DEEP GENERATIVE MODELS FOR SOLVING GEOPHYSICAL INVERSE PROBLEMS

A Dissertation
Presented to
The Academic Faculty

By

Ali Siahkoohi

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

August 2022

DEEP GENERATIVE MODELS FOR SOLVING GEOPHYSICAL INVERSE PROBLEMS

Thesis committee:

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

Dr. Justin Romberg
School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Tobin Isaac
School of Computational Science and Enginnering
Georgia Institute of Technology

Dr. Yao Xie
School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Edmond Chow
School of Computational Science and Enginnering
Georgia Institute of Technology

Date approved: July 15, 2022

To my family

ACKNOWLEDGMENTS

Thanks are due to my adviser, Professor Felix J. Herrmann, for everything I learned from you, for having the flexibility to pursue my research interests, and for the frequent and inspiring discussions we had, which usually involved copious amounts of coffee.

Thanks are also due to my committee members, Professor Tobin Isaac, Professor Edmond Chow, Professor Justin Romberg, and Professor Yao Xie for serving on my Ph.D. committee, reviewing my thesis, and providing valuable feedback.

Thanks to my current and former colleagues at the Seismic Laboratory for Imaging and Modeling (SLIM) for being open to collaboration. I would like to thank Miranda Joyce for her administrative support over the first few years. Special thanks to Henryk Modzelewski, Mathias Louboutin, and Philipp Witte who extensively contributed to the SLIM software environment, which I tremendously benefited from. I would also like to thank Rajiv Kumar and Gabrio Rizzuti for their mentorship and discussing valuable research ideas.

Many thanks to Michael Chinen, Tom Denton, W. Bastiaan Kleijn, and Jan Skoglund for being great mentors during my research internship at Google.

I would like to acknowledge the support of my wonderful family. Thank you, Mom and Dad. Thanks to my amazing wife, Sonia Sargolzaei, for your love, patience, support, and for always being interested in hearing about my research. I am also grateful to Lou and Betty, my feline companions, who taught me the importance of tenacity and sound sleep, respectively. Last but not least,

> “I wanna thank me. I wanna thank me for believing in me. I wanna thank me for doing all this hard work. I wanna thank me for having no days off. I wanna thank me for, for never quitting. I wanna thank me for always being a giver and tryna give more than I recieve. I wanna thank me for tryna do more right than wrong. I wanna thank me for just being me at all times.”
>
> — Calvin Cordozar Broadus Jr.

My studies and this research was partially funded by the University of British Columbia, the Natural Sciences and Engineering Research Council of Canada Collaborative Research and Development Grant (DNOISE II), members of the SINBAD consortium, as well as the Georgia Institute of Technology, the Georgia Research Alliance, and partners of the ML4Seismic Center.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

My thesis presents several novel methods to facilitate solving large-scale inverse problems by utilizing recent advances in machine learning, and particularly deep generative modeling. Inverse problems involve reliably estimating unknown parameters of a physical model from indirect observed data that are noisy. Solving inverse problems presents primarily two challenges. The first challenge is to capture and incorporate prior knowledge into ill-posed inverse problems whose solutions cannot be uniquely identified. The second challenge is the computational complexity of solving inverse problems, particularly the cost of quantifying uncertainty. The main goal of this thesis is to address these issues by developing practical data-driven methods that are scalable to geophysical applications in which access to high-quality training data is often limited.

There are six papers included in this thesis. A majority of these papers focus on addressing computational challenges associated with Bayesian inference and uncertainty quantification, while others focus on developing regularization techniques to improve inverse problem solution quality and accelerate the solution process. These papers demonstrate the applicability of the proposed methods to seismic imaging, a large-scale geophysical inverse problem with a computationally expensive forward operator for which sufficiently capturing the variability in the Earth's heterogeneous subsurface through a training dataset is challenging.

The first two papers present computationally feasible methods of applying a class of methods commonly referred to as deep priors to seismic imaging and uncertainty quantification. I also present a systematic Bayesian approach to translate uncertainty in seismic imaging to uncertainty in downstream tasks performed on the image. The next two papers aim to address the reliability concerns surrounding data-driven methods for solving Bayesian inverse problems by leveraging variational inference formulations that offer the benefits of fully-learned posteriors while being directly informed by physics and data. The

last two papers are concerned with correcting forward modeling errors where the first proposes an adversarially learned postprocessing step to attenuate numerical dispersion artifacts in wave-equation simulations due to coarse finite-difference discretizations, while the second trains a Fourier neural operator surrogate forward model in order to accelerate the qualification of uncertainty due to errors in the forward model parameterization.

CHAPTER 1
INTRODUCTION

The aim of this thesis is to address some of the challenges associated with solving large-scale inverse problems, in particular geophysical inverse problems [1]. An inverse problem involves the estimation of an unknown quantity (model) using noisy indirect measurements, commonly known as observed data. The relationship between the unknown model and observed data is typically described by a physical phenomenon, which can be modeled using a forward operator. In particular, geophysical inverse problems aim to construct an image of the Earth's subsurface by extracting physical properties of subsurface media [1] given geophysical data such as seismic, electrical, electromagnetic, gravitational, and magnetic measurements. These images are of crucial importance for exploration of natural resources [2], monitoring geohazards such as earthquakes [3], as well as carbon control and carbon sequestration [4, 5, 6], which has the potential to mitigate the effects of climate change. The thesis primarily focuses on seismic imaging due to its ability to provide higher resolution images of deep hidden structures in the subsurface of the Earth [1].

1.1 Problem statement

Solving inverse problems, such as seismic imaging, is often challenged by noise, modeling errors, and a nontrivial null-space of the forward operator. These challenges may lead to non-unique solutions in which various models fit the observed data equally well. Instead of relying on a single model estimate, the inverse problem solution non-uniqueness can be characterized by a distribution over the solution space, namely, the posterior distribution [7]. The posterior distribution can be sampled to extract statistical information from the posterior distribution such as an assessment the variability among the possible solutions to the inverse problem, i.e., uncertainty quantification.

