Dahlen et al., 2000, Friederich, 2003, Yoshizawa and Kennett, 2004, Sigloch et al., 2008]. In recent years, successful applications of FWI have been reported in seismic exploration [Sirgue et al., 2010, Prieux et al., 2013, Warner et al., 2013], in regional studies [Chen et al., 2007, Fichtner et al., 2009, Tape et al., 2010, Krischer et al., 2018], and for the whole Earth [French and Romanowicz, 2014, Bozdağ et al., 2016, Fichtner et al., 2018a].

The non-linearity of the inverse problem, i.e. the non-linear relation of waveform fit with respect to the medium parameters, can in principle be handled elegantly by Monte Carlo sampling [Mosegaard and Tarantola, 1995, Sambridge and Mosegaard, 2002]. However, the high-dimensionality of the model space paired with the computational costs of the forward problem, have so far limited its applicability to low-dimensional special cases [Käufl et al., 2013, Afanasiev et al., 2014, Kotsi et al., 2018, Hunziker et al., 2019, Visser et al., 2019]. For the same reasons, resolution and uncertainty analysis in FWI is still mostly local, making the assumption of a Gaussian posterior centered near a hopefully meaningful approximation of the maximum-likelihood model [Fichtner and Trampert, 2011, Bui-Thanh et al., 2013, Fichtner and Leeuwen, 2015, Liu et al., 2019a,b].

With this work, we explore non-linearity and uncertainty quantification in FWI with a high-dimensional model space using a sampling method that has recently been popularised in geophysics, known as Hamiltonian Monte Carlo (HMC) [Duane et al., 1987, Betancourt, 2017, Sen and Biswas, 2017, Fichtner et al., 2018b, Fichtner and Simute, 2018]. Exploiting derivative information, HMC may solve high-dimensional problems where widely-used variants of the Metropolis-Hastings algorithm [Chib and Greenberg, 1995] tend to fail. This work seeks to extend previous Bayesian FWI studies by making no assumptions on the characteristics of the posterior as well as not to perform dimensionality reduction of it, while accelerating sampling by using the HMC algorithm.

2.1.2 Objectives and outline

Our primary objective is a proof of principle that HMC can be successfully applied to two end-member cases of 2D elastic FWI: (1) the non-linear search for coarse and *a priori* poorly known models that may serve as plausible starting points for subsequent spatial refinements, and (2) the probabilistic inversion for smaller-scale variations within more limited bounds, set, for instance, by the previous coarse-scale inversion combined with geologic prior knowlege.

We begin, in section 2.2, with a summary of the necessary theoretical background, including Bayesian inference, Markov chain Monte Carlo, Hamiltonian Monte Carlo, and numerical seismic wave propagation. In section 2.3, we consider comparatively low-dimensional FWI, with 75 free parameters in total. This corresponds to the end-member case (1) described above. Section 2.4 is focused on high-dimensional problems. Specifically, we show that a single HMC chain can provide uncertainty information for $> 30'000$ material parameters locally, while the algorithm also allows us to globally explore a posterior for as many parameters. Finally, in section 2.5, we provide a detailed discussion of advantages and drawbacks of the method. Also, we indicate possible improvements that are likely to increase the efficiency of the sampler, possibly allowing it to address higher-dimensional 3D problems in the future.

2.2 Theoretical background

We consider 2D elastic, isotropic models with density, ρ, S-wave velocity, v_s, and P-wave velocity, v_p, as free parameters. The HMC algorithm can be readily applied to FWI, as gradients of the misfit function can be conveniently computed using adjoint techniques [Lions, 1968, Tarantola, 1988, Liu and Tromp, 2006, Fichtner et al., 2006a,b, Plessix, 2006]. There are, however, important technical details concerning target models, algorithm tuning, and model priors that affect the efficiency of the sampling. The following subsections comprise a short introduction to the theory of basic HMC sampling, and the synthesis of FWI and HMC. For a complete theoretical overview of and possible extensions to HMC we refer to Neal [2011] and Betancourt [2017]. Summaries of FWI theory can be found in Virieux and Operto [2009], Fichtner [2011] and Liu and Gu [2012].

2.2.1 Bayesian inference and Markov Chain Monte Carlo

To set the stage and to establish basic notation, we begin with a brief recapitulation of Bayesian inference [Jaynes, 2003, Tarantola, 2005]. For this we define $\mathbf{m}$ as an n-dimensional vector containing values of the discretized material properties ρ, v_s and v_p. Information on $\mathbf{m}$ available prior to the analysis of any data is decribed by the probability density function (PDF) $p(\mathbf{m})$. Similarly, the prior probability of observing data $\mathbf{d}_{\text{obs}}$ given a specific $\mathbf{m}$ is encoded by a conditional PDF $p(\mathbf{d}_{\text{obs}}|\mathbf{m})$, usually referred to as the likelihood function. Both priors, $p(\mathbf{m})$ and $p(\mathbf{d}_{\text{obs}}|\mathbf{m})$, can be combined into the posterior PDF $p(\mathbf{m}|\mathbf{d}_{\text{obs}})$ using Bayes' theorem,

$$p(\mathbf{m}|\mathbf{d}_{\text{obs}}) = \frac{p(\mathbf{d}_{\text{obs}}|\mathbf{m})p(\mathbf{m})}{p(\mathbf{d}_{\text{obs}})} . \tag{2.1}$$

The evidence $p(\mathbf{d}_{\text{obs}}) = \int p(\mathbf{d}_{\text{obs}}|\mathbf{m})p(\mathbf{m})\,d\mathbf{m}$ normalizes the posterior $p(\mathbf{m}|\mathbf{d}_{\text{obs}})$, which contains all possibly available information on models $\mathbf{m}$ given observations $\mathbf{d}_{\text{obs}}$. In this study, the focus of the Bayesian inference is the unnormalized density, and therefore we do not consider the evidence further. It is, however, a relevant quantity which tests the relative validity of the assumptions made in the inference [Sambridge et al., 2006]. The likelihood function is typically written as exponential of a misfit function $\chi(\mathbf{m},\mathbf{d}_{\text{obs}})$,

$$p(\mathbf{d}_{\text{obs}}|\mathbf{m}) \propto e^{-\chi(\mathbf{m},\mathbf{d}_{\text{obs}})} . \tag{2.2}$$

The misfit function serves as a measure of fit between observed data $\mathbf{d}_{\text{obs}}$ and synthetic data $\mathbf{d}$ computed from $\mathbf{m}$ via the solution of the forward modelling equations.

The posterior $p(\mathbf{m}|\mathbf{d}_{\text{obs}})$ is an n-dimensional PDF that is usually not known explicitly. Therefore, quantities of interest, such as means, (co)variances or marginal PDFs, are typically approximated by Markov chain Monte Carlo (MCMC) sampling of $p(\mathbf{m}|\mathbf{d}_{\text{obs}})$. All MCMC methods suffer from the curse of dimensionality in some form. Widely used variants of the Metropolis-Hastings algorithm [Metropolis et al., 1953, Hastings, 1970, Chib and Greenberg, 1995, Mosegaard and Tarantola, 1995], for instance, require increasingly smaller step sizes as the dimension n grows, in order to ensure reasonable acceptance rates of proposed models. As a consequence, model space exploration is slow, and subsequent samples are highly correlated. Hamiltonian Monte Carlo (HMC), outlined in the following paragraphs, has been designed to overcome this problem, and to enable long-distance moves through model space while maintaining high acceptance rates [Duane et al., 1987, Neal, 2011, Betancourt, 2017].

2.2.2 Hamiltonian Monte Carlo

Originally developed for molecular dynamics under the name hybrid Monte Carlo [Duane et al., 1987], Hamiltonian Monte Carlo (HMC) is now commonly used for the subset of sampling problems where gradients of the posterior $p(\mathbf{m}|\mathbf{d}_{\text{obs}})$ with respect to the model parameters $\mathbf{m}$ are easy to compute. The cost of generating independent samples with HMC under increasing dimension n grows as $\mathscr{O}(n^{5/4})$ [Neal, 2011], whereas it grows as $\mathscr{O}(n^2)$ for standard Metropolis-Hastings [Creutz, 1988].

HMC constructs a Markov chain over an arbitrary n-dimensional probability density function $p(\mathbf{m})$ using classical Hamiltonian mechanics [Landau and Lifshitz, 1976]. The algorithm regards the current state $\mathbf{m}$ of the Markov chain as the location of a physical particle in n-dimensional space $\mathbb{M}$. It moves under the influence of a potential energy, U, which is defined as

$$U(\mathbf{m}) = -\ln p(\mathbf{m}) . \tag{2.3}$$

In the case of a Gaussian probability density p, the potential energy U is up to an additive constant equal to the least-squares misfit $\chi(\mathbf{m})$. To complete the physical system, the state of the Markov chain needs to be artificially augmented with momentum variables $\mathbf{p}$ for every dimension and a generalized mass for every dimension pair. The collection of resulting masses are contained in a positive definite mass matrix $\mathbf{M}$ of dimension $n \times n$. The momenta and the mass matrix define the kinetic energy of a model as

$$K(\mathbf{p}) = \frac{1}{2}\mathbf{p}^T\mathbf{M}^{-1}\mathbf{p} . \tag{2.4}$$

In the HMC algorithm, the momenta $\mathbf{p}$ are drawn randomly from a multivariate Gaussian with covariance matrix $\mathbf{M}$. The location-dependent potential and kinetic energies constitute the total energy or Hamiltonian of the system,

$$H(\mathbf{m},\mathbf{p}) = U(\mathbf{m}) + K(\mathbf{p}). \tag{2.5}$$

Hamilton's equations

$$\frac{d\mathbf{m}}{d\tau} = \frac{\partial H}{\partial \mathbf{p}}, \qquad \frac{d\mathbf{p}}{d\tau} = -\frac{\partial H}{\partial \mathbf{m}} \tag{2.6}$$

determine the position of the particle as a function of the artificial time variable τ. We can simplify Hamilton's equations using the fact that kinetic and potential energy depend only on momentum and location, respectively,

$$\frac{d\mathbf{m}}{d\tau} = \mathbf{M}^{-1}\mathbf{p}, \qquad \frac{d\mathbf{p}}{d\tau} = -\frac{\partial U}{\partial \mathbf{m}}. \tag{2.7}$$

The combination of a model $\mathbf{m}$ and kinetic energy $\mathbf{p}$ is called a state, and is uniquely linked to a potential, kinetic and total energy. Evolving model $\mathbf{m}$ and kinetic energy $\mathbf{p}$ over time τ generates another possible state of the system with new position $\tilde{\mathbf{m}}$, momentum $\tilde{\mathbf{p}}$, potential energy $\tilde{U}$, and kinetic energy $\tilde{K}$. Due to the conservation of energy, the Hamiltonian is equal in both states. Successively drawing random momenta and evolving the system generates a distribution of the possible states of the system. Thereby, HMC samples the joint momentum and model space, referred to as phase space. As we are not interested in the momentum component of phase space, we marginalize over the momenta by simply dropping them. This results in samples drawn from the distribution $\exp(-U(\mathbf{m}))$, i.e. $p(\mathbf{m})$.

If one could solve Hamilton's equations exactly, every proposed state would be a valid sample of $p(\mathbf{m})$. Since Hamilton's equations for non-linear forward models cannot be solved analytically, the system must be integrated numerically. Suitable integrators are symplectic, meaning that time reversibility, phase space partitioning and volume preservation are satisfied [Neal, 2011, Fichtner and Zunino, 2019]. However, the Hamiltonian is generally not preserved exactly when explicit time-stepping schemes are used. In this work, we employ the leapfrog method as described in Neal [2011]. As the Hamiltonian is not preserved, the time evolution generates samples not exactly proportional to the original distribution. A Metropolis-Hastings correction step is therefore applied at the end of numerical integration.

In summary, samples are generated starting from a random model $\mathbf{m}$ in the following way:

1. Propose momenta $\mathbf{p}$ according to the Gaussian with mean $\mathbf{0}$ and covariance $\mathbf{M}$;

2. Compute the Hamiltonian H of model $\mathbf{m}$ with momenta $\mathbf{p}$;

3. Propagate $\mathbf{m}$ and $\mathbf{p}$ for some time τ to $\tilde{\mathbf{m}}$ and $\tilde{\mathbf{p}}$, using the discretized version of Hamilton's equations and a suitable numerical integrator;

4. Compute the Hamiltonian $\tilde{H}$ of model $\tilde{\mathbf{m}}$ with momenta $\tilde{\mathbf{p}}$;

5. Accept the proposed move $\mathbf{m} \rightarrow \tilde{\mathbf{m}}$ with probability

$$p_{\text{accept}} = \min\left(1, \exp(H - \tilde{H})\right) . \tag{2.8}$$

6. If accepted, use (and count) $\tilde{\mathbf{m}}$ as the new state. Otherwise, keep (and count) the previous state. Then return to 1.

The main factor influencing the acceptance rate of the algorithm is the conservation of energy, H, along the trajectory. If the leapfrog integration has too large time steps, or the gradients of the misfit function are computed incorrectly (e.g., by badly discretizing the forward model), H is less well conserved, and the algorithm's acceptance rate decreases.

The main cost of HMC, compared to other MCMC samplers, is the computation of the gradient $\partial U / \partial \mathbf{m}$ at every step in the leapfrog propagation. When gradients can be computed easily, HMC can provide improved performance for two reasons: (1) the reduced cost of generating independent samples, that is, the avoidance of random-walk behaviour [Neal, 2011], and (2) the better scaling of HMC with increasing dimension [Creutz, 1988, Neal, 2011].

The tuning parameters in HMC are simulation time τ and the mass matrix $\mathbf{M}$. HMC has the potential to inject additional knowledge about the distribution p via the mass matrix in order to enhance convergence significantly. At the same time, the abundance of tuning parameters also creates potential for choosing inefficient settings, leading to suboptimal convergence. Fichtner et al. [2018b] and Fichtner and Zunino [2019] both illustrate how to create relevant mass matrices for tomographic inverse problems.

We adapt the specific tuning strategy for the mass matrix in this study depending on the target, as illustrated in the following sections. However, for all targets we choose the size of the discrete time steps empirically such that the acceptance rate is close to the optimum of 65 % [Neal, 2011]. This typically results in needing approximately 10 leap-frog steps per proposal, i.e. requiring this many forward and adjoint solves per proposal.

2.2.3 Numerical seismic wave propagation

Our inversions target 2D vertical cross sections of isotropic wave velocities and density. For this, we consider the P-SV wave system, written in velocity-stress formulation as

$$\begin{aligned}
\partial_t v_x &= \rho^{-1} \left(\partial_x \tau_{xx} + \partial_z \tau_{xz}\right) , && (2.9)\\
\partial_t v_z &= \rho^{-1} \left(\partial_x \tau_{xz} + \partial_z \tau_{zz}\right) , && (2.10)\\
\partial_t \tau_{xx} &= (\lambda + 2\mu)\, \partial_x v_x + \lambda \partial_z v_z , && (2.11)\\
\partial_t \tau_{zz} &= (\lambda + 2\mu)\, \partial_z v_z + \lambda \partial_x v_x , && (2.12)\\
\partial_t \tau_{xz} &= \mu \left(\partial_z v_x + \partial_x v_z\right) && (2.13)
\end{aligned}$$

with the velocity vector (v_x, v_z), the stress tensor components τ_{xx}, τ_{zz} and τ_{xz}, the Lamé coefficients λ and μ, and density ρ. All quantities are a function of position $\mathbf{x} = (x, z)$.

We discretize these differential equations using the fourth-order variant of the staggered-grid finite-difference scheme developed by Virieux [1986]. As free parameters we use the P-wave velocity $v_p = \sqrt{(\lambda + 2\mu)/\rho}$, the S-wave velocity $v_s = \sqrt{\mu/\rho}$, and density, ρ.

For the computation of sensitivity kernels, we use the adjoint method [Lions, 1968, Tarantola, 1988, Liu and Tromp, 2006, Fichtner et al., 2006a,b, Plessix, 2006]. Subsequently, we project the kernels onto the basis functions, used to represent the elastic medium. This yields the gradient needed in the HMC algorithm.

2.2.4 Waveform misfit and tempering

An important choice in the solution of an inverse problem is the misfit used to quantify differences between observed data $\mathbf{d}_{\text{obs}}$ and synthetic data $\mathbf{d}(\mathbf{m})$. It determines, among other things, the extent to which different parameters can be resolved. In the interest of simplicity, we choose the L_2 waveform difference,

$$\chi_{L_2}(\mathbf{m}) = \frac{1}{2}\sum_{\text{i}} \left(\frac{d_{i,\text{obs}} - d_i(\mathbf{m})}{\sigma_i} \right)^2 , \tag{2.14}$$

where the indices i denote time samples. The scalars σ_i are the standard deviations per data point, corresponding to a diagonal data covariance matrix. More complex and in real-data applications more meaningful data covariances can be used; but this is beyond the scope of this synthetic study.

The data variances σ_i^2 can be parameters of the inversion, which may be estimated by hierarchical inversion [Malinverno and Briggs, 2004, Bodin et al., 2012]. They should, however, not be changed to make the HMC sampler behave in a specific way [Scales and Snieder, 1997]. Choosing an identical variance $\sigma_i^2 = \sigma^2$ for all data points, makes σ^2 behave analogously to the temperature parameter T in tempering [Geyer and Thompson, 1995, Sambridge et al., 2013]. A tempered distribution p_T is constructed from the original distribution p as

$$p_{\text{T}}(\mathbf{m}) \quad = \quad p(\mathbf{m}\,|\,\mathbf{d}_{\text{obs}})^{1/T} \propto \exp\left(-\frac{\chi}{T}\right) \tag{2.15}$$

The variable T determines the temperature of the tempered distribution p_T. We analyse the impact of changing temperature as a proxy for changing data variance (i.e., noise levels) in section 2.4.

Though the L_2 waveform difference (2.14) has been used traditionally in FWI studies [Bamberger et al., 1982, Tarantola, 1984, Gauthier et al., 1986, Igel et al., 1996], other, and in practice often more suitable, misfits may be used [Luo and Schuster, 1991, Gee and Jordan, 1992, Fichtner et al., 2008, Leeuwen and Mulder, 2010, Métivier et al., 2016]

HMC does not necessarily require a globally convex misfit function. However, the absence of local minima generally improves convergence, sometimes at the expense of reduced resolution. For an illustration of the behaviour of HMC in multimodal posteriors and possible mitigations, we refer to Neal [2011]. The L_2 waveform misfit (2.14) is dominated by large-amplitude S-waves, whereas lower-amplitude P-waves have a smaller influence. This property will be reflected in the inversion results.

2.2.5 Prior information

All model priors used in this work are uniform distributions within certain bounds. The width of the prior reflects two end-member scenarios and objectives of Bayesian FWI: (1) To find a range of admissible initial models for deterministic FWI, a small number of model parameters will be used with a weak prior, that is, a broad uniform distribution. This mode of operation is related to global optimization in the potential presence of multiple local minima. (2) In contrast, a large number of model parameters with deviations from a well-known background is needed to constrain small-scale deviations. Their priors will be stronger, that is, the uniform distribution will be comparatively narrow. As with the misfit and data covariance model, the uniform prior was chosen for its simplicity. In practice, more complex and meaningful priors can and should be used, which for example introduce correlations (i.e. smoothness constraints on admissible models). Examples of constructing geologically realistic priors can be found in e.g. Linde et al. [2015].

In the HMC sampling, any model with non-zero prior likelihood could be proposed. Consequently, the finite-difference simulations must be numerically stable for any model admitted by the prior. This requires a conservative choice of temporal and spatial sampling for the finite-difference simulations, dependent on the width (bounds) of the prior. In this work, all prior realisations still result in numerically stable wavefield simulations. Additionally, the boundaries of the prior have to be evaluated explicitly when solving Hamilton's equations, as these are not differentiable. As soon as the probability of the prior goes to zero outside of the prior, the particle encounters a potential wall. Thus, when the model passes a boundary of the uniform distributions in one dimension, it is perfectly reflected across this boundary back into the relevant part of model space.

2.3 Low-dimensional model space sampling

We start with the case of a low-dimensional model space where Earth structure is *a priori* poorly known and represented by few basis functions. This scenario is intended to mimic the situation where plausible but yet simple initial models for a deterministic FWI need to be found, often in the presence of limited data (in spatial coverage, bandwidth, or both). The priors used are relatively wide with respect to the next section, with space-invariant uniform distributions in the interval 2000 ± 1000 m/s for v_p, 800 ± 400 m/s for v_s, and 1500 ± 500 kg/m^3 for ρ.

In the interest of simplicity and easy visualization, we consider two checkerboard patterns, shown in Fig. 2.1. Both checkerboard models are embedded in the same physical domain, and they share identical source-receiver setups. The domain is 125 m by 125 m wide, with absorbing boundaries at all sides except the top, where a free surface is implemented. Waves from two moment tensor sources are recorded by six receivers. The source time function is a Ricker wavelet with a central frequency of 50 Hz and is assumed known. The number of checkerboard blocks is 5×5. Having three physical parameters, this corresponds to model space dimension of 75.

The checkerboards, used to compute artificial data, differ in one aspect: The anomaly strength is either 10 % or 25 % relative to the background, and for all model parameters,

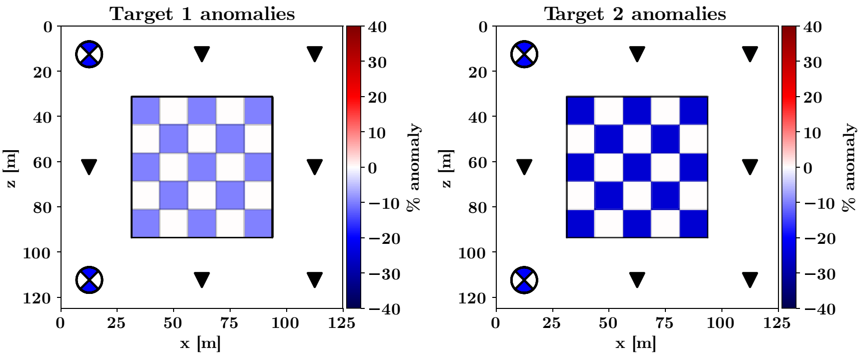


Figure 2.1: Checkerboard patterns in v_p, v_s, and ρ. The source-receiver setup and the domain dimensions are identical, with receivers (▼) present both at depth and at the surface. The source mechanisms, represented by beachballs, are equally oriented.

v_p, v_s, and ρ. Variable anomaly strength allows us to investigate the effect of increasing non-linearity on algorithm performance and the posterior distributions. The weaker perturbations are referred to as checkerboard 1, whereas the stronger perturbations are referred to as checkerboard 2. In all chequerboard inversions, the assumed noise levels are equal, but no actual noise is added.

2.3.1 Tuning strategy and starting models

While the mass matrix $\mathbf{M}$ can in principle be any positive definite matrix, its design determines the effectiveness of HMC by controlling the relative speed of the particle (i.e., the model) for every separate dimension in model space. Ideally, all dimensions are explored equally fast. For linear inverse problems (with Gaussian prior and Gaussian noise model), this can be achieved with a mass matrix that equals the inverse posterior covariance [Fichtner et al., 2018b].

The posterior covariance matrix is, by definition, not known a priori, and our inverse problem is not linear. Therefore, we employ a trial-and-error tuning strategy based on the acceptance rate and trace plots of preliminary runs of the Markov chain. For this, we first simplify the mass matrix (with dimensions 75×75) to three tuning parameters, with one mass for each parameter set (m_{v_p}, m_{v_s}, and m_ρ),

$$\mathbf{M} = \begin{bmatrix} m_{v_p}\mathbf{I}_{25} & & \\ & m_{v_s}\mathbf{I}_{25} & \\ & & m_\rho \mathbf{I}_{25} \end{bmatrix}, \tag{2.16}$$

where $\mathbf{I}_{25}$ stands for the 25×25 identity mass matrix. An added benefit is that the mass matrix is diagonal, which greatly accelerates the computation of kinetic energy (2.4) and the proposal of momenta.

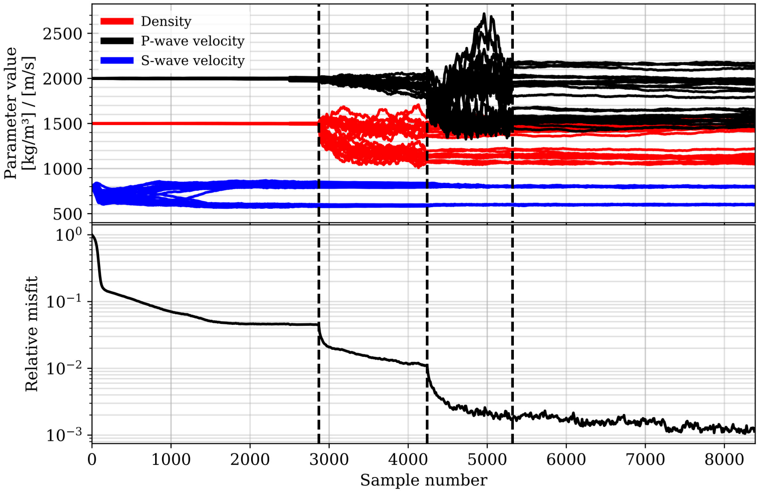


Figure 2.2: Iterative tuning of the mass matrix for checkerboard target 2 as given in Fig. 2.1. Top: Trace plot of an exploratory Markov chain. Each curve corresponds to one of the 75 model parameters. Coloring represents the parameter class. Bottom: The corresponding waveform misfit during this chain. Every vertical line represents an update of the three scalar masses. The parameters with the strongest impact on the misfit stabilize first.

The tuning of m_{v_p}, m_{v_s} and m_ρ is illustrated in Fig. 2.2. We start with a mass matrix equal to the identity matrix. During the first few hundred samples, the values of v_s stabilize into two groups, as expected for the checkerboard. However, the values of v_p and ρ hardly move, suggesting that their masses are too large. Therefore, after around 2900 samples, we decrease m_{v_p} and m_ρ. This leads to larger movement of the respective parameters; and after around 4000 samples also the values of ρ have stabilized. To further increase the movement of v_p, we again decrease m_{v_p}. This sequence can be repeated several times. In this specific example, 3 iterations were sufficient to obtain reasonable values for m_{v_p}, m_{v_s} and m_ρ.

The initial model for all inversions is chosen to be homogeneous. After the preliminary tuning chain, the samples after the last mass matrix update are used to supply initial models for subsequent Markov chains. These models are chosen at random.

2.3.2 Sampling strategy and performance

We sample the posterior by naïve parallelization on a computing cluster, meaning that we simultaneously run 20 to 40 chains with identical settings but different starting models. Autocorrelations of all parameters of properly converging chains fall below 0.1 within 5 % of the total chain length, and subsequently oscillate around zero. We terminate the chains empirically; when no apparent change to means and variances is observed over many samples the chain is terminated. Afterwards, we assess convergence through the joint analysis of parameter autocorrelations, Geweke tests [Geweke, 1991], running means and variances, and trace plots of parameters and misfit. A more detailed discussion about convergence can be found in section 2.5 while the convergence analysis can be found in the electronic supplement. The results are Markov chains with 100'000 samples (target 1) and 600'000 samples (target 2), respectively. Their properties are summarized in Table 2.1. The following subsections highlight interesting results from selected Markov chains. Posterior statistics for all targets are accessible in the electronic supplement. Generating a single proposal for the targets in this section requires about 6 seconds on 12 logical cores of an 36 logical core Intel i9-7980XE CPU @ 2.60GHz.

2.3.3 Marginal moments and maximum poster probability

Though non-linearity implies non-Gaussianity, we begin the characterization of the posterior with an analysis of means and standard deviations, shown in Figs. 2.3a,b for target 2. While the posterior mean v_s model is nearly identical to the target v_s model, larger differences between mean and target are visible for v_p and ρ. This is also reflected in the standard deviations, which are significantly larger for ρ and v_p than for v_s. In this, somewhat limited sense, v_s is better resolved than ρ and v_p. These results are plausible given the relative insensitivity of seismic waveforms to density, and the dominance of larger-amplitude S waves over lower-amplitude P waves in the L_2 waveform misfit [Blom et al., 2017]. For all parameters, the magnitude of the standard deviations is much smaller than the uniform prior standard deviations ($\sigma_{\text{prior}} = \text{width}/\sqrt{12}$), indicating that we were able to reduce dispersion of the marginals using the FWI experiment, i.e. we 'learned something'.

As can be seen in Fig. 2.3, standard deviations depend on the target parameter value, in addition to depending on location relative to sources and receivers. The dependence of posterior standard deviations on the target itself is a consequence of non-linearity and non-Gaussianity, which is more explicitly expressed by the third statistical moment, the skewness

$$S = \frac{E\left[(X-\bar{X})^3\right]}{E[(X-\bar{X})^2]^{3/2}}, \tag{2.17}$$

where X is a physical parameter (e.g., v_s), $E[.]$ denotes the mean, and $\bar{X} = E[X]$. The skewness is a measure of dispersion where positive and negative deviations contribute in opposite magnitude, due to the third power. Thereby, skewness is a measure of lopsidedness or asymmetry of a distribution. Non-zero values indicate whether the distribution

	Target	N_{dim}	$\frac{d}{dm}$	Σ_D	M	N_{samples}	Decorr. length	Parallel chains	Replica exchange	Convergence
A	Chqr.board	75	10%	Low	Scalars	100’000	Medium	20	✗	✓
B	Chqr.board	75	25%	Low	Scalars	100’000	Long	20	✗	✓
C	Chqr.board	300	25%	Low	Scalars	600’000	Long	40	✗	✗
D	Structure	32’400		High	Scalars	10’000	Short	1	✗	✓
E	Structure	32’400		Medium	Scalars	10’000	Medium	1	✗	✗
F	Structure	32’400		Medium	Optimal (E)	13’000	Medium	1	✗	✓
G	Structure	32’400		Med./High	Scalars	10’000	Short	6	✓	✓
H	Structure	32’400		Low	Scalars	10’000	Long	1	✗	✗
I	Structure	32’400		Low	Optimal (H)	30’000	Long	1	✗	

Table 2.1: Overview of all Markov chain-sampled inverse problems. Inversions which use “Optimal” mass re-use the standard deviations of a previous chain on the same target for the mass matrix, indicated in the brackets. The target for all "Structure" targets is constant, and as such perturbation strength is not reported. High, medium, and low data variance correspond to σ^2 values of respectively 10 μm^2, 1 μm^2, and 0.1 μm^2. Chain G is a parallel tempering chain with data variances logarithmically spaced between $\sigma^2 = 10\ \mu\text{m}^2$ and $\sigma^2 = 8.1\ \mu\text{m}^2$. Note that the number of samples indicates approximately 1/10th of the required forward and adjoint simulations. With convergence, we mean convergence of the quantities of interest, i.e. means, standard deviations and marginals. Convergence is assessed by trace-plots, autocorrelations, running estimates and Geweke scores.

is leaning heavily to one side, i.e., having asymmetric tails. A (multivariate) normal distribution, corresponding to the posterior of a linear inverse problem, has zero skewness. For illustration, a skew-normal distribution with two different skewness values is shown in Fig. 2.4. Note that the mean, median, and mode are all different. This in turn implies that plotting the mean model is not a sufficient characterization for highly skewed or non-normal distributions.

Skewness in target 2, shown in Fig. 2.3, is non-zero for many parameters, indicating a non-Gaussian posterior. This suggests that the use of a Hessian approximation for uncertainty quantification in full-waveform inversion may not be sufficient.

Additionally shown in Fig. 2.3 is the model (sample) with the maximum posterior probability (MAP) evaluated during sampling. Although occasionally interpreted as *the* solution, it is most likely not the absolute global minimum. In this case, the MAP model actually reflects the target worse than the means of the Markov chain, especially for P-wave velocity. However, from a wave-physical point of view, the MAP model (as well as all other samples) do explain the observations adequately. To make the posterior draws and the MAP model more geologically relevant one would need to encode this in the prior (e.g. [Linde et al., 2015]), as the wave physics (likelihood) doesn't require this. We deliberately make our prior unrestrictive, such that we prevent artificially reducing the effective dimension of the model space and making the inverse problem easier to solve.

2.3.4 Joint distributions and inter-parameter moments

In addition to the marginal statistical moments, sampling also allows us to visualize marginal or conditional distributions. An example of a 2D marginal for the checkerboard targets is shown in Fig. 2.5. The two parameters visualized are P-wave velocities in neighboring blocks centered at (78 m, 62 m) and (78 m, 46 m). As a consequence of non-linearity, the marginal is significantly non-Gaussian, and the parameters are strongly dependent, even for target 1 with the lower-amplitude perturbations. This, again, highlights that uncertainty analysis based on a Gaussian approximation may have limited meaning.

Posterior samples provide information on covariance, which allows us to compute correlations between model parameters in the posterior. The correlation matrix for target 2 is shown in Fig. 2.6. A more physical interpretation can be done by selecting one column/row from the correlation matrix and plotting the values in the corresponding basis functions. This has been done for parameter 8 of target 2 in Fig. 2.7. In the electronic supplement correlations between all parameters are available. Higher-order co-moments (e.g. co-skewness, co-kurtosis) are also available from the samples, but are more difficult to interpret.

2.4 High-dimensional model space sampling

Following the consideration of a relatively low-dimensional case with large prior uncertainties, we continue with a high-dimensional model space that is more suitable for the representation of detailed geologic structures. Combined with lower prior uncertainties, this corresponds to a scenario where we seek smaller-scale variations relative to a coarse

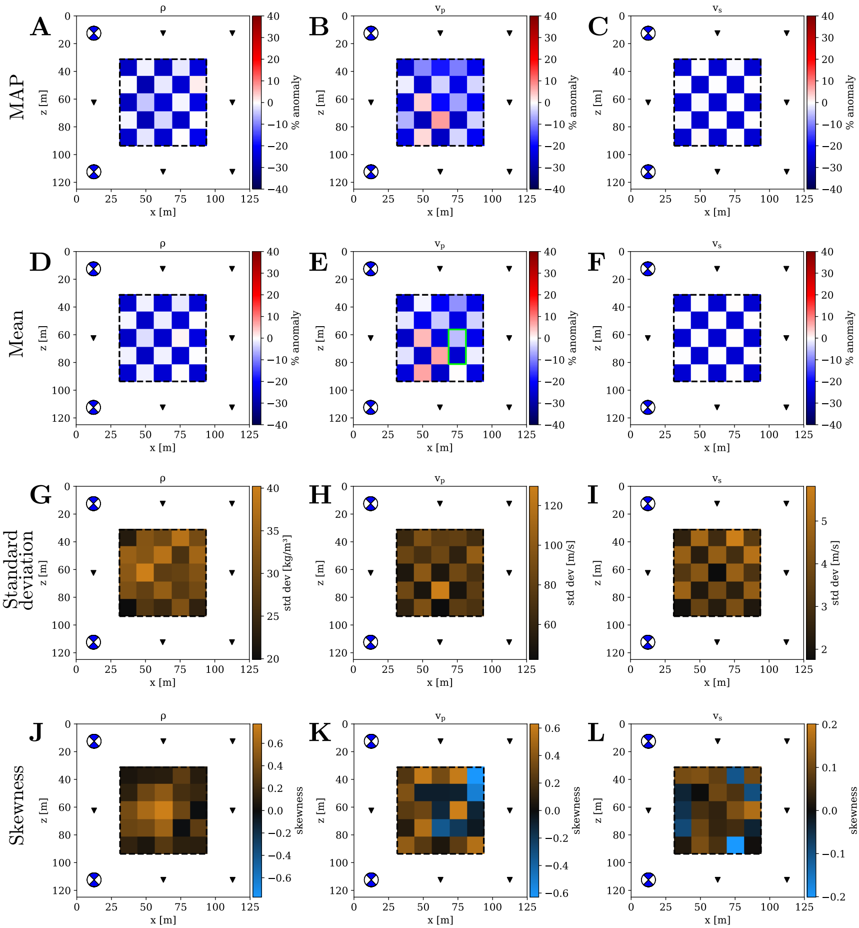


Figure 2.3: Summary of the posterior distribution, including the MAP point and first three statistical moments for checkerboard target 2 shown in Fig. 2.1. These quantities are computed from chain B in Table 2.1. While S-wave velocity in the MAP model (subfigure C) is virtually indistinguishable from the respective means (subfigure F), density and P-wave velocity show deviations. Density and S-wave velocity are well resolved in the sense of having small standard deviations, but P-wave velocity is not as close to the true model for both the means as well as the MAP model. As expected, the smallest standard deviations (subfigures G through I) for all parameters occur close to the sources. A large portion of the parameters has non-zero skewness (subfigures J through L), indicating that these are non-Gaussian. The green box plotted in subfigure E refers to the parameters visualized in Fig. 2.5.

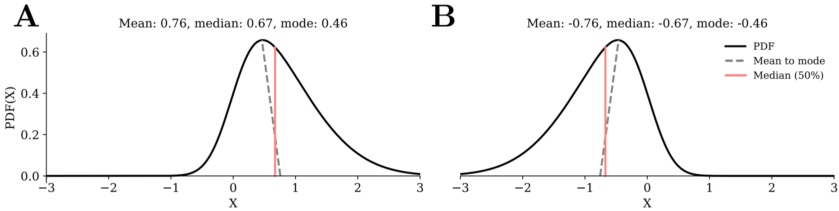


Figure 2.4: Examples of two skew-normal distributions with positive (subfigure A) and negative (subfigure B) skew. The two lines plotted are: a line connecting the mean on the x-axis with the mode (maximum of the distribution) and a line dividing the PDF area in half (median). The skew-normal distributions are described with location 0, scale 1, and skewness 5 (subfigure A) or -5 (subfigure B).

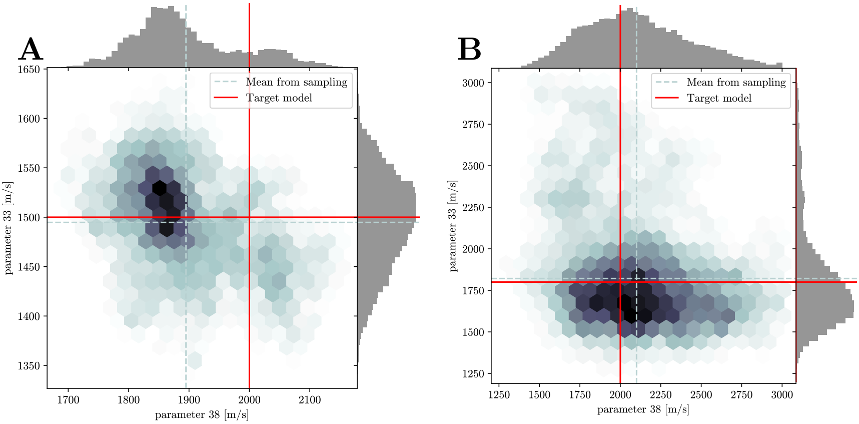


Figure 2.5: Examples of 2D joint distributions for two highly skewed and correlated parameters of checkerboard targets 1 (subfigure A) and 2 (subfigure B), shown in Fig. 2.1. The two parameters are P-wave velocity of adjacent basis functions, shown by the green rectangle in Fig. 2.3. These 2D marginals are extreme cases of non-Gaussian behaviour in the obtained posteriors. Note, however, that for target 1 (smaller perturbations), the mean and samples of the entire marginal distribution are closer to the target model compared to those for target 2 (larger perturbations).

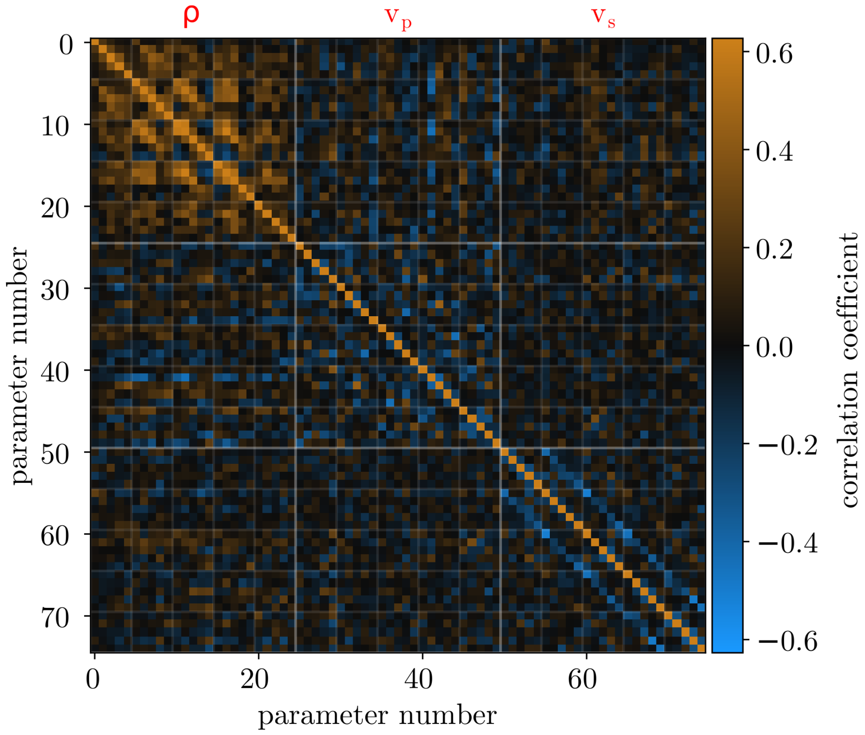


Figure 2.6: Correlation matrix for checkerboard target 2 shown in Fig. 2.1. Parameter correlations are significant within all parameter groups, but also occur between parameter groups. Density is positively correlated to surrounding densities, while velocities are negatively correlated to surrounding velocities.

