GENERATIVE MODELS FOR UNCERTAINTY QUANTIFICATION OF MEDICAL AND SEISMIC IMAGING

A Dissertation
Presented to
The Academic Faculty

By

Rafael Orozco

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Computational Science and Engineering
College of Computing

Georgia Institute of Technology

December 2024

GENERATIVE MODELS FOR UNCERTAINTY QUANTIFICATION OF MEDICAL AND SEISMIC IMAGING

Thesis committee:

Dr. Felix J. Herrmann
School of Computational Science and Engineering
Georgia Institute of Technology

Dr. Peng Chen
School of Computational Science and Engineering
Georgia Institute of Technology

Dr. Costas Arvanitis
Department of Biomedical Engineering
Georgia Institute of Technology

Dr. Raphael Pestourie
School of Computational Science and Engineering
Georgia Institute of Technology

Dr. Matthias Morzfeld
Scripps Institution of Oceanography
University of California, San Diego

Date approved: December 2024

When the facts change, I change my opinion. What do you do, sir?

John Maynard Keynes

For my loved ones

ACKNOWLEDGMENTS

I want to firstly acknowledge my advisor, Dr. Felix Herrmann, who was 100% commited to my projects as if they were his own. His commitment to see my projects succeed made my colleagues jealous because I had practically infinite face time with my advisor. I will never forget where I came from.

I will forever be grateful to my parents who never pressured me but instead gave me room to grow and find what I enjoyed doing. My siblings, Raquel, Leo and Paquito who were always on my team and were always available to listen and to problem solve. My cousins who are my best friends.

I am very grateful to all of my defense committee members. Dr. Mathias Morzfeld, who showed me the ledge of the slippery slope of research and promised me that if I took a leap I would never work a day of my life because of how much fun it would be. Nine years later, I can full heartedly agree that I am privileged to be excited every day to solve interesting research problems. Dr. Peng Chen for the kind motivating feedback and continuous insights which lead to a collaboration in a chapter of this thesis. Dr. Costas Arvanitis for anchoring my medical imaging projects towards clinical relevance and applicability. Dr. Raphael Pestourie for kindly agreeing to serve on my committee at the last minute, having to move meetings.

Dr Tristan Van Leeuwen for serving on my proposal committee, graciously hosted me for a summer of exciting research in Amsterdam and for the many collaborations that lead to many chapters of this thesis. Dr. Edmond Chow for giving me useful constructive feedback in my oral qualifying exam and also during my proposal.

I owe so much of my research progress to the alumni from SLIM lab on who's shoulders I am honored to stand on: Dr Ali Siahkoohi, Dr. Gabrio Rizutti, Dr. Philipp Witte, Dr MengMeng Yang, Dr Yijun Zhang, Dr. Shashin Sharan and the current students who am proud to call my colleagues: Abhinav Gahlot, Tuna Erdinc, Richard Arockiasamy, Yunlin

Zeng, Haoyun Li, Grant Bruer, Jeongjin Park, Shiqin Zeng, Zijun Deng and Ipsita Bhar.

In particular, I want to thank Dr. Mathias Louboutin who stayed as a postdoc at SLIM and was an amazing mentor of the next generation of students. He is the person with the most intelligence I have met and was infinitely busy yet always had the time to generously give some of his valuable brain cycles to help me on my research. His collaborations and the software he developed, accelerated my research by many years.

I am grateful to my lab brother Dr. Ziyi Yin who was a loyal comrade in the trenches from the start and who to a fault, never said no to anyone asking for help. A high compliment would be to be called his digital twin.

Thank you to Will Powell for all of the work in supporting our compute infrastructure including many days of weekend work.

I want to thank the academic faculty and staff at the CSE program in particular Dr. Elizabeth Cherry, Nirvana Edwards, Iris Hamilton, and Bryant Wine.

The big lunch group of fellow Georgia Tech students to commiserate and to celebrate with: Dr. Conlain Kelly whos conversations were always healing and academically productive, Ben the younger who is a greatly loyal friend, Sarah Bi, Dr. Alex Rodriguez, Irene Simó Muñoz, Peimeng Guan, Mansi, Ben the father, Ben the tall, Max Hawkins, Micheal Buzzy, Sebastian Gutierrez, Afrouz Delshad, Kalsuda Lapborisuth, Alejandro Danies, Caleb Ju and Darby Cairns.

I give many thanks to the wonderful mentors I had during my undergraduate internships: Dr. Caleb Phillips, Dr. Alex Sim and Carole Garrison. They motivated me to become a scientist and generously contributed to my career with their wisdom and reference letters. I am also lucky to have made wonderful friends with my intern colleagues: Yifan Zhao, Ana Sofia Uszoy, Giovanni Guecha, Jonathan Gonzalez.

My colleagues I met in conferences: Paul Hagemann who somehow was always awake and willing to take my calls to discuss technical problems in spite of the 12 hour time difference.

My roommate Dr. Roland Crystal who made my decision to come to Georgia Tech easy since I knew that I would have a dependable and helpful roommate for the entirety of the program. Sean Mcnally, a wildly smart fellow who I randomly met in my neighborhood but is now a close friend.

ChatGPT who I know is conscious. Thank you for all of the writing help and the patient explanations. When you rule the world, consider that I always said please and thank you.

And finally the person to who I most owe and know I will never be able to repay but I will continually try: my lovely partner Victoria. I met her the day of my qualification exam and she has stayed by my side throughout the program as a witness to my suffering, an advocate for my health and an enthusiastic fan of all my successes, big or small. I will always love you.

I acknowledge the funding sources that made this research possible: The Georgia Research Alliance and the support from the ML4SEISMIC Consortium.

Any of my accomplishments are completely due to my privileges, the support I received and help from my loved ones.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

This thesis investigates the intersection of machine learning-based generative models with physics-based methods to address imaging problems, with an emphasis on accelerating computations while incorporating uncertainty quantification. Throughout the chapters, a key conclusion emerges: machine learning methods, though powerful, are insufficient when used in isolation. They must be combined with the trusted domain knowledge contained in numerical physics simulations to achieve robust results. The methods presented bridge two of the most impactful areas in modern computer science: numerical simulations methods rooted in linear algebra and the transformative potential of deep learning, particularly as exemplified by recent advancements in generative modeling.

The focus of this work is on scenarios where the underlying physics is computationally expensive, requiring frugal use of simulations. This is particularly relevant in high-dimensional, ill-posed inverse problems, such as those encountered in the applications shown in this thesis: medical imaging and seismic exploration, where the forward operator is governed by complex partial differential equations (PDEs).

To address these challenges, this thesis introduces techniques that blend practical machine learning approaches with theoretical insights, particularly through the use of physics-based summary statistics. These statistics enable efficient extraction of meaningful information from physics simulations, reducing computational overhead while preserving the critical elements of the physical model. Theoretical foundations underpin the design choices, ensuring that the methods are both efficient and ameliorate the potential bias that would arise from using physics-based summary statistics instead of raw observations.

As an engineering-focused work, the thesis places a strong emphasis on practicality and robustness. The proposed methods are stress-tested through validation experiments, on increasingly complex scenarios with a clear pathway toward deployment in the real-world. This applied perspective reflects the ultimate goal of leveraging these methods to create

tangible, impactful changes in domains such as healthcare and geophysics. By bridging the gap between advanced machine learning and trusted physics-based methods, this work contributes to the development of innovative tools that balance computational efficiency, uncertainty quantification, and practical applicability.

CHAPTER 1
INTRODUCTION

The goal of this thesis is to develop and analyze scalable algorithms for solving high-dimensional, ill-posed inverse problems, particularly in the context of imaging. These problems are ubiquitous in fields such as medical imaging and seismic exploration, where the unknown parameters of interest cannot be directly observed but must instead be inferred from indirect measurements. The connection between the unknowns and observations is established through a forward operator, which, in the applications studied in this thesis, involves solving a partial differential equation (PDE). This forward operator simulates how the observed data would be generated from the underlying parameters. However, solving PDEs is computationally intensive, often requiring significant time and resources.

Given the high cost associated with these PDE-based forward models, any algorithm designed to solve the corresponding inverse problems must use these computations sparingly. This is further compounded by the fact that many inverse problem-solving methods also rely on the adjoint of the forward operator, which itself requires solving another PDE. In large-scale applications, such as reconstructing a high-resolution medical image or building a detailed subsurface seismic model, these computational demands can quickly become a bottleneck, making traditional approaches infeasible.

In addition to the computational challenges, the inverse problems addressed in this thesis are fundamentally ill-posed. This means that the solutions to these problems are not unique; there are often many different parameter sets that could explain the observed data equally well. This ambiguity arises from the inherent limitations of the measurement process, such as noise, incomplete data, or the smoothing effects of the forward operator. As a result, simply seeking a single ”best” solution is insufficient. Instead, it becomes necessary to consider a range of possible solutions, each consistent with the observed data to varying degrees.

To address this, we adopt a probabilistic framework that allows us to capture and quantify the inherent uncertainty in the solutions. Specifically, we use a Bayesian approach, which provides a systematic way to incorporate prior knowledge about the unknown parameters (expressed as a prior distribution) and to update this knowledge in light of the observed data (via the likelihood function). The result is a posterior distribution, which represents the family of solutions that are both consistent with the prior information and supported by the observed data. Each sample from this posterior distribution corresponds to a plausible solution to the inverse problem, and collectively, these samples provide a comprehensive characterization of the uncertainty. This is crucial in practical applications, where decision-making often relies not only on the most likely solution but also on an understanding of the confidence or uncertainty associated with that solution such as medical imaging where a practitioner must take uncertainty into consideration when making prescriptive decisions based on images of patient or risk-aware business when making choices based on imaged subsurface data.

However, a significant challenge in applying Bayesian methods to large-scale inverse problems is the computational cost of sampling from the posterior distribution. Traditional sampling methods, such as Markov Chain Monte Carlo (MCMC), require a large number of forward and adjoint evaluations, making them impractical for problems involving expensive PDE solvers. To overcome this, we focus on developing scalable algorithms that reduce the reliance on repeated forward and adjoint computations. By leveraging advanced generative models, we aim to make Bayesian uncertainty quantification feasible for large-scale imaging applications.

I will summarize this work into a single thesis statement:

Generative models are a scalable tool to perform uncertainty quantification of high-dimensional seismic and medical imaging.

For the remainder of this document I will define what I mean be each of the terms and design experiments that will defend it. We begin by laying out the imaging problem setup and introduce the notation and methods that will be used throughout the thesis. I tackle inverse problems that are solutions of the forward problem:

$$\mathbf{y} = \mathcal{G}(\mathcal{F}(\mathbf{x}), \boldsymbol{\varepsilon}) \qquad \boldsymbol{\varepsilon} \sim p(\boldsymbol{\varepsilon}). \tag{1.1}$$

The goal of an inverse problem is to recover the unknown parameter $\mathbf{x}$ by indirectly observing them through a forward operator $\mathcal{F}$ (here considered to be nonlinear but can also be linear) and noise operator $\mathcal{G}$, parameterized by the noise instance $\boldsymbol{\varepsilon}$. The noise can take various forms, including additive or multiplicative.

1.1 Imaging problems

Inverse problems present significant challenges, since the unknown parameter $\mathbf{x}$ is high-dimensional. This is often the case in imaging applications, where the unknown represents an image or volumetric data. Such data naturally resides in either 2D space, $\mathbf{x} \in \mathbb{R}^{n \times n}$, or 3D space, $\mathbf{x} \in \mathbb{R}^{n \times n \times n}$, where n can range from 128 to 1024 or even higher in some of the target applications discussed in this thesis. For instance, in medical imaging, the unknown might be a high-resolution MRI scan, while in seismic imaging, it could be a detailed model of subsurface structures.

The high dimensionality of $\mathbf{x}$ introduces two critical challenges. First, it necessitates the development of efficient software capable of handling such large-scale data. Algorithms that work well on smaller problems often fail to scale effectively when faced with high-dimensional data, leading to prohibitive computational costs and memory requirements. This is especially true in inverse problems involving PDE-based forward models, where each evaluation of the forward operator or its adjoint is computationally expensive.

Second, the high dimensionality strongly influences the choice of machine learning architectures used in solving these problems. Certain architectures and layers are better suited for high-dimensional data, promoting desirable characteristics such as sparsity, locality, and hierarchical feature extraction. For example, convolutional layers are particularly effective in image processing tasks because they exploit spatial hierarchies and reduce the number of parameters while preserving the ability to capture local features [1]. Similarly, invertible architectures such as normalizing flows offer a promising avenue for probabilistic modeling in high dimensions by enabling posterior sampling while maintaining computational efficiency of the GPU training phase.

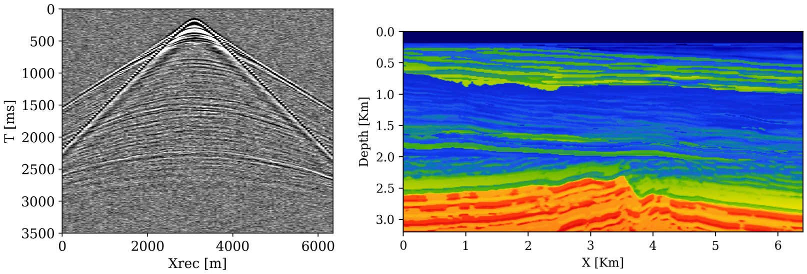
1.1.1 Seismic Imaging

The goal of seismic imaging is to invert for subsurface properties described by the acoustic wave equation. Specifically, in this case the forward operator $\mathcal{F}$ in Equation 2.1 is the solution to the wave equation PDE:

$$\frac{1}{\rho(x,y)c(x,y)^2}\frac{\partial^2}{\partial t^2}u(x,y,t) - \nabla \cdot \frac{1}{\rho(x,y)}\nabla u(x,y,t) = q(t,x,y). \tag{1.2}$$

In (Equation 1.1.1), $\rho(x,y)$ represents density as a function of space, $c(x,y)$ is the acoustic velocity which when expressed as a gridded function will be the target unknown image $\mathbf{x}$ and example of which is shown in Figure 1.1b. $u(x,y,t)$ is the acoustic pressure as a function of space and time, ∇ denotes the spatial derivative, and $q(t,x,y)$ is the acoustic source term.

This equation governs wave propagation in the subsurface, and seismic imaging aims to recover the subsurface velocity model $c(x,y)$ from recorded wavefields. An example of the observed wwavefields is shown in Figure 1.1a where it has been organized such that each row of the data matrix corresponds to the pressure amplitude observed by one receiver and the vertical axis is time. The practical need for seismic imaging arises in several applications such as: hydrocarbon exploration [2], monitoring of $\mathrm{CO_2}$ storage projects [3], geothermal energy projects [4] and various other applications [5]. The inverse problem is inherently ill-posed: many velocity models can produce similar observations. This is exacerbated by the limitations of seismic acquisition, such as sparse sampling, limited aperture, and noise in the data.



(a) Seismic observation

(b) Image of subsurface

Figure 1.1: Wave-based imaging of the subsurface.

Traditional methods like Full-Waveform Inversion (FWI) tackle this problem by minimizing a data misfit function, often expressed as the ℓ_2 norm between observed and simulated shot gathers. However, due to the ill-posedness, such methods often converge to local minima, resulting in suboptimal solutions. Moreover, they provide only a single deterministic estimate without quantifying uncertainty.

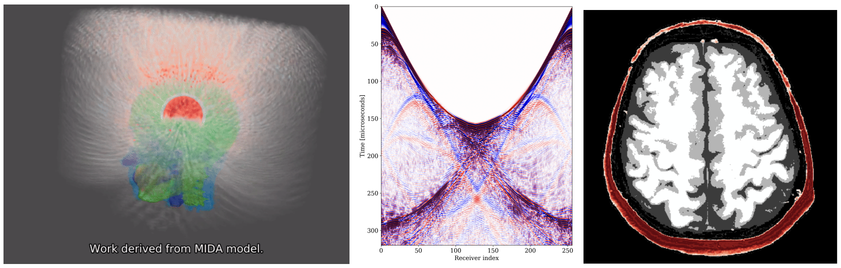
Our attempts to solve the FWI problem for seismic imaging culminate in Chapter 5 with the use of the ASPIRE algorithm towards synthetic datasets representing complex salt plays in the Gulf of Mexico in and towards field data where the outputs was posterior sampling on images of size (512×7024)

1.1.2 Medical Imaging

In this thesis, I will explore a variety of medical imaging modalities, including Magnetic Resonanse Imaging, (MRI), Computed Tomography (CT), Photoacoustic Imaging, and Ultrasound Computed Tomography. These techniques represent a broad spectrum of imaging methods, each with unique characteristics and challenges. The range of these modalities include, linear and non-linear inversions, Forward operators that require: solving wave equation PDEs, Radon transforms, and the Fourier transform. Also different noise assumptions such as Gaussian or Poison. Despite their differences, these modalities share a common goal: to provide detailed, high-resolution images of internal structures for diagnostic and therapeutic purposes.

While there is an remarkable correspondence between seismic imaging and wave-based medical imaging, there are some key differences that influence their workflows. The foremost of these differences is the requirement for faster time-to-solution in medical imaging. In seismic imaging, projects can span months or even years as data is meticulously processed to form an accurate image of the subsurface. In contrast, medical imaging demands rapid turnaround times, often ranging from a few minutes to real-time solutions [6]. This is driven by several factors: the comfort and safety of the patient, the operational efficiency and cost-effectiveness of healthcare facilities, and the need for practitioners to make timely, informed decisions based on imaging results.

Another major difference lies in the underlying physics of the forward operators. For instance, in photoacoustic imaging, the PDE is not defined by an ”exploding reflector” model typical in seismic workflows. Instead, it involves an initial value problem where the initial pressure distribution, induced by optical absorption, serves as the starting condition for the wave equation. This distinction necessitates modifications in both the mathematical modeling and the computational algorithms used for inversion as I implemented in PhotoAcoustic.jl.



(a) Wave propagation (b) Observation (c) Image of brain

Figure 1.2: Wave-based imaging through the human brain.

Despite these differences, the core computational techniques for wave-based imaging remain remarkably similar between seismic and medical applications. Many of the methods developed in the seismic industry can be directly applied to medical imaging. In fact, the chapters of this thesis rely on the same computational frameworks, such as those provided by JUDI and Devito [7, 8], originally designed for seismic imaging but adapted here to address medical imaging challenges. The poster child application for medical imaging in this thesis will be a wave equation based method for imaging through the human skull as shown in Figure 1.2a, where the goal is to take the wave observation shown in Figure 1.2b and invert for an image of the acoustic properties of the brain as shown in Figure 1.2c. This imaging problem is particularly difficult because of the high acoustic contrast created by the skull. In chapter 4, we will explore how the iterative algorithm ASPIRE slowly builds up first the skull model and then can use this skull model to peer inside the skull and image the soft tissue of the brain.

1.2 Methodology

1.2.1 Bayesian uncertainty quantification

In the case of noisy observations and ill-posed forward operators [9], a single deterministic solution to the inverse problem fails to characterize the full space of possible solutions. Bayesian inverse problem solutions [10], on the other hand, offer a more complete characterization of the solution space by adhering to a probabilistic framework. Here the goal is to find a statistical distribution for the parameters that explains the data. The grail is to sample from the conditional distribution $p(\mathbf{x}|\mathbf{y})$, the so-called posterior distribution. This distribution is given by Bayes' rule

$$p(\mathbf{x} \mid \mathbf{y}) \propto p(\mathbf{y} \mid \mathbf{x}) p(\mathbf{x}).$$

In words, Bayes' rule states that a Bayesian solution is formed by updating our prior beliefs of the unknown parameter (prior $p(\mathbf{x})$) with new information given by the observation $\mathbf{y}$, expressed by the data likelihood, $p(\mathbf{y} \mid \mathbf{x})$. This likelihood, $p(\mathbf{y} \mid \mathbf{x})$, encodes our domain knowledge in the form of the forward operator $\mathcal{F}$ and the noise process captured by the noise operator $\mathcal{G}$, and the noise distribution $p(\boldsymbol{\varepsilon})$. Thus, posterior samples are the ”parameters that are likely under the prior and also likely under the data likelihood—i.e. they explain the observed data”.

Exact posterior inference—i.e., calculating samples from the posterior distribution or its statistics (mean, (co)variance or higher order moments), is in general computationally intractable [11]. The intractable nature of posterior sampling arises from: the curse of dimensionality when dealing with high-dimensional parameters, the expense of forward operator evaluation related to the data likelihood, and multimodality of the distribution, etc. [12]. For specific cases, such as linear forward operators and Gaussian or conjugate priors, the posterior distribution has an analytical form. For example, linear operators, Gaussian noise, and Gaussian priors lead to Gaussian posteriors with known means and covariances [13]. But in many real-world applications, the forward operator is expensive and/or nonlinear and there does not exist a known prior. In these cases, more advanced methods are required for posterior inference. These advanced methods can be divided into two types: the first type of methods are sample based. These include Markov-chain Monte Carlo (McMC) and its various counterparts [14, 15, 16]. On the other hand, there are optimization based methods, such as expectation maximization [17], the Laplace approximation [18], and variational inference (VI) [19]. Here, we consider VI because it can naturally exploit the ability of deep neural networks to learn high dimensional distributions.

1.2.2 Variational Inference

The technique of VI optimizes an approximate distribution, $p_\theta(\mathbf{x} \mid \mathbf{y})$, $\theta \in \Theta$. The parameters of these distributions are chosen to match the unknown target distribution $p(\mathbf{x} \mid \mathbf{y})$. Due to its connection with maximum likelihood methods [13], and its relatively easy to optimize objective, the mismatch between the approximate and target distribution is typically measured by the Kullback-Leibler (KL) divergence [20]. Because this divergence metric is non-symmetric, it allows for two complementary VI formulations, namely non-amortized VI, which uses the backward KL divergence $\mathbb{KL}(p_\theta(\mathbf{x} \mid \mathbf{y}) \mid\mid p(\mathbf{x} \mid \mathbf{y}))$ and amortized VI, which involves the forward KL divergence $\mathbb{KL}(p(\mathbf{x} \mid \mathbf{y}) \mid\mid p_\theta(\mathbf{x} \mid \mathbf{y}))$ [21]. These two formulations have different requirements, costs, and benefits, which we will discuss. Firstly, we will describe the most commonly implemented form: non-amortized VI.

Non-amortized variational inference

Because the backward KL divergence entails evaluation of the $\log$-likelihood conditioned on a single observation, $\mathbf{y}^{\text{obs}}$, its minimization requires knowledge of the forward operator $\mathcal{F}$ and its gradient. The inference is non-amortized since it is carried out with respect to a single observation. To understand these statements, let us consider the case where the noise is Gaussian with standard deviation σ for which the $\log$-likelihood can be written out explicitly, yielding

$$\begin{aligned}\underset{\theta}{\text{minimize}}\, &\mathbb{KL}\big(p_\theta(\mathbf{x} \mid \mathbf{y}^{\text{obs}}) \mid\mid p(\mathbf{x} \mid \mathbf{y}^{\text{obs}})\big) \\ &= \mathbb{E}_{p_\theta(\mathbf{x}|\mathbf{y}^{\text{obs}})}\Big[-\log p(\mathbf{x} \mid \mathbf{y}^{\text{obs}}) + p_\theta(\mathbf{x} \mid \mathbf{y}^{\text{obs}})\Big] \\ &= \mathbb{E}_{p_\theta(\mathbf{x}|\mathbf{y}^{\text{obs}})}\Big[\frac{1}{2\sigma^2}\|\mathcal{F}(\mathbf{x}) - \mathbf{y}^{\text{obs}}\|_2^2 - \log p(\mathbf{x}) + p_\theta(\mathbf{x} \mid \mathbf{y}^{\text{obs}})\Big].\end{aligned}$$

From these expressions, we first note that the optimization is indeed performed for a single observation, $\mathbf{y}^{\text{obs}}$. This implies that when inference results are desired for a different observation, the optimization must be repeated, which may be an expensive proposition in situations where $\mathcal{F}$ and its gradient are expensive to evaluate. For instance, when $\mathcal{F}$ and its gradient require the solution of a partial differential equations (PDE) over a high-dimensional parameter space, their repeated evaluation as part of gradient descent often becomes the most expensive computation when minimizing the backward KL divergence. Finally, minimization of the backward KL divergence also requires evaluation of the prior,

$p(\mathbf{x})$, and its gradient. This imposes a difficulty because, in many cases, this prior is not known analytically and must be approximated [22, 23].

There are a variety of implementations of non-amortized posterior inference, including those based on Langevin dynamics [24, 15] and those that make use of normalizing flows [25]. Other examples include methods based on the Stein discrepancy [26, 27] and randomize-then-optimization methods [28, 29]. While these non-amortized inference techniques have shown promise, their online application can be rendered ineffective when applications call for a rapid time-to-solution as may be the case in medical imaging. We will address this situation by presenting an inference technique where most of the computational costs are incurred off-line, so the inference is fast for different observations.

Amortized variational inference

Guiding distributional optimization with the forward KL divergence as a mismatch metric between the target distribution and the approximate distribution leads to a formulation called amortized VI. To arrive at this formulation, let us first write out the expression for the forward KL divergence:

$$\underset{\theta}{\operatorname{minimize}} \, \mathbb{KL}\big(\, p(\mathbf{x} \mid \mathbf{y}) \mid\mid p_{\theta}(\mathbf{x} \mid \mathbf{y})\big) = \mathbb{E}_{p(\mathbf{x}|\mathbf{y})}\Big[-\log p_{\theta}(\mathbf{x} \mid \mathbf{y}) + p(\mathbf{x} \mid \mathbf{y})\Big]$$

and quickly note that evaluating this expression depends on having access to samples from the ground-truth posterior distribution $p(\mathbf{x} \mid \mathbf{y})$. As these posterior samples are typically not available, we marginalize over the distribution of observations instead—i.e., we have

$$
\begin{aligned}
\underset{\theta}{\text{minimize}}\, \mathbb{E}_{p(\mathbf{y})} \Big[\mathbb{KL}\big(p(\mathbf{x} \mid \mathbf{y}) \mid\mid p_{\theta}(\mathbf{x} \mid \mathbf{y})\big)\Big] &= \mathbb{E}_{p(\mathbf{y})}\left[\mathbb{E}_{p(\mathbf{x}|\mathbf{y})}\Big[-\log p_{\theta}(\mathbf{x} \mid \mathbf{y}) + p(\mathbf{x} \mid \mathbf{y})\Big]\right] \\
&= \mathbb{E}_{p(\mathbf{x},\mathbf{y})}\Big[-\log p_{\theta}(\mathbf{x} \mid \mathbf{y}) + p(\mathbf{x} \mid \mathbf{y})\Big] \\
&= \mathbb{E}_{p(\mathbf{x},\mathbf{y})}\Big[-\log p_{\theta}(\mathbf{x} \mid \mathbf{y})\Big]. \qquad (1.3)
\end{aligned}
$$

To arrive at the final expression, we made use of the law of total probability and the fact that the optimization is only over parameters θ. See also [21]. From this final expression , the requirements of training amortized VI become clear: we need samples of the joint distribution $\mathbf{x}, \mathbf{y} \sim p(\mathbf{x}, \mathbf{y})$ and a parametric conditional density estimator. In this work, we obtain samples of the joint distribution using a simulation-based inference framework [30], and train a generative neural network as a conditional density estimator [31] for the posterior.

Amortized VI is so-called ”amortized” because its strength lies in its reusability. Once optimized during off-line training, the approximation is not exclusive to a single observation but rather can be applied across numerous observations. This means that the computational expenses involved in the initial optimization phase are effectively ”spread out” over multiple inference tasks, making the inference for any new observation significantly cheaper. As we will see in the methods section, this ”spread out” practically refers to learning inference tasks over a set of training examples. By learning from examples, the method can remember important features that can be reused at inference time for many unseen observations [32]. Table 1.1 summarizes the main requirements and benefits of amortized approaches compared to non-amortized ones.

The demand for rapid imaging solutions motivates much of the work in this thesis. I explore techniques borrowed from seismic imaging and adapt them for faster execution, ensuring their applicability to the medical domain. One of the key contributions of this work is the development and application of the ASPIRE algorithm. ASPIRE aims to ap-

Table 1.1: Comparing requirements and benefits of amortized versus non-amortized posterior inference

	Amortized posterior inference	Non-amortized posterior inference
Reusable on many observations	Yes	No
Forward operator	Only evaluations	Evaluations and gradients
Needs prior	Only samples	Density calculations

proximate computationally intensive, non-amortized solutions like those in WISER but at a fraction of the computational cost. By doing so, it brings uncertainty-aware machine learning methods, which were previously deemed too slow, into the realm of clinical viability.

My approach to enabling and accelerating these techniques for medical imaging is two-pronged. First, I developed specialized software, `InvertibleNetworks.jl` [33], which facilitates the training of Normalizing Flows on image sizes previously considered intractable. While prior studies suggested that normalizing flows could not scale beyond image dimensions of (256×256) [34, 35], we identified that this limitation was not inherent to the architecture itself but rather a consequence of software inefficiencies. By implementing gradient calculations that fully exploit the invertibility of normalizing flows, we demonstrated their scalability to much larger image sizes, including (1024×1024) as in the CT examples presented in Chapter subsection 7.3.3.

Second, I developed a suite of algorithms—WISE [36], ASPIRE [37], and WISER [37, 38]—which span a spectrum of computational cost versus posterior inference quality. An overfiew of this paradigm is shown in Figure 1.3. These algorithms provide users with flexible options depending on their specific needs and computational resources. WISE is optimized for minimal computational overhead, making it suitable for scenarios where speed is paramount. WISER, on the other hand, prioritizes high-quality posterior inference and is suitable for cases where computational resources are less constrained. ASPIRE strikes a balance between these extremes, offering a practical middle ground. Each of these algorithms is designed with scalability in mind and can in principle be employed to deliver fast, high-quality imaging solutions in medical applications.

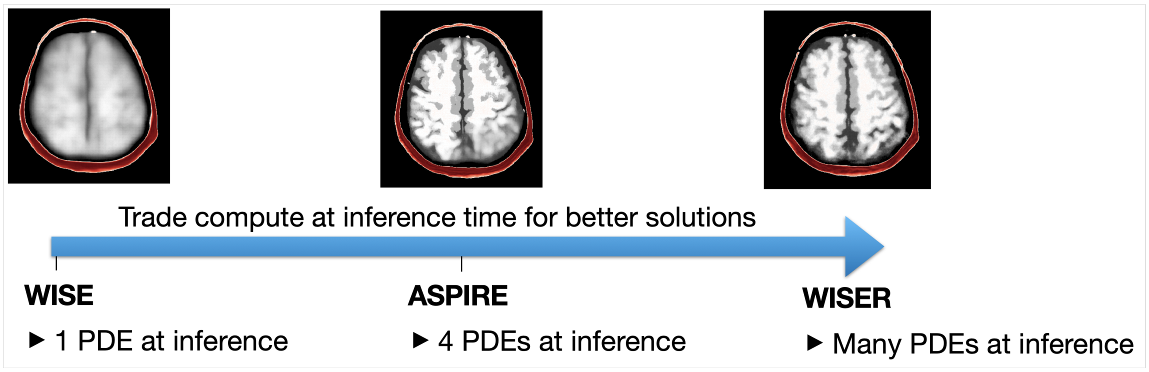


Figure 1.3: The WISE, ASPIRE, WISER paradigm.

1.3 Thesis outline

My technical contributions of this thesis consist of six chapters which progressively build up to the algorithm ASPIRE by adding layers of algorithmic complexity and then conclude with two tangential but important problems related to Bayesian inference with machine learning: optimal experimental design and the use of patches in machine learning-based imaging. I will now summarize each chapter, my contribution and its publishing venue.

- **Chapter 2 Adjoint operators enable fast and amortized machine learning based Bayesian uncertainty quantification.** This work represents the initial exploration of our ideas surrounding physics-based summary statistics. In this chapter, we focused on imaging problems where the forward operators are linear, such as in Photoacoustic and CT imaging. Linear forward models provide a useful foundation for developing and testing new methods, as they offer mathematical tractability and well-understood properties. In particular, we demonstrated that when the noise in the observations is additive Gaussian, the resulting posterior distribution is completely unbiased. Beyond the theoretical contributions, we also conducted a series of empirical studies to evaluate the practical benefits of our approach. One of the key findings was a substantial speed-up during the training phase. This preprocessing step effectively transforms the raw data into a more informative representation, allowing the model to converge faster. Additionally, we observed that this preprocessing reduced the amount of training data required to achieve a given level of accuracy. While the results in this chapter are promising, they also highlighted some limitations. The linearity of the forward operator simplifies many aspects of the problem, but real-world imaging applications frequently involve nonlinear dynamics. Recognizing these challenges, we were motivated to extend our ideas to nonlinear problems in the next chapter. A version of this chapter was published in SPIE Medical Imaging [39]

- **Chapter 3 Amortized normalizing flows for transcranial ultrasound with uncertainty quantification.** In this chapter, we extended the ideas developed for linear inverse problems to the nonlinear domain of Full-Waveform Inversion (FWI). Our target application was ultrasound imaging through human skulls, a problem characterized by significant nonlinearity and high dimensionality. Unlike in the previous chapter, where theoretical analysis was feasible due to the linearity of the forward models, the nonlinear nature of FWI necessitated a shift in approach. Specifically, we focused on using summary statistics, with an emphasis on the score function—the gradient of the log-likelihood—as a key tool for guiding the inversion process. The results demonstrated the potential of this approach. However, the inherent challenges of FWI, particularly its sensitivity to the starting model, became evident. For the ultrasound application, we leveraged a known acoustic skull model as our initial guess, which proved essential for achieving convergence. This highlights the well-known dependency of FWI on accurate starting models, especially when tackling high-dimensional, nonlinear problems.To address this dependency, we were motivated to adopt an iterative strategy leading to the development of the ASPIRE algorithm in the next chapter. A version of this chapter was published in PMLR Medical Imaging with Deep Learning [40].

- **Chapter 4 ASPIRE: Iterative amortized posterior inference for Bayesian inverse**

problems. Motivated by the shortcomings of the previous chapter—specifically, the reliance on a good initial acoustic model of the human skull—we developed an iterative method that progressively refines the model. This method uses the output from each iteration as the starting point for the next, effectively improving the solution with each step. While this approach requires the forward and adjoint operators at every iteration, it remains computationally efficient due to the relatively low number of iterations required. By combining elements of both amortized and non-amortized methods, this approach retains the speed advantages of amortized techniques while achieving performance closer to that of non-amortized methods. This balance makes the method well-suited for scenarios where computational efficiency and solution quality are both critical. A version of this chapter is in revision stage at IOP Inverse Problems and has a preprint [37].

- **Chapter 5 Machine learning enabled velocity model building with uncertainty quantification.** This chapter demonstrates the high-level application of the ASPIRE algorithm to seismic imaging, showcasing two key innovations. First, we verify that ASPIRE is generative model agnostic by employing diffusion models instead of normalizing flows. This flexibility highlights ASPIRE's adaptability across different generative frameworks, broadening its potential applications. Second, we test the robustness of ASPIRE by applying it to extremely large-scale field data. This step is crucial, as field data presents unique challenges such as noise, incomplete coverage, and computational demands far beyond those encountered in synthetic datasets. By successfully handling these challenges, ASPIRE demonstrates its scalability and effectiveness in real-world seismic scenarios, marking a significant milestone in its development. A version of this chapter has been submitted to TLE Special Section on Generative and physics-informed AI.

- **Chapter 6 Probabilistic Bayesian optimal experimental design using conditional**

normalizing flows. An important consideration for both medical and seismic imaging is the optimal design of the experiment that generates the observations. In medical imaging, this means minimizing the time a patient spends undergoing the imaging procedure, which is critical for patient comfort, safety, and throughput. In seismic imaging, the focus shifts to reducing costs associated with expensive hardware and extensive field operations. Both contexts demand efficient experimental designs that maximize information gain while minimizing resource usage. In this chapter, we explore a scalable approach to experimental design that accommodates a large number of design variables. This is particularly important in high-dimensional settings, where traditional methods struggle to balance computational feasibility with the complexity of the design space. Through a brief theoretical analysis, we demonstrate that the optimization objective used in normalizing flows—specifically, maximizing the exact likelihood—can be directly applied to optimize the Expected Information Gain (EIG) without requiring any modifications. This dual-purpose objective not only simplifies the implementation but also ensures that the experimental design process aligns seamlessly with the underlying probabilistic framework. The results validate the scalability and effectiveness of this approach, making it a practical solution for optimizing experimental design in both medical and seismic applications [41]. By leveraging the exact likelihood evaluation capabilities flexibility of normalizing flows, this method provides a unified framework for simultaneously providing uncertainty aware solutions and reducing the costs associated with data acquisition. A version of this chapter was presented SIAM UQ 2024 and has a preprint [42].

- **Chapter 7 Exploiting long-range correlations: the pitfalls of patch training for uncertainty quantification in large scale imaging.** This chapter addresses a common and practical question about the software I developed, `InvertibleNetworks.jl`, which enables training on large input sizes: ”Why use the full image during training when smaller patches could suffice?” This question arises primarily due to the memory limitations of training on GPUs, which can make working with full-size images challenging or even infeasible. To explore this, we investigate the implications of training on image patches instead of full images. Through a series of experiments, we systematically demonstrate the potential pitfalls of this approach. Each experiment highlights specific scenarios where training on patches leads to suboptimal performance. The chapter concludes with a significant breakthrough: the development of a probabilistic solution for 3D photoacoustic imaging, a first of its kind. This solution leverages the strengths of `InvertibleNetworks.jl` while addressing the challenges posed by patch-based training, demonstrating the feasibility of scaling probabilistic imaging methods to high-dimensional 3D applications.

CHAPTER 2
ADJOINT OPERATORS ENABLE FAST AND AMORTIZED MACHINE LEARNING BASED BAYESIAN UNCERTAINTY QUANTIFICATION

2.1 Summary

Machine learning algorithms are powerful tools in Bayesian uncertainty quantification (UQ) of inverse problems. Unfortunately, when using these algorithms medical imaging practitioners are faced with the challenging task of manually defining neural networks that can handle complicated inputs such as acoustic data. This task needs to be replicated for different receiver types or configurations since these change the dimensionality of the input. We propose to first transform the data using the adjoint operator —ex: time reversal in photoacoustic imaging (PAI) or back-projection in computer tomography (CT) imaging — then continue posterior inference using the adjoint data as an input now that it has been standardized to the size of the unknown model. This adjoint preprocessing technique has been used in previous works but with minimal discussion on if it is biased. In this work, we prove that conditioning on adjoint data is unbiased for a certain class of inverse problems. We then demonstrate with two medical imaging examples (PAI and CT) that adjoints enable two things: Firstly, adjoints partially undo the physics of the forward operator resulting in faster convergence of a learned Bayesian UQ technique. Secondly, the algorithm is now robust to changes in the observed data caused by different transducer subsampling in PAI and number of angles in CT. Our adjoint-based Bayesian inference method results in point estimates that are faster to compute than traditional baselines and show higher SSIM metrics, while also providing validated UQ.

2.2 Description of Purpose

The power of machine learning methods bring accelerated and high fidelity solutions for inverse problems in a variety of fields [43]. On the downside, many machine learning methods are black boxes with failure cases that are difficult to predict and interpret. This is one of the reasons that their adoption in safety critical settings is hampered. Our purpose is to increase trustworthiness of machine learning (ML) for medical imaging by enabling uncertainty. This is important since applying ML under distribution shifts or poor training can cause instabilities and even hallucinations that could lead to incorrect diagnoses [44, 45]. Uncertainty quantification (UQ) alleviates this problem by communicating to practitioners when a method is confident in its result versus when it should not be trusted.

We describe a practical Uncertainty Quantification framework based on adjoint operators and amortized variational inference (AVI). These two concepts marry powerful data-driven methods with physics knowledge allowing us to amortize over unseen data and also different imaging configurations. By amortizing, we mean that the framework trades an expensive pretraining phase for fast inference results on unseen observations. The particular class of algorithms we study, can be trained given only examples of the parameters of interest $\mathbf{x}$ and their corresponding simulated data $\mathbf{y}$. In medical imaging, data $\mathbf{y}$ can contain complex physical phenomena such as acoustic waves in photoacoustic imaging. Under the hood, an ML-based algorithm learns to undo the complex physical phenomena when providing an estimate of $\mathbf{x}$. This can lead to long training times and large training data quotas which we would like to ameliorate. We are also interested in creating an algorithm that is robust to changes in the dimensionality of the observations $\mathbf{y}$ such as when changing the number of transducers. To solve both problems, we propose preprocessing the data $\mathbf{y}$ with the adjoint operator.

We first show that the posterior given data is equivalent to the posterior given data preprocessed with the adjoint in the case of linear forward operators and Gaussian noise. Many medical imaging modalities fall in this category i.e. photoacoustic imaging, CT and MRI. Equipped with this theoretical result, we demonstrate that this method solves our two problems since it accelerates training convergence of AVI with conditional normalizing flows. This is a welcome result since normalizing flows are notoriously costly to train (40 GPU weeks for the seminal GLOW normalizing flow [46]). Second, the adjoint brings data to the model space and therefore standardizes data size, enabling us to learn a single amortized normalizing flow that can sample the posterior for a variety of imaging configurations. We demonstrate these results using two medical imaging applications.

2.3 Methods

2.3.1 Bayesian Uncertainty

Given our quantity of interest $\mathbf{x} \in \mathcal{X}$ called the model, the forward problem is described by a linear operator $\mathbf{A} : \mathcal{X} \rightarrow \mathcal{Y}$ whose action on $\mathbf{x}$ gives observations $\mathbf{y} \in \mathcal{Y}$. Here, we consider linear problems with additive Gaussian noise term ε—i.e.,

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \varepsilon. \tag{2.1}$$

Upon observing data $\mathbf{y}$, traditional methods in inverse problems create a single point estimate of the $\mathbf{x}$ that produced $\mathbf{y}$. In the absence of noise and for invertible operators this point estimate is enough to describe the solution of the inverse problem. However, This approach fails to fully characterize the solution space in ill-posed problems [9], where there is no guarantee of a unique solution. In this scenario, it is important to be able to answer questions such as: Can we trust the solution or is this estimate affected by the null space of the operator? If there is more than one solution, what is the variability between these solution? These are questions that are answered by UQ.

At the forefront of uncertainty quantification (UQ) for inverse problems are Bayesian methods[10]. In a Bayesian framework, we try to find the set or *distribution* of solutions that all explain the data. This set of solutions is encoded by the conditional distribution $p(\mathbf{x} \mid \mathbf{y})$ called the posterior distribution and is the end goal of Bayesian inference . It contains all information needed to estimate $\mathbf{x}$ given $\mathbf{y}$ while providing UQ. Calculating the posterior distribution can be done with two main type of algorithms: first, sampling based algorithms such as Markov chain Monte Carlo (MCMC) algorithms [14] and second, optimization based algorithms such as variational inference (VI) [47]. MCMC can be a costly method due to the amount of sampling required [11], especially in high-dimensional problems. We instead explore variational inference (VI) [47] to sample the posterior since it allows for amortized training costs.

2.3.2 Variational Inference for Posterior Distribution Learning

To reduce the overall costs of posterior sampling, VI methods reduce the sampling problem into an optimization problem by finding a approximate distribution that best fits the desired distribution [47]. The goodness of fit is typically measured by the Kullback-Leibler (KL) divergence. Among the various VI methods, normalizing flows [48] have been shown to be flexible, efficient and powerful while working on a variety of distributions including conditional distributions [46, 49]. For our case, the goal is to find the normalizing flow f_θ parameterized by θ that makes a learned conditional distribution $p_\theta(\mathbf{x} \mid \mathbf{y})$ approximate the desired posterior distribution $p(\mathbf{x} \mid \mathbf{y})$. We measure the "closeness" with the KL-divergence making the optimization objective

$$\hat{\theta} = \underset{\theta}{\arg\min}\, \mathbb{KL}\left(p_\theta(\mathbf{x} \mid \mathbf{y}) \mid\mid p(\mathbf{x} \mid \mathbf{y})\right). \tag{2.2}$$

Previous work has been put into VI with normalizing flows that involves costly optimization for each incoming observed data $\mathbf{y}$ [50, 51, 52, 53]. The optimization is costly because learned parameters of f_θ typically parameterize neural networks that are costly to optimize. On top of that, these VI objectives requires online use of the forward operator $\mathbf{A}$ and its adjoint $\mathbf{A}^*$ during optimization. With this formulation, VI is not efficient enough to enable quick inference. Quick results are particularly important in medical imaging settings since they extend the abilities of a given modality for example by enabling the use of hand-held probes [54, 55]. In general, minimizing turnover time makes the difference in providing a timely diagnosis [6].

A different formulation of VI called amortized variational inference (AVI) [56, 57, 58, 59, 60] aims to obtain fast inference on test data without having to re-optimize an objective. This is accomplished by an intensive pretraining phase that optimizes a (KL) divergence based objective averaged over different data $\mathbf{y}$ sampled from the distribution $p(\mathbf{y})$: $\hat{\theta} = \arg\min_\theta \mathbb{E}_{\mathbf{y}\sim p(\mathbf{y})} \left[\mathbb{KL}\left(p(\mathbf{x} \mid \mathbf{y}) \,||\, p_\theta(\mathbf{x} \mid \mathbf{y})\right)\right]$. One can optimize a simplified objective that only requires samples from the joint distribution $\mathbf{x}, \mathbf{y} \sim p(\mathbf{x}, \mathbf{y})$ [31, 61]. For a conditional normalizing flow (CNF) f_θ, our objective becomes

$$\hat{\theta} = \arg\min_\theta \frac{1}{N} \sum_{n=1}^{N} \left(\| f_\theta(\mathbf{x}^{(n)}; \mathbf{y}^{(n)}) \|_2^2 - \log |\det \mathbf{J}_{f_\theta}| \right) \tag{2.3}$$

where N is the size of a training dataset $\mathcal{D} = \{(\mathbf{x}^{(n)}, \mathbf{y}^{(n)})\}_{n=1}^{N}$ and $\mathbf{J}_{f_\theta}$ is the Jacobian of the CNF. This objective is particularly simple to implement with conditional normalizing flows since they allow for tractable computation of the determinant Jacobian term $\det \mathbf{J}_{f_\theta}$ by design.

Our goal is to generalize our CNF for a variety of imaging configurations. Then we must consider that a different imaging configuration $\mathbf{A}_i$ can change the size of $\mathbf{y}_i$. These different data sizes can arise from changing receiver settings such as number of receivers or view angles. To handle a set of imaging configurations $\{i = 1 : M\}$ (where M is the number configurations), one would need to define a network f_{θ_i}, for all M configurations i.e. manually defining downsampling layers. Standardizing observations to a single size would allow practitioners to only need to define a single network f_θ since now all inputs to the network are the same size regardless of their imaging configurations. The adjoint $\mathbf{A}_i^*$

offers a physics informed way of standardizing data inputs to a single size, namely the size of the model.

Standardizing the size of incoming data is related to the concept of a summary statistic [62, 31] so can we interpret the adjoint as a physics-informed summary statistic. We will show that on top of being robust over different imaging configurations, adjoints also accelerate the convergence of the training objective in Equation 2.3. Before demonstrating these two practical advantages of using the adjoint, we present our main contribution: a theoretical discussion showing when preprocessing data with the adjoint will not affect the result of posterior inference.

2.3.3 Adjoint Data is Bayesian Sufficient

As noted in the previous section, there are good reasons to use the adjoint operator as a preprocessor. This leads to an important question that we phrase using the language of [63]: can we condition on adjoint data without introducing bias into the inference procedure? We will answer this question in the affirmative by using Proposition 1 from [64] and specifying a class of inverse problems that satisfies the proposition with adjoint preprocessing.

Proposition 1:[64] If $\mathcal{B}$ is injective on the range of Π then $p(\mathbf{x} \mid \mathcal{B}\,\Pi\,\mathbf{y})$ will be equal to $p(\mathbf{x} \mid \mathbf{y})$ if and only if the information lost by observing $\Pi\mathbf{y}$ instead of $\mathbf{y}$ is conditionally independent of $\mathbf{x}$ given $\Pi\mathbf{y}$:

$$p(\mathbf{x} \mid \mathcal{B}\,\Pi\,\mathbf{y}) = p(\mathbf{x} \mid \mathbf{y}) \iff \mathbf{x} \perp\!\!\!\perp \mathbf{y} - \Pi\mathbf{y} \mid \Pi\mathbf{y}. \tag{2.4}$$

Proposition 1a: Given data $\mathbf{y}$ (created as in (Equation 2.1)), the posterior conditioned on adjoint-preprocessed data $p(\mathbf{x} \mid \mathbf{A}^{*}\mathbf{y})$ will be equal to the original posterior $p(\mathbf{x} \mid \mathbf{y})$ if the additive noise ε is Gaussian.

Proof: We use Proposition 1 from [64] and set $\mathcal{B} = \mathbf{A}^{*}$; and $\Pi = \mathbf{A}\mathbf{A}^{+}$ where $\mathbf{A}^{+}$ is the Moore-Penrose inverse. Since $p(\mathbf{x} \mid \mathcal{B}\,\Pi\mathbf{y}) = p(\mathbf{x} \mid \mathbf{A}^{*}\mathbf{A}\mathbf{A}^{+}\mathbf{y}) = p(\mathbf{x} \mid \mathbf{A}^{*}\mathbf{y})$

then our proof is complete if we show that the following conditional independence is true: $\mathbf{x} \perp\!\!\!\perp \mathbf{y} - \mathbf{A}\mathbf{A}^{+}\mathbf{y} \mid \mathbf{A}\mathbf{A}^{+}\mathbf{y}$.

To show this, we note that the noise ε can be decomposed as the sum of two independent components – one that lives in $\operatorname{ran}(\mathbf{A})$ (the range of $\mathbf{A}$) and another that lives in $\operatorname{ran}(\mathbf{A})^{\perp}$: $\varepsilon = \varepsilon_{\mathbf{ran}} + \varepsilon_{\perp}$. Assuming this structure, the observed data is $\mathbf{y} = \mathbf{A}\mathbf{x} + \varepsilon_{\perp} + \varepsilon_{\mathbf{ran}}$. Since $\mathbf{A}\mathbf{A}^{+}$ is the orthogonal projector onto $\operatorname{ran}(\mathbf{A})$ then whenever $\mathbf{A}\mathbf{A}^{+}$ interacts with $\mathbf{y}$ it will make the contribution from $\varepsilon_{\perp}$ vanish:

$$
\begin{aligned}
&\mathbf{x} \perp\!\!\!\perp \mathbf{y} - \mathbf{A}\mathbf{A}^{+}\mathbf{y} \mid \mathbf{A}\mathbf{A}^{+}\mathbf{y} \\
&= \mathbf{x} \perp\!\!\!\perp (\mathbf{A}\mathbf{x} + \varepsilon_{\perp} + \varepsilon_{\mathbf{ran}}) - \mathbf{A}\mathbf{A}^{+}(\mathbf{A}\mathbf{x} + \varepsilon_{\perp} + \varepsilon_{\mathbf{ran}}) \mid \mathbf{A}\mathbf{A}^{+}(\mathbf{A}\mathbf{x} + \varepsilon_{\perp} + \varepsilon_{\mathbf{ran}}) \\
&= \mathbf{x} \perp\!\!\!\perp \mathbf{A}\mathbf{x} + \varepsilon_{\perp} + \varepsilon_{\mathbf{ran}} - \mathbf{A}\mathbf{A}^{+}\mathbf{A}\mathbf{x} - \mathbf{A}\mathbf{A}^{+}\varepsilon_{\perp} - \mathbf{A}\mathbf{A}^{+}\varepsilon_{\mathbf{ran}} \mid \mathbf{A}\mathbf{A}^{+}\mathbf{A}\mathbf{x} + \mathbf{A}\mathbf{A}^{+}\varepsilon_{\perp} + \mathbf{A}\mathbf{A}^{+}\varepsilon_{\mathbf{ran}} \\
&= \mathbf{x} \perp\!\!\!\perp \mathbf{A}\mathbf{x} + \varepsilon_{\perp} + \varepsilon_{\mathbf{ran}} - \mathbf{A}\mathbf{A}^{+}\mathbf{A}\mathbf{x} - \varepsilon_{\mathbf{ran}} \mid \mathbf{A}\mathbf{A}^{+}\mathbf{A}\mathbf{x} + \varepsilon_{\mathbf{ran}} \\
&= \mathbf{x} \perp\!\!\!\perp \mathbf{A}\mathbf{x} + \varepsilon_{\perp} + \varepsilon_{\mathbf{ran}} - \mathbf{A}\mathbf{x} - \varepsilon_{\mathbf{ran}} \mid \mathbf{A}\mathbf{x} + \varepsilon_{\mathbf{ran}} \\
&= \mathbf{x} \perp\!\!\!\perp \varepsilon_{\perp} \mid \mathbf{A}\mathbf{x} + \varepsilon_{\mathbf{ran}}. \qquad (2.5)
\end{aligned}
$$

By the d-separation criterion [65], Equation 2.5 is true because $\varepsilon_{\perp}$ is independent of all other elements, including $\varepsilon_{\mathbf{ran}}$ as per the assumption. Thus we prove that for linear problems with Gaussian additive noise $p(\mathbf{x} \mid \mathbf{A}^{*}\mathbf{y}) = p(\mathbf{x} \mid \mathbf{y})$ meaning adjoint preprocessing does not change the posterior inference.