In general, posterior inference presents two challenges. The first step to formulating the posterior density is selecting a prior distribution that encodes prior knowledge about the unknown quantity. The selection of prior distributions is essential in Bayesian approaches as it biases the inference. Prior knowledge has traditionally been incorporated into the solution of inverse problems through the construction of prior distributions that satisfy regularity conditions regarding parameters of models or derivatives of these parameters [8, 9, 10, 11, 12, 13, 14]. Despite its popularity in controlled settings due to its simplicity and applicability, this type of priors may lead to undesirable biases in the outcome of Bayesian inference. In contrast, recent approaches based on deep learning [15, 16, 17, 18, 19, 18, 20, 21, 22] learn a prior distribution from available model samples. These methods provide a better understanding of the available prior information in comparison with generic handcrafted priors, however, their usage is typically limited to domains in which high-quality training data already exists. In reality, such an assumption rarely holds for certain applications. For example, unlike in medical imaging in which variability between patients is limited, in geophysical applications Earth's heterogeneity across geological scenarios might limit the feasibility of data-driven approaches that heavily rely on pretraining.

The second challenge associated with posterior inference relates to its computational cost. Extraction of information from the posterior distribution typically involves high-dimensional sampling. Sampling is computationally expensive, and existing methods, such as Markov chain Monte Carlo [MCMC, 23] techniques, often scale poorly. MCMC methods sample the posterior distribution by taking a series of random walks in the probability space, where the posterior probability density function needs to be approximated or evaluated at each step. In general, these sampling methods require numerous steps to traverse the probability space [8, 9, 11, 12, 24, 25, 26, 27, 28, 29, 30, 31], thereby limiting their applicability to large-scale problems. Deep learning methods provide a means to train generative models—a family of deep neural networks—which can be used to accelerate the sampling process from the posterior distribution [32, 33, 34, 16, 35, 18, 36, 37]. Even though most of these methods reduce the costs of Bayesian inference in high-dimensional settings, they rely on having access to a training dataset that captures the full training distribution. While this assumption may hold for certain domains, such as medical imaging, where variability among patients is generally low, in geophysical applications, due to the strong heterogeneity of Earth, it is difficult to capture the full prior or posterior distribution using a training dataset.

1.2 Objective

The computational cost of large-scale Bayesian inference, challenges associated with choosing a prior distribution, and lack of access to training pairs, highlight the importance of developing data-driven methods for solving inverse problems that can withstand changes in data distribution during inference. The aim of this thesis is to develop generative-model-based methods for enhancing the solution quality of inverse problems, shortening the time to solution, and reducing the cost of uncertainty quantification when access to training data is limited. This is significant since it allows us to utilize deep generative models to reliably solve inverse problems in real-world scenarios.

1.3 Seismic imaging

Seismic imaging deals with, extracting physical properties of subsurface media, e.g., seismic velocity and density from large volumes of measurements that are recorded at the Earth's surface [2]. The wave equation represents the underlying physical data generation phenomena, linking the mentioned media properties of the Earth subsurface to observed data.

1.3.1 Data acquisition

Observed data in seismic imaging is obtained by firing several active sources on the Earth surface, e.g., a vibroseis truck used on land or an air-gun used offshore, and recording the Earth's response via multiple receivers. The receivers are typically spread out on the surface of the survey area, and they record the wave pressure (offshore) or particle motion (in land) for each source experiment.

The collection of the recorded pressure or particle motion at all receiver locations for each source experiment is referred to as a shot record, denoted by $\mathbf{d}_i \in \mathbb{R}^{n_d}$, $i = 1, \ldots, n_s$, where n_s is the number of sources and $n_d = n_r {\cdot} n_t$ represents the data dimensionality, which is comprised of n_r receivers each recording the data with n_t discretized time samples.

1.3.2 The forward problem

The relationship between the observed data and the unknown squared-slowness model $\mathbf{m} \in \mathbb{R}^n$, defined on a discretized spatial gird, is encoded in the following forward model,

$$\mathbf{d}_i = \mathcal{F}(\mathbf{m}, \mathbf{q}_i) + \boldsymbol{\epsilon}_i, \quad i = 1, \ldots, n_s, \tag{1.1}$$

where $\mathbf{q}_i \in \mathbb{R}^N$ with $N = n \cdot n_t$ denotes the source signature defined on a temporal and spatial grid, $\boldsymbol{\epsilon}_i \in \mathbb{R}^{n_d}$ measurement noise, and $\mathcal{F}(\mathbf{m}, \mathbf{q}_i)$ is the forward operator, which is parameterized by the squared-slowness model and source signature. In the simplest acoustic form, the forward operator is related to the solution of the acoustic wave equation via

$$\mathcal{F}(\mathbf{m}, \mathbf{q}_i) = \mathbf{P}\mathbf{A}(\mathbf{m})^{-1}\mathbf{q}_i. \tag{1.2}$$

In the expression above, $\mathbf{P} : \mathbb{R}^N \to \mathbb{R}^{n_d}$ is the linear receiver projection operator, which restricts the observed data to the location of the receivers, and the linear operator $\mathbf{A}(\mathbf{m}) : \mathbb{R}^N \to \mathbb{R}^N$ represents the discretized (in time and space) acoustic wave equation.

The general workflow of imaging the subsurface structure is a two-step process. The first step is to reconstruct a background squared-slowness model, $\mathbf{m}_0 \in \mathbb{R}^n$, which describes the long-wavelength characteristics of the subsurface and correctly predicts the kinematics of wave propagation in the true subsurface. The second step is to image the short-wavelength subsurface features $\delta\mathbf{m} \in \mathbb{R}^{n}$—i.e., an image of the perturbation with respect to the background squared-slowness model—using the squared-slowness model from the first step [38, 39]. The ensemble of the long and short-wavelength subsurface features provides an detailed description of the subsurface structures.

1.3.3 Linearized forward problem

The seismic imaging linearized forward model can be obtained by linearizing the nonlinear relationship between seismic data and the unknown squared-slowness mode in equation 1.1. This process involves linearly approximating the nonlinear forward operator $\mathcal{F}(\mathbf{m}, \mathbf{q}_i)$ via Taylor's series expansion around a known smooth background squared-slowness model $\mathbf{m}_0$,

$$\mathcal{F}\left(\mathbf{m}, \mathbf{q}_i\right) \approx \mathcal{F}\left(\mathbf{m}_0, \mathbf{q}_i\right) + J(\mathbf{m}, \mathbf{q}_i)\delta\mathbf{m} + \mathcal{O}\left(\delta\mathbf{m}^{\top}\delta\mathbf{m}\right), \tag{1.3}$$

with $\delta\mathbf{m} = \mathbf{m} - \mathbf{m}_0$ being the unknown perturbation model, commonly referred to as the seismic image, and $J(\mathbf{m}, \mathbf{q}_i)$ the linearized Born scattering operator derived as

$$\begin{aligned} J(\mathbf{m}, \mathbf{q}_i) &= \frac{\partial \mathcal{F}\left(\mathbf{m}_0, \mathbf{q}_i\right)}{\partial \mathbf{m}}\bigg|_{\mathbf{m}=\mathbf{m}_0} \\ &= -\mathbf{P}\mathbf{A}(\mathbf{m}_0)^{-1}\mathrm{diag}\left(\frac{\partial \mathbf{A}(\mathbf{m})}{\partial \mathbf{m}}\bigg|_{\mathbf{m}=\mathbf{m}_0}\mathbf{A}(\mathbf{m}_0)^{-1}\mathbf{q}_i\right). \end{aligned} \tag{1.4}$$