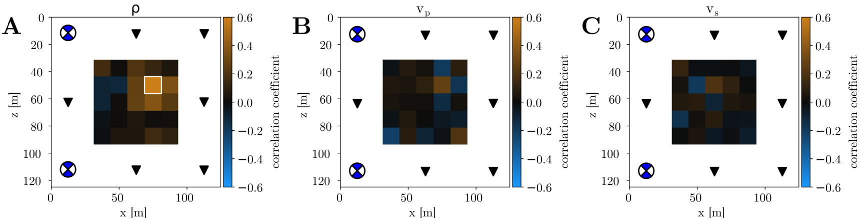


Figure 2.7: Correlations to parameter 8 for checkerboard target 2 shown in Fig. 2.1. Parameter 8 is highlighted by the white box. The positive correlation between densities is focused around parameter 8, as seen in subfigure A. Thus, if one of these densities is found to be higher (lower), neighboring densities are likely to be higher (lower). This is opposed to the (weak) negative correlation to some of the surrounding P-wave velocities as seen in subfigure B. If the density is found to be higher (lower), the surrounding velocities are expected to be lower (higher).

background model that is already well constrained.

Specifically, we construct a $32'400$-dimensional target that mimics a geological structure set in a transmission-dominated experiment. The free parameters are v_p, v_s and ρ defined on the $180 \times 60 = 10'800$ finite-difference grid points. The sources have random moment tensors and are positioned near the bottom of the domain. As source-time function we again use a Ricker wavelet with dominant frequency of 50 Hz. Note that at the dominant frequency, the spatial structure is sub-wavelength. The structural target and source-receiver setup are shown in Fig. 2.8.

While the increased model space dimension acts to decelerate convergence relative to the low-dimensional checkerboard models, this is balanced by stronger prior knowledge, that is, smaller variations with respect to the background. The prior distributions are uniform in the interval 2000 ± 100 m/s for v_p, 800 ± 50 m/s for v_s, and 1500 ± 100 kg/m^3 for ρ.

Additionally, we vary data variance to investigate the influence of the (assumed) noise level. The data variance is given by σ^2 in Eq. (2.14), and varied from $\sigma^2 = 10\ \mu\text{m}^2$ to $\sigma^2 = 1\ \mu\text{m}^2$ and finally to $\sigma^2 = 0.1\ \mu\text{m}^2$. We henceforth describe these values qualitatively as high (10 μm^2), medium (1 μm^2), and low (0.1 μm^2) data variance. The data noise described by it's variance was not added as synthetic noise to the data. As before, important characteristics of all chains are summarized in Table 2.1 and posterior statistics of all targets are accessible through the electronic supplement. The time needed for generating a single proposal for the target in Figure 2.8 is approximately 12.5 seconds on 12 logical cores of an 36 logical core Intel i9-7980XE CPU @ 2.60GHz.

2.4.1 Updated tuning strategy

The tuning strategy from section 2.3.1, where the mass matrix is simplified to include only three parameters, works well when the data variance is high, i.e., when the typical set oc-

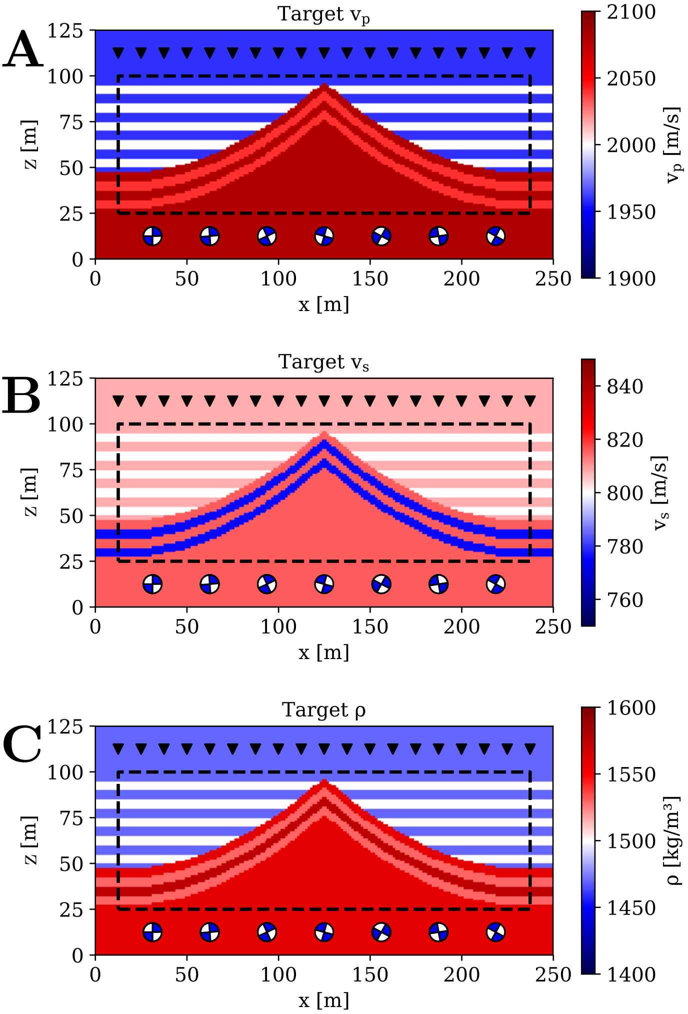


Figure 2.8: Structural target with 10′800 free parameters for v_p, v_s and ρ plotted, respectively, s in subfigures A, B and C. The dashed black line indicates the region within which parameters are allowed to vary. Sources and receivers are indicated by beachballs and black ▼ symbols, respectively.

cupies a large model space volume. However, for lower data variances, the volume of the typical set shrinks quickly, and a more elaborate mass matrix tuning becomes necessary to ensure acceptable convergence.

Our strategy is based on the analysis of linear inverse problems, where the posterior covariance matrix can be shown to be the optimal mass matrix [Fichtner and Zunino, 2019]. Since the posterior covariance is by definition a priori unknown, we choose an approximate approach. For this, we first run a shorter preliminary chain using the simple mass matrix tuning introduced in section 2.3.1. Based on these samples, we compute a rough estimate of the diagonal entries of the posterior covariance, which then serves as a more suitable mass matrix. Though this approach could in principle be repeated multiple times, we only use a single estimate from a previous chain.

Implicitly, this tuning strategy rests on the assumptions that the posterior is roughly Gaussian and that the posterior covariance matrix can be reasonably approximated by its diagonal estimated from a limited number of samples. The extent to which these assumptions hold, determines the effectiveness of the enhanced tuning strategy. Empirically, we find that this strategy accelerates convergence significantly.

In the case of high data variance ($\sigma^2 = 10\ \mu\text{m}^2$), the Markov chain converges relatively fast. Decorrelation length for many parameters is only 3 to 5 samples, after which on average the sampler has generated an independent sample. Satisfactory convergence of 1D and 2D marginals is achieved using 10’000 samples, though more samples would certainly be needed for higher-dimensional marginals or the full posterior. Re-tuning the mass matrix, as described above, allows for the chain with $\sigma^2 = 1\ \mu\text{m}^2$ to converge in approximately 3 times as many samples as in the $\sigma^2 = 10\ \mu\text{m}^2$ case. Although the chain with $\sigma^2 = 0.1\ \mu\text{m}^2$ is not run until means and variances appear stable, the convergence seems to be equally enhanced by re-tuning the mass matrix.

2.4.2 Analysis of the posterior

The means and standard deviations for the converged chains are shown in Figs. 2.9 and 2.11, respectively. As expected, the means differ strongly between the cases of high and medium data variance. Because the amount of effective samples differs per chain, some posterior quantities appear ‘noisy’ for the case of lower variance. This is effectively undersampling to a low degree.

While the means of v_s only delineate the strongest discontinuities for $\sigma^2 = 10\ \mu\text{m}^2$, they provide a remarkably accurate image of the target model for the medium data variance of $\sigma^2 = 1\ \mu\text{m}^2$. The posterior mean values of ρ resemble the target density mostly near discontinuities, which is plausible given that seismic waveforms are primarily sensitive to density gradients. In contrast to v_s and ρ, the posterior mean of v_p is hardly similar to the target model. This is partly due to the pronounced non-Gaussianity of the v_p posterior, exemplified by the 2D posterior marginals shown in Fig. 2.10. In addition, the fact that the true parameters are relatively much closer to the edge of the prior for v_p (due to the increased uncertainty), probably influences the means of the posterior such that they coincide less with the respective mode.

The standard deviations show a strong influence of the data variance σ^2. P-wave ve-

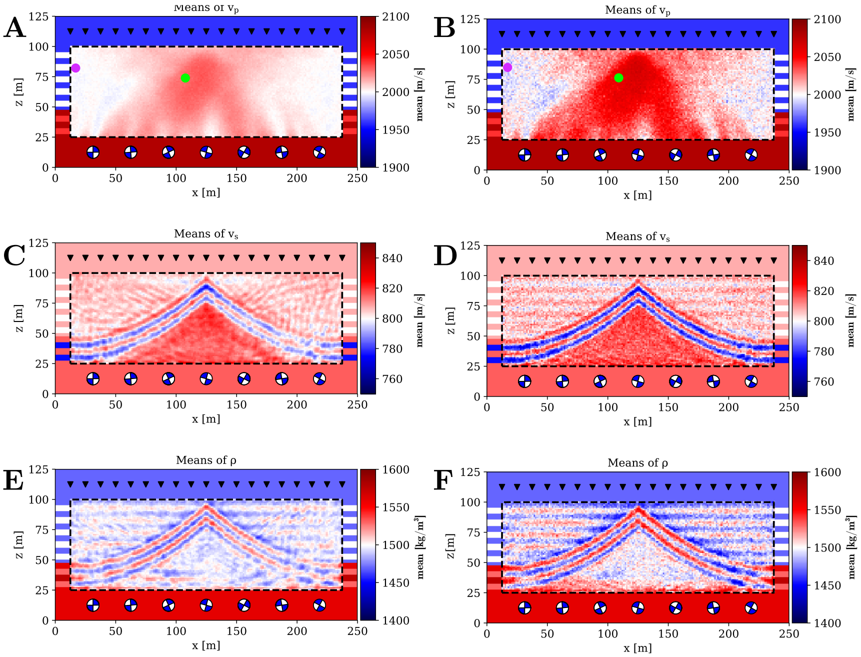


Figure 2.9: Posterior means of chains with $\sigma^2 = 10\ \mu\text{m}^2$ (subfigures A, C and E) and $\sigma^2 = 1\ \mu\text{m}^2$ (subfigures B, D and F). The means for the lower data variance ($\sigma^2 = 1\ \mu\text{m}^2$) show a closer resemblance to the target model. However, these means appear less smooth due to a stronger dependence of samples (undersampling). The v_p mean (subfigure A and B) is hardly similar to the v_p target because the posterior is strongly non-Gaussian, as illustrated in Fig. 2.10. The green and purple dot respectively indicate parameter 8000 and 9720, the parameters visualized in Fig. 2.10.

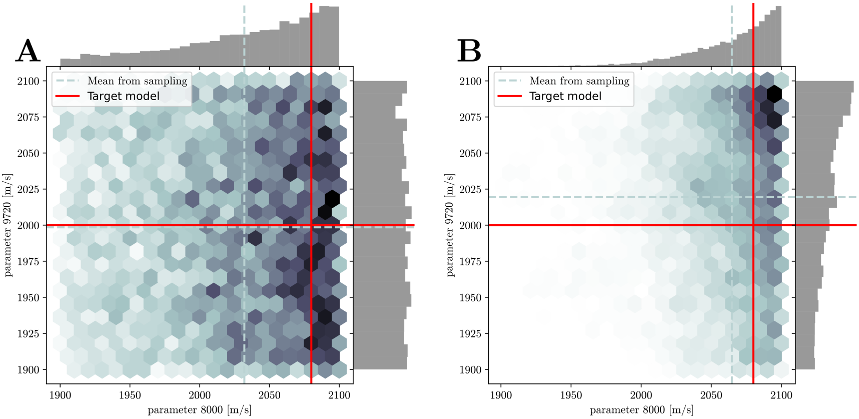


Figure 2.10: Examples of 2D posterior marginal distributions for two neighboring v_p parameters. Locations are indicated in Fig. 2.9. The prior is represented by the limits of the axes. The posterior marginals appear truncated by the prior, adding to the non-Gaussianity of the distribution. As expected, the marginal for $\sigma^2 = 1\ \mu\text{m}^2$ (subfigure B) is more localized than for $\sigma^2 = 10\ \mu\text{m}^2$ (subfigure A). The data is especially non-informative for parameter 9720 in subfigure A, where the posterior marginal strongly resembles the prior.

locity shows relatively low standard deviation along the direct wave paths. As expected, standard deviations of density are lower at discontinuities, and standard deviations for v_s are lowest in regions of elevated v_s. For all parameters, posterior standard deviations are strongly model dependent, again highlighting the non-linear nature of the inverse problem. Changing assumed data noise, i.e., the data variance σ^2, not only modifies the magnitude of the posterior variance, but also its spatial distribution. The white regions of standard deviation in Figure 2.11 indicate parameters on which the standard deviation was not decreased with respect to the prior, ie. no knowledge was gained from the FWI experiment.

2.5 Discussion

In the following paragraphs we discuss further details of our method, including the analysis and acceleration of MCMC convergence, the recovery of density structure, and the future extension to real-data applications in 3D.

2.5.1 Convergence diagnostics

While a large number of convergence diagnostics for MCMC methods have been developed [Gelman and Rubin, 1992, Geweke, 1991, Raftery and Lewis, 1991], none of these is universally applicable or useful [Cowles and Carlin, 1996]. Thus, convergence is relative

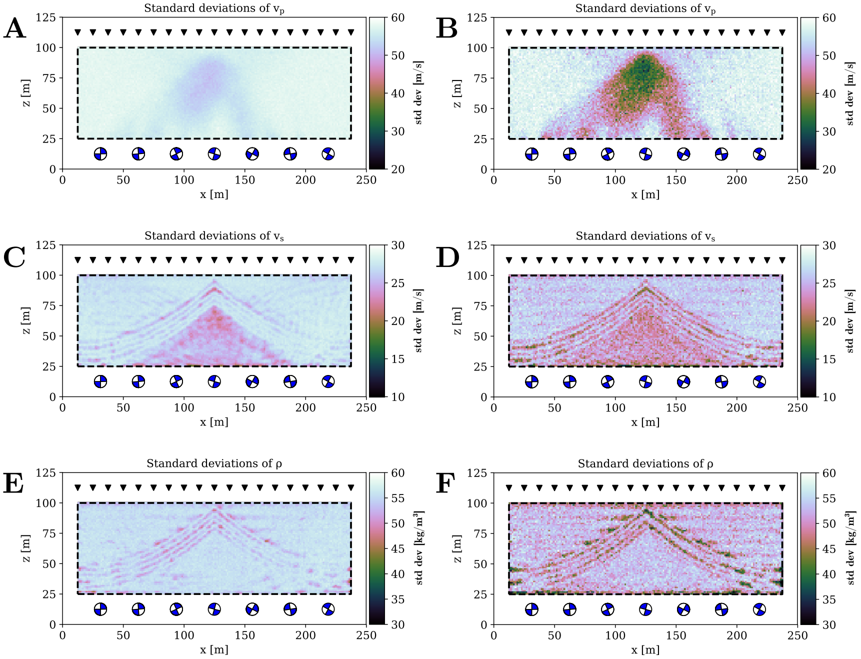


Figure 2.11: Posterior standard deviations for $\sigma^2 = 10\ \mu\text{m}^2$ (subfigures A, C and E) and $\sigma^2 = 1\ \mu\text{m}^2$ (subfigures B, D and F). Not only the magnitude but also the spatial distribution of posterior standard deviations depends strongly on σ. Density clearly has smaller standard deviation at discontinuities, whereas v_s standard deviation is strongly dependent on the v_s target itself. The maximum of the colorbars represent the prior standard deviations, thus white regions indicate parameters where no information was gained by the FWI experiment.

to the information that one wishes to extract from the prior, or to the decisions one needs to make on its basis. Here we subjectively chose to assess convergence by monitoring means and covariances, parameter autocorrelations, trace-plots and Geweke scores. To imply convergence, the Geweke scores are required to lie between -2 and 2. This may be replaced by other metrics, depending on the application.

Furthermore, we ran chains from different initial models for the chequerboard models in order to detect trapping in a local minimum. This was the case in an enlarged checkerboard inversion with 10×10 basis functions, corresponding to 300 model parameters (see chain C in Table 2.1). We therefore conclude that chain C has not converged.

2.5.2 Parallel tempering

A sampling approach that tries to mitigate local minima is parallel tempering or replica exchange, where the states of two Markov chains at different temperatures are randomly swapped [Geyer, 1991, Sambridge, 2014]. Since one of the chains has a higher temperature, it has a higher probability of escaping local minima.

As a proof of concept, we implemented parallel tempering for the high-dimensional target shown in Fig. 2.8. We linked six Markov chains differing in temperature, i.e., data variance. The highest variance was at $\sigma^2 = 10\ \mu\text{m}^2$, and the spacing of the lower variances was chosen empirically, such that an acceptance rate between 20-80 % was achieved. The result is chain G in Table 2.1. With six chains we were able to bridge variances from $\sigma^2 = 10\ \mu\text{m}^2$ to $\sigma^2 = 8.1\ \mu\text{m}^2$ using a logarithmic temperature spacing. On extrapolation, this would mean approximately 50-60 chains for an order of magnitude decrease in data variance. It has been suggested that chains need not be adjacent for swaps to occur, possibly reducing the number of required chains. We expect parallel tempering to also benefit convergence of slowly mixing chains like chains H and I.

2.5.3 Tuning and the mass matrix

The tuning of the mass matrix largely controls the efficiency of HMC. The mass matrix determines the relative speed in the different coordinate directions along a Hamiltonian trajectory, and it may be used for preferential sampling, e.g., of particularly smooth or rough models [Fichtner and Zunino, 2019].

In this work, we applied an intuitive tuning where the diagonal elements of the mass matrix are adjusted using either a visual analysis of trace plots or a rough approximation of the variances using a small number of samples. Ideally, the mass matrix should adapt dynamically to the local curvature of the posterior, e.g., using second-derivative information from the local Hessian [Dahlin et al., 2015, Fu et al., 2016]. The testing of such Hessian-aware algorithms in the context of FWI is work in progress.

2.5.4 Recovery of density structure and avoiding parametrization induced regularization

A particularly noteworthy result of the HMC-based FWI is the recovery of a probabilistic density model that does not suffer from artificial biases introduced by regularization, needed to stabilize inversions for density using gradient-driven optimization [Köhn et al., 2010, Prieux et al., 2013, Blom et al., 2017]. MCMC sampling allows density to vary regularization-free, as required by the data. This, in turn, avoids biases in seismic velocity variations that are often scaled *a priori* to density variations using empirical, but not universally valid, $v_{p,s}/\rho$ ratios [Brocher, 2005].

Additionally, because HMC scales well with increasing posterior dimension, one is able to use many parameters to discretize the inverse problem. This in turn may avoid parametrization induced regularization of the inverse problem, avoiding bias in the posterior.

2.5.5 Towards real-data applications

This work constitutes a pilot study, intended to establish a 2D proof of concept for probabilistic full-waveform inversion based on Hamiltonian Monte Carlo sampling. The transition to 3D real-data applications will require several improvements and additions to the current method.

Most importantly, the numerical modelling must be extended to 3D and optimized for modern high-performance computers. GPU-enabled wave propagation codes are available for this purpose [Komatitsch and Tromp, 2002a,b, Peter et al., 2011, Gokhberg and Fichtner, 2016, Afanasiev et al., 2019]. Additionally, the added free parameters relative to the added data will likely result in the inverse problem becoming non-linear and non-unique in 3D applications, resulting in increased computational cost.

In this synthetic study, we have deliberately chosen a simple misfit functional, i.e., the L_2 waveform misfit (Eq. 2.14). In practice, this would be replaced by other measures of waveform similarity that are more robust and potentially introduce less trade-offs between seismic source parameters and Earth structure [Luo and Schuster, 1991, Gee and Jordan, 1992, Fichtner et al., 2008, Brossier et al., 2010, Leeuwen and Mulder, 2010, Bozdağ et al., 2011, Métivier et al., 2016]. Furthermore, the observational error statistics associated with a specific misfit will need to be analyzed carefully to ensure that the model space posterior is meaningful.

2.6 Conclusions

We have provided a proof of concept for a Bayesian elastic full-waveform inversion in 2D. This was intended to establish the methodological and computational basis for future extensions to real-data applications.

Key ingredients of our method are (1) a Hamiltonian Monte Carlo sampler that explores the full posterior distribution, (2) the computation of misfit derivatives with the help of adjoint techniques, and (3) a tuning strategy that adjusts the diagonal elements of the mass matrix to accounts for the different sensitivities of seismic velocities and density.

The method successfully works for two synthetic end-member scenarios with 75 and $32'400$ dimensions, respectively. In both cases, the algorithm recovers important aspects of the posterior, which can be significantly non-Gaussian. In addition to P-wave and S-wave velocity, the sampling provides constraints on density structure that are free from subjective regularization artifacts, yielding the prior when the data itself is uninformative on a parameter

The most important conclusion is that further improvements listed in section 2.5.5 seem feasible, while certainly not being trivial. This suggests that 3D probabilistic FWI is within reach.

2.7 Acknowledgements

We gratefully acknowledge fruitful discussions with our colleagues Andrew Curtis, Alison Malcolm, Klaus Mosegaard, Julien Thurin, Jean Virieux and Andrea Zunino. Additionally, we thankfully acknowledge the constructive feedback from Niklas Linde and one anonymous reviewer. This work was supported by the Swiss National Supercomputing Center (CSCS) in the form of computing time grants c13 and s868. The code used to produce the synthetic data and perform Hamiltonian Monte Carlo is available through Gebraad et al. [2019].

Chapter 3

HMCLab: a numerical laboratory for algorithmic research in Seismology

Chapter submitted as A. Zunino, L. Gebraad, A. Ghirotto, and A. Fichtner. HMCLab: a framework for solving diverse geophysical inverse problems using the Hamiltonian Monte Carlo method. *Geophys. J. Int.*, 235(3):2979–2991, 10 2023. ISSN 0956-540X. doi: 10.1093/gji/ggad403. URL https://doi.org/10.1093/gji/ggad403. Currently in revision for Geophysical Journal International. A. Z. and L. G. contributed equally to this work, with A. Z. developing the Julia variant, and L. G. the Python variant.

Abstract

The use of the probabilistic approach to solve inverse problems is becoming more popular in the geophysical community, thanks to its ability to address nonlinear forward problems and to provide uncertainty quantification. However, such strategy is often tailored to specific applications and therefore there is a lack of a common platform for solving a range of different geophysical inverse problems and showing potential and pitfalls of the methodology. In this work, we demonstrate a common framework within which it is possible to solve such inverse problems ranging from, e.g, earthquake source location to potential field data inversion and seismic tomography. This allows us to fully address nonlinear problems and to derive sophisticated but useful information about the subsurface, including uncertainty estimation. This approach, in fact, can provide probabilities related to certain properties or structure of the subsurface, such as histograms of the value of some physical property, the expected volume of buried geological bodies or the probability of having boundaries defining different layers. Thanks to its ability to address high-dimensional problems, the Hamiltonian Monte Carlo (HMC) algorithm has emerged as the state-of-the-art tool for solving geophysical inverse problems within the probabilistic framework. HMC requires the computation of gradients, which can be obtained by adjoint methods. This unique combination of HMC and adjoint methods is what makes the solution of tomographic problems ultimately feasible. These results can be obtained

with “HMCLab”, a numerical laboratory for solving a range of different geophysical inverse problems using sampling methods, focusing in particular on the HMC algorithm. HMCLab consists of a set of samplers (HMC and others) and a set of geophysical forward problems. For each problem its misfit function and gradient computation are provided and, in addition, a set of prior models can be combined to inject additional information into the inverse problem. This allows users to experiment with probabilistic inverse problems and also address real-world studies. We show how to solve a selected set of problems within this framework using variants of the HMC algorithm and analyze the results. HMCLab is provided as an open source package written both in Python and Julia, welcoming contributions from the community.

3.1 Introduction

Historically, many new methods have been developed to solve geophysical problems. However, broader impact has typically only resulted from generic and easy-to-use software implementations. In such regards, development of methodologies and implementation of related codes, together with their public availability, can have a major influence to foster progress in geophysics. Some notable examples from the literature which are currently widely employed in the geophysical community are, e.g., the implementation of the spectral element method for seismic wave propagation [Komatitsch and Vilotte, 1998, Komatitsch and Tromp, 1999], the relatively recent implementation of Obspy [Beyreuther et al., 2010, Krischer et al., 2015], a toolbox for seismology, allowing the users to download and process seismic data (among other things), AxiSEM [Nissen-Meyer et al., 2007, 2008, 2014], a spectral element code for global axisymmetric seismic problems, and SimPEG [Cockett et al., 2015], a framework to perform simulations and solve geophysical inverse problems, with focus on electromagnetic methods, just to name a few.

One area where generic and easy to use tools are still missing is the solution of geophysical inverse problems with the Hamiltonian Monte Carlo (HMC) method [e.g., Duane et al., 1987, Neal, 2011, Fichtner and Zunino, 2019], which has recently gained attention in solid Earth geophysics because of its peculiar properties. In this work we aim at filling this gap by providing a framework where a range of different geophysical inverse problems can be solved using the same toolbox.

The HMC method belongs to the framework of the probabilistic approach to inverse problems. Within such framework, an inverse problem essentially represents an indirect measurement where the knowledge about the observed data and model parameters is completely expressed in terms of probabilities [Tarantola, 2005]. Within such formalism, the general solution to the inverse problem is a probability density function (PDF), i.e., the posterior PDF (see Tarantola and Valette [1982] and Mosegaard and Sambridge [2002] for a detailed explanation).

The posterior PDF is constructed from the combination of two pieces of separate information: 1) the prior knowledge on the model parameters, expressed by the PDF $\rho(\mathbf{m})$, where $\mathbf{m}$ represents the model parameters and 2) the information provided by the experiment, described by $L(\mathbf{m})$. The posterior distribution, under certain fairly wide assumptions, is then given by [Mosegaard and Sambridge, 2002, Tarantola, 2005]:

$$\sigma(\mathbf{m}) = k\rho(\mathbf{m})L(\mathbf{m}). \tag{3.1}$$

Since $\sigma(\mathbf{m})$ is a PDF, it requires to evaluate the relevant integrals to find features of interest. For example, calculating the expected model given the data requires evaluating the following integral

$$E_M[\mathbf{m}] = \int_M \mathbf{m}\,\sigma(\mathbf{m})\,\mathrm{d}\mathbf{m}, \tag{3.2}$$

where M represents the whole model space.

This is mainly because of two reasons: to I) simplify the mathematical framework (linear algebra) and make the interpretation of the results easier (e.g., one single “optimal” solution, uncertainty fully quantified by a covariance matrix) and II) tractability in terms of computational requirements, since Monte Carlo methods typically require a very large amount of forward model evaluations for a large number of unknowns. Because of that, plain Monte Carlo approaches quickly become intractable for large scale problems. However, thanks to both algorithmic and computational advances it is now becoming possible to address several nonlinear geophysical problems by means of sampling the posterior distribution, thereby avoiding linearizations and incorporating more sophisticated prior information. The result of this is a fully nonlinear appraisal of inverse problems where the output probabilities can show significantly non-Gaussian properties, the cases where linearization-based approaches may potentially fail.

Selecting algorithms that generate samples from a specific posterior efficiently is central to doing efficient Bayesian inference and, in this context, we now introduce HMC.

In this work we aim at unifying these diverse applications under a common framework, showing how the probabilistic approach can be useful in multiple contexts in geophysics and, on the practical side, providing a numerical laboratory in the form of a software package.

In the particular case of linear forward models and Gaussian uncertainty, the probabilistic formalism provides the same closed-form solution than the classical least squares approach. In general, however, such high-dimensional integrals cannot be computed. Therefore, we resort to *sampling*, a technique to approximate the computation of high-dimensional integrals. By generating points in the model space whose density (number of points per unit volume) is proportional to the posterior $\sigma(\mathbf{m})$, i.e., samples, one greatly reduces the amount of computations required to estimate statistics (such as $E_M[\mathbf{m}]$) compared to a systematic grid search. Markov Chain Monte Carlo (MCMC) methods provide a clever way to construct a Markov chain that produces samples drawn for the target distribution, i.e., the posterior PDF. Statistical analysis of the samples obtained with MCMC then provides the answer to any inquiry in terms of probability of certain events, i.e., specific features of the solution. Practically, this means we can compute probabilities related to particular properties or structures of the solution. For instance, we might be interested in the probability of a certain geological body to have a certain volume, or the probability that there is a continuous permeable layer connecting two locations in the subsurface.
Instead, quantities such are these can be computed using Monte Carlo integration, which is the basis of appraising Bayesian posteriors using MCMC methods.

MCMC to sample target distributions has evolved from the appearance of the original Metropolis algorithm in physics [e.g., Metropolis and Ulam, 1949, Metropolis et al., 1953] into a plethora of variants in many different scientific fields, including geophysics [see Sambridge and Mosegaard, 2002, for a review]. The MCMC method to solve inverse problems in geophysics has a long history, which dates back to the pioneering work of Keilis-Borok and Yanovskaja [1967] and Press [1968], which were contemporary to the classic work of Backus and Gilbert [1967] on geophysical inverse theory. The formalization of a comprehensive theoretical framework based on probability theory for geophysical problems was started in the 80s by Tarantola and Valette [1982] and extended later on to include MCMC sampling methods to solve nonlinear inverse problems [e.g., Mosegaard and Tarantola, 1995, Mosegaard and Sambridge, 2002, Tarantola, 2005, Mosegaard, 2011]. An extensive review can be found in Dębski [2010].

The main reason for this is that MCMC methods have long been thought to be unfeasible for medium to large scale problems due to the high computational requirements. However, when the forward problem is nonlinear, i.e., the calculated data depend on the model parameters in a nonlinear fashion, Monte Carlo methods represent one of the strategies capable of properly addressing the consequences of such nonlinearity and, at least theoretically, to explore the set of different solutions compatible with the observed data [Mosegaard and Tarantola, 1995, Mosegaard and Sambridge, 2002, Sambridge and Mosegaard, 2002]. Moreover, MCMC methods provide uncertainty quantification for such problems, an essential tool to appraise the found solution.

There has been a recent increase in the use of Monte Carlo methods to solve (geophysical) inverse problems, essentially for two reasons: I) relatively recent advances in the sampling algorithms and II) substantial increase in the available computational resources. These advances now allow us to tackle medium (hundreds of model parameters) to relatively large-scale (tens of thousands model parameters) inverse problems within the probabilistic framework.

Regarding I), the literature provides a large collection of generic algorithms to perform Monte Carlo sampling, including classic MCMC [e.g., Hastings, 1970], slice sampling [e.g., Neal, 2003], Gibbs sampling [e.g., Geman and Geman, 1984], rejection sampling [e.g., Gilks et al., 1995], sequential Monte Carlo [e.g., Liu and Chen, 1998] and trans-dimensional Monte Carlo [e.g., Green, 1995], just to cite some of the main categories. The amount of literature dedicated to such algorithms from different fields is so vast that it would be pointless to attempt to give an overview here. More specifically, in geophysics there has been a constant increase in the number of algorithms proposed both for (pseudo-)sampling the posterior PDF and performing global optimization. These include, for instance, an extension of the classic Metropolis-Hasting sampler [e.g., Mosegaard and Tarantola, 1995], simulating annealing for global optimization [e.g., Stoffa and Sen, 1991], a strategy based on the nearest neighbour (Voronoi) partition of the model space [e.g., Sambridge, 1999], joint inversion of different geophysical data with a cascade MCMC [e.g., Bosch, 1999], trans-dimensional MCMC inversion [e.g., Malinverno, 2002, Bodin and Sambridge, 2009] and samplers with geostatistical-based priors [e.g., Hansen et al., 2012, Zunino et al., 2015]. Regarding II), the geophysicist's arsenal for solving inverse problems now includes large high-performance computing resources and more easily accessible computation accelerators such as GPUs, TPUs and high-thread-count CPUs.

Traditional algorithms such as the random walk Metropolis may be regarded difficult to setup for large problems because of two reasons. The first is the property of generating correlated samples which requires a large number of iterations to obtain reliable statistics. The second is the difficulty for the proposal mechanism to generate high-probability models. In high-dimensional spaces, simply randomly perturbing a model will very unlikely produce another model with a higher posterior PDF. The reason is that high-dimensional spaces tend to be very empty ("curse of dimensionality") and so the vast majority of search directions (random perturbations) will point to low probability regions, making the overall algorithm inefficient [Curtis and Lomax, 2001]. This tends to produce a slow exploration of the model space, making the overall algorithm inefficient.

If strong prior information is available and if sampling directly the prior is a possibility, the extended Metropolis algorithm [Mosegaard and Tarantola, 1995] can offer a substantial improvement in terms of efficiency. Nevertheless, defining a geologically realistic prior and being able to sample it may not be an easy task in practice.

Other approaches are based on an adaptive algorithm which, for instance, computes an approximate local covariance matrix and then samples such information to increase the chance of moving in a direction of higher probability [e.g., Gilks et al., 1996]. A recently proposed methodology for improving the proposal strategy is that of constructing an ad hoc proposal PDF based on the results of a simplified deterministic inversion [Khoshkholgh et al., 2021, 2022], which will make the sampling more efficient in practice, without altering the final equilibrium distribution. This may enable the solution of large problems, although such methodology requires solving an additional inverse problem beforehand and performing some sort of interpretation in addition to the estimation of the modeling error.

Moreover, the more the samples are uncorrelated, the better the statistical estimations. Because of that, methods which tend to produce more independent samples are desirable because they provide the same accuracy with a smaller number of samples compared to methods producing highly correlated samples.

An alternative algorithm that has recently gained popularity is the HMC method [e.g., Duane et al., 1987, Neal, 2011, Fichtner and Zunino, 2019]. HMC combines sampling with ideas from the field of optimization, where the proposal mechanism exploits also information coming from the gradient of the posterior distribution. This unique combination enables a more efficient solution of problems when the calculation of gradients is not computationally too expensive with respect to the cost of simulating the forward problem, e.g., when adjoint methods [Tarantola, 1984, Tromp et al., 2005, Fichtner et al., 2006a, Plessix, 2006] come into play [e.g., Zunino and Mosegaard, 2018, Fichtner and Zunino, 2019, Gebraad et al., 2020]. This is effectively a result of the No Free Lunch-theorem described by Wolpert and Macready [1997]: one applies prior knowledge to the objective function and its properties (i.e., cheaply available gradient information), whereby it becomes possible to select a relatively efficient algorithm. HMC is capable of generating more uncorrelated samples compared to traditional purely random-walk based algorithms such as the random walk Metropolis algorithm [Neal, 2011], producing more accurate statistical estimations with a smaller number of samples. Thanks to this property, HMC is more suitable to address high-dimensional inverse problems than traditional derivative-free sampling methods. In fact, the cost of generating independent samples with HMC under increasing dimension n grows as $O(n^{5/4})$ [Neal, 2011], whereas it grows as $O(n^2)$ for standard Metropolis-Hastings [Creutz, 1988]

An alternative and recently popularized approach using also information from the gradient is Stein Variational Gradient Descent [SVGD, Liu and Wang, 2019] a variational inference algorithm, which aims at approximate inference by minimizing the Kullback-Leibler divergence between the proposed and target distributions.

By using the Hamiltonian dynamics as a proposal mechanism, at each iteration all the model parameters are perturbed based on the information from the gradient of the posterior PDF and the momentum. This enables a faster convergence to equilibrium (shorter burn-in) and longer moves compared to traditional pure random walk strategies. Moreover it does not need any particular intervention from the user apart from setting the tuning parameters and does not require the ability to draw realizations from the prior PDF. Thanks to the above mentioned features and modern computational resources, HMC can now address problems with thousands of unknowns.

Although the theoretical formalism and the infrastructure to perform intensive computations are there, a common framework to address different geophysical inverse problems has not emerged yet. Implementations of the HMC algorithm are typically application-specific and often not easily accessible to non-specialists. In addition, as these methods are nascent in the field of solid Earth geophysics, the community as a whole has not had time to acquire substantial expertise in the usage of these methods in order to evaluate their potential and routinely apply them to realistic problems. Our work aims at facilitating at least part of the generation of this expertise, specifically in applying gradient-based sampling methods to inverse problems and analyzing their results. In this work we show how HMC can be used to obtain useful information from a set of diverse geophysical data sets through some illustrative selected examples from seismology and potential fields problems. All problems are addressed within the same framework, where generic samplers and data structures allows us to easily experiment with different data, priors and possibly to combine them.

These results are obtained with “HMCLab”, a tool to solve research problems and a numerical laboratory for experimenting with inverse algorithms such as HMC for a variety of geophysical topics. HMCLab provides software for a set of geophysical problems, for which functions to solve the forward problem, compute gradients of the misfit function and several kinds of priors and samplers for the HMC method are provided. This package is currently written partly in Python [van Rossum, 1995] and partly in Julia [Bezanson et al., 2017], depending on the specific problem. It is, however, in constant evolution as new geophysical problems are added or translated into the other one of the two languages. Moreover, users can supply their own forward model functions and priors which can easily be used with the HMC samplers. In addition, several Jupyter Notebooks are provided that guide the user through the various aspects of applying MCMC algorithms and analyzing their results, for various inverse problems.

In the following, we first give a brief overview of the core of HMC algorithms and illustrate what kind of information can obtained by solving some selected example problems using HMCLab.

3.2 Theoretical background

The HMC algorithm can be efficient in sampling a broad class of posterior PDFs compared to sampling algorithms that do not exploit gradient information and rely on purely random walk behavior (e.g. the “Random Walk Metropolis-Hastings” (RWMH)).

3.2.1 The original HMC sampler

HMC constructs a Markov chain over an n-dimensional probability density function $\sigma(\mathbf{m})$ using classical Hamiltonian mechanics. The algorithm regards the current state $\mathbf{m}$ of the Markov chain as the location of a physical particle in an n-dimensional space $\mathscr{M}$ (i.e., model or parameter space). It moves under the influence of a potential energy, U, which is defined as

$$U(\mathbf{m}) = -\ln(\sigma(\mathbf{m})). \tag{3.3}$$

To complete the physical system, the state of the Markov chain needs to be artificially augmented with momentum variables $\mathbf{p}$ and a generalized mass for every dimension pair. The collection of resulting masses is contained in a symmetric positive definite mass matrix $\mathbf{M}$ of dimension $n \times n$. The momenta and the mass matrix define the kinetic energy of the particle as

$$K(\mathbf{p}) = \frac{1}{2}\mathbf{p}^T\mathbf{M}^{-1}\mathbf{p}. \tag{3.4}$$

In the HMC algorithm, the momenta $\mathbf{p}$ are drawn randomly from a multivariate Gaussian with covariance matrix $\mathbf{M}$ (the mass matrix). The sum of the location-dependent potential and momentum-dependent kinetic energy constitute the total energy, or Hamiltonian, of the system

$$H(\mathbf{m},\mathbf{p}) = U(\mathbf{m}) + K(\mathbf{p}). \tag{3.5}$$

The Hamiltonian dynamics are governed by the following equations,

$$\frac{\partial \mathbf{m}}{\partial \tau} = \frac{\partial H}{\partial \mathbf{p}}, \quad \frac{\partial \mathbf{p}}{\partial \tau} = -\frac{\partial H}{\partial \mathbf{m}}, \tag{3.6}$$

which determine the position and momentum of the particle as a function of time τ. This time τ is artificial just like the mass matrix, it has no connection to the actual physics of the inverse problem at hand.

We can simplify Hamilton's equations using the fact that kinetic and potential energy depend only on momentum and location, respectively, to obtain

$$\frac{\partial \mathbf{m}}{\partial \tau} = \mathbf{M}^{-1}\mathbf{p}, \quad \frac{\partial \mathbf{p}}{\partial \tau} = -\frac{\partial U}{\partial \mathbf{m}}. \tag{3.7}$$

Evolving $\mathbf{m}$ over time τ generates another possible state of the system with new position $\tilde{\mathbf{m}}$, momentum $\tilde{\mathbf{p}}$, potential energy $\tilde{U}$, and kinetic energy $\tilde{K}$. Due to the conservation of energy, the Hamiltonian is equal in both states, i.e., $U + K = \tilde{U} + \tilde{K}$. Successively drawing random momenta and evolving the system generates a distribution of the possible states of the system. Thereby, HMC samples the joint momentum and model space, referred to as phase space. As we are not interested in the momentum component of phase space, we marginalize over the momenta by simply dropping them. This results in samples drawn from $\sigma(\mathbf{m})$.