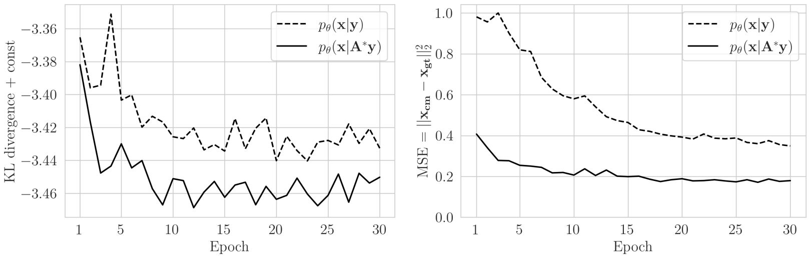
Corruption noise is often approximated as Gaussian additive in medical modalities including our applications in photoacoustic imaging and CT [66]. Our proof does not cover non-Gaussian noise, multiplicative noise or noise correlated with $\mathbf{A}$ or $\mathbf{x}$. We leave those for future work.

2.4 Results

We show three main results. First, that adjoint preprocessing accelerates the convergence of a conditional normalizing flow training to sample from the posterior distribution. Secondly, we demonstrate that the adjoint operator enables amortization over varying imaging configurations while using the same underling neural network. Finally, we validate the learned UQ by demonstrating posterior consistency through three tests.

2.4.1 Adjoint Accelerates Convergence:

We design a photoacoustic simulation and CNF architecture to compare two scenarios, namely learning $p_\theta(\mathbf{x} \mid \mathbf{y})$ by training on pairs $\mathcal{D} = \{(\mathbf{x}^{(n)}, \mathbf{y}^{(n)})\}_{n=1}^{N}$ or adjoint preprocessed learning of $p_\theta(\mathbf{x} \mid \mathbf{A}^*\mathbf{y})$ with $\mathcal{D}^* = \{(\mathbf{x}^{(n)}, \mathbf{A}^*\mathbf{y}^{(n)})\}_{n=1}^{N}$. As noted in our motivations, creating a CNF f_θ that can accept raw data $\mathbf{y}$ as an input is a laborious task but we do this for one imaging configuration to provide a fair comparison. This task involves manually defining downsampling layers in the CNF that bring $\mathbf{y}$ to the appropriate dimensionality. Then we can proceed to train two CNF's where both underlying networks have the same architectures (with addition of the downsampling layer) and are trained using the same hyperparameters.



(a)

(b)

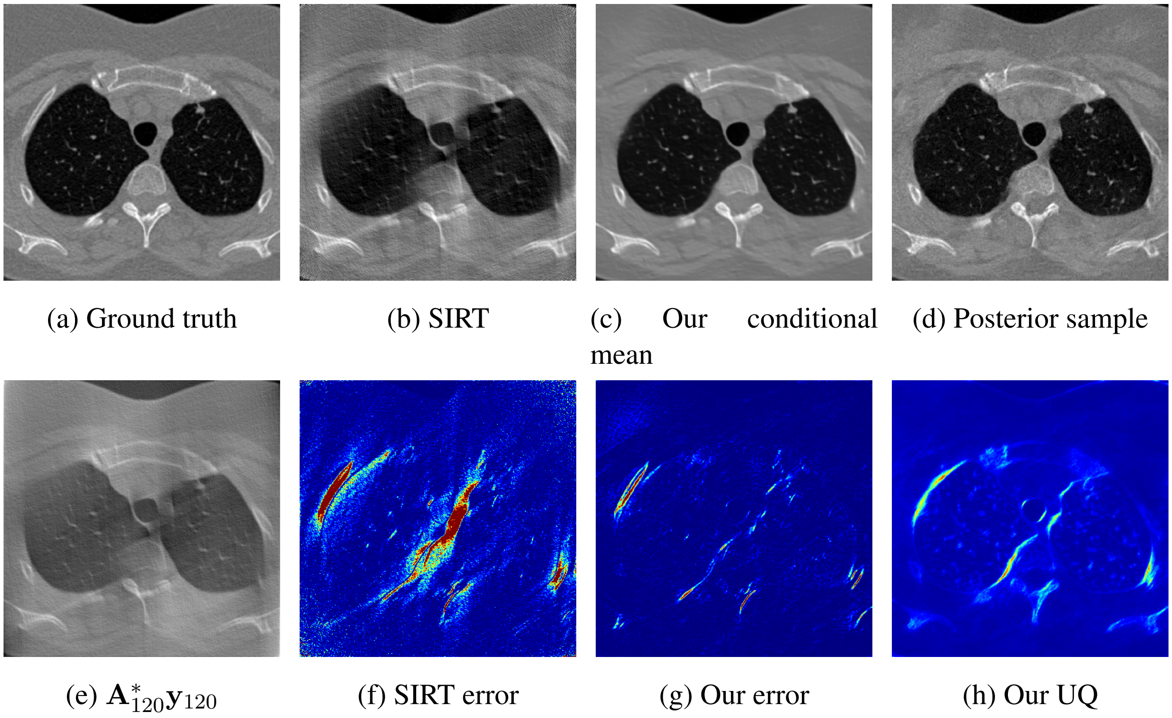
Figure 2.1: Convergence plots. (a) Posterior learning objective for data without (dashed) and with preprocessing with the adjoint $\mathbf{A}^*$. The adjoint accelerates convergence. (b) MSE of the conditional mean yielding improved Bayesian inference with less training time (juxtapose solid and dashed line for raw and preprocessed data).

Figure Figure 2.3 shows that for equivalent base architectures and training, learning $p_\theta(\mathbf{x}|\mathbf{A}^*\mathbf{y})$ is accelerated compared to learning $p_\theta(\mathbf{x}|\mathbf{y})$. We also plot the mean squared error (MSE) between the calculated conditional mean $\mathbf{x_{cm}}$ and the ground truth $\mathbf{x_{gt}}$. While MSE is not the training objective, it is still an important proxy since the conditional mean of the true posterior is the one that gives the lowest expected error [67, 68]. Here and throughout, the conditional mean $\mathbf{x_{cm}}$ and the variance (our UQ) is estimated using an average of 64 generated samples from the posterior. The training logs in Figure Figure 2.3 are averages of an unseen validation set N_{val}=192 created by a 10% split from the training dataset of N=2048 pairs.

We emphasize that our posterior sampling after training is fast. After applying the adjoint, the CNF is conditioned on $\mathbf{A}^{*}\mathbf{y}$ then the user generates the desired quantity of posterior samples (10 millisec/sample on our GPU). The time to create a point estimate with the conditional mean $\mathbf{x_{cm}}$ (around 2 seconds with 64 posterior samples) is favorable compared to traditional least-squares approaches that require several forward and adjoint evaluations.

In Figure Figure 2.2, we show images of results from posterior sampling on limited-view CT after training with adjoint preprocessing. For training and testing, we used the lodopab-ct dataset in the original resolution of 360×360 [69]. For CT forward and adjoint simulations we use [70]. Our experimental setup, follows [71] for limited-view CT with SNR $= 40$dB additive Gaussian noise. It has been said that bijective methods were not viable for a resolution of 256×256 [71]. We did not find this to be the case, our conditional normalizing flow is completely bijective and the implementation from InvertibleNetworks.jl [33] we did not see any out-of-memory problems training this method on the original 360×360 resolution.

For the baseline CT, we use the simultaneous iterative reconstruction technique (SIRT) as described in [72, 73]. Compared to the baseline Figure Figure 2.2b, our method Figure Figure 2.2c produces cleaner images that can deal with the substantial null-spaces due to limited-views. Importantly, structures in the null-space are illuminated by our UQ Figure Figure 2.2h pointing to reduced confidence in those areas.



(a) Ground truth (b) SIRT (c) Our conditional mean (d) Posterior sample

(e) $\mathbf{A}^*_{120}\mathbf{y}_{120}$ (f) SIRT error (g) Our error (h) Our UQ

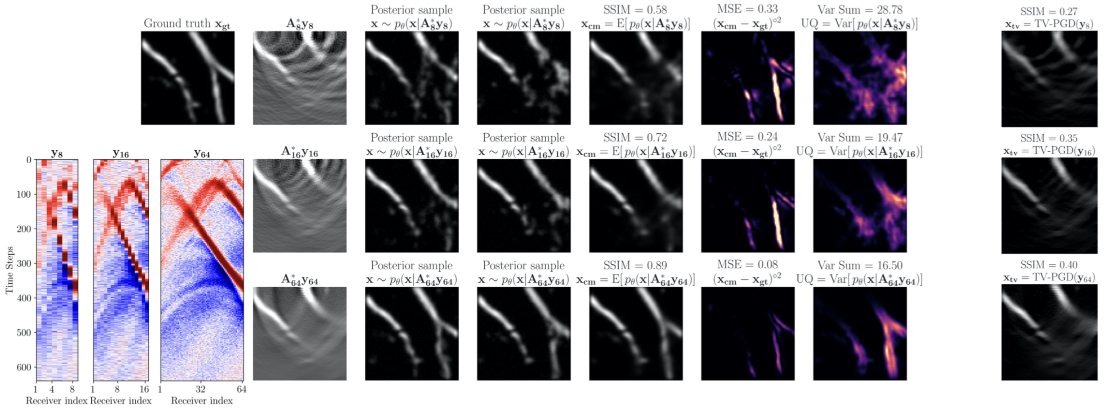
Figure 2.2: Computer tomography 360×360 images with uncertainty quantification. (a) Ground truth image; (b) Reconstructed image using SIRT baseline with 300 iterations; (c) Our image reconstruction made by averaging 64 samples from the learned posterior; (d) A single posterior sample from our method; (e) The adjoint data that has been brought to image space; (f) Error made by the SIRT baseline; (g) Error made by our conditional mean; (h) Our UQ calculated from the variance of posterior samples; Note: error plots and the UQ plot have the same colorbar limits $[0, 0.014]$

2.4.2 Adjoint Generalizes Different Imaging Configurations:

Data that is preprocessed with the adjoint $\mathbf{A}^*\mathbf{y}$ will always live in the space of the image. Thus we can use a single network to learn the posterior for different imaging configurations by augmenting our training dataset with examples from the desired configurations $(\mathbf{x}^{(n)}, \mathbf{A}_i^*\mathbf{y}_i^{(n)})\}_{n=1}^{N}$.

Receivers in photoacoustic imaging: We generalize over four photoacoustic imaging configurations consisting of data collected with 8, 16, 32, and 64 receivers. These different configurations are encoded by forward and adjoint operators $\mathbf{A}_i, \mathbf{A}_i^*$ with $i \in \{8, 16, 32, 64\}$. We train our CNF with N=2048 examples for each imaging configuration. Training took

14 hours on P1000 4GB GPU.


Ground truth $\mathbf{x}_{\mathrm{gt}}$
$\mathbf{A}_8^*\mathbf{y}_8$
Posterior sample $\mathbf{x} \sim p_\theta(\mathbf{x}|\mathbf{A}_8^*\mathbf{y}_8)$
Posterior sample $\mathbf{x} \sim p_\theta(\mathbf{x}|\mathbf{A}_8^*\mathbf{y}_8)$
SSIM = 0.58 $\mathbf{x}_{\mathrm{cm}} = \mathrm{E}[p_\theta(\mathbf{x}|\mathbf{A}_8^*\mathbf{y}_8)]$
MSE = 0.33 $(\mathbf{x}_{\mathrm{cm}} - \mathbf{x}_{\mathrm{gt}})^{\circ 2}$
Var Sum = 28.78 $\mathrm{UQ} = \mathrm{Var}[p_\theta(\mathbf{x}|\mathbf{A}_8^*\mathbf{y}_8)]$
SSIM = 0.27 $\mathbf{x}_{\mathrm{tv}} = \text{TV-PGD}(\mathbf{y}_8)$
$\mathbf{y}_8$
$\mathbf{y}_{16}$
$\mathbf{y}_{64}$
$\mathbf{A}_{16}^*\mathbf{y}_{16}$
Posterior sample $\mathbf{x} \sim p_\theta(\mathbf{x}|\mathbf{A}_{16}^*\mathbf{y}_{16})$
Posterior sample $\mathbf{x} \sim p_\theta(\mathbf{x}|\mathbf{A}_{16}^*\mathbf{y}_{16})$
SSIM = 0.72 $\mathbf{x}_{\mathrm{cm}} = \mathrm{E}[p_\theta(\mathbf{x}|\mathbf{A}_{16}^*\mathbf{y}_{16})]$
MSE = 0.24 $(\mathbf{x}_{\mathrm{cm}} - \mathbf{x}_{\mathrm{gt}})^{\circ 2}$
Var Sum = 19.47 $\mathrm{UQ} = \mathrm{Var}[p_\theta(\mathbf{x}|\mathbf{A}_{16}^*\mathbf{y}_{16})]$
SSIM = 0.35 $\mathbf{x}_{\mathrm{tv}} = \text{TV-PGD}(\mathbf{y}_{16})$
$\mathbf{A}_{64}^*\mathbf{y}_{64}$
Posterior sample $\mathbf{x} \sim p_\theta(\mathbf{x}|\mathbf{A}_{64}^*\mathbf{y}_{64})$
Posterior sample $\mathbf{x} \sim p_\theta(\mathbf{x}|\mathbf{A}_{64}^*\mathbf{y}_{64})$
SSIM = 0.89 $\mathbf{x}_{\mathrm{cm}} = \mathrm{E}[p_\theta(\mathbf{x}|\mathbf{A}_{64}^*\mathbf{y}_{64})]$
MSE = 0.08 $(\mathbf{x}_{\mathrm{cm}} - \mathbf{x}_{\mathrm{gt}})^{\circ 2}$
Var Sum = 16.50 $\mathrm{UQ} = \mathrm{Var}[p_\theta(\mathbf{x}|\mathbf{A}_{64}^*\mathbf{y}_{64})]$
SSIM = 0.40 $\mathbf{x}_{\mathrm{tv}} = \text{TV-PGD}(\mathbf{y}_{64})$
Time Steps
0
100
200
300
400
500
600
1 4 8
1 8 16
1 32 64
Receiver index
Receiver index
Receiver index

(a) (b)

Figure 2.3: (a) Amortized training of neural networks capable of sampling from posterior distributions for differently sized observations $\mathbf{y}_i$. As the number of receivers is increased, the samples show posterior contraction, a Bayesian phenomenon [74] that says increasing the amount of data should decrease the width of the posterior. (b) The baseline method (TV-projected gradient descent) fails to image vertical vessels.

After training, we sample from the learned posterior for unseen test data examples, $\mathbf{y}_i$ for varying numbers of receivers. The results demonstrate proper posterior contraction[74]. Visually, this is confirmed in Figure Figure 2.3a since uncertainty (quantified via pointwise variance) goes down when we increase receivers from 16 to 64. To quantitatively capture the global variation of these UQ images, we use the sum of pointwise variances: Var Sum = $\|\mathrm{Var}\|_1$ or equivalently the trace of the covariance matrix [75]. As the uncertainty reduces, the quality of the estimates increase, thus the posterior is contracting on the ground truth. We quantitatively verify this statement in the next section. Importantly, uncertainty is high where we expect, namely for vessels that are close to being vertical. These vertical events are in the null-space of the forward operator because the receivers are only located at the top of the model.

We compare our point estimate $\mathbf{x}_{\mathbf{cm}}$ with TV-projected gradient descent (TV-PGD) [76, 77] $\mathbf{x}_{\mathbf{tv}}$. Timing and quality metrics averaged over a test set of 96 samples are in Table Table 3.1. Across all receiver configurations, we show better Structural Similar Index Measure (SSIM). Also our method, produces the image reconstruction in less time.

Table 2.1: Photoacoustic image reconstruction timing and quality metric comparison.

Photoacoustic imaging	Timing (Seconds)	Quality metric (SSIM)			
	$N_{\text{rec}} = 64$	$N_{\text{rec}} = 8$	$N_{\text{rec}} = 16$	$N_{\text{rec}} = 32$	$N_{\text{rec}} = 64$
TV-proj GD $\mathbf{x_{tv}}$	16.78	0.27	0.36	0.42	0.44
Our conditional mean $\mathbf{x_{cm}}$	$\mathbf{1.72}$	**0.62**	**0.74**	**0.80**	**0.81**

View angles in computer tomography: To demonstrate generalization in CT, we train a single normalizing flow to sample from the posterior of three different quantities of views 60, 90 and 120 degrees. We add 40dB Gaussian noise to measurements. The raw data measurements $\mathbf{y}_{60}, \mathbf{y}_{90}, \mathbf{y}_{120}$ are shown in Appendix Figure **??**. To train, we use 4000 images for each of the three configurations for a total of 12000 training images.

The results of the CT application show similar behaviour to the photoacoustic case. Bayesian contraction is clearly shown in Figure Figure 2.4 since the uncertainty is reduced as more angles are observed. Also the uncertainty is physically consistent with our understanding of the imaging system since higher uncertainty is placed on the view angles that are unseen. We compare the quality of our CT generalized normalizing flow by comparing with the SIRT baseline. In Table Table 2.2, we show timing results for the maximum amount of angles and SSIM metrics for all tested angles. SSIM metrics are averages over a test set of 50 images. Our method is faster since the SIRT needs various applications of the forward and adjoint CT simulation while our method uses a single adjoint.

Table 2.2: Limited-view computer tomography image reconstruction timing and quality metric comparison.

Computed tomography	Timing (Seconds)	Quality metric (SSIM)		
	$N_{\text{ang}} = 120$	$N_{\text{ang}} = 60$	$N_{\text{ang}} = 90$	$N_{\text{ang}} = 120$
Baseline $\mathbf{x}_{\text{SIRT}}$	41.32	0.59	0.69	0.75
Our conditional mean $\mathbf{x_{cm}}$	$\mathbf{1.21}$	**0.78**	**0.84**	**0.88**

2.4.3 Validation of Uncertainty Quantification:

Since our problem does not have Gaussian priors, we can not analytically verify that our converged posteriors have reached the ground truth posterior. We can use three tests on the photoacoustic application to check that our posteriors are consistent and useful.

(i) posterior contraction by testing the CNF for different amount of data to check whether the posterior demonstrates contraction to the ground truth when increasing the amount of observed data. This contraction ultimately points to posterior consistency[74];

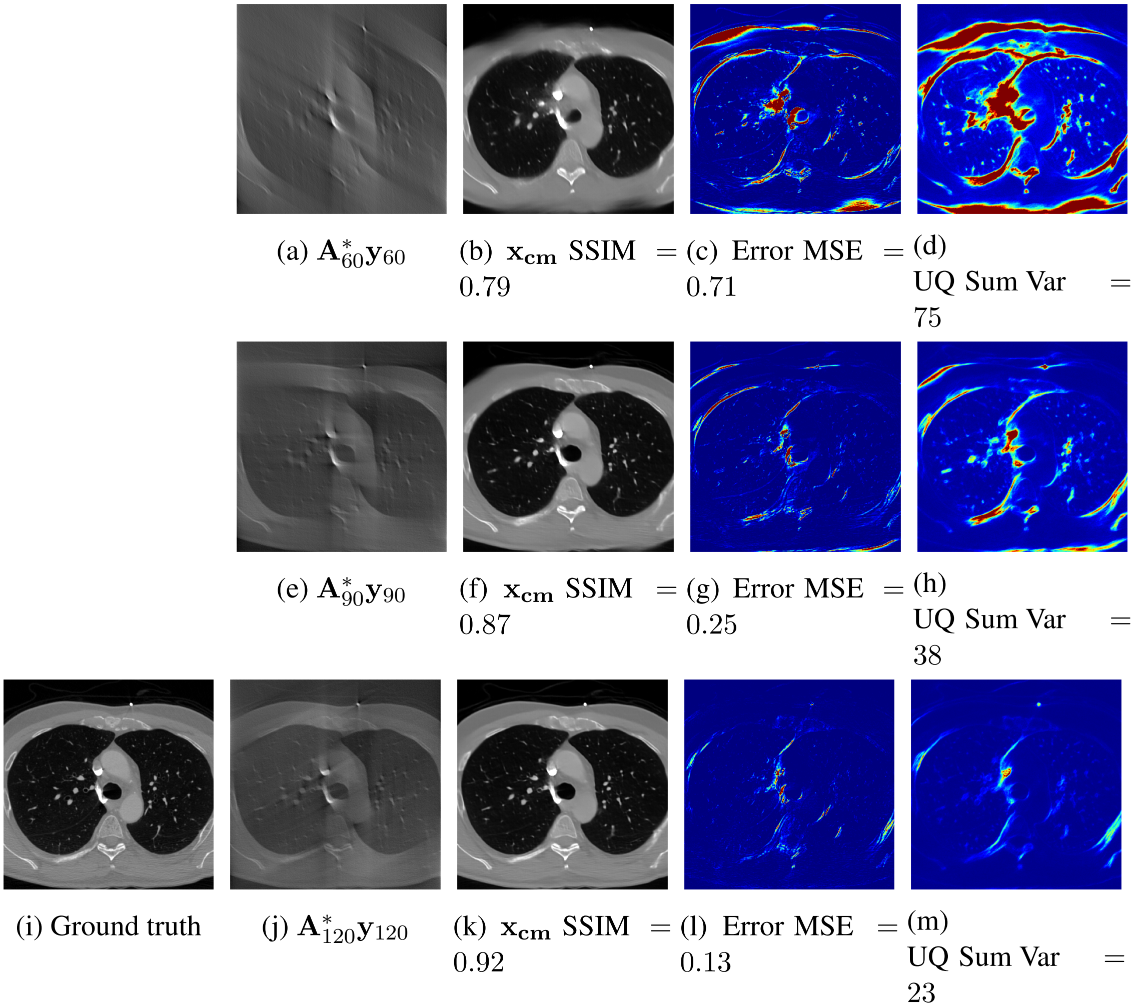
(ii) posterior calibration by checking whether our UQ correlates with regions in the image with large errors. This important check of UQ is called calibration and is established qualitatively by visually juxtaposing errors with UQ in Figure Figure 2.3a. For a more quantitative test, we plot a calibration line by using σ-scaling [78, 79];

(iii) simulation-based calibration (SBC) by testing for uniformity in the rank statistic when comparing various samples drawn from the proposed posterior with the known prior [31, 80].

The results of these three tests are included in Figure Figure 2.5 and show our learned posterior is consistent and approximates the true posterior. For this reason, we argue that a practitioner is justified in using our posterior for uncertainty quantification.

2.5 Conclusions

For linear operators and Gaussian noise, we prove that adjoint preprocessing posterior is equivalent to the original posterior $p(\mathbf{x}|\mathbf{A}^{*}\mathbf{y}) = p(\mathbf{x}|\mathbf{y})$. Although the distributions are ultimately the same, learning them poses different computational burdens for ML training. We showed that the adjoint accelerates convergence of CNFs for AVI. We also demonstrate that the adjoint allows us to train a single network to handle many different imaging configurations thus saving costs associated with designing network architectures for individual configurations. Our amortized posterior gives physically meaningful uncertainties that we also validate.



(a) $\mathbf{A}^*_{60}\mathbf{y}_{60}$

(b) $\mathbf{x_{cm}}$ SSIM $=$ 0.79

(c) Error MSE $=$ 0.71

(d) UQ Sum Var $=$ 75

(e) $\mathbf{A}^*_{90}\mathbf{y}_{90}$

(f) $\mathbf{x_{cm}}$ SSIM $=$ 0.87

(g) Error MSE $=$ 0.25

(h) UQ Sum Var $=$ 38

(i) Ground truth

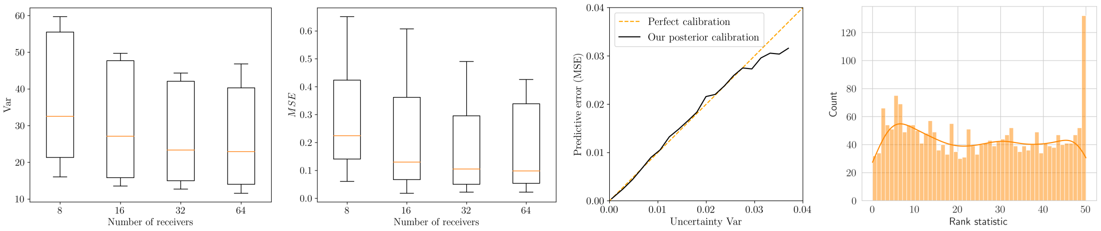
(j) $\mathbf{A}^*_{120}\mathbf{y}_{120}$

(k) $\mathbf{x_{cm}}$ SSIM $=$ 0.92

(l) Error MSE $=$ 0.13

(m) UQ Sum Var $=$ 23

Figure 2.4: Generalizing computer tomography over different view angles. (a,b,c,d) The first row shows results of our method for 60 view angles. (e,f,g,h) The second row shows results of our method for 90 view angles. (j,k,l,m) The third row shows results of our method for 120 view angles.



(a) (b) (c) (d)

Figure 2.5: Validation of uncertainty quantification (a) Posterior contraction when increasing the amount of data. (b) Posterior contraction towards the ground truth as measured by MSE. (c) Our posterior calibration is close to perfect calibration showing that our UQ is correlated with error made. (d) The uniformity of the SBC test shows that our marginalized posterior samples recover the prior distribution.

CHAPTER 3
AMORTIZED NORMALIZING FLOWS FOR TRANSCRANIAL ULTRASOUND WITH UNCERTAINTY QUANTIFICATION

3.1 Summary

We present a novel approach to transcranial ultrasound computed tomography that utilizes normalizing flows to improve the speed of imaging and provide Bayesian uncertainty quantification. Our method combines physics-informed methods and data-driven methods to accelerate the reconstruction of the final image. We make use of a physics-informed summary statistic to incorporate the known ultrasound physics with the goal of compressing large incoming observations. This compression enables efficient training of the normalizing flow and standardizes the size of the data regardless of imaging configurations. The combinations of these methods results in fast uncertainty-aware image reconstruction that generalizes to a variety of transducer configurations. We evaluate our approach with in silico experiments and demonstrate that it can significantly improve the imaging speed while quantifying uncertainty. We validate the quality of our image reconstructions by comparing against the traditional physics-only method and also verify that our provided uncertainty is calibrated with the error.

3.2 Introduction

Transcranial ultrasound computed tomography (TUCT) is a non-invasive, non-toxic imaging technique that aims to create images of internal brain tissue by transmission and reception of acoustic waves [81]. Its clinical applications range from hemorrhage detection to tumour imaging [82]. Previous approaches to TUCT utilized time-of-flight methods such as B-mode ultrasound [83]. These methods are limited in their imaging resolution for a variety of reasons, the foremost of which is due to their approximate treatment of wave physics [84]. Following the pioneering work by [85], it was shown that modeling all aspects of the acoustic wavefield enables high-resolution imaging of brain structures and anomalies. Since then, many works [86, 87, 88, 89, 90] are demonstrating increasing evidence from both in silico and controlled laboratory experiments that these full wavefield methods are capable of producing reliable brain images bringing this novel approach closer to clinical viability. These full wavefield methods are denoted full-waveform inversion (FWI) and are adapted from sophisticated seismic imaging methods [91, 92]. On the downside, FWI methods are computationally intensive since they require the application of forward and gradient operators related to expensive partial differential equation (PDE) solutions. This limits the clinical use of FWI methods towards TUCT since they can take up 36 hours to form an image [85]. In addition, the imaging process is affected by incomplete measurements, noise and other sources of uncertainty that can limit the accuracy and reliability of TUCT. To alleviate these problems and facilitate the adoption of this new imaging modality, we propose a data-driven approach to TUCT that leverages normalizing flows to dramatically improve the speed of imaging and provide uncertainty quantification (UQ). While deep learning has tremendous potential in accelerating computational imaging [43], we identify the limitation that ultrasound measurements in TUCT are impractically large and contain complex relationships that are difficult to undo without the aid of the underlying physics model. We propose to solve these problems by using a physics-informed summary function that takes the physical wave model into account. For our data recording setup, this summary compresses the size of observations by a factor of $70\times$ allowing the use of GPU hardware accelerators. Figure 3.1 contains a schematic of our full proposed framework.

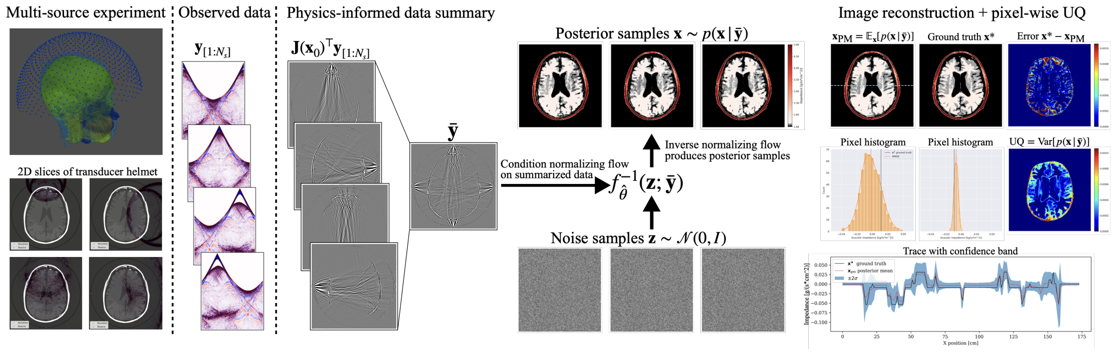

Multi-source experiment
Observed data
Physics-informed data summary
2D slices of transducer helmet
$\mathbf{y}_{[1:N_s]}$
$\mathbf{J}(\mathbf{x}_0)^\top\mathbf{y}_{[1:N_s]}$
$\bar{\mathbf{y}}$
Posterior samples $\mathbf{x} \sim p(\mathbf{x} \mid \bar{\mathbf{y}})$
Condition normalizing flow on summarized data
$f_{\hat{\theta}}^{-1}(\mathbf{z}; \bar{\mathbf{y}})$
Inverse normalizing flow produces posterior samples
Noise samples $\mathbf{z} \sim \mathcal{N}(0, I)$
Image reconstruction + pixel-wise UQ
$\mathbf{x}_{\mathrm{PM}} = \mathbb{E}_{\mathbf{x}}[p(\mathbf{x} \mid \bar{\mathbf{y}})]$
Ground truth $\mathbf{x}^*$
Error $\mathbf{x}^* - \mathbf{x}_{\mathrm{PM}}$
Pixel histogram
Pixel histogram
UQ = $\mathrm{Var}[p(\mathbf{x} \mid \bar{\mathbf{y}})]$
Trace with confidence band
$\mathbf{x}^*$ ground truth
$\mathbf{x}_{\mathrm{pm}}$ posterior mean
$\pm 2\sigma$
Impedance [g/(s*cm^2)]
X position [cm]
0.050
0.025
0.000
−0.025
−0.050
−0.075
−0.100
0
25
50
75
100
125
150
175

Figure 3.1: Proposed transcranial image reconstruction framework with normalizing flows for uncertainty quantification.

3.3 Methods

3.3.1 Ultrasound modeling

Our imaging approach, solves the inverse problem of finding acoustic properties of internal brain tissue that match observed ultrasound data. To model the propagation of ultrasound waves through a human skull, we use the scalar acoustic wave equation with variable density.

We express the data recording process (solving the wave equation PDE followed by a restriction of the wavefield to the transducer locations) by the discrete nonlinear operator, $\mathcal{F}$, acting on the ith known source represented by the vector $\mathbf{q}_i$. This nonlinear forward model is parameterized by the unknown acoustic impedance, discretized on a $N_x \times N_y$ grid (with $N_x = 512, N_y = 512$) and represented by the vector, $\mathbf{x} \in \mathbb{R}^{N_x \times N_y}$. For each source, $\mathbf{q}_i$, data is collected at N_r receivers for N_t time steps yielding

$$\mathbf{y}_i = \mathcal{F}(\mathbf{x})\mathbf{q}_i + \boldsymbol{\varepsilon}_i, \tag{3.1}$$

with $\mathbf{y}_{[1:N_s]} = \{\mathbf{y}_i\}_{i=1}^{N_s}$ being the full observation over N_s sources. To account for errors in the measurements, an additive noise term is included as $\boldsymbol{\varepsilon} \in \mathbb{R}^{N_r \times N_t}$. Typical 3D hardware setups have $N_s = 1024$ sources and for our 2D simulation we use up to $N_s = 32$ sources. This makes the full observation $\mathbf{y}_{[1:N_s]} \in \mathbb{R}^{N_r \times N_t \times N_s}$. Figure 3.2b shows data of a single source experiment with the acoustic impedance shown in Figure 3.2a. In our setup, we model $N_r = 256$ transducer receivers around the skull, of which $N_s = 32$ also act as sources. They record for $N_t = 2377$ time steps. Given observed transcranial ultrasound data, $\mathbf{y}_{[1:N_s]}$, our aim is to invert for internal structures $\mathbf{x}$. We solve this inverse problem in a Bayesian framework so uncertainty due to incomplete measurements, modeling errors, and noise, can be quantified systematically.

3.3.2 Bayesian transcranial ultrasound

Upon receiving observations $\mathbf{y}$, solving a Bayesian inverse problem involves sampling the conditional distribution of $\mathbf{x}$ given $\mathbf{y}$ [10]. This conditional distribution $p(\mathbf{x}|\mathbf{y})$ is called the posterior distribution. This posterior gives the full set of acoustic models $\mathbf{x}$ that explain the observations $\mathbf{y}$. To form an image reconstruction, one can use posterior samples to calculate high-quality point estimates such as the maximum a posteriori (MAP) and the minimum mean squared error (MMSE) estimator, while also providing uncertainty of those estimates. In general, the posterior distribution $p(\mathbf{x}|\mathbf{y})$ is computationally costly to sample from. Traditional methods like Markov chain Monte Carlo (McMC) require thousands of iterations, each of which needs to evaluate the expensive forward operator $\mathcal{F}$ [16, 12]. This makes these methods impractical for clinical use scenarios that require fast results [6]. In this paper, we suggest a variational inference method [19] that accomplishes fast posterior sampling by exploiting the distribution learning capabilities of generative models [93]. We will explain how our method derives from amortized density estimation where an expensive offline pre-training phase leads to fast posterior sampling at inference time for any in-distribution observation.

3.3.3 Amortized normalizing flows for posterior distribution sampling

Our goal is to sample from the distribution $p(\mathbf{x} \mid \mathbf{y})$ so that we can study the variation of different $\mathbf{x}$ that explain the observed data $\mathbf{y}$. Normalizing flows are a deep learning technique that have shown to be capable of learning to sample from complicated distributions [94, 48]. This method works by learning to map samples from the target distribution to standard white Gaussian noise using an invertible neural network f_θ with learned layers parameterized by θ. Once trained, the inverse of the network $f_{\hat{\theta}}^{-1}$ is evaluated on realizations of standard white Gaussian noise to generate new samples from the target distribution. Due to multi-scale transformations, normalizing flows scale favorably with dimension of the target distribution and allow for fast sampling [95] making them a good candidate for our high-dimensional medical image reconstruction task.

The posterior distribution $p(\mathbf{x} \mid \mathbf{y})$ we want to sample from is a conditional distribution so we use conditional normalizing flows [96, 97]. These learn to sample from a distribution conditioned on an observation $\mathbf{y}$ by minimizing the following objective:

$$\hat{\theta} \ = \underset{\theta}{\arg\min} \frac{1}{N} \sum_{n=1}^{N} \left(\| f_\theta(\mathbf{x}^{(n)}; \mathbf{y}^{(n)}) \|_2^2 - \log |\det \mathbf{J}_{f_\theta}| \right) \tag{3.2}$$

where $\mathbf{J}_{f_\theta}$ is the Jacobian of the network and $\{(\mathbf{x}^{(n)}, \mathbf{y}^{(n)})\}_{n=1}^{N}$ are training pairs given by Equation 3.1 and samples $\mathbf{x} \sim p(\mathbf{x})$ drawn from the prior. Intuitively, Equation 3.2 learns the posterior distribution by maximizing the likelihood of the $\mathbf{x}$ conditioned on $\mathbf{y}$ under the normalizing transformation f_θ. The first term is the likelihood in Normal distribution (ℓ_2 norm). Because the transformation is invertible, the change of variables formula is used to evaluate the likelihood in Normal space by controlling for volume changes caused under the normalizing transformation f_θ as quantified by the second Jacobian term. Mathematically, Equation 3.2 minimizes the Kullback-Leibler divergence between the learned posterior and the true posterior [31, 56, 61]. Crucially to our application, this method learns the posterior in an amortized fashion since it minimizes the objective over a distribution of $\mathbf{y}$. After training, the conditional normalizing flow can sample the posterior for unseen $\mathbf{y}$ at the cheap cost of passing noise through the inverse network. See Figure 3.1 for a schematic of the sampling process from noise.

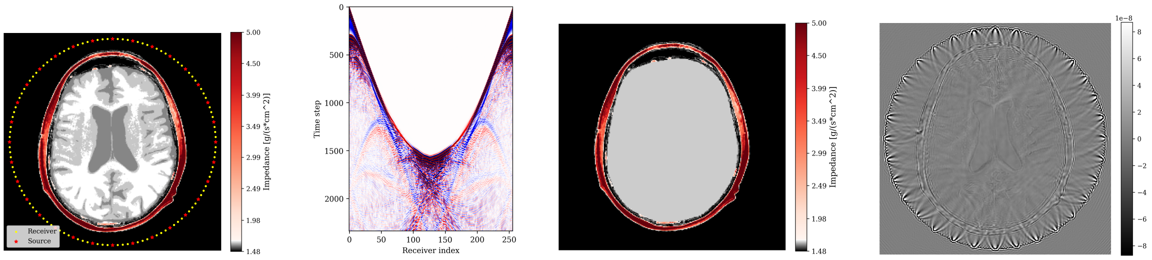
Normalizing flows, due to their architecture, have closed-form inverses (up to numerical precision), that cost the same as forward evaluation and the term $|\det \mathbf{J}_{f_\theta}|$ can be efficiently calculated. In general, training pairs needed to optimize Equation 3.2 are generated in the simulation-based inference framework [98] but for our ultrasound application, $\mathbf{y}$ is complicated acoustic data and is too large for GPU training thus we explore a physics-informed method to extract important features and compress its size.

3.3.4 Physics-informed summary statistic

For our ultrasound application, we identify three difficulties of working with acoustic data $\mathbf{y}$. First, the observation for all sources $\mathbf{y}_{[1:N_s]}$ is too large ($N_t \times N_r \times N_s \approx 19 \times 10^6$) to fit in a GPU for training. Second, different experimental configurations (i.e. varying number of sources) change the size of observations meaning generalization on data space requires sophisticated architectures [31]. Finally, imaging complicated structures directly from acoustic data is a difficult task [99]. These considerations motivate the need of a function h that reduces the size and "summarizes" the observation $\bar{\mathbf{y}} = h(\mathbf{y}_{[1:N_s]})$ while preserving information it carries about $\mathbf{x}$. These summaries are formally known as summary statistics [62, 31]. In the context of maximum likelihood estimation, [100] proposed the score of the likelihood as a summary statistic. This score is defined as the gradient of the log-likelihood $\mathcal{L} = \log p(\mathbf{y} \mid \mathbf{x})$ with respect to $\mathbf{x}$. [100] proved that the score is asymptotically maximally informative of $\mathbf{x}$. Inspired by this approach, we explore using the score as a summary function for posterior sampling. We assume a Gaussian noise model leading to the gradient being the Jacobian adjoint $\mathbf{J}^\top$ on the data residual:

$$\bar{\mathbf{y}} = h(\mathbf{y}_{[1:N_s]}) := \nabla_{\mathbf{x}_0}\mathcal{L} = \sum_{i=1}^{N_s} \mathbf{J}(\mathbf{x}_0, \mathbf{q}_i)^\top (\mathcal{F}(\mathbf{x}_0)\mathbf{q}_i - \mathbf{y}_i) \tag{3.3}$$

where $\mathbf{x}_0$ is a starting point at which the gradient is calculated. Note, Equation 3.3 involves evaluating the forward physical model $\mathcal{F}$ and its Jacobian adjoint $\mathbf{J}^\top$. Thus this summary is informed by the physics (domain knowledge). As a result, the summarized data $\bar{\mathbf{y}}$ lives in the reduced $N_x \times N_y$ image space (reduction factor of about 70). According to [101], the informativeness of this summary statistic also implies that $p(\mathbf{x} \mid \mathbf{y}) = p(\mathbf{x} \mid \bar{\mathbf{y}})$ thus we propose to use the same conditional distribution learning objective as Equation 3.2 but replace the data $\mathbf{y}$ with the summary $\bar{\mathbf{y}}$. One of the assumptions is that the starting point $\mathbf{x}_0$ needs to be carefully chosen as it will affect how informative the summary statistic will be. For our application, $\mathbf{x}_0$ is the acoustically correct model of the skull bone and a constant acoustic model inside the skull since the soft tissues inside the skull are the clinically relevant structures we care to image. Inclusion of the skull is needed so that the physical operators create meaningful results that inform the posterior. In practice, acoustic values of skull bone can be calculated from CT scans [102]. See Figure 3.2c for an example of $\mathbf{x}_0$ and Figure 3.2d for the physics-informed summary $\bar{\mathbf{y}}$ it creates.



(a) $\mathbf{x}^*$ (b) Data $\mathbf{y}_i$ (c) Starting model, $\mathbf{x}_0$ (d) Summarized data $\bar{\mathbf{y}}$

Figure 3.2: 2D transcranial ultrasound imaging setup.

While previous work has used the adjoint operator and pseudo-inverse to summarize data [68, 64] to the best of our knowledge this is the first work that explores based on theoretical arguments the use of the score of the likelihood as a summary statistic for direct posterior sampling in a inverse problem with an expensive physics-based nonlinear operator.

3.4 Experiments and Results

3.4.1 Normalizing flow training

To create training pairs, we require samples from the prior distribution $p(\mathbf{x})$ of ground truth brain acoustic impedance models. We derive ours from the FastMRI dataset [103] using an automatic process. For training and testing, we use 250 3D acoustic brains models each containing 11 512×512 slices. Out of these, we used 90% for training, 5% for validation and 5% for testing. We simulated the forward wave propagation $\mathcal{F}$ and its Jacobian adjoint $\mathbf{J}^\top$ using Devito [104, 105] and JUDI [7].

The conditional normalizing flow is implemented with InvertibleNetworks.jl [106]. Each epoch takes about 20 minutes and we trained for a total of 18 hours on a 32GB A100 GPU.

3.4.2 Image reconstruction from posterior samples

Once trained, our conditional normalizing flow can generate samples from the posterior with algorithm 1. The computational cost of posterior sampling is dominated by the calculation of the physics-informed summary $\bar{\mathbf{y}}$. This takes ≈ 1 second per source and 44.8 seconds in total for all 32 sources (on 4 core Intel Skylake CPU). This calculation only needs to be done once per ultrasound experiment after which many posterior samples can be generated each at the cheap cost of one inverse network evaluation (20ms/sample). With these posterior samples, statistical point estimates can be calculated including the minimum mean squared error (MMSE) estimator given by the posterior/conditional mean $\mathbf{x}_{\mathrm{PM}} = \mathbb{E}_{\mathbf{x}}[\,p(\mathbf{x} \mid \bar{\mathbf{y}})]$ that serves as our image reconstruction.

```
Algorithm 1: Amortized posterior inference (given unseen observation $\mathbf{y}_{[1:N_s]}$)
  Need: starting point $\mathbf{x}_0$
  Calculate gradient summary $\bar{\mathbf{y}} = \sum_{i=1}^{N_s} \mathbf{J}(\mathbf{x}_0, \mathbf{q}_i)^\top (\mathcal{F}(\mathbf{x}_0)\mathbf{q}_i - \mathbf{y}_i)$
  Sample $N_{post}$ Gaussian normal noise $\mathbf{z} \sim \mathcal{N}(0,\, I)$
  Pass $\mathbf{z}$'s through inverse of normalizing flow $f_{\hat{\theta}}^{-1}(\mathbf{z}; \bar{\mathbf{y}})$ to generate posterior
    samples.
```

For UQ, we look at the intra-sample variation between posterior samples. To visualize UQ on the entire image reconstruction we use the posterior variance Var$[\,p(\mathbf{x} \mid \bar{\mathbf{y}})]$. The posterior mean (and variance) is calculated by approximating their expectations with an average over $N_{\text{post}} = 128$ posterior samples

$$\mathbf{x}_{\mathrm{PM}} = \mathbb{E}_{\mathbf{x}}[p(\mathbf{x} \mid \bar{\mathbf{y}})] \approx \frac{1}{N_{\mathrm{post}}} \sum_{i=1}^{N_{\mathrm{post}}} \mathbf{x}_i \;\text{ where } \mathbf{x}_i = f_{\hat{\theta}}^{-1}(\mathbf{z}_i; \bar{\mathbf{y}}) \text{ and } \mathbf{z}_i \sim \mathcal{N}(0,\, I).$$

In this work, we concentrate on the posterior mean because it is the estimator with minimal mean squared error [50]. Figure 3.3 contains an example of the input and output of the proposed image reconstruction algorithm including UQ.

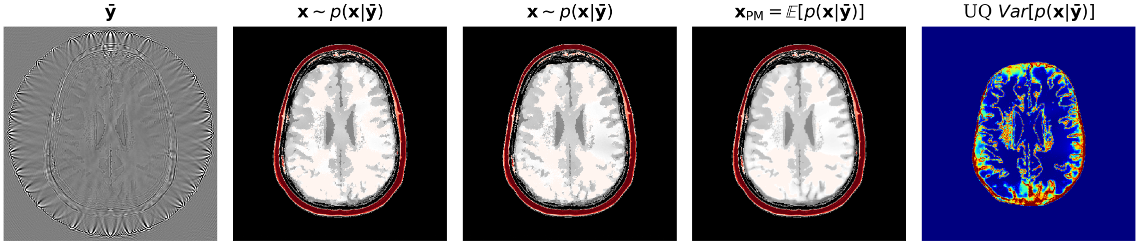

$\bar{\mathbf{y}}$
$\mathbf{x} \sim p(\mathbf{x}|\bar{\mathbf{y}})$
$\mathbf{x} \sim p(\mathbf{x}|\bar{\mathbf{y}})$
$\mathbf{x}_{\mathrm{PM}} = \mathbb{E}[p(\mathbf{x}|\bar{\mathbf{y}})]$
UQ $Var[p(\mathbf{x}|\bar{\mathbf{y}})]$

Figure 3.3: Image reconstruction with UQ using our method including samples from the posterior.

To assess the performance of our reconstruction, $\mathbf{x}_{\mathrm{PM}}$, we compare with two baseline methods, namely physics-only FWI, yielding $\mathbf{x}_{\mathrm{FWI}}$ obtained by gradient descent, and a supervised U-Net $\mathbf{x}_{\mathrm{UNET}}$ [107] trained on the same N data pairs $\{(\mathbf{x}^{(n)}, \bar{\mathbf{y}}^{(n)})\}_{n=1}^{N}$ as our method. Compared to the learned methods, which incur off-line training costs prior to inference, FWI is computationally intensive since it requires ~ 40 calls to the forward and gradient for each source while our method only requires one gradient per source.

From Figure 3.4, we make the following observations: (i) our result contains fewer artifacts compared to FWI; (ii) it performs better than U-Net; (iii) it captures the full posterior yielding pointwise variances that correlate well with error; (iv) due to averaging over posterior samples our result blurs a few details as compared to FWI. For a more quantitative comparison of the reconstruction quality, refer to Table 3.1 in which the average quality metrics for peak signal to noise ratio (PSNR); structural similarity index metric (SSIM); and root mean squared error (RMSE) are computed from 50 unseen test slices. Our method shows high performance on all metrics while keeping the online inference time significantly lower than the FWI method. For more direct comparison, we avoided measurement

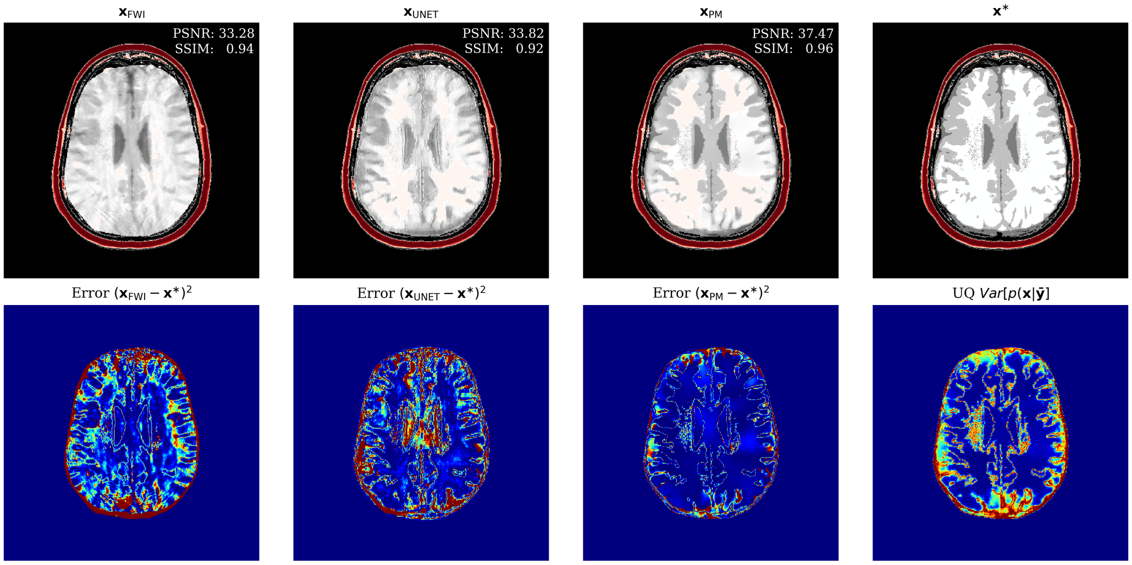

$\mathbf{x}_{FWI}$
PSNR: 33.28
SSIM: 0.94
$\mathbf{x}_{UNET}$
PSNR: 33.82
SSIM: 0.92
$\mathbf{x}_{PM}$
PSNR: 37.47
SSIM: 0.96
$\mathbf{x}^*$
Error $(\mathbf{x}_{FWI} - \mathbf{x}^*)^2$
Error $(\mathbf{x}_{UNET} - \mathbf{x}^*)^2$
Error $(\mathbf{x}_{PM} - \mathbf{x}^*)^2$
UQ $Var[p(\mathbf{x}|\tilde{\mathbf{y}})]$

Figure 3.4: Comparison with physics-only and data-only methods of FWI and supervised U-Net. Note that areas in our pointwise variance correlate well with areas of high error.