Due to discarding the rest of the terms in the Taylor's series expansion in equation 1.3, in addition to measurement noise, seismic imaging forward model also include linearization errors $\boldsymbol{\eta}_i \in \mathbb{R}^{n_d}$,

$$\delta\mathbf{d}_i = J(\mathbf{m}, \mathbf{q}_i)\delta\mathbf{m} + \boldsymbol{\epsilon}_i + \boldsymbol{\eta}_i, \quad i = 1, \ldots, n_s, \tag{1.5}$$

with $\delta\mathbf{d}_i = \mathbf{d}_i - \mathcal{F}\left(\mathbf{m}_0, \mathbf{q}_i\right)$ referred to as linearized data.

Estimating $\delta \mathbf{m}$ from linearized data $\delta \mathbf{d}_i, \ i = 1, \ldots, n_s$ is challenged by noise, linearization errors, and a nontrivial null-space of the Born scattering forward operator. These challenges may result in non-unique solutions where different models may fit the observed data equally well.

1.4 The outline of my contributions

This thesis six papers, each of which partially addresses the challenges associated with solving seismic imaging. Each chapter follows the general structure of a technical journal article and begins with a summary, followed by an introduction into the respective topic, and then describes the main contribution and provides numerical examples. Below I outline these chapters by summarizing the problem and my contribution.

- Chapter 2 (published in [40]) proposes regularizing seismic imaging with deep priors and describes an approach to translate uncertainty in seismic imaging to uncertainty in tasks performed on the image, such as horizon tracking. My contribution includes a systematic approach to translate uncertainty due to noise in the data to confidence intervals of automatically tracked horizons in the image. In this approach I characterized the uncertainty a convolutional neural network (CNN) and to assess these uncertainties, I propose sampling from the posterior distribution of the CNN weights, used to parameterize the image. I employ the stochastic gradient Langevin dynamics sampling method to sample from the posterior distribution, which is designed to handle large-scale Bayesian inference problems with computationally expensive forward operators as in seismic imaging. Using the these posterior samples, aside from obtaining a robust alternative to maximum a posteriori estimate that is prone to overfitting, I use these samples to translate uncertainty in the image, due to noise in the data, to uncertainty on the tracked horizons.

- Chapter 3 (published in [41]) proposes weak deep priors, a computationally convenient formulation that relaxes deep priors. My contribution includes relaxing the requirement that the seismic image must lie in the deep prior CNN range, and letting the unknowns deviate from the network output according to a Gaussian distribution. To solve the inverse problem, I propose jointly solving for the reflectivity model and CNN weights. The chief advantage of this approach is that the gradient with resect to the CNN weights no longer involves the action of the forward operator Jacobian adjoint, making regularization with deep priors computationally feasible in seismic imaging, while still providing the regularization benefits of deep priors.

- Chapter 4 (published in [42]) proposes a reliable variational-inference-based multi-fidelity scheme for accelerating sampling from the posterior distribution of large-scale inverse problems by being tied to the physics. My contribution includes a data-driven preconditioning scheme that accelerates the minimization of an variational inference problem, which involves approximating the posterior distribution with a normalizing flow. I pretrain a conditional normalizing flow that is able to approximate the posterior distribution over a family of observed data for a specific inverse problem. I accelerate the variational inference problem including approximating the posterior distribution of interest by finetuning the weights of the pretrained conditional normalizing flow such that it better approximates the posterior distribution. Through experiments I indicate that the preconditioned variational inference method accelerates the inference process compared to the non-preconditioned approach, which is important for high-dimensional Bayesian inference settings.

- Chapter 5 proposes techniques from variational inference to train a deep neural network that accelerates Bayesian inference while being robust to data distribution shifts. My contribution includes learning a physics-based correction to the conditional normalizing flow latent distribution, which is capable of approximating the posterior distribution for previously unseen data, to provide a more accurate approximation to the posterior distribution for the observed data at hand. I accomplish this by parameterizing the conditional normalizing flow latent distribution by a Gaussian distribution with an unknown mean and diagonal covariance, which are estimated by minimizing the Kullback-Leibler divergence between the corrected posterior distribution estimate and the true posterior distribution. By means of a realistic seismic imaging example I show that for a certain class of shifts to the data likelihood and prior distribution, this approach provides reliable posterior samples with a limited computational cost.

- Chapter 6 (publised in [43]) shows the importance of transfer learning for data-driven solutions in domains such as geophysical inverse problems where access to high-quality training data is limited. My contribution includes pretraining a CNN within the generative adversarial networks framework to conduct complex data processing tasks including removal of the free-surface effects and mitigation of the effects of numerical dispersion during reverse-time migration and wave simulations. As part of dealing with limited training data challenge, I expose the CNNs to only a small percentage of training data pertinent to the task at hand. By means of numerical experiments, I demonstrates that as long data from a neighboring survey is available, finetuning this network with only a few low- and high-fidelity pairs pertinent to the current model results can provide a good performance for the task at hand. This strategy may lead to improvements in efficiency where computationally expensive (e.g., wave-equation driven) processing can partly be replaced by a potentially numerically more efficient neural network.

- Chapter 7 (published in [44]) proposes a data-driven method to reduce the computational cost of quantifying seismic imaging uncertainty due to errors in the background model parameterization of the forward operator. My contribution includes a survey-specific Fourier neural operator surrogate to velocity continuation that maps seismic images associated with one background model to another virtually for free. While being trained with only 200 background and seismic image pairs, this surrogate is able to accurately predict seismic images associated with new background models, thus accelerating seismic imaging uncertainty quantification. I support this method with a realistic data example in which I quantify seismic imaging uncertainties using a Fourier neural operator surrogate, illustrating how variations in background models affect the position of reflectors in a seismic image.