If one could solve Hamilton's equations exactly, every proposed state (after burn-in) would be a valid sample of $\sigma(\mathbf{m})$. Since Hamilton's equations for non-linear forward models cannot be solved analytically, the system must be integrated numerically. Suitable integrators are symplectic, meaning that time reversibility, phase space partitioning and volume preservation are satisfied [Neal, 2011, Fichtner and Zunino, 2019]. In this work, we employ the leapfrog method as described in Neal [2011], with higher order symplectic integrators also implemented. However, the Hamiltonian is generally not preserved exactly when explicit time-stepping schemes are used [e.g., Simo et al., 1992]. Therefore, the time evolution generates samples not exactly proportional to the original distribution. A Metropolis-Hastings correction step is therefore applied at the end of numerical integration.

In summary, at each iteration, samples are generated starting from a randomly drawn model $\mathbf{m}$ in the following way:

1. Propose momenta $\mathbf{p}$ according to the Gaussian with mean $\mathbf{0}$ and covariance matrix $\mathbf{M}$;
2. Compute the Hamiltonian H of model $\mathbf{m}$ with momenta $\mathbf{p}$;
3. Propagate $\mathbf{m}$ and $\mathbf{p}$ for some time τ to $\tilde{\mathbf{m}}$ and $\tilde{\mathbf{p}}$, using the discretized version of Hamilton's equations and a suitable numerical integrator;
4. Compute the Hamiltonian $\tilde{H}$ of model $\tilde{\mathbf{m}}$ with momenta $\tilde{\mathbf{p}}$;
5. Accept the proposed move $\mathbf{m} \to \tilde{\mathbf{m}}$ with probability
$$p_{\text{accept}} = \min\left(1, \exp(H - \tilde{H})\right) . \tag{3.8}$$
6. If accepted, use (and count) $\tilde{\mathbf{m}}$ as the new state. Otherwise, keep (and count) the previous state. Then return to 1.

The mass matrix $\mathbf{M}$ is one of the important tuning parameters of the HMC algorithm; details on its meaning and suggestions for tuning can be found in Fichtner and Zunino [2019], Fichtner et al. [2021]. Moreover, employing the discrete leapfrog integrator implies that there are two additional parameters that need to be tuned, namely the time step ε and the number of iterations L [Neal, 2011].

3.2.2 HMC variants

The algorithm described so far is the simplest version of HMC, however, several variants of the original algorithm exist, which mostly aim at automatically tuning some of the parameters or improving mixing [Neal, 2011, Sambridge, 2014, Fichtner et al., 2021].

A notable example is the No U-Turn Sampler (NUTS) [Hoffmann and Gelman, 2014], which aims at providing an automatic tuning of the two leapfrog integrator-related parameters, ε and L. NUTS finds a suitable value for ε during the burn-in and then fixes it for the following iterations to avoid breaking the detailed balance property. The number of iterations L instead is dynamically adjusted ("dynamic HMC") at each iteration in a way such that there is no doubling back of the trajectory. This allows for long or short moves depending on the region of the model space which the algorithm is visiting. NUTS is implemented in HMCLab following Hoffmann and Gelman [2014].

Additionally, methods to investigate inverse problems that might show strongly isolated modes exist. By running multiple chains with tempered (i.e. smoothed) posterior PDFs and letting these samplers exchange states, the exploration of local minima might be accelerated [Sambridge, 2014]. Tempered trajectories following Neal [2011] may also help discovering isolated modes. This variation does not require multiple Markov chains, nonetheless, it is able to more easily transition between local minima, at the expense of a reduced acceptance rate.

3.2.3 Gradient computations

As mentioned above, one important aspect of a successful HMC strategy is the capability to efficiently compute the gradient of the potential energy $\nabla U(\mathbf{m}) = \nabla(-\log(\sigma(\mathbf{m})))$. The first method that proves powerful for relatively simple models is to evaluate derivatives analytically. This typically allows for cheap computation of the gradients, but is only applicable to models that can be analytically differentiated. This is most notably used for the joint non-linear source location and medium velocity estimation as mentioned later in this manuscript. For larger problems, a tool that can provide a substantial help in making gradient calculations efficient is the adjoint method [e.g., Lions, 1971, Tarantola, 1984, Talagrand and Courtier, 1987, Tromp et al., 2005, Fichtner et al., 2006a, Plessix, 2006, Hinze et al., 2008]. This strategy allows us to compute the gradient ∇U with a computational cost of about two (three in practice) forward simulations, much cheaper than other approaches such as finite difference methods. We employ the adjoint technique in some of our geophysical problems, namely in the case of acoustic and elastic full waveform inversion and for the nonlinear traveltime problem (eikonal solver).

Another useful tool to compute the gradient for certain problems is automatic differentiation [e.g., Sambridge et al., 2007, Griewank and Walther, 2008], a computational technique where derivatives of a user-coded function are provided automatically by the software in the form of a function. This technique can be convenient for problems where it is difficult to derive the adjoint equations (e.g., when the forward operator is not self-adjoint) or where the forward model needs to be adapted for each specific case because it depends, e.g., on the specific rock types present in the area under study, requiring a re-derivation of the analytical derivatives (such as rock physics models [Mavko et al., 2003]). We use this tool, e.g., for the problem of inversion of amplitude-versus-angle (AVA) seismic reflection data, where the forward modelling is a combination of a rock physics model [e.g., Mavko et al., 2003] and a convolutional seismic model.

3.2.4 Prior information

Prior information plays an important role in solving inverse problems by providing additional information directly on the model parameters to better constrain plausible values for the solution and helping to mitigate the non-uniqueness [e.g., Curtis and Lomax, 2001, Scales and Tenorio, 2001, Hansen et al., 2012, Zunino et al., 2015, Hansen et al., 2016]. In the probabilistic approach, prior information is represented by a PDF $\rho(\mathbf{m})$ on the model parameters.

HMCLab provides a set of common PDFs for the prior, ranging from simple multivariate Gaussian distributions to more complex distributions such as a combination of Beta PDF-based marginals with a Gaussian copula to correlate the marginals. Another interesting prior is based on the Laplace distribution (related to the L1-norm) which promotes sparse (or blocky) models. Moreover, the user can provide his/her own prior by simply implementing functions with the appropriate signature (see the code documentation for more details). Any of the available priors can be combined with any of the available or user-generated forward models.

3.3 Inferring complex information about the subsurface with HMCLab

The HMCLab framework allows us to solve diverse inverse problems using sampling methods under a common platform. The software package includes a set of pre-defined geophysical forward and inverse problems, a set of prior distributions and allows the user to supply his/her own forward problem. In the following we show some examples of how to extract useful information about the subsurface for a set of selected geophysical inverse problems in the framework of the HMC method.

Once a collection of samples from the posterior distribution has been obtained, in order to calculate some arbitrary function $\phi(\mathbf{m})$ of $\mathbf{m}$, we can use the following relationship:

$$\int_{\mathrm{M}} \phi(\mathbf{m})\sigma(\mathbf{m})\,\mathrm{d}\mathbf{m} \approx \frac{1}{N}\sum_{i=1}^{N}\phi(\mathbf{m}_i) \tag{3.9}$$

where N is the number of available samples and $\mathbf{m}_i$ represents one of the posterior models.

3.3.1 2D full waveform acoustic inversion of reflection data

The first example is a 2D inversion of a seismic dataset based on the acoustic approximation. The forward problem is represented by the constant-density acoustic wave equation:

$$\frac{1}{v^2}\frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial z^2} + s \tag{3.10}$$

where t is time, x and z the spatial coordinates, u is the pressure field, v the acoustic velocity and s the source term. Forward calculations are carried out using a finite-difference scheme [Bunks et al., 1995, Pasalic and McGarry, 2010], where the model parameters are velocity at a set of grid points with size $(N_x \times N_z) = (160 \times 90)$ for the x- and z-direction respectively, for a total of 14400 model parameters. The grid spacing is 10 m in both directions. We assume the observational errors to be Gaussian distributed. Therefore, we use an L_2-norm potential energy function. The gradient of such a misfit function with respect to velocity is computed by means of the adjoint method for the acoustic wave equation,

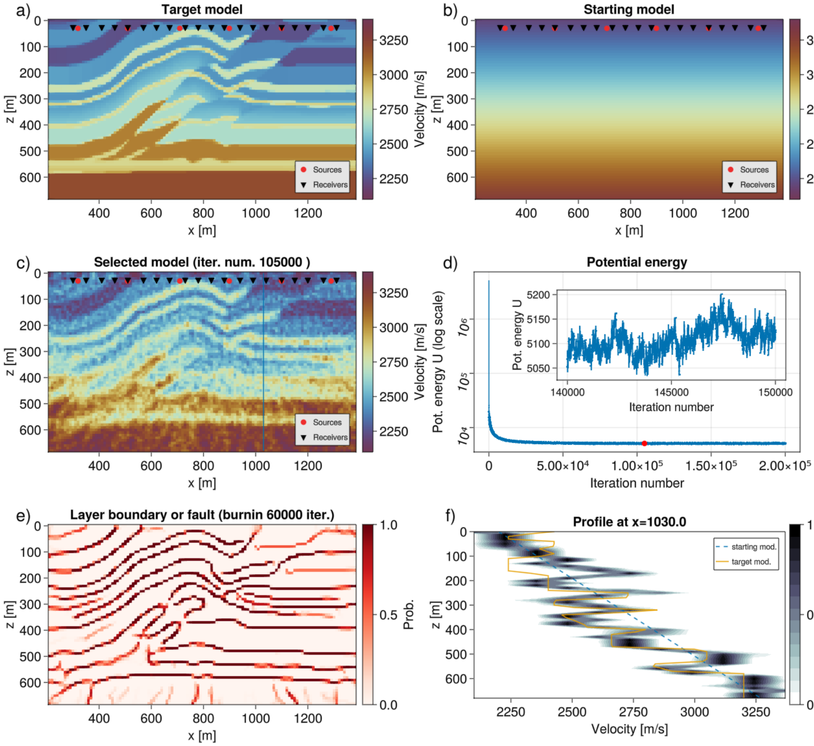


Figure 3.1: Acoustic waves inversion for velocity. a) Target model of velocity, i.e., the one used to calculate the synthetic data. b) The starting velocity model used in the inversion. c) A randomly selected model from the collection of posterior models. d) A plot of the potential energy (misfit) as a function of iteration number, where the red dot indicates the potential energy of the model shown in panel d). e) A map of the probability of having a layer boundary of a fault computed using the collection of posterior models. f) A vertical profile of velocity showing the probability as computed from the collection of posterior models. The profile location is shown by a vertical line in panel c).

as described in Bunks et al. [1995]. The use of the adjoint method enables us to efficiently evaluate the gradient, an essential prerequisite for being able to perform an HMC inversion. The geometry of the problem resembles the one typically found in exploration seismology, where active sources and receivers are located near the surface of the Earth, as shown in Fig. 3.1. The top boundary condition is a free surface, while the other sides are absorbing boundaries implemented as C-PML layers [Komatitsch and Martin, 2007]. We use a set of 6 sources to generate synthetic data, add correlated Gaussian noise (standard deviation 0.05, correlation length 0.01 s) and use the result as the observed data to be inverted for the velocity model. To perform the inversion we use the NUTS algorithm [Hoffmann and Gelman, 2014], part of HMCLab.

We ran 2×10^5 iterations of the NUTS algorithm, collecting about 45000 samples after thinning the chain and removing the models resulting from the burn-in phase. The starting velocity model is laterally homogeneous (see Fig. 3.1b). The target model is a modified version of the SEG/EAGE overthrust model [Aminzadeh and Brac, 1997].

Fig. 3.1c shows a randomly chosen model from the collection of the posterior models. The model resembles the target model well, and all the different layers are visible. The potential energy decreases rapidly within the first few hundreds of iterations, when the algorithm attempts to find a model which fits the large-scale structures (see Fig. 3.1d). Subsequently, the misfit keeps decreasing relatively slowly for much longer. We suspect this is due to the algorithm slowly adjusting the fine-scale structures, until it reaches a relatively stable misfit value. From the resulting collection of posterior models we can extract different pieces of information. One practical example is, e.g., calculating the probability of having a layer boundary or fault at any given node of the grid. To do so, we exploit eq. 3.9 using an indicator function $h(\mathbf{m})$ and compute

$$\int_{\mathrm{M}} h(\mathbf{m})\sigma(\mathbf{m})\,\mathrm{d}\mathbf{m} \approx \frac{1}{N}\sum_{i=1}^{N} h(\mathbf{m}_i) \tag{3.11}$$

which produces a value of one in case a boundary/edge is detected and zero if not. The function $h(\mathbf{m})$ in this case is represented by a Canny edge detection filter [Canny, 1986], which is applied to each velocity model in the posterior collection, so that the model is transformed into a binary image of zeros and ones. Fig. 3.1e shows the results of such calculations. The large majority of the boundaries present in the target model appear as high probability structures in Fig. 3.1e, particularly at shallow depths. Finally, Fig. 3.1f shows a map of probability for a vertical profile of velocity at $x = 1030.0$ m, showing the spread of the solutions. The profiles of the starting and target model are shown for comparison.

3.3.2 First arrival traveltime hypocenter location using fiber-optic sensing

An archetypal seismological inverse problem is the estimation of earthquake hypocenters in a medium with unknown structure and velocity using the first arrival times of the seicmic waves excited by the events. HMCLab supplies a simple approach to the hypocenter location problem that uses first arrival data, assuming a homogeneous medium. Arrival times are modelled by straight rays propagating through this homogeneous medium. The inverse problem only requires relative arrival times of a single phase.

Although the physics of this model seem simple, the strong trade-offs between location, origin time and medium velocity make it ideal for Bayesian inference methods. Especially in the presence of trade-offs, one would expect strongly correlated posteriors, for which the HMC algorithm works particularly well.

To illustrate how HMC performs on this inverse problem, we applied the algorithm to a dataset acquired on Grimsvötn volcano in Iceland using a 12.5 km long Distributed Acoustic Sensing (DAS) fiber [Fichtner et al., 2022, Klaasen et al., 2022], for which the acquisition geometry is shown in Fig. 3.2. Due to the use of DAS, this dataset has approximately 1500 separate channels. We simultaneously infer the location of multiple events for which the effective medium velocity is assumed to be the same. The events are selected based on similarity in the observed move-out, as we expect these events to be relatively close. By simulatenously inferring the location of multiple events the data better constrains the medium velocity, which reduces the trade-off between origin time and medium velocity compared to inferring location, origin time and medium velocity for a single event. We define an L_2 misfit on the relative first arrivals of the picked phases, and only include those channels where the phase is picked. This means that some events have relatively less data points and therefore less importance within the inference. As prior, we use uniform distributions on location in a bounded cube of 20 km by 20 km by 10 km (width by length by depth) centered around the DAS cable. As the medium below the field site is unkown, we construct a prior on the P-wave velocity with a logarithmic uniform distribution (to take into account the fact that velocity is a positive parameter) between 340 and 7000 m/s, the extreme ends of possible medium material, i.e. air to relatively fast rock.

The results of a parallel tempered appraisal with 10 chains using HMC are given in Fig. 3.2. The posterior on medium velocity, seen in Fig. 3.2c, shows how, despite having a model with strong trade-offs in the parameters (namely origin time and medium P-wave velocity), one is still able to infer knowledge on one of these parameters, adding knowledge compared to the prior. Event 3, which was recorded on relatively few channels of the fiber-optic cable, features a high uncertainty of its location in the subsurface. This is in contrast to event 1 and 2, which seem to have a more concentrated volume of uncertainty. These results show that the posterior PDF of the location of these events is neither unimodal nor Gaussian, something which would have been difficult to deal with using deterministic methods. In general, one can see in Fig. 3.2b that events with fewer picks are constrained less. Examples of relevant indicator functions [Arnold and Curtis, 2018] for this inference would be the expected average depth of an event, or the probability of the medium velocity lying within a limited interval.

3.3.3 First arrival tomography based on the eikonal equation

Traveltime tomography is a popular approach for seismic inversion where the arrival time of seismic waves at given locations is used to infer the velocity structure of the subsurface.

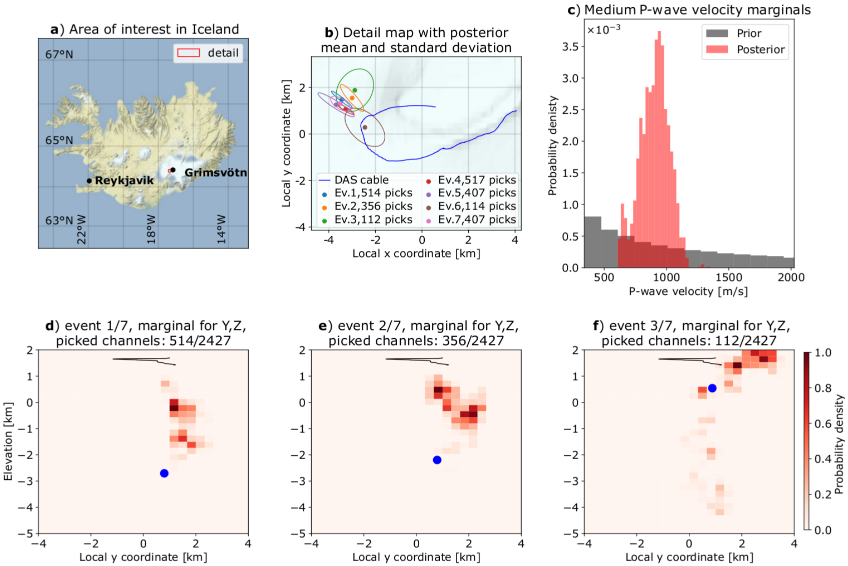


Figure 3.2: Source hypocenter location for data recorded on the Grimsvötn volcano. a) Overview of Iceland and location of the inset b). b) Geometry of the DAS acquisition fiber with posterior means and standard deviation ellipses oriented along principal axes. c) Marginal distribution of medium velocity prior and posterior to the inference. Note that the prior on medium velocity extends beyond the range of the plot. d)-f) Marginal distributions for the Y and Z components of the first 3 events. Note how for event 1 and 2, the volumes of uncertainty are more concentrated than for event 3.

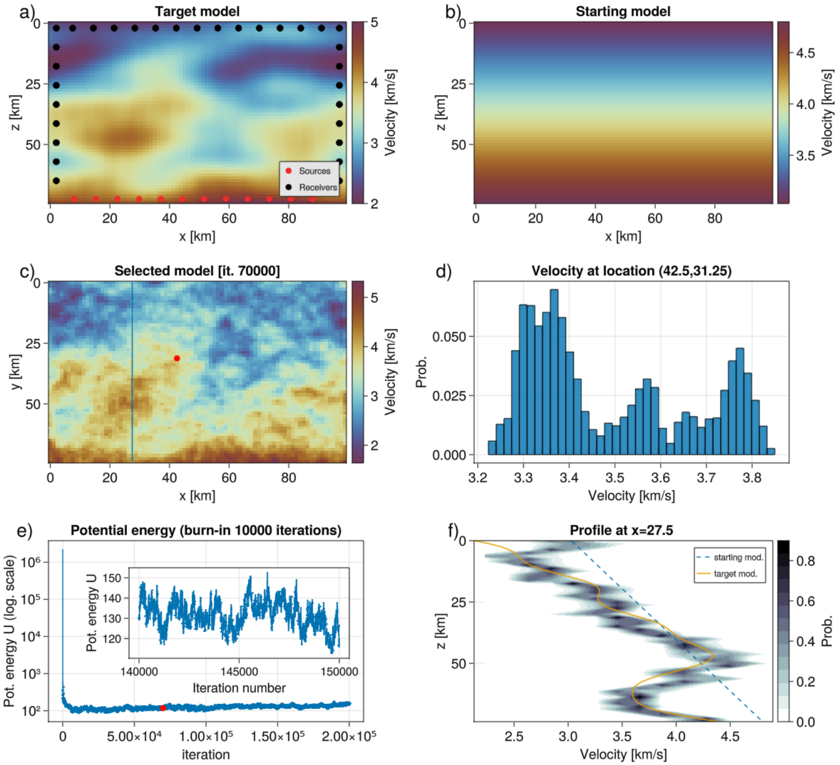


Figure 3.3: Nonlinear traveltime inversion. a) Target model used to generate the "observed" data, b) starting model, c) a selected model from the posterior collection, d) histogram of velocity at $(x,y) = (92,5,26.25)$ (marked by a red dot in c)), e) potential energy, f) profile of velocity showing a probability map as computed from the posterior collection of models (the profile location is shown by a vertical line in panel c)).

In this case, the forward model is represented by the eikonal equation:

$$\sum_d \left(\frac{\partial t}{\partial x_d}\right)^2 = c^{-2}(x)\,, \tag{3.12}$$

where t is the traveltime, c the velocity and $d = 2\,\text{or}\,3$ is the number of dimensions. Since the forward model is nonlinear, typically ray paths are computed a priori in a reference Earth model and fixed, to linearize the problem and hence solve the inverse problem with gradient-based deterministic methods. However, such strategy where a single solution is sought might miss some important information. Because of the nonlinearity of the problem, the misfit functional may feature multiple minima which cannot be detected with deterministic methods. On the contrary, a probabilistic approach may reveal different plausible velocity models to be consistent with the observed data, as we show in the following example.

The example we address here is to solve a 2D inverse problem with a geometry depicted in Fig. 3.3, where we have a velocity model described by 4800 cells (80 in the x-direction and 60 in the y-direction), a set of sources randomly distributed close to the bottom of the model and a set of receivers near the surface and along the left and right sides of the model. The grid spacing is 1.25 km in both directions. The forward problem is solved using a fast marching method (FMM) [e.g., Sethian, 1996, Rawlinson and Sambridge, 2004, Treister and Haber, 2016] which computes the traveltimes at each point of a grid using a finite-difference strategy. The gradient of the Gaussian misfit functional with respect to velocity is computed by means of the adjoint method [e.g., Leung and Qian, 2006, Taillandier et al., 2009, Zunino and Mosegaard, 2018]. To solve the inverse problem, we ran 2×10^5 iterations with the NUTS algorithm. The target model is shown in Fig. 3.3a, while the starting model is depicted in Fig. 3.3b. The latter is a laterally homogeneous velocity model. Panels c and d of Fig. 3.3 show a randomly selected model from the posterior collection and a histogram of the velocity at a given location as obtained by looking at all samples after the burn-in period. Interestingly, the histogram shows a multi-modal distribution, where probable velocity values cluster near three different values. This means that there are three different ranges of velocity values which are highly probable, i.e., they are all compatible with the observed data. Such finding would not be possible with an optimization method where the solution is represented by a single velocity model. The potential energy (misfit) as a function of iterations (Fig. 3.3) features a sharp decrease in the first hundreds of iterations and then a slow descent until equilibrium is reached around 5×10^5 iterations. Fig. 3.3f shows a map of probability for a vertical profile at $x = 27.5$ km, together with the profile for the starting and target model.

3.3.4 Magnetic anomaly inversion with polygonal bodies

The last synthetic example presented is an inversion of magnetic anomaly data using a 2.75D parameterization in terms of polygonal bodies [e.g., Rasmussen and Pedersen, 1979, Campbell, 1983]. The design and detailed description of the method to construct and solve this inverse problem is the subject of another paper [Zunino et al., 2022]. In this setup, each polygonal body has a homogeneous magnetization and its shape is controlled

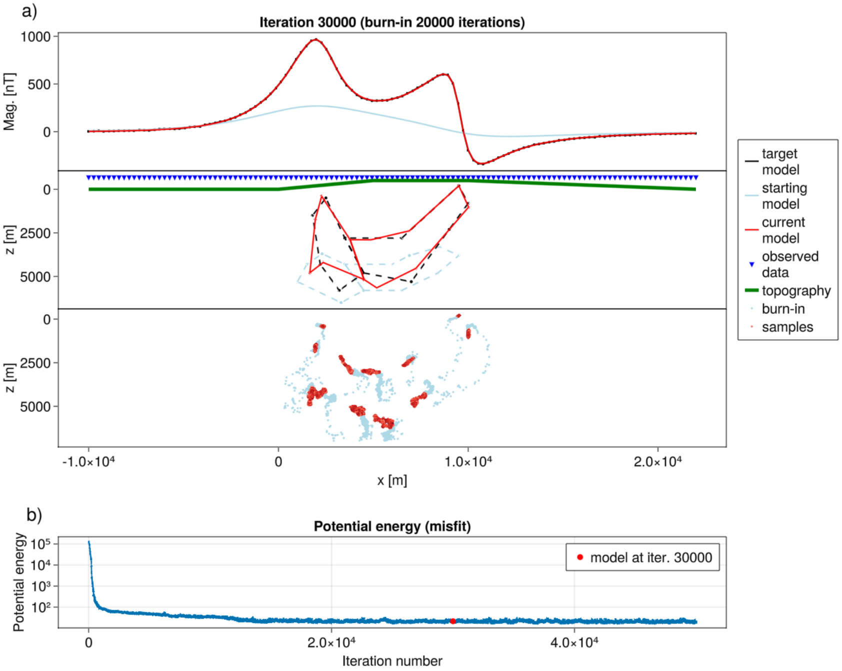


Figure 3.4: Magnetic anomaly problem. a) Three panels showing (from top to bottom): 1) the observed magnetic anomaly, the one calculated from the starting and selected models, 2) the starting, target and selected polygonal bodies including topography and position of measurements and 3) a scatter plot of the position of the vertices before and after the burn-in phase (every 10 iterations). b) A plot of the potential energy (log. scale) as a function of iteration number.

by the position of its vertices, which, in this example, are the unknowns of the inverse problem. In this setup, the relation between the position of vertices (model parameters) and the magnetic response is nonlinear. The label 2.75D means that the polygonal bodies have a given finite lateral extent in the y direction that can be different in the $+y$ and $-y$ directions.

The observed data are represented by a profile along the x direction with about 130 observation points (see Fig. 3.4a). To solve the inverse problem, we ran 5×10^4 iterations using the NUTS algorithm. The potential energy decreases rapidly in the first few hundreds of iterations (see Fig. 3.4b) when the polygonal bodies move from the starting position to a more likely configuration, similar to the target model. Afterwards, the algorithm samples the posterior distribution, producing a set of posterior models. In this example the mass matrix contains non-zero off-diagonal elements in order to force the algorithm to avoid creating geologically implausible shapes. Such off-diagonal elements control the correlation of the momentum variables and hence indirectly the correlation shown in the models visited by the algorithm.

A randomly selected model from the posterior collection is shown in Fig. 3.4a, which fits the observed data well and is also close to the shape of the target model. In the last panel of Fig. 3.4a a set of position of the vertices before and after the burn-in phase are depicted, showing the relatively low uncertainty in the positions for the anomaly.

3.4 The HMCLab framework as a software package

The HMCLab framework is practically implemented in a set of open-source software packages written in the Python [van Rossum, 1995] and Julia [Bezanson et al., 2017] programming languages. HMCLab is, in fact, a numerical laboratory to allow the user to experiment with sampling algorithms on different geophysical problems, ranging from purely educational examples to research-oriented studies. The aim is to provide a user-friendly framework where it is possible to experiment with various problems and algorithms and solve realistic inverse problems, either provided by HMCLab or created by the user.

Table 3.1 summarizes the geophysical problems and prior models which are currently available in HMCLab. Both Julia and Python implementations are modular in that, forward modelling, gradient calculations and sampler (or optimizer) can be combined arbitrarily. Moreover, in addition to the listed problems, the sampler (or optimizer) can be used on user-defined problems.

The main categories of inverse problems currently addressed by HMCLab are:

- linearized (straight rays) and nonlinear (eikonal equation) traveltime tomography,
- full waveform inversion in 2D in the acoustic and elastic formulations (P-SV),
- earthquake source location in 3D based on the straight-ray approximation,
- joint (or independent) gravity and magnetic anomaly inversion in 2.75D using polygonal bodies,

Inverse problem	Julia	Python
Traveltime tomography (eikonal eq., nonlinear)	2D, 3D (parall.)	
Traveltime tomography (refracted rays, nonlinear)		1D (parall.)
Traveltime tomography (straight rays, linear)		2D (parall.)
Full-waveform inversion	acoustic 1D, 2D, 3D (parall., GPU)	elastic 2D (parall.)
Earthquake source location		2D, 3D
AVA seismic reflection data + rock physics	1D, 2D, 3D (parall.)	
Magnetic anomalies (polygons)	2D, 2.75D	
Gravity anomalies (polygons)	2D, 2.75D	
Joint gravity and magnetic anomalies (polygons)	2D, 2.75D	
Arbitrary full and sparse matrix equations		✓
User-provided forwards	✓	✓

Prior distribution	Julia	Python
Gaussian (L2)	✓	✓
Laplace (L1)		✓
Uniform	✓	✓
Beta marginals + Gaussian copula	✓(parall.)	
Arbitrary mixture		✓
User defined marginals	✓	✓

Table 3.1: Snapshot of currently available inverse problems and priors. HMCLab is constantly evolving, therefore changes are expected. 1-2-3D refers to the physical dimensions of the problem, "parall." means a parallelized (multi-core) implementation is available.

- amplitude versus angle (AVA) seismic data including a rock physics model in 3D.

In addition, a set of priors is provided, which can be combined with any problem. For all the above mentioned geophysical problems, HMCLab provides functions to solve forward problems and to compute the gradient of misfit functions with respect to model parameters. For all these physics, samplers are available to appraise the inverse problem. However, the functions of these physics can be used independently of the sampler, hence the user can construct his/her own inversion scheme. Also, the user may supply his/her own forward model, prior and gradient calculation code and subsequently use the available inversion algorithms by providing a minimum set of functions with the appropriate signature as described in the documentation.

HMCLab is not limited to flavors of the HMC algorithm, but includes other more traditional algorithms such as the random walk Metropolis-Hastings algorithm [Metropolis et al., 1953, Hastings, 1970], where gradients are not used. Moreover, we provide an interface [Gebraad, 2022b] to the Stein Variational Gradient Descent (SVGD, [Liu and Wang, 2019]) variational inference algorithm, providing an alternative probabilistic appraisal algorithm to the included MCMC algorithms. As already mentioned, HMCLab includes functions to compute gradients of the misfit functional, and, as such, deterministic inversions are also possible [e.g., Zunino and Mosegaard, 2019]. An example is basic gradient descent, where the modes (local minima) of the defined posterior distribution can be found deterministically. As such, HMCLab also facilitates the usage of Python [e.g., Virtanen et al., 2020] and Julia optimization libraries, including popular algorithms such as the (Limited Memory) Broyden-Fletcher-Goldfarb-Shanno and Newton Conjugate Gradient algorithms [Nocedal and Wright, 2006].

Noteworthy is the inclusion of several notebooks that illustrate the basic concepts of MCMC sampling in general, applied to HMCLab in particular. This ranges from investigating the basic properties of a Markov chain, such as number of proposals, stepsizes and resulting acceptance rates, to tuning the various included algorithms and even implementing one's own inverse problems. These notebooks are available to all users to run out-of-the-box in our supplied Docker environment, but are also available without a Python or Julia interpreter as plain HTML. HMCLab homepage can be found at `https://hmclab.science`, while the Julia and Python versions at `https://gitlab.com/JuliaGeoph` and `https://python.hmclab.science/index.html`, respectively.

Finally, HMCLab is constantly updated and expanded and contributions from the community are welcomed.

3.5 Conclusions

In this work we have shown how a common framework for probabilistic inversion can be utilized to solve a diverse range of geophysical problems. In particular, the HMC method can be used to solve a variety of nonlinear geophysical inverse problems. In contrast to other MCMC algorithms, HMC exploits the information derived from the gradient of the posterior PDF to drive the sampling towards regions of high probability, hence being able to traverse the model space more efficiently. This property is very beneficial, especially for problems which allow efficient computation of the gradients, e.g., when using the adjoint method, analytical derivatives or automatic differentiation. In this paper we have shown a set of example problems including full-waveform acoustic inversion, hypocenter location, traveltime tomography and magnetic anomaly inversion using polygonal bodies. The examples presented have been solved using the software implementation HMCLab, a numerical laboratory for better understanding and solving inverse problems in a probabilistic manner, written in the high-level languages Julia and Python. By providing a collection of models as the solution of the inverse problem, HMCLab enables the user to perform a statistical analysis and retrieve desired probabilistic information. HMCLab is in constant evolution, and hopefully will be augmented by contributions from interested users. In addition to the available forward problems and priors, we made it accessible to the user to construct their own inverse problem and easily apply the methods provided by HMCLab. Moreover, several types of prior information are available, allowing the user to adequately describe prior certainties and uncertainties.

3.6 Data Availability

The general webpage of the project HMCLab can be found at https:/hmclab.science, containing all the relevant info and links to the software repositories. The Julia version of HMCLab is available in the public repository Gitlab at https://gitlab.com/JuliaGeoph, while the Python version is available in the public repository Github at https://github.com/larsgeb/hmclab.

3.7 Acknowledgements

(∗) The first two authors contributed equally to this work.

We wish to thank Klaus Mosegaard, who, over the years, has contributed much to the authors' understanding of Markov chain Monte Carlo methods and their use in Geosciences. We also would like to thank Sara Klaasen for sharing the Grimsvötn DAS dataset. L. G. would like to thank Xin Zhang and Andrew Curtis for fruitful discussions on Bayesian inference and Stein Variational Gradient Descent. Furthermore, thanks go out to anyone who dared to test the software in the experimental state before this publication to help us improve and debug HMCLab; Cyrill, Marc, Runa, Sara, Ariane, and to anyone who will contribute in the future. This work has been supported by the Swiss National Science Foundation ("Hamiltonian Monte Carlo Full-Waveform Inversion", grant number 200021_192236).

Chapter 4

Interrogating algorithms for free lunch

Chapter in preparation for submission to Inverse Problems, by L. Gebraad, X. Zhang, A. Zunino, A. Fichtner, and A. Curtis.

Abstract

This study investigates optimal appraisal algorithms within the context of Bayesian inference problems in geophysics, aiming to establish a concise, quantifiable performance metric. For Bayesian inference problems of small dimensionality, it is found that the entropy of an algorithm's output can be compared to the expected entropy of an inverse problem. The disparity between these entropies facilitates a relative numerical measure of algorithmic performance.

We apply this metric to an variety of widely used inference algorithms from the Markov chain Monte Carlo and variational inference categories. As our metric evaluates increasingly complex and non-linear inference problems, it reveals the inherent limitations of the tested algorithms. In these cases, clear performance differences for the tested algorithms can be observed, demonstrating that a specific algorithm can consistently and objectively outperform others within a subset of problems.

While optimal algorithms are identifiable for certain cases, it is discovered that due to the complexities of calculating entropies for high-dimensional distributions, our scoring method is inherently constrained to evaluating algorithms for low-dimensionality appraisal problems. This is due to the methods for calculating entropy becoming more resource-intensive in line with the curse of dimensionality—an effect that most sampling algorithms seek to circumvent—leading to a circular limitation.

4.1 Introduction

This study examines the implications of the No-Free Lunch theorems [Wolpert and Macready, 1997] in the context of geophysical problems. The No-Free Lunch theorems state that, averaged across all conceivable problems, all algorithms exhibit the same performance, with performance defined as the average across all possible metrics.

Our work aims to construct a relevant metric for assessing the quality of samples generated by Bayesian appraisal in the context of geophysical inversion. Using this metric, we explore the potential for certain algorithms to consistently outperform others within a specific class of problems.

Although the theory of Bayesian inference in parameter estimation settings seems elegant, in practical terms the analytical descriptions of the posterior density at any point are often inadequate. To effectively leverage Bayesian inference, meaningful questions need to be posed to the data. For example, a relevant question might be, "What are the most probable parameter values describing our physical system given the data?"

However, to answer such questions, integrals over the posterior distributions generated by Bayesian inference need to be calculated. As the complexity of these integrals increases with the number of parameters, standard techniques for evaluating integrals such as numerical quadrature fail. Over the past 70 years, a multitude of methods for computing these high-dimensional integrals has been developed. The aim of this work is to determine if any algorithm can be tested to consistently outperform others.

4.2 Bayesian inference

We start by reviewing the core elements of Bayesian inference and providing an extended interpretation of its components. This work assumes that the experimental design, numerical implementation, and the relationship between the model and data are all fixed.

Models are a collection of parameters that describe the state of nature. These parameters can be either continuous or discrete, while the collection itself can be finite or infinite in number of parameters. For the rest of this work, we won't consider models of infinite dimension. Typical examples in geophysics include discrete velocity maps of the subsurface, where each pixel, or voxel in three dimensions, represents a cell with a constant velocity, or the coefficients and spatial coordinates describing a moment tensor. In this work, we represent models as vectors **m** residing in the in linear model space $\mathscr{M}$, although in general one should consider models in manifolds [Tarantola, 2005]. This vector resides. The dimension of this space is often referred to as the dimension of the inverse problem, as it represents the space in which algorithms search for models that explain the recorded data.

Data is defined as a collection of observations, which are similarly organised in this work into a vector. This vector, denoted as **d**, could be a concatenation of multiple seismograms or the measured gravity at certain locations organised into a vector. Unless specified, the vector **d** can refer to observed, computed or hypothesized data. This vector exists in data space $\mathscr{D}$. Combined, a given model and data reside in the **joint model-data space**, $\mathscr{J}$, which has a dimensionality equal to $dim(\mathscr{M})+dim(\mathscr{D})$.

The forward relationship is the embodiment of a geophysicist's understanding and description of the process at play in the inverse problem. It is the mathematical or numerical operation that transforms a model into predictions. Examples might include a numerical wave propagation code or a lookup table for precomputed waveforms. Often referred to as the forward modelling relationship or simply the forward model, it is often denoted as $\mathbf{d} = F(\mathbf{m})$, although implicit relationships such as solution to differential equations might also be used. To complete the Bayesian interpretation of the data, data noise should be included, which may or may not depend on the model but is generally considered a stochastic rather than deterministic quantity, modifying the previous equation to $\mathbf{d} = F(\mathbf{m}) + \varepsilon(\mathbf{m})$.

Prior beliefs encapsulated in $p(\mathbf{m})$ summarise pre-experimental presumptions about what the subsurface might look like, in terms of the model parametrisation. It assigns a probability to any potential model $\mathbf{m} \in \mathscr{M}$. However, a significant knowledge gap might exist here; the understanding of a geological setting, for instance, does not straightforwardly translate into a discretisation and hence into a distribution over the used model space. As such, transposing complete prior beliefs into a proper mathematical formulation often proves challenging.

4.2.1 Data likelihood

The probability of any given datum being produced by a specific model, while adequately accounting for error statistics ε, is determined by the data likelihood. In an ideal scenario without noise, this likelihood is represented by:

$$p(\mathbf{d}|\mathbf{m}) \propto \delta\left(F(\mathbf{m}) - \mathbf{d}\right) \tag{4.1}$$

In this equation, δ represents the Dirac delta distribution. It should be noted that in this equation, $\mathbf{d}$ is a free parameter.

If one were to condition this equation with some observation ($\mathbf{d} = \mathbf{d}_{\text{observed}}$), this relationship implicates that a model is only admissible if it perfectly explains this data. This is often not realistic, thus demanding the introduction of error statistics. When ε is independent of $\mathbf{m}$ and uniformly distributed across all datapoints according to $\mathscr{N}(0, \sigma)$, a more accurate representation of the data likelihood becomes:

$$p(\mathbf{d}|\mathbf{m}) \propto \exp\left(-\frac{F(\mathbf{m}) - \mathbf{d}}{\sigma^2}\right) \tag{4.2}$$

Variations of this relationship exist, typically based on the assumed statistics of ε. The likelihood distribution can be seen as a function of $\mathbf{m}$, $\mathbf{d}$, or both.

Given that all likelihoods are defined up to a multiplicative constant, denoted by the 'proportional to' symbol $\propto$, the likelihood can be interpreted in model space, data space, or joint space. In practical terms, this means a conditional probability can be considered in the joint space or with either $\mathbf{d}$ or $\mathbf{m}$ fixed to a certain value, giving interpretations in the model space or data space, respectively.

Examples of these quantities include $p(\mathbf{d}|\mathbf{m}=\mathbf{m}_0)$, which denotes the probability of any given datapoint given a fixed model $\mathbf{m}_0$, and $p(\mathbf{d}=\mathbf{d}_0|\mathbf{m})$, representing the probability that a specific datum $\mathbf{d}_0$ is produced by any specific model $\mathbf{m}$. The latter interpretation is commonly used in parameter estimation problems. The normalisation constant is calculated by integrating the entire likelihood over the relevant space and ensuring the integral equals 1 by introducing a multiplicative constant.

4.2.2 The post-experimental knowledge on model and data

In this section, a deviation from the typical interpretation of Bayesian inference is done. Rather than applying Bayes' theorem directly, the joint knowledge derived from prior information and the forward relationship is considered first. The $\mathbf{d}$ vector is not fixed to any potential observations at this stage. The joint knowledge is obtained by multiplying the prior and the data likelihood, in accordance with the chain rule of probability:

$$p(\mathbf{m},\mathbf{d}) = p(\mathbf{d}|\mathbf{m})\,p(\mathbf{m}) \tag{4.3}$$

This distribution truly is only defined in the joint space. It describes the likelihood of any datum and model combination, integrating both the physics and the prior. If noise is Gaussian with standard deviation σ, the probability of any noise realisation is non-zero (since $\varepsilon \sim \mathscr{N}(0,\sigma)$ and thus $\varepsilon \in (-\inf,\inf)$), making the probability for any datum-model combination non-zero, irrespective of its absolute likelihood.