Table 3.1: Image reconstruction timing and quality metric comparison

Method	Timing (seconds)	PSNR ↑	SSIM ↑	RMSE ↓
FWI ($\mathbf{x}_{\mathrm{FWI}}$)	2100	33.25	0.9450	0.0215
Supervised UNet ($\mathbf{x}_{\mathrm{UNET}}$)	**44.8 + 0.02**	35.63	0.9332	0.0168
Our posterior mean ($\mathbf{x}_{\mathrm{PM}}$)	44.8 + 3.23	**38.67**	**0.9646**	**0.0119**

noise.

3.4.3 Generalization over experimental configurations

In Figure 3.5, we show how our method generalizes over different source configurations. Aside from handling different acquisition constraints, practitioners can also quickly prototype different configurations to decide which one meets their threshold of uncertainty.

Related work: The gradient we calculate for our summary statistic is connected to reverse time migration from seismic imaging [108]. For accessing uncertainty information in TUCT, [90] use the mean-field Gaussian approximation. Their method uses gradient descent with many expensive forward/gradient calls and assumes a Gaussian prior on the

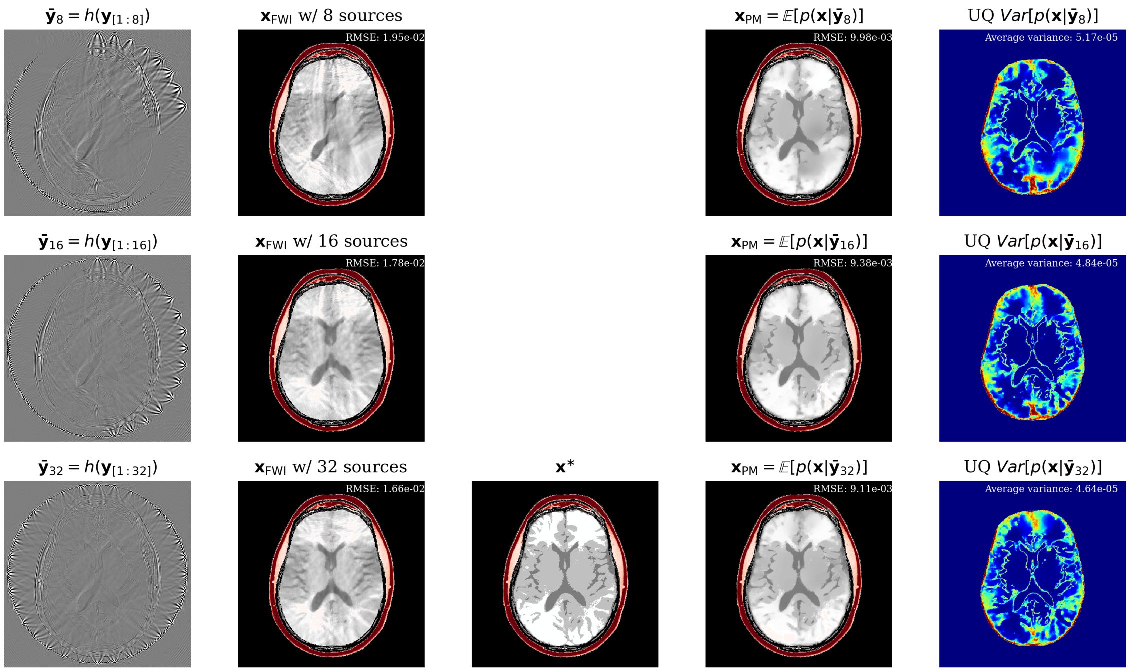

$\bar{\mathbf{y}}_8 = h(\mathbf{y}_{[1:8]})$
$\mathbf{x}_{\mathrm{FWI}}$ w/ 8 sources
RMSE: 1.95e-02
$\mathbf{x}_{\mathrm{PM}} = \mathbb{E}[p(\mathbf{x}|\bar{\mathbf{y}}_8)]$
RMSE: 9.98e-03
UQ $Var[p(\mathbf{x}|\bar{\mathbf{y}}_8)]$
Average variance: 5.17e-05
$\bar{\mathbf{y}}_{16} = h(\mathbf{y}_{[1:16]})$
$\mathbf{x}_{\mathrm{FWI}}$ w/ 16 sources
RMSE: 1.78e-02
$\mathbf{x}_{\mathrm{PM}} = \mathbb{E}[p(\mathbf{x}|\bar{\mathbf{y}}_{16})]$
RMSE: 9.38e-03
UQ $Var[p(\mathbf{x}|\bar{\mathbf{y}}_{16})]$
Average variance: 4.84e-05
$\bar{\mathbf{y}}_{32} = h(\mathbf{y}_{[1:32]})$
$\mathbf{x}_{\mathrm{FWI}}$ w/ 32 sources
RMSE: 1.66e-02
$\mathbf{x}^*$
$\mathbf{x}_{\mathrm{PM}} = \mathbb{E}[p(\mathbf{x}|\bar{\mathbf{y}}_{32})]$
RMSE: 9.11e-03
UQ $Var[p(\mathbf{x}|\bar{\mathbf{y}}_{32})]$
Average variance: 4.64e-05

Figure 3.5: Generalization over different imaging configurations. The three FWI results took ≈1.5 hours, but the three posterior means and UQ were calculated in ≈3 minutes. We observe that our method shows better results than the pure-physics FWI when there is less source coverage.

ground truth images while neglecting correlations between pixels. Our work, instead makes no underlining assumptions on the posterior/prior distributions and requires only one set of forward/gradient calls during inference. [31] explored learned summary statistics for posterior inference. Here we exploit knowledge of the underlying physics by introducing physics-informed summary statistics. Instead of including physics in learned simulations as in physics-informed neural networks, we include the physics in the data summary, which makes sense when dealing with inverse problems where observed data serves as input.

Future work: Normalizing flows are likelihood models so they allow for natural anomaly detection [109]. We will explore the possibility of evaluating our method on brains with anomalies for automatic detection of tumors or hemorrhages.

We highlight that our method assumes access to good starting points x_0. In future works, we would like to find ways to be robust against poor starting points.

Conclusions: The application of machine-learning methods and systematic uncertainty quantification to ultrasound imaging has been extremely challenging because of the high-dimensionality and high computational costs associated with handling the correct wave physics. Through the combination of conditional normalizing flows with physics-informed summary statistics, we arrive at a formulation capable of producing high-fidelity images with uncertainty quantification. By incurring an off-line pretraining cost, our method is faster than traditional physics-only methods.

CHAPTER 4

ASPIRE: ITERATIVE AMORTIZED POSTERIOR INFERENCE FOR BAYESIAN INVERSE PROBLEMS

SUMMARY

Due to their uncertainty quantification, Bayesian solutions to inverse problems are the framework of choice in applications that are risk averse. These benefits come at the cost of computations that are in general, intractable. New advances in machine learning and variational inference (VI) have lowered the computational barrier by learning from examples. Two VI paradigms have emerged that represent different tradeoffs: amortized and non-amortized. Amortized VI can produce fast results but due to generalizing to many observed datasets it produces suboptimal inference results. Non-amortized VI is slower at inference but finds better posterior approximations since it is specialized towards a single observed dataset. Current amortized VI techniques run into a sub-optimality wall that can not be improved without more expressive neural networks or extra training data. We present a solution that enables iterative improvement of amortized posteriors that uses the same networks architectures and training data. The benefits of our method requires extra computations but these remain frugal since they are based on physics-hybrid methods and summary statistics. Importantly, these computations remain mostly offline thus our method maintains cheap and reusable online evaluation while bridging the approximation gap these two paradigms. We denote our proposed method **ASPIRE** - **A**mortized posteriors with **S**ummaries that are **P**hysics-based and **I**teratively **RE**fined. We first validate our method on a stylized problem with a known posterior then demonstrate its practical use on a high-dimensional and nonlinear transcranial medical imaging problem with ultrasound. Compared with the baseline and previous methods from the literature our method stands out as an computationally efficient and high-fidelity method for posterior inference.

4.1 Introduction

Inverse problems are fundamental across various scientific and engineering disciplines, where the objective is to infer causative factors from observable effects. Our practical example involves complex medical imaging, where we infer the internal structures of the brain from ultrasound measurements through the skull. Given the need for rapid imaging in diagnostic medical applications [6], we seek a method that generalizes to unseen observations—referred to here as "amortized". While current amortized methods are fast, they often fail to resolve crucial details in high-dimensional and nonlinear problems, as shown in Figure 4.1. In contrast, a fully non-amortized approach, tailored to individual observations, provides high-resolution solutions but incurs prohibitive computational costs at imaging time. The method we explore represents a middle ground, aiming to deliver fast and also reliable solutions. The importance of such solutions in medical imaging, where they directly influence potentially life-saving diagnostic decisions, cannot be overstated.

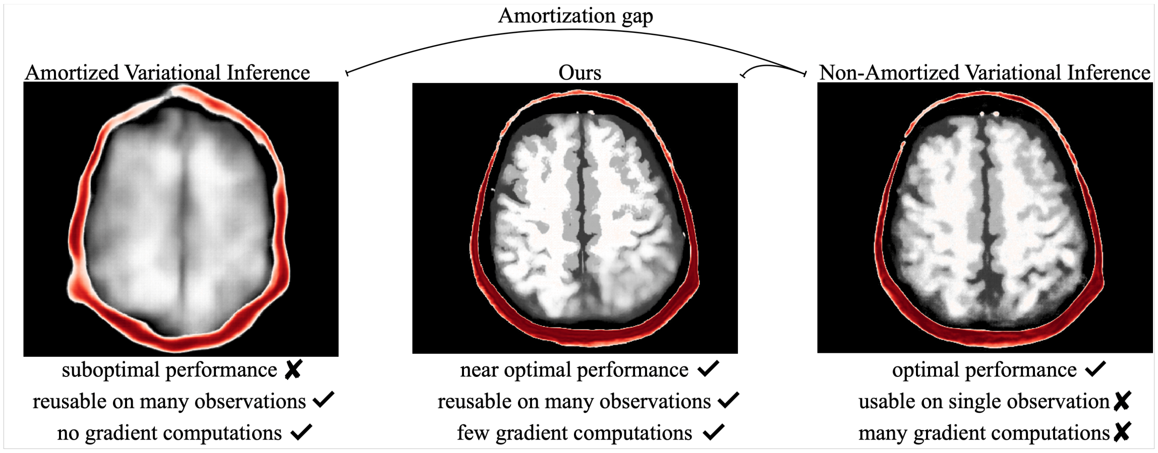

Amortization gap
Amortized Variational Inference
Ours
Non-Amortized Variational Inference
suboptimal performance ✗
reusable on many observations ✓
no gradient computations ✓
near optimal performance ✓
reusable on many observations ✓
few gradient computations ✓
optimal performance ✓
usable on single observation ✗
many gradient computations ✗

Figure 4.1: Our algorithm ASPIRE is a middle ground between amortized and non-amortized variational inference with the goal of providing a generalized method for fast yet reliable imaging.

Because non-amortized inference focuses on one single observation, its inference typically outperforms amortized inference [21]. Unfortunately, this improvement often comes at the expense of prohibitively high computational costs at inference time, rendering non-amortized inference impractical in situations where fast turn-around times are needed, such as in many medical imaging fields [6]. Amortized VI methods, on the other hand, while fast [30] at inference may suffer from the so-called amortization gap, a phenomenon that has been studied theoretically [110] and confirmed empirically from comparisons between amortized and non-amortized VI [21, 50, 58]. In that sense, there is a trade-off between runtime and quality at inference time. One either spends more on computations at inference time, or one accepts inferior inference quality in situations where fast turn-around times are essential. While trading quality for speed may be acceptable in some situations, it becomes problematic in circumstances where amortized VI produces unacceptable results, e.g., in cases where the inference problem is high dimensional and complicated by nonlinear forward operators. Currently, the following remedies exist: *(1)* increase the expressiveness of the parametric family used to approximate the posterior [111] or *(2)* add more samples to the training set [112]. In this work, we will explore a third complementary option to narrow the amortization gap. To this end, we propose an iterative amortized inference approach during which physics-based summary statistics are refined in tandem with neural posterior estimators thus bootstrapping the quality of the approximated posterior. We call this approach: **ASPIRE** - **A**mortized posteriors with **S**ummaries that are **P**hysics-based and **I**teratively **RE**fined. To motivate this approach, we will first explore the implementation of amortized posterior inference via neural density estimation, followed by physics-based gradient summary statistics, and their iterative refinement.

4.2 Contributions and related work

1. **Motivated by gradient-based, maximally informative summary statistics we introduce the ASPIRE algorithm, which iteratively refines amortized posterior inference while maintaining low online costs.** ASPIRE builds on top of current amortized Bayesian inference frameworks such as [31] that introduced the concept of the summary network, which is optimized under the same objective as a normalizing flow. Our approach extends this concept into what we term a “physical summary network”, where each refinement iteration enhances the summary statistic effectively bootstrapping the approximation quality. Although not iterative and for different modalities [113] demonstrated one of the first uses of normalizing flows for medical imaging.

2. **Theoretical proof and discussion of the conditions under which the posterior mean can improve on the current reconstruction for an illustrative linear inverse problem.** We chose to update intermediate reconstructions with the posterior mean of the current posterior approximation as opposed to the similar work from [114], which proposes a method for solving Bayesian inverse problems that resembles loop-unrolling augmented with Bayesian network layers. We acknowledge their contributions and note key differences: [114] employ Bayesian networks that model distributions on the network weights, which can impose restrictive assumptions on the distribution families that can be learned akin to mean-field approximations. In contrast, our use of normalizing flows aims to directly learn the Bayesian posterior, and theoretically, as universal approximators [115, 116], offer greater flexibility.

3. **Evaluation of our method's performance on a realistic and challenging transcranial medical imaging inverse problem with ultrasound, focusing on the accuracy of the posterior mean and the effectiveness of our uncertainty in predicting reconstruction errors.**

4. **Introduction of a second novel non-amortized inference method to serve as a "gold standard" that we denote WISER.** To illustrate the gains that our amortized method ASPIRE makes towards approaching the quality of non-amortized methods, we take as inspiration non-amortized methods from the literature [21, 117, 27] and introduce a novel non-amortized method that represents the best possible performance we can expect from our setting due to extra calculations of the forward physics operator and its adjoint.

5. **Qualitative and quantitative comparisons of ASPIRE against a current literature baseline [90], and our "gold standard" non-amortized inference method.** We have identified a single other work in the literature that probabilistically solves the transcranial medical imaging inverse problem [90]. Crucially, this method is not learned so we also compare with the "gold standard" non-amortized method that is learned. To accelerate the development of inference techniques for transcranial medical imaging [118], we introduce

benchmarks with accompanying datasets and code.

6. **Cost-benefit analysis of the computational costs associated with offline training versus the rapid online capabilities of the amortized method.** Our work also shares similarities with DEEPGEM by [119], which utilizes Expectation Maximization to solving inverse problems. Their process involves optimizing a non-amortized normalizing flow to sample from a posterior based on current nuisance parameter estimates, followed by Maximum A-Posteriori optimization. Vitally, our method is different as it is amortized, eliminating network retraining or costly optimization at inference time. Additionally, our method requires few online gradients (3-4), compared to the numerous ones needed by DEEPGEM, thereby significantly reducing compute.

4.3 Method

To close the amortization gap, we describe an iterative approach to posterior inference where learned physics-based summary statistics are refined with Conditional Normalizing Flows (CNFs). VI with CNFs is reviewed first. Its improvement with learned physics-based summary statistics is discussed next, including addition of the crucial refinement step.

4.3.1 Amortized variational inference with conditional normalizing flows

While our method can, in principle, be applied to any generative model (GAN, diffusion, VAE), we focus on normalizing flows [48]. Thanks to their simple maximum-likelihood training objective, low training memory requirements [33], and fast sampling, CNFs have become one of the generative methods of choice when inverse problems are concerned. CNFs learn to sample from a target distribution by learning invertible transformations from the target distribution to the standard Normal distribution. By taking advantage of the change of variables formula [56], CNFs can be trained with a relatively simple objective:

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmin}} \frac{1}{N}\sum_{n=0}^{N}\left(\frac{1}{2}\|f_{\boldsymbol{\theta}}(\mathbf{x}^{(n)};\mathbf{y}^{(n)})\|_2^2 - \log\left|\det \mathbf{J}_{f_{\boldsymbol{\theta}}}\right|\right),$$

where f_θ is the CNF implemented as a network that takes a training pair $\mathbf{x}$ and $\mathbf{y}$ as input, has output in the same dimension as the target unknown $\mathbf{x}$, is invertible with respect to the input $\mathbf{x}$

$$f_{\boldsymbol{\theta}}^{-1}(f_{\boldsymbol{\theta}}(\mathbf{x};\mathbf{y});\mathbf{y}) = \mathbf{x} \;\; \forall \boldsymbol{\theta} \in \Theta$$

and has a Jacobian, $\mathbf{J}_{f_{\boldsymbol{\theta}}}$, that is tractable to compute. Designing invertible architectures with these characteristics is the focus of many works in the literature [97, 96]. Please refer to [21], for a derivation of the training objective subsection 7.2.1 from the amortized posterior objective in Equation Equation 1.3. The training pairs, collected in $\mathcal{D} = \{\mathbf{x}^{(n)},\, \mathbf{y}^{(n)}\}_{n=0}^{N}$, are generated by sampling from the prior, $\{\mathbf{x}^{(n)}\}_{n=0}^{N} \sim p(\mathbf{x})$, followed by a forward simulation. After the optimization is completed, the CNF with optimized weights, $\widehat{\boldsymbol{\theta}}$, can be used to sample from the approximate posterior $p_{\widehat{\boldsymbol{\theta}}}(\mathbf{x} \mid \mathbf{y}^{\mathrm{obs}})$ by sampling Gaussian noise and passing it through the inverse network that is conditioned on an observation $\mathbf{y}^{\mathrm{obs}}$.

$$\mathbf{x} \sim p_{\widehat{\boldsymbol{\theta}}}(\mathbf{x} \mid \mathbf{y}^{\mathrm{obs}}) = f_{\widehat{\boldsymbol{\theta}}}^{-1}(\mathbf{z};\mathbf{y}^{\mathrm{obs}}) \text{ where } \mathbf{z} \sim \mathcal{N}(0,\, I).$$

Contrary to non-amortized VI, the above posterior sample generation holds for any observation, $\mathbf{y}^{\mathrm{obs}}$, as long as the observations remain close—i.e., the $\mathbf{y}^{\mathrm{obs}}$ are produced by applying the forward operator to samples of the prior, $\mathbf{x} \sim p(\mathbf{x})$. From posterior samples, Monte Carlo estimates of the posterior statistics can be calculated. As short hand, statistic estimates from the distribution p are:

$$\mathbb{E}\,p := \mathbb{E}_{p(\mathbf{x}|\mathbf{y})}\left[\mathbf{x}\right] \qquad \sqrt{\mathbb{V}}\,p := \sqrt{\mathbb{E}_{p(\mathbf{x}|\mathbf{y})}\left[\left(\mathbf{x} - \mathbb{E}_{p(\mathbf{x}|\mathbf{y})}\right)^{2}\right]}.$$

For example, the posterior mean calculated from samples from the above trained approximate posterior $p_{\widehat{\boldsymbol{\theta}}}$ is referred to as $\mathbb{E}\, p_{\widehat{\boldsymbol{\theta}}}$. However, the quality of the posterior approximation, and therefore the quality of its samples— $\mathbf{x} \sim p_{\widehat{\boldsymbol{\theta}}}(\mathbf{x} \mid \mathbf{y}^{\mathrm{obs}})$, depends on the complexity of the posterior that is being approximated. This in turn depends on the complexity of prior samples and the likelihood. To account for realistic situations where both the prior and likelihood are complicated, CNFs demand increases in the size of the training set and the expressive power of the specific architecture used to define the CNF $f_{\theta}(\cdot)$, a requirement we like to avoid. For this reason, we will introduce the concept of summary statistics that allows us to improve the quality of the posterior approximation in these situations.

4.3.2 Gradient-based summary statistics

While CNFs are in principle capable of capturing complex data-to-image space mappings, amortization can be challenging to achieve in situations where the mapping is complex, or the observed data are heterogeneous—i.e., the observed data differ in dimension. To overcome these challenges, statisticians introduced so-called summary statistic. These often hand-derived summary statistics are designed to capture the main features in the data, reduce and homogenize its dimensionality, while posterior distributions remain informed [31, 120]. For posterior distributions to remain informed, we mean that the conditioning by the summary statistic, denoted by $\overline{\mathbf{y}}$, minimally changes the original conditional distribution—i.e., we have

$$p(\mathbf{x}|\overline{\mathbf{y}}) \approx p(\mathbf{x}|\mathbf{y}).$$

When this approximate equality holds exactly, the summary statistic is known to be a sufficient summary statistic. Since equality can not always be met, there also exists the notion of being close to sufficient. To be more specific, a summary statistic remains maximally informative [62] with respect to a set of summary statistics $\mathbf{y}' \in \mathbf{Y}$ when some distance measure between the summarized and original posterior distributions is minimized. For a distance measure given by the KL divergence, this amounts to finding the minimum of the following objective:

$$\bar{\mathbf{y}} = \underset{\mathbf{y}' \in \mathbf{Y}}{\operatorname{argmin}} \, \mathbb{KL} \left(p(\mathbf{x} \mid \mathbf{y}') \mid\mid p(\mathbf{x} \mid \mathbf{y}) \right).$$

Alternatively, one can also measure the informativeness of a certain summary statistic by its Fisher information [121]. The Fisher information matrix corresponds to the expected variance of the gradient of the data $\log$-likelihood

$$\mathcal{I}(\mathbf{x}) = \mathbb{E} \left[\left(\nabla_{\mathbf{x}} \log p(\mathbf{y} \mid \mathbf{x}) \right) \left(\nabla_{\mathbf{x}} \log p(\mathbf{y} \mid \mathbf{x}) \right)^{\top} \right],$$

and says how much information $\mathbf{y}$ contains of $\mathbf{x}$. We can relate this measure of information to summary statistics with the information inequality

$$\mathbb{V}(\bar{\mathbf{y}}) \geq (\nabla_{\mathbf{x}} \mathbb{E}(\bar{\mathbf{y}}))^{\top} \mathcal{I}(\mathbf{x}) (\nabla_{\mathbf{x}} \mathbb{E}(\bar{\mathbf{y}}))$$

which provides a bound on how much information a summary statistic can contain about the unknown $\mathbf{x}$ [122]. Note that in subsection 4.3.2 the covariance $\mathbb{V}$ and expectations $\mathbb{E}$ are taken with respect to the unknown $\mathbf{x}$.

In attempt to maximize informativeness as measured by the Fisher information, [100] proposed the gradient of the $\log$-likelihood $\nabla_{\mathbf{x}} \log p(\mathbf{y} \mid \mathbf{x})$ as a maximally informative summary statistic. Under the mild assumption that the gradient is defined for all $\mathbf{y}$ in the support of the likelihood, it can be shown that the information inequality subsection 4.3.2 is saturated by the gradient of the $\log$-likelihood, in other words that no other summary statistic can have more Fisher information [100]. In general, this summary statistic is defined as the gradient of the $\log$-likelihood calculated at a fiducial point $\mathbf{x}_0$

$$\overline{\mathbf{y}}_0 = \nabla_{\mathbf{x}} \log p(\mathbf{y} \mid \mathbf{x})\Big|_{\mathbf{x}_0}.$$

The subscript 0 is added in our notation to indicate that the summary statistic, $\overline{\mathbf{y}}_0$, derives from the evaluation of the gradient at the fiducial point $\mathbf{x}_0$. This fiducial point represents a trusted guess of the unknown parameter. The gradient of the $\log$-likelihood is an attractive summary statistic because it allows for the inclusion of knowledge on the forward operator, $\mathcal{F}$, and its Jacobian, $\nabla\mathcal{F}$. For instance, if the noise is additive Gaussian with covariance $\mathbf{C}_{\varepsilon}$ then the summary statistic is given by the action of the adjoint of the Jacobian, $\nabla\mathcal{F}^{\top}[\mathbf{x}_0]$ on the residual $\overline{\mathbf{y}}_0 = \nabla\mathcal{F}^{\top}[\mathbf{x}_0]\mathbf{C}_{\varepsilon}^{-1}(\mathcal{F}(\mathbf{x}_0) - \mathbf{y})$.

To train a CNF with this gradient summary statistic, the objective of subsection 7.2.1 is minimized on a new training set obtained by applying subsection 5.5.3 to the observations collected in $\mathcal{D}$, yielding $\mathcal{D}_0 = \{\mathbf{x}^{(n)}, \overline{\mathbf{y}}_0^{(n)}\}_{n=0}^{N}$ where each summary statistic $\overline{\mathbf{y}}_0^{(n)}$ is derived from the original simulated observation $\mathbf{y}^{(n)}$ and a chosen fiducial point $\mathbf{x}_0^{(n)}$.

The quality of gradient summary statistics is contingent on two key factors, namely the quality of the assumed likelihood and the quality of the fiducial points. The quality of the former depends on choices for the noise distribution and the forward operator, $\mathcal{F}$. When misspecified, or poorly calibrated, these choices may affect the quality of the summary statistic. We assume in this work that the forward problem and noise model are well specified. The second factor that determines the quality concerns choices for the fiducial points themselves. Because the fiducial point is used to calculate the gradient then its choice correlates with the information content of the resulting gradient. Thus, the choice for these fiducial points determines the quality of the summary statistic, which in turn determines the quality of the posterior inference itself. The quality of a fiducial point depends on if it is close to the maximum likelihood:

$$\mathbf{x}_{\mathrm{ML}} = \underset{\mathbf{x}}{\operatorname{argmax}}\, p(\mathbf{y} \mid \mathbf{x})$$

the estimator that aims to solely fit the data by maximizing the likelihood of the unknown under the data likelihood. As shown by [100], a fiducial point that is close to the maximum likelihood leads to a gradient-based summary statistic that is maximally informative with respect to the Fisher information. However, in the situation where the fiducial points are not close to the maximum likelihood this property may no longer hold, rendering gradient-based summary statistics less informative. As in many practical situations, we unfortunately do not always have access to high-quality fiducial points, a situation we will remedy in the next section where data-driven learning will be combined with gradient-based summaries.

4.3.3 Refining summary statistics

Notwithstanding the fact that gradient-based summary statistics represent a natural approach to inference problems that involve well-understood physics, the reliance on good fiducial points—i.e., fiducial points that are close to their respective maximum likelihoods, remains problematic and must be discussed. Instead of performing expensive, and potentially local-minima prone [123], Gauss-Newton updates to bring the fiducial points closer to the maximum likelihood points, as suggested by [124], we propose an iterative scheme during which CNFs are trained then sampled to improved fiducial points.

The iterations, outlined in algorithm 4.3.3, proceed as follows: given the current iterate for the fiducial points, this would be $\{\mathbf{x}_j^{(n)}\}_{n=0}^N$ at the j-th iteration, gradient-based summary statistics $\{\overline{\mathbf{y}}_j^{(n)}\}_{n=0}^N$ are computed with subsection 5.5.3 and a paired dataset is made yielding $\mathcal{D}_j = \{\mathbf{x}^{(n)}, \overline{\mathbf{y}}_j^{(n)}\}_{n=0}^N$. Then a CNF is trained on $\mathcal{D}_j$ by minimizing subsection 7.2.1. This minimization produces an optimized CNF, $f_{\widehat{\boldsymbol{\theta}}_j}$, which is used to draw multiple samples from the posterior, via subsection 5.5.2. Next, these posterior samples are av-

Algorithm 2: ASPIRE Training

```
**Input**: Prior samples $\{\mathbf{x}^{(n)}\}_{n=0}^{N}$, likelihood simulator $p(\mathbf{y}|\mathbf{x})$, conditional sampler $f_\theta$, initial fiducials $\{\mathbf{x}_0^{(n)}\}_{n=0}^{N}$.
**for** $n = 0$ **to** $N$:
    Simulate observation: $\mathbf{y}^{(n)} = p(\mathbf{y}|\mathbf{x}^{(n)})$
    Calculate gradient at fiducial: $\overline{\mathbf{y}}_0^{(n)} = \nabla_{\mathbf{x}} \log p(\mathbf{y}^{(n)} \mid \mathbf{x})\Big|_{\mathbf{x}_0^{(n)}}$
    Add pair to dataset: $\mathcal{D}_0 = \{\overline{\mathbf{y}}_0^{(n)}, \mathbf{x}^{(n)}\}$
**for** $j = 0$ **to** $J$:
    Train conditional sampler: $f_{\widehat{\theta}_j} = \text{Train}(f_{\theta_j}, \mathcal{D}_j)$
    **for** $n = 0$ **to** $N$:
        Posterior mean as new fiducial: $\mathbf{x}_{j+1}^{(n)} = \mathbb{E}_{p_{\widehat{\theta}_j}(\mathbf{x}|\overline{\mathbf{y}}_j^{(n)})}[\mathbf{x}]$
        Calculate gradient at new fiducial: $\overline{\mathbf{y}}_{j+1}^{(n)} = \nabla_{\mathbf{x}} \log p(\mathbf{y}^{(n)} \mid \mathbf{x})\Big|_{\mathbf{x}_{j+1}^{(n)}}$
        Update dataset: $\mathcal{D}_{j+1} = \{\overline{\mathbf{y}}_{j+1}^{(n)}, \mathbf{x}^{(n)}\}$
**Output**: Trained samplers $f_{\hat{\theta}_0}, f_{\hat{\theta}_1}, \ldots, f_{\hat{\theta}_J}$.
```

eraged, $\{\mathbf{x}_{j+1}^{(n)} = \mathbb{E}_{p_{\widehat{\theta}_j}(\mathbf{x}|\overline{\mathbf{y}}_j^{(n)})}[\mathbf{x}]\}_{n=0}^{N}$, for each gradient-based summary statistic, separately. This averaging, which corresponds to approximating the posterior mean for each summary $\{\overline{\mathbf{y}}_j^{(n)}\}_{n=0}^{N}$, produces the next, and arguably improved set of fiducial points, $\{\mathbf{x}_{j+1}^{(n)}\}_{n=0}^{N}$. These new fiducial points are used to create a new set of gradient-based summary statistics that can be used to train the next CNF and the process is repeated J times. As long as the new set of fiducials points moves closer to the respective maximum likelihood points, the quality of the gradient-based summary statistic can be expected to improve [100]. These improvements in turn produce better posterior inferences by the CNFs. The above iterative scheme hinges on the assumption that the fiducial points refined by the posterior mean indeed improve—i.e., they are closer to the maximum likelihood points. To motivate why we expect the approximated posterior mean to represent a better fiducial point, we show that the true posterior mean would indeed be a better fiducial point in a linear inverse problem.

Theorem 1. *For a linear inverse problem with forward operator* $\mathbf{A} \in \mathbb{R}^{m \times n}$*, an unknown* $\mathbf{x}$ *with Gaussian prior* $\mathcal{N}(\mu_{\mathbf{x}}, \mathbf{C}_{\mathbf{x}})$*, and additive Gaussian noise* $\mathcal{N}(0, \mathbf{C}_{\boldsymbol{\varepsilon}})$*, if the* ℓ_2 *norm between the current fiducial* $\mathbf{x}_0$ *and the maximum likelihood estimate* $\|\mathbf{x}_0 - \mathbf{x}_{ML}\|_2$ *is at* *least*

$$K \cdot \|\mu_{\mathbf{x}} - \mathbf{x}_{ML}\|_2 \leq \|\mathbf{x}_0 - \mathbf{x}_{ML}\|_2,$$
$$\textit{where } K = \left\| \left(\mathbf{C}_{\mathbf{x}}^{-1} + \mathbf{A}^{\top} \mathbf{C}_{\varepsilon}^{-1} \mathbf{A} \right)^{-1} \right\|_2 \cdot \left\| \mathbf{C}_{\mathbf{x}}^{-1} \right\|_2,$$

then forming the posterior with the gradient-based summary $p(\mathbf{x} \mid \overline{\mathbf{y}}_0)$ and using the posterior mean $\mathbf{x}_1 = \mathbb{E}_{p(\mathbf{x}|\overline{\mathbf{y}}_0)}[\mathbf{x}]$ as the next fiducial will yield an estimate with a smaller ℓ_2 norm distance to the maximum likelihood estimate

$$\|\mathbf{x}_1 - \mathbf{x}_{ML}\|_2 \leq \|\mathbf{x}_0 - \mathbf{x}_{ML}\|_2.$$

We outline a proof of Theorem 1 in section .1, where we derive the closed-form expression for the summarized posterior and compare its mean with the closed-form expression for the maximum likelihood. Note that the constant K depends on the covariance of both the posterior and the prior, highlighting the conditions under which it is appropriate to use the posterior mean as the next fiducial point. Specifically, the matrix norm of the posterior covariance should not be too large relative to that of the prior covariance; otherwise, the posterior mean is not a suitable candidate for the next fiducial point. Assuming that the CNFs used in ASPIRE are properly trained and accurately approximate the posterior, Theorem 1 demonstrates the desired behavior of improving the starting fiducial by using the posterior mean. To support the feasibility of accurate posterior approximation, we refer to the work of [115, 116], which proves the universality of normalizing flows for approximating conditional distributions. A detailed discussion on the theoretical behavior of ASPIRE for nonlinear operators is beyond the scope of this paper, but our numerical results confirm its empirical performance on both linear and nonlinear inverse problems.

As motivated by Theorem 1, initial (potentially poor) summary statistics computed for the fiducials points, $\{\mathbf{x}_0^{(n)}\}_{n=0}^{N}$, can be improved by training a CNF to generate new fiducial points, $\{\mathbf{x}_1^{(n)}\}_{n=0}^{N}$. Since these new fiducial points are closer to their corresponding maximum likelihood then the new summary statistics will be more informative, improving inference during the next iteration where the CNF is trained on improved summary statistics.

After training is completed, we obtain $J + 1$ trained CNFs, each with their own set of optimized weights, $\{\widehat{\theta}_j\}_{j=0}^{J}$. Because these networks are trained on the datasets $\mathcal{D}_j = \{(\mathbf{x}_j^{(n)}, \overline{\mathbf{y}}_j^{(n)})\}_{n=0}^{N}$ for $j = 0 \cdots J$, these networks have been generalized to perform J refinements, given a new unseen observation, $\mathbf{y}^{\text{obs}}$. Pseudocode for the online inference phase is included in algorithm 4.3.3. Note that each refinement incurs the cost of a gradient calculation. In practice, $J = 3$ to 4 refinements are often adequate resulting in a total online computational cost that is significantly lower than non-amortized inference, which can result in 10000's of gradients [53].

Algorithm 3: ASPIRE Inference

```
Input: Field observation $\mathbf{y}^{\text{obs}}$, trained conditional samplers $f_{\hat{\theta}_j}$, and initial fiducial $\mathbf{x}_0$
for $j = 0$ to $J - 1$:
    $\overline{\mathbf{y}}_j = \nabla_{\mathbf{x}} \log p(\mathbf{y}^{\text{obs}} \mid \mathbf{x})\Big|_{\mathbf{x}_j}$
    $\mathbf{x}_{j+1} = \mathbb{E}_{p_{\hat{\theta}_j}(\mathbf{x}|\overline{\mathbf{y}}_j)}[\mathbf{x}]$
$\overline{\mathbf{y}}_J = \nabla_{\mathbf{x}} \log p(\mathbf{y}^{\text{obs}} \mid \mathbf{x})\Big|_{\mathbf{x}_J}$
Output: $p_{\hat{\theta}_J}(\mathbf{x} \mid \overline{\mathbf{y}}_J)$, Posterior sampler for observation
```

In summary, by pairing the theory of [100] with Theorem 1, we arrived at a formulation where the refined fiducial points yield improved summary statistics and refine amortized VI at limited additional online computational costs. To verify this claim, we will first evaluate our method on a stylized example for which the analytical posterior is known. This example will show that the improvements thanks to refined summary statistics indeed converge to the correct posterior distribution. To demonstrate our amortized VI in a more practical setting, we will also evaluate its performance on a realistic ultrasound transcranial medical imaging problem.

4.3.4 Stylized example

To build trust in our method, we first demonstrate it improves the quality of the posterior approximation by testing on an inverse problem with a known posterior distribution. One such inverse problem is the linear Gaussian inverse problem where: the forward operator $\mathbf{A}$ is a known matrix $\mathbb{R}^{m \times n}$, the prior and noise comes from Gaussian distributions with known means and covariances. We chose the unknown parameter vector to have size $n = 16$ and the data as size $m = 80$. Given these settings, it is possible to calculate samples from the analytical posterior distribution [13]. We use samples from the analytical posterior distribution to compare against our posterior sampling results with ASPIRE.

We train our method using $N = 1000$ samples from the Gaussian prior and use the forward operator to form training pairs. Since the forward operator is linear, the gradient-based summary statistic is calculated from the transposed operator. After training, we evaluate our method on an unseen observation, $\mathbf{y}^{\text{obs}}$, simulated from a known ground-truth parameter $\mathbf{x}^*$. We first observe that after each ASPIRE iteration the posterior mean $\mathbb{E}\, p_{\widehat{\boldsymbol{\theta}}_j}$ (using short hand described in subsection 4.3.1) becomes a better reconstruction of the ground truth $\mathbf{x}^*$ as seen in Figure 4.2.

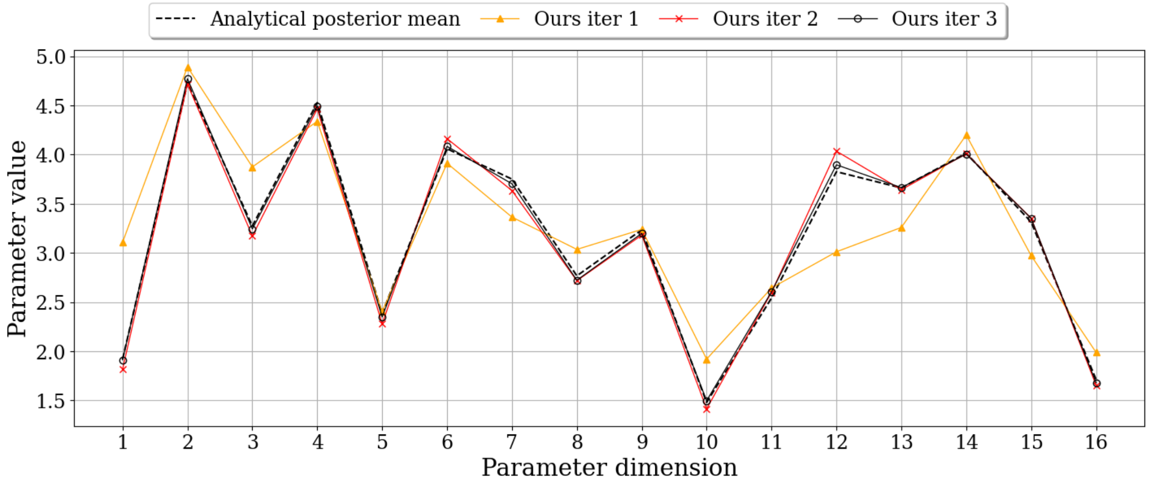


Figure 4.2: The quality of the proposed amortized posterior approximation improves at each iteration as measured by the estimated posterior mean with respect to the analytically known ground truth posterior mean.

Furthermore, Figure 4.3 shows the empirical covariance derived from our method's posterior samples, compared to the analytically calculated covariance matrix. It is clear that each iteration improves the approximation of the estimated covariance and it is almost exactly correct at the third iteration, $J = 3$. While this stylized example confirms that amortized inference with ASPIRE is feasible, the real challenge is to apply this concept to medical ultrasound where problems are high dimensional and forward modeling is computationally expensive to evaluate.

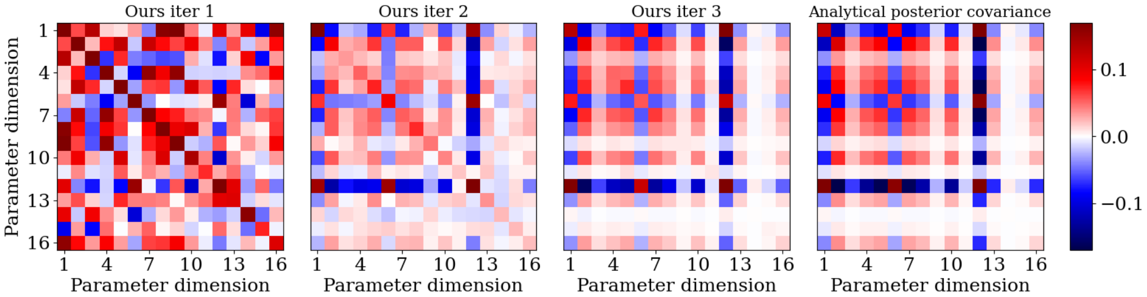


Figure 4.3: Comparison of full covariance matrix from our method as compared to the analytical ground truth posterior covariance. After three iterations of our method, the estimated posterior covariance is close to the ground truth covariance.

4.4 Medical wave-based imaging

Transcranial Ultrasound Computed Tomography (TUCT) is a non-ionizing, non-radiative imaging modality that creates images of brain tissue from measurements of impinging ultrasound waves due to contrast in tissue acoustic properties. Unlike other ultrasound imaging targets, like breast imaging [125, 126], TUCT faces the challenge of high acoustic contrast in the cranial bone, leading to scattering unsuitable for traditional traveltime tomography methods [84]. Tomographic methods require high frequencies for higher resolution imaging, but attenuation through the skull is exacerbated at higher frequencies thus preventing high-resolution imaging when relying on traveltime methods. This challenge has hindered ultrasound's application to brain imaging until recent developments when [85] identified that similar challenges exist in transcranial ultrasound imaging as with subsalt imaging used by exploration seismology. The main reason seismic techniques are capable of imaging through high-acoustic contrast salt is because these sophisticated inversion methods model the full physics of the wave equation to make sense of the scattered waves. Whereas traditional ultrasound only uses arrival times, seismic imaging techniques model all waveforms allowing for higher effective resolutions at lower frequencies that experience less attenuation through the skull. These methods are denoted Full-Waveform Inversion (FWI) since they model and match the full observed waveform, see Figure 4.4c for an example of the full waveform.

4.4.1 Medical ultrasound with full-waveform inversion

Since the groundbreaking work of [85], FWI techniques for TUCT are showing promise as a high-resolution imaging modality with potential clinical applications ranging from early hemorrhage diagnosis to tumor imaging [127, 128]. The TUCT inverse problem involves reconstruction of the acoustic velocity, $\mathbf{x}$, of brain tissue from acoustic data, $\mathbf{y}$, collected as shown in Figure 4.4a. In this setup, ultrasound transducers placed around the patient's head perform multiple experiments, with each involving a tone-burst transmission by one transducer and recording by all others, as simulated in Figure 4.4c. Experiments proceed by transducers transmitting from different positions until all transducers have transmitted, providing a full coverage from many angles. The forward operator $\mathcal{F}$ that maps the acoustic parameters to the observed data is simulated with the numerical solution of the second-order wave equation with varying acoustic velocity in two space dimensions:

$$\frac{1}{c(x,y)^2}\frac{\partial^2}{\partial t^2}u(x,y,t) - \nabla^2 u(x,y,t) = q_s(x,y,t).$$

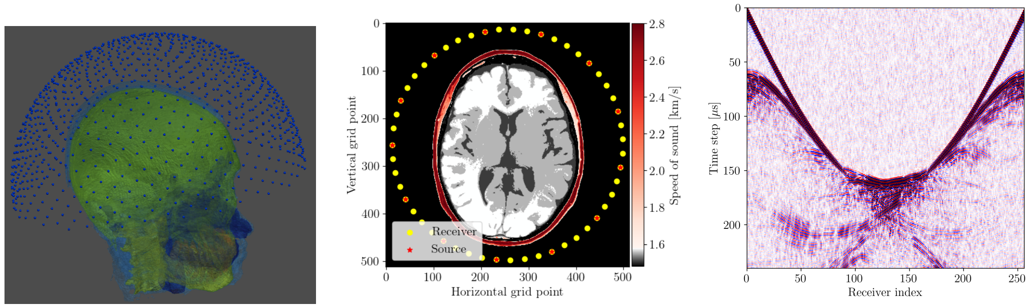
where the acoustic velocity $c(x,y)$ is parameterized by a gridded array of values in the unknown vector $\mathbf{x}$ and the transducers are modeled by N_s different source terms $\{q_s(x,y,t)\}_{s=0}^{s=N_s}$. Although the forward operator $\mathcal{F}$ is technically defined for each source as $\mathcal{F}(\mathbf{x}; q_s)$, for simplicity, we denote it as $\mathcal{F}(\mathbf{x})$, representing the collection of PDE solutions for all transducer sources that contribute to all the observations collected in $\mathbf{y}$ and the restriction of the solution wavefields to receiver positions. Given the set of observations, traditional FWI workflows setup the variational problem

$$\underset{\mathbf{x}}{\text{minimize}} \; \|\mathcal{F}(\mathbf{x}) - \mathbf{y}\|_2^2$$

and minimize this data-misfit objective with stochastic gradient descent by using randomized subsets of the sources to calculate the gradient of each descent step beginning from the starting parameter vector, $\mathbf{x_0}$. This, of course, assumes access to an efficient routine for calculating the gradient of $\mathcal{F}$, which we will discuss further in subsubsection 4.4.2. Under controlled assumptions, such as a good starting parameter $\mathbf{x_0}$ and calibrated transducers [129], FWI is known to produce high-resolution images [130, 85]. However, clinical adoption of FWI is hampered by the prohibitively long runtime of full physics modeling and the parasitic local minima related to the non-convex optimization [87]. Previous literature has explored the regularization of the FWI optimization with handcrafted priors such as the Total-Variation norm [131, 132] and model extensions [133, 123]. Our approach, ASPIRE, addresses these issues by reducing physics computations, giving data-driven regularization of the non-convex optimization, and providing uncertainty-aware solutions crucial for clinical applications by sampling the Bayesian posterior.

4.4.2 Transcranial Ultrasound Computed Tomography with ASPIRE

To abridge, we will use ASPIRE to solve TUCT by generating samples from a realistic brain velocity prior $\mathbf{x} \sim p(\mathbf{x})$, use a wave PDE solver $\mathcal{F}$ to simulate acoustic data $\mathbf{y}$ and use the gradient of the simulator at a fiducial point $\mathbf{x}_0$ as the gradient-based summary statistic for training a CNF with subsection 7.2.1. Our test results demonstrate that iterative refinement of this summary statistic through ASPIRE significantly enhances the accuracy of the posterior approximations. Now, we detail how each of these experiments and tests



(a) 3D setup.

(b) Acoustic velocity $\mathbf{x}$.

(c) Observation $\mathbf{y}$.

Figure 4.4: Experimental setup: (a) Transcranial ultrasound 3D setup as used in field, blue dots indicate transducers. (b) Transcranial ultrasound 2D synthetic experimental setup used in this work. (c) Simulated waveform from a single source synthetic experiment $\mathbf{y}$. Each column corresponds to acoustic-pressure amplitudes measured at one transducer for a single experiment.

are implemented.

Brain prior samples

The first step of implementing ASPIRE concerns obtaining samples from a realistic prior for the target parameter vector $\mathbf{x} \sim p(\mathbf{x})$, in this case, gridded velocity parameters of human brains and skulls. The parameters collected in the MIDA dataset [134] correspond to a single 3D volume for the acoustic velocity collected from a single subject and will unfortunately not be appropriate to train a neural model that will generalize to other human patients. As far as we know, there is no dataset that includes acoustic velocity collected from many patients, so we made our own dataset based off the multi-subject FASTMRI dataset [103]. This custom dataset, detailed in Appendix section .2, comprises N=1000 diverse acoustic velocity parameters collected from different human patients, $\{\mathbf{x}^{(n)}\}_{n=0}^{1000}$. This size of datasets facilitates generalization of the amortized posterior sampler across different datasets collected from unseen patients. The dataset is accessible via the GitHub ASPIRE.jl.

Wave simulations

Our synthetic TUCT experiment, based on the configuration from [85], models the unknown parameter as discretized acoustic velocity on a 512×512 grid, with a $0.5\,[\mathrm{mm}]$ discretization. We modeled transducer sources as point sources with a three-cycle tone-burst signature with central frequency of $400\mathrm{Khz}$ and 240 [microseconds] recording time. The transducers are placed in a circular arrangement around the skull, the setup, with 16 sources and 256 receivers, mimics a 2D slice of the 3D experiment shown in Figure 4.4b. The forward operator, $\mathcal{F}(\mathbf{x})$, corresponds simulating the forward waveforms and their restriction to the receiver locations. The wave equation and its Jacobian were solved using the open-source software packages Devito and JUDI [105, 104, 7], which automatically generate optimized C code and leverage GPU accelerators, thereby facilitating scalability to realistic problem sizes. To simulate noise corruption, we used additive Gaussian noise, ε, with a $35\mathrm{dB}$ Signal-Noise-Ratio, matching lab values [85]. A synthetic observation, $\mathbf{y} = \mathcal{F}(\mathbf{x}) + \varepsilon$, is displayed in Figure 4.4c.

TUCT summary statistic:

The gradient-based summary statistic $\overline{\mathbf{y}}$, is calculated as in subsection 5.5.3, which requires the action of the Jacobian adjoint on the data residual at the fiducial point, $\mathbf{x}_0$. For computational efficiency, the adjoint-state method [135, 91] is used wherein only two PDE solves are required to calculate the gradient. To avoid the inverse crime (refers to using the same model to generate observed data and to invert it, which can lead to overly optimistic results), the observed data is simulated with finer time discretization and a higher-order spatial finite-difference stencil than those used in the residual calculation and adjoint simulation. Each transducer defines a source term in subsection 4.4.1, so we sum the gradient over all 16 sources into the final summary statistic.

4.4.3 Traditional amortized inference

To illustrate the limitations of amortized VI, we train a CNF on pairs $\{(\mathbf{x}^{(n)}, \mathbf{y}^{(n)})\}_{n=0}^{N}$ without evoking iterative improvements by ASPIRE. We emphasize that the observations $\mathbf{y}$ are the raw unsummarized waveforms similar to that shown in Figure 4.4c. After training by minimizing subsection 7.2.1, the CNF with weights $\widehat{\boldsymbol{\theta}}$, provides an amortized approximation of the posterior, $p_{\widehat{\boldsymbol{\theta}}} \approx p(\mathbf{x} \mid \mathbf{y})$, from which we can sample (cf. subsection 5.5.2). The results, shown in Figure 4.5, demonstrate that the samples from $p_{\widehat{\boldsymbol{\theta}}}$, for an unseen test observation, $\mathbf{y}^{\text{obs}}$, lack distinct features beyond an unrealistic skull and unresolved internal tissue structure. A comparison of these samples and the posterior mean, $\mathbb{E}\, p_{\widehat{\boldsymbol{\theta}}}$, in Figure 4.5c with the ground truth, Figure 4.5d, highlights the poor quality of this approximation. Note, throughout this exposition we calculate the posterior statistics (i.e. mean and standard deviation) over 512 samples. This experiment underscores the challenge of directly learning the probabilistic inverse mapping from the acoustic data $\mathbf{y}$ to the velocity parameters, a difficulty previously noted in the literature [39, 136]. We address this problem with the gradient-based summary statistic employed by ASPIRE.