1.5 References

[1] R. E. Sheriff and L. P. Geldart, *Exploration seismology*. Cambridge University Press, 1995 (page 1).

[2] O. Yilmaz, *Seismic data analysis: Processing, inversion, and interpretation of seismic data*. Society of Exploration Geophysicists, 2001 (pages 1, 3).

[3] S. R. McNutt, "Seismic monitoring and eruption forecasting of volcanoes: A review of the state-of-the-art and case histories," *Monitoring and mitigation of volcano hazards*, pp. 99–146, 1996 (page 1).

[4] D. E. Lumley, "Time-lapse seismic reservoir monitoring," *Geophysics*, vol. 66, no. 1, pp. 50–53, 2001 (page 1).

[5] O. Eiken, I. Brevik, R. Arts, E. Lindeberg, and K. Fagervik, "Seismic monitoring of co2 injected into a marine acquifer," in *SEG Technical Program Expanded Abstracts 2000*, Society of Exploration Geophysicists, 2000, pp. 1623–1626 (page 1).

[6] R. Arts, O. Eiken, A. Chadwick, P. Zweigel, L. van der Meer, and B. Zinszner, "Monitoring of co2 injected at sleipner using time lapse seismic data," in *Greenhouse Gas Control Technologies - 6th International Conference*, J. Gale and Y. Kaya, Eds., Oxford: Pergamon, 2003, pp. 347–352, ISBN: 978-0-08-044276-1 (page 1).

[7] A. Tarantola, *Inverse problem theory and methods for model parameter estimation*. SIAM, 2005, ISBN: 978-0-89871-572-9 (page 1).

[8] A. Malinverno and V. A. Briggs, "Expanded uncertainty quantification in inverse problems: Hierarchical bayes and empirical bayes," *GEOPHYSICS*, vol. 69, no. 4, pp. 1005–1016, 2004 (page 2).

[9] 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. eprint: http://epubs.siam.org/doi/pdf/10.1137/110845598 (page 2).

[10] G. Ely, A. Malcolm, and O. V. Poliannikov, "Assessing uncertainties in velocity models and images with a fast nonlinear uncertainty quantification method," *GEOPHYSICS*, vol. 83, no. 2, R63–R75, 2018. eprint: https://doi.org/10.1190/geo2017-0321.1 (page 2).

[11] Z. Fang, C. D. Silva, R. Kuske, and F. J. Herrmann, "Uncertainty quantification for inverse problems with weak partial-differential-equation constraints," *GEOPHYSICS*, vol. 83, no. 6, R629–R647, 2018 (page 2).

[12] G. K. Stuart, S. E. Minkoff, and F. Pereira, "A two-stage Markov chain Monte Carlo method for seismic inversion and uncertainty quantification," *GEOPHYSICS*, vol. 84, no. 6, R1003–R1020, Nov. 2019 (page 2).

[13] F. J. Herrmann and X. Li, "Efficient least-squares imaging with sparsity promotion and compressive sensing," *Geophysical Prospecting*, vol. 60, no. 4, pp. 696–712, Jul. 2012 (page 2).

[14] T. van Leeuwen, A. Y. Aravkin, and F. Herrmann, "Seismic Waveform Inversion by Stochastic Optimization," *International Journal of Geophysics*, vol. 2011, pp. 1–18, 2011 (page 2).

[15] A. Bora, A. Jalal, E. Price, and A. G. Dimakis, "Compressed sensing using generative models," in *Proceedings of the 34th International Conference on Machine Learning*, D. Precup and Y. W. Teh, Eds., ser. Proceedings of Machine Learning Research, vol. 70, PMLR, Jun. 2017, pp. 537–546 (page 2).

[16] J. Adler and O. öktem, "Deep Bayesian Inversion," *arXiv preprint arXiv:1811.05910*, 2018 (page 2).

[17] P. Putzky and M. Welling, "Invert to Learn to Invert," in *Advances in Neural Information Processing Systems*, vol. 32, Curran Associates, Inc., 2019 (page 2).

[18] M. Asim, M. Daniels, O. Leong, A. Ahmed, and P. Hand, "Invertible generative models for inverse problems: Mitigating representation error and dataset bias," in *Proceedings of the 37th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 119, PMLR, Jul. 2020, pp. 399–409 (page 2).

[19] A. Sriram *et al.*, "End-to-End Variational Networks for Accelerated MRI Reconstruction," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2020, pp. 64–73 (page 2).

[20] A. Hauptmann and B. T. Cox, "Deep Learning in Photoacoustic Tomography: Current approaches and future directions," *Journal of Biomedical Optics*, vol. 25, no. 11, p. 112 903, 2020 (page 2).

[21] 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 (page 2).

[22] S. Mukherjee, M. Carioni, O. öktem, and C.-B. Schönlieb, "End-to-end reconstruction meets data-driven regularization for inverse problems," *arXiv preprint arXiv:2106.03538*, 2021 (page 2).

[23] C. Robert and G. Casella, *Monte Carlo statistical methods*. Springer-Verlag, 2004 (page 2).

[24] A. Malinverno and R. L. Parker, "Two ways to quantify uncertainty in geophysical inverse problems," *GEOPHYSICS*, vol. 71, no. 3, W15–W27, 2006 (page 2).

[25] A. Ray, S. Kaplan, J. Washbourne, and U. Albertin, "Low frequency full waveform seismic inversion within a tree based Bayesian framework," *Geophysical Journal International*, vol. 212, no. 1, pp. 522–542, Oct. 2017. eprint: https://academic.oup.com/gji/article-pdf/212/1/522/21782947/ggx428.pdf (page 2).

[26] Q. Liu, "Stein variational gradient descent as gradient flow," in *Advances in Neural Information Processing Systems*, I. Guyon *et al.*, Eds., vol. 30, Curran Associates, Inc., 2017 (page 2).

[27] Z. Zhao and M. K. Sen, "A gradient based MCMC method for FWI and uncertainty analysis," in *89th Annual International Meeting, SEG*, Expanded Abstracts, 2019, pp. 1465–1469 (page 2).

[28] M. Kotsi, A. Malcolm, and G. Ely, "Uncertainty quantification in time-lapse seismic imaging: a full-waveform approach," *Geophysical Journal International*, vol. 222, no. 2, pp. 1245–1263, May 2020 (page 2).