4

4.2.3 Bayes' theorem

Through Bayes' theorem, a third equality can be added to the previous equation that aids in answering questions about models under a given observation:

$$p(\mathbf{m},\mathbf{d}) = p(\mathbf{d}|\mathbf{m})\,p(\mathbf{m}) = p(\mathbf{m}|\mathbf{d})\,p(\mathbf{d}),$$

which is commonly written as:

$$p(\mathbf{m}|\mathbf{d}) = \frac{p(\mathbf{d}|\mathbf{m})\,p(\mathbf{m})}{p(\mathbf{d})} = \frac{p(\mathbf{m},\mathbf{d})}{p(\mathbf{d})}. \tag{4.4}$$

Here, $p(\mathbf{m}|\mathbf{d})$ denotes the probability of any model $\mathbf{m}$ under observation $\mathbf{d}$, considering prior beliefs.

4.2.4 The interpretation of $p(\mathbf{d})$

It's important to note that when $\mathbf{d}$ is fixed to a specific observation, the term $p(\mathbf{d})$ becomes a constant known as the evidence, or (data) normalisation constant. However, if one is interpreting distributions in the joint space $\mathscr{J}$ or data space $\mathscr{D}$, this term is the data distribution. In either scenario, it contains a priori information about the data. This can be clarified by expressing the term as a marginalisation of $p(\mathbf{m},\mathbf{d})$:

$$
\begin{aligned}
p(\mathbf{d}) &= \int_{\mathscr{M}} p(\mathbf{m},\mathbf{d})\, d\mathbf{m} \\
&= \int_{\mathscr{M}} p(\mathbf{d}|\mathbf{m})\, p(\mathbf{m})\, d\mathbf{m} \\
&= E_{\mathscr{M}}\left[p(\mathbf{d}|\mathbf{m})\right] \qquad (4.5)
\end{aligned}
$$

Here, $E_{\mathscr{M}}(\square)$ denotes the expected value of the term over model space, weighted using the prior distribution. This final line indicates that the term represents the expected probability of observing a given data point across all models, factoring in the prior distribution on $p(\mathbf{m})$. When fixed to a particular observation, it specifies the likelihood of observing that particular data point, averaged over all models, appropriately weighted with the prior. Conversely, when interpreted within the $\mathscr{D}$ space, it expresses the likelihood of any data point being observed across all a priori likely models.

As $p(\mathbf{d})$ is expressed as an integral, it's generally not feasible to find a direct expression for it. This can be intuitively understood by acknowledging that computing this quantity directly would require an inverse relationship for the forward model operator. This operator is in generally not available, with only the simplest of models allowing for the direct inverse relationship. It becomes wholly impossible to acquire when the forward mapping is not one-to-one, a scenario easily caused by the consideration of error statistics.

4.3 The Equality of Entropy

In order to evaluate the results of Bayesian inference appraisal algorithms, a mathematical descriptor is sought that can quantify the 'goodness' of a result relative to a baseline. In this context, we introduce the concept of entropy of a distribution:

$$
\begin{aligned}
\operatorname{Ent} p(\mathbf{m}) &= E_{\mathscr{M}}\left[-\log(p(\mathbf{m}))\right] \\
&= \int_{\mathscr{M}} -\log\left(p(\mathbf{m})\right) p(\mathbf{m}) d\mathbf{m}. \qquad (4.6)
\end{aligned}
$$

This differential entropy can be understood as a measure of the information content within a distribution. A distribution that is 'sharper' or more 'peaked' possesses lower entropy. It is worth noting that a distribution need not be localised to a single region in domain to have low entropy; even a 'sharp' multimodal distribution can exhibit low entropy as long as the information is concentrated at its modes.

4.3.1 Estimating the Entropy of a Distribution

The differential entropy for distributions of continuous random variables can be estimated through various methods. One such approach is through numerical quadrature. However, this method requires knowledge of the normalised probability density function, a condition that is not always met in practical applications .

An alternative method for entropy estimation is the histogram-based calculation. This technique requires only samples drawn from the distribution, which can be obtained from the unnormalised posterior density. The method bins these samples and determines the average probability density for each bin. Subsequently, a discretised version of Equation 4.6 is used to estimate the entropy

$$\mathrm{Ent} = -\sum_i p_i \log(p_i)\,\Delta V_i, \tag{4.7}$$

where p_i is the probability density of each bin, and ΔV represents the hypervolume of each bin. The accuracy of the histogram-based method largely dependent on the selection of the bin width and the total number of samples drawn. It should be noted that the memory requirements for this approach scales exponentially with the distribution's dimensionality, thus practically limiting the method to distributions of around 5-8 dimensions for 1TB of working memory, i.e. on high-end workstations or HPC resources. Under increasing number of samples drawn from the posterior, this method scales linearly.

Kernel density estimation (KDE) presents a third method for entropy estimation, which offers computational stability even in scenarios with fewer samples. This method fits a mixture model to the observed samples, after which their individual normalised probabilities are estimated. By taking the mean of the logarithmic probabilities of all N samples, the entropy is estimated

$$\mathrm{Ent} = -\frac{1}{N}\sum_i \log(p_i). \tag{4.8}$$

The accuracy of the KDE-based method depends strongly on the choice of bandwidth of the kernels. Unlike the histogram-based method, the KDE-based method can compute entropy efficiently in high-dimensional spaces, without the substantial memory requirements inherent to the histogram-based approach. It does, however, experience an exponential increase in the computational time required as the number of samples used for entropy estimation grows, which is necessary to maintain accuracy in high dimensional distributions [Silverman, 1986].

It's important to note that both histogram- and KDE-based entropy calculations are significantly constrained by the curse of dimensionality. While the memory requirements for the histogram-based method scale exponentially with the distribution's dimensionality, the KDE method is limited by the exponential increase in computational time needed to fit a larger number of samples from the distribution. Due to the dimensionality of the distributions involved, the calculations of the entropy for the joint distribution $p(\mathbf{m}, \mathbf{d})$ are the first ones to go out of scale.

4.3.2 Shewry and Wyn: A Litmus Test for Algorithms

Having defined the entropy, it is useful to introduce the following relation between the information of the posteriors, the data, and the joint distribution as stated by Shewry and Wynn [1987]:

$$E_{\mathscr{D}}\left[\mathrm{Ent}\, p(\mathbf{m}|\mathbf{d}, \xi)\right] + \mathrm{Ent}\, p(\mathbf{d}|\xi) = \mathrm{Ent}\, p(\mathbf{m}, \mathbf{d}|\xi). \tag{4.9}$$

Previous studies [van Den Berg et al., 2003] have utilised this equation to find the optimal experimental design ξ, by maximising the entropy in the prior data $\operatorname{Ent} p(\mathbf{d}|\xi)$. The term $\operatorname{Ent} p(\mathbf{m}|\mathbf{d},\xi)$ denotes the amount of information contained in the posterior distribution for a specific datum and experimental design. The calculation of this term's expectation over the entire data space $\mathscr{D}$ yields the expected information in a posterior, regardless of the data realisation. This, in turn, enables the evaluation of information in an experiment without necessitating actual observations.

One can adapt this equation to consider a single inverse problem at a time:

$$E_{\mathscr{D}}\left[\operatorname{Ent} p(\mathbf{m}|\mathbf{d})\right] + \operatorname{Ent} p(\mathbf{d}) = \operatorname{Ent} p(\mathbf{m},\mathbf{d}). \tag{4.10}$$

This equation implies that the entropy of a posterior $p(\mathbf{m}|\mathbf{d})$, averaged over the data distribution, equals the difference in information between the joint and data distributions. Notably, because the expectation over $\mathscr{D}$ is desired, re-expanding into the integral would require weighting by $p(\mathbf{d})$ again in the first term:

$$\int_{\mathscr{D}} \operatorname{Ent} p(\mathbf{m}|\mathbf{d})\, p(\mathbf{d})\, d\mathbf{d} + \operatorname{Ent} p(\mathbf{d}) = \operatorname{Ent} p(\mathbf{m},\mathbf{d}).$$

The crux of the proposed metric is this: the algorithms examined in this work aim to appraise posteriors $p(\mathbf{m}|\mathbf{d})$. In the typical implementations of these algorithms, this is achieved either by generating samples from the posterior or by generating samples for a kernelised parametrisation of the posterior, for Monte Carlo and SVGD methods, respectively. Both these outputs allow for the computation of the entropy of the appraisal. By substituting this entropy into Equation 4.10 and quantifying the discrepancy, a performance metric is constructed:

$$EoE = E_{\mathscr{D}}\left[\operatorname{Ent} p(\mathbf{m}|\mathbf{d})\right]_{\text{algorithm}} - \left[\operatorname{Ent} p(\mathbf{m},\mathbf{d}) - \operatorname{Ent} p(\mathbf{d})\right]_{\text{baseline}}. \tag{4.11}$$

This Error of Entropy (EoE), which is the difference between the entropy produced by the algorithms and the baseline entropy, measures the quality of the posteriors produced by an algorithm. Importantly, this approach does not require any actual data, since it quantifies an algorithm's performance under all prior likely data. Another way to understand why the EoE measures algorithm performance involves considering the baseline entropy $\left[\operatorname{Ent} p(\mathbf{m},\mathbf{d}) - \operatorname{Ent} p(\mathbf{d})\right]_{\text{baseline}}$. This quantity encapsulates how much information should, on average, be present in the posterior, as this is precisely the amount of information the data can extract from the joint distribution.

Importantly, Equation 4.9's terms, $\operatorname{Ent} p(\mathbf{d})$ and $\operatorname{Ent} p(\mathbf{m},\mathbf{d})$, can be calculated with relative ease provided the prior distribution and data error distribution are directly sampleable. To achieve this, a closer look at Equation 4.5 is warranted. By drawing a sample $\mathbf{m}_0$ directly from the prior distribution $p(\mathbf{m})$, one generates samples proportional to the second term of the integral. Using that sample $\mathbf{m}_0$ as input for the forward model, including the stochastic modelling of the errors, one generates a data sample proportional to $p(\mathbf{d}|\mathbf{m}_0)$. Repeating this process with multiple draws from the prior distribution generates samples $(\mathbf{m}_i, \mathbf{d}_i)$ with probability $p(\mathbf{m}_i)$ in the joint space $\mathscr{J}$. These samples are proportional to $p(\mathbf{d}|\mathbf{m})\, p(\mathbf{m})$, and by virtue of Equation 4.3, proportional to $p(\mathbf{m},\mathbf{d})$. From these samples, the entropy of the joint distribution $\operatorname{Ent} p(\mathbf{m},\mathbf{d})$ can be obtained.

Furthermore, marginalising these samples with respect to the components in $\mathbf{m}$ gives samples proportional to the data distribution. This marginalisation, or the process of taking the integral over $\mathscr{M}$, is as simple as disregarding the $\mathbf{m}$-component(s) of each sample, which yields samples proportional to $p(\mathbf{d})$. From these samples, the entropy of the data $\operatorname{Ent} p(\mathbf{d})$ can be readily computed.

The average entropy of the posteriors, $E_{\mathscr{D}}[\operatorname{Ent} p(\mathbf{m}|\mathbf{d})]$, can be calculated by drawing samples from $p(\mathbf{d})$ using the previously outlined method. Serving as mock-observed data, these samples aid in the generation of the posteriors, $p(\mathbf{m}|\mathbf{d})$. The algorithm under review evaluates these posteriors, and the entropy of each is determined based on the results. The mean entropy of these posteriors is subsequently calculated. The difference between this average and the baseline entropy is named the Error of Entropy (EoE).

The EoE method for testing an algorithm $\mathbb{A}$ is encapsulated in the following steps, where all tuning parameters of the EoE method are denoted by a capital N and an associated subscript:

1. Draw N_{prior} samples from the prior distribution $p(\mathbf{m})$.

2. Use each sample from the model prior as input for the forward model with appropriate error statistics $F(\mathbf{m}) + \varepsilon$ to draw samples from $p(\mathbf{d})$.

3. Calculate the entropy of the data distribution $\operatorname{Ent} p(\mathbf{d})$ from its samples, setting a fixed number of bins N_{bins}.

4. Direct sum each prior sample with its corresponding data sample, $\mathbf{m} \oplus \mathbf{d}$, to create an ensemble of samples proportional to $p(\mathbf{m}, \mathbf{d})$.

5. Calculate the entropy of the joint distribution $\operatorname{Ent} p(\mathbf{m}, \mathbf{d})$ from its samples, setting a fixed number of bins N_{bins}.

6. For each of the N_{copies} samples from $p(\mathbf{d})$:

   (a) Create the posterior $p(\mathbf{m}|\mathbf{d}_i)$ using data sample $\mathbf{d}_i$

   (b) Assess the posterior $p(\mathbf{m}|\mathbf{d}_i)$ using the algorithm $\mathbb{A}$, optionally limiting the number of evaluations to N_{eval}.

   (c) From the resulting samples for each posterior, calculate the entropy $\operatorname{Ent} p(\mathbf{m}|\mathbf{d}_i)$, setting a fixed number of bins N_{bins}.

   (d) From these entropies, compute the average entropy for all posterior appraisals $E_{\mathscr{D}}[\operatorname{Ent} p(\mathbf{m}|\mathbf{d})]$.

This process enables the generation of samples from all three terms in Equation 4.9. For the calculation of entropy, either the histogram or KDE-based method can be used used.

4.4 Testing algorithms

What one understands as an algorithm depends on context. When a study on geophysical inverse problems discusses the used methods, it might be common for it to state which algorithm was used, e.g. "for this appraisal we used the Gibbs sampling algorithm". However, in the context of the No-Free Lunch theorem, an algorithm is understood as a combination of the operations of what is typically understood as an algorithm, along with all of its settings, or colloquially, its tuning parameters.

In this work three algorithms are investigated. Below one will find the details on the operations of each algorithm, but each test of algorithm will also include associated tuning parameters.

4.4.1 Monte Carlo autotuning

The performance of MCMC algorithms depends on specific tuning settings. A commonly recurring tuning parameter across many MCMC algorithms is stepsize, which controls the distance of proposed moves from the current sample. By increasing the stepsize, the algorithm's mixing performance can be improved, as it allows the chain to attempt longer distance moves. However, thus also reduces the acceptance rate of proposals, as larger proposed moves are more likely to fall outside the region of probability, thereby being rejected by the target distribution. This trade-off between stepsize and acceptance rate is an important challenge in Monte Carlo methods, with the optimal stepsize for a given problem often not immediately apparent [Neal, 2011].

One heuristic for setting the stepsize is to aim for an acceptance rate that is neither excessively high nor low. This can be accomplished by iteratively adjusting the stepsize during the sampling process, based on the observed acceptance rate. However, this method can inject memory into the sampling process, thereby violating the Markov property and potentially leading to biased estimates. To mitigate this issue, one can use separate tuning stages in which an isolated chain is used to fine-tune the stepsize and other parameters prior to generating samples for analysis.

This study utilises asymptotic tuning to automate the tuning and analysis process within a single chain, by adjusting the stepsize over time to approach an optimal value [Zunino et al., 2023]. Asymptotic tuning offers improved efficiency over iterative tuning, as it negates the need for additional sampling or user input to estimate the optimal stepsize. However, it can introduce further bias if the asymptotic behaviour of the chain is not well understood, or if the algorithm is finely tuned to the local geometry of the distribution.

4.4.2 Random Walk Metropolis-Hastings

The Random Walk Metropolis-Hastings algorithm (RWMH) is a variant of the Metropolis-Hastings (MH) algorithm, initially proposed by Metropolis et al. in 1953 [Metropolis et al., 1953]. The Metropolis algorithm introduced a symmetric proposal distribution. The Random Walk Metropolis-Hastings algorithm, as we know it today, incorporates developments from the 1970 paper by Hastings [Hastings, 1970]. In this work, Hastings generalized the algorithm to allow for a wider class of proposal distributions, including non-symmetric ones. The key innovation was the introduction of the ’Hastings correction’ to account for the asymmetry in the proposal distribution. Specifically, RWMH generates new samples by proposing a move around the current sample using a Gaussian distribution. The mean of this distribution, denoted as μ, is set to the model coordinates of the current sample, and the covariance, denoted as S, serves as a tuning parameter for the algorithm. This covariance defines a hyper-ellipsoid proposal volume.

The random walk version of the MH algorithm is powerful as it can generate new samples with a relatively high posterior likelihood by conducting a local search around the current sample. This might not be the case when using the prior distribution as the proposal distribution for the MH algorithm in scenarios where the prior is considerably different from the posterior, as this would result in a relatively low acceptance rate in regions with large discrepancy.

The covariance S is typically simplified to $s^2 I$, where s represents the step size, and I is the identity matrix corresponding to the target’s dimensionality. By utilising a full covariance matrix, one could introduce prior knowledge on the Hessian structure of the posterior. While this does not alter the MCMC result in the limit, it can typically expedite sampling. If only stepsize is used to tune RWMH, the proposal volume becomes a unit hypersphere scaled by the stepsize.

A notable drawback of the local proposal strategy is that RWMH exhibits random walk behaviour, leading to poor mixing and slow convergence. The chain might take a substantial amount of time to explore the entire parameter space, particularly in the presence of multiple local minima. The adverse effects of this behaviour are tied to the stepsize, or covariance, of the proposal distribution. Using a larger stepsize generates distant proposals that ameliorate these unfavorable properties, while using a smaller stepsize exacerbates their effects. Unfortunately, increasing the stepsize also leads to a lower acceptance rate, as the sampler seeks new likely models in a significantly larger volume in model space.

As the model space dimension increases, the exponetially growing volume of the proposal distribution results in a decrease in acceptance rate for a specific stepsize. As such, to maintain acceptance rates, stepsizes are reduced for RWMH applied to high-dimensional posteriors. This results in an unfavorable scaling property of the algorithm as the problem size increases [Creutz, 1988, Neal, 2011].

4.4.3 Hamiltonian Monte Carlo

The Hamiltonian Monte Carlo (HMC) [Duane et al., 1987] algorithm is a complex variation of the standard MH algorithm. It implements two significant alterations to the traditional methodology. Firstly, it expands the target distribution with a distribution over auxiliary variables. This resultant distribution is referred to as the canonical distribution. Secondly, instead of using a proposal distribution for the canonical distribution, it directly generates new coordinates for the auxiliary variables (analogous to blocked Gibbs sampling) and then applies Hamiltonian dynamics to evolve the original and auxiliary variables together, proposing a new state.

The proposal process can be interpreted within the framework of Hamiltonian dynamics, drawing parallels to physical phenomena. The original variables of the posterior distribution are perceived as the coordinates of an N-dimensional particle, while the auxiliary variables are considered the particle's momentum. The negative logarithmic probabilities of the posterior and momentum variables, $-\log p(\mathbf{m})$ and $-\log p(\mathbf{q})$, can then be interpreted as the potential and kinetic energies of this system, $P(\mathbf{m})$ and $K(\mathbf{q})$ respectively. Consequently, this system models a particle moving in a potential field described by P, and its behaviour can be predicted using Hamilton's equations.

This physical interpretation leads to efficient exploration of the parameter space because the Hamiltonian dynamics naturally move the particle towards regions of higher posterior probability. The HMC algorithm generates candidate moves by evolving the system for a fixed amount of "time" according to Hamilton's equations, using numerical integration methods.

The key advantage of HMC over e.g. RWMH is its ability to propose distant moves while still maintaining a high acceptance rate, which leads to more efficient exploration of the parameter space. This property makes it particularly suited for high-dimensional problems where the acceptance rate of RWMH drops.

The algorithm introduces three key tuning parameters: the step size for the numerical integration, the "time" for which the system is evolved, and the arbitrary distribution on the auxiliary variables. The step size for the numerical integration functions much like the stepsize of RWMH, and can be tuned accordingly.

The numerical requires the gradient of the posterior distribution to be available. The integation time, defined by the number of steps taken by the numerical integration, controls how many gradient evaluations are performed during a proposal.

Lastly, the distribution of the auxiliary variables functions exactly the same as the preconditioning matrix in RWMH. It allows one to correct for large variations in curvature and correlation. In the physical intepretation of HMC, it can be regarded as the multidimensional mass matrix of the hypothetical particle.

4.4.4 Stein Variational Gradient Descent

Stein Variational Gradient Descent [SVGD, Liu and Wang, 2019] belongs to the class of variational inference algorithms. Instead of generating samples from the posterior, SVGD represents the target distribution through a set of particles. These particles are updated using optimisation techniques, specifically by minimising the Kullback-Leibler divergence between the particles and the true posterior, evaluated at the particle locations. This methodology allows SVGD to generate globally dispersed samples, mitigating some of the issues of local exploration that traditional MCMC methods encounter. Additionally, since all samples are updated simultaneously without requiring communication, their computations can be executed in parallel during each iteration of the algorithm, making SVGD well suited for modern HPC systems.

The key tuning parameters in SVGD are the choice of kernel function, its bandwidth, the number of samples that initialise the algorithm's ensemble, and finally the tuning parameters of the optimisation algorithm. The kernel function and its bandwidth define the similarity between particles. The bandwidth parameter is particularly crucial as it determines the range of influence or the spread of the kernel, and is therefore integral to the performance of the algorithm. In many respects, it is analogous to the stepsize in MCMC methods, as it scales the ensemble to match the posterior. The number of samples in the ensemble determines the final size of the posterior samples. A larger sample size enables more precise estimators of integrands of interest. However, it also linearly impacts the computational time required for SVGD. Finally, the parameters for the optimisation algorithm are essential as well. This depends on the choice of optimisation algorithm.

The tuning parameter for bandwidth can be autotuned using the 'median trick'. This approach involves computing the pairwise Euclidean distances between all pairs of particles and selecting the median of these distances as the bandwidth parameter. The use of median distance as the bandwidth allows the RBF kernel to adaptively change size as the local structure, or Hessian, of the posterior changes when the samples are updated [Liu and Wang, 2019].

The SVGD algorithm also has a physical interpretation, in that pulls the particles towards the high-probability regions of the target distribution as if the posterior were a potential, counterbalanced by the particles repelling static forces defined by their kernels.

4.4.5 The symmetries between the algorithms tuning parameters

We posit that the parameters used to tune each of the three algorithms in this manuscript have strong similarities. Firstly, all three algorithms can be autotuned to account for the unknown scale of the posterior. For both Monte Carlo algorithms - RWMH and HMC - this is achieved by autotuning based on the acceptance rate. For SVGD, this is done using the median trick. Therefore, it is expected that the effect of these parameters and their automatic tuning methods help decouple the performance of the algorithms from the scale of the distribution.

Secondly, the curvature of a target distribution can also be accounted for in all three algorithms. The curvature is a multivariate effect that goes beyond the absolute scaling of the posterior in all dimensions equally. Curvature describes how the individual parameters are dispersed and correlated. This can be described locally by the Hessian, or second derivative, of the posterior, but can vary in a non-Gaussian way across the posterior. Accounting for this effect is done by controlling the proposal distribution, auxiliary distribution or kernel distribution for RWMH, HMC and SVGD respectively.

The basic variations of each algorithm can account for a multivariate Gaussian curvature of the posterior, that is, a negative log of the distribution that behaves like a quadratic polynomial. However, any distribution which can be sampled from can in principle be incorporated as a preconditioner through either of the three preconditioning distributions. The effects of preconditioning on non-Gaussian posteriors is not investigated in this work. For SVGD, non-Gaussian kernels can be tuned for automatically [Ai et al., 2022]. Similar approaches for the mass matrix in HMC can be constructed using (L-)BFGS style gradient accumulationFichtner et al. [2021].

	RWMH	HMC-3	HMC-10	SVGD-50	SVGD-500
Iterations	As limited by the numerical test, based on allowed $f(\mathbf{m})$ evaluations				
Preconditioner	Unit	Unit	Unit	Unit	Unit
Stepsize	Autotuned	Autotuned	Autotuned	—	—
Integration steps	—	3	10	—	—
Burn-in	20%	20%	20%	—	—
Kernel size	—	—	—	Median trick	
Optimiser	—	—	—	Gradient descent	
$f(\mathbf{m})$ eval. per iteration	1	4	11	50	500
Final ensemble size	Eval.	$\sim$ Eval./4	$\sim$ Eval./11	50	500

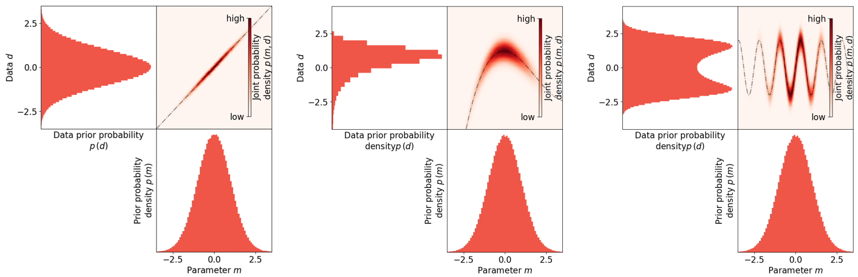
Table 4.1: Tuning settings for the five different instances of the algorithms under test. For RWMH, HMC, and SVGD, the preconditioner corresponds to the proposal distribution, mass matrix, and kernel function, respectively. "Unit" refers to the unit normal distribution.

4.5 Algorithm evaluations

To evaluate the performance of RWMH, HMC, and SVGD algorithms, numerical Expected Posterior Entropy (EoE) tests on inverse problems of varying dimensionality and complexity are performed. A comparison is conducted between one instance of the RWMH algorithm, two instances of the HMC algorithm, and two instances of the SVGD algorithm. The HMC instances differ in the number of integration steps, while the SVGD instances vary in initial ensemble size. A detailed description of all tuning parameters can be found in Table 4.1.

In the context of the 'no free lunch' theorems, the number of iterations an algorithm performs is considered a tuning parameter. In these tests, the number of iterations of each algorithm is limited by a maximum number of model $f(\mathbf{m})$ evaluations. The idea is, that with an equal number of model evaluations, algorithms are evaluated based on their performance within a given computational budget. Since SVGD and HMC require the calculation of the gradients of the posteriors, $\frac{\partial}{\partial \mathbf{m}} f(\mathbf{m})$, these evaluations are considered additional model evaluations, and thus detract from the overall budget. This one-to-one cost is rationalised by the fact that, in cases of analytical derivatives, the gradients typically present similar complexities. In more practical cases of numerical models, the adjoint method allows for the computation of gradients at roughly the same order of magnitude of computational cost as a forward evaluation. In adapting the presented method, one can scale the weighting of the evaluation of the gradient of the forward model to correspond to relatively costlier gradients.

To calculate the expected posterior entropy, each algorithm is run multiple times with different realisations of the data for a specified number of forward model evaluations. The number of allowed evaluations varies in logarithmic spacing from 10^0 to 10^7 across 40 grid points. At each point, all algorithms are evaluated 50 times for different realisations of the data. This amounts to a total of 10'000 algorithm runs per forward model for the five algorithms tested, with runs at higher evaluation numbers being significantly more computationally demanding than runs with fewer evaluation numbers. This computational



(a) Linear $f(\mathbf{m})$

(b) Cubic $f(\mathbf{m})$

(c) Sinusoid $f(\mathbf{m})$

Figure 4.1: The prior, joint and data distributions of the three tested models of dimensionality $n = 1$. If one were to condition on a specific datum, i.e. create a posterior $p(\mathbf{m}|\mathbf{d})$, that would mean to create a slice of the joint distribution at that given datum. The dotted lines in the panel of the joint densities are the forward models. This shows why posteriors can be non-unique, i.e. through noise in all cases, and through the absence of one-to-one mappings in the non-linear cases.

demand is facilitated through parallel evaluation of the algorithms on high-performance computing systems and local workstations.

4.5.1 Linear models

First, we test the method on linear models of the form

$$\mathbf{d} = G\mathbf{m} \tag{4.12}$$

where G is an arbitrary matrix of dimensions $n \times n$. We simplify this even further by setting G equal to the unit matrix I_n. A Gaussian noise model with a standard deviation of $\sigma = 0.1$ is used, yielding the prior, joint, and data distributions as given in Figure 4.1a for $n = 1$. This results in a 1-dimensional inverse problem, with a 1-dimensional data space and a 2-dimensional joint space. The same linear model (i.e., $G = I_n$) is tested for dimensionality 2 and 3. Of note is that the combination of a Gaussian prior with a linear model and a Gaussian noise model creates a Gaussian posterior, with characteristic function equal to the least squares solution.

The resulting entropies for dimensionality 1 are shown in Figure 4.2a. It can be clearly seen that as the number of iterations increases, all models significantly alter their average results. All three Monte Carlo methods asymptotically converge to the expected posterior entropy, with the RWMH algorithm showing the best convergence, at around 10^3 number of evaluations. Furthermore, the HMC algorithms show a similar level of convergence after an order of magnitude more model evaluations, around 10^4 model evaluations. The comparison of the two variants of HMC illustrate that in the case of a 1 dimensional linear inverse problem, the expenditure of more computational power per sample, in the form of more integration steps, is not beneficial to the rate of convergence of the entropy. As all Monte Carlo algorithms sequentially grow the ensemble, their initial entropies start out at low values and increase towards the expected posterior entropy.

For the two SVGD variants, one sees that these start out with a relatively high entropy. As these algorithms are initialised from the prior distribution, they start out more dispersive than the posteriors. The SVGD variant with 50 samples naturally has a lower entropy, as it spans less of the prior with fewer samples in the same volume of model space. Interestingly, for either SVGD configuration, the asymptotic entropy value as the number of evaluations tends to infinity is not the expected posterior entropy. This means that in the limit, the SVGD algorithm does not converge to the true solution. However, the SVGD algorithms seem to converge faster to their asymptotic value than the HMC algorithms, with an ensemble of 50 samples reaching convergence after 10^3 model evaluations. The SVGD algorithm seems to underestimate the dispersion of the posterior in the limit but does transition through the correct entropy. This additionally means that from a single evaluation of the EoE, it might be impossible to tell if an algorithm has performed well or simply found the correct entropy by chance. This highlights a limitation of our method: the correct entropy cannot fully capture the entire posterior, as no other quantity except the full posterior can.

As the SVGD algorithm only uses a predetermined number of samples in its ensemble to describe the posterior, the concern existed that this limited number of samples might actually describe the posteriors better than was shown from the EoE test, which do not take into account the kernel function. To test for this, the entropy on posterior sample ensembles created by multiplexing the SVGD results was computed. Using the kernel trick, the bandwidth of the kernel was estimated for the final state of the SVGD-produced ensemble, after which the generation of derived samples from the ensemble was made possible by repeatedly selecting ensemble members at random and adding Gaussian noise proportional to the bandwidth. However, the resulting multiplexed ensembles showed no improvement in EoE score towards the target posteriors.

As the dimensionality of the inverse problem increases, the asymptotic behaviour of the Monte Carlo algorithms occurs at an exponentially higher number of model evaluations, meaning the algorithms need more evaluations to converge. Interestingly, increasing dimensionality of the inverse problems does not seem to influence the convergence point of the SVGD algorithms noticeably in the linear forward model case. However, for the SVGD instances, the EoE score does drop with increasing dimensionality.

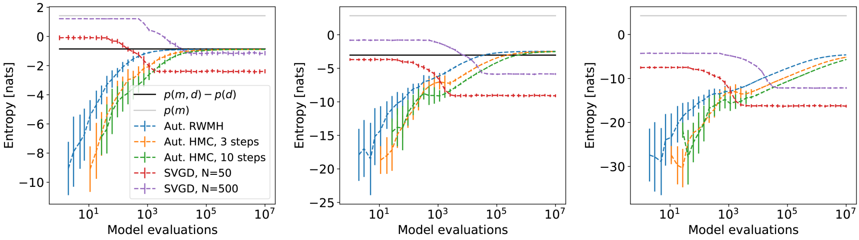
4.5.2 Non-linear models

The algorithms are subsequently tested on models with non-Gaussian curvature, potentially having local minima. The two models considered are a cubic polynomial, defined as

$$f(\mathbf{m}) = 0.1\mathbf{m}^3 + 0.1\mathbf{m}^2 - 1.5\mathbf{m} - 0.9, \tag{4.13}$$

and a sinusoid, defined as

$$f(\mathbf{m}) = 2\sin(5\mathbf{m}). \tag{4.14}$$



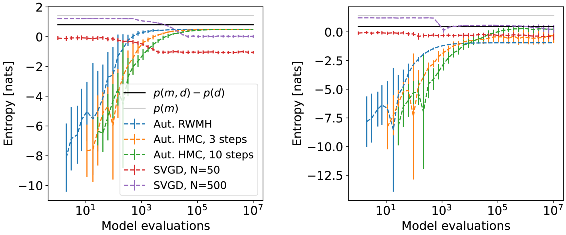
(a) $n = 1$ (b) $n = 2$ (c) $n = 3$

Figure 4.2: Entropy terms for the linear models of dimensionality $n = 1$, 2, and 3, calculated using the histogram-based method with 200 bins. The EoE is not directly shown, however, the left and right-hand side of the required equality are shown, meaning that a well-performing algorithm converges with a minimum amount of model evaluations to the black line of expected posterior entropy $p(\mathbf{m}, \mathbf{d}) - p(\mathbf{d})$. The vertical bars indicate the variability in the posterior entropy. For $n = 3$, the computation of the entropy of the joint distribution in 6-dimensional space with high enough precision becomes intractable.

These two forward models and their associated probability density functions can be found in Figures 4.1b and 4.1c, respectively.

The results of the cubic polynomial forward model are very similar to those of the linear models, showing convergence of the Monte Carlo algorithms around the same number of iterations. Similarly, the SVGD algorithms show convergence after approximately 10^4 model evaluations, but to entropies that are significantly lower than the expected posterior entropy from the EoE expression.

Interesting behaviour appears for algorithms appraised on the sinusoid model. The Monte Carlo algorithms seem to stagnate far below the expected posterior entropy. Notably, both HMC configurations come closer to the expected entropy than RWMH, outperforming RWMH for a given number of model evaluations for the first time. As the number of evaluations increases beyond 10^5 evaluations, the HMC algorithm using 10 integration steps diverges from its previous asymptote, converging closer to the expected posterior entropy. In contrast, both the RWMH algorithm and the HMC algorithm with 3 integration steps fail to exhibit this behaviour before the maximum number of evaluations tested. This suggests that these algorithms appear to be limited by the strong multimodality of the posteriors in this model, as the lower entropy asymptote seems to indicate. The HMC algorithm with 10 integration steps can escape these local minima and performs a more efficient global search, in the number of evaluations tested. It is expected that all Monte Carlo algorithms would show this same behaviour in the limit of infinite samples, as predicted by the theoretically guaranteed convergence of MCMC methods. For the SVGD algorithm with an ensemble size of 50, it is difficult to estimate the number of model evaluations required to reach an asymptote, as the algorithm seems to decrease its entropy very slowly as the number of evaluations increases. On the other hand, the configuration with an ensemble size of 500 moves to its asymptote relatively quickly. For an ensemble



(a) Cubic $f(\mathbf{m})$

(b) Sinusoid $f(\mathbf{m})$

Figure 4.3: The entropy terms for the EoE tests on the non-linear and sinusoid models.

that spans the prior well, SVGD is expected to handle multimodal posteriors effectively, as it can initially capture much of the multimodality by initialising the ensemble over a large volume of the prior.

4.6 Discussion

It is rather unfortunate that the very curse of dimensionality that motivates the development of appraisal algorithms also limits the extent to which it is possible to characterise their performance. When characterising inverse problems it becomes apparent that algorithms show differing behaviour. RWMH excels in posterior distributions similar to Gaussians, whereas gradient-based algorithms clearly outperform RWMH on multimodal posteriors and posteriors without constant curvature. The limit on entropy estimation, unfortunately, precludes a thorough investigation of the scaling relationships of these algorithms as the dimensionality increases.

While Monte Carlo methods are developed for the generation of independent and identically distributed samples, variational inference methods are designed around the aim of minimising the KL-divergence between the approximation and true posterior. This difference in algorithm intent is not clearly reflected in the entropies produced for the distributions investigated. However, the way the Monte Carlo and SVGD algorithms approach the expected posterior entropy, from either a less or a more dispersive initial ensemble respectively, can be used to design conservative experiments, where either under- or overestimation of uncertainty is preferred in case of any errors due to limited evaluations from computational constraints.

4.6.1 Computational limitations

Estimating the entropy of the joint distribution $\operatorname{Ent} p(\mathbf{m}, \mathbf{d})$ exacerbates the limitations of entropy estimation. As the highest dimensional object in EoE calculations, this distribution is the first to hit computational resource limits. The memory scaling of the histogram method is exponential. For instance, for a 6-dimensional joint distribution, this requirement is about 90 terabytes.

The use of a fitted KDE model only seems to mitigate this issue initially. Calculating entropy on a 6-dimensional distribution requires very limited memory, but to achieve sufficient precision, the logarithmic probabilities for each drawn sample must be calculated, requiring the normalisation of the distribution. The computational time for KDE estimates of entropy scales exponentially with an increasing ensemble size.

To apply the EoE method to larger inverse problems, alternative methods for calculating entropies for high-dimensional distributions should be considered. For example, adaptive or sparse binning strategies in histogram methods [e.g. as included in Brun and Rademakers, 1997] might resolve the memory issues encountered in the histogram-based approach.

4.6.2 Parallelisation in Bayesian inference

Although the quantification of performance could only be done for small-scale problems in this work, it is important to note that all the algorithms discussed in this work have been successfully applied to more computationally demanding practical cases [Fichtner et al., 2018b, Gebraad et al., 2020, Zhang and Curtis, 2021, Zhang et al., 2023a, Zunino et al., 2023]. For large-scale problems, this process typically involves a high degree of multi-node parallelisation to effectively utilise modern supercomputing resources. In this sense, the curse of dimensionality does not limit the practical usage of the discussed algorithm, solely the practical scoring.

MCMC methods produce i.i.d. samples in the limit of many iterations, allowing separate MCMC results to be combined into a single ensemble of samples without altering the original distribution. This is best done by ensuring similar burn-in and mixing characteristics for all separate chains. Multiple works demonstrate how parallel non-communicating Markov chains can be used to accelerate convergence on high-dimensional or computationally demanding targets [Murray, 2010, Laskey and Myers, 2003]. MCMC mixing can be further enhanced in parallelisation using methods such as replica exchange [Earl and Deem, 2005] or data-partitioning [Neiswanger et al., 2014, De Souza et al., 2022].

SVGD can be parallelised without altering the algorithm, with each update to the ensemble requiring multiple independent evaluations of the posterior likelihood. These evaluations can be performed on separate compute nodes, and the results used for an update on a single node. As a result, this method scales well on multi-node resources.

However, despite both sampling and variational methods being applicable to large-scale geophysical inverse problems, the actual quantification of entropies in these large dimensionality problems remains out of scale. SVGD and HMC might specifically avoid sampling large low-probability spaces by utilising gradients of the posterior. However, the required histogram operations as well as the sampling of the joint distributions make our current approach out of scale for problems with dimensionalities (prior model and prior data combined) higher than approximately 4, even on HPC resources.

4.6.3 Conclusion

This study provides insights into how algorithm performance can be quantified. It highlights the different behaviours of these algorithms, with RWMH excelling in simple Gaussian distributions, and gradient-based methods performing better for multimodal posteriors and those lacking constant curvature.

What is striking is the immediate reduction in accuracy of any method as the dimensionality of the inference problems increases. Per the defined EoE metric, the number evaluations required to reach similar conservation of entropy with MCMC algorithms seems to grow exponentially with the number of dimensions. This effect is not seen for variational methods, but these methods never reach similar performance.

The study also underlines the challenges in assessing algorithm performance in high-dimensional spaces, especially due to the computational limitations of entropy estimation. The curse of dimensionality is an inherent limit that pervades all aspects of optimisation theory. Therefore, future research should focus on developing methods to efficiently estimate entropy and approximate posteriors in high dimensions, as these are crucial for evaluating complex, real-world inverse problems.

4.7 Acknowledgements

The authors would like to extensively thank the HPC-Europa3 programme for enabling the collaboration between co-authors. The work has been performed under the Project HPC-EUROPA3 (INFRAIA-2016-1-730897), with the support of the EC Research Innovation Action under the H2020 Programme; in particular, the author gratefully acknowledges the support of Andrew Curtis, of the School of GeoSciences at the University of Edinburgh and the computer resources and technical support provided by the High Performance Computing facilities at EPCC at the University of Edinburgh.

Chapter 5

psvWave: elastic wave propagation in 2d for Python and C++

Chapter submitted as preprint L. Gebraad and A. Fichtner. psvWave: elastic wave propagation in 2d for Python and C++. *EarthArXiv*, Feb. 2022b. doi: 10.31223/X5R91Q. This chapter is not intended to be published in a peer-reviewed fashion, but serves to document the development of a sizeable software package.

Abstract

We present 'psvWave', a basic numerical finite difference solver for Python and C++, specifically targeted at seismologists. The solver is based on the well-established staggered grid approaches developed for the P-SV elastic wave equation. Although its functionality is limited (solely moment tensor sources, only Ricker wavelets source time functions), it does possess the ability to perform adjoint simulations, and its performance has so far allowed the development of Bayesian sampling for Full-Waveform Inversion using the Hamiltonian Monte Carlo algorithm. We present this as an open source project, and invite anyone to contribute.