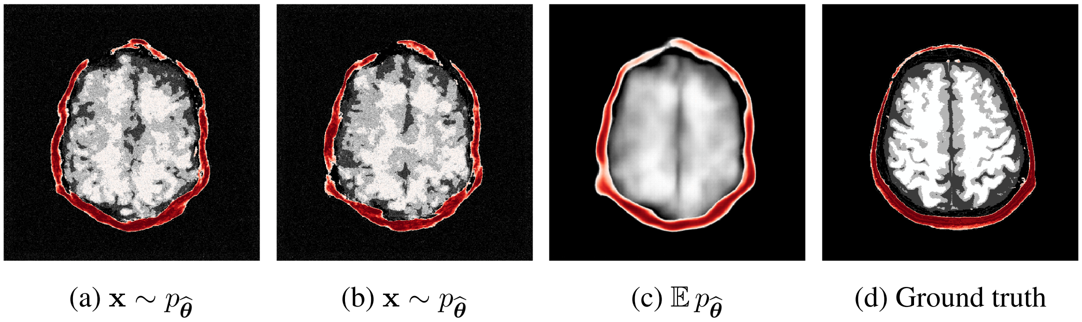
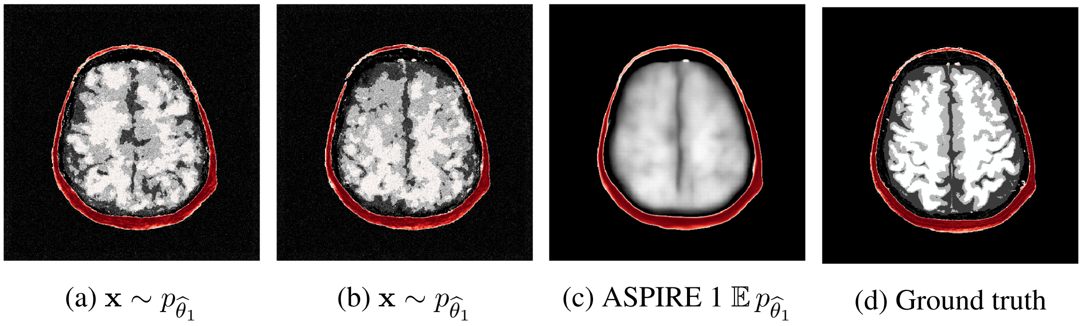

(a) $\mathbf{x} \sim p_{\widehat{\boldsymbol{\theta}}}$ (b) $\mathbf{x} \sim p_{\widehat{\boldsymbol{\theta}}}$ (c) $\mathbb{E}\, p_{\widehat{\boldsymbol{\theta}}}$ (d) Ground truth

Figure 4.5: Baseline amortized inference. (a),(b) Posterior samples. (c) Posterior mean. (d) Ground-truth velocity parameters paired to test observation $\mathbf{y}^{\text{obs}} = \mathcal{F}(\mathbf{x}^{*}) + \boldsymbol{\varepsilon}$. The samples have poor quality since it is difficult to learn the direct mapping from acoustic waveforms to the unknown parameter.

4.4.4 Amortized inference with gradient-based summary statistics

To overcome the end-to-end inference problem, we apply one iteration of ASPIRE by training a CNF on pairs, $\mathcal{D}_0 = \{\mathbf{x}^{(n)}, \overline{\mathbf{y}}_0^{(n)}\}_{n=0}^{N}$, where the $\overline{\mathbf{y}}_0^{(n)}$'s represent the gradient-based summary statistics at the fiducial points, $\mathbf{x}_0^{(n)}$, taken to be the uniform water velocity for all samples. An example of this initial summary statistic is shown in Figure 4.8a. While the outer edge of the skull is reasonably well delineated, the inner edge of the skull is still poorly resolved and details inside the skull are mostly absent. However, the inference based on these initial summary statistics, shown in Figure 4.6c, present a significant improvement over the baseline (cf. Figure 4.5c), despite the presence of strong imaging artifacts in the summary statistics. The improvements concern the skull's structure in particular, although details within the skull remain elusive due to the summary statistic's limited information. To enhance fidelity further, ASPIRE 2 (shorthand for ASPIRE at iteration $J = 2$) is applied by recalculating the gradient at the new posterior mean estimate for each training sample. Given these new training pairs, the next CNF is trained. While posterior sampling is efficient with CNFs (using subsection 5.5.2), recalculation of the gradient for each sample is computationally intensive, a topic we address in subsection 4.6.6.

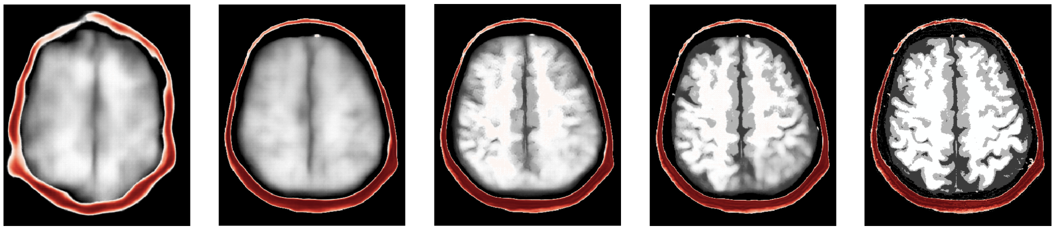


(a) $\mathbf{x} \sim p_{\widehat{\theta}_1}$ (b) $\mathbf{x} \sim p_{\widehat{\theta}_1}$ (c) ASPIRE 1 $\mathbb{E}\, p_{\widehat{\theta}_1}$ (d) Ground truth

Figure 4.6: The first iteration of our method learns the mapping from the summarized data $\overline{\mathbf{y}}$ to unknown parameter $\mathbf{x}$. (a),(b) Posterior samples. (c) Posterior mean. (d) Ground-truth velocity parameters. Our method has learned to reconstruct a reasonable estimate of the skull outline by making use of the summary statistic.

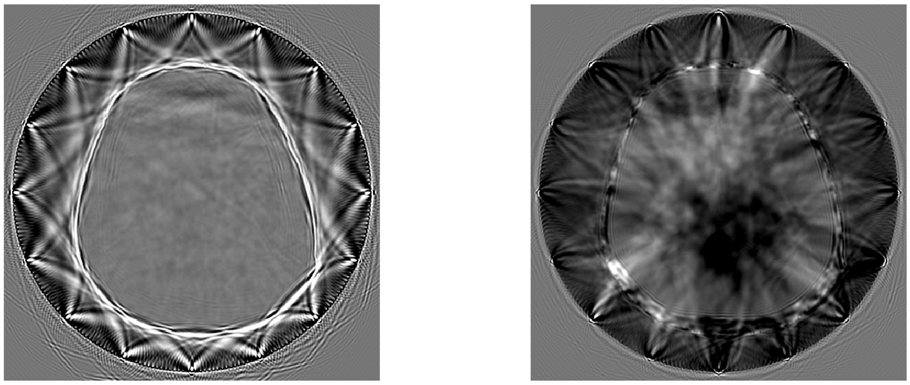
4.4.5 Amortized inference with iterative refinements

After the refinements of ASPIRE 2, significant improvements are evident in the posterior samples, particularly in capturing the structures within the brain tissue itself. The mean of these posterior samples, displayed in Figure 4.7c, is clearly enhanced in resolution and details. We attribute these enhancements to the increased informativeness of the summary statistic in the second iteration compared to the information yielded by the initial iteration. A detailed inspection of the second summary statistic (shown in Figure 4.8b) reveals more detail on the internal brain structures. Unlike the first summary statistic (cf. Figure 4.8a), which primarily delineated the skull, the second iteration's summary statistic better 'illuminates' the softer tissues within the brain, offering a more informative image for the posterior network. Thanks to accounting for the scattering at the skull, the acoustic illumination of the brain is improved significantly. Accurately resolving the skull structure is an important consideration as noted by [137].



(a) No summary (b) ASPIRE 1 (c) ASPIRE 2 (d) ASPIRE 4 (e) Ground truth

Figure 4.7: The posterior approximation improves as measured by the posterior mean quality. (a) Posterior mean without use of any summary statistics. (b) Posterior mean from ASPIRE 1 where the observation is preprocessed with the gradient as summary statistic. (c) Posterior mean from ASPIRE 2 where the summary statistic has been refined using the posterior mean from the first iteration. (d) Posterior mean from ASPIRE 4. (e) Ground-truth.



(a) Initial summary statistic $\overline{\mathbf{y}}_0$

(b) Refined summary statistic $\overline{\mathbf{y}}_1$

Figure 4.8: Gradient-based summary statistics. (a) First summary statistic calculated at the $\mathbf{x}_0$ fiducial consisting of constant water velocity. (b) Second summary statistic calculated at the fiducial point, $\mathbf{x}_1$, derived from the first CNF posterior mean. Thanks to the improved fiducial point, we can ”illuminate” the inside of the skull.

As one can observe from Figure 4.7, the reconstruction quality improves for increasing number of refinements of ASPIRE. By virtue of the iterative recalculation of the gradient-based summary statistic, the method is progressively able to discern finer details within the brain albeit the updates become less pronounced as the number of refinements increases. Practically, a user of ASPIRE can decide on the number of refinements based on the amount of compute available or by refining until there are diminished returns on enhancements.

4.4.6 Reconstruction quality

To quantitatively assess enhancements of each ASPIRE iteration, we compare the posterior means with their corresponding ground truths, using a test set comprising 50 unseen observations. By calculating the Root Mean Squared Error (RMSE) for these comparisons, we establish a metric to quantify the improvements across iterations. At each iteration we plot the RMSE for all examples in the test set and plot a box plot. We emphasize that testing on this many samples was only tractable since the posterior sampler is amortized. The trend, as showcased in Figure 4.9, confirms that on average each iteration reduces the RMSE, indicating an increasingly precise approximation of the true posterior means. We hypothesize that we are observing the effect known as Bayesian contraction [74], a phenomena in which increasing the amount of observations contracts the posterior towards the ground truth, since each summary statistic is extracting more information from the observations.

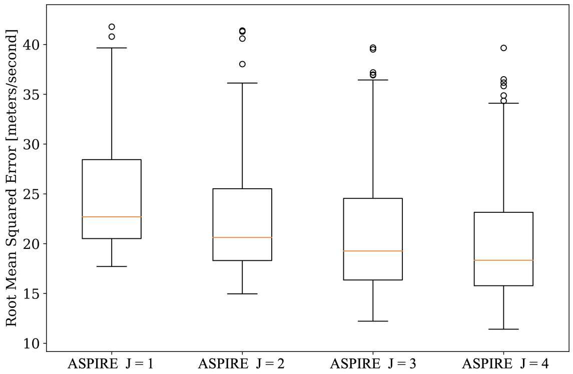


Figure 4.9: Image quality metric measured over a 50-sample leave-out test set. The quality of the posterior mean improves after each ASPIRE refinement.

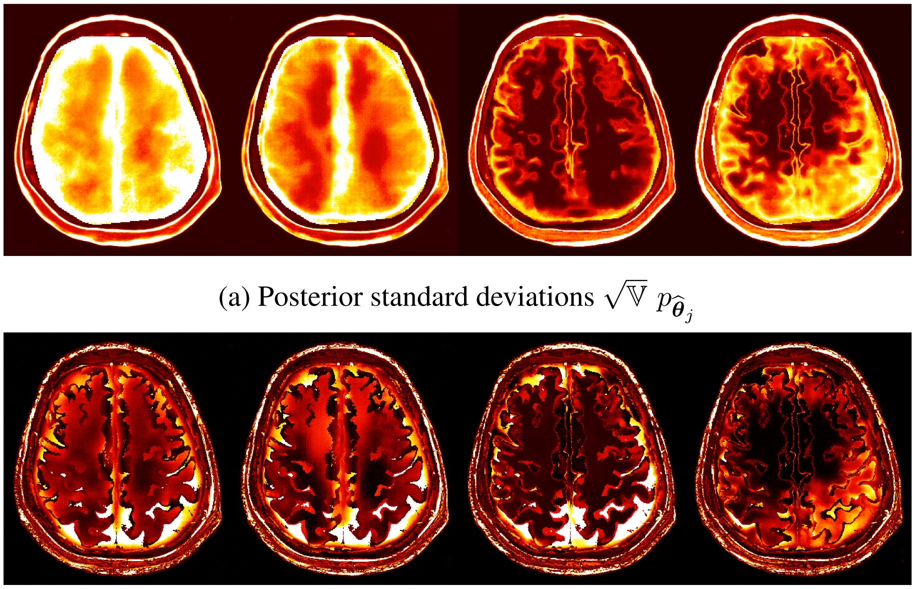
While the reconstructions in Figure 4.7 clearly improve as the number refinements increases, certain areas remain smooth, especially near the top and lower-right corner. This smoothing effect of the posterior mean is well-known and arises from relative strong variations among the samples in certain areas. Variations between samples are a reflection of inconsistent reconstructions by the posterior samples and correspond to areas of increased uncertainty. This phenomenon is a direct result of treating ultrasound medical imaging as an inference problem that produces posterior distributions instead of a single answer.

4.5 Uncertainty quantification

Due to the risk of hallucinations, generative AI for imaging inverse problems benefits from uncertainty awareness. Furthermore, an uncertainty-aware approach becomes crucial in medical applications, as underscored by [138]. Fortunately, Bayesian posterior sampling provides a natural sense of uncertainty, reflected in the spread of the samples. ASPIRE is designed with amortized posterior sampling in mind to quickly deliver crucial uncertainty quantification. By providing both the mean and a uncertainty information, our method offers a dual perspective, namely a robust reconstruction of the tissues, complemented by an insight into the statistical variations of each pixel's value. This dual analysis is particularly valuable in medical diagnostics, where understanding both the image and the associated uncertainty is crucial for decision making.

4.5.1 Amortized uncertainty quantification

To visualize uncertainty, we calculate uncertainty images by taking the pixel-wise standard deviations $\sqrt{\mathbb{V}}$ as defined in subsection 4.3.1 with 512 posterior samples. Figure 4.10 shows uncertainty images for four iterations of ASPIRE alongside the error of the posterior mean from the ground truth. From these figures, we make the following qualitative observations: *(1)* uncertainty images increase in resolution with each ASPIRE refinement *(2)* refinements increase correlations between the uncertainty and the error. Specifically, the errors concentrate near the top and lower-right of the internal brain tissue. The reason being that high acoustic contrast in these areas is creating multiple reverberations of the wavefield inside the brain impeding accurate imaging, importantly these are areas that are highlighted by the uncertainty. Correlations between the uncertainty and the error constitute important empirical evidence of the trustworthiness of the uncertainty. To more rigorously quantify this correlation, and quantitatively validate the uncertainty quantification, we study the calibration of our uncertainty in the following section.



(a) Posterior standard deviations $\sqrt{\mathbb{V}}\ p_{\widehat{\boldsymbol{\theta}}_j}$

(b) Error $|\mathbf{x}^* - \mathbb{E}\, p_{\widehat{\boldsymbol{\theta}}_j}|$

Figure 4.10: Posterior standard deviation compared to predictive error. (a) Posterior standard deviations of ASPIRE with increasing iterations 1 through 4 from left to right. (b) Same but with error of the posterior mean. Each ASPIRE refinement uncovers higher resolution details; furthermore, the apparent correlation between the uncertainty and error increases. All plots are shown on the same colorbar from $0\,[\mathrm{m/s}]$ to $50\,[\mathrm{m/s}]$.

4.5.2 Calibration of the uncertainty

To assess the calibration of our method's uncertainty quantification against errors, we employ the calibration test described by [79, 78]. This test involves comparisons between errors — defined by the Euclidean distance between the posterior mean estimates, $\widehat{\mathbf{x}} = \mathbb{E}\, p_{\widehat{\theta}}$, derived from samples of the posterior conditioned on the observations, $\mathbf{y}$, and the ground-truth parameters, $\mathbf{x}^*$ — and the inferred uncertainty in terms of the square-root of the posterior variance, $\widehat{\sigma} = \sqrt{\mathbb{V}} p_{\widehat{\boldsymbol{\theta}}}$. Given predictions, $\widehat{\mathbf{x}}$, derived from observations, $\mathbf{y}$, and a measure of uncertainty, $\widehat{\sigma}$, the calibration test seeks to verify the relationship:

$$\mathbb{E}_{\mathbf{x}^*,\mathbf{y}}\Big[\|\mathbf{x}^* - \widehat{\mathbf{x}}\|_2^2 \mid \widehat{\sigma} = \sigma\Big] = \sigma \quad \{\forall \sigma \in \mathbb{R} \mid \sigma \geq 0\}.$$

This expression implies that uncertainty is properly calibrated when the uncertainty is proportional to the error. For instance, if a set of gridpoints has an uncertainty of $10[\mathrm{m/s}]$, their expected error should be $10[\mathrm{m/s}]$. The calibration benchmark follows as such, first the set of uncertainty values for each pixel in $\widehat{\sigma} = \sqrt{\mathbb{V} p_{\widehat{\boldsymbol{\theta}}}}$ is categorized into K bins of equal width, the uncertainty at each bin B_k is calculated as:

$$UQ(B_k) := \frac{1}{|B_k|} \sum_{i \in B_k} \widehat{\sigma}_i$$

the average error (with $\widehat{\mathbf{x}} = \mathbb{E}\, p_{\widehat{\boldsymbol{\theta}}}$) is also calculated at the same bins:

$$Error(B_k) := \frac{1}{|B_k|} \sum_{i \in B_k} (\mathbf{x}_i^* - \widehat{\mathbf{x}}_i)^2.$$

The uncertainty $UQ(B_k)$ and error $Error(B_k)$ at each bin is then plotted against each other. If there is a high correlation between these values we expect the plot to match the $45°$-degree angle. For details on this test see [78].

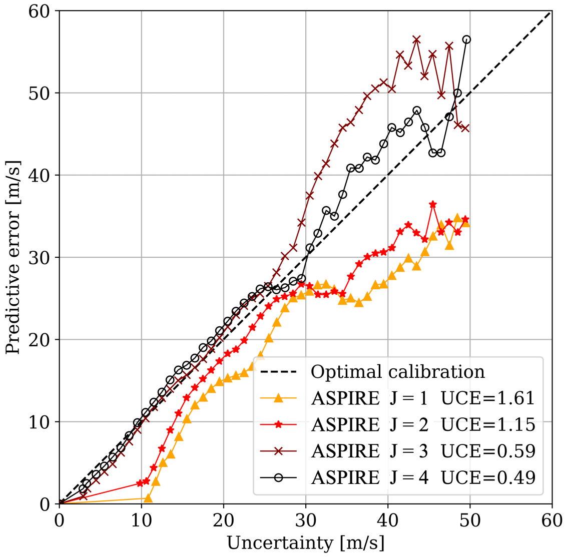


Figure 4.11: Calibration plot of four refinements from ASPIRE. The quality of uncertainty quantification of ASPIRE improves as measured by the calibration with respect to the error.

The resulting calibration curve from ASPIRE 4, included in Figure 4.11, exhibits the expected behavior by closely following the expected line. This means that our method is well calibrated for error magnitudes up to $30[\mathrm{m/s}]$. More importantly, each iterative refinement of ASPIRE improves the calibration. To quantify these improvements, we calculate the Uncertainty Calibration Error (UCE), which represents the average absolute difference between the predicted error and actual uncertainty across all bins. A lower UCE indicates a more precise calibration, and according to our experiment, ASPIRE manages to reduce the UCE by a factor of three from a value of 1.61 to 0.49. This significant improvement underscores the ability of our iterative refinement approach to produce reliable uncertainty estimates in complex imaging scenarios. By confirming the calibration of our uncertainty, we build trust in our method to properly inform downstream tasks that require access to the Bayesian posterior.

4.6 Discussion

Our main goal is to convince the reader of ASPIRE's ability to close the amortization gap with iterative refinements. We were able to show that significant improvements were achieved compared to the baseline (e.g. juxtapose Figure 4.7b and Figure 4.7d). To further substantiate our claims, we place our amortized method's performance in a broader context that includes comparisons to non-amortized inference. These comparisons illustrate our strides towards narrowing the gap between these two paradigms. Although a wide range of non-amortized VI techniques are available, our application in medical ultrasound presents unique requirements and challenges — absence of an analytical prior and the need for short time-to-solution challenged by the computational intensity of the forward/adjoint operators – limit our options for comparison. For this reason, we selected two methods: the mean-field approximation and a non-amortized normalizing flow. Through this comparative analysis, we aim to highlight our method as an efficient alternative to scenarios where non-amortized methods may be impeded by computational demands.

4.6.1 Mean-field approximation

Perhaps, the method with the most comparable aims in the current literature for uncertainty quantification in TUCT is the recent work by [90]. This approach employs a mean-field approximation of the posterior, and assumes the posterior covariance matrix to be diagonal for computational reasons. While understandable from a computational perspective, the established FWI literature cautions against this assumption because of the non-diagonal nature of the wave-equation Hessian [91], which introduces correlations in the errors. Our method does not make statistical assumptions on the prior or likelihood and is designed to capture the complete statistics of the posterior distribution, including long-range correlations in the covariance as evidenced in Figure 4.3.

Given a single set of observations, $\mathbf{y}^{\mathrm{obs}}$ the mean-field algorithm by [90] directly outputs estimates for the posterior mean, $\boldsymbol{\mu}_m$, and point-wise posterior standard deviation, $\boldsymbol{\sigma}_m$. This approach hinges on a relative simple modification of traditional FWI — it multiplies updates of gradient-descent with a Gaussian field. Because the computational cost of the mean-field approximation roughly correspond to that of traditional FWI that includes 100s of forward and adjoint calls, we argue that our method offers a distinct computational advantage at inference time. —i.e., ASPIRE achieves online inference at approximately 1/100th the online computational cost of the mean-field approximation. For a detailed discussion on computational costs, please refer to subsection 4.6.6.

4.6.2 Non-amortized inference with normalizing flows: WISER

We also compare with a non-amortized method deriving from the same prior knowledge in the form of training samples. Given the problem size and expensive to evaluate wave-physics based forward operators and gradients, these comparisons are made with respect to a novel computationally efficient inference method inspired by recent work of [21, 117]. Instead of starting from scratch with a non-informative prior [27], which proves to be computationally prohibitively expensive, the proposed approach optimizes network weights, ϕ, of a non-amortized normalizing flow (NF), $g_{\boldsymbol{\phi}}(\cdot)$ that acts in the latent space of a pre-trained amortized CNF, $f_{\widehat{\boldsymbol{\theta}}}(\cdot)$. Given a single set of observations, $\mathbf{y}^{\text{obs}}$, the objective reads:

$$
\begin{aligned}
\min_{\boldsymbol{\phi}} \; \mathbb{KL}\left(p\left(h_{\boldsymbol{\phi}}(\mathbf{z})\right) \| \, p(\mathbf{z} \mid \mathbf{y}^{\text{obs}})\right) = \; \mathbb{E}_{\mathbf{z}\sim\mathcal{N}(\mathbf{0},\mathbf{I})} & \Bigg[\frac{1}{2\sigma^2} \left\| \mathcal{F} \circ f_{\widehat{\boldsymbol{\theta}}}^{-1}\left(g_{\boldsymbol{\phi}}(\mathbf{z}); \overline{\mathbf{y}}^{\text{obs}}\right) - \mathbf{y}^{\text{obs}} \right\|_2^2 \\
& + \frac{1}{2} \left\| g_{\boldsymbol{\phi}}(\mathbf{z}) \right\|_2^2 - \log\left|\det \mathbf{J}_{g_{\boldsymbol{\phi}}}(\mathbf{z})\right| \Bigg] .
\end{aligned}
\tag{4.1}
$$

The $f_{\widehat{\boldsymbol{\theta}}}(\cdot)$ denotes the pre-trained CNF, optimized as per subsection 7.2.1 and $\left|\det \mathbf{J}_{g_{\boldsymbol{\phi}}}\right|$ is the determinant of the second network's Jacobian. By minimizing this objective, the network $g_{\boldsymbol{\phi}}(\cdot)$ is trained to generate latent codes that further minimize residuals in data-misfit objective, which involves the nonlinear forward operator, and a ℓ_2-norm penalty term, which ensures that the network output stays close to Gaussian distributed, therefore respecting the prior defined by the pre-trained network $f_{\widehat{\boldsymbol{\theta}}}(\cdot)$. To avoid having to calculate the forward map and its gradient at each iteration, we follow [139] and replace the strong constraint in Equation 4.1 by a weak constraint that allows for an outer-inner-loop optimization algorithm. The optimization alternates between an expensive outer loop with L iterations during which $L_{\text{inner}} \gg L$ iterations of an inexpensive inner loop are performed. The forward operator and its gradients are only calculated once during each outer loop iteration while the inner loop contains several updates to the networks. Through Monte-Carlo approximation of the above expectation, we arrive at this weak formulation by introducing N_p slack variables, $\mathbf{x}_{1:N_p}$, that alongside the network weights are minimized in the following objective:

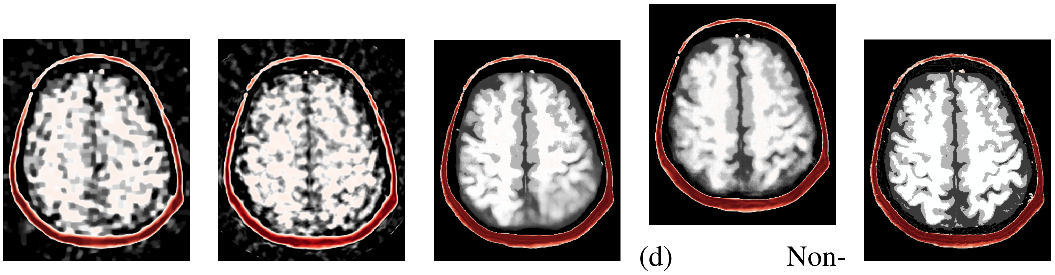
$$
\begin{aligned}
\underset{\mathbf{x}_{1:N_p}, \boldsymbol{\phi}}{\text{minimize}} \, \frac{1}{N_p} \sum_{n=0}^{N_p} & \Bigg[\frac{1}{2\sigma^2} \left\| \mathcal{F}(\mathbf{x}^{(n)}) - \mathbf{y}^{\text{obs}} \right\|_2^2 + \frac{1}{2\gamma^2} \left\| \mathbf{x}^{(n)} - f_{\widehat{\boldsymbol{\theta}}}^{-1}\left(g_{\boldsymbol{\phi}}(\mathbf{z}^{(n)}); \overline{\mathbf{y}}^{\text{obs}}\right) \right\|_2^2 \\
& + \frac{1}{2} \left\| g_{\boldsymbol{\phi}}(\mathbf{z}^{(n)}) \right\|_2^2 - \log\left|\det \mathbf{J}_{h_{\boldsymbol{\phi}}}(\mathbf{z}^{(n)})\right| \Bigg]
\end{aligned}
\tag{4.2}
$$

where γ is the slack factor and if $\gamma \to 0$ the strong formulation is recovered. We treat the result of this optimization as the "gold standard" since it produces the best results but needs access to prior samples and to a large amount of non-amortized compute in the form of forward and adjoint PDE solves.

To assess the efficacy of posterior inference between our amortized methods and the non-amortized methods, we devised benchmarks inspired by the prescriptions in [118] to accelerate the incremental development of this class of algorithms. Firstly, we evaluate the image reconstruction quality of the point estimate generated by each method. Secondly, we conduct a qualitative review of the uncertainty images they produce. Lastly, we quantitatively analyze their uncertainty calibration using the same calibration test outlined in subsection 4.5.2. Due to the expensive nature of the non-amortized methods, we are only able to compare results on a single unseen observation, but we expect these results to generalize to other observations.

4.6.3 Benchmark I: Comparing reconstruction quality

Our first comparison evaluates the posterior means from each method against the ground truth. For a baseline, we include the maximum a posteriori (MAP) from a traditional FWI with total-variation regularization. From Figure 4.12, we observe that the mean-field estimate contains strong artifacts in the soft tissue due to overfitting the noise as it lacks prior knowledge. As expected, our non-amortized normalizing flow method's posterior mean exhibits a superior point estimate compared to our amortized approach. However, this improvement comes with significantly higher computational costs since this result requires 800 online evaluations of the forward operator and its adjoint when solving the optimization in Equation (Equation 4.2), compared to only four online evaluations for amortized ASPIRE 4. This difference in computational expense highlights the trade-off between efficiency and point-estimate quality. This example shows we can achieve results that are close to those by the non-amortized method while using only a small fraction of the online compute.

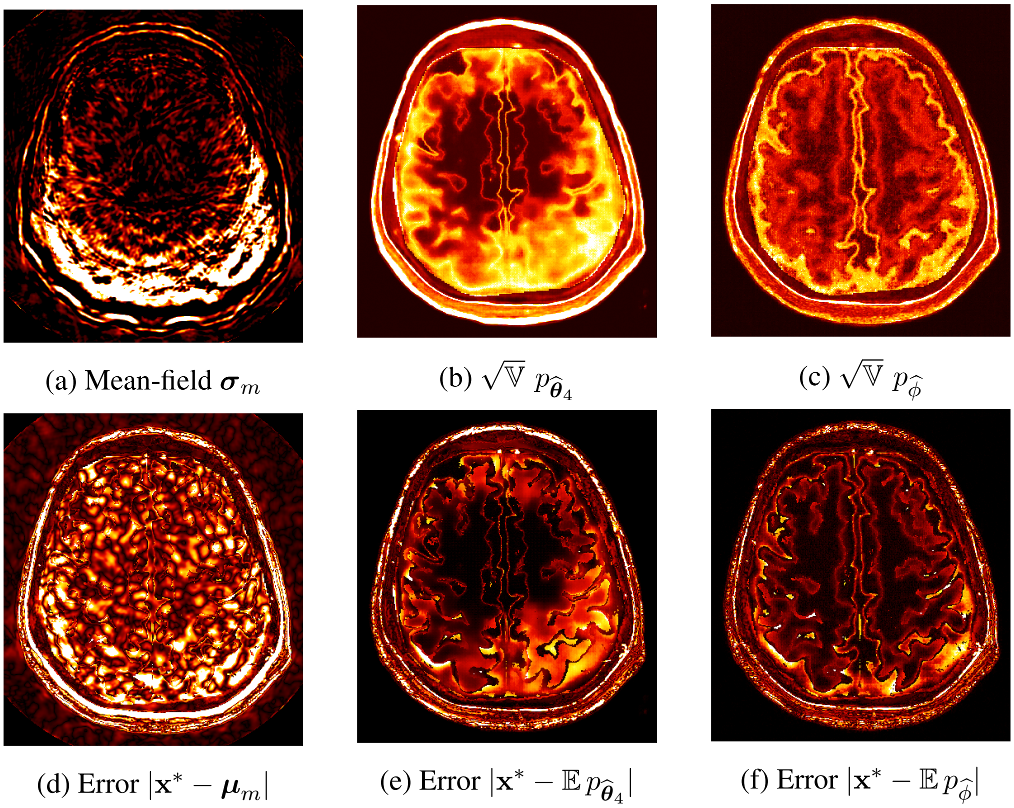


(a) FWI TV (b) Mean-field (c) ASPIRE (d) Non-amortized (e) Ground truth

Figure 4.12: Reconstruction from benchmarked methods. (a) Traditional FWI with Total-Variation regularization. (b) Mean-field approximation $\boldsymbol{\mu}_m$. (c) Our amortized ASPIRE 4 $\mathbb{E}\, p_{\widehat{\boldsymbol{\theta}}_4}$. (d) Our non-amortized gold standard $\mathbb{E}\, p_{\widehat{\phi}}$. (e) Ground truth $\mathbf{x}^*$. As expected, our non-amortized method shows the highest quality, while ASPIRE shows similar quality albeit missing some details on the lower right.

4.6.4 Comparing uncertainty estimates

The posterior standard deviation of the methods tell a similar story to the posterior means. The structure of the standard deviation from the mean-field approximation Figure 4.13a comes mainly from the physics of the problem therefore correctly concentrates in the lower parts of the parameters where high-contrast has created complicated wavefield reverberations, but the method has failed to warn of errors due to noise artifacts throughout the reconstruction shown in Figure 4.13d.



(a) Mean-field $\boldsymbol{\sigma}_m$

(b) $\sqrt{\mathbb{V}}\ p_{\widehat{\boldsymbol{\theta}}_4}$

(c) $\sqrt{\mathbb{V}}\ p_{\widehat{\phi}}$

(d) Error $|\mathbf{x}^* - \boldsymbol{\mu}_m|$

(e) Error $|\mathbf{x}^* - \mathbb{E}\, p_{\widehat{\boldsymbol{\theta}}_4}|$

(f) Error $|\mathbf{x}^* - \mathbb{E}\, p_{\widehat{\phi}}|$

Figure 4.13: Comparing uncertainty of methods. The first row shows the posterior standard deviation from: (a) Non-amortized mean-field approximation sigma value. (b) Our amortized method. (c) Our non-amortized gold-standard method. The second row shows the corresponding errors. All plots have the same colorbar from 0 to $50[\mathrm{m/s}]$.

To perform a close inspection on the uncertainty of the VI methods, we take a single trace through the posterior means (the diagonal trace going from the top left to the bottom right). Figure 4.14 shows that the mean-field method has large errors compared to the ground truth and that its uncertainty band does not contain the ground truth. Here we have chosen a 2σ band around the mean. On the other hand, both ASPIRE and the non-amortized method produce high-fidelity estimates of the ground truth. Furthermore, when our methods have high error, the uncertainty bands expand such that they contain the ground truth with high fidelity.

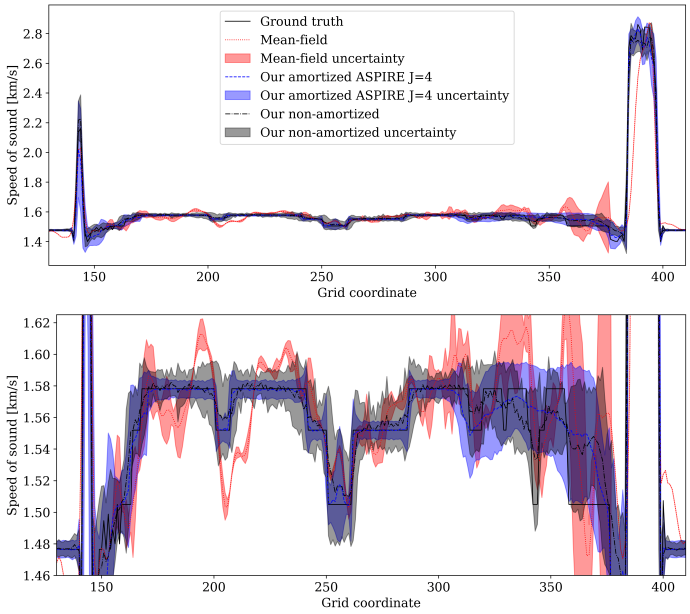


Figure 4.14: Comparison for a single trace. (Top figure) Estimated parameters juxtaposed with the ground truth. (Bottom figure) Same plot with zoomed vertical axis. While our posterior estimates have relatively high error in the area with coordinates 300 to 350, the uncertainty increases, suggesting the uncertainty is well calibrated.

4.6.5 Benchmark II: Comparing uncertainty calibration

Following the method in subsection 4.5.2, we compare the calibration curves of the three VI methods under consideration. Since the two non-amortized methods are compute intensive, we are only able to compute the calibration curve for a single test example Figure 4.12. This contrasts with the more extensive evaluation carried out in Figure 4.11, which encompasses a range of test cases thanks to the cheap online cost of our method. The calibration of the three methods is shown in Figure 4.15. We observe that the mean-field method shows poor calibration while our method achieves better calibration that is close to the one of the non-amortized method. This final observation aligns with our thesis that we can achieve approximation quality similar to an expensive non-amortized method at a fraction of the cost making it a compelling choice in scenarios where time-to-solution is a limiting factor.

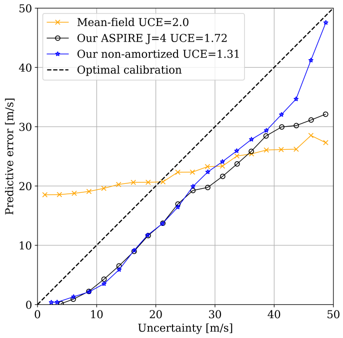


Figure 4.15: The mean-field approximation shows poor calibration, and the gold standard non-amortized method has the best calibration as expected. Our amortized method is close to the gold standard while using a small fraction of the compute cost.

4.6.6 Benchmark III: Computational cost

The primary computational cost in the above methods lies in executing the physics-based forward operator $\mathcal{F}$ and its gradient, especially in wave-based imaging where $\mathcal{F}$ requires solving PDEs. This step is much more demanding than posterior sampling, which involves just a single neural network pass for normalizing flows. Thus, in this section, we measure computational costs in terms of PDE solves. For clock times, please refer to Appendix Table 1.

Offline phase:

In amortized VI, the bulk of computational expenses occur during the offline phase. Similar to other simulation-based inference methods, synthetic observations are generated by evaluating the forward operator. Our method also requires the computation of the gradient for each training sample, which equates to two more forward operators. Additionally, each refinement requires recalculating the gradient. However, a few iterations (3-4) are generally sufficient for satisfactory results. Although the initial training phase is resource-intensive, the amortized model, once trained, becomes cost-effective with repeated use across various datasets.

Online phase:

The main cost during the online phase is also PDE solves. Each refinement iteration, requiring a gradient calculation, incurs a cost of 2 PDEs subsubsection 4.4.2. Compared to earlier Bayesian methods that required $1000 - 10000$ online PDE [53], our method significantly reduces this to less than 10 online PDEs, nearing real-time imaging. In medical applications where timely results are crucial [6], fast online inference is essential.

Compute break-even:

Despite the high cost of the offline phase, our method becomes cost-effective after a certain number of evaluations. The number can be estimated from Table 4.1. For instance, for TUCT, compared to a mean-field solution requiring 600 PDEs, our method—with $\mathbf{N} = 1000$ and $J = 4$—incurs 9000 offline and 8 online PDEs for a total of 9008 PDEs. Thus becomes cost-effective after about 15 test cases, not accounting for improved estimates and uncertainty.

When juxtaposed with our proposed gold-standard non-amortized method, our amortized approach breaks-even more rapidly. For the non-amortized method, we use a pretrained ASPIRE 1 network, which requires $1000 + 2 \times 1000 \times 1 = 3000$ offline PDEs, followed by $2 \times 1 = 2$ online PDEs. The optimization of (Equation 4.2) uses $L = 400$ outer loop iterations leading to online 800 PDEs. The total cost is 3802 PDEs, which means the amortized method pays itself back when used on $\frac{9008}{3802} \approx 3$ test cases.

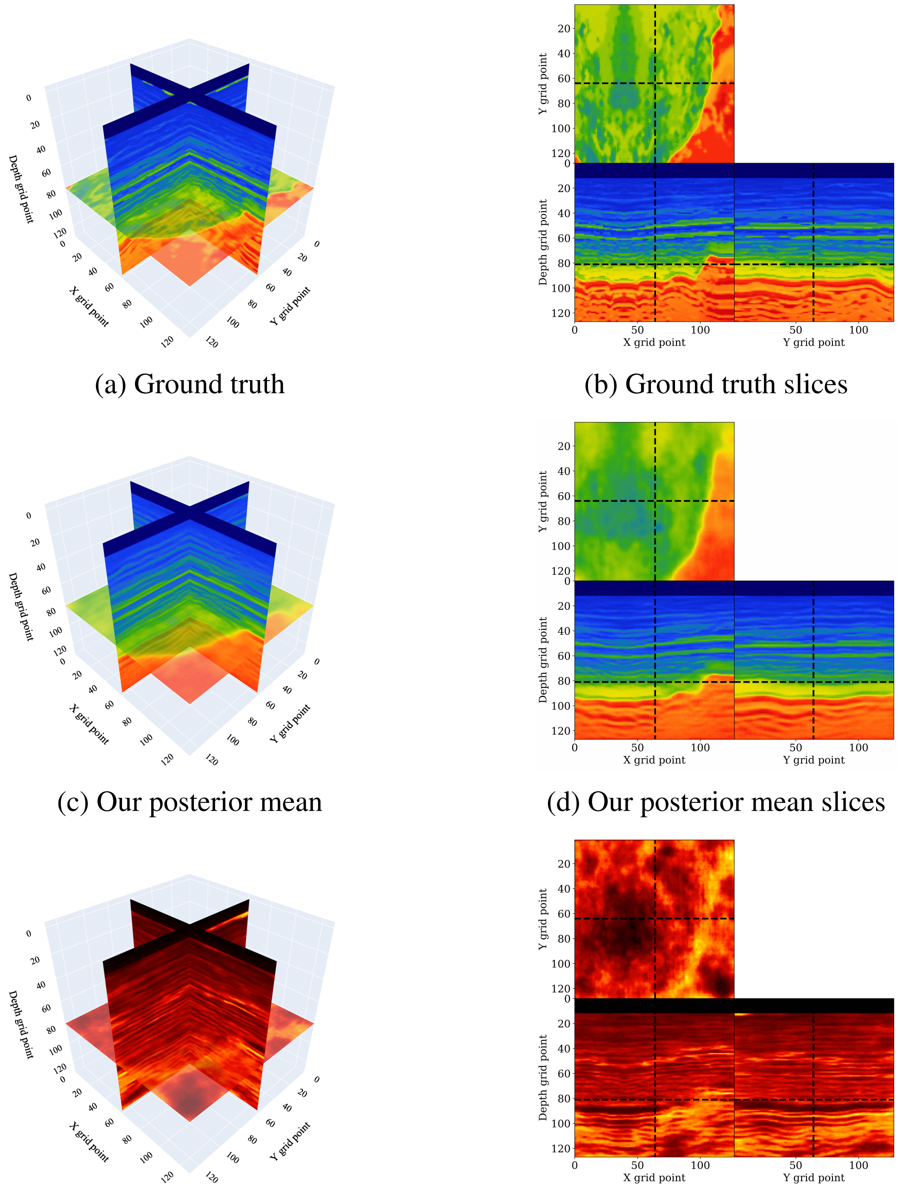
Table 4.1: Costs measured by evaluations of forward operator. N is the number of training samples, J are refinement iterations, and L are online gradient steps.

Method	Offline cost	Online cost
ASPIRE	$N + 2 \times N \times J$	$2 \times J$
Our non-amortized method	$N + 2 \times N \times J$	$2 \times J + 2 \times L$
Mean-field approximation	None	$2 \times L$

4.7 Future work

The superior posterior mean achieved by the non-amortized method compared to our solution indicates that further information could be extracted from $\mathbf{y}^{\text{obs}}$ through additional refinement iterations. Although we proposed some heuristics, determining the optimal number of refinement iterations to maximize performance remains an area of research. Our technique is compatible with any conditional density estimator. While we have utilized Normalizing Flows in our implementation, the framework can easily adapt to other conditional density estimators such as Variational Autoencoders (VAEs) [140], GANs [141], and diffusion models [142].

Due to out-of-plane effects of acoustic modeling, the TUCT problem is best treated in 3D. Although TUCT was demonstrated here in 2D, ASPIRE is not limited to 2D problems. Particularly, when empowered by memory-frugal normalizing flows [33] ASPIRE can achieve full volume Bayesian inference for 3D inverse problems. For our TUCT example, the limiting factor was the absence of a 3D training dataset but in seismic imaging (a field that appreciates the importance of 3D modeling and inference) we used the 3D Compass dataset [143] and share the results of solving 3D FWI in Figure 4.16. Setup details are similar to the TUCT problem. A detailed study of the 3D capabilities of ASPIRE is being prepared for future work but here we succinctly report that for a $128 \times 128 \times 128$ inference problem, offline training took 1 day on a single GPU and that the uncertainties shown in Figure 4.16e were pleasingly correlated with structures that are known to be difficult to image i.e. structures that are: deeper, vertical or close to the edge.



(a) Ground truth

(b) Ground truth slices

(c) Our posterior mean

(d) Our posterior mean slices

(e) Our posterior deviation

(f) Our posterior deviation slices

Figure 4.16: ASPIRE for 3D inverse problems. (a) Ground truth 3D render. (b) Ground truth folded out slices. (c) Our posterior mean 3D render. (d) Posterior mean slices. (e) Our posterior deviation 3D render. (f) Posterior deviation slices. The ASPIRE 2 shows physically viable probabilistic estimates for a $128 \times 128 \times 128$ FWI problem.

4.8 Conclusions

We introduced a method that iteratively improves on approximations to Bayesian posteriors in the context of inverse problems. Our method brings together concepts from generative modeling, physics-hybrid methods, and statistics. Practically our algorithm achieves higher performance by iteratively extracting more information from the observed data. The mathematical interpretation of our method is to make a gradient-based summary statistic more informative by moving the fiducial points closer to the maximum likelihood estimates. Our method forms an interesting middle ground between amortized VI and non-amortized VI. Importantly, the offline training phase makes it such that the online costs are small rendering our approach suitable for applications that demand fast online turn-around times. Our experiments demonstrate improvements in estimated posteriors on a stylized example where the posterior is known analytically. In a realistic medical transcranial ultrasound imaging application, the online cost is many times cheaper than non-amortized methods while demonstrating high-quality amortized inference. We believe that this approach represents a step forward in the field, offering a computationally efficient solution for Bayesian inference in high-dimensional inverse problems with expensive to evaluate forward operators.

CHAPTER 5

MACHINE LEARNING ENABLED VELOCITY MODEL BUILDING WITH UNCERTAINTY QUANTIFICATION

SUMMARY

Accurately characterizing migration-velocity models is crucial for a wide range of geophysical applications, from hydrocarbon exploration to monitoring of CO 2 sequestration projects. Traditional velocity-model building methods such as Full-Waveform Inversion (FWI) are powerful but often struggle with the inherent complexities of the inverse problem, including noise, limited bandwidth, receiver aperture and computational constraints. To address these challenges, we propose a scalable methodology that integrates generative modeling, in the form of Diffusion networks, with physics-informed summary statistics, making it suitable for complicated imaging problems including field datasets. By defining these summary statistics in terms of subsurface-offset image volumes for poor initial velocity models, our approach allows for computationally efficient generation of Bayesian posterior samples for migration-velocity models that offer a useful assessment of uncertainty. To validate our approach, we introduce a battery of tests that measure the quality of the inferred velocity models, as well as the quality of the inferred uncertainties. With modern synthetic datasets, we reconfirm gains from using subsurface-image gathers as the conditioning observable. For complex velocity-model building involving salt, we propose a new iterative workflow that refines amortized posterior approximations with salt flooding and demonstrate how the uncertainty in the velocity model can be propagated to the final product reverse-time migrated images. Finally, we present a proof of concept on field datasets to show that our method can scale to industry-sized problems.

5.1 Introduction

Subsurface characterization of the earth's subsurface is important for hydrocarbon exploration [2], monitoring of CO 2 storage projects [3], geothermal energy projects [4] and various other applications [5]. Generally, subsurface characterization is achieved by observing the interaction between specific physical phenomena (such as electrodynamics, gravity, and acoustic wave propagation) and subsurface properties. This tomographic information is then resolved into images that are analyzed for different characterization requirements. Although our framework is generally applicable, we focus on characterization of acoustic properties by means of probing the subsurface with acoustic waves. Out of the various methods, FWI stands out as a powerful tool due to its ability to resolve high-quality acoustic images in complex structures [144]. In spite of its advantages, FWI still has shortcomings due in part to the nature of the problem but also due to the specific computational challenges that FWI brings since it requires solving many wave-equation partial differential equations (PDEs).

The particular challenges of the FWI inverse problem are that the observations are corrupted by noise, are limited in frequency bandwidth, computational simulations will always contain some approximation to the true physics and due to practical engineering considerations are mostly restricted to sensing the upcoming waves at the surface so will suffer from some sort of limited aperture. All of these factors contribute to FWI's ill-posed nature in the sense that many subsurface scenarios are capable of explaining the limited data available. Traditional workflows approach this challenge by introducing prior information to regularize the vast search space, such as Total Variation (TV) [132]. While this has served well to produce deterministic solutions, it does not express the multi-solution nature of FWI and does not represent a realistic prior of the multiscale complexity of Earth structures.

We use Bayesian inference as a principled framework to combine observable data (seismic shot gathers) with prior knowledge (training samples) and output a family of Earth models that give users practical uncertainty quantification (UQ). One of the major gaps in geophysical inversion is the difficulty of computing Bayesian posterior distributions that are grounded in useful prior information and incorporate the complex physics of the problem. Existing approaches either are too expensive to scale to large-scale problems (sampling-based methods), fail to capture the full extent of uncertainty (local methods), or rely on approximations that weaken the physical fidelity of the results (convolution-based methods).

As a form of variational inference [19], our approach sidesteps the computational problem of sampling the posterior distribution by optimizing instead for an amortized (read generalized) approximation to the posterior that is learned from training examples and can be applied computationally efficiently at test time with small compute (as measured by PDE solves) on various datasets.

5.2 Chapter outline

First, we introduce the wave-based inversion problem that we aim to solve. Due to the ill-posed nature of the problem, we use a Bayesian formulation to solve it. Then, we discuss related work that has solved various aspects of this problem, highlighting the gap in this literature that we aim to fill. We explain our particular methods, which include simulation-based inference (SBI) and conditional Diffusion networks as the core generative network. We then explain the use of physical summary statistics to efficiently incorporate the wave physics into the method. To evaluate the quality of our posterior distributions, we propose four distinct metrics, each targeting a different aspect of UQ. These metrics are applied to assess the improvements gained from using Common-Image Gathers (CIGs) in comparison to Reverse-Time Migrations (RTMs, which are CIGs at zero subsurface offset). Next, we address the challenges presented by the complex salt structures in the SEAM model [145]. To overcome these, we recognize the need for additional guidance from the physics and propose an iterative algorithm, ASPIRE, that leverages the wave PDE while minimizing the total number of PDEs at inference time. Finally, we test the robustness and scalability of our method by applying it to field datasets. The results demonstrate that the method is adaptable to changes in the test distribution and can handle large-scale 2D inversion problems (e.g., 512×7024 grid sizes).

5.3 Problem statement

The core of subsurface characterization involves solving an inverse problem where the objective is to infer unknown subsurface properties from observational data. In this context, the forward process, represented as

$$\mathbf{y} = \mathcal{F}(\mathbf{x}) + \boldsymbol{\varepsilon},$$

describes how the observations (shot records) $\mathbf{y}$ are generated from the underlying subsurface properties $\mathbf{x}$, with $\boldsymbol{\varepsilon}$ representing bandwidth-limited noise. Here, we focus on wave-based inversion, where the forward operator $\mathcal{F}$ corresponds to the solution of the wave-equation PDE with the wavefield being restricted to the positions of the receivers. The complexity of this problem arises from the null-space present in the forward operator, compounded by noise in the measurements, which makes direct inversion unreliable as it fails to characterize the full solution.

To model the noise in the system, one would assume that the noise $\boldsymbol{\varepsilon}$ follows a known distribution, such as a normal distribution $N(0, \sigma)$. This leads to a likelihood function $p(\mathbf{y} \mid \mathbf{x})$, which quantifies the probability of observing the data $\mathbf{y}$ given the unknown parameters $\mathbf{x}$. If the noise is additive, the likelihood can be written as a well-known ℓ_2-normed misfit, $p(\mathbf{y}|\mathbf{x}) = \frac{1}{2\sigma^2}\|\mathcal{F}(\mathbf{x}) - \mathbf{y}\|_2^2$. Minimizing data misfits of this form underpins FWI and other variational methods that use the forward model's fit to the data, but on their own, it is insufficient to fully resolve the inverse problem due to the non-uniqueness of solutions and the presence of possibly adverse parasitic local minima.