[29] X. Zhang and A. Curtis, "Seismic tomography using variational inference methods," *Journal of Geophysical Research: Solid Earth*, vol. 125, no. 4, e2019JB018589, 2020 (page 2).

[30] X. Zhao, A. Curtis, and X. Zhang, "Bayesian seismic tomography using normalizing flows," *Geophysical Journal International*, vol. 228, no. 1, pp. 213–239, Jul. 2021. eprint: https://academic.oup.com/gji/article-pdf/228/1/213/40348424/ggab298.pdf (page 2).

[31] G. Rizzuti, A. Siahkoohi, P. A. Witte, and F. J. Herrmann, "Parameterizing uncertainty by deep invertible networks, an application to reservoir characterization," in *90th Annual International Meeting, SEG*, Sep. 2020, pp. 1541–1545 (page 2).

[32] A. Bora, A. Jalal, E. Price, and A. G. Dimakis, *Compressed sensing using generative models*, 2017. arXiv: 1703.03208 `[stat.ML]` (page 2).

[33] I. Goodfellow *et al.*, "Generative Adversarial Nets," in *Proceedings of the 27th International Conference on Neural Information Processing Systems*, ser. NIPS'14,

Montreal, Canada, 2014, pp. 2672–2680. eprint: http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf (page 2).

[34] L. Mosser, O. Dubrule, and M. Blunt, “Stochastic Seismic Waveform Inversion Using Generative Adversarial Networks as a Geological Prior,” *Mathematical Geosciences*, vol. 84, no. 1, pp. 53–79, 2019 (page 2).

[35] D. Rezende and S. Mohamed, “Variational inference with normalizing flows,” ser. Proceedings of Machine Learning Research, vol. 37, PMLR, Jul. 2015, pp. 1530–1538 (page 2).

[36] J. Kruse, G. Detommaso, R. Scheichl, and U. Köthe, “HINT: Hierarchical Invertible Neural Transport for Density Estimation and Bayesian Inference,” *Proceedings of AAAI-2021*, 2021 (page 2).

[37] K. Kothari, A. Khorashadizadeh, M. de Hoop, and I. Dokmanić, “Trumpets: Injective Flows for Inference and Inverse Problems,” *arXiv preprint arXiv:2102.10461*, 2021 (page 2).

[38] J. K. Cohen and N. Bleistein, “Velocity inversion procedure for acoustic waves,” *Geophysics*, vol. 44, no. 6, pp. 1077–1087, 1979 (page 5).

[39] P. Jaiswal, C. A. Zelt, A. W. Bally, and R. Dasgupta, “2-D traveltime and waveform inversion for improved seismic imaging: Naga Thrust and Fold Belt, India,” *Geophysical Journal International*, vol. 173, no. 2, pp. 642–658, 2008 (page 5).

[40] A. Siahkoohi, G. Rizzuti, and F. J. Herrmann, “Deep Bayesian inference for seismic imaging with tasks,” *Geophysics*, vol. 87, no. 5, Jun. 2022 (page 6).

[41] A. Siahkoohi, G. Rizzuti, and F. J. Herrmann, “Weak deep priors for seismic imaging,” in *90th Annual International Meeting, SEG*, Expanded Abstracts, Sep. 2020, pp. 2998–3002 (page 6).

[42] A. Siahkoohi, G. Rizzuti, M. Louboutin, P. Witte, and F. J. Herrmann, “Preconditioned training of normalizing flows for variational inference in inverse problems,” in *3rd Symposium on Advances in Approximate Bayesian Inference*, Jan. 2021 (page 7).

[43] A. Siahkoohi, M. Louboutin, and F. J. Herrmann, “The importance of transfer learning in seismic modeling and imaging,” *GEOPHYSICS*, vol. 84, no. 6, A47–A52, Nov. 2019 (page 8).

[44] A. Siahkoohi, M. Louboutin, and F. J. Herrmann, “Velocity continuation with Fourier neural operators for accelerated uncertainty quantification,” in *2nd International Meeting for Applied Geoscience & Energy*, 2022 (page 8).

CHAPTER 2

DEEP BAYESIAN INFERENCE FOR SEISMIC IMAGING WITH TASKS

2.1 Summary

We propose to use techniques from Bayesian inference and deep neural networks to translate uncertainty in seismic imaging to uncertainty in tasks performed on the image, such as horizon tracking. Seismic imaging is an ill-posed inverse problem because of bandwidth and aperture limitations, which is hampered by the presence of noise and linearization errors. Many regularization methods, such as transform-domain sparsity promotion, have been designed to deal with the adverse effects of these errors, however, these methods run the risk of biasing the solution and do not provide information on uncertainty in the image space and how this uncertainty impacts certain tasks on the image. A systematic approach is proposed to translate uncertainty due to noise in the data to confidence intervals of automatically tracked horizons in the image. The uncertainty is characterized by a convolutional neural network (CNN) and to assess these uncertainties, samples are drawn from the posterior distribution of the CNN weights, used to parameterize the image. Compared to traditional priors, it is argued in the literature that these CNNs introduce a flexible inductive bias that is a surprisingly good fit for a diverse set of problems, including medical imaging, compressive sensing, and diffraction tomography. The method of stochastic gradient Langevin dynamics is employed to sample from the posterior distribution. This method is designed to handle large scale Bayesian inference problems with computationally expensive forward operators as in seismic imaging. Aside from offering a robust alternative to maximum a posteriori estimate that is prone to overfitting, access to these samples allow us to translate uncertainty in the image, due to noise in the data, to uncertainty on the tracked horizons. For instance, it admits estimates for the pointwise standard deviation on

the image and for confidence intervals on its automatically tracked horizons.

2.2 Introduction

Due to the presence of shadow zones, coherent linearization errors, and noisy finite-aperture measured data, seismic imaging involves an ill-conditioned linear inverse problem [1, 2, 3]. Relying on a single estimate for the model may be subject to data overfit [4] and negatively impacts the quality of the obtained seismic image and tasks performed on it. Casting the seismic imaging problem into a probabilistic framework allows for a more comprehensive description of its solution space [5]. The "solution" of the inverse problem is then a probability distribution over the model space and is commonly referred to as the posterior distribution.