5.1 Introduction

This software was born out of the need to simulate many small wavefields quickly, from a Python environment. As such, we implemented Virieux's Virieux [1986] seminal work in a C++ OpenMP enabled code that interfaces with Python. Configuration and functionality is minimal, and therefore so is overhead. Its fast performance has so far allowed it to be used for Hamiltonian Monte Carlo sampling with hundreds of thousands of simulations performed in a short time Gebraad et al. [2020]. The software can perform 'forward' as well as 'adjoint' computations, thereby facilitating the computation of sensitivity kernels relevant to various inverse theory methods Fichtner et al. [2006a]. It should be noted that the main aim of psvWave is research, and we do not consider it suited for production.

5.2 The solver and Python interface

The psvWave package contains a forward and an adjoint 2d elastic wave equation solver. The staggered grid as well as the leap-frog time integration are equal to that described in Virieux [1986]. Additionally, the C++ core allows one to compute misfits w.r.t. some observed data, and to subsequently calculate sensitivity kernels using the resulting adjoint sources. The Python interface gives access to all core functionality, but also extends the C++ functionality by providing plotting functions.

The simulations performed make a few basic assumptions about the medium, wavefield and sources, as given below.

5.2.1 Assumptions

All sources propagate waves through the same medium / domain in the x,z-plane, and are recorded by the same network of receivers. The physics are for in-plane shear waves and defined in a right-handed coordinate system. However, one can interpret the simulations in any unit and orientation. One should make sure that all units used result in wavefields that are within the range of C++ doubles.

All sources are normal / reverse faults (with strike parallel to the y-axis) using a Ricker wavelet as source time function. Every source can have a different dip angle. This source time function can be altered in both the Python and C++ API, the focal mechanism / source type not.

Simulations are divided in shots, i.e. a single time length in which data is recorded and some sources fire. The code allows for time staggering of sources, i.e. firing multiple sources in a single simulation.

The domain is truncated on all 4 sides by absorbing boundary conditions. It's width is variable, but as of yet, the same on all sides. This does not directly allow for free boundary conditions, but this is planned to change. When measuring distance or counting gridpoints, the zero-point is the first points not inside the boundary layer but in the actual simulation medium. When updating medium properties within the domain, the boundary copies the medium properties closest to it, to avoid creating reflectors.

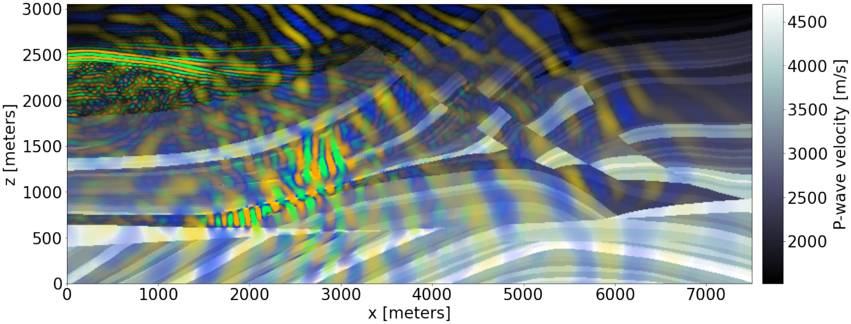


Figure 5.1: Snapshot at time index 120 for P-wave (yellow-blue) and S-wave (orange-green) particle motion from Notebook 1 superimposed on the P-wave velocity model of the elastic Marmousi model.

The location of the sources and receivers is not expressed in distance, but in gridpoint numbering. Because the actual indexing starts within the medium, and not the absorbing boundary, sources and receivers can only be placed inside the medium. However, the and variables determine how many gridpoints are not considered free parameters. The idea behind this is that this allows us to place sources/receivers in regions of the domain that are not inverted for, and are also not inside the boundary. This to avoid near-source and near-receiver effects.

5.2.2 Known limitations

The solver does not employ checkpointing for the forward simulation when it is saved for the imaging condition in the adjoint equation, but simply stores the the forward dynamical fields fields (particle motion, strain) at specific intervals. This setting, found in the configuration files as `inversion.snapshot_interval` allows one to reduce the storage in host machines RAM, but also deteriorates the precision of the computed sensitivity kernels if it is chosen too high.

Additionally, all simulations store their wavefields separately. By growing the number of simulations in a single modeller (by e.g. adding more sources), the memory required to store the wavefields (for subsequent adjoint simulations) grows linearly. This could be circumvented by running all forward and adjoint simulations sequentially and re-using the memory used between sources, but is as of yet not implemented.

The misfit that is implemented in the package, with its adjoint source, is currently only the L2 misfit. The documentation illustrated how one can smooth the gradients for this misfit.

5.3 Availability, instructions and tutorials

In the project repository [Gebraad, 2022a], we provide installation instructions for PyPi installation, as well as a dedicated Docker image. The PyPi installation only supports Linux AMD64 architectures. The Docker is fully multiplatform and requires no set-up. Additionally, the repository provides two notebooks. Notebook 1 concerns itself with the configuration file and performing forward simulations, as well as the visualization of the outputs. Notebook 2 demonstrates how psvWave combined with the L-BFGS [Nocedal, 1980] algorithm allows one to perform deterministic full-waveform inversion.

Acknowledgements

We would like to thank all the open source software projects that allow psvWave to function. Specifically, we make heavy use of Eigen [Guennebaud et al., 2010], PyBind11 [Jakob et al., 2017], NumPy [Harris et al., 2020], SciPy [Virtanen et al., 2020] and Matplotlib [Hunter, 2007]. Our code runs on any platform thanks to Docker. This document is a non-peer reviewed work published on EarthArXiv.org.

Chapter 6

GPU physics with Apple M-series

Chapter published as L. Gebraad and A. Fichtner. Seamless GPU Acceleration for C++-Based Physics with the Metal Shading Language on Apple's M Series Unified Chips. *Seismol. Res. Lett.*, 94(3):1670–1675, Feb. 2023. doi: 10.1785/0220220241.

Abstract

The M series of chips produced by Apple have proven a capable and power-efficient alternative to mainstream Intel and AMD x86 processors for everyday tasks. Additionally, the unified design integrating the central processing and graphics processing unit, have allowed these M series chips to excel at many tasks with heavy graphical requirements without the need for a discrete graphical processing unit (GPU), and in some cases even outperforming discrete GPUs.

In this work, we show how the M series chips can be leveraged using the Metal Shading Language (MSL) to accelerate typical array operations in C++. More importantly, we show how the usage of MSL avoids the typical complexity of CUDA or OpenACC memory management, by allowing the central processing unit (CPU) and GPU to work in unified memory. We demonstrate how performant the M series chips are on standard one-dimensional and two-dimensional array operations such as array addition, Single-Precision A·X Plus Y and finite difference stencils, with respect to serial and OpenMP accelerated CPU code. The reduced complexity of implementing MSL also allows us to accelerate an existing elastic wave equation solver (originally based on OpenMP accelerated C++), while retaining all CPU and OpenMP functionality without modification.

The resulting performance gain of simulating the wave equation is near an order of magnitude for large domain sizes. This gain attained from using MSL is similar to other GPU-accelerated wave-propagation codes with respect to their CPU variants, but does not come at much increased programming complexity that prohibits the typical scientific programmer to leverage these accelerators. This result shows how unified processing units can be a valuable tool to seismologists and computational scientists in general, lowering the bar to writing performant codes that leverage modern GPUs.

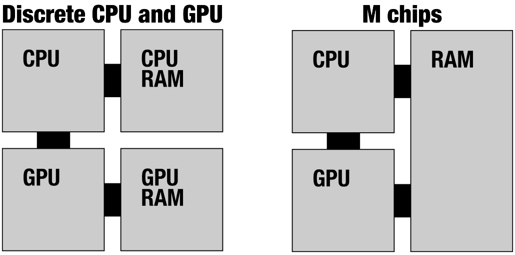


Figure 6.1: Schematic design of discrete processing unit systems versus the M series chips. On M systems, both processing units can talk to the same random access memory (RAM) without moving data locations.

6.1 Introduction

Scientific computing has always been at the forefront of technological developments in computing. On the point of the latest ARM chips produced by Apple, the M series, the scientific community should act no different. This series of chips is a relatively new component of MacBooks produced by Apple. The fact that these are ARM based chips means that their instruction sets are fundamentally different from typical notebook, workstation and HPC suited processors from Intel or AMD. According to Apple, the usage of ARM chips in high performance notebooks and workstations increased performance and power efficiency[Apple, 2020].

More importantly, however, is that the design of the M series chips is such that they combine the central and graphical processing units (CPU and GPU) onto a single chip, creating a 'Unified' Processing Unit, something atypical for modern systems. This means that both processing units can communicate with the same memory, simplifying many operations. The differences in design are schematically illustrated in Figure 6.1.

The Metal Shading Language (MSL), originally developed for the iPhone iOS operating system, is a programming language tailored to address GPU hardware present on Apple ARM chips in general. Although it has been in development for approximately a decade, it only saw its first stable release in 2019, and its first non-mobile usage in 2020, when Apple released their first ARM-based MacBooks, fitted with M1 chips. Prior to this, MSL was only used to perform GPU operations on mobile devices. Although MSL is also able to control various GPUs from other manufacturers, we do not focus on systems with these GPUs in this work.

Although MSL is mostly focused on enabling graphics-oriented operations on Apple ARM GPUs, it also possesses functionality to instruct these chips for mathematical operations. This aspect of the M chip has, as it currently seems, not received much attention in the computational sciences, outside of the development of an MSL version of the popular TensorFlow software [Apple, 2022b]. This might be in part due to the fact that most documentation provided by Apple is focused on the Objective-C and Swift programming languages, typically favoured for developing general-purpose software on MacOS. The focus of these documentations is of course not the computational sciences, and as such, the exposure of the community to MSL has so far been limited.

In this work, we illustrate the usage of MSL in general C++ array operations, and analyse how and when MSL provides a benefit over running “plain” (multithreaded) CPU code. Special attention will be given to GPU operations in existing scientific C++ codes, specifically for numerical simulations of partial differential equations (PDEs) using finite differences. A case study focusing on accelerating elastic wave propagation in two dimensions using the M1 GPU illustrates the potential performance and ease of use of the M chip and unified processing units in general. All our simulations are run using single-precision decimal numbers (floats). Additionally, we provide a web portal that both links to our research codes as well as material helpful to the computational scientist to get started using MSL in C++ [Gebraad and Fichtner, 2022a].

6.2 MSL execution and memory model

MSL code itself compiles to instructions that are purely run on the GPU. These compiled functions are called shaders, or kernels. To orchestrate the execution of these kernels from the CPU, the instructions need to be fed from CPU code, i.e., any program we write in, e.g., C++. The communication of these instructions, as well as the scheduling of multiple operations and other “steering” tasks, are performed by the Metal Framework, an Objective-C library callable directly from C++ using ‘metal-cpp‘ [Apple, 2022c]. Performing operations on data with the Metal Framework on a GPU follows a set collection of steps [Apple, 2022c]:

1. Create a command buffer and encoder. These objects respectively receive instructions (buffer), and encode them into machine language for the appropriate GPU (encoder);
2. Place instructions and data addresses in encoder;
3. Encode instructions with the encoder;
4. Execute instructions.

During the execution of the command buffer, the CPU can resume operation and synchronize with GPU execution at a later stage. Although one does need to instruct the GPU which data to use for these operations, the data itself does not need to be communicated to the GPU, by virtue of the unified design. The Metal Framework allows one to create this shared data in existing applications, and then simply get a standard C++ pointer to the underlying data such that it can be used in existing CPU code. This greatly simplifies exposing existing codes’ arrays to new GPU operations, compared to using e.g. NVidia’s CUDA [NVidia et al., 2022]. Trying to access the data from both the CPU and GPU simultaneously creates a race condition, and should be avoided.

Although encoding the commands and data in the buffer is more intricate than typical C++ operations, we provide a simple interface to these operations for both one-dimensional and two-dimensional data on our web portal.

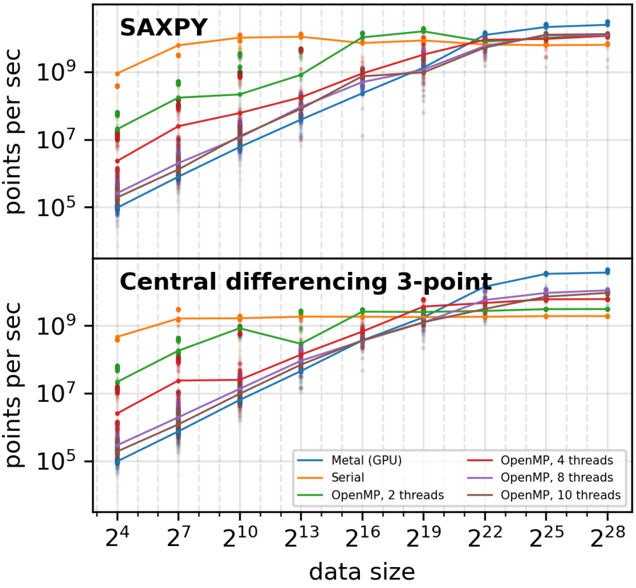


Figure 6.2: Runtime of one-dimensional operations on data of various sizes. Note that the median runtime is connected with the line, but all individual runs (i.e. multiple per data size per configuration) are indicated by circles, demonstrating variability of runtimes. Note that the runtime of OpenMP using 8 threads is always shorter than that of 10 threads, likely because the 2 performance cores of the M1 chip slow down OpenMP scheduling.

6.3 Basic array operations

To introduce how performant the M GPU is with respect to (M) CPU configurations, we benchmark various "single instruction, multiple data", or SIMD, operations on both processing units for one-dimensional data. We test the performance of the SAXPY operation (Single-Precision $a \cdot x + y$) and a 3-point central differencing scheme. By testing these operations on arrays of different sizes, we can thoroughly demonstrate the overhead required for using both OpenMP and MSL configurations. To generate these numbers, we ran the operations repeatedly on a 2021 MacBook Pro equipped with the M1 Max 10 core CPU (of which 8 performance cores) and a 32 core GPU.

Figure 6.2 shows how operations in GPU configurations, as well as operations in multithreaded CPU configuration, only provide benefit when the data is relatively large. Both array operations show fastest performance in serial mode up to an approximate data size of 2^{16} (= 65536). The exact speed-up of using various configurations at largest data size is

Operation	8T vs 1T	MSL vs 1T	MSL vs 8T
1D SAXPY	2.0x	3.9x	1.9x
1D CD	3.4x	19x	5.6x
2D EW	1.4x	4.7x	3.3x
2D LA	1.4x	12x	8.3x
2D LA9p	4.3x	47x	11x

Table 6.1: Speed-up of various array operations at maximum tested array size. For the one-dimensional and two-dimensional operations respectively, these sizes are 2^{28} and $2^{14} \times 2^{14}$ elements. Speed-ups are calculated for OpenMP with 8 threads (8T) versus serial (1T), the Metal Shading Language (MSL) versus serial, and MSL versus OpenMP with 8 threads. The operations summarized are those of the largest data sizes in Figures 6.2 and 6.3; one-dimensional $a \cdot x + y$ (1D SAXPY), one-dimensional central differencing using a 3-point stencil (1D CD), two-dimensional element wise function (2D EW), two-dimensional Laplacian using a 5-point stencil (2D LA) and two-dimensional Laplacian using a 9-point stencil (2D LA9p).

summarized in Table 6.1. These sepeed-ups demonstrate that even for simple operations, the usage of the GPU does accelerate one-dimensional array operations. The benefit of using the GPU for these relatively simple array operations in one dimension however only manifests itself at large data sizes, i.e. at 2^{22} elements and up.

6.4 Multidimensional operations

One staple of physical modelling are spatial derivatives, especially when solving PDEs (e.g. the wave equation) in two or three dimensions. MSL enables one to execute kernels specifically with two- or three-dimensional layouts, greatly simplifying the implementation of kernels that operate on two-dimensional or three-dimensional arrays.

We detail the performance in CPU and GPU configurations by again performing benchmarks of various operations on a 2021 M1 Max chip. Specifically, we investigate element-wise operations as well as 5 and 9-point Laplacian finite-difference stencils.

The runtimes of these benchmarks are given in Figure 6.3. The operations again show dominance of the scheduling overhead (for both CPU and GPU) for multithreaded configurations at small data sizes compared to the serial configurations. The crossover point (for all operations) where GPU becomes the most performant configuration seems to be around a data size of $2^{10} \times 2^{10}$ elements, close to the same number of effective elements for the crossover points in one dimension.

As the number of instructions per operation increases (i.e. top to bottom in Figure 6.3), so does the benefit of both the multithreaded CPU and GPU configurations with respect to serial configuration. Table 6.1 summarizes the speedup of the MSL, serial and optimal OpenMP configurations, showing how for large data sizes, MSL becomes more than an order of magnitude faster than OpenMP if the operation is complex.

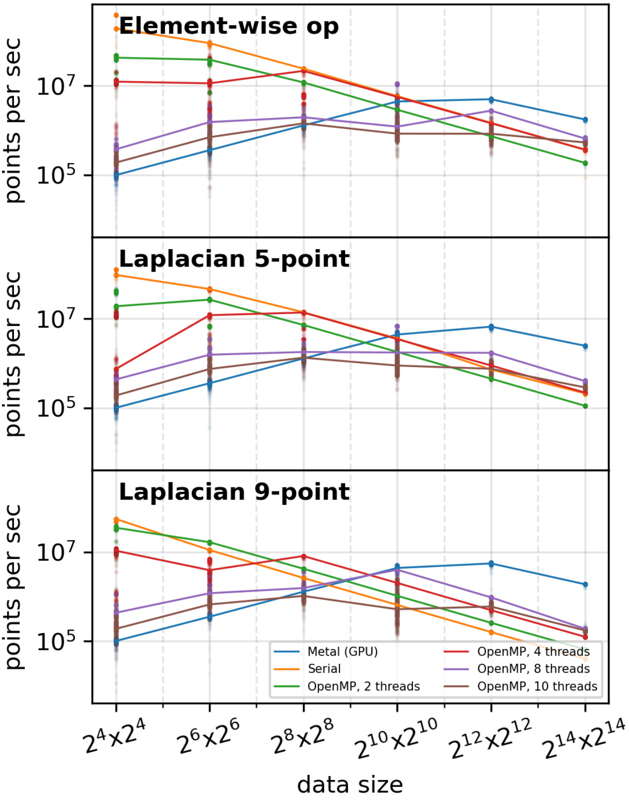


Figure 6.3: Runtime of two-dimensional operations on data of various sizes, always of equal width and height. As in Figure 6.2, separate runtimes are plotted using points, and the medians are connected by lines. The largest data sizes show almost an order of magnitude speed-up between CPU (OpenMP 8 threads) and MSL GPU.

6.5 Accelerating existing codes: elastic wave propagation on the M series GPU

Possibly the biggest strength of using MSL is its drop-in capabilities. Because the CPU and GPU have unified memory, data created in this unified pool can be readily accessed by both processing units. Thus, MSL allows a user to readily modify existing C++ to be GPU-capable.

Specifically, when creating an MSL buffer (i.e. an array that can be seen by the CPU and GPU), one can easily obtain a plain C++ pointer to the underlying data. This way, integrating MSL into an existing C++ application simply requires two additional lines per array: the declaration of the buffer and retrieving the raw C++ pointer.

To demonstrate this capability, we took an existing two-dimensional elastic wave propagation code [Gebraad and Fichtner, 2022b, Gebraad, 2022a] based on Virieux's seminal paper [Virieux, 1986]. This wave propagation code was developed to perform Full-Waveform Inversion (FWI), an approach to fit recorded vibrations to interior structure of materials. This method finds applications in seismology when imaging the Earth [Virieux and Operto, 2009, Lei et al., 2020, Thrastarson et al., 2022], in non-destructive testing when imaging man-made structures [Nguyen and Modrak, 2018, Kordjazi et al., 2020], and in medical tomography when imaging the human body [Guasch et al., 2020, Marty et al., 2021].

We implement the integration of the dynamic fields (material velocity in x and z direction, i.e. v_x, v_z, and vertical, horizontal and shear strains $\tau_{xx}, \tau_{zz}, \tau xz$) in MSL, but perform the rest of the operations required for forward and adjoint simulations[Lions, 1968, Tarantola, 1988, Plessix, 2006, Fichtner et al., 2006a] of wavefields on CPU. These operations include recording the entire forward dynamical wavefields (for later use in the computation of sensitivity kernels), injecting point sources, recording wavefields at receivers, and the cross-correlation between forward and adjoint dynamical fields.

Figure 6.4 demonstrates a surprising result with respect to our preceding results. At all domain sizes, the GPU configuration outperforms the 8-threaded OpenMP configuration. Where we were seeing cross-over points between the two configurations only at larger data sizes for simpler array operations, it now seems that the complexity of integrating wavefields means that MSL always outperforms OpenMP.

As domain size grows, so does the speed-up of MSL with respect to OpenMP. The simulation code fails at domain sizes above approximately 2000×2000, as system RAM runs out on our benchmarking machine. We were able to run a limited number of benchmarks at a domain size of 5000×5000, where the speed-up of MSL with respect to 8-threaded OpenMP was approximately a factor 10.

6.6 Discussion

Although the acceleration of computational physics by graphical processing units has long been acknowledged to yield performant codes, the complexity of implementation does create a barrier to its actual usage. The M chip series might herald the start of a paradigm-shift in computational sciences towards usage of unified chips and their programming lan-

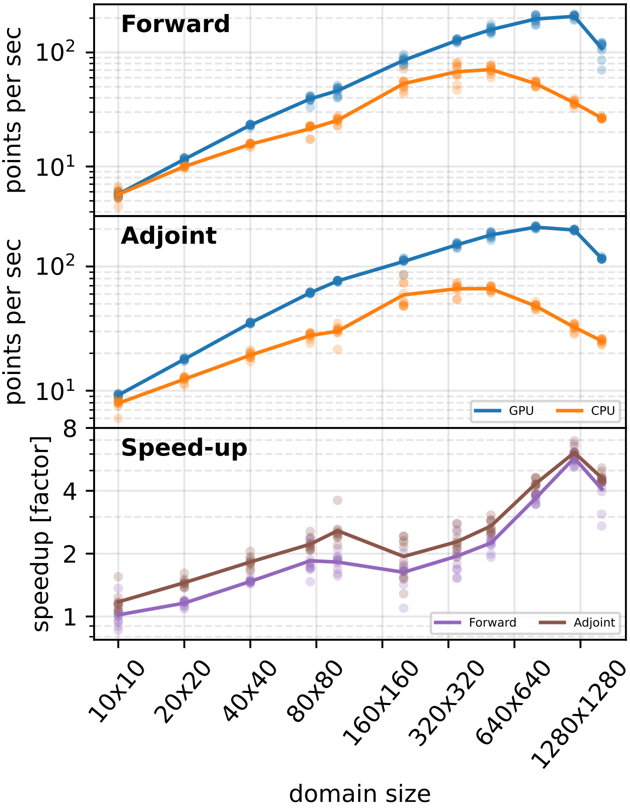


Figure 6.4: Runtimes of two-dimensional elastic wave propagation in media of various sizes for both OpenMP 8 threads configuration and MSL GPU configuration. The domain size indicates the amount of grid-points used for the parameter and dynamical fields.

guages. Although the M1 chip is not the first unified chip to be commercially available, nor is our implementation the first GPU-accelerated physical simulation, its ease of implementation does demonstrate the first implementation on a widely available chip using minimal effort. The upcoming release of the M2 chip promises increased unified memory bandwidth, thus further enlarging the potential for M series chips [Apple, 2022a].

Although for simple operations the speed-up of MSL with respect to OpenMP only becomes apparent at relatively large data sizes, in practical use for computational physics this threshold is much lower, due to the amount of instructions of the operations. This means that practically for our studied example (elastic wave propagation), it is typically worthwhile to use MSL over OpenMP. It is, however, not limited to our specific physics or numerical solver; these concepts of acceleration translate well to other PDEs solved with finite differences as well as to other numerical methods such as the finite element method [Rietmann et al., 2012, Kiss et al., 2012].

6.6.1 ThreadGroupSize dimensions

When using MSL shader functions, one needs to define in what shape the GPU traverses the operation. This access pattern is designated ThreadGroupSize in MSL, and is defined by a one-, two- or three-dimensional vector, depending on the input data. These vectors define in what pattern the multiple parallel cores on the GPU operate on the input data.

As an example, we show how we define the traversing of two-dimensional array operations for arrays with n_x rows and n_y columns. C++ arrays have linear memory layout and are indexed in two dimensions using the following linear index i:

$$i = i_x \cdot n_y + i_y \tag{6.1}$$

where i_x is the index of the row, i_y is the index of the column of the array. The ThreadGroupSize to launch two-dimensional kernels is defined by (t_x, t_y), where t_x and t_y indicate the dimensions of the tile of cores working on the data. The total amount of cores operating on the data in this tile is $t_x \cdot t_y$.

The usage of ThreadGroupSize is similar to CUDA scheduling of thread blocks. The access patterns of the data influence the performance of MSL and CUDA code alike. As an example, consider finite differencing schemes accessing neighboring elements. In these cases it is often beneficial to process multidimensional data in the order it is laid out in memory. Specifically, our two-dimensional data is laid out with strides of 1 element in the second dimension (y), and strides of n_y in the first dimension (x). Therefore, we launch our kernels with ThreadGroupSizes of (t_x,t_y), where $t_y >> t_x$. We find that this way our MSL kernels are the most performant. These optimal memory access patterns are well-known for general CPU and GPU programming, where the practice of optimizing access is typically known as memory coalescing [NVidia, 2013, Davidson and Jinturkar, 1994].

6.6.2 Asynchronous operation

As the CPU and GPU on the M series chips can operate asynchronously, there exists a potential further speed-up of array operations and computational tasks in general. We implemented this hybrid configuration for the two-dimensional elastic wave propagation.

In this configuration, approximately half the workload is shifted from the GPU back to the CPU on the M chip by letting the CPU integrate the vertical velocity field and the shear strain field, while the other 3 fields are integrated by the GPU. As the strain fields depend on the velocity fields and vice-versa, synchronization between the GPU and CPU is performed twice per time-step.

The results, however, are disappointing. For any configuration, the runtime of the hybrid configuration is approximately half that of the CPU configuration. This is because the M1 GPU significantly outperforms the M1 CPU for the elastic wave propagation, and spends most of the time waiting for synchronization. To actually attain a practical speed-up, the workload needs to be divided proportionally to the performance of both processing units, i.e. a larger part of the compute task needs to be allocated to the GPU. This was not implemented for our work.

6.7 Data and Resources

All code used in this work is accessible through the accompanying portal [Gebraad and Fichtner, 2022a]. All data used in this work is generated by these codes.

6.8 Acknowledgements

We would like to thank the various members of the Seismology and Wave Physics Group for providing valuable feedback on and accommodating a proving ground for our MSL implementation. We gratefully acknowledge the support of the Swiss National Science Foundation throught the grant 'Hamiltonian Monte Carlo Full-Waveform Inversion', grant number 200021 192236.

Chapter 7

Structure-from-Motion for seismological fieldwork

Chapter in preparation for submission to Seismica Field Reports, by L. Gebraad, I. Naets, P. Marty, and A. Fichtner.

Abstract

This report documents the application of photogrammetry in diverse seismological fieldwork settings to reconstruct the field site using off-the-shelf products and open-source software with high precision and relative ease. The photogrammetry method was employed in three distinct cases: (1) to survey a high-risk avalanche valley to verify locations of mass movements that potentially generated seismic signals, (2) to document the field state of a complex Distributed Acoustic Sensing (DAS)-fibre deployment, and (3) to survey a structure and surrounding topography for seismic simulation to predict behaviour under seismic movement. Owing to its easy integration with fieldwork, and the wide availability of required resources and skills, photogrammetry has proven to be accessible for a broad range of applications. Our findings underscore that photogrammetry is a potent tool for digitizing field environments, holding potential to inform and augment future seismological research.

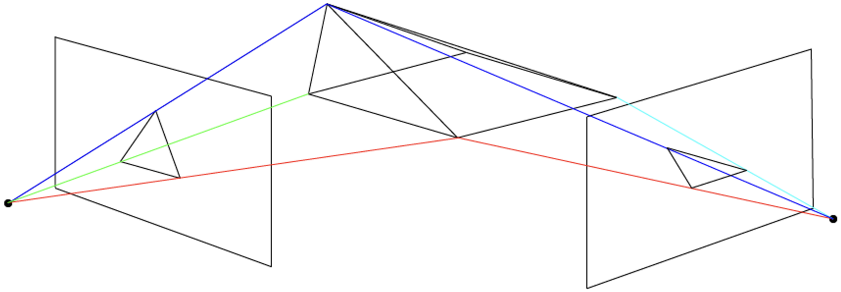


Figure 7.1: Schematic of common points triangulation from two images with known camera properties (3D location, orientation, and projection).

7.1 Introduction

The popularity of unmanned aerial vehicles (UAVs) for consumers has led to an explosive growth in the capabilities of these devices. These UAVs, typically quadcopter drones, are now regularly equipped with high-quality imaging instruments, accurate global positioning sensors and advanced and accessible flight controls.

In parallel, the development of 3D vision algorithms and software has led to the creation of a vast collection of user-friendly photogrammetry products with high performance. Contemporary photogrammetry software handles large volumes of imagery to produce highly accurate models. The Structure-from-Motion (SfM) [Ullman, 1979, Bolles et al., 1987, Özyeşil et al., 2017] technique is based on the identification of common points in multiple images, allowing triangulation of these points in 3D space and creation of surfaces or volumes representing the investigated subject. In the ideal case, if the camera properties, i.e. location, rotation, and projection, of every image are known, common points can be triangulated in 3D space by projecting them out from the images, as illustrated in Figure 7.1.

The availability of SfM algorithms poses a great benefit to the geosciences, which often concerns itself with large outdoor areas. Especially under transient conditions, the fact that only consumer electronics are required make UAV-based SfM much more accessible than other airborne remote sensing methods. Additionally, the ease of performing acquisition for large survey areas make it much faster than conventional laser scanning approaches. Because of this, it is argued that UAV-based reconstructions fill a niche of digital twinning for resource-light studies. Examples in geosciences include reconstructions for permafrost [Kaiser et al., 2022] and forestry [Iglhaut et al., 2019] research. In seismology specifically, studies that leverage SfM surveying typically are focused on transient sites [Johnson et al., 2014, Kayen et al., 2018, Pierce et al., 2020]. Carrivick et al. [2016] provides an overview of SfM applications in geosciences.

In this report, we detail our process for the acquisition, pre-processing and reconstruction of the SfM imagery at three field sites, as well as further analysis of the spatial reconstructions. We provide a detailed description of the field sites. The use of photogrammetry for each field site served a unique scientific purpose, and accordingly, this study aims to distill the three field sites to archetypes of application. Each field site possessed unique properties and required a slight adjustments in our approach to accurately capture and reconstruct the environments.

7.2 The Structure-from-Motion Technique for Seismology

The capacity to capture high-resolution, georeferenced imagery and transform it into detailed 3D models of the environment presents a powerful tool for seismologists. This section delves into the fundamental components of the SfM technique, detailing the acquisition, processing, and analysis of data.

7.2.1 Acquisition

Though professional survey planning and flight control software are available, all surveys for this work were conducted manually, using the complementary flight app from DJI. This app allows for hyperlapse captures, which provide templates for flying in straight lines and circles, during which imagery is acquired at fixed intervals. This feature makes the app ideal for our desired level of control. Two field sites were surveyed using a quadcopter drone. One field site, located near a military airport with flight restrictions, was surveyed using a mirrorless digital camera.

A successful survey requires capturing imagery of all desired reconstruction surfaces from multiple viewpoints. Operating the drone in hyperlapse mode along various lines with varying camera operations proved to provide optimal footage. Hyperlapse mode on consumer drones enables flights in set lines, acquiring imagery at consistent intervals. Although most available software can also process video, and video can be transformed into individual frames, we found no benefit over still images. By using still images, we reduced redundancy in imagery, preventing the generation of overly similar frames. An additional advantage of using still images is that DJI drones record geolocation to these files, unlike with video files. The significance of this in the following reconstruction process cannot be overstated. Lastly, we strongly advise capturing imagery on overcast days, or conducting surveys within as short a timeframe as possible. Preliminary tests revealed that inconsistent lighting in images can impair the reconstruction, to the point where reconstructions become disjointed. For the field site where imagery was captured using a handheld camera, we employed ground control points (GCP) to georeference the reconstruction. Further details of this are provided in the description of the field site.

7.2.2 Processing of the datasets

Successful processing of imagery from aerial or handheld surveys necessitates that image selection is performed prior to the reconstruction. Consequently, only images that are free of rolling shutter effects, which are difficult to correct for in SfM, or inconsistent shadows, which might lead to poor matching of common points, are selected.

To perform the reconstruction, numerous free and open-source softwares (FOSS) [Griwodz et al., 2021, Wu, 2013, Vacca and others, 2019], as well as commercial services, are available. We opted to use WebODM [Toffanin and WebODM Authors, 2023], for several reasons. Firstly, the developers of WebODM provide a Docker image, which requires no additional effort beyond installing Docker to set up the software on any system. Secondly, WebODM's interface is accessible via the internet from any location. This feature allows us to test partial reconstructions in the field, offering rapid prototyping of the acquisition, identifying blind spots, and testing the resolution of the reconstruction. Lastly, we found that the quality of reconstruction more than meets our needs. Potential users might also appreciate the broad community support on their forums, the command-line interface to OpenDroneMap (ODM) itself, or the paid cloud processing access through ODM's Lightning Network. All reconstructions in this report were made with WebODM.

WebODM has multiple settings that allow alteration of the quality of the reconstruction. As a baseline, it is suggested to simply use the default preset, with one important alteration. Given that the extent of the desired reconstruction is typically already known, it is recommended to use a boundary polygon to limit the reconstruction only to areas of interest. This potentially saves a significant amount of computational time by skipping areas that are present in the imagery but not of interest to the field investigation. Boundary polygons can be easily created on geojson.io.

7.2.3 Data products

The reconstructions completed with WebODM produce various data outputs. The raw output of most SfM algorithms, including WebODM, are coloured point clouds of the surfaces in the survey area. Furthermore, WebODM processes these point clouds into surface meshes, which are continuous triangular surfaces. WebODM also constructs two digital elevation models (DEMs) by default: a digital surface and a digital terrain model, respectively with and without buildings, for use in GIS applications. The last data product that is highlighted is the orthophoto. This is a georeferenced re-projection of the acquired imagery such that the scale is uniform and viewed top-down. All data products generated for this study are available in the dataset repository [Gebraad et al., 2023].

7.3 Field sites and data acquisition

Drone- and handheld acquisitions were performed at three field sites. The method was used in three distinct cases:

- to digitise the transient state of a valley at high risk of avalanches, thereby verifying seismic monitoring solutions;

- to digitise a highly complex Distributed Acoustic Sensing (DAS)-fibre deployment that differed from the proposed array geometry, thereby documenting the deployment;

- to digitise a complex structure and surrounding topography, thereby allowing the creation of a spectral element mesh for simulations.

These three field sites have distinct acquisition, processing, and reconstruction attributes, which are given in Table 7.1. Notably, the resolution of the reconstructions is reflected in the average ground sampling distance (GSD) values, highlighting the varying fidelity of the reconstruction, as necessitated by the differing objectives of the fieldwork.

	Field site 1	Field site 2	Field site 3
Location	Flüela pass	Field in Zurich	Contra Dam
Date	2022-03-25	2022-08-06	2023-03-17
Acquisition method	UAV	Handheld	UAV
Instrument used	DJI Mavic Air 2	Nikon Z6	DJI Mavic Air 2
Images used	128	365	1143
Reconstructed area	2.2 km^2	840 m^2	0.16 km^2
Point cloud size	2.2M points	16M points	100M points
Average GSD	12cm	0.25 cm	2.9cm
Reconstruction use	surface state	deployment documentation	mesh generation

Table 7.1: Acquisition, processing, and reconstruction attributes for the three field sites.

7.3.1 Field site 1: Reconstruction for spatial analysis in transient high-risk areas

The first field site was located on the upper parts of the Flüela Pass, a road in Eastern Switzerland that reaches above 2300 metres and is highly susceptible to avalanches in the winter. Our objective was to verify the locations of mass movements that might have generated seismic signals, as studied in Edme et al. [2023]. The rugged terrain and potential hazards made it challenging to access the site directly, particularly as the road is closed in winter and is not cleared of snow. Figure 7.2 indicates the exact location, just east of the highest point of the pass.

To safely fly the drone within line-of-sight over the hazardous field site, we launched on the 25th of March, 2022, from the summit of the pass, ascending from the safer north side. The image acquisition was performed over the course of 1 hour and 20 minutes, during which multiple hyperlapses were collected.

The processing of this dataset yielded a continuous, low-resolution reconstruction of the large area of the valley, which strongly correlates with existent satellite imagery and digital markers from the Swiss Map Vector 25 [Bundesamt für Landestopografie, 2018]. The reconstruction places topographic markers with near-perfect consistency, solely based on the GPS data embedded in the UAV-imagery.

As the primary objective of the study was the verification of mass movements, the data product of most interest was the orthophoto. This orthophoto is visualised in Figure 7.2.

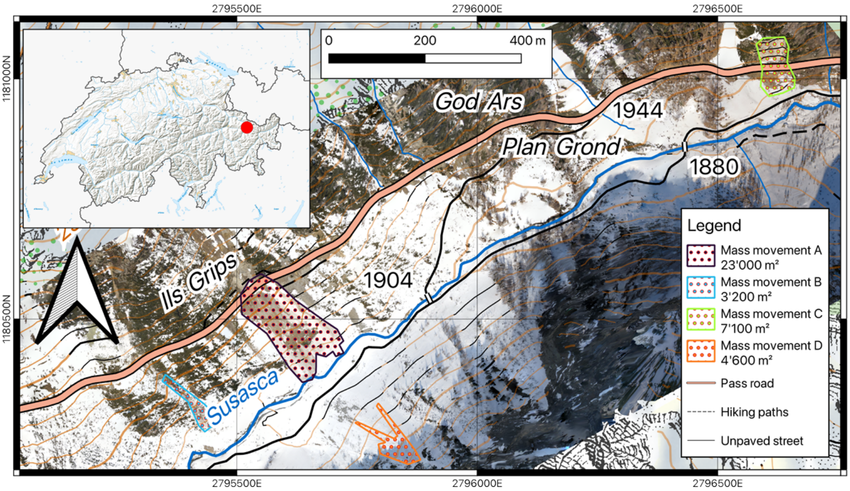


Figure 7.2: The field site on the Flüela Pass in Eastern Switzerland. Pictured here is the orthophoto generated during the construction, overlain with features from SwissTopo. The suspected mass movements are indicated by polygons.

We were able to confidently mark 4 mass movements in this orthophoto, showing a mix of snow and soil for A through C, and only snow for D in the transported material.

7.3.2 Field site 2: Reconstruction for DAS-fibre deployment geometry

The second field site, situated in the city of Zurich, focused on a complex Distributed Acoustic Sensing (DAS)-fibre deployment. The main objective of the photogrammetry campaign was to document the field state of the DAS-fibre system, a task presenting unique challenges due to the system's complexity and the need for accurate spatial configuration capture. Owing to the site's proximity to a military airport, obtaining the necessary flight authorisation for UAV operation was not feasible within the required timeframe. Consequently, a mirrorless camera was used, with imagery being collected by manually moving the camera across the field site grid from multiple orientations on 17 March, 2023, over a span of one hour.

In SfM, camera positions are not directly recorded in the photographs, often resulting in initial incorrect camera positions that necessitate optimisation during the reconstruction process. SfM algorithms are invariant under affine transformations such as scale and rotation, posing challenges for obtaining to-scale, georeferenced reconstructions. Whereas drones usually record GPS positions, providing a more reliable starting point for the algorithms and enhancing their speed and accuracy, the mirrorless camera employed in this study lacked GPS capabilities. Consequently, ground control points (GCPs) were used for georeferencing the reconstruction.

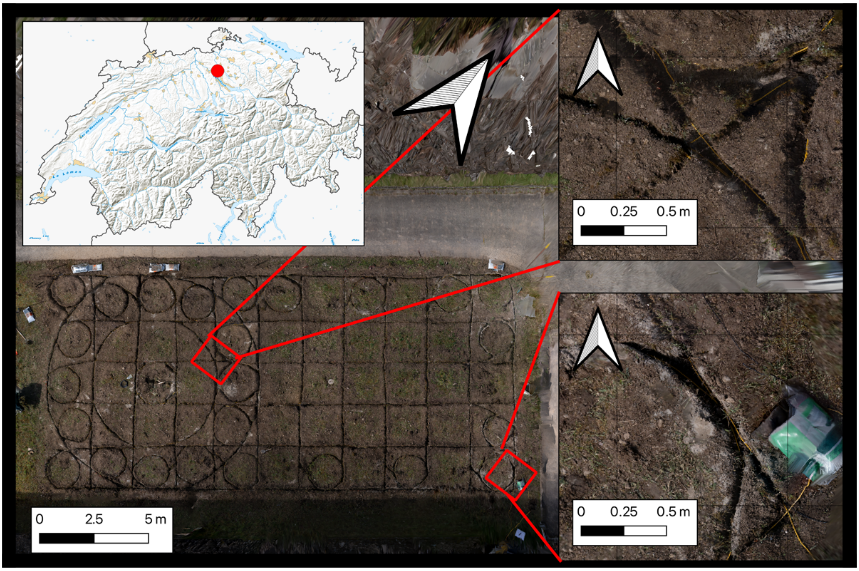


Figure 7.3: The field site for the DAS fibre optic cable deployment geometry study. The pictured reconstruction is the high-resolution orthophoto, with two callouts illustrating the high-resolution georeferenced reconstruction obtained in this reconstruction.