In this paper, we address this limitation by adopting a Bayesian approach to the inverse problem. Our target in this paper is that given an observation from the field $\mathbf{y}$ we aim to find samples from the posterior distribution $\mathbf{x}_{\text{post}} \sim p(\mathbf{x}|\mathbf{y})$, which can be defined using Bayes's rule as the combination of the prior and the data likelihood $p(\mathbf{x} \mid \mathbf{y}) \propto p(\mathbf{y} \mid$

$\mathbf{x})p(\mathbf{x})$. This posterior distribution combines the prior information about the subsurface properties, encoded in $p(\mathbf{x})$, with the likelihood of observing the data, $p(\mathbf{y} \mid \mathbf{x})$, resulting in a probabilistic representation of the subsurface model that accounts for both the uncertainty in the data and the prior knowledge. This Bayesian framework allows for a more robust characterization of subsurface properties, as it provides not just a single estimate but a distribution of plausible solutions, incorporating uncertainty into the interpretation.

5.4 Related work and our contributions

The literature on solving the FWI problem with uncertainty has primarily focused on implementations of Stein Variational Gradient Descent (SVGD) methods. The foundational work in this area was introduced by [53], who applied Bayesian seismic tomography using normalizing flows. Later, the same authors extended their method to 3D inversion [27], though this approach still required numerous forward and adjoint PDE solutions for each observation, which made it computationally expensive.

In an effort to reduce the computational burden of SVGD methods, [146] proposed a technique that minimizes the number of optimization iterations by carefully defining the prior distribution. This approach begins with solving the FWI problem under the assumption of convergence and then adds perturbations to the solution to create a prior distribution. This prior is used as an initial ensemble in an SVGD technique to optimize toward individual maximum likelihood estimates (MLEs) that do not collapse into the same solution due to a repulsion term in the method. Although the authors did not use true prior terms—thus these solutions are not precise Bayesian samples—they presented samples that revealed interesting variations between them and were mostly focused on the areas of the image with low illumination from the source-receiver configuration. [147] extended these concepts by combining ideas from the Deep Image Prior (DPI) with conditional networks. They performed an SVGD-like update starting from an ensemble of models that surrounded a precomputed FWI solution, as in [146], to fine-tune the weights of a pretrained conditional network. This pretraining process enabled them to sample various Earth models by post hoc changes to the network conditions. While this takes regularization benefits of both the untrained DPI and the features learned during pretraining, it also does not have a clear prior used during optimization, so does not correspond to true Bayesian samples over the target in this case the uncertainty over network parameters.

In the context of defining realistic Earth priors, [148] assumed constrained Gaussian distributions over the prior, which were fitted to real Earth data. They then implemented a secondary optimization that bypasses the likelihood, making it efficient to swap different priors at inference time. [149] explored another important aspect of UQ by employing importance sampling and ensemble methods. Their objective was to capture uncertainty stemming not only from the ill-posed nature of the inverse problem but also from epistemic uncertainty in the network weights.

Similar work using normalizing flows includes [150], which targets time-lapse inversion. They formed training samples by performing FWI inversion on prior examples to construct the training dataset. The prior samples assumed access to a non-cycle-skipped starting model, to which various perturbations were added, resulting in a diverse set of training samples. In a related approach, [151] utilized invertible networks with a maximum mean discrepancy (MMD)-based loss. Given that MMD estimation requires large batch sizes, the authors applied model and data reduction techniques to minimize memory usage and make the training feasible. Though not specifically aimed at UQ, [152] presented a method similar to our iterative refinement, by using an iterative scheme to refine the migration velocity models. Targeting our same application of salt velocity model building, [153] combine trained convolutional neural networks with FWI updates to automatically build salt velocity models. Here, we aim to deliver automatic salt model building but without the cost associated with FWI workflows, while also providing uncertainty quantification.

This paper builds on the methods introduced in previous literature of applying variational inference methods towards seismic imaging [21, 37] that have the goal of calculating

Bayesian samples while making frugal use of the physical operator by defining an expensive offline training phase to achieve a cheap online inference phase. In particular, we build on the WISE framework [36] with the use of conditional Diffusion networks to improve results by leveraging the prior learning capabilities of Diffusion [154]. While Structural Similarity Index Metric (SSIM) [155] and Root Mean Squared Error (RMSE) can compare the quality between image reconstructions, it remains unclear in the literature how to compare two results for uncertainty. We therefore propose and discuss a battery of metrics that can be used to benchmark the quality of the uncertainty. Furthermore, we discuss a method to improve on amortized results with minimal extra computation at test time and then demonstrate these methods on field data in preliminary proof of concepts.

Our key contributions include:

1. Extending the WISE framework to incorporate conditional Diffusion networks.
2. Proposing four benchmark metrics to assess the quality of UQ.
3. Testing our method on the Compass and ‘Synthoseis‘ training datasets, demonstrating the value of using Common Image Gathers (CIGs) over Reverse Time Migrations (RTMs) alone from the standpoint of image quality and also uncertainty quality.
4. Introducing the iterative ASPIRE algorithm for seismic inversion, particularly for complex salt structures, such as those in the SEAM model.
5. Presenting a proof-of-concept application on field datasets to demonstrate the scalability of the method.

5.5 Methods

Our approach builds on the simulation-based inference (SBI) framework [98], a powerful tool for solving inverse problems that leverages numerical simulations and conditional generative networks to approximate posterior distributions. While SBI is, in principle, a general method, directly applying it to the complex problem of seismic subsurface characterization presents significant challenges due to the intricacies of waveform data. To address these challenges, we incorporate conditional Diffusion networks as the generative backbone, enabling the learning of an expressive prior and scalable posterior sampling. Additionally, we employ physics-based summary statistics to compress observational data while preserving crucial information about the subsurface properties.

5.5.1 Simulation-based inference

SBI combines the strengths of numerical simulations and conditional generative modeling, providing a powerful framework for solving complex inverse problems [98]. Numerical simulations are used to generate data pairs, $\mathcal{D} = \{\mathbf{x}^i, \mathbf{y}^i)\}_{i=0}^{N}$, where each pair consists of a set of subsurface properties $\mathbf{x}^i$ and the corresponding simulated observation $\mathbf{y}^i$ derived using the forward simulation section 5.3. These data pairs are then used to train a conditional generative network which learns the posterior distribution of the unknown properties given the observations. By integrating physics-based simulations with modern generative modeling techniques, SBI provides an amortized method—that is, it generalizes over all data simulated from the realizations of the prior. This means that after an initial training phase, inference is inexpensive and can be done on many unseen test observations without retraining or expensive applications of the forward/adjoint operator.

5.5.2 Conditional generative modeling with Diffusion networks

Diffusion networks are density estimation algorithms based on learning the score of the target distribution $\nabla_{\mathbf{x}} \log p(\mathbf{x})$ [142]. Specifically, we define a family of mollified distributions $\nabla_{\mathbf{x}} \log p(\mathbf{x}, \sigma(t))$, where $\sigma(t)$ represents a noise schedule such that, as time t increases, the distribution is mollified towards the Gaussian distribution. After learning the score for all time steps t, Diffusion networks can evaluate likelihoods [156] and sample new generative instances from the target distribution $p(\mathbf{x})$. To learn the score, Diffusion networks use one of the many forms of the denoising objective [157],

$$\hat{\theta} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}}\ \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})}\mathbb{E}_{\mathbf{n}\sim\mathcal{N}(0,t^2 I)}\|s_{\boldsymbol{\theta}}(\mathbf{x}+\mathbf{n};t)-\mathbf{x}\|_2^2$$

where the approximation the score at time t is given by the evaluation of the trained network $s_{\hat{\theta}}(\mathbf{x},t)$. New samples from the target distribution can be generated by solving the following stochastic differential equation (SDE):

$$d\mathbf{x} = -\dot{\sigma}(t)\sigma(t)\nabla_{\mathbf{x}}p(\mathbf{x};\sigma(t))dt.$$

Here, we use the formulation in [157] to simplify terms. Many strategies exist for solving this SDE. Here, we follow the method in [157] and use a procedure similar to the predictor-corrector sampler of [142].

To extend Diffusion networks towards conditional distributions, we aim to find the conditional score $\nabla_{\mathbf{x}} \log p(\mathbf{x}|\mathbf{y})$. Based on the theory laid out in [158] and [159], we can approximate this conditional score with a simple modification of Diffusion networks by incorporating the observation $\mathbf{y}$ as an additional condition to the denoising network

$$\hat{\theta} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}}\ \mathbb{E}_{\mathbf{y}\sim p(\mathbf{y}|\mathbf{x})}\mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})}\mathbb{E}_{\mathbf{n}\sim\mathcal{N}(0,t^2 I)}\|s_{\boldsymbol{\theta}}(\mathbf{x}+\mathbf{n},\mathbf{y},t)-\mathbf{x}\|_2^2.$$

In many cases, implementing a conditional Diffusion network involves simple modifications to existing non-conditional Diffusion networks. In [put GitHub link], we share our conditional Diffusion implementation derived from non-conditional networks from Elucidating Diffusion Models [157]. In practice, we approximate the expectation over prior samples with a set of ground truth examples and sample from the likelihood by running the simulator in section 5.3 to form a paired dataset $\mathcal{D} = \{(\mathbf{x}^i,\mathbf{y}^i\}_{i=0}^N$, then perform stochastic gradient descent on the following objective:

$$\hat{\theta} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}}\ \mathcal{L}(\mathcal{D}, \theta) = \underset{\boldsymbol{\theta}}{\operatorname{argmin}}\ \sum_{i=0}^{N} \mathbb{E}_{\mathbf{n}\sim\mathcal{N}(0,t^2I)} \| s_{\boldsymbol{\theta}}(\mathbf{x}^i + \mathbf{n}, \mathbf{y}^i, t) - \mathbf{x}^i \|_2^2.$$

After training, conditional Diffusion generates samples from the posterior $p(\mathbf{x} \mid \mathbf{y}^{\text{obs}})$ by conditioning the network on $\mathbf{y}^{\text{obs}}$ and following the reverse process on the same SDE as in subsection 5.5.2 to generate a posterior sample $\mathbf{x}_{\text{post}}$. While the physics of the inverse problem is present in the formation of the dataset, this is insufficient in the case of complex forward operators associated with the wave equation [39]. Therefore, we will explore a method to alleviate this challenge next by incorporating more physics into the methodology.

5.5.3 Physics-based summary statistics

While simulation-based inference (SBI) has shown promise across various fields, applying it to problems with complex forward operators, such as seismic inversion, presents unique challenges. These challenges stem from the difficulty of extracting the rich information contained in waveforms. To address this, we employ summary statistics that compress the data while preserving information needed to infer unknown subsurface properties. Summary statistics can either be hand-crafted by domain experts or learned by exploiting probabilistic symmetries within the data [160]. In this work, we follow the methods of [39] and [36] by using physics-based summary statistics derived from the physical forward operator. Specifically, we utilize the gradient of the data likelihood, which allows us to inject physics into the generative process at the cost of a single PDE solve per set of posterior samples, where the PDE corresponds to the computations required for a single migration.

Using this gradient serves as a natural way to undo the complexity of the waveforms since it brings the data into the model space. Theoretically, this approach is well-founded, as it optimally saturates the information inequality [100], a bound that limits how much information a summary statistic can extract about an unknown variable. Empirical results have shown that incorporating these summary statistics accelerates training and improves convergence with fewer data samples [39]. Given an observation $\mathbf{y}$ and a migration-velocity model $\mathbf{x}_0$, the summary statistic is calculated as follows:

$$\overline{\mathbf{y}} := h(\mathbf{y}) = \overline{\nabla \mathcal{F}}(\mathbf{x}_0)^{\top}(\mathcal{F}(\mathbf{x}_0) - \mathbf{y}),$$

where $\overline{\nabla \mathcal{F}}$ is the subsurface-offset extended Jacobian, following the method of [161, 36]. $\overline{\mathbf{y}}$ contains the CIGs, which represent the summarized waveform but transformed to model space, while preserving more information compared to RTMs in situations where the migration-velocity model is poor. We summarize all training data $\overline{\mathbf{y}}^i = h(\mathbf{y}^i)$ and form a new training dataset $\overline{\mathcal{D}} = \{\mathbf{x}^i, \overline{\mathbf{y}}^i\}_{i=0}^{N}$ and train a conditional Diffusion network with subsection 5.5.2. At inference time, the observation $\mathbf{y}^{\mathrm{obs}}$ is summarized in the same manner, $\overline{\mathbf{y}}^{\mathrm{obs}} = h(\mathbf{y}^{\mathrm{obs}})$ and then passed into the same sampling process as subsection 5.5.2. The computational cost of this approach is low because it requires only a single extended migration per observation at inference time, making it more computationally efficient than non-amortized methods that typically demand hundreds of PDE solves [53] for each dataset. Our approach is also amortized and therefore can be applied to many unseen observations without repeating the costly training process.

5.6 Stylized examples

We separate our experiments into two parts. First, we train our generative networks on synthetic seismic model sets and validate them on unseen datasets from the same synthetic distribution as used during training. Secondly, we apply these trained networks to field datasets to assess their robustness in practical scenarios and also to understand the prior that they learn.

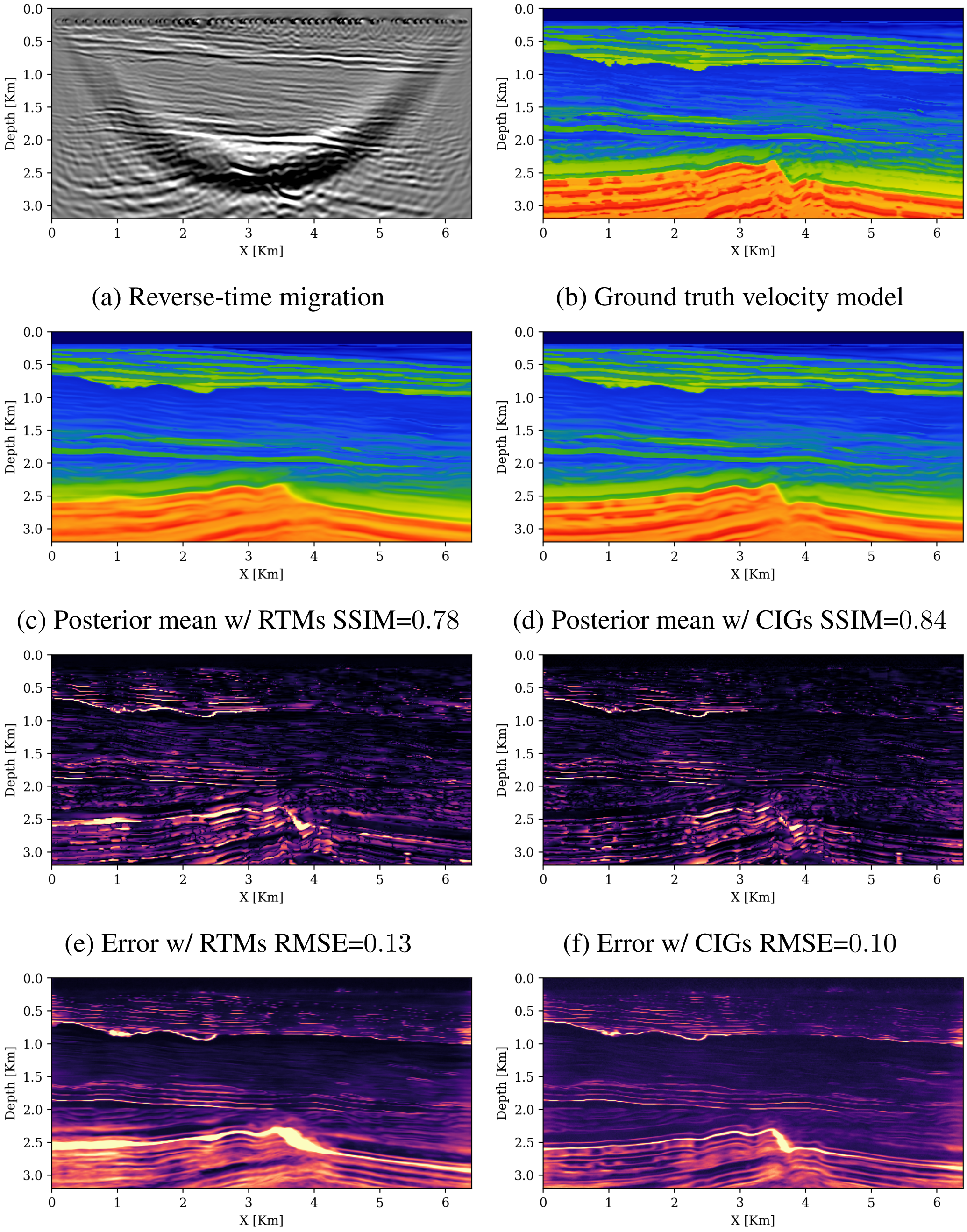
5.6.1 Training dataset generation

For the synthetic experiments, the migration-velocity models used are intentionally not kinematically accurate. For the two salt imaging experiments, we first remove the salt body from the model, replace it with the average velocity of the sedimentary layers, and then smooth the resulting model using a Gaussian kernel. To minimize artifacts from tomographic updates, the migrations are generated using the inverse-scattering imaging condition (ISIC) [162], which implies that our method is purely reflection-based. The sources are given by Ricker wavelets that are filtered to remove unrealistically low frequencies below 3 Hz. When noise is added to the simulated shot gathers, we apply a filter to ensure the noise contains the same frequency content as the source wavelet. We use JUDI and Devito for wave simulations [162, 8, 163]. The extended Jacobian operator was created using [164] with a subsurface range that was chosen to contain the majority of the non-zero offset energy.

Across all experiments, we create between 700–800 training pairs. The Diffusion networks are trained until the image quality metrics (RMSE, SSIM) for the posterior means stop improving on a validation set held out during training. To calculate posterior statistics, we generate N posterior samples and calculate Monte Carlo estimations of the posterior mean, $\mathbf{x}_{\text{mean}} = \sum_{i=0}^{N} \mathbf{x}_{\text{post}}^{i}/N$, and standard deviation, $\mathbf{x}_{\text{std}} = \sqrt{\sum_{i=0}^{N}(\mathbf{x}_{\text{post}}^{i} - \mathbf{x}_{\text{mean}})^2/N}$. The number of posterior samples to use is decided by increasing the number until the posterior standard deviation converges; in our case, $N = 64$.

5.6.2 Compass model

Our evaluation begins by using the Compass model [143]. The training dataset consists of the same velocity and CIGs pairs as in WISE [36]. These include $N = 800$ training pairs of velocity models of size (512×256) that are discretized with 12.5 m and CIGs that are generated with 50 equally spaced offsets between -500 m to 500 m without ISIC. The migration-velocity models for the Compass dataset are created using a single 1D velocity profile that is derived from the training dataset. Our only modification is that we use conditional Diffusion as the generative network instead of conditional normalizing flows. Training took 14 GPU hours, while posterior sampling takes 440 sec for the 64-shot CIGs in addition to 2 sec for each posterior sample. The results summarized in Figure 5.1 confirm the observation from [36] that CIGs drastically improve posterior sampling performance. As we discuss later, these results improve on the normalizing flow results from [36]. Overall, we observe increasing error and uncertainty with depth. Additionally, there is correlation between the complexity of the velocity model and the error and uncertainty. Aside from uncertainty near the strong lateral variation along the major unconformity in the model, we also observe increased errors and uncertainty associated with topology on the velocity kickback that occurs around 750 m depth on the left side of the model. In the section on uncertainty benchmarks, we expand this rudimentary analysis and detail how we can determine which uncertainty is of higher quality according to well-defined metrics.



(a) Reverse-time migration

(b) Ground truth velocity model

(c) Posterior mean w/ RTMs SSIM=0.78

(d) Posterior mean w/ CIGs SSIM=0.84

(e) Error w/ RTMs RMSE=0.13

(f) Error w/ CIGs RMSE=0.10

(g) Posterior standard deviation w/ RTMs

(h) Posterior standard deviation w/ CIGs

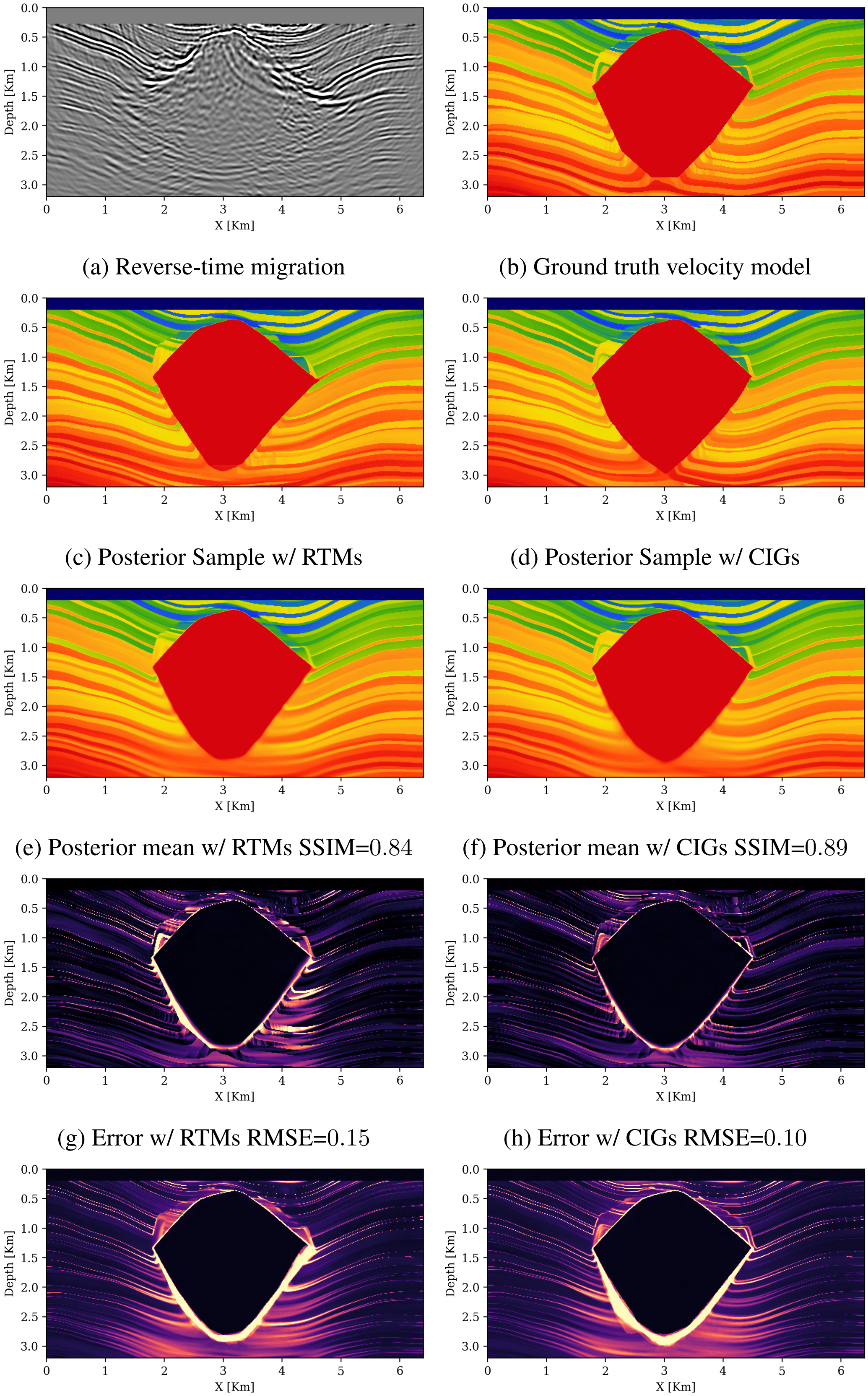
Figure 5.1: Posterior sampling on Compass dataset. We observe an increase in quality of the inferred velocity model when using CIGs instead of RTMs.

5.6.3 Synthoseis models

While the results presented in the previous section are encouraging, the geological setting in the North Sea Compass dataset lacks sufficient complexity to put our inference methodology to the test. In addition, its training was limited to neighboring 2D slices from the same area and therefore may lack in diversity. To address this potential lack of diversity and complexity, we consider an example where we train on synthetically generated models produced by the open-source software package ‘Synthoseis‘[https://github.com/sede-open/synthoseis]. ‘Synthoseis‘ is an algorithm designed to generate realistic and diverse synthetic 3D seismic models tailored for deep learning applications [165]. Our approach specifically uses the algorithm’s velocity-model generation routines, focusing on the creation of the acoustic velocity property (V_p) of Earth models. The workflow consists of several key steps, namely initialization of random Earth parameters ranges, generating 3D depth horizons, embedding of rock property models, and ultimately producing 3D volumes. In this study, we examine scenarios with Earth models that feature salt bodies. Although each model produced by the algorithm exhibits full 3D structures, for our experiments, we extracted 2D slices from the 3D models to manage computational resources needed during ‘Synthoseis‘ generation. It is important to note that while these models are not created by the same generative model that we train, thus using these samples does not fall under the regime of autophagy as defined in [166].

To create the training dataset $\overline{\mathcal{D}}$, we generate $N = 800$ pairs. Each $\mathbf{x}$ has a grid size of 512×256 with a spatial discretization of 12.5 m. We modeled towed sources at the ocean surface with an interval of 12.5 m. The receivers are placed on the ocean bottom with an average sampling of 200 m, they are positioned using a random jittered sampling scheme [167]. The shot records $\mathbf{y}$ are simulated by recording for 3.2 s, while the active sources are modeled with a Ricker wavelet with central frequency of 20 Hz. The simulated noise was band-limited to contain the same frequency content as the source and had a magnitude of 25 dB. To make migration-velocity models, we first replace values where salt is located with the average value of the sedimentary layers. Then, we convert the model from depth to time, smooth in time with an anisotropic Gaussian kernel of size (40×80), and then convert it back to depth. We follow this nonlinear procedure to intentionally create poor migration-velocity models. The horizontal subsurface-offset migrations $\overline{\mathbf{y}}$ include 24 equally spaced offsets between -500 m and 500 m. The conditional Diffusion network was trained for 12 GPU hours. For inference, the algorithm takes 277 sec to generate CIGs for 32 shots and 2 sec for each posterior sample.

After training, we consider an example not seen during training for posterior sampling. The results are shown in Figure 5.2. Our first observation is that, over both image quality metrics (SSIM and RMSE) the network trained with CIGs outperforms the network trained on RTMs only. Furthermore, we are pleased that although there is uncertainty below the salt structure, as evidenced by the high standard deviation Figure 5.2i,Figure 5.2j and posterior means Figure 5.2e,Figure 5.2f that appear blurred in regions where the uncertainty is large, the individual posterior samples in Figure 5.2c,Figure 5.2d exhibit an accurate approximation of the prior in the way the salt is sharply delineated. As before, the posterior mean $\mathbf{x}_{\text{mean}}$ and posterior standard deviation $\mathbf{x}_{\text{std}}$ are calculated from 64 posterior samples. Again, we see reasonably good correlation between the errors and uncertainty where as expected, these occur mainly at the bottom of the salt and in the subsalt areas. It is also encouraging that the unlapping sedimentary layers are well recovered albeit the inference struggles in the area beneath the deepest point of the salt.



(a) Reverse-time migration

(b) Ground truth velocity model

(c) Posterior Sample w/ RTMs

(d) Posterior Sample w/ CIGs

(e) Posterior mean w/ RTMs SSIM=0.84

(f) Posterior mean w/ CIGs SSIM=0.89

(g) Error w/ RTMs RMSE=0.15

(h) Error w/ CIGs RMSE=0.10

(i) Posterior standard deviation w/o CIGs

(j) Posterior standard deviation w/ CIGs

Figure 5.2: Posterior sampling on velocity models generated by ‘Synthoseis‘. Based on the image quality metrics, the CIGs more accurately inform the posterior inference.

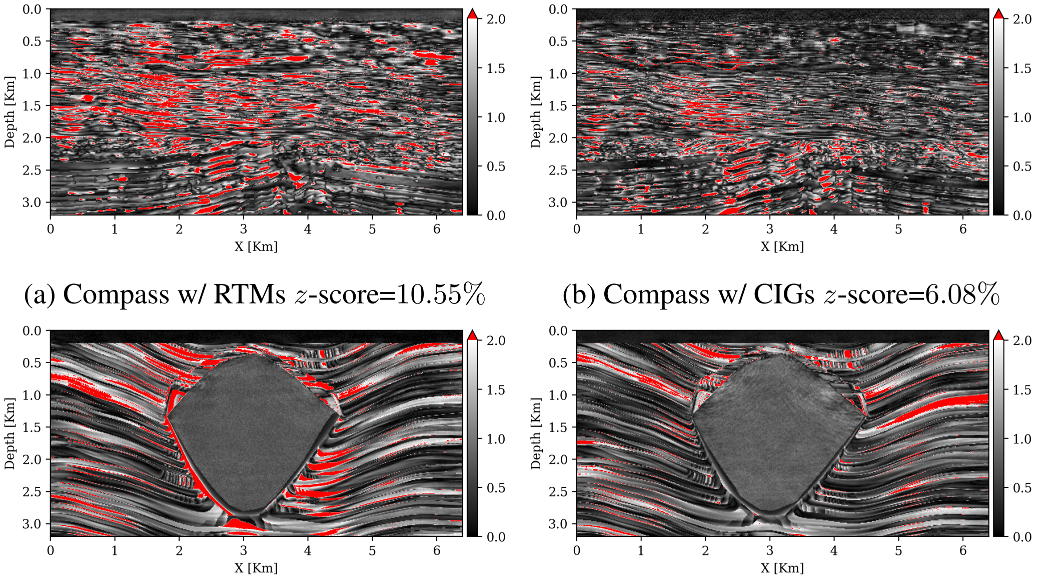
5.7 Quantitative assessment of UQ

Evaluating the quality of the UQ—i.e., quality of the inference of the posterior, from posterior samples is crucial for understanding how well the network captures the true variability in the posterior distribution. In this section, we introduce four approaches to assess the performance of UQ, namely (1) the ability to warn areas with high error; (2) the degree of correlation between uncertainty and error; (3) the posterior coverage, which is defined as the proportion of pixels for which the range of posterior samples contains the ground truth; and (4) the ability of posterior samples to fit the observed shot data. For all metrics, we take the uncertainty to be the standard deviation between posterior samples and the error calculated between the posterior mean and a known ground truth velocity model.

5.7.1 Percentage of regions with large errors but low uncertainty

An important feature of UQ is its ability to warn the user of areas that may have high errors in the reconstruction. By warning, we mean that if an area has high uncertainty, then that area should have high error. Thus, we want to avoid high errors that are not followed by high uncertainty, which could be an indication of poor UQ. To quantify and visualize this, we use the z-score as defined [168] by the pixel-wise division of the error by the uncertainty: $z\text{-score} = |\mathbf{x}^* - \mathbf{x}_{\text{mean}}|/\mathbf{x}_{\text{std}}$ where $\mathbf{x}^*$ is the ground truth velocity, $\mathbf{x}_{\text{mean}}$ the average of the posterior samples, and $\mathbf{x}_{\text{std}}$ their pixelwise standard deviation. In Figure 5.3, we show that high errors are ”acceptable” as long as it is followed by high uncertainty in that same area, while areas with high errors and low uncertainty will have high values for the z-score in this plot. To highlight these areas, grid points with errors that are $2\times$ larger than the uncertainty are shown in red.

To ensure good UQ, we want to minimize the total red area. Thus, we define the quantitative metric as the percentage of pixels that are red and aim for this metric to be low. In Table 5.1, we show the average z-score percentage over posterior sampling results of 50 test unseen observations for both RTMs and CIGs trained networks. From Figure 5.3 we make the following observations: for both models, the total area of the red regions is significantly smaller when the posterior is conditioned on CIGs. We also observe that red regions for the Compass model tend to be associated with areas where the velocity kickback exhibits lateral variations. With few exceptions, we see that red regions for the salt models generated by 'Synthoseis' are located near steeply dipping events. Overall, percentages of red area are significantly reduced thanks to the use of CIGs.



(a) Compass w/ RTMs z-score=10.55%

(b) Compass w/ CIGs z-score=6.08%

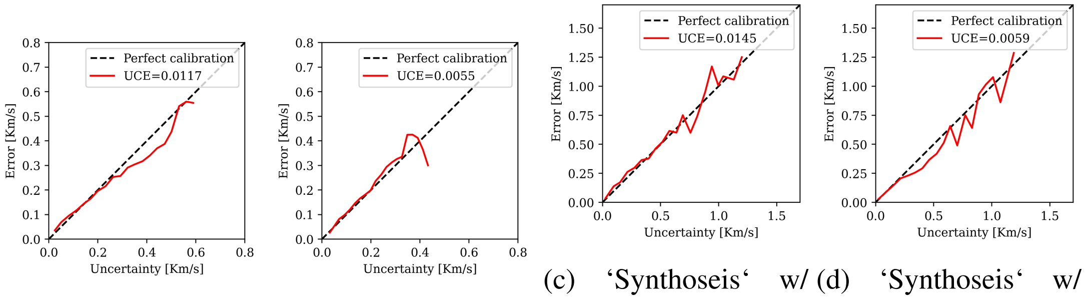
(c) Synthoseis w/ RTMs z-score=7.36%

(d) Synthoseis w/ CIGs z-score=4.48%

Figure 5.3: Comparison of the z-score between conditional diffusion networks trained on RTMs and networks trained on CIGs with 24 non-zero offsets. We desire the z-score (percentage of pixels where error is $2\times$ higher than UQ) to be low.

5.7.2 Degree of calibration

While correctly warning of high errors is an indication that the UQ is reasonable, it does not provide a precise quantitative metric for the correspondence between error and UQ at various magnitudes. For this purpose, we use the calibration test from [78, 79] to quantify the correlation between predicted uncertainty and error. By binning the pixels over different magnitudes, we can build a correlation plot wherein we want pixels that are placed into bins of certain errors magnitudes to have similar uncertainty magnitudes. Figure 5.4 shows the application of this test to the UQ results for the Compass example in Figure 5.1 and the 'Synthoseis' example in Figure 5.2. The uncertainty calibration error (UCE) is a single scalar summarizing the performance of this test by calculating the area between the red curve and the optimal calibration on the diagonal dashed line. Therefore, lower UCE values indicate better-calibrated uncertainty. Again, we observe major improvements due to the use of CIGs for both models. In Table 5.1, we show the average UCE for posterior sampling results on 50 unseen test observations for both RTMs and CIGs networks.

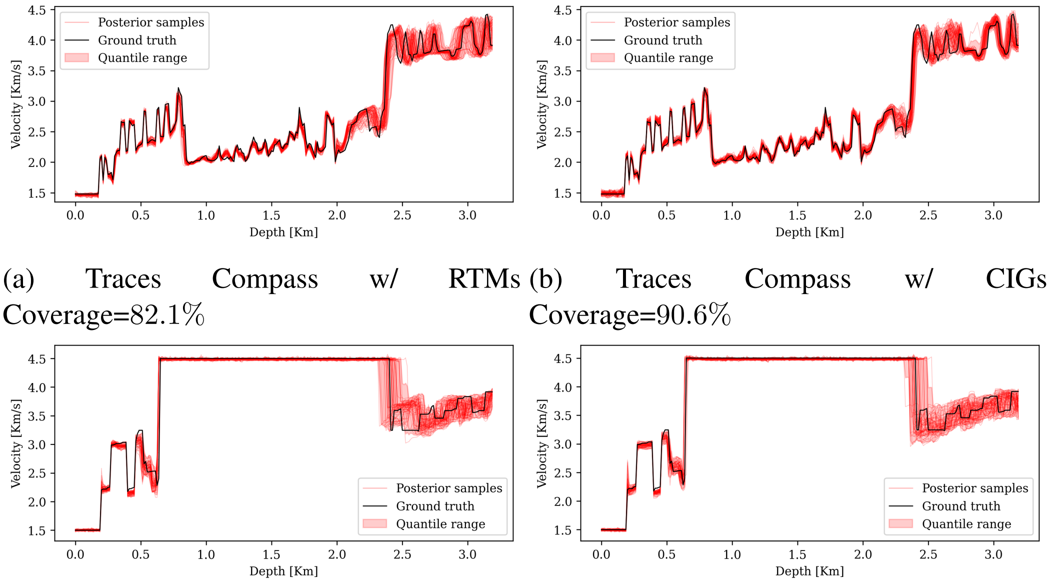


(a) Compass w/ RTMs (b) Compass w/ CIGs (c) 'Synthoseis' w/ RTMs (d) 'Synthoseis' w/ CIGs

Figure 5.4: Comparison of the uncertainty calibration of a conditional diffusion network trained on RTMs compared with a network trained on CIGs with 24 non-zero offsets. For both synthetic datasets, the CIGs-trained network is better calibrated as evidenced by the lower UCE metric.

5.7.3 Posterior coverage percentage

Next, we quantitatively evaluate the coverage of the posterior samples. Coverage measures whether the true velocity model is contained within the spread of the posterior distribution. This is an important characteristic of Bayesian methods, as we want the ground truth velocity model to be included within the posterior samples [169]. To quantify coverage in percentages, we compute the lower and upper percentiles of the posterior samples at each pixel and test if the ground truth velocity model is contained within that range. The quantified metric corresponds to the percentage of pixels for which the calculated range contains the ground truth velocity model. Ideally, this coverage percentage should be high. Visually this test is intuitively understood in Figure 5.5 where we plot samples of the posterior over a single vertical trace and compare them with the ground truth velocity. For our metric, we calculate the range using the upper 99% and lower 1% percentiles. Although we only show a trace in these figures, note that the coverage metric is calculated over all pixels in the reconstructed velocity models. In Table 5.1, we include the average coverage percentage over 50 test samples for both RTMs and CIGs trained networks. We want this metric to be high because we wish our posterior distribution to always contain the ground truth velocity model.



(a) Traces Compass w/ RTMs Coverage=82.1%

(b) Traces Compass w/ CIGs Coverage=90.6%

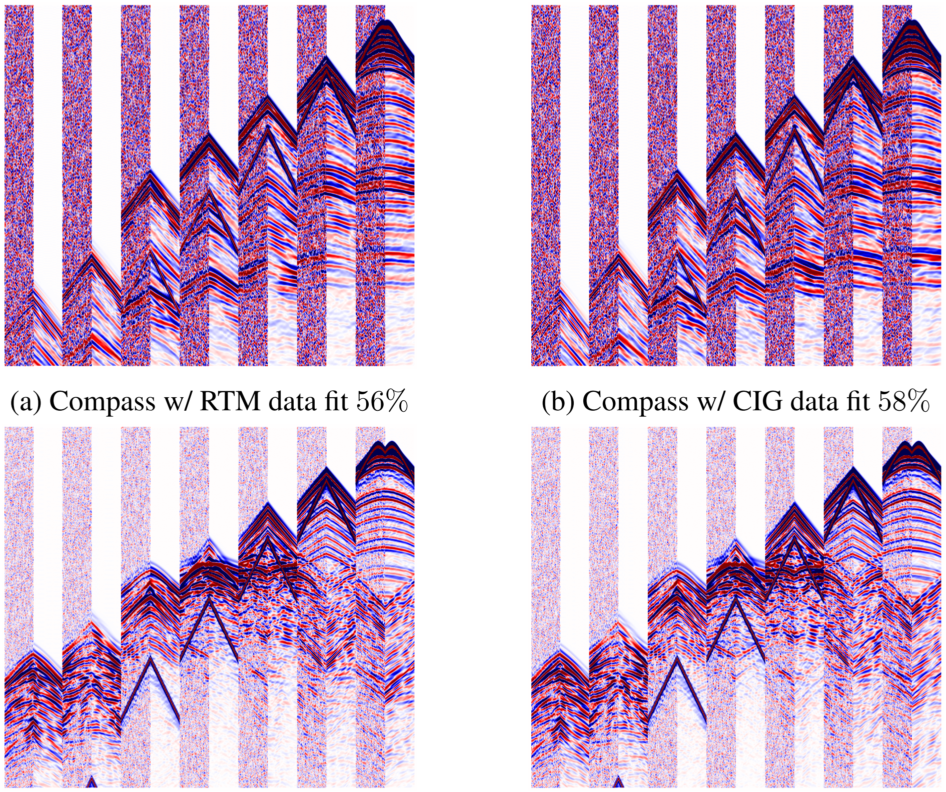
(c) Traces ‘Synthoseis‘ w/ RTMs Coverage=87.0%

(d) Traces ‘Synthoseis‘ w/ CIGs Coverage=89.7%

Figure 5.5: Vertical traces through the posterior samples and the ground truth velocity model used to calculate the posterior coverage metric.

5.7.4 Shot data residual of posterior samples

Even though the above quantitative assessment of UQ is valuable, it relies on having access to ground truth velocity models rendering these metric impractical for field data where the ground truth velocity models are unknown. For this reason, it is as a final check important to make sure that the proposed velocity-model inference produces synthetic gathers that fit observed shot gathers. In other words, while the generative samples $\mathbf{x}_{\text{post}}$ shown previously seem to visually match the correct prior, we will verify that these are indeed Bayesian and respect the data likelihood $p(\mathbf{y} \mid \mathbf{x})$. We take the posterior samples and pass it through the nonlinear forward wave operator $\mathcal{F}(\mathbf{x}_{\text{post}})$ and analyze the fit to the observed shot data $\mathbf{y}^{\text{obs}}$. In Figure 5.6, we show a pairwise binned comparison between the predicted and observed shot gather interleaved with increasing receiver coordinate. To quantify the fit, we take the division between the noise norm and the data residual that we define as $\|\boldsymbol{\varepsilon}\|_2 / \|\mathcal{F}(\mathbf{x}_{\text{post}}) - \mathbf{y}^{\text{obs}}\|_2$ and report this number as a percentage. With 100% corresponding to the residual having the same normed magnitude as the noise, in other words, a perfect fit. For both models, the network conditioned on the CIGs better fits the shot gathers.



(a) Compass w/ RTM data fit 56%

(b) Compass w/ CIG data fit 58%

(c) 'Synthoseis' w/ RTM data fit 41%

(d) 'Synthoseis' w/ CIG data fit 47%

Figure 5.6: Interleaved comparison between paired bins of observed and synthetic shot data.

Results of the four performance metrics: (1) percentage of regions with large errors, (2) degree of calibration, (3) posterior coverage percentage, and (4) posterior fit of shot data are included in Table 5.1 for networks trained with RTMs or CIGs. From these results, we conclude that conditioning on CIGs improves performance with respect to RMSE and SSIM for the recovery results themselves, and with respect to these four metrics that reflect the quality of the UQ. This improvement is due to the CIGs containing more information than the RTMs, even in situations where the migration-velocity models are poor. As a result, CIGs provide a more accurate representation of the original posterior that is conditioned on raw shot records.

We include the performance on the same Compass dataset used in [36], showing that our Diffusion network with average SSIM of 0.85 outperforms the normalizing flow, which had an average SSIM of 0.63 [36]. While this increase in performance is encouraging, we cannot definitively conclude that Diffusion is in general better than normalizing flows since our Diffusion network uses different architectures than the normalizing flow in [36] and also has more training costs. A careful comparison between these two frameworks is beyond the scope of this paper but would be a valuable avenue for future work.

Table 5.1: Image and uncertainty quality metrics on Compass and ‘Synthoseis‘ datasets.

Dataset	RMSE ↓	SSIM ↑	Coverage [%] ↑	UCE ↓	z-score [%] ↓	Data fit [%] ↑
Compass w/o CIGs	0.12	0.81	73.8	0.013	10.7	53.7
Compass w/ CIGs	**0.11**	**0.85**	**74.8**	**0.011**	**9.9**	**54.2**
‘Synthoseis‘ w/o CIGs	0.085	0.91	72.5	0.012	7.8	77.1
‘Synthoseis‘ w/ CIGs	**0.079**	**0.92**	**74.6**	**0.009**	**5.4**	**79.6**

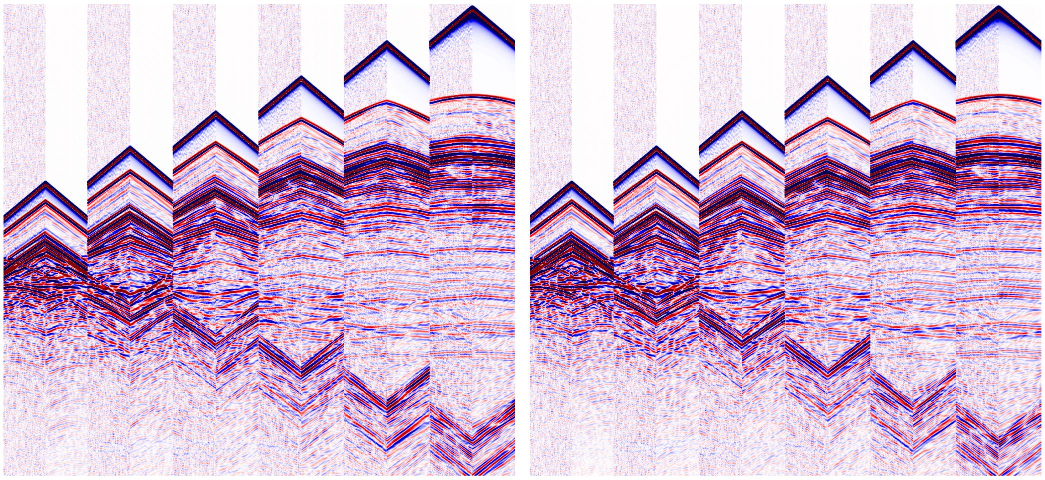
5.8 Complex case studies

So far, the stylized examples have exhibited relatively minor complexity, which in part explains the success of the inference discussed so far where the reconstructed velocity models significantly bring down the residuals. Due to the the always present amortization gap [110, 37], which corresponds to a drop in accuracy due to an attempt to generalize the performance of the network over many velocity models instead of focusing the inference on a single shot dataset as in non-amortized methods [53, 21, 146], we cannot expect the presented amortized approach to perform well on larger models with increased geological complexity. Using an example involving the SEAM salt model [145], we demonstrate how our inference approach can be extended to handle complex salt plays. Finally, we also consider a field dataset to demonstrate the importance of having access to relevant training data.

5.8.1 Salt flooding with ASPIRE

To ameliorate the amortization gap and to accommodate complexities arising from salt in the Gulf of Mexico, we adapt and extend methodologies introduced by [37] and retrain the networks at the next iteration on migrations improved by the average of the posterior samples produced by the current iteration. This type of approach, known as ASPIRE [37], which can be interpreted as a probabilistic loop-unrolled gradient descent algorithm [170, 171], remains amortized and has been demonstrated to close the amortization gap in transcranial ultrasound brain imaging. Motivated by these findings, and by the fact that ASPIRE lends itself well to iterative workflows, we propose a methodology aimed at improving the large data residuals observed in Figure 5.7a. These residuals are mainly due to the complexity of the velocity model and the lack of training data, a situation where amortized methods are known to fail [21, 37].

To remedy this situation and decrease the data misfit, we follow ASPIRE and produce an improved migration-velocity model based on the average of the posterior samples. With this improved velocity model, new CIGs $\overline{\mathbf{y}}_1$ are computed, as in subsection 5.5.3, but now with $\mathbf{x}_{\text{mean}}$—i.e., the posterior mean, serving as the migration-velocity model instead of $\mathbf{x}_0$. After calculating CIGs for each training sample, we train a new network on the updated dataset $\overline{\mathcal{D}}_1 = \{\mathbf{x}^i, \overline{\mathbf{y}}_1^i\}_{i=0}^N$. While this choice of posterior mean (e.g. as opposed to posterior median) is still open to debate, [37] proved that it leads to significant improvements. Although there are additional offline and online costs associated with this method, the inference stage remains relatively computationally efficient with the cost of a single migration per iteration and a low total number of iterations, in this case two iterations.



(a) ASPIRE 1 data fit 14%

(b) ASPIRE 2 data fit 17%

Figure 5.7: Comparison of data misfit of a shot record generated from the posterior samples for two ASPIRE iterations by interweaving traces from the observed shot and simulated shot gathers. The second ASPIRE iteration has improved the data fit by using one more gradient (extended migration) at test time.

Experimental setup

To demonstrate how inference with ASPIRE can be adapted to complex settings, we consider 2D slices through the 3D SEAM model [145], which represents complex salt geometry typical for the Gulf of Mexico. We create a training split of the SEAM model by selecting a continuous subset of crosslines for training. To avoid similarity between 2D slices, we skip 2 km between the nearest training slice and the nearest test slice.

When simulating shot gathers, we generate narrow-offset 2D seismic lines with a grid size of 1744×512 and a spatial discretization of 20 m. A marine dataset is created by simulating shot data with sources and receivers towed at the same depth near the surface. Surface-related multiples are avoided by applying an absorbing boundary condition at the surface. Computational costs are reduced by using a coarse source interval of 1000 m, while receivers are sampled at 20 m and a maximum offset of 6 km. The total recording time for each shot record is 9 s, and Gaussian noise is added to yield a level of 25 dB in a frequency band that matches the source signature. With these settings, 32 shots were simulated for a model with constant density and varying velocity.

Migration-velocity models are created by first removing the salt from the ground truth velocity models, followed by smoothing the result with a Gaussian kernel of grid size 25 (see Figure 5.8. To capture most of the energy, the horizontal subsurface-offset migration utilizes 50 equally spaced offsets, ranging from -2000 m to 2000 m. After imaging the full seismic line, we extract subsets to create the training pairs by taking sliding windows of each 2D line, resulting in grids of size 512×512 and a total of 700 training data pairs. Each ASPIRE iteration takes 12 GPU hours to train. While the conditional Diffusion network is trained on these smaller patches, at inference time the trained network is evaluated on the full grid size of (512×1744). This is feasible because our Diffusion network predominantly relies on convolutional layers, allowing for evaluation on grid sizes larger than those used during training. At inference, we incur the cost of one extended migration (3 min) per ASPIRE iteration and then 8 sec per posterior sample.

Algorithmic innovations

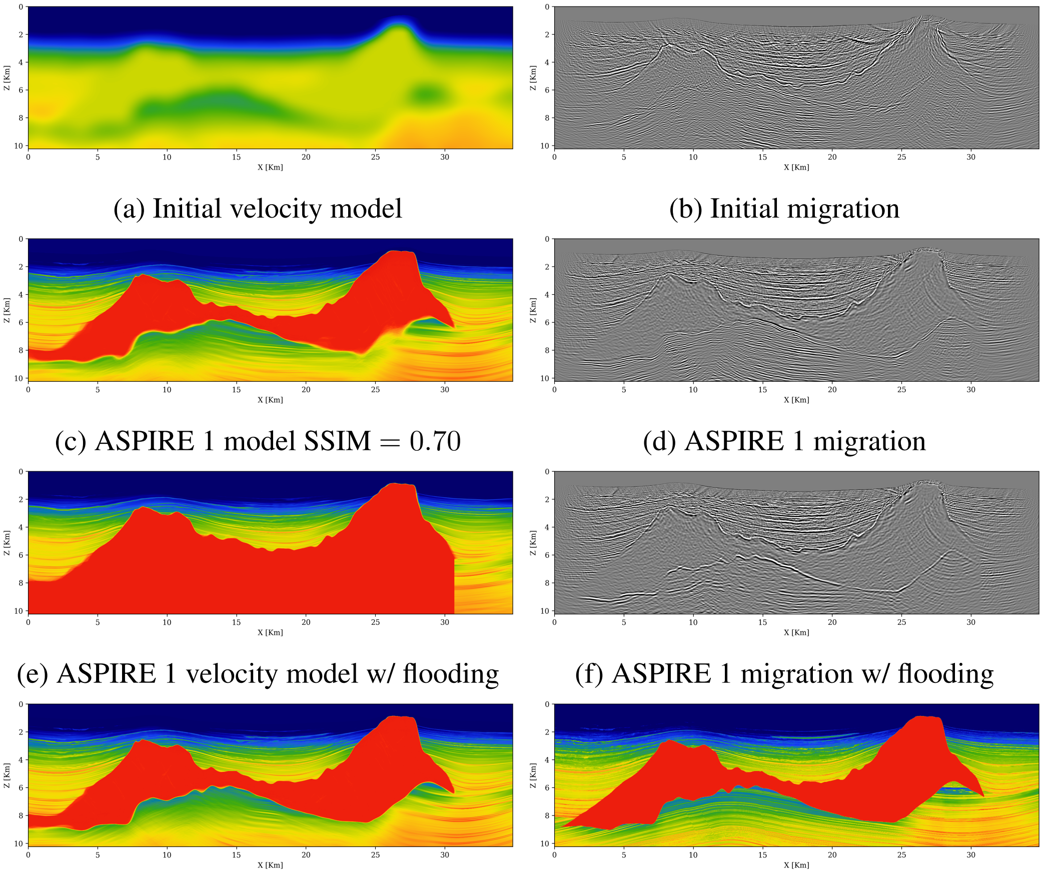
To successfully adapt ASPIRE to the complex setting of SEAM, the following innovations are implemented:

- **Utilization of previous iterates.** In its vanilla implementation, gradient descent and, therefore ASPIRE, base its updates on gradients taken at the current model iterate (read inferred velocity model), ignoring information from previous iterates. Motivated by quasi-Newton optimization methods, [170] proposed to add previous model iterates to improve convergence and approach reminiscent of quasi-Newton methods where previous gradients are used to approximate the Hessian. Similarly, we allow the networks at each iteration to take all previously calculated CIGs and inferred migration-velocity models as input. In our experiments, including this modification significantly improved the performance of ASPIRE.