Aside from the computational challenges associated with uncertainty quantification (UQ) in geophysical inverse problems [4, 5, 6, 7, 8, 9, 10, 11, 12], the choice of prior distributions in Bayesian frameworks is crucial. Recent attempts mostly rely on handcrafted priors, i.e., priors chosen solely based on their simplicity and applicability. For example, restricting feasible solutions to layered media with specific orientations [13, 14, 15], or satisfying regularity conditions related to model parameters or derivatives thereof [4, 6, 16, 17, 7, 18, 19, 20, 21, 9, 10, 11, 22, 12]. While effective in controlled settings, handcrafted priors might introduce unwanted bias to the solution. Recent deep-learning based approaches [23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34], on the other hand, learn a prior distribution from available data[1]. While certainly providing a better description of the available prior information when compared to generic handcrafted priors, they may affect the outcome of Bayesian inference more seriously when out-of-distribution data is considered, e.g., when the training data is not fully representative of a given scenario. Unfortunately, unlike deep-learning based inversion approaches in other imaging modalities,

[1] We use the word "data" interchangeably. In the context of inverse problems "data" refers to observed data. In the context of machine learning data-driven priors refer to priors derived from available samples from an unknown distribution. These are also commonly referred to as "data". The meaning of the word "data" will be clear from the context.

e.g., medical imaging [35, 36, 37, 38, 39, 40], we generally do not have access to high-fidelity information about the Earth's subsurface. This, together with the Earth's strong heterogeneity across geological scenarios, might limit the scope of data-driven approaches that heavily rely on pretraining [41, 42, 43, 44, 45, 46, 47, 48, 49, 50].

In this work, we take advantage of a novel prior recently deployed in computer vision and geophysics [51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63], known as the deep prior, which utilizes the inductive bias [64] of untrained convolutional neural networks (CNNs) as a prior. This approach is tantamount to restricting feasible models to the range of an untrained CNN with a fixed input and randomly initialized weights. Via this reparameterization the weights of the CNN become the new unknowns in seismic imaging and this change of variable leads to a "prior" on the image space that excludes noisy artifacts, as long as overfitting is prevented [51]. This has the potential benefit of being less restrictive than handcrafted priors while not needing training data as approaches based on using pretrained networks [51]. To formally cast the deep prior into a Bayesian framework, we impose a Gaussian distribution on the CNN weights, which is a common regularization strategy in training deep CNNs [65, 66]. To perform uncertainty quantification for seismic imaging, we sample from the posterior distribution of the CNN weights by running preconditioned stochastic gradient Langevin dynamics [SGLD, 67, 68], a gradient-based Markov chain Monte Carlo (MCMC) sampling method developed for Bayesian inference of deep CNNs with large training datasets.

A crucial objective of our study is translating the uncertainty in seismic imaging to uncertainty in downstream tasks such as horizon tracking, semantic segmentation, and tracking CO_2 plumes in carbon capture and sequestration projects. Horizon tracking, which this chapter focuses on, is a task performed after imaging that leads to a stratigraphic model. Horizon trackers use well data and seismic images to delineate stratigraphy and are typically sensitive to structural and stratigraphic unconformities. In these challenging areas, the horizons do not continuously extend spatially, e.g., may be discontinuous due to vertical displacement via faults, hence tracking horizons across unconformities may not be trivial. Failure to include uncertainty on tracked horizons have major implications on the identification of risk. Since the accuracy of horizon tracking is directly linked to the quality of the seismic image, we systematically incorporate uncertainties of seismic imaging into horizon tracking. We achieve this by feeding samples from the imaging posterior to an automatic horizon tracker [69] and obtain an ensemble of likely horizons in the image. Compared to conventional imaging and manual tracking of horizons, our approach allows us to rigorously quantify uncertainty in the location of the horizons due to noise in shot records and modeling errors, e.g., linearization errors. Our probabilistic framework also admits nondeterministic horizon trackers, e.g., uncertain control points or multiple human interpreters. There are parallels between the probabilistic framework we propose for quantifying uncertainty in downstream tasks and the interrogation theory [70]. The purpose of this theory is to answer questions about an unknown quantity by designing experiments (inverse problems) that facilitate answering the question. The probabilistic framework we developed can be described as an application of interrogation theory in that the seismic survey and shot records are provided with no need to design further experiments, and the question involves quantifying uncertainty in horizon tracking. Our probabilistic framework differs fundamentally from other recently developed automatic seismic horizon trackers based on machine learning [see e.g., 71, 72, 73, 74] because horizon uncertainty is ultimately driven by data (through the intermediate imaging distribution), and not from label (control point) uncertainty alone.

In the following sections, we first mathematically formulate deep-prior based seismic imaging, by introducing the likelihood function and the deep prior approach. Next, we describe our proposed SGLD-based sampling approach and its challenges. Subsequently, we introduce a framework to tie uncertainties in imaging to uncertainties in horizon tracking, which allows for deterministic and nondeterministic horizon tracking. We present two realistic examples derived from real seismic image volumes obtained in different geological settings. These numerical experiments are designed to showcase the ability of the proposed deep-prior based approach to produce seismic images with limited artifacts. We conclude by demonstrating our probabilistic horizon tracking approach, which includes estimates for confidence intervals associated with the imaged horizons in the two aforementioned examples.

2.3 Theory

The goal of this chapter is to understand how errors in the data due to noise and linearization assumptions affect the uncertainty of seismic images and typical tasks carried out on these images. We begin with an introduction of the linearized forward model, which forms the basis of seismic imaging via reverse-time migration and discuss Bayesian imaging with regularization via so-called deep priors.

2.3.1 Seismic imaging

In its simplest acoustic form, reverse-time migration follows directly from linearizing the acoustic wave equation around a known, slowly varying background model—i.e., the spatial distribution of the squared slowness. Traditionally, the process of seismic imaging is concerned with estimating the short-wavelength components of the squared-slowness, denoted by $\delta \mathbf{m}$, from n_s processed shot records collected in the vector $\mathbf{d} = \{\mathbf{d}_i\}_{i=1}^{n_s}$. In most cases, these indirect measurements are recorded along the surface or ocean bottom with sources $\{\mathbf{q}_i\}_{i=1}^{n_s}$ located at or near the surface. The placement of the sources and receiver near the surface leads to more uncertainty in the deeper areas of the image.