GCPs are identifiable landmarks depicted in the imagery with known geolocations. They aid the reconstruction algorithm in counteracting the scale, translational, and rotational invariance of the imagery. In this instance, the GCPs were defined retrospectively by selecting identifiable features such as postboxes, traffic markings, and building features based on satellite imagery, which were then tagged in the acquired survey imagery.

Close-range photogrammetry and the high number of images facilitated the high-resolution reconstruction of this field site, as demonstrated in the call-outs in Figure 7.3. The resultant digital model provided valuable documentation of the field state, thereby aiding in the analysis and assessment of the DAS-fibre system's performance.

7.3.3 Field site 3: Reconstruction for simulation mesh generation

The third field site was devoted to the generation of a mesh for elastic wave simulations using the Spectral Element Method (SEM). The aim was to target a specific structure and its surrounding topography for which no mesh existed, and at a low capital cost and with reduced man-hours, to generate a mesh suitable for seismological research. The survey was performed on the Contra Dam in the Verzasca Valley of Southern Switzerland.

Given that the dam operates as an active power station and is within the control zone of Locarno airport, extensive planning and communication were undertaken with the dam

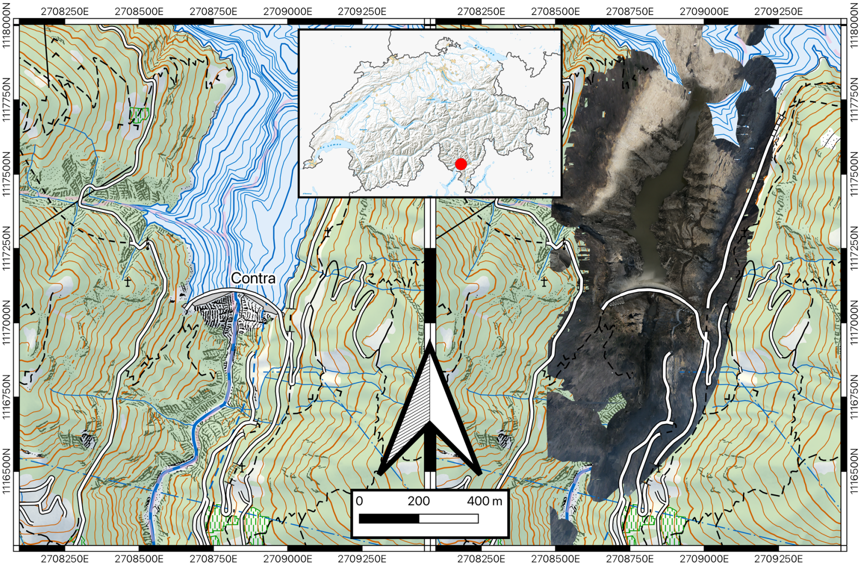


Figure 7.4: The field site for the geometry reconstruction of the Contra Dam in the Verzasca Valley, Southern Switzerland. The left panel illustrates the field setting, while the right panel shows the reconstructed orthophoto. The lake was partially drained at the time of the acquisition.

operators and Locarno Air Traffic Control. Imagery was acquired on March 2, 2022, over approximately three hours, utilising multiple circular and linear hyperlapse flight templates. Coincidentally, the reservoir behind the dam was drained for maintenance at the time of the survey.

The result of the WebODM reconstruction, in conjunction with a map of the field site, is presented in Figure 7.4. To simulate the wave physics of the investigated structure and its surrounding topography, a volumetric description of the field site is required. The process to achieve this involves several steps, detailed below.

Firstly, it is necessary to ensure the surface mesh is continuous. As demonstrated in Figure 7.5, the reconstruction exhibits exceptional quality in the well-surveyed areas, such as near the dam. However, the imagery might be inadequate near the edges of the domain of interest, resulting in defects, typically manifested as holes in the surface. These defects were corrected using 3D sculpting tools in Blender, as shown in Figure 7.6, Panel B. Secondly, the mesh needs to be extruded both horizontally and downwards. The profiles along one bound of the surface mesh were used for horizontal extrusion to square dimensions. These new sides were then extruded downwards to such a degree that the final surface mesh roughly enclosed a cube. This process is depicted in Figure 7.6, Panel C. Finally, the volume enclosed by the newly generated surface mesh required a volumetric meshing

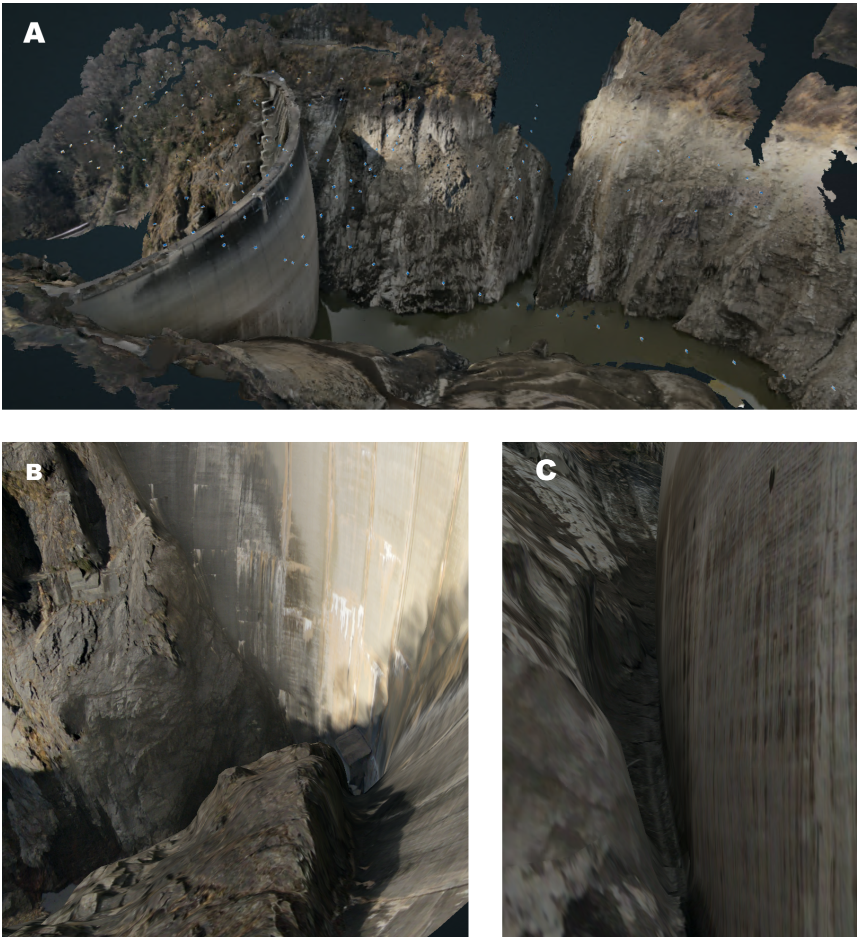


Figure 7.5: The coloured surface mesh from field site 3. Panel A presents an overview of the field site, with camera locations indicated by the blue dots. Unconformities can be observed at various locations near the edges of the mesh, such as downstream (i.e. left side of the image), near the bottom of the valley. Panels B and C show close-ups of the structure itself, respectively downstream and upstream of the dam. These close-ups illustrate the high accuracy achieved for well-surveyed areas.

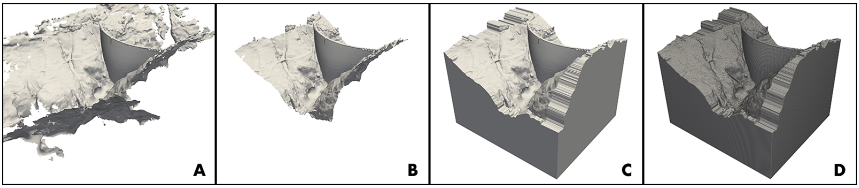


Figure 7.6: The process of creating a volumetric spectral element mesh from the WebODM output, as illustrated for field site 3. Panel A displays the original surface mesh from WebODM. This surface extends beyond the main area of interest, as well as having non-conforming geometry. Panel B shows the surface mesh after cut-out and correction of reconstruction errors. Panel C highlights how the horizontal and vertical extrusion create a surface that encloses a roughly cubic volume. Panel D visualises the volumetric mesh that is generated from the surface mesh using Cubit®.

process to make it compatible with the SEM simulation software Salvus[Afanasiev et al., 2019]. This was achieved using Cubit®, with the result displayed in Panel D of Figure 7.6.

In order to perform simulations using the generated mesh, material properties need to be assigned. For illustrative purposes, typical elastic properties of granite were assigned throughout the entire domain, irrespective of whether the medium was likely part of the structure or surrounding bedrock. As the lake was mostly drained at the time of the survey, the usually submerged topography was digitised. As part of the modelling process, a body of water was virtually reintroduced, consistent with the markings left by the original lake prior to drainage.

Showcase simulations were performed for this report to demonstrate the validity of the approach. A single point source was added at the bottom-centre of the domain, and the effects of this source were propagated through the domain using a coupled elastic-acoustic wave equation for the solid and fluid medium, respectively. An example wavefield snapshot is presented in Figure 7.7.

7.4 Data availability

The input and output data for the SfM reconstructions, as well as the generated SEM meshes for field site 3, are available on Zenodo[Gebraad et al., 2023]. WebODM, Blender, QGIS and ParaView are software programs that are free to use. Cubit and Salvus are software programs that require a license.

7.5 Conclusion

This work presented three diverse field scenarios, representing unique archetypes, where Structure from Motion (SfM) photogrammetry was effectively utilised. These scenarios showcased the robustness and flexibility of SfM, emphasising its accessibility and adaptability in the context of geoscientific research.

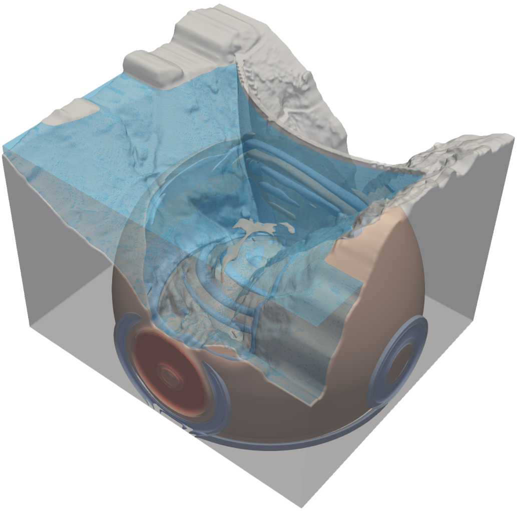


Figure 7.7: Snapshot of wavefields generated from a point source below the digitised Verzasca Dam, visualised using ParaView.

The first scenario highlighted the utility of SfM in enabling inspection of field sites under hazardous conditions. The technology served as a reliable tool to generate detailed 3D reconstructions, providing critical insights without compromising safety. This use-case underscores the potential of SfM as a risk-mitigation tool in diverse geoscience applications.

In the second scenario, the versatility of SfM was demonstrated through the documentation and georeferencing of a complex Distributed Acoustic Sensing (DAS)-fibre deployment. In an urban setting, where traditional measurement and documentation techniques may struggle, SfM provided accurate, detailed data, offering a valuable tool in the precise documentation of instrument deployments.

The third scenario showcased the ability of SfM to generate simulation meshes of field sites. This highlighted the role of SfM not only as a documentation tool but also as a crucial asset in generating data for sophisticated scientific simulations, bridging the gap between field data and computational models.

In conclusion, Structure from Motion photogrammetry is not only accessible but presents exciting possibilities in various geoscience research contexts. This paper illustrates its utility across a range of applications, from inspection of hazardous sites, through precise documentation of deployments, to the generation of sophisticated simulation meshes. The results highlight the potential of SfM as a versatile, low-cost, and accessible tool that can significantly contribute to the advancement of geoscientific research. As its use becomes more widespread, further development and innovation in its application are anticipated, opening the door to new possibilities in geoscientific investigation.

Part III

Scientific collaborations

Chapter 8

Inverse problems in seismology

Within my direct network, we oftentimes come across new inverse problems for which we can readily use HMCLab. Examples of this is normal mode tomography [van Tent et al., 2020, 2021a,b, 2022], hypocenter location[Klaasen et al., 2022, 2023] and surface wave dispersion [Lanteri et al., 2022, 2023]. However, some inverse problems require extensions to the existing functionality of HMCLab, or are not suited for MCMC sampling. This chapter documents two of those cases.

8.1 A bisection algorithm for ray-tracing in layered media

In support of a study with aiming to characterise the vertical structure of ice streams, a ray-tracing algorithm was developed to invert for wave speeds in layered media in a performant way. With field data from the EastGRIP site in Greenland, where a DAS cable was deployed in a vertical borehole to learn more about the mechanical properties of the ice, the following work was recently submitted to arXiv [Fichtner et al., 2023] and the Geophysics Journal International.

> A. Fichtner, C. Hofstede, L. Gebraad, A. Zunino, D. Zigone, and O. Eisen. Borehole fibre-optic seismology inside the Northeast Greenland Ice Stream. *arXiv*, July 2023. doi: 10.48550/arXiv.2307.05976.
>
> **Original abstract** Ice streams are major contributors to ice sheet mass loss and sea level rise. Effects of their dynamic behaviour are imprinted into seismic properties, such as wave speeds and anisotropy. Here we present results from the first Distributed Acoustic Sensing (DAS) experiment in a deep ice-core borehole in the onset region of the Northeast Greenland Ice Stream. A series of active surface sources produced clear recordings of the P and S wavefield, including internal reflections, along a 1500 m long fibre-optic cable that was lowered into the borehole. The combination of nonlinear traveltime tomography with a firn model constrained by multi-mode surface wave data, allows us to invert for P and S wave speeds with depth-dependent uncertainties on the order of only 10 m/s, and vertical resolution of 20–70 m. The wave speed model in conjunction with the regularly spaced DAS data enable a straightforward separation of internal upward reflections followed by a reverse-time migration that provides a detailed reflectivity image of the ice. While the differences between P and S wave

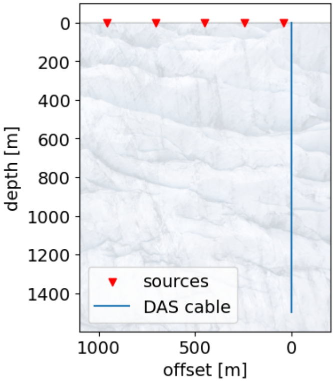


Figure 8.1: Acquisition geometry for the Northeast Greenland Ice Stream DAS experiment.

> speeds hint at anisotropy related to crystal orientation fabric, the reflectivity image seems to carry a pronounced climatic imprint caused by rapid variations in grain size. Currently, resolution is not limited by the DAS channel spacing. Instead, the maximum frequency of body waves below ∼ 200 Hz, low signal-to-noise ratio caused by poor coupling, and systematic errors produced by the ray approximation, appear to be the leading-order issues. Among these, only the latter has a simple existing solution in the form of full-waveform inversion. Improving signal bandwidth and quality, however, will likely require a significantly larger effort in terms of both sensing equipment and logistics.

For this work I developed the algorithm to obtain vertical seismic profiles for P and S wave speeds from the respective refracting arrivals for both phases. The first arrivals were obtained by analysing the active source generated data in the acquisition geometry given in Figure 8.1. The developed machinery is applicable to all vertical seismic profiles with source a varying offset at the surface from the borehole.

To invert for the wave speeds it was assumed that the medium was only varying up to a small degree, such that upward propagation of the phases was negligible. It was further assumed that the medium is only vertically varying, a good approximation in large glaciers. To invert for the medium velocity, a ray has to be traced through this layered medium, after which the travel time along this ray can be compared to the observed first arrival, and subsequently minimised. The ray-tracing for varying take-off angles from a source is demonstrated in Figure 8.2.

Although a code for repeatedly calculating Snell's law in layered media is relatively straightforward to implement for a given take-off angle at the source, the complexity of this inverse problem arises from the fact that it is unknown which take-off angle should be used to connect a specific source and receiver. To find and trace the connecting ray is thus a non-trivial problem. To solve this, I developed a parallel search algorithm based on the bisection algorithm [see e.g. Arfken et al., 2011] that progressively becomes more

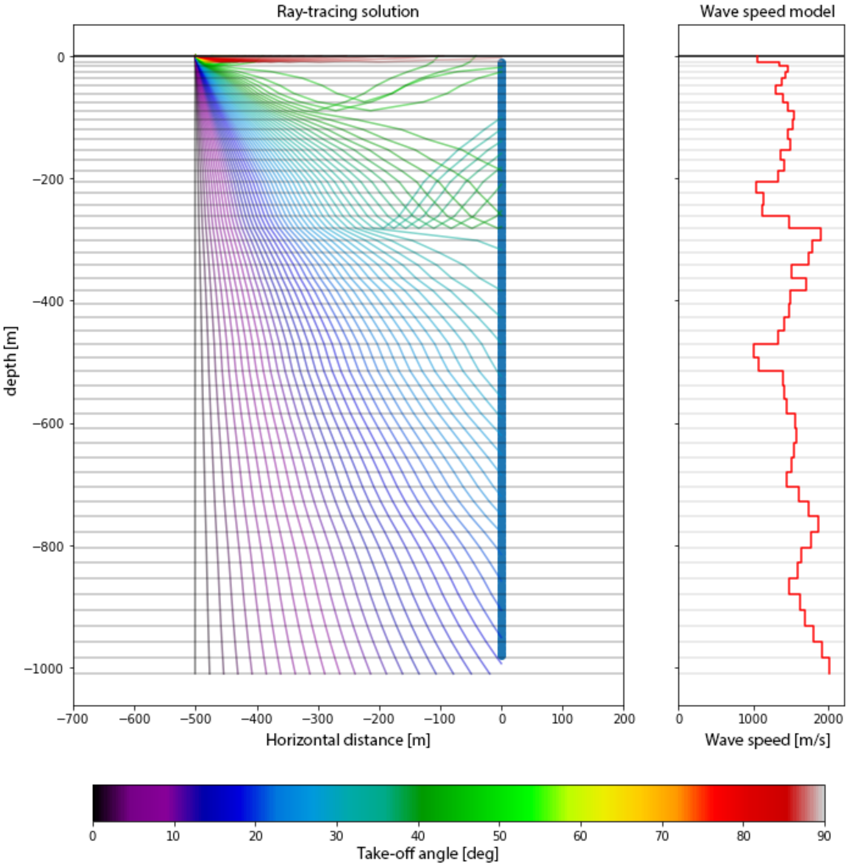


Figure 8.2: Ray-tracing through a random wave speed model for take-off angles linearly spaced between 0 and 90 degrees from vertical. Note that only rays taking off over 20 degrees from vertical arrive at the DAS cable at all, while most of the rays between 40 and 80 degrees do not propagate to the DAS cable.

performant. It initialises itself using linearly spaced take-off angles between 0 and 90 degrees from vertical. The rays originating from these angles are traced in parallel, after the ray with the closest approach to each channel is refined by a bisection of the take-off angles. Furthermore, once enough close approaches are calculated, new bisection are placed at take-off angles predicted by interpolation of close approaches. The addition of a stochastic component to each bisection enables the algorithm to avoid local minima. Further acceleration of the algorithm is achieved by memory of the take-off angles for the

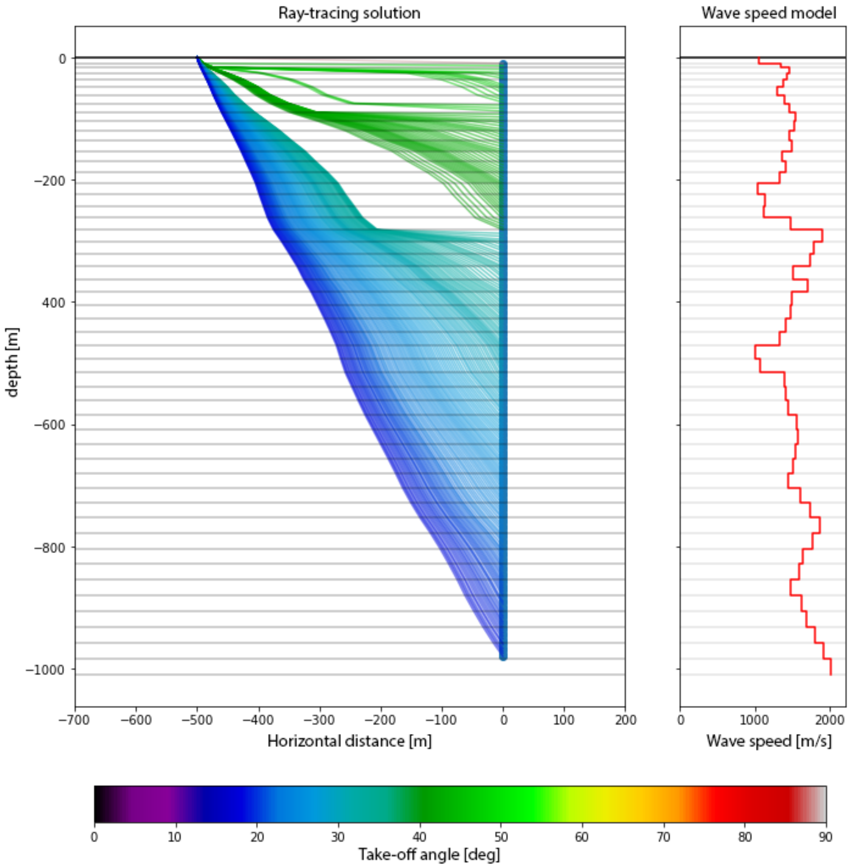


Figure 8.3: Ray-tracing using the solved take-off angles for the same wave speed. Every ray connects the source to a DAS channel.

last computed wave speed model, such that updates in, e.g., gradient based optimisation require much fewer iterations of the search algorithm to execute accurate ray-tracing.

Tracing rays through each layer in the wave speed models allows the computation of the gradient of the observables, i.e., the first arrivals, on the fly. As such, this inverse problem becomes suited to non-linear optimisation strategies. The total dataset for P and S wave speed is inverted for using the L-BFGS algorithm. By bootstrapping the data with the observational noise and repeating these optimizations, a proxy for the posterior uncertainty is created from the ensemble of final models, as shown in Figure 8.4.

The generated model is an example how with relatively simple physics a very interest-

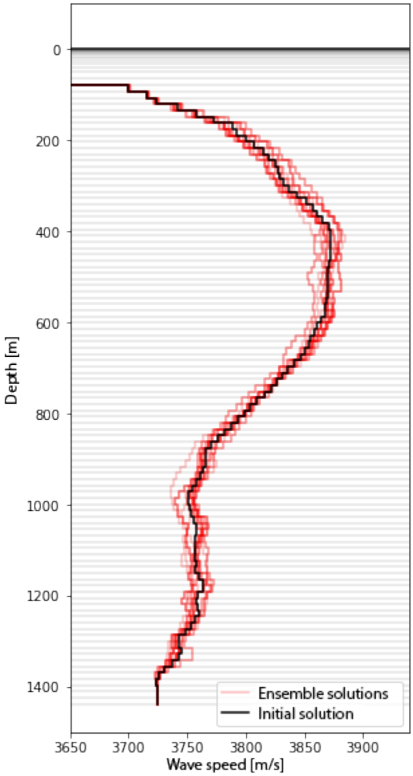
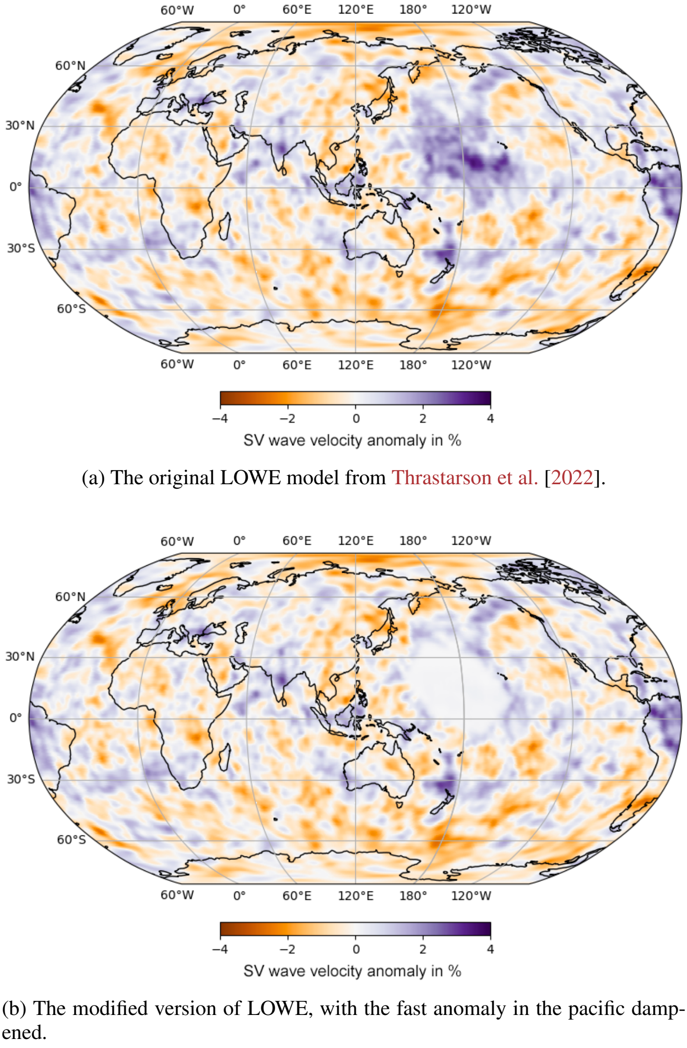


Figure 8.4: Final P wave speed from the first arrivals of all 5 sources.

ing inverse problem can be created. Further extensions using existent data from varying azimuth sources is the incorporation of azimuthal anisotropy. To fully understand the trade-offs and uncertainties of this inverse problem, further probing of the model with Bayesian appraisal algorithms such as HMC is recommended. This code will be made available through the HMCLab Python interface [Zunino et al., 2023].

8.2 Hypothesis testing in large-scale inversions

The following work is a multidisciplinary effort to understand an fast wave speed anomaly that consistently appears across contemporary global tomography models. This extremely large anomaly present in the Pacific Ocean from 800-1100km depth is present in both GLAD-25 [Lei et al., 2020] and LOWE [Thrastarson et al., 2022], which are both FWI-based reconstructions. Currently, the following work is in preparation for submission.



(a) The original LOWE model from Thrastarson et al. [2022].

(b) The modified version of LOWE, with the fast anomaly in the pacific dampened.

Figure 8.5: Global depth slices of SV-wave speed anomalies in LOWE at a depth of 1100 km, with and without the Pacific Ocean Anomaly.

> T. Schouten, L. Gebraad, S. Noe, A. Gülcher, B. Vaes, S. Thrastarson, D.-P. van Herwaarden, and A. Fichtner. Global full-waveform inversion reveals mid-mantle structure beneath the remote Earth: evidence for Cretaceous to Cenozoic intraoceanic Pacific-Panthalassa subduction? In preparation. T. S., L.G. and S.N. contributed equally to this manuscript.
>
> **Original abstract** The plate tectonic cycle is the continuous process of the creation, motion, and destruction of tectonic plates that constitute the Earth's lithosphere. As plates subduct into the mantle, they leave only few hints of their former existence in the rock record. Earth's lithosphere is composed of tectonic plates whose motions can be determined through records of oceanic spreading and quasi-stationary hotspots. However, reconstructing the evolution of convergent plate boundaries is more challenging due to the destructive nature of subduction, which removes much of the evidence from the surface.
>
> Seismic tomographic imaging of the mantle provides an important archive of these plates by detecting fast wavespeed anomalies interpreted as subducted lithosphere. Classical ray-based seismic tomography relies on P- and S-wave arrival times and requires stations near areas where resolving power is desired. This renders it impossible to infer mid-mantle structure beneath remote regions such as the Pacific Ocean.
>
> Recent developments in global-scale full-waveform inversion (FWI) enable tomographic imaging with unprecedented resolution, illuminating the mantle even in regions with poor coverage. Here we present evidence of a large, flat-lying positive wavespeed anomaly at a depth of 800-1200 km below the Western Pacific, which we name the Nemo anomaly. This anomaly lies directly below the area lost to subduction between the Australian, Eurasian and Pacific plates lost to subduction since $\sim$ 120 Ma. We identify several exotic terranes in western Pacific orogens that may hold key geologic information on the origin of the Nemo slab. These findings demonstrate FWI's potential in exploring the mantle in the remote Earth to find the missing pieces of the plate tectonic puzzle.

The anomaly observed in the models, shown in Figure 8.5a, is a massive anomaly. So massive, that we at all costs wanted to avoid the over-interpretation of the anomaly in the case it would be hallucinated by the inversion method or other unforeseen factors, and thus to prevent outrageous, unfounded claims. We furthermore suspect the reason this anomaly has gone unnoticed until now is the recent advent of global-scale FWI, which gives much better spatial sensitivities, especially in regions with as poor station coverage as the Pacific.

My contribution to this manuscript was the extensive testing on whether this anomaly is required by the data. Our hypothesis was that such a coherent feature within one model [Thrastarson et al., 2022, i.e., LOWE], that is furthermore present in multiple models, is a feature imposed by the data and not an inversion artefact.

To test this, we resorted to creating a twin of the final model of LOWE, with the SV-wave speed model dampened to the starting model over the extent of the anomaly, shown in Figure 8.5b. With the two variants of the model, we would be able to see the influence of the anomaly on the different parts of the data fit. Thus we could test the hypothesis that this anomaly is actually present.

We selected all events likely to directly probe the anomaly and computed the synthetic data for the newly created dampened version of LOWE. Compared to the original LOWE model, the anomaly-dampened version produced a worse total data fit. Subsequently, the

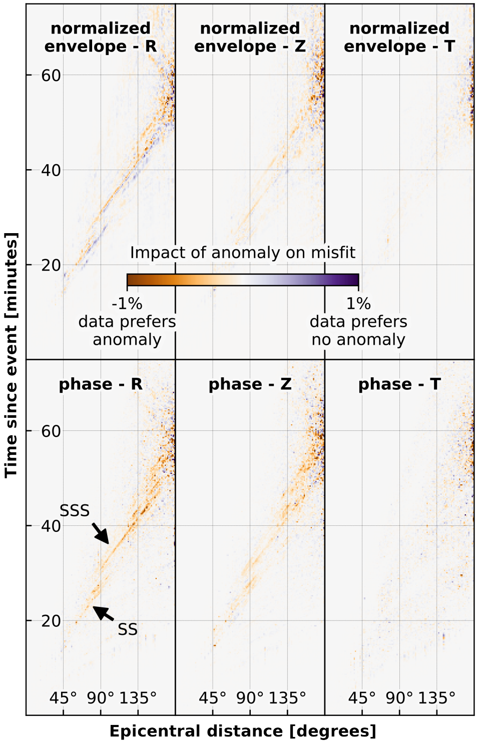


Figure 8.6: The difference between the original residuals in envelope and phase for the LOWE model, compared to the residuals calculated with the anomaly-dampened version of LOWE, stacked for epicentral distances across all simulated events. Visualised are the R, Z and T components observed on at the seismic station, respectively the radial, vertical and transverse component, oriented with respect to the source location. Note that the reduced sensitivity of the transverse component is due to the fact the anomaly removed was SV wave speed anomaly, thus any impact on the fit of the transverse component must be from converted phases in the coda. Almost no change can be observed in the residuals for any phase other than multiples of the S phase.

difference of the synthetics with the observations, i.e. the residuals, in both envelope and phase, were compared to the original residuals created by the LOWE model.

The result shown in Figure 8.6 confirm our suspicion that the phases most sensitive to an anomaly in such a coverage-poor region are mostly complex phases not integrated in pre-FWI global-scale tomographies such as Amaru [2007], Ritsema et al. [2011], Hosseini et al. [2019]. What should furthermore be noted is the LOWE model is optimised using a time and frequency dependent phase misfit [Fichtner et al., 2008].

Furthermore, the overall increase in misfit from the LOWE model to anomaly-dampened LOWE model, along with the propensity of the difference in phase residual (Figure 8.6) to prefer the anomaly, lead us to believe the Pacific Ocean anomaly is in fact present. This claim was recently further reinforced by a study seemingly identifying the reflecting top of the anomaly [Zhang et al., 2023b].

I believe that simplified but targeted probing of misfits liek the approach demonstrated here can bridge the gap between elegant Bayesian inference and practical non-linear optimization approaches. The treatment of the global-scale FWI problem such as the one in Thrastarson et al. [2022] as a statistical one remains heavily out of scale for current computational resources, but targeted hypothesis testing offers a reasonable simplification. This approach could possibly be further enhanced by limited interrogation-theory based sampling [Arnold and Curtis, 2018] of a low dimensional parametrisation of a hypothesis. An example of such a parametrisation might be a scalar-valued feature strength, where an anomaly is dampened by a scalar factor, which is subsequently sampled over to produce posterior beliefs on the likelihood of the feature to be present at a given strenght. This can be naturally extended to all PDE-based inverse problems. Further geological and geodynamical interpretation of this anomaly are left out of this text, as this is still a matter of debate, even between the authors of this study.

Chapter 9

Gradient-based sampling beyond geophysics

Specialising in inverse problems and sampling methodologies not only provides expertise in a specific research area but also equips one with tools applicable across a variety of scientific disciplines. Through fortunate connections, I was able to collaborate with scientists from fields outside of Earth Sciences also dealing with optimisation problems that could potentially benefit from a performant sampler capable of handling high dimensionality. This chapter highlights two collaborations where my experience with HMC has proven advantageous.

9.1 Metamaterial design

The design of metamaterials presents a domain where global optimisation of parameters is of critical importance. Preliminary work conducted in collaboration with Cyrill Boesch showcased the utility of HMC sampling in the context of coupled oscillator design problems. The coupled oscillator are intended to exhibit a specific spectral response, for which the design is attempted to be optimised. Here, HMC successfully bypassed local minima that had limited non-linear optimisation methods. Currently, we are applying HMC to metamaterial design challenges, as outlined in the recent work Dubček et al. [2023].

T. Dubček, D. Moreno-Garcia, T. Haag, P. Omidvar, H. R. Thomsen, T. S. Becker, L. Gebraad, C. Bärlocher, F. Andersson, S. D. Huber, D.-J. van Manen, L. G. Villanueva, J. O. A. Robertsson, and M. Serra-Garcia. Binary classification of spoken words with passive phononic metamaterials. *arXiv*, July 2023. doi: 10.48550/arXiv.2111.08503

Original abstract Mitigating the energy requirements of artificial intelligence requires novel physical substrates for computation. Phononic metamaterials have a vanishingly low power dissipation and hence are a prime candidate for green, always-on computers. However, their use in machine learning applications has not been explored due to the complexity of their design process: Current phononic metamaterials are restricted

to simple geometries (e.g. periodic, tapered), and hence do not possess sufficient expressivity to encode machine learning tasks. We design and fabricate a non-periodic phononic metamaterial, directly from data samples, that can distinguish between pairs of spoken words in the presence of a simple readout nonlinearity; hence demonstrating that phononic metamaterials are a viable avenue towards zero-power smart devices.

9.2 Bayesian neural network training

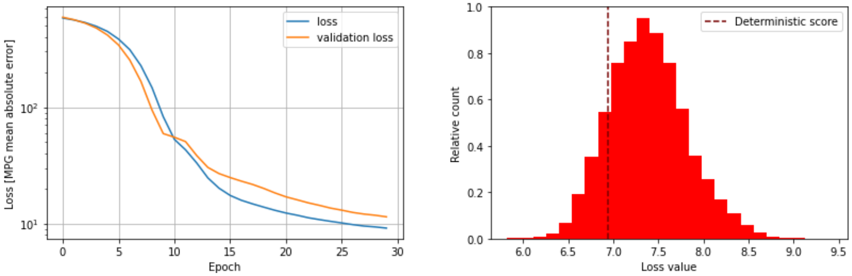
Despite its feasibility for Monte Carlo sampling of neural network parameters being demonstrated as early as the mid-nineties [Neal, 1996], Hamiltonian Monte Carlo (HMC) has not garnered as much popularity as other algorithms such as Adam [Kingma and Ba, 2014] and SGD [Robbins and Monro, 1951]. A Bayesian approach to neural network training, however, offers two significant advantages: it enables quantification of inherent uncertainty—often overlooked by deterministic optimisation—resulting in probabilistic outputs, and it can potentially mitigate the common issue of local minima encountered during neural network training. These advantages of HMC were put to use in the recent work of Louvet et al. [2023].

T. Louvet, V. Maillou, F. Bohte, L. Gebraad, and M. Serra-Garcia. Training elastic neural networks with the Hamiltonian Monte Carlo sampling algorithm. *Bull. Am. Phys. Soc.*, Mar. 2023

Original abstract Because of their low damping and highly non-linear characteristics, artificial neural networks (ANNs) made of nonlinear elastic resonators are promising candidates for low-power computing, as illustrated by recent demonstrations of passive speech recognition. However, designing information-processing elastic structures is a hard optimization problem: While the training of software-based ANNs can be facilitated by increasing the network size (converting local minima into saddle points), and by choosing activation functions with beneficial properties, there are usually hard limits on the size and activation functions in physically-implemented neural networks. Here we train resource-constrained elastic ANNs by applying the Hamiltonian Monte Carlo method, a variant of the Metropolis-Hastings algorithm used in statistical physics to sample probability distributions presenting a large number of local minima. While our work focuses on computers consisting of physical elastic resonators, our conclusions can be applied to general low power/resource constrained machine learning.

To showcase the application of HMCLab on a neural network, we performed a case study with a neural network functioning as a regressor.

The trainable weights $\mathbf{m}$ of a neural network are typically optimised utilising stochastic methods. The loss function $\chi(\mathbf{m})$ along with its derivative, are provided to an optimisation routine such as gradient descent. These routine then attempt to find a vector $\mathbf{m}$ that minimises χ. Considering the typical overparametrisation of neural networks, numerous local minima exist. Algorithms such as gradient descent encounter issues with these minima, as they may become trapped without any hint of local or global convergence. This scenario necessitates the development of techniques to supplement optimisation routines, ensuring they converge as closely as possible to the global minimum - the optimal set of parameters from all possible combinations. During the machine learning surge of the 2010s, Adam, with its various methods to bypass local minima and accelerate convergence, was a popular variant.



(a) Loss and validation loss values throughout 30 epochs of Adam optimisation.

(b) All losses sampled during subsequent HMC sampling.

Figure 9.1: Loss functions of the trained Bayesian neural network during Adam burn-in and the sampling phase. The burn-in phase epoch number was selected to achieve maximum convergence. Note that despite the Adam algorithm typically yielding models with low loss value, the HMC algorithm discovers alternatives with both higher and lower loss values.

Interpreting these loss functions within a Bayesian framework enables us to consider the collection of probable trainable weights that adequately explain the training dataset. In this scenario, the loss function must be explored to characterise all probable models. This exploration equates to navigating the probability distribution denoted by

$$p(\mathbf{m}) = \exp\left(-\chi\left[\mathbf{m}\right]\right). \tag{9.1}$$

This distribution spans the N-dimensional parameter space occupied by the neural networks's trainable weights.

This approach is demonstrated on the Miles-per-Gallon dataset [Quinlan, 1993]. The Miles-per-Gallon (MPG) dataset is a commonly used dataset for regression analysis. The data comes from the 1970s and 1980s and consists of several attributes of automobiles, such as cylinders, displacement, horsepower, weight, acceleration, model year, and origin. The target variable is the car's fuel efficiency measured in miles per gallon (MPG).

We create a dense neural net using Keras to acts a a multivariate regressor between car properties and miles-per-gallon. We utilise a rudimentary network architecture comprising a normalisation layer, followed by a sequence of 3 ReLU layers with 64, 32, and 16 nodes before converging into a single output, for a total of 3265 trainable parameters. Following initial training with Adam, the obtained network weights serve as a starting model for the HMC algorithm. Autotuned solely by the stepsize of the time integration scheme, the algorithm then executed 50,000 proposals. Figure 9.1 illustrates this training process.

The resulting set of network parameters can be utilised for neural network predictions. For every input, the neural network can generate a distribution of outputs. Figure 9.2 presents these predictions.