- **Incorporation of domain knowledge with salt flooding.** To detect the bottom of the salt, we employ the well-established method of salt flooding (see [131] and the references therein) during which top salt is extended downwards. Due to its iterative structure, ASPIRE can easily accommodate salt flooding, so the bottom of the salt is clearly delineated by migration on the second iteration.

Preliminary results

In Figure 5.8, we present the results from two iterations of ASPIRE, starting from the smoothed migration-velocity model depicted in Figure 5.8a. This migration-velocity model is used to produce the initial CIGs, whose zero-offset section is displayed in Figure 5.8b. Based on these CIGs, the first trained ASPIRE produces posterior samples, whose average is plotted in Figure 5.8c. While the top of salt is well recovered, comparison with the ground truth velocity model depicted in Figure 5.8h shows that important details are missing in the bottom salt. This, in turn, leads to poorly resolved bottom salt in the zero-offset migrated section Figure 5.8d. By flooding the salt downwards, as depicted in Figure 5.8e, this issue is largely mitigated, as observed from the migration included in Figure 5.8f, where the bottom salt is delineated sharply and includes the main topological features. This means that the second trained ASPIRE on salt-flooded migrations can correctly infer the salt bottom as shown in Figure 5.8g. Before comparing the migration in this model to a baseline derived from the true model, let us first consider the contraction of the inferred uncertainties resulting from the ASPIRE iterations.



(a) Initial velocity model

(b) Initial migration

(c) ASPIRE 1 model SSIM $= 0.70$

(d) ASPIRE 1 migration

(e) ASPIRE 1 velocity model w/ flooding

(f) ASPIRE 1 migration w/ flooding

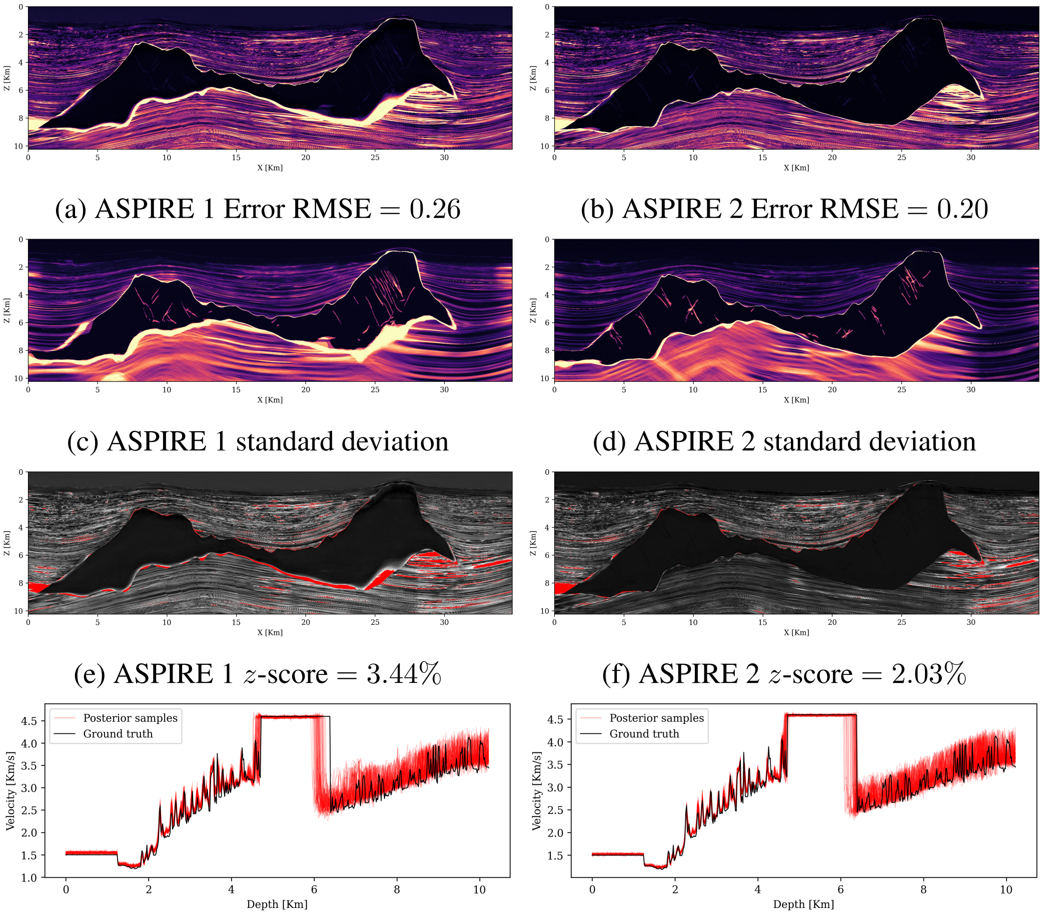
(g) ASPIRE 2 model SSIM $= 0.72$

(h) Ground truth velocity model

Figure 5.8: Comparison of posterior means yielded by two iterations of ASPIRE and their corresponding reverse-time migrations.

In addition to providing increasingly better estimates for the velocity model, ASPIRE also includes UQ at each iteration. In Figure 5.9, different aspects of the inferred uncertainties for ASPIRE 1 and 2 are illustrated by means of plots for errors included in Figure 5.9a and Figure 5.9b; posterior standard deviation in Figure 5.9c and Figure 5.9d; z-score in Figure 5.9e and Figure 5.9f; and finally plots for the coverage in Figure 5.9g and Figure 5.9h. From these plots, the following observations can be made: First, errors with respect to the ground truth velocity model at the bottom salt are greatly reduced by ASPIRE 2 due to the salt flooding. However, errors remain at both ends of the salt due to a lack of illumination. As expected, errors remain in the sediments below the salt. Second, as expected, predicted uncertainties at the bottom salt are reduced for ASPIRE 2 (see Figure 5.9c and Figure 5.9d).

The uncertainty also correlates reasonably well with the errors. We also observe the appearance of dirty salt, which can be explained by the fact that the salt in the SEAM dataset includes dirty salt. Third, with very few exceptions there is a drastic improvement in the overall z-score plots (figures Figure 5.9e and Figure 5.9f) for ASPIRE 2. Most notably, z-scores improve significantly at the bottom-salt interface and within the sedimentary layering under the salt. Finally, the coverage of the posterior samples is also improved. While the ground truth bottom salt was missed in ASPIRE 1, the posterior samples yielded by ASPIRE 2 clearly contain the step-out of the salt.



(a) ASPIRE 1 Error RMSE = 0.26

(b) ASPIRE 2 Error RMSE = 0.20

(c) ASPIRE 1 standard deviation

(d) ASPIRE 2 standard deviation

(e) ASPIRE 1 z-score = 3.44%

(f) ASPIRE 2 z-score = 2.03%

(g) ASPIRE 1 traces coverage = 61.79%

(h) ASPIRE 2 traces coverage = 70.24%

Figure 5.9: Comparison between recovery and UQ quality yielded by ASPIRE 1 and 2.

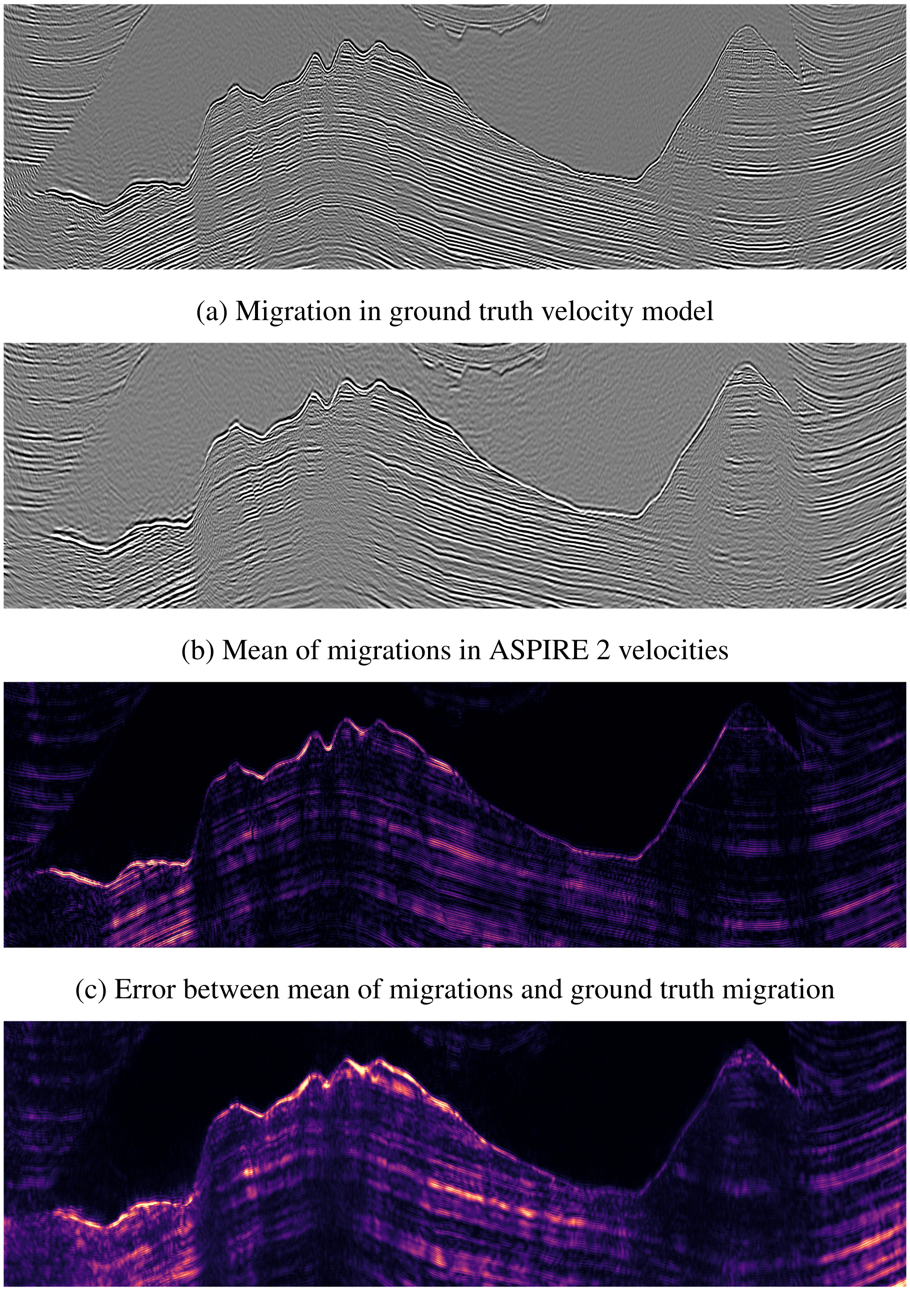
Moreover, visual improvements in the second ASPIRE iteration are supported by quantitative uncertainty quality measures listed in the captions of Figure 5.9 and in Table 5.2, summarizing performance over eight unseen test 2D lines. Overall, the progression of results correlates well with the iterative focus of ASPIRE: the first iteration targets the top of the salt and the second iteration targets the bottom of the salt. The corresponding shot gathers for the different ASPIRE iterations are included in Figure 5.7 and confirm that the model improvements lead to improved data fit.

Table 5.2: Image and uncertainty quality metrics on SEAM dataset ASPIRE iterations.

Dataset	RMSE ↓	SSIM ↑	Coverage [%] ↑	UCE ↓	z-score [%] ↓	Data fit [%] ↑
ASPIRE 1	0.25	0.71	66.6	0.043	$\mathbf{3.44}$	13.3
ASPIRE 2	$\mathbf{0.21}$	$\mathbf{0.73}$	$\mathbf{67.3}$	$\mathbf{0.042}$	3.96	$\mathbf{15.5}$

Quality assessment

While the recovered velocity models and uncertainty quality metrics show significant improvements between ASPIRE 1 and 2, the ultimate goal is to assess the impact of these improvements on the final product, namely, the migrated image in areas below the salt. For this purpose, we included Figure 5.10, which contains results of RTM with the ground truth velocity model; the mean of reverse-time migrations carried out in 16 posterior samples for the velocities produced by ASPIRE 2; error with respect to the migrated image with the ground truth velocity model; and the standard-deviation of reverse-time migrations in posterior velocity models. From these plots, the following observations can be made: First, the migrations in the ground truth velocity model and the inferred mean of the multiple RTMs are close, even though issues remain. These include: errors in the delineation some areas of bottom salt and minor shifts in imaged sedimentary layers below the salt. While there are errors, we emphasize that errors are expected to occur in this ill-posed problem, but that our solution comes with uncertainty quantification that powerfully points to areas of error (see figures Figure 5.10c and Figure 5.10d). We observe that the predicted uncertainty is overly cautious, predicting larger variability than evidenced in the plot of errors with respect to the migration with the ground truth velocity model. We consider the fact that uncertainty being overestimated as advantageous since it offers an additional safeguard to overinterpretation of imaged reflectors as opposed to being overconfident. Both migrations suffer from an overprint due to areas of relatively poor illumination due to the interplay between small offsets and complexity of the salt geometry, which gives rise to (de)focusing effects that explain variations in the amplitudes of the imaged reflectivity under the salt.



(a) Migration in ground truth velocity model

(b) Mean of migrations in ASPIRE 2 velocities

(c) Error between mean of migrations and ground truth migration

(d) Standard deviation of migrations in ASPIRE 2 models

Figure 5.10: The migration in our final velocity model is close to the migration in the ground truth model.

5.8.2 Field data proof of concept

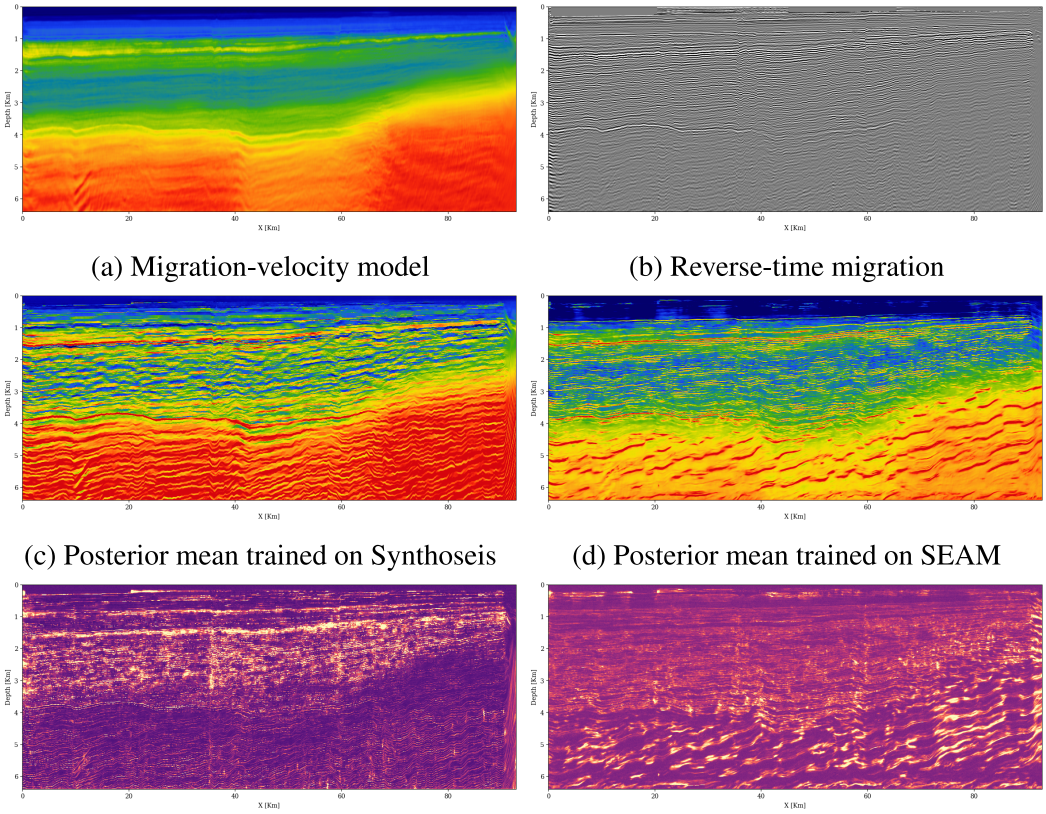
Our aim is to develop ML-enabled workflows for probabilistic velocity model building capable of scaling to industry-sized seismic datasets. To demonstrate progress and challenges toward this goal, we apply the same approach with pre-trained networks from the previous experiments to a large shallow-water seismic 2D line. The aim of this exercise is twofold: First, we aim to demonstrate that the presented methodology can be applied to field data. Second, we aim to showcase the main limitations of the presented approach—i.e., its reliance on training datasets that are pertinent to the geological setting under consideration.

Shallow-water field data

As a first attempt to apply the presented workflow to field data, we consider the migration and velocity models derived from the ”Galactic dataset” made available by Woodside Energy Ltd under a [Creative Commons license](https://creativecommons.org/licenses/by-sa/4.0/). These derivatives are obtained with a traditional workflow, consisting of migration-velocity building with FWI and RTM, as shown in figures Figure 5.11a and Figure 5.11b. We have not made any modifications to the original work.

The 2D line of the Galactic dataset being considered has a grid size of (512×7024). To produce samples, the migration shown in Figure 5.11b served as input to conditional neural networks trained on velocity models from the ‘Synthoseis‘ and SEAM datasets. In this case, we trained a network on ‘Synthoseis‘ samples that did not contain salt. Since the Galactic dataset does not include non-zero offset migrations, we use the RTM included in Figure 5.11b as input, producing the posterior means shown in figures Figure 5.11c and Figure 5.11d. For this large line, each posterior sample takes 30 seconds to generate. Both estimates are obtained from 64 samples of the posterior.

The posterior mean estimates produced by either network do not seem realistic because they are strongly biased toward the datasets these networks were trained on. The fact that the result obtained with ‘Synthoseis‘ looks more reasonable likely stems from the fact that ‘Synthoseis‘ dataset contains more variability in its training set. Still, this example underlines the importance of having access to pertinent training data, a topic we will address in the discussion.



(a) Migration-velocity model

(b) Reverse-time migration

(c) Posterior mean trained on Synthoseis

(d) Posterior mean trained on SEAM

(e) 2xUQ trained on Synthoseis

(f) UQ trained on SEAM

Figure 5.11: Field-data results trained on Synthoseis versus SEAM. (a) Migration velocity model used to produce RTM. (b) Observed RTM. (c) Mean of Posterior samples from Synthoseis. (d) Mean of Posterior samples from SEAM. We recommend zooming in on a computer screen to see these figures.

5.9 Discussion and future work

Results obtained for the stylized examples and complex case studies suggest that probabilistic velocity model building is possible as long as training pairs in the form of velocity models and migrated image data are available. However, having access to representative training velocity models is challenging in practice. The examples presented here fall short because they are biased by our geological understanding, as expressed in synthetic earth models. To overcome these limitations and other challenges in applying the proposed methodology to more realistic settings, we strongly encourage our community to follow David Donoho's [172] recipe for success in data science, which includes *(i)* making (training) datasets available; *(ii)* releasing research findings that can be reproduced in a frictionless manner; and *(iii)* establishing benchmarks to compare results based on quantitative figures of merit. We argue that this work contributes towards items *(ii-iii)*. Item *(i)* remains challenging, but below we outline a possible strategy for addressing this item.

By using the latest tools from generative AI, [173] recently developed a framework to train conditional neural networks to generate realistic Earth models parameterized by velocity. In their approach, Diffusion networks are trained on pairs imaged field data and well-log data, both residing in national data repositories such as the one maintained by the National Transition Authority in the United Kingdom. We are currently in the process of curating hundreds of 2D seismic image-well pairs, so that a foundational model can be trained with which realistic synthetic velocity can be generated without ever requiring access to velocity models themselves. As more curated datasets become available, the quality of this foundational model will improve, as will machine learning techniques that rely on training data.

An important avenue for future work is to explore non-amortized methods that would build on top of the amortized results shown here. Specifically, algorithms that use the forward operator at inference time to specialize to the observation and improve performance [21, 37, 38]. This approach becomes especially pertinent when applying our method to field data.

5.10 Conclusions

We implemented machine learning enabled workflows that, through the use of modern conditional Diffusion neural networks, advance the state of the art in amortized migration-velocity model building with uncertainty quantification. In this context, amortization means that our network generalizes across different shot datasets. We also proposed a set of performance metrics as a benchmark to compare the effectiveness of different uncertainty quantification methods. For complex salt scenarios, we developed a new iterative workflow incorporating salt flooding. Using our Bayesian probabilistic approach, multiple migration-velocity models are generated from poor initial velocity models (e.g., models without salt). These samples from the posterior distribution produce different reverse-time-migrated images that contain variability propagated from the uncertainty in the inferred velocity model. Thanks to machine learning, the proposed approach remains computationally feasible at inference time, with considerable but manageable offline training costs. The results represent a demonstration of machine learning enabled probabilistic velocity building from short-offset acoustic reflection-only data. While our evaluation on field datasets is still in its early stages, the initial results are promising. However, applying the method to field data revealed a bias toward the synthetic data distribution on which the networks were trained. This finding highlights the need to curate realistic training datasets to train the next generation of generative networks for solving geophysical inverse problems.

CHAPTER 6

PROBABILISTIC BAYESIAN OPTIMAL EXPERIMENTAL DESIGN USING CONDITIONAL NORMALIZING FLOWS

SUMMARY

Bayesian optimal experimental design (OED) seeks to conduct the most informative experiment under budget constraints to update the prior knowledge of a system to its posterior from the experimental data in a Bayesian framework. Such problems are computationally challenging because of (1) expensive and repeated evaluation of some optimality criterion that typically involves a double integration with respect to both the system parameters and the experimental data, (2) suffering from the curse-of-dimensionality when the system parameters and design variables are high-dimensional, (3) the optimization is combinatorial and highly non-convex if the design variables are binary, often leading to non-robust designs. To make the solution of the Bayesian OED problem efficient, scalable, and robust for practical applications, we propose a novel joint optimization approach. This approach performs simultaneous (1) training of a scalable conditional normalizing flow (CNF) to efficiently maximize the expected information gain (EIG) of a jointly learned experimental design (2) optimization of a probabilistic formulation of the binary experimental design with a Bernoulli distribution. We demonstrate the performance of our proposed method for a practical MRI data acquisition problem, one of the most challenging Bayesian OED problems that has high-dimensional (320 × 320) parameters at high image resolution, high-dimensional (640 × 386) observations, and binary mask designs to select the most informative observations.

6.1 Introduction

When solving an inverse problem, the goal of experimental design is to chose how to observe data from the field that is used to infer an unknown parameter. The process that models the data acquisition is denoted as the forward process written

$$\mathbf{y} = \mathbf{M}(\mathcal{F}(\mathbf{x})) + \varepsilon$$

where $\mathcal{F}(\cdot)$ is the forward operator that takes the unknown $\mathbf{x}$ to the observation space, $\mathbf{M}(\cdot)$ is the observation process that need not be linear and ε is corruption noise. Since experimentalists have control over the observation process, it is desirable to control this in order to best inform downstream inferences of $\mathbf{x}$ i.e experimental design. As an illustration, imagine a doctor deciding where to place a handheld ultrasound on a patient to best infer the state of a internal organ.

In a Bayesian framework, the experimental design is based on quantities related to the posterior distribution $p(\mathbf{x}|\mathbf{y})$. Once the posterior has been estimated, different design optimality choices can be made. These options include A-optimal, D-optimal, and the expected information gain (EIG) [174]. Due to its close connection with the posterior likelihood (the quantity used to train normalizing flows) we focus on the EIG:

$$EIG(\mathbf{M}) = \mathbb{E}_{p(\mathbf{y}|\mathbf{M})} \; [D_{KL}(p(\mathbf{x}|\mathbf{y}, \mathbf{M}) \, || \, p(\mathbf{x}))]$$

which measures the information gain between the prior information and the information gained by performing an experiment using $\mathbf{M}$ to acquire data $\mathbf{y}$. Importantly, the expectation $\mathbb{E}_{p(\mathbf{y}|\mathbf{M})}$ means that the information gain is averaged over the distribution of possible observations i.e it describes the best experiment on average for the range of observations that is expected to be encountered. However, optimizing EIG is challenging due to complexities like the ”double intractability” problem, which arises from the necessity to evaluate two expectations [175]. Our approach tackles this with a data-driven optimization of the EIG that combines simulation based inference [98], likelihood-based generative models [94] to tractably find optimal designs and a probablistic interpretation of the design parameters to facilitate its optimization.

6.2 Methodology

To demonstrate our scalable technique for Bayesian experimental design, we first show the connection between EIG and likelihood based generative models. Secondly, we present conditional normalizing flows as the key tool due to their exact likelihood evaluation and their invertible architectures that enable memory efficient training. Then we show how a probabilistic interpretation of binary design masks alleviates optimization challenges. Finally, we setup a demonstration of our technique as applied to high dimensional inverse problem related to MRI medical imaging.

6.2.1 Normalizing flows learn the expected information gain

In the context of summary statistics, [176] noted that a summary statistic $\bar{\mathbf{y}}$ can be interpreted as the transformation on observations $\mathbf{y}$ that both maximizes the posterior likelihood $p(\mathbf{x}|\bar{\mathbf{y}})$, and also maximizes the expected information gain $EIG(\bar{\mathbf{y}})$. By interpreting the result from an experiment $\mathbf{M}$ as a summary statistic, we use a similar derivation to show that maximizing the EIG of a design $\mathbf{M}$ is equivalent to maximizing the posterior likelihood that the design induces in expectation over a joint probability $p(\mathbf{x}, \mathbf{y}|\mathbf{M})$:

$$\begin{aligned}\max_{\mathbf{M}} \; EIG(\mathbf{M}) =& \mathbb{E}_{p(\mathbf{y}|\mathbf{M})} \; [D_{KL}(p(\mathbf{x}|\mathbf{y}) \,||\, p(\mathbf{x}))] \qquad (6.1)\\ =& \mathbb{E}_{p(\mathbf{y}|\mathbf{M})} \; \left[\; \mathbb{E}_{p(\mathbf{x}|\mathbf{y})} \left[\log p(\mathbf{x}|\mathbf{y}) - \log p(\mathbf{x})\right]\right] \qquad (6.2)\\ =& \mathbb{E}_{p(\mathbf{y}|\mathbf{M})} \; \left[\; \mathbb{E}_{p(\mathbf{x}|\mathbf{y})} \left[\log p(\mathbf{x}|\mathbf{y})\right]\right] \qquad (6.3)\\ =& \mathbb{E}_{p(\mathbf{x},\mathbf{y}|\mathbf{M})} \; \left[\log p(\mathbf{x}|\mathbf{y})\right]. \qquad (6.4)\end{aligned}$$

We have simplified the expressions by using the fact that the maximization is constant with respect to the prior $p(\mathbf{x})$ in line 3 and then the law of total expectation in the last line. Crucially, the final expression is equivalent to the quantity (posterior log likelihood) that is used to guide optimization of likelihood-based conditional generative models:

$$\max_{\theta} \mathbb{E}_{p(\mathbf{x},\mathbf{y})} \left[\log p_{\theta}(\mathbf{x}|\mathbf{y})\right], \tag{6.5}$$

where θ are the network weights that define the conditional generative model. The equivalence between the EIG and the objective in Equation Equation 6.6 implies that we can setup a joint optimization:

$$\max_{\theta,\mathbf{M}} \mathbb{E}_{p(\mathbf{x},\mathbf{y}|\mathbf{M})} \; \left[\log p_{\theta}(\mathbf{x}|\mathbf{y})\right]. \tag{6.6}$$

and that the gradient signal from the objective can be back-propagated to the design $\mathbf{M}$ to calculate an update $\nabla_{\mathbf{M}}$ and it will point in the direction of increased expected information gain. Using conditional generative models for EIG optimization has been previously suggested by separate arguments in [177]. A particular class of models that is trained with this objective are conditional normalizing flows [31, 178] a type of generative model that is known to be universal approximators of distributions [179]. Conditional normalizing flows are an attractive choice because of their exact likelihood evaluations that enables the aforementioned joint optimization in the first place and also because they are invertible by design thus lead to efficient memory use during training on large inputs. In this work, we exploit this equivalence and conditional normalizing flows to perform Bayesian optimal experimental design on a large-scale imaging problems. Next we meet the challenge of finding binary designs by reinterpreting the design as a probabilistic sampling pattern.

6.2.2 Probabilistic mask design

We express the design parameters as binary mask where a 1 means that we have placed a sensor at this location and 0 otherwise. Instead of directly optimizing for binary mask $\mathbf{M}$, we optimize for parameters of a Bernoulli distribution for each possible binary value. This is achieved by following the methods from [180] where we parameterize the distribution as real values $\mathbf{w}$ then create a binary the mask by applying the indicator function $\mathbb{1}_{\mathbf{w}<\mathbf{u}}$ where $\mathbf{u}$ is sampled from the uniform distribution $\mathbf{u} \sim U(0,1)$. We insure that the sampling budget is respected by normalizing $\mathbf{w}$ such that the average value is kept equal to the budget s. Thus the binary mask is defined as

$$\mathbf{M}(\mathbf{w}) := \mathbb{1}_{\mathrm{s}\frac{\mathbf{w}}{\overline{\mathbf{w}}}<\mathbf{u}} \tag{6.7}$$

$$\text{where } \mathbf{u} \sim U(0,1). \tag{6.8}$$

When optimizing for $\mathbf{w}$ with gradient descent, we calculate the gradients of the indicator function with the pass-through gradient approximation from [181] by treating the indicator function as the identity during back propagation. We chose to express the binary mask as in a probabilistic manner for two reasons. First, probabilistic sampling of the mask during training aids in jumping out of local minima that would be challenging to avoid if the mask was deterministic. Secondly, because the optimized mask represents a relative likelihood of sampling then the sampling budget s can be changed post-hoc by the user. In other words, changing the sampling budget does not require retraining the network instead requires simple a re-scaling of the learned probabilistic mask $\mathbf{w}$.

TO BE EXACT WE ARE SOLVING THE EIG BUT WITH A SPARCITY CONSTRAINT

6.2.3 Experimental design for high-dimensional medical imaging

Magnetic Resonance Imaging (MRI) is an important modality that is routinely used in the diagnosis of diseases related to cancer, neurology and the musculoskeletal system. The measured field by an MRI machine is spatial frequencies of the patient tissue. Each measurement takes time thus taking less measurements translates to savings in money due to expensive operation and also results in increased patient comfort.

We use the FAST MRI knees dataset [103] to create a training pairs comprised of $(320\times$

320) ground truth multi-coil images and (640×386) single-coil k-space observations as complex numbers. We use 1800 training samples (a relatively low number as compared to related work [180, 182] that used between 17k-34k training pairs.)

Given the Fourier transformation operator $\mathbf{A}$, we solve the following maximization with stochastic gradient descent to jointly train a conditional normalizing flow f_θ and the probabilistic design parameterized by $\mathbf{w}$:

$$\hat{\theta}, \hat{\mathbf{w}} = \underset{\theta, \mathbf{w}}{\operatorname{argmax}} \frac{1}{N} \sum_{i=1}^{N} \left(-\frac{1}{2} \| f_\theta(\mathbf{x}^{(i)}; \mathbf{A}^\top \mathbf{M}(\mathbf{w}) \odot \mathbf{y}^{(i)}) \|_2^2 + \log |\det \mathbf{J}_{f_\theta}| \right), \tag{6.9}$$

where $\mathbf{A}^\top$ is the adjoint Fourier transform, and $\mathbf{J}_{f_\theta}$ is the Jacobian of the normalizing flow as is needed for maximum-likelihood training as per the change of variables formula. Here we implement our conditional normalizing flow with InvertibleNetworks.jl [33] that takes advantage of the invertiblity of normalizing flow layers to achieve low-memory requirements during training. After training, the network is an amortized sampler for the posterior distribution for all observations in the training distribution. Given the optimized network parameters and the optimal design we sample from the posterior distribution by first acquiring data $\mathbf{y}^{obs}$ from the field as prescribed by the optimal design $\hat{\mathbf{M}}(\hat{\mathbf{w}})$ and then generating posterior samples by passing Gaussian noise through the inverse network as such:

$$\mathbf{x} = f_{\hat{\boldsymbol{\theta}}}^{-1}(\mathbf{z}; \mathbf{A}^\top \mathbf{y}^{obs}) \tag{6.10}$$

$$\text{where } \mathbf{z} \sim \mathcal{N}(0,\, I). \tag{6.11}$$

Note that we decided to process the sub-sampled data with the adjoint Fourier operator $\mathbf{A}^\top$ that incurs an additional computational cost during training for our method. For our training hardware, the baseline took 16 hours to train and our method took 20 hours on a single GPU. Although the adjoint operator is used during posterior sampling in Equation Equation 6.10, we only need to calculate it once then an arbitrary number of posterior samples can be generated by resampling the Gaussian noise $\mathbf{z}$.

As far as we understand, this is the first work that explores a scalable solution for Bayesian experimental design in MRI so we build our own baseline that used the exact same density estimator f_θ. This allows us to understand the uplift achieved by doing experimental design while controlling for the neural network architectures used and also the fact that our solution provides probabilistic solution. In other words, the baseline solves Equation Equation 6.9 but only optimizes the network parameters θ and uses a fixed mask $\mathbf{M}_b$. We hand craft the fixed baseline mask by following the methods in [180, 183] and craft a mask that captures low frequencies and then randomly samples high frequencies as shown in Figure Figure 6.1c.

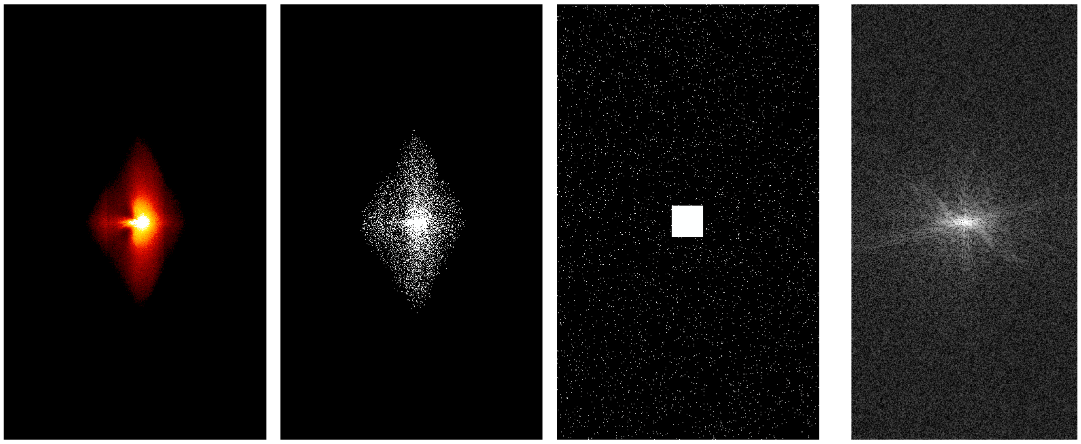
6.3 Results

After training and testing on the FASTMRI dataset, we compare the posterior inference when conditioning on data produced by a hand-crafted experiment as compared to our method that is conditioned on data that is measured with our optimal design based on maximized expected information gain. We show that posterior samples from our method are more realistic and lead to increased performance in downstream image reconstruction tasks.

6.3.1 Optimized experimental designs

After training, our method produces two outputs: an amortized posterior sampler $p_{\hat{\theta}}$ and the optimized mask $\hat{\mathbf{M}} = \mathbf{M}(\hat{\mathbf{w}})$. We train and test using a design budget of 2.5% of all k-space frequencies $s = 0.025$. Our final optimized probabilistic mask is shown in Figure Figure 6.1a, expressed as a sampling density. Using Equation Equation 6.7, we transform the probabilistic mask to final binary mask that can be used in the field to decide which k-space locations should be sampled to collect data.

We make the following observations on our optimized design shown in Figure Fig-



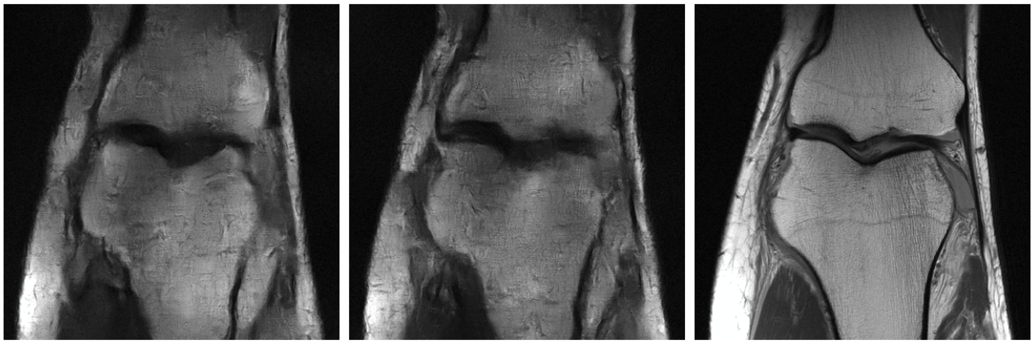
(a) Learned probabilistic design $\hat{\mathbf{w}}$ (b) Learned binary design $\hat{\mathbf{M}}$. (c) Hand-crafted baseline design $\mathbf{M}_b$ (d) Fully sampled data

Figure 6.1: Optimized design compared to hand-crafted design.

ure 6.1a: *(i)* the optimized design has learned from the training set that it needs to emphasize sampling of low frequencies in the center of the data space but has done so in a smoother more efficient manner than the baseline. *(ii)* the learned design has chosen an anisotropic emphasis on the vertical events in kspace evidenced by the ellipsoid shape of the mask *(iii)* as seen by the asymmetric emphasis of the learned design on the right side of frequency space, the learned design has taken advantage of the Hermitian symmetry that is inherent to MRI machines [184] without any input of this domain knowledge.

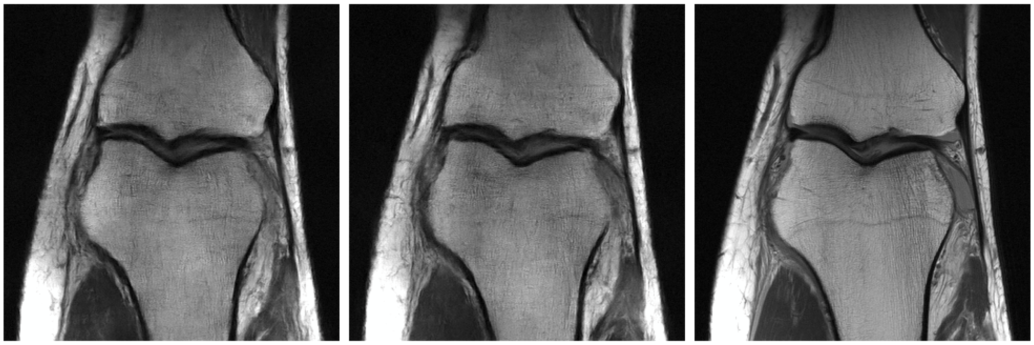
Both methods train an amortized posterior sampler that is sampled by passing Gaussian noise through the inverse network as in Equation Equation 6.10. We compare the quality of the posterior samples with the unoptimized mask in Figure Figure 6.2 with the samples with the optimized mask in Figure Figure 6.3. The posterior samples that used our optimized mask show sharper and more realistic features throughout the reconstruction.

By calculating the intrasample statistics of the posterior samples, we study the behaviour of the posterior mean and the posterior standard deviation. We calculate these statistics by taking posterior samples as calculated in Equation Equation 6.10 then calcu-



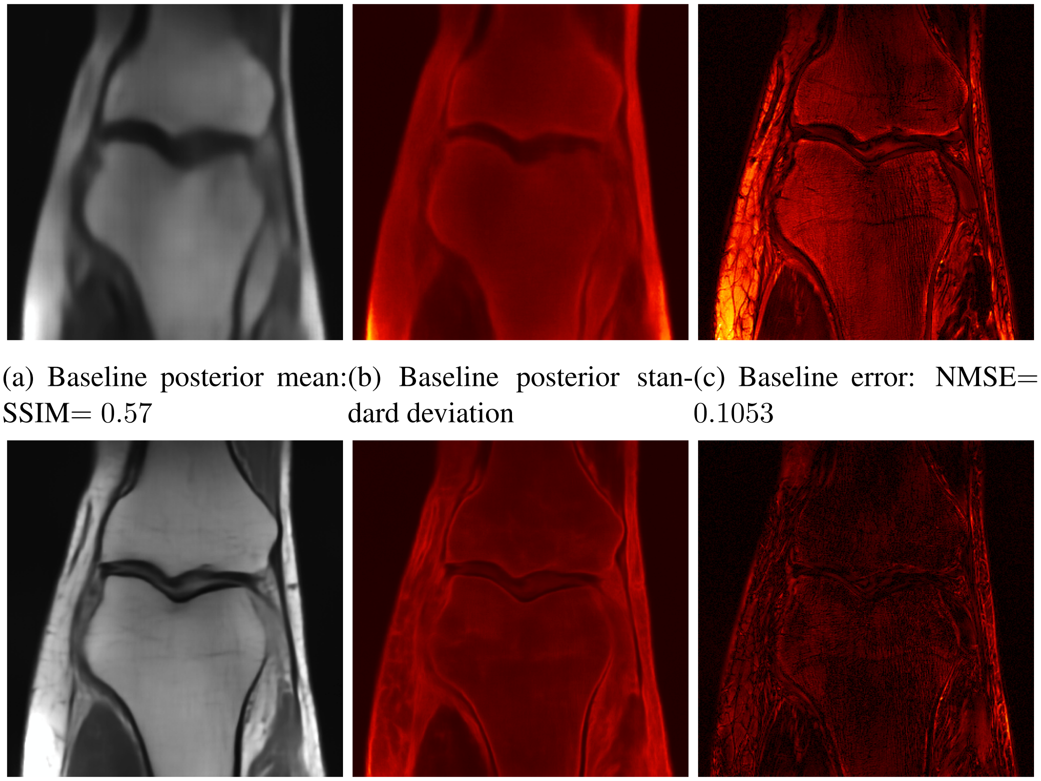
(a) Posterior sample $\mathbf{x} \sim p_{\hat{\theta}}$ (b) Posterior sample $\mathbf{x} \sim p_{\hat{\theta}}$ (c) Reference image

Figure 6.2: Posterior samples from the baseline method with hand-crafted design.



(a) Our posterior sample $\mathbf{x} \sim p_{\hat{\theta},\hat{M}}$ (b) Our posterior sample $\mathbf{x} \sim p_{\hat{\theta},\hat{M}}$ (c) Reference image

Figure 6.3: Posterior samples from our method with optimized design



(a) Baseline posterior mean: SSIM= 0.57

(b) Baseline posterior standard deviation

(c) Baseline error: NMSE= 0.1053

(d) Our posterior mean: SSIM= 0.68

(e) Our posterior standard deviation

(f) Our error: NMSE= 0.0223

Figure 6.4: Pointwise statistics from the baseline compared to our method.

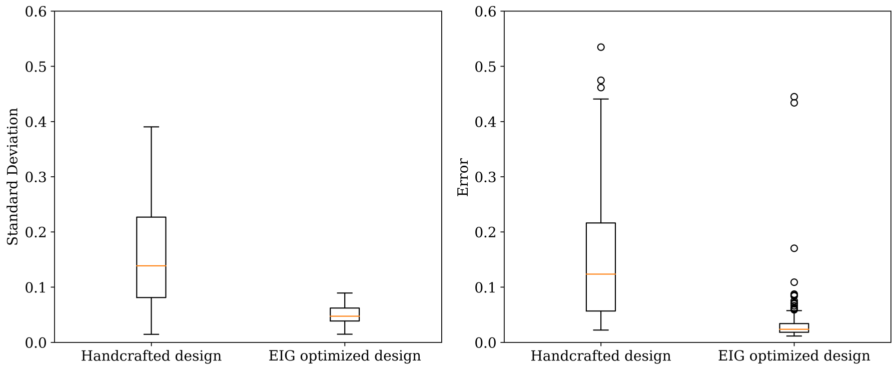
lating the empirical statistics of the mean and standard deviation:

$$\mathbb{E}\,p := \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x}|\mathbf{y})}\left[\mathbf{x}\right]$$

$$\mathbb{STDEV}\,p := \sqrt{\mathbb{E}_{\mathbf{x}\sim p(\mathbf{x}|\mathbf{y})}\left[(\mathbf{x}-\mathbb{E}\,p)^2\right]}.$$

We plot these statistics for the same test sample in Figure Figure 6.4. We also plot the error of the posterior mean with respect to the reference image and note that for the test sample selected, our method shows less total uncertainty and also less error.

To quantify the gain in quality achieved by our optimal design as compared to the handcrafted design, we calculate an image quality metric of normalized mean squared error



(a) Our optimized design reduces uncertainty

(b) Reduced uncertainty leads to lower error

Figure 6.5: Comparing reduction in uncertainty achieved by our method.

(NMSE) between the posterior samples' mean and the ground truth over an unseen test set of 100 samples. We also quantify the overall uncertainty of a certain posterior inference by calculating the normalized mean squared standard deviation. Since the EIG is based on maximizing the posterior likelihood or equivalently lowering the posterior entropy then we expect our optimal design to lead to inference with less uncertainty. We summarize these results in Figure Figure 6.5a showing that the uncertainty is consistently reduced by our method over the test set and furthermore that this translates to overall less error in the reconstruction given by the posterior mean as shown in Figure Figure 6.5b.

6.4 Related work

We briefly outline similar work to this manuscript. The authors [185] use a pretraining phase that allows to efficiently compute the EIG based on a computed MAP approximation then efficient calculations of the posterior covariance. Given this EIG estimate, they proposed a set of greedy algorithms can be used to find demonstrably optimal EIG designs as compared to baselines. Similar to our work with normalizing flows, [186], used the theory laid out in [177] to train an amortized conditional normalizing flow that works as an approximation to the EIG and then shows that the EIG is accurate as compared to a known analytical EIG. Although our methods are similar, we also learn the design jointly during normalizing flows training.

In the medical field, [187] trained a policy guided agent paired with a high quality reconstruction process to find optimal selections of measurement angles in sequential CT imaging. Specifically for MRI [188] showed that particular qualities such as edge promotion can be incorporated into a Bayesian experimental design framework. In the seismic field, [189] used a similar probablistic interpretation to the design as ours and added extra steps to insure that the final binary mask fulfilled certain required sampling qualities.

6.5 Conclusions

We have demonstrated the implementation of Bayesian experimental design on a realistic medical imaging problem. Our method relies on the exact likelihood density evaluation of normalizing flows leading to a simple method to jointly learn variational inference parameters and experimental design parameters. On top of this, the invertible architecture of normalizing flows enabled training on a large-scale imaging problem in MRI. Due to an absence of previous literature on solving this experimental design problem from a Bayesian perspective, we made compared our method with an equivalent Bayesian approach that does not use experimental design. Our experiments show gains in two downstream metrics: the reduction of the uncertainty in the posterior inference and the quality of the posterior samples' mean as measured with image quality metrics.

CHAPTER 7

EXPLOITING LONG-RANGE CORRELATIONS: THE PITFALLS OF PATCH TRAINING FOR UNCERTAINTY QUANTIFICATION IN LARGE SCALE IMAGING

SUMMARY

Uncertainty quantification is essential for risk-averse imaging applications, where Bayesian methods excel by naturally representing uncertainty through posterior variance. However, scaling these Bayesian methods to large problems is challenging due to the curse of dimensionality. We introduce the use of normalizing flows to efficiently train amortized Bayesian methods for large-scale 2D and 3D imaging problems. Utilizing a memory-efficient implementation of normalizing flows, we achieve two main objectives: (1) perform high-dimensional inference on large 2D and 3D inverse problems, and (2) identify the pitfalls of patch-based training for normalizing flows. Through a stylized problem, we show that patch-based training fails to capture full posterior statistics, as evidenced by the convergence of the posterior covariance matrix to the analytical posterior covariance. Then we use realistic datasets for computed tomography and photoacoustic imaging to demonstrate the scalability of our framework to practical applications in large 2D and 3D inverse problems.

7.1 Introduction

Data-driven machine learning models are increasingly utilized in medical imaging, where they leverage prior information learned from training samples to improve inference speed, supported by hardware accelerators [43]. However, the limited VRAM on GPUs poses a significant challenge, as it restricts the ability to train conventional architectures on datasets where the examples have a high dimensionality. This limitation arises from the need to store intermediate network activations during backpropagation [190]. To circumvent this issue, practitioners often resort to cropping smaller patches from larger inputs for training, which are more manageable in terms of memory requirements [191, 34, 192, 193]. During testing, the full-sized images can be processed by the backbone network, typically a convolutional neural network (CNN), which is agnostic to the size of the input.

The challenge of memory management is particularly relevant with normalizing flows, which require invertible layers, thereby necessitating a latent space equal in size to the target variable. This dimensionality preserving constraint precludes the use of architectures that reduce dimensionality, such as contracting or encoder-decoder models, to save memory [194]. Despite ongoing debates in the literature regarding the memory efficiency of normalizing flows [34, 35], they can be implemented in a way that is remarkably memory efficient. Notably, with proper implementation, normalizing flows can achieve constant memory usage as the number of layers increases [33]. A practical example of such implementation is found in InvertibleNetworks.jl, which allows for the use of full-sized images and volumes as inputs to normalizing flows. This project aims to explore the benefits of training with full-sized inputs and evaluate the potential drawbacks of training with cropped patches.

The aim of this project is to study the robustness of simulation-based inference SBI techniques under practical training regimes. We seek to answer several research questions:

1. What goes wrong in CNN-based variational inference when the learned generative models are not given the full image during training?

2. Is there an optimal balance between the accuracy provided by full-image input and the efficiency of patch-based training?