The unknown ground truth perturbation model $\delta \mathbf{m}^*$ is linearly related to the data via

$$\mathbf{d}_i = \mathbf{J}(\mathbf{m}_0, \mathbf{q}_i)\delta \mathbf{m}^* + \boldsymbol{\epsilon}_i, \quad \boldsymbol{\epsilon}_i \sim \mathrm{N}(\mathbf{0}, \sigma^2 \mathbf{I}), \tag{2.1}$$

where $\mathbf{J}(\mathbf{m}_0, \mathbf{q}_i)$ corresponds to the linearized Born scattering operator for the ith source and the background squared slowness model $\mathbf{m}_0$. Because of possible errors in the processed data, the presence of noise, and linearization errors, the above expression contains the noise term $\boldsymbol{\epsilon}_i$. While other choices can be made, we assume this noise to be distributed according to a zero-centered Gaussian distribution with known covariance matrix $\sigma^2\mathbf{I}$. For small σ and a kinematically correct background model $\mathbf{m}_0$, the above linear relationship can be inverted by minimizing

$$\min_{\delta\mathbf{m}} \sum_{i=1}^{n_s} \left\| \mathbf{d}_i - \mathbf{J}(\mathbf{m}_0, \mathbf{q}_i)\delta\mathbf{m} \right\|_2^2. \tag{2.2}$$

While this approach is in principle capable of producing high-fidelity true-amplitude images [75, 76, 77], the noise term is in practice never negligible and may adversely affect the image quality [3] especially in situations where the source spectrum is narrow band and the aperture limited. Therefore not only the problem in equation 2.2 requires regularization but also calls for a statistical inference framework that allows us to draw conclusions in the presence of uncertainty.

2.3.2 Probabilistic imaging with Bayesian inference

To account for uncertainties in the image induced by the random noise term $\boldsymbol{\epsilon}_i$ in equation 7.1, we follow the seminal work of [5] and cast our noisy imaging as a Bayesian inverse problem. Instead of calculating a single image by solving equation 2.2, we assign probabilities to a family of images that fit the observed data to various degrees. This distribution is known as the posterior distribution. In this Bayesian framework, the solution to the inverse problem, i.e., the image, and the noise in the observed data are considered random variables. According to Bayes' rule, the conditional posterior distribution, denoted by p_{post}, states that

$$p_{\text{post}}\left(\delta\mathbf{m} \mid \mathbf{d}\right) \propto p_{\text{like}}(\mathbf{d} \mid \delta\mathbf{m})\, p_{\text{prior}}(\delta\mathbf{m}). \tag{2.3}$$

In this expression, p_{like} is the likelihood function, which is related to the probability density function (PDF) of the noise, and p_{prior} is the prior PDF of the image, which encodes prior beliefs on the unknown perturbations $\delta\mathbf{m}$. This prior distribution assigns probabilities to all potential seismic images before incorporating the data via the likelihood. The constant of proportionality in equation 2.3 corresponds to the PDF of the observed data, which is independent of $\delta\mathbf{m}$. Based on the distribution of the noise, the likelihood term measures how well the forward modeled data (equation 7.1) and observed data agree.

As stated by Bayes' rule, the posterior PDF of $\delta\mathbf{m}$, denoted by $p_{\text{post}}\left(\delta\mathbf{m} \mid \mathbf{d}\right)$, is proportional to the product of the likelihood and the prior PDF, given observed data. The log-likelihood function takes the following form:

$$\begin{aligned} -\log p_{\text{like}}\left(\mathbf{d} \mid \delta\mathbf{m}\right) &= -\sum_{i=1}^{n_s} \log p_{\text{like}}\left(\mathbf{d}_i \mid \delta\mathbf{m}\right) \\ &= \frac{1}{2\sigma^2}\sum_{i=1}^{n_s} \left\|\mathbf{d}_i - \mathbf{J}(\mathbf{m}_0, \mathbf{q}_i)\delta\mathbf{m}\right\|_2^2 + \text{const}, \end{aligned} \tag{2.4}$$

where the constant term is independent of $\delta\mathbf{m}$. For the uncorrelated Gaussian noise assumption, this negative log-likelihood function equals the squared ℓ_2-norm of the residual scaled by the noise variance σ^2.

Aside from depending on the residual, i.e., the difference between observed and modeled data, for each shot record, the choice of the prior influences the posterior distribution. Before the advent of data-driven methods involving generative neural networks, the definitions of priors were mostly handcrafted and often based on somewhat ad hoc Gaussian or Laplacian distributions in the physical or in some transformed domain [16, 17, 18, 19]. While these approaches have proven to be useful and are theoretically well understood [78], there is always a risk of a biased outcome something we would like to avoid. On the other hand, using pretrained generative networks as priors has proven to be effective [79, 23, 37, 25, 48]. However, their success hinges on the quality of pretraining and having access to a fully representative training data that accurately captures the prior distribution. Since we are dealing with highly complex heterogeneity of the Earth subsurface to which we have limited access, we will stay away from data-driven methods to train a neural network to act as a prior.

2.3.3 Deep priors

Following recent work by [51], we avoid using the need to have access to realizations of true perturbations by using untrained generative CNNs as priors. We fix a random input latent variable and use a randomly initialized [80] CNN with a special architecture [51] to reparameterize the unknown perturbations $\delta \mathbf{m}$ in terms of CNN weights. Given the shot data, we minimize the data misfit with respect to the CNN weights on which we impose a Gaussian prior. The CNN's architecture (see details in Appendix A) and the Gaussian prior imposed on its weights act as a regularization in the image space that avoids representing incoherent noisy artifacts, as long as overfitting is prevented [51].

To be more specific, let $g(\mathbf{z}, \mathbf{w}) \in \mathbb{R}^N$, with the N the number of gridpoints in the image, denote a untrained, specially designed, CNN [51] with fixed input $\mathbf{z} \sim \mathrm{N}(\mathbf{0}, \mathbf{I})$ with the same size as the image and unknown weights $\mathbf{w} \in \mathbb{R}^M$ with $M \gg N$. Restricting the unknown perturbation model to the output of the CNN, i.e., $\delta \mathbf{m} = g(\mathbf{z}, \mathbf{w})$, corresponds to a nonlinear representation for the image and the following expression for the likelihood function:

$$
\begin{aligned}
-\log p_{\text{like}}\left(\mathbf{d} \mid \mathbf{w}\right) &= -\sum_{i=1}^{n_s} \log p_{\text{like}}\left(\mathbf{d}_i \mid \mathbf{w}\right) \\
&= \frac{1}{2\sigma^2} \sum_{i=1}^{n_s} \left\| \mathbf{d}_i - \mathbf{J}(\mathbf{m}_0, \mathbf{q}_i) g(\mathbf{z}, \mathbf{w}) \right\|_2^2 + \text{const},
\end{aligned}
\tag{2.5}
$$