As the predictions are samples, they allow for the calculation of their statistics, such as means and variances. A more comprehensive comparison of the deterministic and

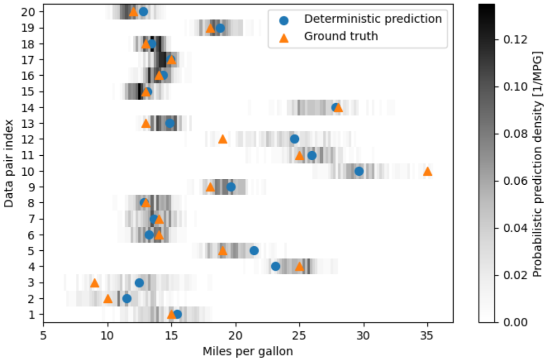
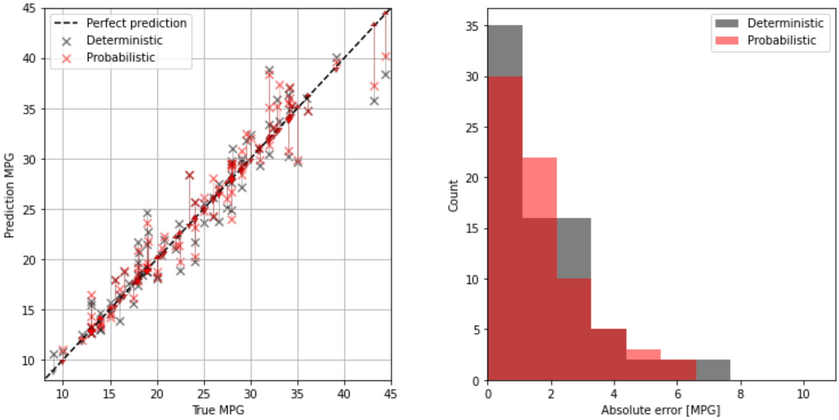


Figure 9.2: Direct comparison of the deterministic and probabilistic neural network predictions. Note the considerable variability in the dispersion of the probabilistic network's output, typically demonstrating high uncertainty when the deterministic result deviates substantially from ground truth.

probabilistic network can be achieved by calculating the mean of the probabilistic neural network output, as done in Figure 9.3. This analysis clearly reveals that even with the complete ensemble, which includes parameter instances with higher loss, the probabilistic predictions yield lower absolute errors.

Though training neural networks using HMC offers unique benefits, it also introduces a significantly higher computational cost during the training phase. As a result, it is generally not recommended for use in common settings. This method is best reserved for situations where a probabilistic output from the neural network is either desired or necessary.



(a) Predicted values comparison for neural networks trained using Adam and HMC, evaluated over the test dataset.

(b) Distribution of Mean Absolute Error (MAE) for the neural networks trained using Adam and HMC, evaluated over the test dataset.

Figure 9.3: Comparison of the prediction accuracy and error rates for the neural networks trained using both methods. Note that the Bayesian capabilities of the network were utilised for the HMC trained net, with predictions evaluated using the ensemble output mean instead of the best sampled set of parameters.

Part IV

Synthesis

Horizons for Bayesian FWI

The pace of development within Bayesian seismology is increasing. We are witnessing a rise in the number of tomographies performed using more advanced samplers and variational algorithms, such as Thurin et al. [2019], Guo et al. [2020], Huang et al. [2020], Zhang and Curtis [2021], Zhang et al. [2023a]. Looking ahead, there are several methodologies that could lead to further advancements in Bayesian FWI capabilities and related inverse problems.

Replica exchange for parallel and multi-fidelity tomography

New MCMC algorithms can be readily constructed from a combination of existing algorithms. As long as the separate transitions leave the distribution invariant, proposal mechanism from e.g. the RWMH, Gibbs and HMC sampler can alternatingly applied. This might affect the convergence behaviour of the chain, but in the limit will still produce i.i.d. samples of the target distribution.

This approach is how replica exchange algorithms are constructed. In these algorithms, multiple instances of a sampler are run in parallel, typically with varying tuning settings, creating multiple Markov chains. After a defined number of iterations by each sampler, exchange proposals are made between the chains. Because during the replica exchange a Metropolis correction step is applied, these exchanges also leave the target distribution invariant.

When multiple Markov chains are constructed over a distribution $p(\mathbf{m})$ that is differently tempered for each Markov chain, that is, it is raised to a different power of $1/T$,

$$p_i(\mathbf{m}) = p(\mathbf{m})^{\frac{1}{T_i}}, \tag{9.2}$$

the approach is known as parallel tempering [Dosso et al., 2012, Sambridge, 2014]. Figure 9.4 illustrated the Himmelblau distribution at various temperatures, sampled with 3 separate HMC instances and 3 replica exchanging HMC instances. It is clear that the instances that communicate by replica exchange achieve much better mixing with a limited number of evaluations. From the perspective of a single instance of a sampler in the parallel tempering scheme, the replica exchanges across temperatures simply emulates a much more informed proposal distribution. Practically, this we can design temperature schedules tailored to the computational resources available. Demonstrated in Klaasen et al.

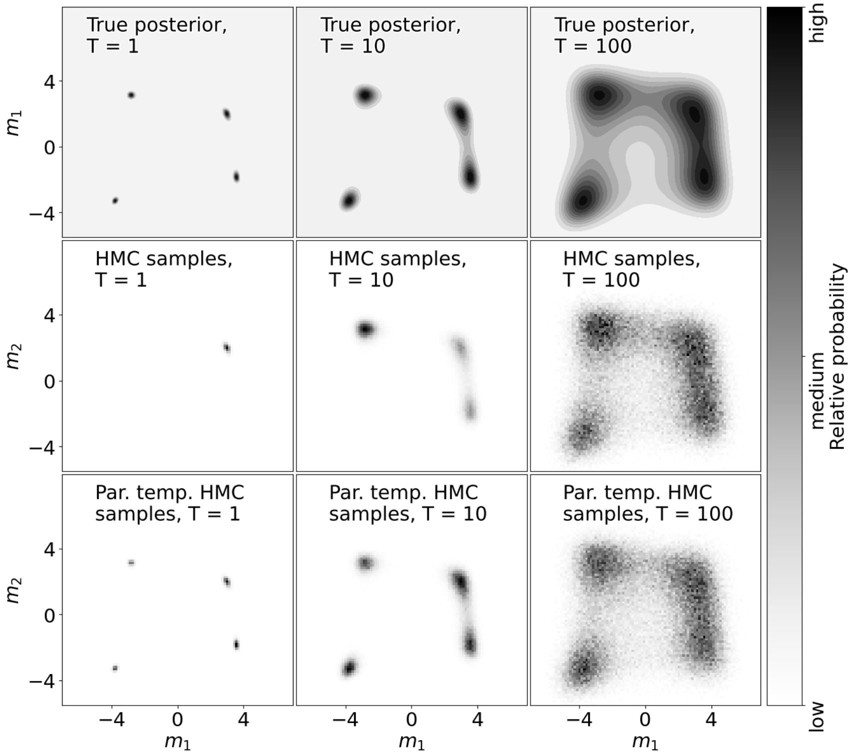


Figure 9.4: The Himmelblau distribution for various temperatures, sampled using unmodified HMC and parallel tempered HMC. Note how the individual sampler at temperature $T = 1$ fails to escape a local minima at all. The sampler appraising the distribution for temperature $T = 10$ is able to appraise more local minima, but still is biased by its initial model. Only after linking the samplers to a sampler that moves more freely through model space, are the distributions adequately sampled. The tuning settings of every algorithm were identical, with the stepsize of the time integration in HMC being estimated by autotuning. Acceptance rates for all samplers were around 88%.

[2023], these schedules should aim to saturate the available multithreaded capabilities of a system, accelerate mixing by including high temperatures, but also increase the number of samples by including multiple instances of the base distribution at $T = 1$, thus yielding samples from multiple instances for population MCMC [Laskey and Myers, 2003].

A further boost in efficiency could be achieved by an unstudied method. Recently, multi-fidelity sampling has gained popularity as a sequential approach to enhance PDE-based Bayesian inference, e.g. Cai and Adams [2022] and Nitzler et al. [2022]. The work by Cai and Adams [2022] demonstrates that transitioning to computationally manageable models can serve as shortcuts to efficiently explore relevant parts of the model space.

This concept can be extended to parallel replica exchange by using sampler instances with varying resolutions of the computational model. The more computationally efficient sampler instances can propose more states within the same computational resources required for the complex model to make one proposal. Consequently, these efficient samplers are likely to traverse greater distances in the model space. Subsequent replica exchanges would then propagate these distant states to the high-resolution sampler instance, leading to overall better convergence. Importantly, this does not alter the final distribution of samples from the separate models, as the Metropolis-Hastings correction steps ensure the convergence of the Markov chains.

In seismic tomography, the computational expense of a model primarily depends on the desired resolution of Earth's structure and the volume of included data, which is influenced by frequency content and the number of source-receiver combinations. High-resolution Bayesian inference in waveform-based tomography is strongly limited by the frequency scaling of computational costs. Although various approaches exist to mitigate some of these costs [Tromp and Bachmann, 2019, van Herwaarden et al., 2020, Thrastarson et al., 2022], multi-fidelity replica exchange allows for further parallelisation on modern high-performance computing hardware. This approach could potentially enable additional scaling of Bayesian FWI. Notably, the communication of model states in multi-fidelity replica exchange is much more demanding on an MPI interface compared to the communication of dynamical fields in wavefield simulations. Performance could be further improved by batching the data, similar to the approach in van Herwaarden et al. [2020], following the method of De Souza et al. [2022]. Although I am currently unaware of any works implementing this, it is possible that any alternative method to MCMC sampling for Bayesian FWI, such as ensemble filters or variational inference, could potentially benefit from the same parallelisation across resolutions.

Memory-HMC

In the HMC algorithm, a significant amount of information generated during its execution is typically discarded. Between the first and last state of a Hamiltonian trajectory, many gradient evaluations of the posterior go to waste. Over the years, I've had the urge to recycle this valuable information and incorporate it back into the sampler, into some sort of memory of traversed model space. However, doing so presents theoretical inconsistencies as it breaks the ergodic property. This, however, might not be a practical issue at all, as demonstrated by the autotuning strategies in Fichtner et al. [2021], which makes a first step in this regard, propagating information to an BFGS-style accumulator to adaptively precondition the posterior during sampling.

However, the variable curvature of arbitrary distributions is not accounted for in this approach. Inverse problems like FWI can exhibit strong non-linearity, which contradicts the assumption of BFGS to approximate the curvature of the target distribution using a single Hessian matrix. An alternative approach may be found in the field of physics-informed neural networks (PINNs), a relatively recent branch of machine learning. PINNs train regressors with the specific purpose of approximating an unknown high-dimensional function. In Bayesian FWI, the PINN regression model would not only be trained to minimise the loss between e.g. predicted and observed wavefields (input-output pairs) but also to satisfy the wave equation. A neural network trained during the Monte Carlo sampling of an inverse problem could be employed online to accelerate new proposals or aid in iterative runs of the algorithm.

Another potential extension of the HMC algorithm involves training adjoint-informed neural networks. Each evaluation of the posterior produces a negative log-likelihood value and its gradient, which has the potential to further accelerate neural network training. Preliminary tests with JAX [Bradbury et al., 2018] have shown that neural networks can be constructed to output a function and its mathematical derivatives. By combining the input model parameters and the output negative log-likelihood and its gradient with respect to the model parameters, these pairs of observations can be used to train these networks. The inclusion of gradients enhances the network's ability to reconstruct a function [Papoulis, 1977], similar to seismic gradiometry [Igel et al., 2007, Langston, 2007a,b].

Concluding remarks

When discussing our attempts to investigate high resolution tomographies using the methods of computational statistics, we are more often than not met with incredulity from statisticians. The scale at which Earth scientists would like to perform analyses does not correspond to the size of computational models statisticians seem to run. I have felt, however, that as Earth scientists, we possess some naive optimism about the possibilities of these mathematical and computational tools. The theme of this thesis has undeniably been computational statistics, and its application as a powerful method of analysis for geophysical inverse problems (Chapter 2), as a didactic tool (Chapter 3), an investigation of the tools of computational statistics (Chapter 4), and two contributions to the efficient solving of PDE-based models we use in this field (Chapter 5 and 6).

Chapter 2 demonstrates that inference on the scale of high-resolution tomographies is possible. This is enabled by the application of the HMC algorithm combined with the adjoint method to compute the derivatives needed for this algorithm. Furthermore, this chapter shows that density does imprint itself into our observation, although only through its spatial changes.

Chapter 3 relates the development of HMCLab software package, a collection of MCMC algorithms tailored to geophysical inverse problems, with a focus on gradient-based sampling. It describes various inverse problems included in the software package, but also encourages the user to probe their own inverse problems using the developed algorithms. This software fills the niche left open by popular sampling software as it ensures the interfaces, tutorials and examples speak the common language of geophysics.

Chapter 4 shows that although we might find ways to quantify the performance of an algorithm, the curse of dimensionality still remains an important limitation when dealing with high dimensional functions and distributions. The method developed in this chapter can be employed the determine which algorithm is appropriate for a specific inverse problem, although only at limited dimensionality.

Chapter 5 briefly documents the development of psvWave, a parallelised finite difference simulation code written in C++ to be used for high-performance FWI in limited settings. Interfaces to Python are developed, to enhance its utility to the seismological community.

Chapter 6 highlights the capabilities of modern unified chip systems, and the benefits it gives researchers wanting the write performant code in C++. The usage of unified chips allows bottlenecks in existing CPU codes to be accelerated by the GPU with minimal code rewriting. It's applicability is demonstrated on the psvWave suite.

Chapter 7 concludes with the reporting of multiple imagery-enhanced fieldworks during which the capabilities of modern Structure-from-Motion are demonstrated. These digitisation methods are shown to be accessible, allowing for the documentation of field sites, the safe surveying of hazardous fields, and finally the ability to create digital twins of structures and topography on which to simulate wave physics. Although the final chapter (Chapter 7) of this thesis is a relative outsider, it does reiterate the need for open accessible and well-designed software for any scientific field. The fact that I, as someone with relatively little experience on drone operation and 3D reconstructions, was able to go from field operation to digitisation and meshing of fields sites highlights both the accessibility of the WebODM software, and its potentials for seismology at large. I hope that with community involvement, HMCLab can be as enabling for Bayesian inference in seismology.

In my work, I have more focused on the how than the why of Bayesian inference in seismology, by enabling larger inferences to be performed faster (with the accelerated GPU physics) and simpler (with HMCLab). Going forward, seismologist will likely be able to do MCMC and variational inference on much bigger scales than demonstrated in this work. This will be great, as the desire for higher and higher resolution tomographies seems here to stay. It is useful in itself to generate large ensembles of solutions to inverse problems, even if it is only to escape local minima and end up at a better singular model in the end. We must, at the same time, also reflect on the why of Bayesian inference, already realised in our field 25 years ago [Scales and Snieder, 1997]; what do we really want to answer when we perform these tomographies? In accordance with that, I think the future lies in performing high-resolution Bayesian inference, and subsequently reducing these results effectively to answer hypotheses.

Acknowledgements

Opportunities multiply as they are seized.

That seems like a sensible quote to conclude my thesis. I won't pretend to have read The Art of War to counteract the corniness of quoting Sun Tzu, but I realise I needed to seize this quote as I chanced upon it while Googling for quotes on opportunity. When I sat in the benches of Andreas' 2017 course on Inverse Theory, I found my niche in geophysics. Two brash requests, one for an internship and the other for a PhD, and my opportunities had multiplied exponentially.

I cannot express enough gratitude to Andreas. First of all, for allowing me the opportunity to conduct research at one of the federal institutes in Switzerland. It's no secret that doctoral students in Switzerland are given much support to help them succeed, and I am extremely grateful for this. I did not know it was common for professors in Switzerland to have such an open house policy – living like kings for three months in your house was surreal.

The flexibility with which Andreas mentors all PhD students under his care is impressive, catering to everyone's personal needs. Although there might have been times when I was happily off randomising somewhere, whenever I required sparring over an idea or help with a problem, you were available. I think most of his students have an idea of how busy Andreas is, which makes his flexibility all the more impressive. The freedom this gave me has allowed me to become a more versatile scientist.

However, I would also like to reproach Andreas. You have made it far too difficult to make the choice to not continue full-time in an academic setting, but only for the right reasons. The resources and opportunities you provide us are incredible, and I believe I made ample use of them during my studies. Multiple research visits to Copenhagen and Edinburgh, attendance at many conferences and workshops, and the means to let a Bayesian, computational seismologist conduct his own fieldwork with drones – because why not? I am very happy that we can maintain a situation in which I can continue to interact with the SWP group, and I am excited about where your group will be in 10 years.

Although Andrea wasn't originally part of the SWP group, I have interacted with him since the inception of my research on HMC. The early invitation to Klaus's group and the idea to combine forces in developing the HMC sampler have laid the foundation for a very relaxed and rewarding collaboration. I am very glad that we got the chance to work together at the same institute, and you have been a Bayesian enrichment for our group. We have many ambitions for what HMC could potentially do for geophysics, and I hope to realise a few together in the future and ultimately make tons of money from the inevitable startup.

When I first walked into the offices of the Seismology and Wave Physics group, it was not called Seismology and Wave Physics. Computational Seismology in its original intent was more singularly focused on, well, computational seismology. During this time, I was allowed to write my master's thesis in H35, in between the doctoral candidates. This was a bit overwhelming and intimidating, but after an EGU Conference, I already felt part of the group. I am grateful for spending my first conference with Sövi and Dirk-Philip, which set us up for an amazing time in the office and outside of it, and I am so glad how this has cultivated our friendships. It can even be said that at times, such as the 2019 annual SWP ski day, we carried the group. You are a dynamic duo for sure, but I think having kids on the same day is a bit much.

Sölvi and I are birds of a feather. Literally, as the number of times we were confused for brothers is ridiculous (Jonas deserves an honorable mention here as well), although we don't really see it. However, our likeness goes beyond appearances. We both enjoy some good kamelåså, are connoisseurs of Sting-infused Reggae and motivational fitness music, and share a fondness for productive terminal applications (typing makes it shake!) and VSCode shortcuts. Thanks again for all the good times, and not least of all, academic guidance on writing and science in general. I recently (June 2023) learned how to pronounce your name, corrected by you as I was starting to pronounce Kári's name wrong. You had six years, just saying.

The enthusiasm with which Dirk-Philip undertakes everything is a huge inspiration to me. His zeal for understanding *anything* might have driven me a little cranky on busy days in the office, going in for the third blackboard session in an hour. At the same time, this is hugely motivating, as you simply refuse to not understand any concept. You show the same drive in your hobbies, and through you, I have gained quite a few, not the least of which are sourdough baking and pizza making.

The other person who stumbled into the SWP group at the same time as I did was Cyrill. For a long time, Cyrill was special as he was the only Swiss person in our office. That is, if you can actually catch him in the office. His questions relating Bayesian inference to fundamental philosophical concepts and his demands for explanations on the HMC sampler have led to a greater understanding of the subject matter for me and continuous collaboration on each other's research.

In the first year or two of my work at ETH, my alternate supervisor was Christian. While Andreas and I were trying to make heads or tails of the HMC method, it seemed that whatever insight we attained was already inherently within you. The depth and breadth of your understanding for any mathematical concept that could be discussed with you was a solid bedrock for me, and I think for many students in our group. You have tried to shun real data and real problems, but it seems that you have ended up in the real world at last. I am very much looking forward to the opportunity to work with you, Lion, and Mike on something exciting.

I would further like to highlight all the other group members with whom I have had the pleasure of overlappingly working in SWP. Nadine and Manolya offered excellent support, and Andrei, Ariane, Daniel, Emanuele, Ines, Iris, Joeri, Jonas, Korbinian, Krystyna, Laura, Lion, Maria, Marie, Marta, Martin, Mike, Naiara, Neda, Pascal, Pat, Patrick, Percy, Sara, Saule, Scott, Sebastian, Vaclav – it has been a pleasure!

Special thanks go out to a few seismologists who have strongly influenced the present work. First, Klaus Mosegaard, for the many insightful discussions and for hosting an academic visit to the Niels Bohr Institute. Similarly, I want to thank Heiner Igel (and naturally the rest of the Igel family) for being such great hosts for many Skience workshops, providing amazing learning environments, and singing a song here and there. These workshops were essential for me in connecting with experts in the field of Bayesian Seismology. Skience 2020 facilitated an extended discussion with Andrew Curtis, which resulted in an extended research visit to the University of Edinburgh and indirectly contributed to the work of Chapter 4. For this, I thank Heiner Igel for the connection, Andrew Curtis for his hospitality, and the HPC3-Europa program for its funding.

The Department of Earth Sciences at ETH is a special place for me, as I believe it is for many graduates. I highly value the social cohesion that is created by events such as the Doctoral Retreat and the Friday Peers. This would not be possible without the continued efforts of selfless people who care about connecting across research groups. To them and all the other amazing people I have gotten to know at D-ERDW, I want to express my appreciation. This makes for a truly great environment.

My social connections outside of the department have been a great support for me in my current work. They provide positive distractions when needed. Michi and Jesper, our trips and nights out with the "Biffles" are always fun. "Actieve Overachievers," you have brought a piece of home to Zurich. The "SCB," what are we even doing, three years later? Isamu and Florian, you know what I would write here. Without jest though, I am happy you are also lingering in Zurich. All the other people I met at D-ERDW, Friday Peers, and around Zurich, you make it a special place.

Lastly, the deepest well of inspiration and motivation comes from Nienke. I could not have a better example of perseverance than what you demonstrate on a regular basis. I don’t think anyone could ever work as hard as you do, initially for science, and now for your athletic ambitions. Certainly not me. Together, it is always the best of times, and you help motivate me to work hard, exercise harder, but also have a lot of fun.

Bibliography

M. V. Afanasiev, R. G. Pratt, R. Kamei, and G. McDowell. Waveform-based simulated annealing of crosshole transmission data: A semi-global method for estimating seismic anisotropy. *Geophys. J. Int.*, 199(3):1586–1607, Dec. 2014. doi: 10.1093/gji/ggu307.

M. V. Afanasiev, C. Boehm, M. v. Driel, L. Krischer, M. Rietmann, D. A. May, M. G. Knepley, and A. Fichtner. Modular and flexible spectral-element waveform modelling in two and three dimensions. *Geophys. J. Int.*, 216(3):1675–1692, Mar. 2019. doi: 10.1093/gji/ggy469.

Q. Ai, S. Liu, L. He, and Z. Xu. Stein Variational Gradient Descent with Multiple Kernels. *Cogn. Comput.*, 15(2):672–682, Nov. 2022. doi: 10.1007/s12559-022-10069-5.

K. Aki, A. Christoffersson, and E. S. Husebye. Determination of three-dimensional seismic structure of the lithosphere. *J. Geophys. Res.*, 81(2):277–296, Jan. 1977. doi: 10.1029/JB082i002p00277.

M. Amaru. *Global travel time tomography with 3-D reference models*, volume 274. Utrecht University, 2007.

F. Aminzadeh and J. Brac. SEG/EAGE 3-D Overthrust Models. *Zenodo*, Jan. 1997. doi: 10.5281/zenodo.4252588.

J. Y. Angela. Adaptive behavior: Humans act as bayesian learners. *Curr. Biol.*, 17(22): R977–R980, 2007.

Apple. Apple unleashes M1, 2020. URL `https://web.archive.org/web/20230707043055/https://www.apple.com/newsroom/2020/11/apple-unleashes-m1/`. Accessed: July 14, 2023.

Apple. Apple unveils M2, taking the breakthrough performance and capabilities of M1 even further, 2022a. URL `https://web.archive.org/web/20230619002026/https://www.apple.com/newsroom/2022/06/apple-unveils-m2-with-breakthrough-performance-and-capabilities/`. Accessed: July 14, 2023.

Apple. Get Started with tensorflow-metal, 2022b. URL https://web.archive.org/web/20230619034325/https://developer.apple.com/metal/tensorflow-plugin/. Accessed: July 14, 2023.

Apple. Getting started with Metal-cpp, 2022c. URL https://web.archive.org/web/20230613072609/https://developer.apple.com/metal/cpp/. Accessed: July 14, 2023.

G. B. Arfken, H. J. Weber, and F. E. Harris. *Mathematical methods for physicists: a comprehensive guide*. Academic press, 2011.

R. Arnold and A. Curtis. Interrogation theory. *Geophys. J. Int.*, 214(3):1830–1846, July 2018. doi: 10.1093/gji/ggy248.

G. E. Backus and J. F. Gilbert. Numerical Applications of a Formalism for Geophysical Inverse Problems. *Geophys. J. Int.*, 13(1-3):247–276, July 1967. doi: 10.1111/j.1365-246X.1967.tb02159.x.

A. Bamberger, G. Chavent, and P. Lailly. *Une application de la théorie du contrôle à un problème inverse sismique*, volume 33 of *Centre de mathématiques appliquées de l'École polytechnique*. École polytechnique, 1977. ISBN 0202-1284.

A. Bamberger, G. Chavent, C. Hemons, and P. Lailly. Inversion of normal incidence seismograms. *Geophysics*, 47(5):757–770, May 1982. doi: 10.1190/1.1441345.

T. Bayes. An essay towards solving a problem in the doctrine of chances. *Philos. Trans. R. Soc. London*, 53:370–418, Dec. 1763.

M. Betancourt. A conceptual introduction to Hamiltonian Monte Carlo. *arXiv*, Jan. 2017. doi: 10.48550/arXiv.1701.02434.

M. Beyreuther, R. Barsch, L. Krischer, T. Megies, Y. Behr, and J. Wassermann. ObsPy: A Python Toolbox for Seismology. *Seismol. Res. Lett.*, 81(3):530–533, May 2010. doi: 10.1785/gssrl.81.3.530.

J. Bezanson, A. Edelman, S. Karpinski, and V. B. Shah. Julia: A Fresh Approach to Numerical Computing. *SIAM Rev.*, 59(1):65–98, 2017. doi: 10.1137/141000671.

N. Blom, C. Boehm, and A. Fichtner. Synthetic inversions for density using seismic and gravity data. *Geophys. J. Int.*, 209(2):1204–1220, Mar. 2017. doi: 10.1093/gji/ggx076.

T. Bodin and M. Sambridge. Seismic Tomography with the Reversible Jump Algorithm. *Geophys. J. Int.*, 178(3):1411–1436, Sept. 2009. doi: 10.1111/j.1365-246X.2009.04226.x.

T. Bodin, M. Sambridge, N. Rawlinson, and P. Arroucau. Transdimensional tomography with unknown data noise: Transdimensional tomography. *Geophys. J. Int.*, 189(3):1536–1556, June 2012. doi: 10.1111/j.1365-246X.2012.05414.x.

R. C. Bolles, H. H. Baker, and D. H. Marimont. Epipolar-plane image analysis: An approach to determining structure from motion. *Int. J. Comput. Vision*, 1(1):7–55, Mar. 1987. doi: 10.1007/BF00128525.

M. Bosch. Lithologic Tomography: From Plural Geophysical Data to Lithology Estimation. *J. Geophys. Res.*, 104(B1):749–766, 1999. doi: 10.1029/1998JB900014.

E. Bozdağ, J. Trampert, and J. Tromp. Misfit functions for full waveform inversion based on instantaneous phase and envelope measurements. *Geophys. J. Int.*, 185(2):845–870, May 2011. doi: 10.1111/j.1365-246X.2011.04970.x.

E. Bozdağ, D. Peter, M. Lefebvre, D. Komatitsch, J. Tromp, J. Hill, N. Podhorszki, and D. Pugmire. Global adjoint tomography: First-generation model. *Geophys. J. Int.*, 207 (3):1739–1766, Dec. 2016. doi: 10.1093/gji/ggw356.

J. Bradbury, R. Frostig, P. Hawkins, M. J. Johnson, C. Leary, D. Maclaurin, G. Necula, A. Paszke, J. VanderPlas, S. Wanderman-Milne, and Q. Zhang. JAX: composable transformations of Python+NumPy programs, 2018. URL https://web.archive.org/web/20230626062913/http://github.com/google/jax. Accessed: July 14, 2023.

T. M. Brocher. Empirical relations between elastic wavespeeds and density in the Earth's crust. *Bull. Seismol. Soc. Am.*, 95(6):2081–2092, Dec. 2005. doi: 10.1785/0120050077.

R. Brossier, S. Operto, and J. Virieux. Which data residual norm for robust elastic frequency-domain full waveform inversion? *Geophysics*, 75(3):R37–R46, May 2010. doi: 10.1190/1.3379323.

R. Brun and F. Rademakers. ROOT—An object oriented data analysis framework. *Nucl. Instrum. Methods Phys. Res., Sect. A*, 389(1-2):81–86, Apr. 1997. doi: 10.1016/S0168-9002(97)00048-X.

T. Bui-Thanh, O. Ghattas, J. Martin, and G. Stadler. A Computational Framework for Infinite-Dimensional Bayesian Inverse Problems Part I: The Linearized Case, with Application to Global Seismic Inversion. *SIAM J. Sci. Comput.*, 35(6):A2494–A2523, Jan. 2013. doi: 10.1137/12089586X.

Bundesamt für Landestopografie. Vektordaten für bessere Karten. *Geomatik Schweiz*, 5 (116):120–126, May 2018.

C. Bunks, F. M. Saleck, S. Zaleski, and G. Chavent. Multiscale seismic waveform inversion. *Geophysics*, 60(5):1457–1473, Sept. 1995. doi: 10.1190/1.1443880.

D. Cai and R. P. Adams. Multi-fidelity monte carlo: a pseudo-marginal approach. *arXiv*, Oct. 2022. doi: 10.48550/arXiv.2210.01534.

D. L. Campbell. BASIC Programs to Calculate Gravity and Magnetic Anomalies for 2 1/2 - Dimensional Prismatic Bodies. Technical Report 83-154, U.S. Geological Survey,, 1983.

J. Canny. A Computational Approach to Edge Detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, PAMI-8(6):679–698, Nov. 1986. doi: 10.1109/TPAMI.1986.4767851.

J. L. Carrivick, M. W. Smith, and D. J. Quincey. *Structure from Motion in the Geosciences*. John Wiley & Sons, July 2016. ISBN 978-1-118-89581-8. doi: 10.1002/9781118895818.

P. Chen, L. Zhao, and T. H. Jordan. Full 3D tomography for the crustal structure of the Los Angeles region. *Bull. Seismol. Soc. Am.*, 97(4):1094–1120, Aug. 2007. doi: 10.1785/0120060222.

S. Chib and E. Greenberg. Understanding the Metropolis-Hastings algorithm. *Am. Stat.*, 49(4):327–335, Nov. 1995. doi: 10.2307/2684568.

A. Clark. Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behav. Brain Sci.*, 36(3):181–204, May 2013. doi: 10.1017/S0140525X12000477.

R. Cockett, S. Kang, L. J. Heagy, A. Pidlisecky, and D. W. Oldenburg. SimPEG: An Open Source Framework for Simulation and Gradient Based Parameter Estimation in Geophysical Applications. *Comput. Geosci.*, 85:142–154, Dec. 2015. doi: 10.1016/j.cageo.2015.09.015.

M. K. Cowles and B. P. Carlin. Markov chain Monte Carlo converge diagnostics: A comparative review. *J. Am. Stat. Assoc.*, 91(434):883–904, 1996. doi: 10.1080/01621459.1996.10476956.

M. Creutz. Global Monte Carlo algorithms for many-fermion systems. *Phys. Rev. D*, 38 (4):1228–1238, Aug. 1988. doi: 10.1103/PhysRevD.38.1228.

A. Curtis and A. Lomax. Prior Information, Sampling Distributions, and the Curse of Dimensionality. *Geophysics*, 66(2):372–378, Mar. 2001. doi: 10.1190/1.1444928.

F. Dahlen, S.-H. Hung, and G. Nolet. Fréchet kernels for finite-frequency traveltimes – I. Theory. *Geophys. J. Int.*, 141(1):157–174, Apr. 2000. doi: 10.1046/j.1365-246X.2000.00070.x.

J. Dahlin, F. Lindsten, and T. B. Schön. Quasi-Newton particle Metropolis-Hastings. *IFAC-PapersOnLine*, 48(28):981–986, Feb. 2015. doi: 10.1016/j.ifacol.2015.12.258.

J. W. Davidson and S. Jinturkar. Memory Access Coalescing: A Technique for Eliminating Redundant Memory Accesses. In *Proceedings of the ACM SIGPLAN 1994 Conference on Programming Language Design and Implementation*, PLDI ’94, pages 186–195. Association for Computing Machinery, 1994. doi: 10.1145/178243.178259.

D. A. De Souza, D. Mesquita, S. Kaski, and L. Acerbi. Parallel MCMC Without Embarrassing Failures. In G. Camps-Valls, F. J. R. Ruiz, and I. Valera, editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151, pages 1786–1804. PMLR, Mar. 2022.

S. E. Dosso, C. W. Holland, and M. Sambridge. Parallel tempering for strongly nonlinear geoacoustic inversion. *J. Acoust. Soc. Am.*, 132(5):3030–3040, Nov. 2012. doi: 10.1121/1.4757639.

S. Duane. Stochastic quantization versus the microcanonical ensemble: Getting the best of both worlds. *Nucl. Phys. B*, 257:652–662, Apr. 1985.

S. Duane, A. D. Kennedy, B. J. Pendleton, and D. Roweth. Hybrid Monte Carlo. *Phys. Lett. B*, 195(3):216–222, Sept. 1987. doi: 10.1016/0370-2693(87)91197-X.

T. Dubček, D. Moreno-Garcia, T. Haag, P. Omidvar, H. R. Thomsen, T. S. Becker, L. Gebraad, C. Bärlocher, F. Andersson, S. D. Huber, D.-J. van Manen, L. G. Villanueva, J. O. A. Robertsson, and M. Serra-Garcia. Binary classification of spoken words with passive phononic metamaterials. *arXiv*, July 2023. doi: 10.48550/arXiv.2111.08503.

A. M. Dziewoński, B. H. Hager, and R. J. O'Connell. Large-scale heterogeneities in the lower mantle. *J. Geophys. Res.*, 82(2):239–255, Jan. 1977. doi: 10.1029/JB082i002p00239.

W. Dębski. Chapter 1 - Probabilistic Inverse Theory. In R. Dmowska, editor, *Adv. Geophys.*, volume 52, pages 1–102. Elsevier, Jan. 2010. doi: 10.1016/S0065-2687(10)52001-6.

D. J. Earl and M. W. Deem. Parallel tempering: Theory, applications, and new perspectives. *Phys. Chem. Chem. Phys.*, 7(23):3910, Sept. 2005. doi: 10.1039/b509983h.

P. Edme, P. Paitz, F. Walter, A. van Herwijnen, and A. Fichtner. Fiber-optic detection of snow avalanches using telecommunication infrastructure. *arXiv*, Feb. 2023. doi: 10.48550/arXiv.2302.12649.

A. Fichtner. *Full Seismic Waveform Modelling and Inversion*. Advances in Geophysical and Environmental Mechanics and Mathematics. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011. ISBN 978-3-642-15806-3 978-3-642-15807-0. doi: 10.1007/978-3-642-15807-0.

A. Fichtner and T. v. Leeuwen. Resolution analysis by random probing. *J. Geophys. Res.*, 120(8):5549–5573, July 2015. doi: 10.1002/2015JB012106.

A. Fichtner and S. Simute. Hamiltonian Monte Carlo inversion of seismic sources in complex media. *J. Geophys. Res.*, 123(4):2984–2999, Mar. 2018. doi: 10.1002/2017JB015249.

A. Fichtner and J. Trampert. Hessian kernels of seismic data functionals based upon adjoint techniques. *Geophys. J. Int.*, 185(2):775–798, May 2011. doi: 10.1111/j.1365-246X.2011.04966.x.

A. Fichtner and A. Zunino. Hamiltonian nullspace shuttles. *Geophys. Res. Lett.*, 46(2): 644–651, Jan. 2019. doi: 10.1029/2018GL080931.

A. Fichtner, H.-P. Bunge, and H. Igel. The adjoint method in seismology - I. Theory. *Phys. Earth Planet. Inter.*, 157(1-2):86–104, Aug. 2006a. doi: 10.1016/j.pepi.2006.03.016.

A. Fichtner, H.-P. Bunge, and H. Igel. The adjoint method in seismology - II. Applications: traveltimes and sensitivity functionals. *Phys. Earth Planet. Inter.*, 157:105–123, 2006b. doi: 10.1016/j.pepi.2006.03.018.

A. Fichtner, B. L. N. Kennett, H. Igel, and H.-P. Bunge. Theoretical background for continental- and global-scale full-waveform inversion in the time-frequency domain. *Geophys. J. Int.*, 175(2):665–685, Nov. 2008. doi: 10.1111/j.1365-246X.2008.03923.x.

A. Fichtner, B. L. N. Kennett, H. Igel, and H.-P. Bunge. Full seismic waveform tomography for upper-mantle structure in the Australasian region using adjoint methods. *Geophys. J. Int.*, 179(3):1703–1725, Dec. 2009. doi: 10.1111/j.1365-246X.2009.04368.x.

A. Fichtner, D.-P. v. Herwaarden, M. Afanasiev, S. Simute, L. Krischer, Y. Cubuk-Sabuncu, T. Taymaz, L. Colli, E. Saygin, A. Villasenor, J. Trampert, P. Cupillard, H.-P. Bunge, and H. Igel. The Collaborative Seismic Earth Model: Generation I. *Geophys. Res. Lett.*, 45(9):4007–4016, May 2018a. doi: 10.1029/2018GL077338.

A. Fichtner, A. Zunino, and L. Gebraad. Hamiltonian Monte Carlo solution of tomographic inverse problems. *Geophys. J. Int.*, 216(2):1344–1363, Feb. 2018b. doi: 10.1093/gji/ggy496.

A. Fichtner, A. Zunino, L. Gebraad, and C. Boehm. Autotuning Hamiltonian Monte Carlo for Efficient Generalized Nullspace Exploration. *Geophys. J. Int.*, 227(2):941–968, July 2021. doi: 10.1093/gji/ggab270.

A. Fichtner, S. Klaasen, S. Thrastarson, Y. Çubuk Sabuncu, P. Paitz, and K. Jónsdóttir. Fiber-Optic Observation of Volcanic Tremor through Floating Ice Sheet Resonance. *The Seismic Record*, 2(3):148–155, July 2022. doi: 10.1785/0320220010.

A. Fichtner, C. Hofstede, L. Gebraad, A. Zunino, D. Zigone, and O. Eisen. Borehole fibre-optic seismology inside the Northeast Greenland Ice Stream. *arXiv*, July 2023. doi: 10.48550/arXiv.2307.05976.

P. C. Fletcher and C. D. Frith. Perceiving is believing: a bayesian approach to explaining the positive symptoms of schizophrenia. *Nat. Rev. Neurosci.*, 10(1):48–58, Dec. 2009.

S. W. French and B. A. Romanowicz. Whole-mantle radially anisotropic shear velocity structure from spectral-element waveform tomography. *Geophys. J. Int.*, 199(3):1303–1327, Sept. 2014. doi: 10.1093/gji/ggu334.

W. Friederich. The S-velocity structure of the East Asian mantle from inversion of shear and surface waveforms. *Geophys. J. Int.*, 153(1):88–102, Apr. 2003. doi: 10.1046/j.1365-246X.2003.01869.x.

T. Fu, L. Luo, and Z. Zhang. Quasi-Newton Hamiltonian Monte Carlo. In *Proc. 32nd Conf. Uncert. Art. Int.*, pages 212–221. AUAI Press, Arlington, U.S., 2016.

O. Gauthier, J. Virieux, and A. Tarantola. Two-dimensional nonlinear inversion of seismic waveforms: numerical results. *Geophysics*, 51(7):1387–1403, July 1986. doi: 10.1190/1.1442188.

L. Gebraad. larsgeb/psvwave: Version 1.0. *Zenodo*, Feb. 2022a. doi: 10.5281/zenodo.6216312.

L. Gebraad. simpleSVGD: A Tiny Interface to Stein Variational Gradient Descent Using Various Optimization Algorithms. *Zenodo*, Jan. 2022b. doi: 10.5281/zenodo.5938430.

L. Gebraad and A. Fichtner. MSL for Scientific C++ portal, 2022a. URL `https://web.archive.org/web/20220703034555/https://larsgeb.github.io/msl-portal/`. Accessed: July 14, 2023.

L. Gebraad and A. Fichtner. psvWave: elastic wave propagation in 2d for Python and C++. *EarthArXiv*, Feb. 2022b. doi: 10.31223/X5R91Q.

L. Gebraad and A. Fichtner. Seamless GPU Acceleration for C++-Based Physics with the Metal Shading Language on Apple's M Series Unified Chips. *Seismol. Res. Lett.*, 94(3):1670–1675, Feb. 2023. doi: 10.1785/0220220241.

L. Gebraad, C. Boehm, and A. Fichtner. Code for performing Bayesian Full-waveform inversion using Hamiltonian Monte Carlo. *Zenodo*, Dec. 2019. doi: 10.5281/zenodo.3565313.

L. Gebraad, C. Boehm, and A. Fichtner. Bayesian Elastic Full-Waveform Inversion Using Hamiltonian Monte Carlo. *J. Geophys. Res.*, 125(3):e2019JB018428, Feb. 2020. doi: 10.1029/2019JB018428.

L. Gebraad, I. Naets, P. Marty, and A. Fichtner. Structure-from-Motion for Seismology: imagery and reconstructions. *Zenodo*, Aug. 2023. doi: 10.5281/zenodo.8105812.

L. S. Gee and T. H. Jordan. Generalized seismological data functionals. *Geophys. J. Int.*, 111(2):363–390, Nov. 1992. doi: 10.1111/j.1365-246X.1992.tb00584.x.

A. Gelman and D. B. Rubin. Inference from iterative simulation using multiple sequences. *Statist. Sci.*, 7(4):457–472, Nov. 1992. doi: 10.1214/ss/1177011136.

S. Geman and D. Geman. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Trans. Pattern Anal. Mach. Intell.*, PAMI-6(6):721–741, Nov. 1984. doi: 10.1109/TPAMI.1984.4767596.