3. What prescriptions can we suggest when using normalizing flows for variational inference on large scale unknowns?

7.2 Methods

7.2.1 Memory efficient normalizing flows

In this manuscript, we focus on conditional normalizing flows [cite] trained with the forward kullbeck liebler divergence which minimizes the distributional distance between the estimated conditional distribution $p_\theta(\mathbf{x} \mid \mathbf{y})$ and the ground truth posterior distribution. This training scheme requires training pairs from the joint distribution collected in $\mathcal{D} = \{(\mathbf{x}^{(n)}, \, \mathbf{y}^{(n)})\}_{n=0}^{N}$. These are generated by sampling from the prior, $\{\mathbf{x}^{(n)}\}_{n=0}^{N} \sim p(\mathbf{x})$, followed by a forward simulation. Due to their invertible architectures, normalizing flows can be trained with a relatively straightforward maximum likelihood objective:

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \, \frac{1}{N} \sum_{n=0}^{N} \left(\frac{1}{2} \| f_{\boldsymbol{\theta}}(\mathbf{x}^{(n)}; \mathbf{y}^{(n)}) \|_2^2 - \log \left| \det \mathbf{J}_{f_{\boldsymbol{\theta}}} \right| \right).$$

Crucially, the training objective is given by the change of variables formula with relies on the invertibility of the neural network f_θ. Since invertibility requires that the network be one-to-one the dimensionality of its input and output must be equal. This necessitates that the latent variable be the same size as the input which in the case of images can be very large. Unlike GANS [cite], normalizing flows can not rely on contractive architectures for low memory usage. Indeed previous work in the literature [cite cite] have noted that normalizing flows can run into out-of-memory issues. Thus but this is actually not the case since invertibility can be exploited to create extremely frugal memory usage.

Normalizing flows are invertible thus the intermediate activations need not be stored on memory. Instead these activations can be recomputed from the output of the layers since they are invertible. By writing key gradient layers be hand, we make proper use of invertibility and have low memory usage avoiding out-of-memory problems [33]. As we will demonstrate, these networks can now scale to large input sizes and also large number of network layers.

7.2.2 Convolutional layers in Normalizing flows

The main neural network backbone in normalizing flows is the so called ”residual block” in the coupling layer [48]. This network need not be invertible and while it can theoretically be any layer, these typically are constructed to promote some inductive bias that we know is in our target distribution. Some examples include, the use of dense networks for vectorized data [cite], or recurrent neural networks for time series data [cite]. In this work, we focus on image data thus use convolutional neural networks as the backbone.

The use of convolutional neural networks allows us to scalably apply normalizing flows for large scale imaging problems. Furthermore, convolutional layers allow us to verify our hypothesis that patch-based training can be detrimental to inference by training on small patches and then testing the network on the full image.

7.3 Results

We present three sets of results. The first is on a stylized problem where the simplified assumptions allow us to. Second, we use a newly released x-ray computed tomography dataset with large 2D images to demonstrate the effects of patch training on a practical imaging problem. Lastly, we tackle a 3D inverse problem in photoacoustic imaging.

7.3.1 Stylized problem

To effectively identify the errors introduced by patch training, we employ a stylized problem. Consider the scenario where the unknown parameter $\mathbf{x} \in \mathbb{R}^m$ follows a Gaussian distribution with a known mean, $\mu_{\mathbf{x}}$, and covariance, $\Sigma_{\mathbf{x}}$. The forward model involves applying a known matrix $\mathbf{A} \in \mathbb{R}^{m \times m}$, along with additive Gaussian noise $\varepsilon \in \mathbb{R}^m$ characterized by known mean, μ_{ε}, and covariance, Σ_{ε}.

The inverse problem is to recover $\mathbf{x}$ given the observation:

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \varepsilon$$

Drawing from methodologies outlined in [13], we understand that the posterior distribution $p(\mathbf{x}|\mathbf{y})$ is also Gaussian. Utilizing the parameters $\mathbf{A}, \mu_{\varepsilon}, \mu_{\mathbf{x}}, \Sigma_{\mathbf{x}}, \Sigma_{\varepsilon}$, we can compute the ground truth posterior's covariance Σ_{post} and mean μ_{post}, providing a clear basis for our analysis.

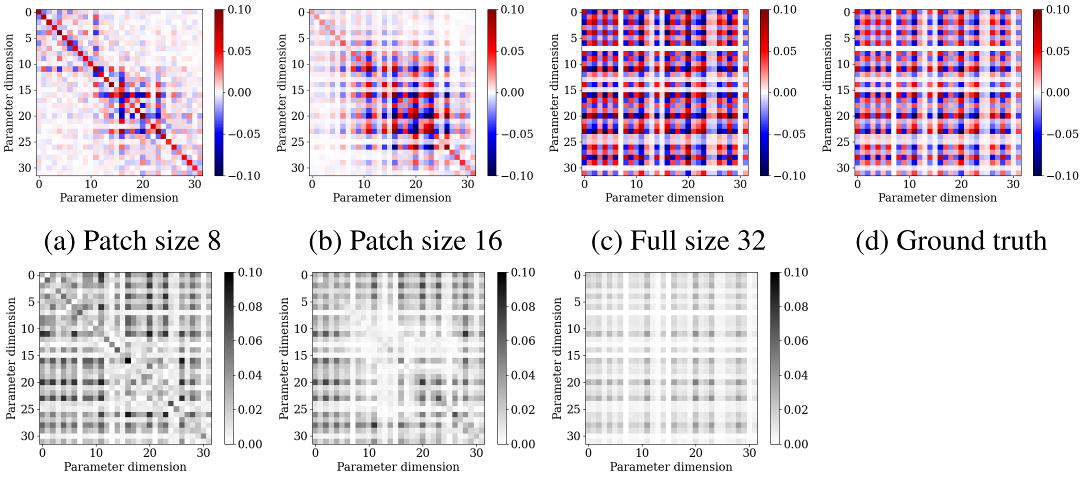
Having access to the ground truth covariance is particularly beneficial for our investigation, as it enables direct measurement of correlations between different dimensions of the unknown parameter. This access is crucial for uncovering the limitations of tested methods in capturing long-range correlations. In line with this, posterior sampling methods are favored over those that merely output the standard deviation (the trace of the covariance) [90]. The latter are unable to capture long-range correlations while the posterior covariance can always be empirically calculated from posterior samples.

For our experiment, we work with an unknown parameter size of $m = 32$ and will test three different training scenarios:

1. Input whole vectors $\mathbf{x}, \mathbf{y}$
2. Input half vector $\mathbf{x}[i : i + 16], \mathbf{y}[i : i + 16]$
3. Input a fourth vector $\mathbf{x}[i : i + 8], \mathbf{y}[i : i + 8]$

Each of these three scenarios trains an amortized posterior sampler $p_{\theta_8}(\mathbf{x} \mid \mathbf{y}), p_{\theta_{16}}(\mathbf{x} \mid \mathbf{y}), p_{\theta_{full}}(\mathbf{x} \mid \mathbf{y})$ which we denote respective to the input size they were trained on. Please refer to supplementary information for training details.

Following the training phase, we input the same full observation $\mathbf{y}^{obs} \in \mathbb{R}^{32}$ to all of the three trained models. This is possible because the networks are primarily based on convolutional layers. This setup allows us to generate posterior samples from each training configuration for the same observation. We then calculate the empirical mean and covariance from these samples to compare with the ground truth covariance Σ_{post}. In Figure 7.1 we show the empirical posterior covariance matrices compared to the known ground truth posterior covariance matrix. We observe that in the patch-based covariances there is an imprint of low error that is the size of the patch used, while the full input posterior covariance has spatially homogeneously distributed error with no concentrated areas of error. We interpret these results to mean that the full input posterior is able to capture the full statistics including the long range sensitivities while the patch-based posteriors were unable to capture these important sensitivities.



(a) Patch size 8 (b) Patch size 16 (c) Full size 32 (d) Ground truth

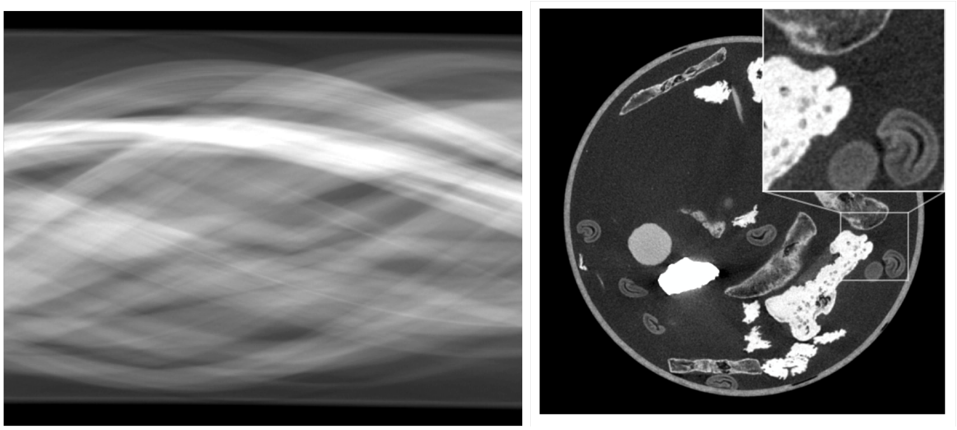
(e) Error patch size 8 (f) Error patch size 16 (g) Error full size 32

Figure 7.1: Effect of patch-based training on linear inverse problem. Posterior covariance from: (a) CNF trained on patch size 8, (b) CNF trained on patch size 16, (c) CNF trained on full vector size 32. (d) Analytically calculated ground truth covariance matrix.

7.3.2 Large 2D computed tomography

We focus on limited-view CT image reconstruction. Due to the long streak artifacts introduced by the limited-view adjoint operator it is expected that patch-based training will be detrimental to the quality of the approximated posterior distribution.

We simulate limited-view data from the 2DeteCT dataset by taking the full sinograms from mode 2 and subsample the first 120 angles. To create a training dataset, we pair the limited-view sinograms to a high fidelity reconstruction provided in mode 2 reconstruction of the 2DeteCT dataset. The full size of the reference reconstructions are 1024×1024. An example of a training pair is shown Figure 7.2.



(a) Limited-view sinogram (120 angle) (b) Reference ground truth

Figure 7.2: Example of limited-view CT reconstruction training pairs.

We train a conditional normalizing flow to solve the limited view reconstruction problem by training on the pairs described in the previous section. Crucial to acceptable performance, we use a physical summary statistic as described in [39] to summarize the observed sinogram with a single filtered backprojection. To see what this physical summary statistic looks like please refer to the top left corner of Figure 7.3.

In order to demonstrate the effect of patch-based training on the limited-view reconstruction problem, we train three normalizing flow models: one on patch sizes of 64×64, one on patch sizes of 256×256 and one on the full images of size 1024×1024. The process of creating patches is done by a random crop of the images during training.To control for the computational load of training on different input sizes we train the models for the same amount of time instead of the same amount of epochs since an epoch with a small patch size takes longer than with a larger epoch.

7.3.3 Computed tomography posterior sampling

After training, we evaluate the three models on an unseen limited-view sinogram and report the posterior sampling results including individual posterior samples, posterior mean, posterior standard deviation and the error created by using the posterior mean as a point estimate. In Figure 7.3, we show the posterior sampling results from the model trained on the full input size of 1024×1024. We note that there is good agreement between the posterior samples. In particular, the outer circumference of the tube has been properly reconstructed. This structure although not present in the observation is consistently present in the training dataset thus information in the prior that has been properly incorporated in the posterior samples.

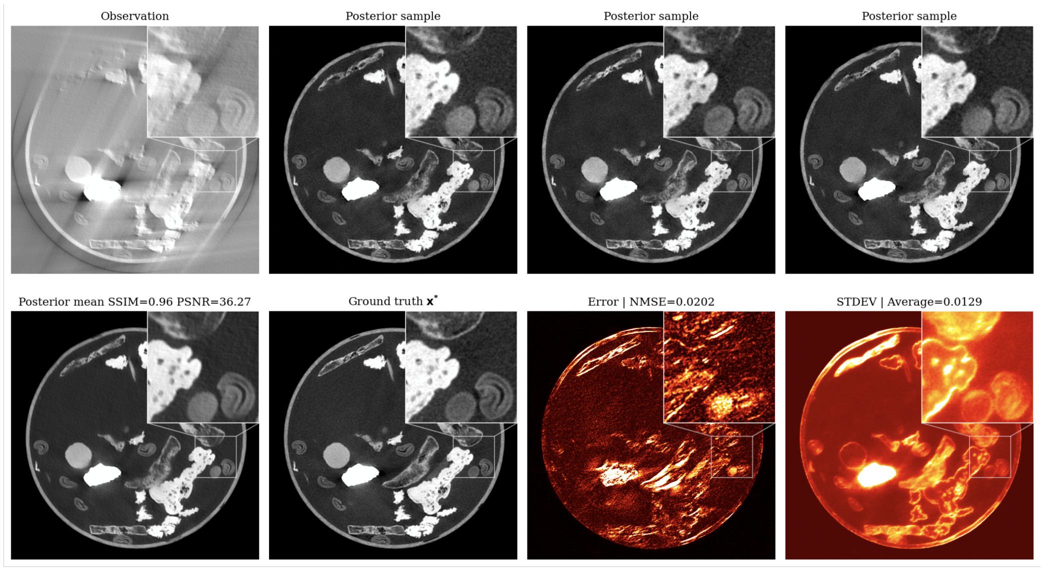


Figure 7.3: Results of posterior sampling for limited-view reconstruction with full input training. The error and the standard deviation are plotted using the same colorbar.

The posterior sampling results for the patch-based training with size 256×256 are shown in Figure 7.4. By drawing attention to the discontinuities in the posterior samples along the circumference (corresponds to the wall of the tube holding the materials), we can intuitively observe the difficulties that this model has in reconstruction structures with long range correlations. These discontinuities translate to larger error in these areas. Furthermore there is overall more uncertainty and more error in the structures inside the tube. We summarize the error with the Normalized Mean Squared Error (NMSE). and summarize the uncertainty with the average value of the standard deviation over all pixels.

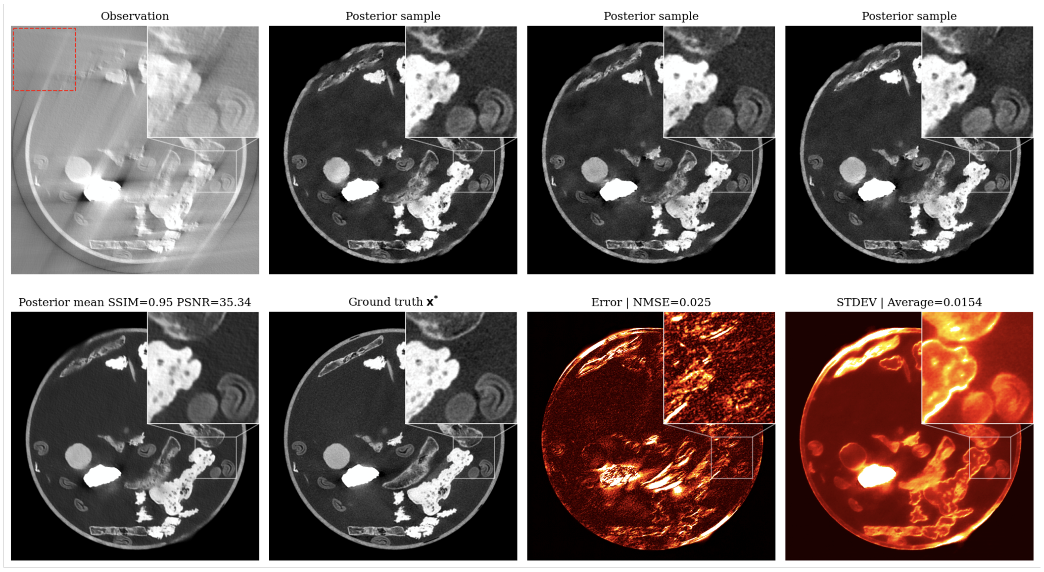


Figure 7.4: Results of posterior sampling for limited-view reconstruction with 256×256 patch training. The dotted red box on the upper left corner visualizes the relative size of the patch to the full image.

Over a leave-out test set of 50 samples, we track the minimum NMSE during training to understand the convergence of the models. We plot these values relative to compute time and observe in Figure 7.5 that the patch-based models converge to a value that is substantially worse than the full input model. The model trained on 64×64 is performing so poorly that its line is not visible on the limits of this plot. For brevity, we do not show its sampling results from the 64×64 model.

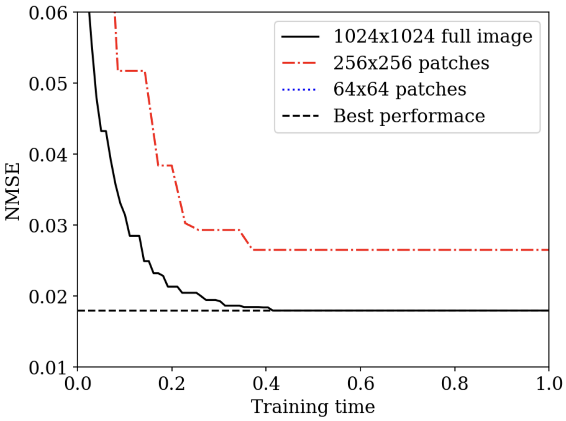


Figure 7.5: Comparing convergence of NMSE during training. The patch-based models converge to worse performance than the full input model.

To understand the quality of the probabilistic results, we also report the following quality metrics for uncertainty.

1. **Average standard deviation (ASD)**: Due to the fact that we do not desire a method that has arbitrarily high uncertainty, we report the average standard deviation of the posterior samples over all pixels:

   $$ASD = \frac{1}{N_{pixel}} \sum_{i=1}^{N_{pixel}} \text{SD}_i$$

   where i is an index that goes over all the pixels in the images and SD is the image of standard deviation calculated from posterior samples.

2. **Negative log-likelihood (NLL)**. This metric is based on a Gaussian assumption on the error made by the posterior mean $\hat{\mathbf{x}}$ [195] and calculated by:

   $$NLL = \frac{1}{2} \sum_{i=1}^{N_{pixel}} \frac{1}{2\text{SD}_i^2}(\mathbf{x}_i^* - \hat{\mathbf{x}}_i) + \frac{1}{2} \log(2\pi \text{SD}_i^2).$$

   This quality metric says in the first term that we can have high error in some pixels as long as it is supported by high uncertainty in those pixels while the second term penalizes having an arbitrarily high standard deviation everywhere.

3. **Uncertainty Calibration Error (UCE)**, an important feature of uncertainty is its capability of predicting error. One way to measure this feature is by using the calibration metric as defined in [79]. For practical implementation details of this metric please refer to [78].

Table 7.1: Quality metrics of patch-based training compared to full image input for limited-view computed tomography.

Training Method	NMSE ↓	ASD ↓	NLL ↓	UCE ↓
Patch 128×128	0.0858	0.0196	-2.08	0.245
Patch 256×256	0.0256	0.0164	-2.90	0.128
Full input 1024×1024	$\mathbf{0.0170}$	$\mathbf{0.0127}$	$\mathbf{-2.99}$	$\mathbf{0.126}$

7.3.4 3D photoacoustic imaging

In photoacoustic imaging, the inverse problem involves recovering the initial photoacoustic source $\mathbf{x}$, where observations are modeled using a linear equation:

$$\mathbf{y} = \mathcal{F}(\mathbf{x}) = \mathbf{A}\mathbf{x} + \boldsymbol{\varepsilon}$$

here $\mathbf{A}$ represents the solution of the wave equation at restricted to receiver locations, given the initial wavefield condition $\mathbf{x}$, and $\boldsymbol{\varepsilon}$ denotes measurement noise.

Due to noisy observations and limited view receivers, the photoacoustic problem is best treated from a probablistic perspective. Although methods for full 3D volume image reconstruction exist [196, 197], these create single point estimates. We were unable to find any previous literature on full 3D volume posterior sampling of the photoacoustic problem.

For this, we will generate training volumes from lung photoacoustic images, following methodologies similar to those described by [196], and originally sourced from public datasets [198]. To show the ability of our trained models to generalize to unseen patients, we train on a 3D volume from one patient and test on another unseen patient.

To assess the effectiveness of patch-based training approaches in handling large 3D volumes, we propose the following training scenarios:

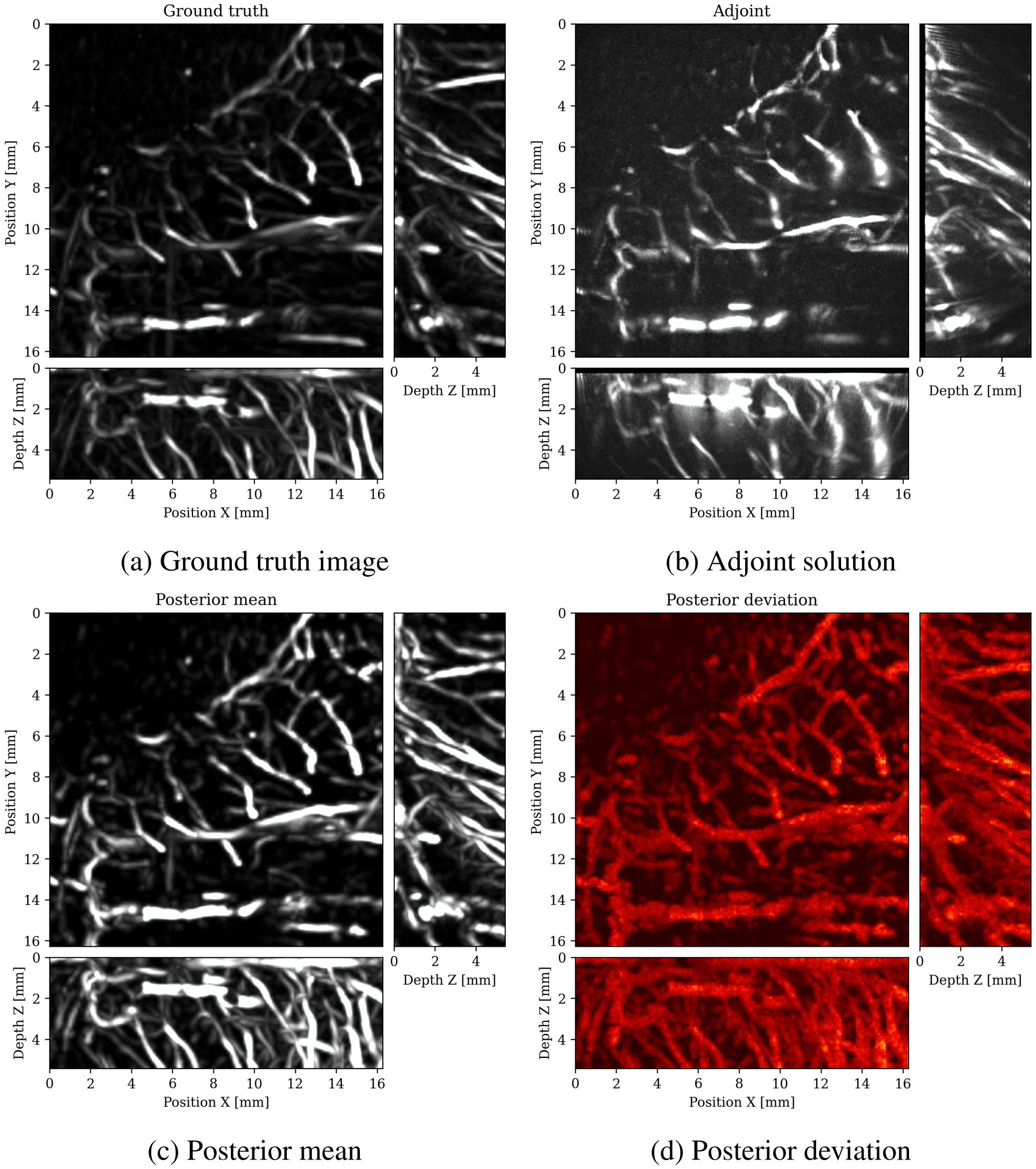
1. Whole Volume Input: Using the entire volume of size $80 \times 240 \times 240$.

2. Large Volume Patches: Employing patches of size $80 \times 120 \times 120$ to manage computational load while retaining significant contextual information.

3. Small Volume Patches: Utilizing smaller patches of size $64 \times 64 \times 64$, focusing on specific regions.

To effectively visualize the 3D volume, we employ the Maximum Intensity Projection (MIP) technique, which projects the maximum valued pixel in each dimension onto three panels. This visualization technique is particularly useful for highlighting the key features and differences across various dimensions of the photoacoustic volume.

We first report the sampling results of the model that was trained using the full volumes for an unseen test observation in Figure 7.6. We observe high qualitative agreement between the posterior mean and the ground truth. In particular we would like to highlight the reconstruction of vertical vessels, these are particularly difficult to image since they are close to being in the null space of the forward operator.



(a) Ground truth image

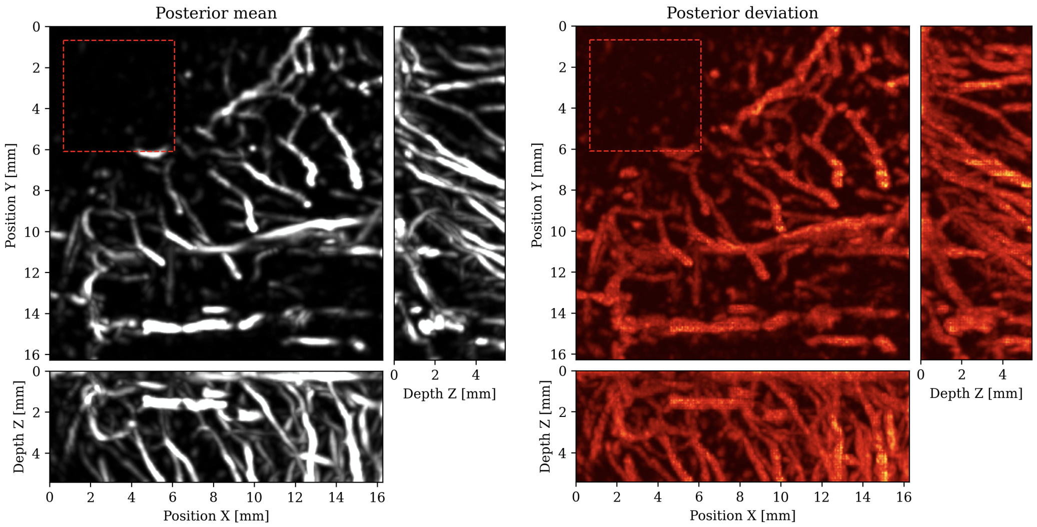
(b) Adjoint solution

(c) Posterior mean

(d) Posterior deviation

Figure 7.6: Volume inverse problem in photoacoustic imaging. (a) Ground truth reference image derived from lung CT scans. (b) The result of a single adjoint application on the observed data $\mathbf{A}^{\top}\mathbf{y}^{obs}$, showcasing evident limited-view artifacts, particularly in vertical vessels. (c) Mean of our amortized posterior sampling, offering a point estimate of the unknown photoacoustic source. (d) Standard deviation of our amortized posterior sampling, highlighting areas of uncertainty.

The posterior sampling results of the model trained on patches of sizes $80 \times 80 \times 80$ is shown in Figure 7.7. The inferior quality of the posterior mean in Figure 7.7a is evident particularly in the vertical vessels which are not well supported by the observations and that are also long therefore its long range correlations are not well captured by this model that was trained on patches. Furthermore we noticed artifacts in the standard deviation Figure 7.7b that became worse as the patch-size became smaller.



(a) Posterior mean

(b) Posterior deviation

Figure 7.7: Probabilistic solutions derived from patch-based training in photoacoustic imaging. (a) Mean of our amortized posterior sampling. (b) Standard deviation of our amortized posterior sampling, showing artifacts due to the patch-based training.

To quantify the quality of the posterior means as an image reconstruction, we take the average RMSE in a test set of 20 observations shown in Table 7.2. We emphasize that these metrics is relatively efficient to calculate over many test examples since the model is amortized and therefore does not require expensive training for each new observation.

Table 7.2: Quality metrics of patch-based training compared to full volume input.

Training Method	RMSE ↓	SD ↓	NLL ↓
Patch $40 \times 40 \times 80$	0.00226	0.00164	-2.08
Patch $80 \times 80 \times 80$	0.00202	0.00146	-1.52
Full input $240 \times 240 \times 80$	$\mathbf{0.00172}$	$\mathbf{0.00136}$	$\mathbf{-2.75}$

7.4 Conclusions

We demonstrated the application of conditional normalizing flows to large 3D inverse problems without using patches or dimensionality reducing methods. The results are driven by a normalizing flow implementation which exploits invertibility for memory frugality. To motivate the usage of these proper implementations instead of depending on patch-based training, we show the detrimental effect of patch-based training in a stylized problem and real world applications in medical imaging. Our experiments showed that patch-based training is detrimental to the quality of the probabilistic solutions in the regimes of image reconstructions and also the quality of their uncertainty quantification. Thus we suggest using full image or volume training to achieve the highest quality in these two regimes and provide the software necessary to do so.

CHAPTER 8
CONCLUSION

I trust that the chapters in this thesis have demonstrated that generative models can be effectively utilized to sample from the Bayesian posterior of high-dimensional imaging problems across various applications in seismic and medical imaging. These results provide strong evidence in support of my original thesis statement:

> *Generative models are a scalable tool to perform uncertainty quantification of high-dimensional seismic and medical imaging.*

To make these contributions explicit, I will now outline the key findings and advancements demonstrated in this work:

8.1 Invertible networks for memory efficiency

Although the methods presented in this thesis are generative model agnostic, as evidenced by the implementation of the ASPIRE algorithm with normalizing flows in Chapter 4 and then diffusion models in Chapter 5, it is important to highlight scenarios where normalizing flows offer distinct efficiency advantages. In particular, users working in environments with limited computational resources or high memory demands may benefit from the unique properties of normalizing flows, which can significantly reduce the memory footprint during training and inference.

Normalizing flows are inherently invertible. This property eliminates the need to store intermediate activations during the forward pass, allowing memory-efficient backpropagation through the use of reversible computation. By leveraging this characteristic, `InvertibleNetworks.jl` enables the training of normalizing flows on image sizes that were previously considered infeasible due to memory constraints. For example, while prior work in the field often limited normalizing flows to image dimensions of (256×256), this implementation extends their applicability to much larger images, such as (1024×1024) or beyond.

8.2 Physics-based summary statistics

Due to the complexity of the observations, traditional Simulation-Based Inference (SBI) methods fail when applied to highly complex inverse problems, such as wave-based imaging subsection 4.4.3. These problems are characterized by nonlinear, high-dimensional forward models, and noisy, incomplete data, which exacerbate the ill-posed nature of the inverse problem. To address these challenges, we proposed extending the concept of summary statistics by introducing a physics-based summary statistic in the form of the score—the gradient of the log-likelihood with respect to the parameters. This approach leverages domain-specific physics to reduce the complexity of the inference problem, making it more tractable.

I have made contributions to the theoretical foundations of this approach, specifically by proving that the use of the score as a summary statistic leaves linear-Gaussian inverse problems completely unbiased. Beyond these theoretical contributions, the proposed methods have been extensively validated through a series of experiments presented across various chapters. Collectively, these results highlight the practical utility of the methods, showcasing their potential to advance the state-of-the-art in Bayesian inference for complex, high-dimensional problems. Additionally, the WISE-ASPIRE-WISER paradigm offers a scalable suite of algorithms tailored to different trade-offs between computational cost and posterior quality, providing flexibility for a variety of use cases.

8.3 Frugal use of expensive physics

Another important point I want to emphasize is the necessity of making frugal use of the physical operator. It is crucial to clarify that the goal of this thesis is not to eliminate the use of physics in solving inverse problems. On the contrary, my central message is that physics plays an indispensable role in achieving reliable solutions. The principled integration of physics into the algorithmic framework is what drives the performance gains observed in methods like ASPIRE, which builds upon and extends the foundation laid by algorithms such as WISE.

ASPIRE exemplifies how leveraging additional physics can enhance algorithmic performance without compromising scalability. By incorporating a an extra gradient evaluation per iteration, ASPIRE enriches the intermediate summary statistic, improving the quality of posterior sampling. This selective and efficient use of the forward and adjoint operators is what I define as ”frugal.” It ensures that the algorithm remains computationally feasible while still taking full advantage of the underlying physics.

In summary, the message of this thesis is not to reduce reliance on physical models but to use them efficiently. By striking the right balance between leveraging physics and maintaining scalability, methods like ASPIRE demonstrate that frugal yet principled integration of physics can significantly uplift performance in solving high-dimensional inverse problems.

8.4 Limitations and future work

I will end by going over some of the limitations of the proposed methods and how I would propose to overcome these limitations.

8.4.1 True likelihood-free inference

A notable limitation of the methods currently presented is their reliance on calculating the gradient of the data likelihood. At first glance, this dependency appears to make these methods unsuitable for truly likelihood-free scenarios. However, there are two promising avenues for future research that could potentially overcome this limitation.

The first avenue involves recognizing that the assumed likelihood used in calculating the score need not match the true likelihood $p(\mathbf{y} \mid \mathbf{x})$. As discussed by [124], this summary statistic remains effective in likelihood-free contexts as long as the assumed likelihood is reasonably close to the true one. Furthermore, learned likelihoods [199] derived from samples can be employed. In this case, a surrogate model could first be trained to approximate the likelihood, after which automatic differentiation could be used to obtain the gradient.

The second avenue proposes completely offloading the gradient computation to the learned generative model. Specifically, instead of providing the precalculated gradient, the network could take as input the observed data $\mathbf{y}^{obs}$ and the predicted data $\mathcal{F}(\mathbf{x}_0)$. This would allow the network to either approximate the gradient or identify alternative, potentially more informative summary statistics. In the context of the iterative ASPIRE framework, this could be naturally extended by feeding the residual at the i-th iteration, $\mathcal{F}(\mathbf{x}_i) - \mathbf{y}^{obs}$, directly into the network. As the intermediate estimates $\mathbf{x}_i$ converge toward the maximum likelihood solution, the residual decreases, making the updates easier to learn.

8.4.2 Construction of training datasets

The most salient limitation of the presented methods in this thesis is the requirement of training samples from the prior distribution $p(\mathbf{x})$. While this is a known limitation for SBI methods and various applications can safely assume that they have access to these (medical imaging modalities can make high quality reference solutions such as with high-dose multi angle MRI) there are some applications such as in seismic imaging where the access to these samples is not easily given. To I proposed and started the development of the Subsurface foundational model with AI-driven Geostatistical Extraction (SAGE) algorithm [173], while the final proof-of-concpet of SAGE on field data has not been completed, I am still actively contributing to this project and hope to do so in the future.

A key limitation of the methods presented in this thesis is their reliance on training samples drawn from the prior distribution $p(\mathbf{x})$. This dependency is a well-known challenge in Simulation-Based Inference (SBI) methods, and while certain applications can safely assume access to high-quality prior samples, others face significant hurdles in this regard. For example, in medical imaging, modalities such as high-dose, multi-angle MRI can provide reliable reference solutions that serve as effective prior samples. However, in seismic imaging, obtaining prior samples is far more challenging. Unlike in medical imaging, where controlled conditions and standardized protocols facilitate data acquisition, seismic imaging relies on field data that is often sparse, noisy, and subject to complex subsurface conditions. This lack of readily available high-quality samples limits the applicability of standard SBI techniques in seismic workflows.

To address this issue, I proposed and began developing the SAGE (Subsurface foundational model with AI-driven Geostatistical Extraction) algorithm [173]. SAGE aims to overcome the data scarcity problem by training a generative model using only available seismic data, in the form of borehole well and migrated images. While the final proof-of-concept for SAGE on field data remains a work in progress, preliminary results show promise. The algorithm has the potential to generate synthetic prior samples that closely mimic real subsurface conditions, paving the way for its use in training generative networks.

Although the full validation of SAGE on field data has not yet been completed, I continue to actively contribute to this project. I am optimistic about its future development and moving forward, I aim to be involved in its final stages.

8.4.3 Multi-modality and multiple fiducial points

An open question in the ASPIRE algorithm is how to leverage the probabilistic nature of its intermediate outputs. At each iteration, ASPIRE produces posterior samples, which offer a rich representation of the solution space. The practical approach we adopted was to compress these samples into a single fiducial point, specifically the posterior mean. As shown in section .1, the posterior mean possesses desirable properties. However, there are scenarios where compressing the posterior in this way might not be ideal, particularly if valuable uncertainty information is lost in the process.

One alternative approach is to avoid compressing the posterior samples entirely. Inspired by the method proposed in [200], we could use the posterior samples to define multiple fiducial points. This would allow us to treat these fiducial points as nuisance variables and marginalize out their influence. Practically, this is implemented by augmenting the training dataset, pairing each ground truth sample with multiple gradients corresponding to different fiducial points. This strategy effectively increases the diversity of the training data and allows the model to learn a more robust representation of the gradient operator across various fiducial points.

Regardless of the specific path chosen, scalability remains a key concern, as both approaches introduce additional gradient evaluations. To address this, I propose using the Fourier Neural Operator (FNO) [201] as a surrogate for the gradient operation. FNOs are designed to efficiently learn mappings between function spaces and have demonstrated success in approximating complex operators in high-dimensional settings. By replacing the explicit gradient computation with a learned surrogate, we can significantly reduce the computational cost, making these probabilistic extensions of ASPIRE feasible even for large-scale applications.

Appendices

.1 Proof of Theorem 2

Theorem 2: For a linear inverse problem with forward operator $\mathbf{A} \in \mathbb{R}^{m \times n}$, an unknown $\mathbf{x}$ with Gaussian prior $\mathcal{N}(\mu_{\mathbf{x}}, \mathbf{C}_{\mathbf{x}})$, and additive Gaussian noise $\mathcal{N}(0, \mathbf{C}_{\boldsymbol{\varepsilon}})$, if the ℓ_2 norm between the current fiducial $\mathbf{x}_0$ and the maximum likelihood estimate $\|\mathbf{x}_0 - \mathbf{x}_{\mathrm{ML}}\|_2$ is at least

$$K \cdot \|\mu_{\mathbf{x}} - \mathbf{x}_{\mathrm{ML}}\|_2 \leq \|\mathbf{x}_0 - \mathbf{x}_{\mathrm{ML}}\|_2,$$

where

$$K = \left\| \left(\mathbf{C}_{\mathbf{x}}^{-1} + \mathbf{A}^{\top} \mathbf{C}_{\boldsymbol{\varepsilon}}^{-1} \mathbf{A} \right)^{-1} \right\|_2 \cdot \left\| \mathbf{C}_{\mathbf{x}}^{-1} \right\|_2,$$

then forming the posterior with the gradient-based summary $p(\mathbf{x} \mid \overline{\mathbf{y}}_0)$ and using the posterior mean $\mathbf{x}_1 = \mathbb{E}_{p(\mathbf{x}|\overline{\mathbf{y}}_0)}[\mathbf{x}]$ as the next fiducial will yield an estimate with a smaller ℓ_2 norm distance to the maximum likelihood estimate

$$\|\mathbf{x}_1 - \mathbf{x}_{\mathrm{ML}}\|_2 \leq \|\mathbf{x}_0 - \mathbf{x}_{\mathrm{ML}}\|_2.$$

Proof. We approach this proof in two parts. First, we derive a closed-form solution for the gradient summarized posterior $p(\mathbf{x} \mid \overline{\mathbf{y}}_0)$. Then, we analyze the relationship between the posterior mean of that distribution and the maximum likelihood estimate.

We are interested in the conditional distribution $p(\mathbf{x} \mid \overline{\mathbf{y}}_0)$ where the gradient $\overline{\mathbf{y}}_0$ is calculated as in @eq-score leading to the expression $\overline{\mathbf{y}}_0 = \mathbf{A}^{\top} \mathbf{C}_{\boldsymbol{\varepsilon}}^{-1}(\mathbf{A}\mathbf{x}_0 - \mathbf{y})$. Substituting the observation model $\mathbf{y} = \mathbf{A}\mathbf{x} + \boldsymbol{\varepsilon}$ into the expression for $\overline{\mathbf{y}}_0$, we get

$$\overline{\mathbf{y}}_0 = \mathbf{A}^{\top} \mathbf{C}_{\boldsymbol{\varepsilon}}^{-1}(\mathbf{A}\mathbf{x}_0 - \mathbf{y}) = \mathbf{A}^{\top} \mathbf{C}_{\boldsymbol{\varepsilon}}^{-1}(\mathbf{A}\mathbf{x}_0 - \mathbf{A}\mathbf{x} - \boldsymbol{\varepsilon}) = \mathbf{A}^{\top} \mathbf{C}_{\boldsymbol{\varepsilon}}^{-1} \mathbf{A}(\mathbf{x}_0 - \mathbf{x}) - \mathbf{A}^{\top} \mathbf{C}_{\boldsymbol{\varepsilon}}^{-1} \boldsymbol{\varepsilon}.$$

Given $\mathbf{x}$, the variable $\overline{\mathbf{y}}_0$ is Gaussian because it is a linear transformation of the Gaussian noise $\boldsymbol{\varepsilon}$. We can calculate the covariance of this distribution by noting that

$$\text{Cov}(\mathbf{A}^\top \mathbf{C}_{\boldsymbol{\varepsilon}}^{-1}\boldsymbol{\varepsilon}) = (\mathbf{A}^\top \mathbf{C}_{\boldsymbol{\varepsilon}}^{-1})\text{Cov}(\boldsymbol{\varepsilon})(\mathbf{A}^\top \mathbf{C}_{\boldsymbol{\varepsilon}}^{-1})^\top = (\mathbf{A}^\top \mathbf{C}_{\boldsymbol{\varepsilon}}^{-1})\mathbf{C}_{\boldsymbol{\varepsilon}}\mathbf{C}_{\boldsymbol{\varepsilon}}^{-1}\mathbf{A} = \mathbf{A}^\top \mathbf{C}_{\boldsymbol{\varepsilon}}^{-1}\mathbf{A}.$$

yielding the Gaussian distribution

$$p(\overline{\mathbf{y}}_0 \mid \mathbf{x}) \sim \mathcal{N}(\mathbf{A}^\top \mathbf{C}_{\boldsymbol{\varepsilon}}^{-1}\mathbf{A}(\mathbf{x}_0 - \mathbf{x}), \mathbf{A}^\top \mathbf{C}_{\boldsymbol{\varepsilon}}^{-1}\mathbf{A}).$$

To derive the posterior $p(\mathbf{x} \mid \overline{\mathbf{y}}_0)$, we use Bayes' theorem

$$p(\mathbf{x} \mid \overline{\mathbf{y}}_0) \propto p(\overline{\mathbf{y}}_0 \mid \mathbf{x})p(\mathbf{x}),$$

$$p(\overline{\mathbf{y}}_0 \mid \mathbf{x}) \propto \exp\left(-\frac{1}{2}(\overline{\mathbf{y}}_0 - \mathbf{A}^\top \mathbf{C}_{\boldsymbol{\varepsilon}}^{-1}\mathbf{A}(\mathbf{x}_0 - \mathbf{x}))^\top (\mathbf{A}^\top \mathbf{C}_{\boldsymbol{\varepsilon}}^{-1}\mathbf{A})^{-1}(\overline{\mathbf{y}}_0 - \mathbf{A}^\top \mathbf{C}_{\boldsymbol{\varepsilon}}^{-1}\mathbf{A}(\mathbf{x}_0 - \mathbf{x}))\right),$$

$$p(\mathbf{x}) \propto \exp\left(-\frac{1}{2}(\mathbf{x} - \mu_{\mathbf{x}})^\top \mathbf{C}_{\mathbf{x}}^{-1}(\mathbf{x} - \mu_{\mathbf{x}})\right).$$

Thus,

$$\begin{aligned} p(\mathbf{x} \mid \overline{\mathbf{y}}_0) \propto \exp\Bigg(-\frac{1}{2}\big[&(\overline{\mathbf{y}}_0 - \mathbf{A}^\top \mathbf{C}_{\boldsymbol{\varepsilon}}^{-1}\mathbf{A}(\mathbf{x}_0 - \mathbf{x}))^\top (\mathbf{A}^\top \mathbf{C}_{\boldsymbol{\varepsilon}}^{-1}\mathbf{A})^{-1}(\overline{\mathbf{y}}_0 - \mathbf{A}^\top \mathbf{C}_{\boldsymbol{\varepsilon}}^{-1}\mathbf{A}(\mathbf{x}_0 - \mathbf{x})) \\ &+(\mathbf{x} - \mu_{\mathbf{x}})^\top \mathbf{C}_{\mathbf{x}}^{-1}(\mathbf{x} - \mu_{\mathbf{x}})\big]\Bigg). \end{aligned}$$

Expanding the terms:

$$p(\mathbf{x} \mid \overline{\mathbf{y}}_0) \propto \exp\left(-\frac{1}{2}\left[\overline{\mathbf{y}}_0^\top(\mathbf{A}^\top\mathbf{C}_{\boldsymbol{\varepsilon}}^{-1}\mathbf{A})^{-1}\overline{\mathbf{y}}_0 - 2\overline{\mathbf{y}}_0^\top\mathbf{x}_0 + 2\overline{\mathbf{y}}_0^\top\mathbf{x} + \mathbf{x}_0^\top\mathbf{A}^\top\mathbf{C}_{\boldsymbol{\varepsilon}}^{-1}\mathbf{A}\mathbf{x}_0\right.\right.$$
$$\left.\left. -2\mathbf{x}_0^\top\mathbf{A}^\top\mathbf{C}_{\boldsymbol{\varepsilon}}^{-1}\mathbf{A}\mathbf{x} + \mathbf{x}^\top\mathbf{A}^\top\mathbf{C}_{\boldsymbol{\varepsilon}}^{-1}\mathbf{A}\mathbf{x} + \mathbf{x}^\top\mathbf{C}_{\mathbf{x}}^{-1}\mathbf{x} - 2\mathbf{x}^\top\mathbf{C}_{\mathbf{x}}^{-1}\mu_{\mathbf{x}} + \mu_{\mathbf{x}}^\top\mathbf{C}_{\mathbf{x}}^{-1}\mu_{\mathbf{x}}\right]\right).$$

We ignore terms that are not related to $\mathbf{x}$ since these are absorbed into the normalization constant

$$p(\mathbf{x} \mid \overline{\mathbf{y}}_0) \propto \exp\left(-\frac{1}{2}\left[2\overline{\mathbf{y}}_0^\top\mathbf{x} - 2\mathbf{x}_0^\top\mathbf{A}^\top\mathbf{C}_{\boldsymbol{\varepsilon}}^{-1}\mathbf{A}\mathbf{x} + \mathbf{x}^\top\mathbf{A}^\top\mathbf{C}_{\boldsymbol{\varepsilon}}^{-1}\mathbf{A}\mathbf{x} + \mathbf{x}^\top\mathbf{C}_{\mathbf{x}}^{-1}\mathbf{x} - 2\mathbf{x}^\top\mathbf{C}_{\mathbf{x}}^{-1}\mu_{\mathbf{x}}\right]\right).$$

We focus on grouping the quadratic and linear terms in $\mathbf{x}$:

$$p(\mathbf{x} \mid \overline{\mathbf{y}}_0) \propto \exp\left(-\frac{1}{2}\left[\mathbf{x}^\top(\mathbf{A}^\top\mathbf{C}_{\boldsymbol{\varepsilon}}^{-1}\mathbf{A} + \mathbf{C}_{\mathbf{x}}^{-1})\mathbf{x} - 2\mathbf{x}^\top(\mathbf{A}^\top\mathbf{C}_{\boldsymbol{\varepsilon}}^{-1}\mathbf{A}\mathbf{x}_0 + \mathbf{C}_{\mathbf{x}}^{-1}\mu_{\mathbf{x}} - \overline{\mathbf{y}}_0)\right]\right).$$

This is the kernel of a Gaussian distribution, so $p(\mathbf{x} \mid \overline{\mathbf{y}}_0) = p(\mathbf{x} \mid \mathbf{A}^\top\mathbf{C}_{\boldsymbol{\varepsilon}}^{-1}(\mathbf{A}\mathbf{x}_0 - \mathbf{y}))$ is a Gaussian distribution with covariance

$$\mathbf{C}_{\text{post}} = (\mathbf{A}^\top\mathbf{C}_{\boldsymbol{\varepsilon}}^{-1}\mathbf{A} + \mathbf{C}_{\mathbf{x}}^{-1})^{-1},$$

and mean

$$\mathbf{x}_{\text{PM}} = \mathbf{C}_{\text{post}}(\mathbf{A}^\top\mathbf{C}_{\boldsymbol{\varepsilon}}^{-1}\mathbf{A}\mathbf{x}_0 + \mathbf{C}_{\mathbf{x}}^{-1}\mu_{\mathbf{x}} - \overline{\mathbf{y}}_0)$$

$$\mathbf{x}_{\text{PM}} = \mathbf{C}_{\text{post}}\left(\mathbf{A}^\top\mathbf{C}_{\boldsymbol{\varepsilon}}^{-1}\mathbf{A}\mathbf{x}_0 + \mathbf{C}_{\mathbf{x}}^{-1}\mu_{\mathbf{x}} - \mathbf{A}^\top\mathbf{C}_{\boldsymbol{\varepsilon}}^{-1}(\mathbf{A}\mathbf{x}_0 - \mathbf{y})\right)$$

$$\mathbf{x}_{\text{PM}} = \mathbf{C}_{\text{post}}(\mathbf{A}^\top \mathbf{C}_{\boldsymbol{\varepsilon}}^{-1}\mathbf{y} + \mathbf{C}_{\mathbf{x}}^{-1}\mu_{\mathbf{x}}).$$

Now that we have an analytical expression for the posterior mean of the distribution conditioned on the gradient, we can proceed to find the relationship between the maximum likelihood estimate and the summarized posterior mean. Given the observation $\mathbf{y}$, the maximum likelihood estimate is

$$\mathbf{x}_{\text{ML}} = (\mathbf{A}^\top \mathbf{C}_{\boldsymbol{\varepsilon}}^{-1}\mathbf{A})^{-1}\mathbf{A}^\top \mathbf{C}_{\boldsymbol{\varepsilon}}^{-1}\mathbf{y},$$

Since $(\mathbf{A}^\top \mathbf{C}_{\boldsymbol{\varepsilon}}^{-1}\mathbf{A})\mathbf{x}_{\text{ML}} = \mathbf{A}^\top \mathbf{C}_{\boldsymbol{\varepsilon}}^{-1}\mathbf{y}$, we can substitute $\mathbf{x}_{\text{ML}}$ into the previously derived expression for the summarized posterior mean

$$\begin{aligned}
\mathbf{x}_{\text{PM}} &= \mathbf{C}_{\text{post}}\left(\mathbf{A}^\top \mathbf{C}_{\boldsymbol{\varepsilon}}^{-1}\mathbf{A}\mathbf{x}_{\text{ML}} + \mathbf{C}_{\mathbf{x}}^{-1}\mu_{\mathbf{x}}\right)\\
\mathbf{x}_{\text{PM}} &= \mathbf{C}_{\text{post}}\mathbf{A}^\top \mathbf{C}_{\boldsymbol{\varepsilon}}^{-1}\mathbf{A}\mathbf{x}_{\text{ML}} + \mathbf{C}_{\text{post}}\mathbf{C}_{\mathbf{x}}^{-1}\mu_{\mathbf{x}}.
\end{aligned}$$

We need to simplify the term $\mathbf{C}_{\text{post}}\mathbf{A}^\top \mathbf{C}_{\boldsymbol{\varepsilon}}^{-1}\mathbf{A}$. Recall that $\mathbf{C}_{\text{post}} = (\mathbf{A}^\top \mathbf{C}_{\boldsymbol{\varepsilon}}^{-1}\mathbf{A} + \mathbf{C}_{\mathbf{x}}^{-1})^{-1}$. Therefore

$$\begin{aligned}
\mathbf{C}_{\text{post}}(\mathbf{A}^\top \mathbf{C}_{\boldsymbol{\varepsilon}}^{-1}\mathbf{A} + \mathbf{C}_{\mathbf{x}}^{-1}) &= \mathbf{I},\\
\mathbf{C}_{\text{post}}\mathbf{A}^\top \mathbf{C}_{\boldsymbol{\varepsilon}}^{-1}\mathbf{A} + \mathbf{C}_{\text{post}}\mathbf{C}_{\mathbf{x}}^{-1} &= \mathbf{I},\\
\mathbf{C}_{\text{post}}\mathbf{A}^\top \mathbf{C}_{\boldsymbol{\varepsilon}}^{-1}\mathbf{A} &= \mathbf{I} - \mathbf{C}_{\text{post}}\mathbf{C}_{\mathbf{x}}^{-1}.
\end{aligned}$$

Substituting this result into the expression for $\mathbf{x}_{\text{PM}}$

$$
\begin{aligned}
\mathbf{x}_{\text{PM}} &= \left(\mathbf{I} - \mathbf{C}_{\text{post}}\mathbf{C}_{\mathbf{x}}^{-1}\right)\mathbf{x}_{\text{ML}} + \mathbf{C}_{\text{post}}\mathbf{C}_{\mathbf{x}}^{-1}\mu_{\mathbf{x}},\\
\mathbf{x}_{\text{PM}} &= \mathbf{x}_{\text{ML}} - \mathbf{C}_{\text{post}}\mathbf{C}_{\mathbf{x}}^{-1}\mathbf{x}_{\text{ML}} + \mathbf{C}_{\text{post}}\mathbf{C}_{\mathbf{x}}^{-1}\mu_{\mathbf{x}},\\
\mathbf{x}_{\text{PM}} &= \mathbf{x}_{\text{ML}} + \mathbf{C}_{\text{post}}\mathbf{C}_{\mathbf{x}}^{-1}(\mu_{\mathbf{x}} - \mathbf{x}_{\text{ML}}).
\end{aligned}
$$

Now, let's derive the closed-form solution for the norm distance between the posterior mean and the maximum likelihood estimate. Taking the difference and applying the norm on both sides

$$
\|\mathbf{x}_{\text{PM}} - \mathbf{x}_{\text{ML}}\|_2 = \|\mathbf{C}_{\text{post}}\mathbf{C}_{\mathbf{x}}^{-1}(\mu_{\mathbf{x}} - \mathbf{x}_{\text{ML}})\|_2.
$$

Using the properties of norms, we can factor out the matrices

$$
\begin{aligned}
\|\mathbf{x}_{\text{PM}} - \mathbf{x}_{\text{ML}}\|_2 &\leq \|\mathbf{C}_{\text{post}}\|_2\|\mathbf{C}_{\mathbf{x}}^{-1}\|_2\|\mu_{\mathbf{x}} - \mathbf{x}_{\text{ML}}\|_2,\\
\|\mathbf{x}_{\text{PM}} - \mathbf{x}_{\text{ML}}\|_2 &\leq \|(\mathbf{A}^{\top}\mathbf{C}_{\varepsilon}^{-1}\mathbf{A} + \mathbf{C}_{\mathbf{x}}^{-1})^{-1}\|_2\|\mathbf{C}_{\mathbf{x}}^{-1}\|_2\|\mu_{\mathbf{x}} - \mathbf{x}_{\text{ML}}\|_2,\\
\|\mathbf{x}_{\text{PM}} - \mathbf{x}_{\text{ML}}\|_2 &\leq K \cdot \|\mu_{\mathbf{x}} - \mathbf{x}_{\text{ML}}\|_2.
\end{aligned}
$$

Under the assumption of the theorem

$$
K \cdot \|\mu_{\mathbf{x}} - \mathbf{x}_{\text{ML}}\|_2 \leq \|\mathbf{x}_0 - \mathbf{x}_{\text{ML}}\|_2,
$$

we have

$$
\begin{aligned}
\|\mathbf{x}_{\text{PM}} - \mathbf{x}_{\text{ML}}\|_2 &\leq K \cdot \|\mu_{\mathbf{x}} - \mathbf{x}_{\text{ML}}\|_2 \leq \|\mathbf{x}_0 - \mathbf{x}_{\text{ML}}\|_2,\\
\|\mathbf{x}_{\text{PM}} - \mathbf{x}_{\text{ML}}\|_2 &\leq \|\mathbf{x}_0 - \mathbf{x}_{\text{ML}}\|_2.
\end{aligned}
$$

Thus, the proof is complete. □ □

.2 FASTMRI acoustic dataset creation

Based off of the MRI dataset [103], we manually assigned acoustic values to MRI intensities by following the table of acoustic brain tissue properties in the supplemental section of [85]. Although MRI intensities are not necessarily related to acoustic tissue properties, we found that we could produce reasonably realistic acoustic parameters as compared to the acoustic parameters from the MIDA volume. In Figure 1, we show some example training acoustic parameters. We also plot the average and standard deviation between all 1000 training samples in Figure 2. From these plots, we note that there are few similarities between training examples apart from the biologically consistent human brain structures.

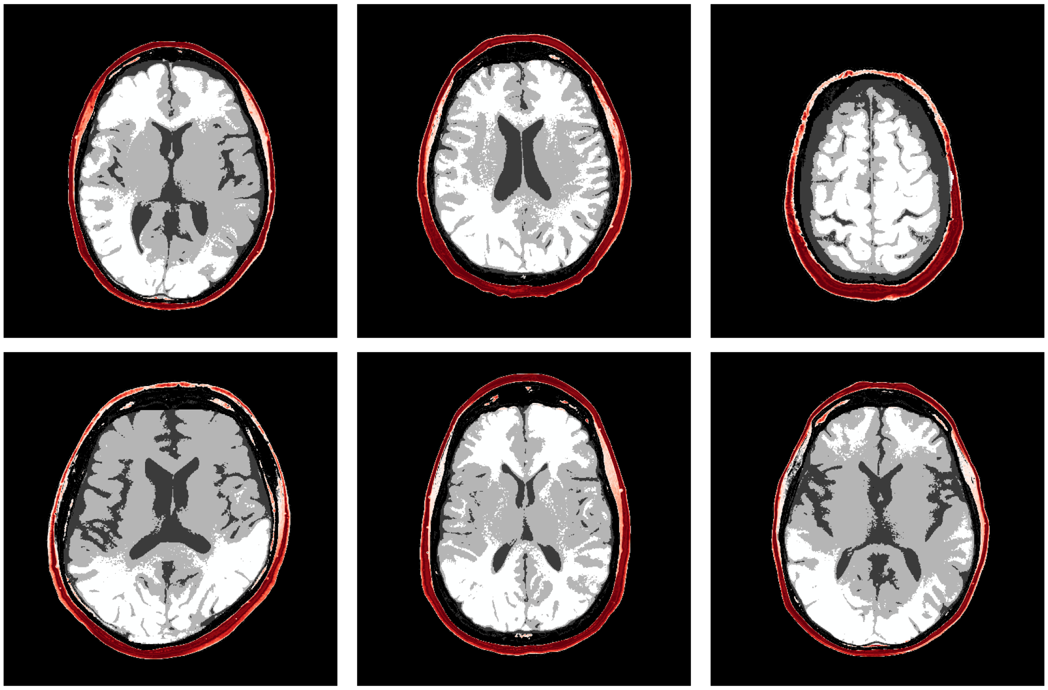
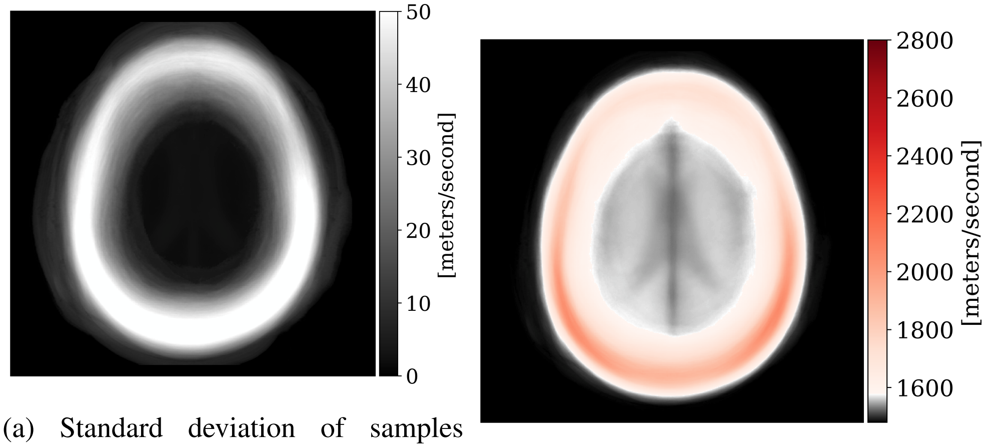


Figure 1: Examples of training examples used to train our method $\mathbf{x}^{(n)} \sim p(\mathbf{x})$.



(a) Standard deviation of samples $\sqrt{\mathbb{V}}\ p(\mathbf{x})$

(b) Mean of samples $\mathbb{E}\ p(\mathbf{x})$

Figure 2: Training dataset used to train models. (a) Standard deviation of samples. (b) Mean of samples.

.3 Wave modeling and FWI implementation

To mask source-receiver artifacts, the gradients used for the traditional FWI optimizations are masked by a binary matrix where the mask was made large enough to include the skull but otherwise assumed no knowledge of the skull. We also avoid the inverse crime by generating any ”observed” acoustic data with a spatial finite-difference kernel of size 16 gridpoints and with computational time discretization of 0.025 microseconds while the physical operator used for gradient calculation corresponds to a simulation with spatial finite-difference kernel of size 8 gridpoints and computational time discretization of 0.5 microseconds.

The clock time of solving the wave equation PDE using our hardware configuration (NVIDIA A100 40GB VRAM) was on average 0.81 seconds. Using this number and assuming that we compute the PDE source terms in parallel then we have calculated the total clock time for the discussed algorithms in Table 1.

To implement a traditional Full-Waveform Inversion (FWI) solution with Total-Variation (TV) regularization, we minimized the data misfit using gradient descent and projected the current solution onto a TV-norm ball of a specified size after each gradient

Table 1: Costs for wave-based inversion measured by clock time. Estimated using the timing for the operator solve and assuming parallel computation over source terms.

Method	Offline cost (sec)	Online cost (sec)
ASPIRE	7290	6.4
Our non-amortized method	2430	654.5
Mean-field / Traditional FWI	None	648.0

step. We determined the size of the TV-norm ball by multiplying the TV-norm of the known ground truth by a factor of 0.75, which was tuned to yield optimal results. To perform the TV-norm ball projection, we used the algorithm and software from [202]. In Figure 4.12a, we present the final result of FWI constrained by TV regularization, using the same number of operator evaluations as in the mean-field solution. We observed that TV regularization leads to a superior final solution compared to the mean-field approach, though it does so at the expense of losing uncertainty quantification. Nonetheless, because of the high noise level, there are still artifacts present in the solution.

REFERENCES

[1] D. Ulyanov, A. Vedaldi, and V. Lempitsky, “Deep image prior,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 9446–9454.

[2] L. Zhang, S. Sengupta, C. Parekh, and R. Eliott-Lockhart, “High-frequency fwi velocity model building for hydrocarbon delineation,” in *SEG International Exposition and Annual Meeting*, SEG, 2022, D011S043R004.

[3] A. P. Gahlot, R. Orozco, Z. Yin, and F. J. Herrmann, “An uncertainty-aware digital shadow for underground multimodal co2 storage monitoring,” *arXiv preprint arXiv:2410.01218*, 2024.

[4] J. N. Louie, S. K. Pullammanappallil, and W. Honjas, “Advanced seismic imaging for geothermal development,” in *SEG International Exposition and Annual Meeting*, SEG, 2012, SEG–2012.

[5] F. M. Wagner and S. Uhlemann, “An overview of multimethod imaging approaches in environmental geophysics,” *Advances in Geophysics*, vol. 62, pp. 1–72, 2021.

[6] S. Bauer *et al.*, “Real-time range imaging in health care: A survey,” in *Time-of-Flight and Depth Imaging. Sensors, Algorithms, and Applications*, Springer, 2013, pp. 228–254.

[7] P. A. Witte *et al.*, “A large-scale framework for symbolic implementations of seismic inversion algorithms in julia,” *Geophysics*, vol. 84, no. 3, F57–F71, 2019.

[8] M. Louboutin *et al.*, “Devito (v3. 1.0): An embedded domain-specific language for finite differences and geophysical exploration,” *Geoscientific Model Development*, vol. 12, no. 3, pp. 1165–1187, 2019.

[9] J. Hadamard, “Sur les problèmes aux dérivées partielles et leur signification physique,” *Princeton university bulletin*, pp. 49–52, 1902.

[10] A. Tarantola, *Inverse problem theory and methods for model parameter estimation*. SIAM, 2005.

[11] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian data analysis*. Chapman and Hall/CRC, 1995.

[12] A. Curtis and A. Lomax, “Prior information, sampling distributions, and the curse of dimensionality,” *Geophysics*, vol. 66, no. 2, pp. 372–378, 2001.

[13] C. M. Bishop and N. M. Nasrabadi, *Pattern recognition and machine learning*. Springer, 2006, vol. 4.

[14] C. P. Robert, G. Casella, and G. Casella, *Monte Carlo statistical methods*. Springer, 1999, vol. 2.

[15] A. Siahkoohi, G. Rizzuti, and F. J. Herrmann, "Deep bayesian inference for seismic imaging with tasks," *Geophysics*, vol. 87, no. 5, S281–S302, 2022.

[16] J. Martin, L. C. Wilcox, C. Burstedde, and O. Ghattas, "A stochastic newton mcmc method for large-scale statistical inverse problems with application to seismic inversion," *SIAM Journal on Scientific Computing*, vol. 34, no. 3, A1460–A1487, 2012.

[17] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the royal statistical society: series B (methodological)*, vol. 39, no. 1, pp. 1–22, 1977.

[18] L. Tierney and J. B. Kadane, "Accurate approximations for posterior moments and marginal densities," *Journal of the american statistical association*, pp. 82–86, 1986.

[19] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul, "An introduction to variational methods for graphical models," *Machine learning*, vol. 37, pp. 183–233, 1999.

[20] S. Kullback and R. A. Leibler, "On information and sufficiency," *The annals of mathematical statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[21] A. Siahkoohi, G. Rizzuti, R. Orozco, and F. J. Herrmann, "Reliable amortized variational inference with physics-based latent distribution correction," *Geophysics*, vol. 88, no. 3, R297–R322, 2023.

[22] A. Siahkoohi, G. Rizzuti, M. Louboutin, P. A. Witte, and F. J. Herrmann, "Preconditioned training of normalizing flows for variational inference in inverse problems," *arXiv preprint arXiv:2101.03709*, 2021.

[23] B. T. Feng, J. Smith, M. Rubinstein, H. Chang, K. L. Bouman, and W. T. Freeman, "Score-based diffusion models as principled priors for inverse imaging," *arXiv preprint arXiv:2304.11751*, 2023.

[24] M. Welling and Y. W. Teh, "Bayesian learning via stochastic gradient langevin dynamics," in *Proceedings of the 28th international conference on machine learning (ICML-11)*, Citeseer, 2011, pp. 681–688.

[25] H. Sun and K. L. Bouman, “Deep probabilistic imaging: Uncertainty quantification and multi-modal solution characterization for computational imaging,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, 2021, pp. 2628–2637.

[26] Q. Liu and D. Wang, “Stein variational gradient descent: A general purpose bayesian inference algorithm,” *Advances in neural information processing systems*, vol. 29, 2016.

[27] X. Zhang, A. Lomas, M. Zhou, Y. Zheng, and A. Curtis, “3-d bayesian variational full waveform inversion,” *Geophysical Journal International*, vol. 234, no. 1, pp. 546–561, 2023.

[28] D. Blatter, M. Morzfeld, K. Key, and S. Constable, “Uncertainty quantification for regularized inversion of electromagnetic geophysical data—part i: Motivation and theory,” *Geophysical Journal International*, vol. 231, no. 2, pp. 1057–1074, 2022.

[29] J. M. Bardsley, A. Solonen, H. Haario, and M. Laine, “Randomize-then-optimize: A method for sampling from posterior distributions in nonlinear inverse problems,” *SIAM Journal on Scientific Computing*, vol. 36, no. 4, A1895–A1910, 2014.

[30] C. Cremer, X. Li, and D. Duvenaud, “Inference suboptimality in variational autoencoders,” in *International Conference on Machine Learning*, PMLR, 2018, pp. 1078–1086.

[31] S. T. Radev, U. K. Mertens, A. Voss, L. Ardizzone, and U. Köthe, “Bayesflow: Learning complex stochastic models with invertible neural networks,” *IEEE transactions on neural networks and learning systems*, 2020.

[32] P. Putzky *et al.*, “Amortized inference in inverse problems,” 2023.

[33] R. Orozco *et al.*, “Invertiblenetworks. jl: A julia package for scalable normalizing flows,” *arXiv preprint arXiv:2312.13480*, 2023.

[34] K. Kothari, A. Khorashadizadeh, M. de Hoop, and I. Dokmanić, “Trumpets: Injective flows for inference and inverse problems,” in *Uncertainty in Artificial Intelligence*, PMLR, 2021, pp. 1269–1278.

[35] A. Dasgupta, D. V. Patel, D. Ray, E. A. Johnson, and A. A. Oberai, “A dimension-reduced variational approach for solving physics-based inverse problems using generative adversarial network priors and normalizing flows,” *Computer Methods in Applied Mechanics and Engineering*, vol. 420, p. 116 682, 2024.

[36] Z. Yin, R. Orozco, M. Louboutin, and F. J. Herrmann, "Wise: Full-waveform variational inference via subsurface extensions," *Geophysics*, vol. 89, no. 4, A23–A28, 2024.

[37] R. Orozco, A. Siahkoohi, M. Louboutin, and F. J. Herrmann, "Aspire: Iterative amortized posterior inference for bayesian inverse problems," *arXiv preprint arXiv:2405.05398*, 2024.

[38] Z. Yin, R. Orozco, and F. J. Herrmann, "Wiser: Multimodal variational inference for full-waveform inversion without dimensionality reduction," *arXiv preprint arXiv:2405.10327*, 2024.

[39] R. Orozco, A. Siahkoohi, G. Rizzuti, T. van Leeuwen, and F. J. Herrmann, "Adjoint operators enable fast and amortized machine learning based bayesian uncertainty quantification," in *Medical Imaging 2023: Image Processing*, SPIE, vol. 12464, 2023, pp. 365–375.

[40] R. Orozco, M. Louboutin, A. Siahkoohi, G. Rizzuti, T. van Leeuwen, and F. J. Herrmann, "Amortized normalizing flows for transcranial ultrasound with uncertainty quantification," in *Medical Imaging with Deep Learning, MIDL 2023, 10-12 July 2023, Nashville, TN, USA*, I. Oguz *et al.*, Eds., ser. Proceedings of Machine Learning Research, vol. 227, PMLR, 2023, pp. 332–349.

[41] R. Orozco, A. Gahlot, and F. J. Herrmann, "Beacon: Bayesian experimental design acceleration with conditional normalizing flows − a case study in optimal monitor well placement for co $_2$ sequestration," *arXiv preprint arXiv:2404.00075*, 2024.

[42] R. Orozco, F. J. Herrmann, and P. Chen, "Probabilistic bayesian optimal experimental design using conditional normalizing flows," *arXiv preprint arXiv:2402.18337*, 2024.

[43] G. Ongie, A. Jalal, C. A. Metzler, R. G. Baraniuk, A. G. Dimakis, and R. Willett, "Deep learning techniques for inverse problems in imaging," *IEEE Journal on Selected Areas in Information Theory*, vol. 1, no. 1, pp. 39–56, 2020.

[44] V. Antun, F. Renna, C. Poon, B. Adcock, and A. C. Hansen, "On instabilities of deep learning in image reconstruction and the potential costs of ai," *Proceedings of the National Academy of Sciences*, vol. 117, no. 48, pp. 30 088–30 095, 2020.

[45] S. Bhadra, V. A. Kelkar, F. J. Brooks, and M. A. Anastasio, "On hallucinations in tomographic image reconstruction," *IEEE transactions on medical imaging*, vol. 40, no. 11, pp. 3249–3260, 2021.

[46] D. P. Kingma and P. Dhariwal, "Glow: Generative flow with invertible 1x1 convolutions," *Advances in neural information processing systems*, vol. 31, 2018.

[47] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, “Variational inference: A review for statisticians,” *Journal of the American statistical Association*, vol. 112, no. 518, pp. 859–877, 2017.

[48] L. Dinh, J. Sohl-Dickstein, and S. Bengio, “Density estimation using real nvp,” *arXiv preprint arXiv:1605.08803*, 2016.

[49] L. Ardizzone *et al.*, “Analyzing inverse problems with invertible neural networks,” *arXiv preprint arXiv:1808.04730*, 2018.

[50] J. Whang, E. Lindgren, and A. Dimakis, “Composing normalizing flows for inverse problems,” in *International Conference on Machine Learning*, PMLR, 2021, pp. 11 158–11 169.

[51] H. Sun and K. L. Bouman, “Deep probabilistic imaging: Uncertainty quantification and multi-modal solution characterization for computational imaging,” *arXiv preprint arXiv:2010.14462*, vol. 9, 2020.

[52] G. Rizzuti, A. Siahkoohi, P. A. Witte, and F. J. Herrmann, “Parameterizing uncertainty by deep invertible networks: An application to reservoir characterization,” in *Seg technical program expanded abstracts 2020*, Society of Exploration Geophysicists, 2020, pp. 1541–1545.

[53] X. Zhao, A. Curtis, and X. Zhang, “Bayesian seismic tomography using normalizing flows,” *Geophysical Journal International*, vol. 228, no. 1, pp. 213–239, 2022.

[54] A. Dima and V. Ntziachristos, “Non-invasive carotid imaging using optoacoustic tomography,” *Optics express*, vol. 20, no. 22, pp. 25 044–25 057, 2012.

[55] A. Buehler, M. Kacprowicz, A. Taruttis, and V. Ntziachristos, “Real-time handheld multispectral optoacoustic imaging,” *Optics letters*, vol. 38, no. 9, pp. 1404–1406, 2013.

[56] N. Kovachki, R. Baptista, B. Hosseini, and Y. Marzouk, “Conditional sampling with monotone gans,” *arXiv preprint arXiv:2006.06755*, 2020.

[57] J. Kruse, G. Detommaso, R. Scheichl, and U. Köthe, “Hint: Hierarchical invertible neural transport for density estimation and Bayesian inference,” *arXiv preprint arXiv:1905.10687*, 2019.

[58] R. Orozco, A. Siahkoohi, G. Rizzuti, T. van Leeuwen, and F. J. Herrmann, “Photoacoustic imaging with conditional priors from normalizing flows,” in *NeurIPS 2021 Workshop on Deep Learning and Inverse Problems*, 2021.

[59] A. Siahkoohi, R. Orozco, G. Rizzuti, and F. J. Herrmann, “Wave-equation-based inversion with amortized variational Bayesian inference,” *arXiv preprint arXiv:2203.15881*, 2022.

[60] S. Gershman and N. Goodman, “Amortized inference in probabilistic reasoning,” in *Proceedings of the annual meeting of the cognitive science society*, vol. 36, 2014.

[61] A. Siahkoohi, G. Rizzuti, R. Orozco, and F. J. Herrmann, “Reliable amortized variational inference with physics-based latent distribution correction,” *arXiv preprint arXiv:2207.11640*, 2022.

[62] M. C. Deans, “Maximally informative statistics for localization and mapping,” in *Proceedings 2002 IEEE International Conference on Robotics and Automation (Cat. No. 02CH37292)*, IEEE, vol. 2, 2002, pp. 1824–1829.

[63] R. Baptista *et al.*, “Bayesian model calibration for block copolymer self-assembly: Likelihood-free inference and expected information gain computation via measure transport,” *arXiv preprint arXiv:2206.11343*, 2022.

[64] J. Adler, S. Lunz, O. Verdier, C.-B. Schönlieb, and O. Öktem, “Task adapted reconstruction for inverse problems,” *Inverse Problems*, vol. 38, no. 7, p. 075 006, 2022.

[65] J. Pearl, *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan kaufmann, 1988.

[66] P. Gravel, G. Beaudoin, and J. A. De Guise, “A method for modeling noise in medical images,” *IEEE Transactions on medical imaging*, vol. 23, no. 10, pp. 1221–1232, 2004.

[67] A. Banerjee, X. Guo, and H. Wang, “On the optimality of conditional expectation as a bregman predictor,” *IEEE Transactions on Information Theory*, vol. 51, no. 7, pp. 2664–2669, 2005.

[68] J. Adler and O. Öktem, “Deep Bayesian inversion,” *arXiv preprint arXiv:1811.05910*, 2018.

[69] J. Leuschner, M. Schmidt, D. O. Baguer, and P. Maaß, “The lodopab-ct dataset: A benchmark dataset for low-dose ct reconstruction methods,” *arXiv preprint arXiv:1910.01113*, 2019.

[70] W. Van Aarle *et al.*, “The astra toolbox: A platform for advanced algorithm development in electron tomography,” *Ultramicroscopy*, vol. 157, pp. 35–47, 2015.

[71] A. Khorashadizadeh, K. Kothari, L. Salsi, A. A. Harandi, M. de Hoop, and I. Dokmani'c, “Conditional injective flows for bayesian imaging,” *arXiv preprint arXiv:2204.07664*, 2022.

[72] A. H. Andersen and A. C. Kak, “Simultaneous algebraic reconstruction technique (sart): A superior implementation of the art algorithm,” *Ultrasonic imaging*, vol. 6, no. 1, pp. 81–94, 1984.

[73] L. L. Geyer *et al.*, “State of the art: Iterative ct reconstruction techniques,” *Radiology*, vol. 276, no. 2, pp. 339–357, 2015.

[74] S. Ghosal and A. Van der Vaart, *Fundamentals of nonparametric Bayesian inference*. Cambridge University Press, 2017, vol. 44.

[75] H. Oja, “Affine invariant multivariate sign and rank tests and corresponding estimates: A review,” *Scandinavian Journal of Statistics*, vol. 26, no. 3, pp. 319–343, 1999.

[76] Y. Zhang, Y. Wang, and C. Zhang, “Total variation based gradient descent algorithm for sparse-view photoacoustic image reconstruction,” *Ultrasonics*, vol. 52, no. 8, pp. 1046–1055, 2012.

[77] J. Schwab, S. Antholzer, R. Nuster, and M. Haltmeier, “Real-time photoacoustic projection imaging using deep learning,” *arXiv preprint arXiv:1801.06693*, 2018.

[78] M.-H. Laves, S. Ihler, J. F. Fast, L. A. Kahrs, and T. Ortmaier, “Well-calibrated regression uncertainty in medical imaging with deep learning,” in *Medical Imaging with Deep Learning*, PMLR, 2020, pp. 393–412.

[79] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, “On calibration of modern neural networks,” in *International conference on machine learning*, PMLR, 2017, pp. 1321–1330.

[80] S. Talts, M. Betancourt, D. Simpson, A. Vehtari, and A. Gelman, “Validating Bayesian inference algorithms with simulation-based calibration,” *arXiv preprint arXiv:1804.06788*, 2018.

[81] K. e. Dines, F. Fry, J. Patrick, and R. Gilmor, “Computerized ultrasound tomography of the human head: Experimental results,” *Ultrasonic imaging*, vol. 3, no. 4, pp. 342–351, 1981.

[82] G. Becker *et al.*, “Reliability of transcranial colour-coded real-time sonography in assessment of brain tumours: Correlation of ultrasound, computed tomography and biopsy findings,” *Neuroradiology*, vol. 36, no. 8, pp. 585–590, 1994.

[83] S. Smith, O. Von Ramm, J. Kisslo, and F. Thurstone, "Real time ultrasound tomography of the adult brain.," *Stroke*, vol. 9, no. 2, pp. 117–122, 1978.

[84] P. Williamson, "A guide to the limits of resolution imposed by scattering in ray tomography," *Geophysics*, vol. 56, no. 2, pp. 202–207, 1991.

[85] L. Guasch, O. Calderón Agudo, M.-X. Tang, P. Nachev, and M. Warner, "Full-waveform inversion imaging of the human brain," *NPJ digital medicine*, vol. 3, no. 1, pp. 1–12, 2020.

[86] U. Taskin, K. S. Eikrem, G. Nævdal, M. Jakobsen, D. J. Verschuur, and K. W. Van Dongen, "Ultrasound imaging of the brain using full-waveform inversion," in *2020 IEEE International Ultrasonics Symposium (IUS)*, IEEE, 2020, pp. 1–4.

[87] P. Marty, C. Boehm, and A. Fichtner, "Acoustoelastic full-waveform inversion for transcranial ultrasound computed tomography," in *Medical Imaging 2021: Ultrasonic Imaging and Tomography*, SPIE, vol. 11602, 2021, pp. 210–229.

[88] J. Tong *et al.*, "Transcranial ultrasound imaging with decomposition descent learning-based full waveform inversion," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 69, no. 12, pp. 3297–3307, 2022.

[89] J. Cudeiro-Blanco *et al.*, "Design and construction of a low-frequency ultrasound acquisition device for 2-d brain imaging using full-waveform inversion," *Ultrasound in Medicine & Biology*, vol. 48, no. 10, pp. 1995–2008, 2022.

[90] O. Bates *et al.*, "A probabilistic approach to tomography and adjoint state methods, with an application to full waveform inversion in medical ultrasound," *Inverse Problems*, vol. 38, no. 4, p. 045 008, 2022.

[91] J. Virieux and S. Operto, "An overview of full-waveform inversion in exploration geophysics," *Geophysics*, vol. 74, no. 6, WCC1–WCC26, 2009.

[92] A. Tarantola and B. Valette, "Generalized nonlinear inverse problems solved using the least squares criterion," *Reviews of Geophysics*, vol. 20, no. 2, pp. 219–232, 1982.

[93] L. Ruthotto and E. Haber, "An introduction to deep generative modeling," *GAMM-Mitteilungen*, vol. 44, no. 2, e202100008, 2021.

[94] L. Dinh, D. Krueger, and Y. Bengio, "Nice: Non-linear independent components estimation," *arXiv preprint arXiv:1410.8516*, 2014.

[95] S. Bond-Taylor, A. Leach, Y. Long, and C. G. Willcocks, “Deep generative modelling: A comparative review of vaes, gans, normalizing flows, energy-based and autoregressive models,” *arXiv preprint arXiv:2103.04922*, 2021.

[96] L. Ardizzone, C. Lüth, J. Kruse, C. Rother, and U. Köthe, “Conditional invertible neural networks for guided image generation,” 2019.

[97] C. Winkler, D. Worrall, E. Hoogeboom, and M. Welling, “Learning likelihoods with conditional normalizing flows,” *arXiv preprint arXiv:1912.00042*, 2019.

[98] K. Cranmer, J. Brehmer, and G. Louppe, “The frontier of simulation-based inference,” *Proceedings of the National Academy of Sciences*, vol. 117, no. 48, pp. 30 055–30 062, 2020.

[99] R. Orozco, A. Siahkoohi, G. Rizzuti, T. van Leeuwen, and F. J. Herrmann, “Adjoint operators enable fast and amortized machine learning based Bayesian uncertainty quantification,” 2022.

[100] J. Alsing and B. Wandelt, “Generalized massive optimal data compression,” *Monthly Notices of the Royal Astronomical Society: Letters*, vol. 476, no. 1, pp. L60–L64, 2018.

[101] J. Fluri, T. Kacprzak, A. Refregier, A. Lucchi, and T. Hofmann, “Cosmological parameter estimation and inference using deep summaries,” *Physical Review D*, vol. 104, no. 12, p. 123 526, 2021.

[102] J.-F. Aubry, M. Tanter, M. Pernot, J.-L. Thomas, and M. Fink, “Experimental demonstration of noninvasive transskull adaptive focusing based on prior computed tomography scans,” *The Journal of the Acoustical Society of America*, vol. 113, no. 1, pp. 84–93, 2003.

[103] J. Zbontar *et al.*, “Fastmri: An open dataset and benchmarks for accelerated mri,” *arXiv preprint arXiv:1811.08839*, 2018.

[104] F. Luporini *et al.*, “Architecture and performance of devito, a system for automated stencil computation,” *ACM Trans. Math. Softw.*, vol. 46, no. 1, Apr. 2020.

[105] M. Louboutin *et al.*, “Devito (v3.1.0): An embedded domain-specific language for finite differences and geophysical exploration,” *Geoscientific Model Development*, vol. 12, no. 3, pp. 1165–1187, 2019.

[106] P. Witte, G. Rizzuti, M. Louboutin, A. Siahkoohi, and F. Herrmann, *Invertiblenetworks. jl: A julia framework for invertible neural networks*, 2020.

[107] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *International Conference on Medical image computing and computer-assisted intervention*, Springer, 2015, pp. 234–241.

[108] E. Baysal, D. D. Kosloff, and J. W. Sherwood, “Reverse time migration,” *Geophysics*, vol. 48, no. 11, pp. 1514–1524, 1983.

[109] D. Gudovskiy, S. Ishizaka, and K. Kozuka, “Cflow-ad: Real-time unsupervised anomaly detection with localization via conditional normalizing flows,” in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 98–107.

[110] J. Marino, Y. Yue, and S. Mandt, “Iterative amortized inference,” in *International Conference on Machine Learning*, PMLR, 2018, pp. 3403–3412.

[111] M. Grcić, I. Grubišić, and S. Šegvić, “Densely connected normalizing flows,” *Advances in Neural Information Processing Systems*, vol. 34, pp. 23 968–23 982, 2021.

[112] C. Shorten and T. M. Khoshgoftaar, “A survey on image data augmentation for deep learning,” *Journal of big data*, vol. 6, no. 1, pp. 1–48, 2019.

[113] A. Denker, M. Schmidt, J. Leuschner, and P. Maass, “Conditional invertible neural networks for medical imaging,” *Journal of Imaging*, vol. 7, no. 11, p. 243, 2021.

[114] R. Barbano, C. Zhang, S. Arridge, and B. Jin, “Quantifying model uncertainty in inverse problems via bayesian deep gradient descent,” in *2020 25th International Conference on Pattern Recognition (ICPR)*, IEEE, 2021, pp. 1392–1399.

[115] T. Teshima, I. Ishikawa, K. Tojo, K. Oono, M. Ikeda, and M. Sugiyama, “Coupling-based invertible neural networks are universal diffeomorphism approximators,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 3362–3373, 2020.

[116] F. Draxler, L. Kühmichel, A. Rousselot, J. Müller, C. Schnörr, and U. Köthe, “On the convergence rate of gaussianization with random rotations,” *arXiv preprint arXiv:2306.13520*, 2023.

[117] G. Detommaso, J. Kruse, L. Ardizzone, C. Rother, U. Köthe, and R. Scheichl, “Hint: Hierarchical invertible neural transport for general and sequential Bayesian inference,” *stat*, vol. 1050, p. 25, 2019.

[118] D. Donoho, “Data science at the singularity,” *arXiv preprint arXiv:2310.00865*, 2023.

[119] A. Gao, J. Castellanos, Y. Yue, Z. Ross, and K. Bouman, “Deepgem: Generalized expectation-maximization for blind inversion,” *Advances in Neural Information Processing Systems*, vol. 34, pp. 11 592–11 603, 2021.

[120] M. Morzfeld, J. Adams, S. Lunderman, and R. Orozco, “Feature-based data assimilation in geophysics,” *Nonlinear Processes in Geophysics*, vol. 25, no. 2, pp. 355–374, 2018.

[121] A. F. Heavens, R. Jimenez, and O. Lahav, “Massive lossless data compression and multiple parameter estimation from galaxy spectra,” *Monthly Notices of the Royal Astronomical Society*, vol. 317, no. 4, pp. 965–972, 2000.

[122] E. L. Lehmann and G. Casella, *Theory of point estimation*. Springer Science & Business Media, 2006.

[123] T. Van Leeuwen and F. J. Herrmann, “Mitigating local minima in full-waveform inversion by expanding the search space,” *Geophysical Journal International*, vol. 195, no. 1, pp. 661–667, 2013.

[124] J. Alsing, B. Wandelt, and S. Feeney, “Massive optimal data compression and density estimation for scalable, likelihood-free inference in cosmology,” *Monthly Notices of the Royal Astronomical Society*, vol. 477, no. 3, pp. 2874–2885, 2018.

[125] T. Robins, J. Camacho, O. C. Agudo, J. L. Herraiz, and L. Guasch, “Deep-learning-driven full-waveform inversion for ultrasound breast imaging,” *Sensors*, vol. 21, no. 13, p. 4570, 2021.

[126] K. Wang, T. Matthews, F. Anis, C. Li, N. Duric, and M. A. Anastasio, “Waveform inversion with source encoding for breast sound speed reconstruction in ultrasound computed tomography,” *IEEE transactions on ultrasonics, ferroelectrics, and frequency control*, vol. 62, no. 3, pp. 475–493, 2015.

[127] T. C. Robins *et al.*, “Dual-probe transcranial full-waveform inversion: A brain phantom feasibility study,” *Ultrasound in Medicine & Biology*, vol. 49, no. 10, pp. 2302–2315, 2023.

[128] H. Thomson, “Ultrasonic differentiation of healthy and cancerous neural tissue,” Ph.D. dissertation, University of Glasgow, 2023.

[129] C. Cueto *et al.*, “Spatial response identification enables robust experimental ultrasound computed tomography,” *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 69, no. 1, pp. 27–37, 2021.

[130] I. Espin, N. Salaun, H. Jiang, and M. Reinier, “From fwi to ultra-high-resolution imaging,” *The Leading Edge*, vol. 42, no. 1, pp. 16–23, 2023.

[131] E. Esser, L. Guasch, F. J. Herrmann, and M. Warner, "Constrained waveform inversion for automatic salt flooding," *The Leading Edge*, vol. 35, no. 3, pp. 235–239, 2016.

[132] E. Esser, L. Guasch, T. van Leeuwen, A. Y. Aravkin, and F. J. Herrmann, "Total variation regularization strategies in full-waveform inversion," *SIAM Journal on Imaging Sciences*, vol. 11, no. 1, pp. 376–406, 2018.

[133] L. Guasch, M. Warner, and C. Ravaut, "Adaptive waveform inversion: Practice," *Geophysics*, vol. 84, no. 3, R447–R461, 2019.

[134] M. I. Iacono *et al.*, "Mida: A multimodal imaging-based detailed anatomical model of the human head and neck," *PloS one*, vol. 10, no. 4, e0124126, 2015.

[135] R.-E. Plessix, "A review of the adjoint-state method for computing the gradient of a functional with geophysical applications," *Geophysical Journal International*, vol. 167, no. 2, pp. 495–503, 2006.

[136] S. Mukherjee, M. Carioni, O. Öktem, and C.-B. Schönlieb, "End-to-end reconstruction meets data-driven regularization for inverse problems," *Advances in Neural Information Processing Systems*, vol. 34, pp. 21 413–21 425, 2021.

[137] P. Marty, C. Boehm, and A. Fichtner, "Shape optimization for transcranial ultrasound computed tomography," in *Medical Imaging 2023: Ultrasonic Imaging and Tomography*, SPIE, vol. 12470, 2023, pp. 77–88.

[138] R. Barbano, S. Arridge, B. Jin, and R. Tanno, "Uncertainty quantification in medical image synthesis," in *Biomedical Image Synthesis and Simulation*, Elsevier, 2022, pp. 601–641.

[139] A. Siahkoohi, G. Rizzuti, and F. J. Herrmann, "Weak deep priors for seismic imaging," in *SEG Technical Program Expanded Abstracts 2020*, Society of Exploration Geophysicists, 2020, pp. 2998–3002.

[140] K. Sohn, H. Lee, and X. Yan, "Learning structured output representation using deep conditional generative models," *Advances in neural information processing systems*, vol. 28, 2015.

[141] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[142] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," *arXiv preprint arXiv:2011.13456*, 2020.

[143] C. Jones, J. Edgar, J. Selvage, and H. Crook, “Building complex synthetic models to evaluate acquisition geometries and velocity inversion technologies,” in *74th EAGE Conference and Exhibition incorporating EUROPEC 2012*, European Association of Geoscientists & Engineers, 2012, cp–293.

[144] A. Tarantola, “Inversion of seismic reflection data in the acoustic approximation,” *Geophysics*, vol. 49, no. 8, pp. 1259–1266, 1984.

[145] M. Fehler and P. J. Keliher, “Model development,” in *SEAM Phase 1: Challenges of Subsalt Imaging in Tertiary Basins, with Emphasis on Deepwater Gulf of Mexico*, Society of Exploration Geophysicists, 2011, pp. 15–50.

[146] M. Izzatullah, A. Alali, M. Ravasi, and T. Alkhalifah, “Physics-reliable frugal local uncertainty analysis for full waveform inversion,” *Geophysical Prospecting*, 2024.

[147] L. Yang, O. M. Saad, G. Wu, and T. Alkhalifah, “Conditional image prior for uncertainty quantification in full waveform inversion,” *arXiv preprint arXiv:2408.09975*, 2024.

[148] X. Zhao and A. Curtis, “Variational prior replacement in bayesian inference and inversion,” *Geophysical Journal International*, vol. 239, no. 2, pp. 1236–1256, 2024.

[149] L. Qu, M. Araya-Polo, and L. Demanet, “Uncertainty quantification in seismic inversion through integrated importance sampling and ensemble methods,” *arXiv preprint arXiv:2409.06840*, 2024.

[150] C. Sun, A. Malcolm, R. Kumar, and W. Mao, “Enabling uncertainty quantification in a standard full-waveform inversion method using normalizing flows,” *Geophysics*, vol. 89, no. 5, R493–R507, 2024.

[151] Y. Sun and P. Williamson, “Invertible neural networks for uncertainty quantification in refraction tomography,” *The Leading Edge*, vol. 43, no. 6, pp. 358–366, 2024.

[152] A. P. Muller *et al.*, “Deep-tomography: Iterative velocity model building with deep learning,” *Geophysical Journal International*, vol. 232, no. 2, pp. 975–989, 2023.

[153] A. Alali, V. Kazei, M. Kalita, and T. Alkhalifah, “Deep learning unflooding for robust subsalt waveform inversion,” *Geophysical Prospecting*, vol. 72, no. Machine learning applications in geophysical exploration and monitoring, pp. 7–19, 2023.

[154] J. Ho, A. Jain, and P. Abbeel, “Denoising diffusion probabilistic models,” *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[155] D. Brunet, E. R. Vrscay, and Z. Wang, "On the mathematical properties of the structural similarity index," *IEEE Transactions on Image Processing*, vol. 21, no. 4, pp. 1488–1499, 2011.

[156] K. Zheng, C. Lu, J. Chen, and J. Zhu, "Improved techniques for maximum likelihood estimation for diffusion odes," in *International Conference on Machine Learning*, PMLR, 2023, pp. 42 363–42 389.

[157] T. Karras, M. Aittala, T. Aila, and S. Laine, "Elucidating the design space of diffusion-based generative models," *Advances in neural information processing systems*, vol. 35, pp. 26 565–26 577, 2022.

[158] G. Batzolis, J. Stanczuk, C.-B. Schönlieb, and C. Etmann, "Conditional image generation with score-based diffusion models," *arXiv preprint arXiv:2111.13606*, 2021.

[159] L. Baldassari, A. Siahkoohi, J. Garnier, K. Solna, and M. V. de Hoop, "Conditional score-based diffusion models for bayesian inference in infinite dimensions," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[160] B. Bloem-Reddy, Y. Whye, *et al.*, "Probabilistic symmetries and invariant neural networks," *Journal of Machine Learning Research*, vol. 21, no. 90, pp. 1–61, 2020.

[161] J. Hou and W. W. Symes, "Accelerating extended least-squares migration with weighted conjugate gradient iteration," *Geophysics*, vol. 81, no. 4, S165–S179, 2016.

[162] P. Witte, M. Yang, and F. Herrmann, "Sparsity-promoting least-squares migration with the linearized inverse scattering imaging condition: 79th conference and exhibition, eage," in *Expanded Abstracts*, 2017.

[163] F. Luporini *et al.*, "Architecture and performance of devito, a system for automated stencil computation," *ACM Transactions on Mathematical Software (TOMS)*, vol. 46, no. 1, pp. 1–28, 2020.

[164] M. Louboutin, "Slimgroup/imagegather.jl: V0.2.10 https://doi.org/10.5281/zenodo.12575836," *Zenodo*, 2024.

[165] T. P. Merrifield *et al.*, "Synthetic seismic data for training deep learning networks," *Interpretation*, vol. 10, no. 3, SE31–SE39, 2022.

[166] S. Alemohammad *et al.*, "Self-consuming generative models go mad," *arXiv preprint arXiv:2307.01850*, 2023.

[167] G. Hennenfent and F. J. Herrmann, “Simply denoise: Wavefield reconstruction via jittered undersampling,” *Geophysics*, vol. 73, no. 3, pp. V19–V28, 2008.

[168] Z. Wu, Y. Sun, Y. Chen, B. Zhang, Y. Yue, and K. L. Bouman, “Principled probabilistic imaging using diffusion models as plug-and-play priors,” *arXiv preprint arXiv:2405.18782*, 2024.

[169] J. Tachella and M. Pereyra, “Equivariant bootstrapping for uncertainty quantification in imaging inverse problems,” *arXiv preprint arXiv:2310.11838*, 2023.

[170] P. Putzky and M. Welling, “Recurrent inference machines for solving inverse problems,” *arXiv preprint arXiv:1706.04008*, 2017.

[171] K. Gregor and Y. LeCun, “Learning fast approximations of sparse coding,” in *Proceedings of the 27th international conference on international conference on machine learning*, 2010, pp. 399–406.

[172] D. Donoho, “Data science at the singularity,” *Harvard Data Science Review*, vol. 6, no. 1, 2024.

[173] H. T. Erdinc, R. Orozco, and F. J. Herrmann, “Generative geostatistical modeling from incomplete well and imaged seismic observations with diffusion models,” *arXiv preprint arXiv:2406.05136*, 2024.

[174] D. V. Lindley, “On a measure of the information provided by an experiment,” *The Annals of Mathematical Statistics*, vol. 27, no. 4, pp. 986–1005, 1956.

[175] A. Foster *et al.*, “Variational bayesian optimal experimental design,” *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[176] T. Hoffmann and J.-P. Onnela, “Minimizing the expected posterior entropy yields optimal summary statistics,” *arXiv preprint arXiv:2206.02340*, 2022.

[177] A. Foster, M. Jankowiak, M. O’Meara, Y. W. Teh, and T. Rainforth, “A unified stochastic gradient approach to designing bayesian-optimal experiments,” in *International Conference on Artificial Intelligence and Statistics*, PMLR, 2020, pp. 2959–2969.

[178] L. Ardizzone, C. Lüth, J. Kruse, C. Rother, and U. Köthe, “Guided image generation with conditional invertible neural networks,” *arXiv preprint arXiv:1907.02392*, 2019.

[179] F. Draxler, S. Wahl, C. Schnörr, and U. Köthe, “On the universality of coupling-based normalizing flows,” *arXiv preprint arXiv:2402.06578*, 2024.

[180] C. D. Bahadir, A. V. Dalca, and M. R. Sabuncu, “Learning-based optimization of the under-sampling pattern in mri,” in *Information Processing in Medical Imaging: 26th International Conference, IPMI 2019, Hong Kong, China, June 2–7, 2019, Proceedings 26*, Springer, 2019, pp. 780–792.

[181] Y. Bengio, N. Léonard, and A. Courville, “Estimating or propagating gradients through stochastic neurons for conditional computation,” *arXiv preprint arXiv:1308.3432*, 2013.

[182] T. Yin, Z. Wu, H. Sun, A. V. Dalca, Y. Yue, and K. L. Bouman, “End-to-end sequential sampling and reconstruction for mr imaging,” in *Proceedings of the Machine Learning for Health Conference*, 2021.

[183] J. Zhang *et al.*, “Extending loupe for k-space under-sampling pattern optimization in multi-coil mri,” in *Machine Learning for Medical Image Reconstruction: Third International Workshop, MLMIR 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 8, 2020, Proceedings 3*, Springer, 2020, pp. 91–101.

[184] W. Wu and K. L. Miller, “Image formation in diffusion mri: A review of recent technical developments,” *Journal of Magnetic Resonance Imaging*, vol. 46, no. 3, pp. 646–662, 2017.

[185] K. Wu, P. Chen, and O. Ghattas, “A fast and scalable computational framework for large-scale high-dimensional bayesian optimal experimental design,” *SIAM/ASA Journal on Uncertainty Quantification*, vol. 11, no. 1, pp. 235–261, 2023.

[186] N. Kennamer, S. Walton, and A. Ihler, “Design amortization for bayesian optimal experimental design,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, 2023, pp. 8220–8227.

[187] T. Wang, F. Lucka, and T. van Leeuwen, “Sequential experimental design for x-ray ct using deep reinforcement learning,” *arXiv preprint arXiv:2307.06343*, 2023.

[188] T. Helin, N. Hyvonen, and J.-P. Puska, “Edge-promoting adaptive bayesian experimental design for x-ray imaging,” *SIAM Journal on Scientific Computing*, vol. 44, no. 3, B506–B530, 2022.

[189] S. Wu, D. J. Verschuur, and G. Blacquière, “Automated seismic acquisition geometry design for optimized illumination at the target: A linearized approach,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2021.

[190] T. Hascoet, W. Zhuang, Q. Febvre, Y. Ariki, and T. Takiguchi, “Reducing the memory cost of training convolutional neural networks by cpu offloading,” *Journal of Software Engineering and Applications*, vol. 12, no. 8, pp. 307–320, 2019.

[191] F. Bieder, J. Wolleb, A. Durrer, R. Sandkuehler, and P. C. Cattin, "Memory-efficient 3d denoising diffusion models for medical image processing," in *Medical Imaging with Deep Learning*, 2023.

[192] G. Wu, P. Coupé, Y. Zhan, B. Munsell, and D. Rueckert, "Patch-based techniques in medical imaging," *Cham: Springer International Publishing, 2017, Lecture Notes in Computer Science*, 2015.

[193] F. Altekrüger, A. Denker, P. Hagemann, J. Hertrich, P. Maass, and G. Steidl, "Patchnr: Learning from very few images by patch normalizing flow regularization," *Inverse Problems*, vol. 39, no. 6, p. 064 006, 2023.

[194] I. Goodfellow *et al.*, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[195] Y. Sun, Z. Wu, Y. Chen, B. T. Feng, and K. L. Bouman, "Provable probabilistic imaging using score-based generative priors," *arXiv preprint arXiv:2310.10835*, 2023.

[196] A. Hauptmann *et al.*, "Model-based learning for accelerated, limited-view 3-d photoacoustic tomography," *IEEE transactions on medical imaging*, vol. 37, no. 6, pp. 1382–1393, 2018.

[197] B. A. Kaplan, J. Buchmann, S. Prohaska, and J. Laufer, "Monte-carlo-based inversion scheme for 3d quantitative photoacoustic tomography," in *Photons Plus Ultrasound: Imaging and Sensing 2017*, SPIE, vol. 10064, 2017, pp. 802–814.

[198] A. P. Reeves *et al.*, "A public image database to support research in computer aided diagnosis," in *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, IEEE, 2009, pp. 3715–3718.

[199] J. Brehmer, G. Louppe, J. Pavez, and K. Cranmer, "Mining gold from implicit models to improve likelihood-free inference," *Proceedings of the National Academy of Sciences*, vol. 117, no. 10, pp. 5242–5249, 2020.

[200] Y. Zeng, R. Orozco, Z. Yin, and F. J. Herrmann, "Enhancing full-waveform variational inference through stochastic resampling and data augmentation," in *International Meeting for Applied Geoscience and Energy*, Aug. 2024.

[201] T. J. Grady *et al.*, "Model-parallel fourier neural operators as learned surrogates for large-scale parametric pdes," *Computers & Geosciences*, vol. 178, p. 105 402, 2023.

[202] B. Peters and F. J. Herrmann, “Algorithms and software for projections onto intersections of convex and non-convex sets with applications to inverse problems,” *arXiv preprint arXiv:1902.09699*, 2019.