J. Geweke. Evaluating the accuracy of sampling-based approaches to the calculations of posterior moments. *Bayesian Stat.*, 4:641–649, Dec. 1991.

C. J. Geyer. Markov chain Monte Carlo maximum likelihood. In *Comp. Sci. Stat., Proc. 23rd Symposium on the Interface*. Interface Foundation of North America, 1991.

C. J. Geyer and E. A. Thompson. Annealing Markov Chain Monte Carlo with applications to ancestral inference. *J. Am. Stat. Assoc.*, 90(431):909–920, Sept. 1995. doi: 10.1080/01621459.1995.10476590.

W. R. Gilks, N. G. Best, and K. K. C. Tan. Adaptive Rejection Metropolis Sampling within Gibbs Sampling. *Appl. Statist.*, 44(4):455–472, 1995. doi: 10.2307/2986138.

W. R. Gilks, S. Richardson, and D. Spiegelhalter, editors. *Markov Chain Monte Carlo in Practice*. Chapman and Hall/CRC, Boca Raton, Fla, 1st edition edition, Jan. 1996. ISBN 978-0-412-05551-5.

M. Girolami and B. Calderhead. Riemann manifold langevin and hamiltonian monte carlo methods. *J. R. Stat. Soc. B*, 73(2):123–214, Mar. 2011. doi: 10.1111/j.1467-9868.2010.00765.x.

A. Gokhberg and A. Fichtner. Full-waveform inversion on heterogeneous HPC systems. *Comput. Geosci.*, 89:260–268, Apr. 2016. doi: 10.1016/j.cageo.2015.12.013.

A. Gorbatov and B. L. N. Kennett. Joint bulk-sound and shear tomography for Western Pacific subduction zones. *Earth Planet. Sci. Lett.*, 210(3-4):527–543, May 2003. doi: 10.1016/S0012-821X(03)00165-1.

S. Grand, R. VanDerHilst, and S. Widiyantoro. Global seismic tomography: A snapshot of convection in the Earth. *Geol. Soc. Am. Today*, 7, No.4:1–7, Jan. 1997.

P. J. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732, Dec. 1995. doi: 10.1093/biomet/82.4.711.

A. Griewank and A. Walther. *Evaluating Derivatives: Principles and Techniques of Algorithmic Differentiation*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2nd ed edition, 2008. ISBN 978-0-89871-659-7.

C. Griwodz, S. Gasparini, L. Calvet, P. Gurdjos, F. Castan, B. Maujean, G. D. Lillo, and Y. Lanthony. AliceVision Meshroom: An open-source 3D reconstruction pipeline. In *Proceedings of the 12th ACM Multimedia Systems Conference - MMSys '21*. ACM Press, 2021. doi: 10.1145/3458305.3478443.

L. Guasch, O. Calderón Agudo, M.-X. Tang, P. Nachev, and M. Warner. Full-waveform inversion imaging of the human brain. *npj Digit. Med.*, 3(1):1–12, Mar. 2020. doi: 10.1038/s41746-020-0240-8.

G. Guennebaud, B. Jacob, and others. Eigen v3, 2010. URL https://web.archive.org/web/20230706031120/https://eigen.tuxfamily.org/index.php?title=Main_Page. Accessed: July 14, 2023.

P. Guo, G. Visser, and E. Saygin. Bayesian trans-dimensional full waveform inversion: synthetic and field data application. *Geophys. J. Int.*, 222(1):610–627, Mar. 2020. doi: 10.1093/gji/ggaa201.

T. M. Hansen, K. S. Cordua, and K. Mosegaard. Inverse Problems with Non-Trivial Priors: Efficient Solution through Sequential Gibbs Sampling. *Comput. Geosci.*, 16(3):593–611, June 2012. doi: 10.1007/s10596-011-9271-1.

T. M. Hansen, K. S. Cordua, A. Zunino, and K. Mosegaard. Probabilistic Integration of Geo-Information. In *Integrated Imaging of the Earth*, pages 93–116. American Geophysical Union (AGU), 2016. ISBN 978-1-118-92906-3. doi: 10.1002/9781118929063.ch6.

C. R. Harris, K. J. Millman, S. J. v. d. Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. v. Kerkwijk, M. Brett, A. Haldane, J. F. d. Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, Sept. 2020. doi: 10.1038/s41586-020-2649-2.

W. K. Hastings. Monte Carlo sampling methods using Markov Chains and their applications. *Biometrika*, 57(1):97–109, Apr. 1970. doi: 10.1093/biomet/57.1.97.

R. D. v. d. Hilst, S. Widiyantoro, and E. R. Engdahl. Evidence for deep mantle circulation from global tomography. *Nature*, 386:578–584, Apr. 1997. doi: 10.1038/386578a0.

M. Hinze, R. Pinnau, M. Ulbrich, and S. Ulbrich. *Optimization with PDE Constraints*. Springer, New York, 2009th edition edition, Nov. 2008. ISBN 978-1-4020-8838-4.

M. D. Hoffmann and A. Gelman. The No-U-Turn sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res.*, 15(1):1593–1623, Jan. 2014. doi: 10.5555/2627435.2638586.

K. Hosseini, K. Sigloch, M. Tsekhmistrenko, A. Zaheri, T. Nissen-Meyer, and H. Igel. Global mantle structure from multifrequency tomography using P, PP and P-diffracted waves. *Geophys. J. Int.*, 220(1):96–141, Sept. 2019. doi: 10.1093/gji/ggz394.

X. Huang, K. S. Eikrem, M. Jakobsen, and G. Nævdal. Bayesian full-waveform inversion in anisotropic elastic media using the iterated extended Kalman filter. *Geophysics*, 85 (4):C125–C139, June 2020. doi: 10.1190/geo2019-0644.1.

J. D. Hunter. Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.*, 9(3):90–95, May 2007. doi: 10.1109/MCSE.2007.55.

J. Hunziker, E. Laloy, and N. Linde. Bayesian full-waveform tomography with application to crosshole ground penetrating radar data. *Geophys. J. Int.*, 218(2):913–931, Apr. 2019. doi: 10.1093/gji/ggz194.

H. Igel, H. Djikpesse, and A. Tarantola. Waveform inversion of marine reflection seismograms for P impedance and Poisson's ratio. *Geophys. J. Int.*, 124(2):363–371, Feb. 1996. doi: 10.1111/j.1365-246X.1996.tb07026.x.

H. Igel, A. Cochard, J. Wassermann, A. Flaws, U. Schreiber, A. Velikoseltsev, and N. Pham Dinh. Broad-band observations of earthquake-induced rotational ground motions. *Geophys. J. Int.*, 168(1):182–196, Jan. 2007.

J. Iglhaut, C. Cabo, S. Puliti, L. Piermattei, J. O'Connor, and J. Rosette. Structure from motion photogrammetry in forestry: A review. *Curr. For. Rep.*, 5:155–168, July 2019. doi: 10.1007/s40725-019-00094-3.

W. Jakob, J. Rhinelander, and D. Moldovan. pybind11 – Seamless operability between C++11 and Python, 2017.

E. T. Jaynes. Information theory and statistical mechanics. *Phys. Rev.*, 106(4):620–630, May 1957. doi: 10.1103/PhysRev.106.620.

E. T. Jaynes. How does the brain do plausible reasoning? In *Maximum-entropy and Bayesian methods in science and engineering: Foundations*, pages 1–24. Springer, 1988.

E. T. Jaynes. *Probability Theory - The Logic of Science*. Cambridge University Press, Cambridge, UK, 2003.

H. Jeffreys. *Theory of Probability*. The International Series of Managraphs on Physics. Oxford University Press, 1939.

K. Johnson, E. Nissen, S. Saripalli, J. R. Arrowsmith, P. McGarey, K. Scharer, P. Williams, and K. Blisniuk. Rapid mapping of ultrafine fault zone topography with structure from motion. *Geosphere*, 10(5):969–986, Oct. 2014. doi: 10.1130/GES01017.1.

S. Kaiser, J. Boike, G. Grosse, and M. Langer. The Potential of UAV Imagery for the Detection of Rapid Permafrost Degradation: Assessing the Impacts on Critical Arctic Infrastructure. *Remote Sens.*, 14(23):6107, Dec. 2022. doi: 10.3390/rs14236107.

R. E. Kayen, S. Gori, B. Lingwall, F. Galadini, E. Falcucci, K. Franke, J. Stewart, and P. Zimmaro. Mt. Vettore fault zone rupture-LIDAR-and UAS-based Structure-from-Motion computational imaging. In *Proceedings 16th European Conference on Earthquake Engineering (16ECEE)*, 2018.

V. I. Keilis-Borok and T. B. Yanovskaja. Inverse Problems of Seismology (Structural Review). *Geophys. J. Int.*, 13(1-3):223–234, July 1967. doi: 10.1111/j.1365-246X.1967.tb02156.x.

S. Khoshkholgh, A. Zunino, and K. Mosegaard. Informed Proposal Monte Carlo. *Geophys. J. Int.*, 226(2):1239–1248, Aug. 2021. doi: 10.1093/gji/ggab173.

S. Khoshkholgh, A. Zunino, and K. Mosegaard. Full-Waveform Inversion by Informed-Proposal Monte Carlo. *Geophys. J. Int.*, 230(3):1824–1833, Sept. 2022. doi: 10.1093/gji/ggac150.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv*, Dec. 2014. doi: doi.org/10.48550/arXiv.1412.6980.

I. Kiss, S. Gyimothy, Z. Badics, and J. Pavo. Parallel realization of the element-by-element FEM technique by CUDA. *IEEE Trans. Magn.*, 48(2):507–510, Jan. 2012. doi: 10.1109/TMAG.2011.2175905.

S. Klaasen, S. Thrastarson, A. Fichtner, Y. Çubuk Sabuncu, and K. Jónsdóttir. Sensing Iceland's Most Active Volcano with a "Buried Hair". *Eos*, 103, Jan. 2022. doi: 10.1029/2022EO220007.

S. Klaasen, S. Thrastarson, Y. Çubuk-Sabuncu, K. Jónsdóttir, L. Gebraad, P. Paitz, and A. Fichtner. Subglacial volcano monitoring with fiber-optic sensing: Grímsvötn, Iceland. *Volcanica*, 6(2):301–311, 2023.

D. Komatitsch and R. Martin. An Unsplit Convolutional Perfectly Matched Layer Improved at Grazing Incidence for the Seismic Wave Equation. *Geophysics*, 72(5):SM155–SM167, Sept. 2007. doi: 10.1190/1.2757586.

D. Komatitsch and J. Tromp. Introduction to the Spectral Element Method for Three-Dimensional Seismic Wave Propagation. *Geophys. J. Int.*, 139(3):806–822, Dec. 1999. doi: 10.1046/j.1365-246x.1999.00967.x.

D. Komatitsch and J. Tromp. Spectral-element simulations of global seismic wave propagation, Part II: 3-D models, oceans, rotation, and gravity. *Geophys. J. Int.*, 150:303–318, 2002a. doi: 10.1046/j.1365-246X.2002.01716.x.

D. Komatitsch and J. Tromp. Spectral-element simulations of global seismic wave propagation, Part I: validation. *Geophys. J. Int.*, 149(2):390–412, May 2002b. doi: 10.1046/j.1365-246X.2002.01653.x.

D. Komatitsch and J.-P. Vilotte. The Spectral Element Method: An Efficient Tool to Simulate the Seismic Response of 2D and 3D Geological Structures. *Bull. Seismol. Soc. Am.*, 88(2):368–392, Apr. 1998. doi: 10.1785/BSSA0880020368.

A. Kordjazi, J. T. Coe, and M. Afanasiev. The Use of the Spectral Element Method for Modeling Stress Wave Propagation in Non-Destructive Testing Applications for Drilled Shafts. In *Geo-Congress 2020: Modeling, Geomaterials, and Site Characterization*, pages 434–443. American Society of Civil Engineers Reston, VA, Feb. 2020.

M. Kotsi, A. E. Malcolm, and G. Ely. 4D full-waveform Metropolis-Hastings inversion using a local acoustic solver. *SEG Expanded Abstracts*, pages 5323–5327, Aug. 2018. doi: 10.1190/segam2018-2997858.1.

L. Krischer, T. Megies, R. Barsch, M. Beyreuther, T. Lecocq, C. Caudron, and J. Wassermann. ObsPy: A Bridge for Seismology into the Scientific Python Ecosystem. *Comput. Sci. Discovery*, 8(1), May 2015. doi: 10.1088/1749-4699/8/1/014003.

L. Krischer, H. Igel, and A. Fichtner. Automated large-scale full seismic waveform inversion for North America and the North Atlantic. *J. Geophys. Res.*, 123(7):5902–5928, Apr. 2018. doi: 10.1029/2017JB015289.

P. Käufl, A. Fichtner, and H. Igel. Probabilistic full waveform inversion based on tectonic regionalisation - Development and application to the Australian upper mantle. *Geophys. J. Int.*, 193(1):437–451, Apr. 2013. doi: 10.1093/gji/ggs131.

D. Köhn, A. Kurzmann, A. Przebindowska, D. d. Nil, and T. Bohlen. 2D elastic full waveform tomography of synthetic marine reflection seismic data. In *72nd EAGE Conference & Exhibition 2010*, volume 1/2010, pages 23–27. European Association of Geoscientists & Engineers, June 2010. ISBN 978-90-73781-86-3. doi: 10.3997/2214-4609.201401242.

P. Lailly. The seismic inverse problem as a sequence of before stack migrations. In J. Bednar, R. Redner, E. Robinson, and A. Weglein, editors, *Conference on Inverse Scattering: Theory and Application*. Soc. Industr. Appl. Math., Philadelphia, PA., 1983.

L. D. Landau and E. M. Lifshitz. *Course of Theoretical Physics, Volume 1, Mechanics, 3rd edition*. Elsevier Butterworth Heinemann, Amsterdam, 1976.

C. A. Langston. Spatial gradient analysis for linear seismic arrays. *Bull. Seismol. Soc. Am.*, 97(1B):265–280, Feb. 2007a.

C. A. Langston. Wave gradiometry in two dimensions. *Bull. Seismol. Soc. Am.*, 97(2): 401–416, Apr. 2007b.

A. Lanteri, L. Gebraad, A. Zunino, and A. Fichtner. Hamiltonian Monte Carlo inversion of surface wave dispersion to evaluate their potential to constrain the density distribution in the Earth. In *EGU General Assembly Conference Abstracts*, pages EGU22–6399, May 2022. doi: 10.5194/egusphere-egu22-6399.

A. Lanteri, L. Gebraad, A. Zunino, S. Klaasen, K. Jonsdottir, C. Hofstede, O. Eisen, D. Zigone, and A. Fichtner. Bayesian surface wave dispersion inversion of glaciated environments. In *General Assembly of the International Union of Geodesy and Geophysics (IUGG)*. GFZ German Research Centre for Geosciences, July 2023. doi: 10.57757/IUGG23-1912.

P. S. Laplace. Mémoire sur la probabilité de causes par les évenements. *Mémoire de l'académie royale des sciences*, 1774.

P. S. Laplace. *Théorie analytique des probabilités*. Courcier, 1814.

K. B. Laskey and J. W. Myers. Population Markov chain Monte Carlo. *Mach. Learn.*, 50: 175–196, Jan. 2003. doi: 10.1023/A:1020206129842.

T. v. Leeuwen and W. A. Mulder. A correlation-based misfit criterion for wave-equation traveltime tomography. *Geophys. J. Int.*, 182(3):1383–1394, Sept. 2010. doi: 10.1111/j.1365-246X.2010.04681.x.

W. Lei, Y. Ruan, E. Bozdağ, D. Peter, M. Lefebvre, D. Komatitsch, J. Tromp, J. Hill, N. Podhorszki, and D. Pugmire. Global adjoint tomography—model GLAD-M25. *Geophys. J. Int.*, 223(1):1–21, May 2020. doi: 10.1093/gji/ggaa253.

S. Leung and J. Qian. An Adjoint State Method For Three-dimensional Transmission Traveltime Tomography Using First-Arrivals. *Commun. Math. Sci.*, 4(1):249–266, Mar. 2006.

N. Linde, P. Renard, T. Mukerji, and J. Caers. Geological realism in hydrogeological and geophysical inverse modeling: A review. *Adv. Water Resour.*, 86:86–101, Dec. 2015. doi: 10.1016/j.advwatres.2015.09.019.

J.-L. Lions. *Contrôle optimal de systèmes gouvernés par des équations aux dérivées partielles*. Dunod Gauthier-Villars, 1968.

J.-L. Lions. *Optimal Control of Systems Governed by Partial Differential Equations*. Springer-Verlag, 1971. ISBN 978-0-387-05115-4.

J. S. Liu and R. Chen. Sequential Monte Carlo Methods for Dynamic Systems. *J. Am. Stat. Assoc.*, 93(443):1032–1044, Sept. 1998. doi: 10.1080/01621459.1998.10473765.

L. Liu, D. Peter, and C. Tape. Square-root variable metric based elastic full-waveform inversion - Part 1: theory and validation. *Geophys. J. Int.*, 218(2):1100–1120, Aug. 2019a. doi: 10.1093/gji/ggz188.

L. Liu, D. Peter, and C. Tape. Square-root variable metric based elastic full-waveform inversion - Part 2: uncertainty estimation. *Geophys. J. Int.*, 218(2):1121–1135, Aug. 2019b. doi: 10.1093/gji/ggz137.

Q. Liu and Y. Gu. Seismic imaging: from classical to adjoint tomography. *Tectonophysics*, 566-567:31–66, 2012. doi: 10.1016/j.tecto.2012.07.006.

Q. Liu and J. Tromp. Finite-frequency kernels based on adjoint methods. *Bull. Seismol. Soc. Am.*, 96(6):2383–2397, Dec. 2006. doi: 10.1785/0120060041.

Q. Liu and D. Wang. Stein Variational Gradient Descent: A General Purpose Bayesian Inference Algorithm. *arXiv*, Sept. 2019. doi: 10.48550/arXiv.1608.04471.

T. Louvet, V. Maillou, F. Bohte, L. Gebraad, and M. Serra-Garcia. Training elastic neural networks with the Hamiltonian Monte Carlo sampling algorithm. *Bull. Am. Phys. Soc.*, Mar. 2023.

Y. Luo and G. T. Schuster. Wave-equation traveltime inversion. *Geophysics*, 56(5):645–653, May 1991. doi: 10.1190/1.1443081.

A. Malinverno. Parsimonious Bayesian Markov Chain Monte Carlo Inversion in a Nonlinear Geophysical Problem. *Geophys. J. Int.*, 151(3):675–688, Dec. 2002. doi: 10.1046/j.1365-246X.2002.01847.x.

A. Malinverno and V. A. Briggs. Expanded uncertainty quantification in inverse problems: Hierarchical Bayes and empirical Bayes. *Geophysics*, 69(4):1005–1016, July 2004. doi: 10.1190/1.1778243.

R. Mallet and J. W. Mallet. *The Earthquake Catalogue of the British Association: With the Discussion, Curves, and Maps, Etc.* Taylor & Francis, 1858.

P. Marty, C. Boehm, and A. Fichtner. Acoustoelastic full-waveform inversion for transcranial ultrasound computed tomography. In *Medical Imaging 2021: Ultrasonic Imaging and Tomography*, volume 11602, page 1160211. International Society for Optics and Photonics, Feb. 2021. doi: 10.1117/12.2581029.

G. Mavko, T. Mukerji, and J. Dvorkin. *The Rock Physics Handbook: Tools for Seismic Analysis of Porous Media*. Cambridge University Press, Oct. 2003. ISBN 978-0-521-54344-6. doi: 10.1017/CBO9780511626753.

N. Metropolis and S. Ulam. The Monte Carlo Method. *J. Am. Stat. Assoc.*, 44(247): 335–341, 1949. doi: 10.1080/01621459.1949.10483310.

N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equations of state calculations by fast computing machines. *J. Chem. Phys.*, 21:1087–1092, June 1953. doi: 10.1063/1.1699114.

K. Mosegaard. Quest for Consistency, Symmetry, and Simplicity — The Legacy of Albert Tarantola. *Geophysics*, 76(5):W51–W61, Sept. 2011. doi: 10.1190/geo2010-0328.1.

K. Mosegaard and M. Sambridge. Monte Carlo Analysis of Inverse Problems. *Inverse Probl.*, 18(3):R29–R54, Apr. 2002. doi: 10.1088/0266-5611/18/3/201.

K. Mosegaard and A. Tarantola. Monte Carlo sampling of solutions to inverse problems. *J. Geophys. Res.*, 100(B7):12431–12447, July 1995. doi: 10.1029/94JB03097.

L. Murray. Distributed markov chain monte carlo. In *Proceedings of neural information processing systems workshop on learning on cores, clusters and clouds*, volume 11, 2010.

L. Métivier, R. Brossier, Q. Mérigot, E. Oudet, and J. Virieux. Measuring the misfit between seismograms using an optimal transport distance: application to full waveform inversion. *Geophys. J. Int.*, 205(1):345–377, Apr. 2016. doi: 10.1093/gji/ggw014.

R. Neal. Slice Sampling. *Ann. Stat.*, 31(3):705–767, June 2003. doi: 10.1214/aos/1056562461.

R. Neal. *MCMC Using Hamiltonian Dynamics*. Chapman and Hall/CRC, May 2011. doi: 10.1201/b10905-6.

R. M. Neal. *Bayesian learning for neural networks*, volume 118 of *Lecture Notes in Statistics*. Springer Science & Business Media, 1996. doi: 10.1007/978-1-4612-0745-0.

W. Neiswanger, C. Wang, and E. P. Xing. Asymptotically Exact, Embarrassingly Parallel MCMC. In *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence*, UAI'14, pages 623–632, Arlington, Virginia, USA, 2014. AUAI Press. ISBN 978-0-9749039-1-0.

L. T. Nguyen and R. T. Modrak. Ultrasonic wavefield inversion and migration in complex heterogeneous structures: 2D numerical imaging and nondestructive testing experiments. *Ultrasonics*, 82:357–370, 2018. doi: 10.1016/j.ultras.2017.09.011.

T. Nissen-Meyer, A. Fournier, and F. A. Dahlen. A two-dimensional spectral-element method for computing spherical-earth seismograms - I. Moment-tensor source. *Geophys. J. Int.*, 168(3):1067–1092, Mar. 2007. doi: 10.1111/j.1365-246X.2006.03121.x.

T. Nissen-Meyer, A. Fournier, and F. A. Dahlen. A two-dimensional spectral-element method for computing spherical-earth seismograms - II. Waves in solid-fluid media. *Geophys. J. Int.*, 174(3):873–888, Sept. 2008. doi: 10.1111/j.1365-246X.2008.03813.x.

T. Nissen-Meyer, M. van Driel, S. C. Stähler, K. Hosseini, S. Hempel, L. Auer, A. Colombi, and A. Fournier. AxiSEM: Broadband 3-D Seismic Wavefields in Axisymmetric Media. *Solid Earth*, 5(1):425–445, June 2014. doi: 10.5194/se-5-425-2014.

J. Nitzler, J. Biehler, N. Fehn, P.-S. Koutsourelakis, and W. A. Wall. A generalized probabilistic learning approach for multi-fidelity uncertainty quantification in complex physical simulations. *Comput. Methods Appl. Mech. Eng.*, 400:115600, Oct. 2022. doi: 10.1016/j.cma.2022.115600.

J. Nocedal. Updating quasi-Newton matrices with limited storage. *Math. Comp.*, 35(151): 773–782, 1980. doi: 10.2307/2006193.

J. Nocedal and S. J. Wright. *Numerical optimization*. Springer series in operations research. Springer, New York, 2nd ed edition, 2006. ISBN 978-0-387-30303-1.

NVidia. How to Access Global Memory Efficiently in CUDA C/C++ Kernels, 2013. URL `https://web.archive.org/web/20230530054523/https://developer.nvidia.com/blog/how-access-global-memory-efficiently-cuda-c-kernels/`. Accessed: July 14, 2023.

NVidia, P. Vingelmann, and F. H. Fitzek. CUDA, release: 11.7, 2022. URL `https://web.archive.org/web/20230702025457/https://developer.nvidia.com/cuda-toolkit`. Accessed: July 14, 2023.

A. Papoulis. Generalized sampling expansion. *IEEE Trans. Circuits Syst.*, 24(11):652–654, 1977.

D. Pasalic and R. McGarry. Convolutional Perfectly Matched Layer for Isotropic and Anisotropic Acoustic Wave Equations. In *SEG Technical Program Expanded Abstracts 2010*, pages 2925–2929. Society of Exploration Geophysicists, Jan. 2010. doi: 10.1190/1.3513453.

D. Peter, D. Komatitsch, Y. Luo, R. Martin, N. L. Goff, E. Casarotti, P. L. Loher, F. Magnoni, Q. Liu, C. Blitz, T. Nissen-Meyer, P. Basini, and J. Tromp. Forward and adjoint simulations of seismic wave propagation on fully unstructured hexahedral meshes. *Geophys. J. Int.*, 186(2):721–739, Aug. 2011. doi: 10.1111/j.1365-246X.2011.05044.x.

I. Pierce, A. Williams, R. D. Koehler, and C. Chupik. High-resolution structure-from-motion models and orthophotos of the southern sections of the 2019 Mw 7.1 and 6.4 Ridgecrest earthquakes surface ruptures. *Seismol. Res. Lett.*, 91(4):2124–2126, 2020. doi: 10.1785/0220190289.

R.-E. Plessix. A Review of the Adjoint-State Method for Computing the Gradient of a Functional with Geophysical Applications. *Geophys. J. Int.*, 167(2):495–503, Nov. 2006. doi: 10.1111/j.1365-246X.2006.02978.x.

F. Press. Earth Models Obtained by Monte Carlo Inversion. *J. Geophys. Res.*, 73(16):5223–5234, Aug. 1968. doi: 10.1029/JB073i016p05223.

V. Prieux, R. Brossier, S. Operto, and J. Virieux. Multiparameter full waveform inversion of multicomponent ocean-bottom-cable data from the Valhall field. Part 1: Imaging compressional wave speed, density and attenuation. *Geophys. J. Int.*, 194(3):1640–1664, Sept. 2013. doi: 10.1093/gji/ggt177.

R. Quinlan. Auto mpg. *UCI Machine Learning Repository*, 1993. doi: 10.24432/C5859H.

A. E. Raftery and S. Lewis. How many iterations in the Gibbs sampler? *Bayesian Stat.*, 4:763–773, Apr. 1991.

R. Rasmussen and L. B. Pedersen. End Corrections in Potential Field Modeling. *Geophys. Prospect.*, 27(4):749–760, Dec. 1979. doi: 10.1111/j.1365-2478.1979.tb00994.x.

N. Rawlinson and M. Sambridge. Wave Front Evolution in Strongly Heterogeneous Layered Media Using the Fast Marching Method. *Geophys. J. Int.*, 156(3):631–647, Mar. 2004. doi: 10.1111/j.1365-246X.2004.02153.x.

M. Rietmann, P. Messmer, T. Nissen-Meyer, D. Peter, P. Basini, D. Komatitsch, O. Schenk, J. Tromp, L. Boschi, and D. Giardini. Forward and adjoint simulations of seismic wave propagation on emerging large-scale GPU architectures. In *SC'12: Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis*, pages 1–11. IEEE, 2012.

J. Ritsema, A. Deuss, H. J. van Heijst, and J. H. Woodhouse. S40RTS: a degree-40 shear-velocity model for the mantle from new Rayleigh wave dispersion, teleseismic traveltime and normal-mode splitting function measurements. *Geophys. J. Int.*, 184(3): 1223–1236, Mar. 2011. doi: 10.1111/j.1365-246X.2010.04884.x.

H. Robbins and S. Monro. A stochastic approximation method. *Ann. Math. Stat.*, pages 400–407, Sept. 1951. doi: 10.1214/aoms/1177729586.

M. Sambridge. Geophysical Inversion with a Neighbourhood Algorithm—I. Searching a Parameter Space. *Geophys. J. Int.*, 138(2):479–494, Aug. 1999. doi: 10.1046/j.1365-246X.1999.00876.x.

M. Sambridge. A Parallel Tempering Algorithm for Probabilistic Sampling and Multimodal Optimization. *Geophys. J. Int.*, 196(1):357–374, Jan. 2014. doi: 10.1093/gji/ggt342.

M. Sambridge, K. Gallagher, A. Jackson, and P. Rickwood. Trans-dimensional inverse problems, model comparison and the evidence. *Geophys. J. Int.*, 167(2):528–542, Nov. 2006. doi: 10.1111/j.1365-246X.2006.03155.x.

M. Sambridge, P. Rickwood, N. Rawlinson, and S. Sommacal. Automatic Differentiation in Geophysical Inverse Problems. *Geophys. J. Int.*, 170(1):1–8, July 2007. doi: 10.1111/j.1365-246X.2007.03400.x.

M. S. Sambridge and K. Mosegaard. Monte Carlo methods in geophysical inverse problems. *Rev. Geophys.*, 40(3):3–1–3–29, 2002. doi: 10.1029/2000RG000089.

M. S. Sambridge, T. Bodin, K. Gallagher, and H. Tkalcic. Transdimensional inference in the geosciences. *Phil. Trans. R. Soc. A*, 371(1984), Feb. 2013. doi: 10.1098/rsta.2011.0547.

J. Scales and L. Tenorio. Prior Information and Uncertainty in Inverse Problems. *Geophysics*, 66(2):389–397, Mar. 2001. doi: 10.1190/1.1444930.

J. A. Scales and R. Snieder. To Bayes or not to Bayes. *Geophysics*, 62(4):1045–1046, July 1997. doi: 10.1190/1.6241045.1.

M. K. Sen and R. Biswas. Transdimensional seismic inversion using the reversible jump Hamiltonian Monte Carlo algorithm. *Geophysics*, 82(3):R119–R134, May 2017. doi: 10.1190/geo2016-0010.1.

J. A. Sethian. A Fast Marching Level Set Method for Monotonically Advancing Fronts. *Proceedings of the National Academy of Sciences*, 93(4):1591–1595, Feb. 1996. doi: 10.1073/pnas.93.4.1591.

M. C. Shewry and H. P. Wynn. Maximum entropy sampling. *J. Appl. Stat.*, 14(2):165–170, 1987. doi: 10.1080/02664768700000020.

K. Sigloch, N. McQuarrie, and G. Nolet. Two-stage subduction history under North America inferred from multiple-frequency tomography. *Nat. Geosc.*, 1:458–462, June 2008. doi: 10.1038/ngeo231.

B. W. Silverman. *Density estimation for statistics and data analysis*, volume 26. CRC press, 1986.

J. C. Simo, N. Tarnow, and K. K. Wong. Exact Energy-Momentum Conserving Algorithms and Symplectic Schemes for Nonlinear Dynamics. *Comput. Methods Appl. Mech. Eng.*, 100(1):63–116, Oct. 1992. doi: 10.1016/0045-7825(92)90115-Z.

L. Sirgue, O. I. Barkved, J. Dellinger, J. Etgen, U. Albertin, and J. H. Kommedal. Full-waveform inversion: the next leap forward in imaging at Valhall. *First Break*, 28:65–70, 2010. doi: 10.3997/1365-2397.2010012.

W. Spakman, S. v. d. Lee, and R. v. d. Hilst. Travel-time tomography of the European-Mediterranean mantle down to 1400 km. *Phys. Earth Planet. Int.*, 79(1-2):3–74, Aug. 1993. doi: 10.1016/0031-9201(93)90142-V.

P. L. Stoffa and M. K. Sen. Nonlinear Multiparameter Optimization Using Genetic Algorithms: Inversion of Plane-wave Seismograms. *Geophysics*, 56(11):1794–1810, Nov. 1991. doi: 10.1190/1.1442992.

C. Taillandier, M. Noble, H. Chauris, and H. Calandra. First-Arrival Traveltime Tomography Based on the Adjoint-State Method. *Geophysics*, 74(6):WCB1–WCB10, Nov. 2009. doi: 10.1190/1.3250266.

O. Talagrand and P. Courtier. Variational Assimilation of Meteorological Observations With the Adjoint Vorticity Equation. I: Theory. *Q. J. R. Meteorolog. Soc.*, 113(478): 1311–1328, Oct. 1987. doi: 10.1002/qj.49711347812.

C. Tape, Q. Liu, A. Maggi, and J. Tromp. Seismic tomography of the southern California crust based upon spectral-element and adjoint methods. *Geophys. J. Int.*, 180(1):433–462, Jan. 2010. doi: 10.1111/j.1365-246X.2009.04429.x.

A. Tarantola. Inversion of Seismic Reflection Data in the Acoustic Approximation. *Geophysics*, 49(8):1259–1266, Aug. 1984. doi: 10.1190/1.1441754.

A. Tarantola. Theoretical background for the inversion of seismic waveforms, including elasticity and attenuation. *Pure Appl. Geophys.*, 128:365–399, Mar. 1988. doi: 10.1007/BF01772605.

A. Tarantola. *Inverse Problem Theory and Methods for Model Parameter Estimation*. Society for Industrial and Applied Mathematics, Jan. 2005. ISBN 978-0-89871-572-9 978-0-89871-792-1. doi: 10.1137/1.9780898717921.

A. Tarantola and B. Valette. Inverse Problems = Quest for Information. *J. Geophys.*, 50 (1):159–170, Oct. 1982.

S. Thrastarson, D.-P. van Herwaarden, L. Krischer, C. Boehm, M. van Driel, M. Afanasiev, and A. Fichtner. Data-adaptive global full-waveform inversion. *Geophys. J. Int.*, 230 (2):1374–1393, Mar. 2022. doi: 10.1093/gji/ggac122.

J. Thurin, R. Brossier, and L. Métivier. Ensemble-based uncertainty estimation in full waveform inversion. *Geophys. J. Int.*, 219(3):1613–1635, Aug. 2019. doi: 10.1093/gji/ggz384.

P. Toffanin and WebODM Authors. OpenDroneMap/WebODM: 2.0.3. *Zenodo*, May 2023. doi: 10.5281/zenodo.7951358.

E. Treister and E. Haber. A Fast Marching Algorithm for the Factored Eikonal Equation. *J. Comput. Phys.*, 324:210–225, Nov. 2016. doi: 10.1016/j.jcp.2016.08.012.

J. Tromp and E. Bachmann. Source encoding for adjoint tomography. *Geophys. J. Int.*, 218(3):2019–2044, June 2019. doi: 10.1093/gji/ggz271.

J. Tromp, C. Tape, and Q. Liu. Seismic Tomography, Adjoint Methods, Time Reversal and Banana-Doughnut Kernels. *Geophys. J. Int.*, 160(1):195–216, Jan. 2005. doi: 10.1111/j.1365-246X.2004.02453.x.

S. Ullman. The interpretation of structure from motion. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 203(1153):405–426, 1979.

G. Vacca and others. Overview of open source software for close range photogrammetry. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 42(4):239–245, Aug. 2019. doi: 10.5194/isprs-archives-XLII-4-W14-239-2019.

J. van Den Berg, A. Curtis, and J. Trampert. Optimal nonlinear Bayesian experimental design: an application to amplitude versus offset experiments. *Geophys. J. Int.*, 155(2): 411–421, Nov. 2003. doi: 10.1046/j.1365-246X.2003.02048.x.

D. P. van Herwaarden, C. Boehm, M. Afanasiev, S. Thrastarson, L. Krischer, J. Trampert, and A. Fichtner. Accelerated full-waveform inversion using dynamic mini-batches. *Geophys. J. Int.*, 221(2):1427–1438, Feb. 2020. doi: 10.1093/gji/ggaa079.

G. van Rossum. Python Tutorial. Technical Report CS-R9526, Centrum voor Wiskunde en Informatica (CWI), Amsterdam, 1995.

R. van Tent, A. F. Deuss, A. Fichtner, L. Gebraad, and J. Trampert. An Analysis of Normal-Mode Based 3-D Mantle Density Models using Hamiltonian Monte Carlo Methods. In *AGU Fall Meeting Abstracts*, volume 2020, Dec. 2020.

R. van Tent, L. Cobden, F. Deschamps, A. Fichtner, L. Gebraad, S. Schneider, J. Trampert, and A. Deuss. Reconciling 3-D mantle density models using recent normal-mode measurements and thermochemical convection modelling. In *AGU Fall Meeting Abstracts*, volume 2021, pages DI14A–05, Dec. 2021a.

R. van Tent, A. Deuss, A. Fichtner, L. Gebraad, S. Schneider, and J. Trampert. A new 3-D mantle density model from recent normal-mode measurements. In *EGU General Assembly Conference Abstracts*, pages EGU21–9852, May 2021b.

R. van Tent, A. Rijal, L. Gebraad, L. J. Cobden, L. Waszek, A. Fichtner, and A. Deuss. Global transition-zone topography and elastic properties from normal-mode tomography and their implications for compositional heterogeneity. In *AGU Fall Meeting Abstracts*, volume 2022, pages DI26A–03, Dec. 2022.

J. Virieux. P-SV wave propagation in heterogeneous media: velocity-stress finite difference method. *Geophysics*, 51(4):889–901, Apr. 1986. doi: 10.1190/1.1442147.

J. Virieux and S. Operto. An overview of full-waveform inversion in exploration geophysics. *Geophysics*, 74(6):WCC1–WCC26, Nov. 2009. doi: 10.1190/1.3238367.

P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, and P. van Mulbregt. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Meth.*, 17(3):261, Feb. 2020. doi: 10.1038/s41592-019-0686-2.

G. Visser, P. Guo, and E. Saygin. Bayesian Transdimensional Seismic Full Waveform Inversion with a Dipping Layer Parameterization. *Geophysics*, 84(6):R845–R858, Nov. 2019. doi: 10.1190/geo2018-0785.1.

M. Warner, A. Ratcliffe, T. Nangoo, J. Morgan, A. Umpleby, N. Shah, V. Vinje, I. Stekl, L. Guasch, C. Win, G. Conroy, and A. Betrand. Anisotropic 3D full-waveform inversion. *Geophysics*, 78(2):R59–R80, Mar. 2013. doi: 10.1190/geo2012-0338.1.

D. Wolpert and W. Macready. No free lunch theorems for optimization. *IEEE Trans. Evol. Computat.*, 1(1):67–82, Apr. 1997. doi: 10.1109/4235.585893.

C. Wu. Towards linear-time incremental structure from motion. In *2013 International Conference on 3D Vision*, pages 127–134. IEEE, June 2013. doi: 10.1109/3DV.2013.25.

K. Yomogida. Fresnel zone inversion for lateral heterogeneities in the Earth. *Pure Appl. Geophys.*, 138:391–406, 1992. doi: 10.1007/BF00876879.

K. Yoshizawa and B. L. N. Kennett. Multi-mode surface wave tomography for the Australian region using a 3-stage approach incorporating finite-frequency effects. *J. Geophys. Res.*, 109(B2), Feb. 2004. doi: 10.1029/2002JB002254.

X. Zhang and A. Curtis. Bayesian full-waveform inversion with realistic priors. *Geophysics*, 86(5):A45–A49, Aug. 2021. doi: 10.1190/geo2021-0118.1.

X. Zhang, A. Lomas, M. Zhou, Y. Zheng, and A. Curtis. 3-d bayesian variational full waveform inversion. *Geophys. J. Int.*, 234(1):546–561, Feb. 2023a. doi: 10.1093/gji/ggad057.

Z. Zhang, J. C. Irving, F. J. Simons, and T. Alkhalifah. Seismic evidence for a 1000 km mantle discontinuity under the pacific. *Nat. Commun.*, 14(1):1714, 2023b.

A. Zunino and K. Mosegaard. Integrating Gradient Information with Probabilistic Traveltime Tomography Using the Hamiltonian Monte Carlo Algorithm. In *80th EAGE Conference & Exhibition 2018 Workshop Programme*. European Association of Geoscientists & Engineers, June 2018. ISBN 978-94-6282-257-3. doi: 10.3997/2214-4609.201801971.

A. Zunino and K. Mosegaard. An Efficient Method to Solve Large Linearizable Inverse Problems under Gaussian and Separability Assumptions. *Comput. Geosci.*, 122:77–86, Jan. 2019. doi: 10.1016/j.cageo.2018.09.005.

A. Zunino, K. Mosegaard, K. Lange, Y. Melnikova, and T. Mejer Hansen. Monte Carlo Reservoir Analysis Combining Seismic Reflection Data and Informed Priors. *Geophysics*, 80(1):31, Jan. 2015. doi: 10.1190/geo2014-0052.1.

A. Zunino, A. Ghirotto, E. Armadillo, and A. Fichtner. Hamiltonian Monte Carlo Probabilistic Joint Inversion of 2D (2.75D) Gravity and Magnetic Data. *Geophys. Res. Lett.*, 49(20):e2022GL099789, Oct. 2022. doi: 10.1029/2022GL099789.

A. Zunino, L. Gebraad, A. Ghirotto, and A. Fichtner. HMCLab: a framework for solving diverse geophysical inverse problems using the Hamiltonian Monte Carlo method. *Geophys. J. Int.*, 235(3):2979–2991, 10 2023. ISSN 0956-540X. doi: 10.1093/gji/ggad403. URL https://doi.org/10.1093/gji/ggad403.

O. Özyeşil, V. Voroninski, R. Basri, and A. Singer. A survey of structure from motion. *Acta Numer.*, 26:305–364, May 2017. doi: 10.1017/S096249291700006X.