with the constant term independent of $\mathbf{w}$. In essence, deep priors correspond to a nonlinear "change of variables" where the unknowns are the CNN weights and the image is constrained to the range of the CNN output for a fixed random input. Compared to data-driven methods, no training samples are needed. While the nonlinearity makes it more difficult to minimize the likelihood term (the likelihood in equation 2.4), a zero-centered Gaussian prior for the weights with covariance $\lambda^{-2}\mathbf{I}$ suffices thanks to the overparameterization of the CNN ($M \gg N$). With this Gaussian prior on the weights, the posterior distribution for the weights given the data reads

$$p_{\text{post}}\left(\mathbf{w} \mid \mathbf{d}\right) \propto \left[\prod_{i=1}^{n_s} p_{\text{like}}\left(\mathbf{d}_i \mid \mathbf{w}\right)\right] \mathrm{N}\big(\mathbf{w} \mid \mathbf{0}, \lambda^{-2}\mathbf{I}\big) \tag{2.6}$$

where $\mathrm{N}\big(\mathbf{w} \mid \mathbf{0}, \lambda^{-2}\mathbf{I}\big)$ stands for the PDF of the zero-centered Gaussian prior. Given equation 2.5, the negative log-posterior distribution becomes

$$\begin{aligned} -\log p_{\text{post}}\left(\mathbf{w} \mid \mathbf{d}\right) &= -\left[\sum_{i=1}^{n_s} \log p_{\text{like}}\left(\mathbf{d}_i \mid \mathbf{w}\right)\right] - \log \mathrm{N}\big(\mathbf{w} \mid \mathbf{0}, \lambda^{-2}\mathbf{I}\big) + \text{const} \\ &= \frac{1}{2\sigma^2} \sum_{i=1}^{n_s} \big\|\mathbf{d}_i - \mathbf{J}(\mathbf{m}_0, \mathbf{q}_i) g(\mathbf{z}, \mathbf{w})\big\|_2^2 + \frac{\lambda^2}{2} \big\|\mathbf{w}\big\|_2^2 + \text{const.} \end{aligned} \tag{2.7}$$

Compared to conventional formulations of Bayesian inference, knowledge of the deep prior resides both in the likelihood term, through the reparameterization of the image as the output of a CNN, and in the traditional λ weighted ℓ_2-norm squared term. This is different from the traditional Bayesian settings where prior information resides exclusively in the prior term. [53] provided a theoretical Bayesian perspective on deep priors, describing them as Gaussian process priors in classical Bayesian terms. Specifically, [53] showed that for infinitely wide CNNs, i.e. CNNs with a large number of channels, the inductive bias of the CNN architecture and the Gaussian prior on its weights are equivalent to a stationary Gaussian process prior in the image space. [53] also explicitly made a connection between the kernel of this Gaussian process and the architecture of a CNN, by characterizing the effects of convolutions, non-linearities, up-sampling, down-sampling, and skip connections, which provides insights on selecting an appropriate CNN architecture. Independently, [57] argue that the weak form of our constrained formulation with deep priors yield the same solutions for the correct Lagrange multiplier. This means there is a direct connection between our formulation and unconstrained variational approaches. The latter permit a straightforward Bayesian interpretation.

Aside from choosing the right CNN architecture [51, 57], random initialization of its weight [80] and fixed input for the latent variable, the above posterior depends on selecting a value for the tradeoff parameter $\lambda > 0$, which weighs the importance of the Gaussian prior against the noise-variance weighted data misfit term in the likelihood function. In the sections below, we will comment how to choose the value for λ.

The above expression for the posterior in equation 2.7 forms the basis of our proposed probabilistic imaging scheme based on Bayesian inference. Before discussing how to sample from this distribution, we first briefly describe how to extract various statistical properties from this posterior distribution on the image. Specifically, we will review how to obtain point estimates [81], including maximum likelihood estimate (MLE) , maximum a posteriori estimate (MAP) and estimates for the mean and pointwise standard deviation, and 99% confidence intervals.

2.3.4 Estimation with Bayesian inference

Based on the expressions for the negative log-likelihood (equation 2.5) and posterior (equation 2.7), we derive expressions for different point and interval estimates [52].

Maximum likelihood estimation

To establish a baseline for image estimates obtained without regularization, we first consider point estimates for the image that correspond to finding an image that best fits the observed data. Since this estimate is obtained by maximizing the likelihood function with respect to the unknown image, $\delta\mathbf{m}$, this estimate is known as the MLE. The corresponding optimization problem can be written as

$$\begin{aligned}
\delta\mathbf{m}_{\text{MLE}} &= \underset{\delta\mathbf{m}}{\arg\min} - \log p_{\text{like}}\left(\mathbf{d} \mid \delta\mathbf{m}\right) \\
&= \underset{\delta\mathbf{m}}{\arg\min}\, \frac{1}{2\sigma^2} \sum_{i=1}^{n_s} \left\|\mathbf{d}_i - \mathbf{J}(\mathbf{m}_0, \mathbf{q}_i)\delta\mathbf{m}\right\|_2^2,
\end{aligned} \tag{2.8}$$

where the last equality follows from equation 2.4. Observe that the MLE corresponds to the deterministic least-squares solution, yielded by equation 2.2. Unfortunately, MLE images are prone to overfitting [81, 82] that results in imaging artifacts [3].

Maximum a posteriori estimation

Adding regularization to inverse problems, including seismic imaging, is known to limit overfitting and is capable of filling in at least part of the null space of the modeling operator. In case of regularization with deep priors, this corresponds to finding the image that maximizes the posterior distribution, i.e., we have

$$\begin{aligned}
\mathbf{w}_{\text{MAP}} &= \underset{\mathbf{w}}{\arg\max}\, p_{\text{post}}\left(\mathbf{w} \mid \mathbf{d}\right) \\
&= \underset{\mathbf{w}}{\arg\min}\, \frac{1}{2\sigma^2} \sum_{i=1}^{n_s} \left\|\mathbf{d}_i - \mathbf{J}(\mathbf{m}_0, \mathbf{q}_i) g(\mathbf{z}, \mathbf{w})\right\|_2^2 + \frac{\lambda^2}{2}\left\|\mathbf{w}\right\|_2^2.
\end{aligned} \tag{2.9}$$

This estimation for the weights $\mathbf{w}$ is known as the MAP estimate. Given this estimate $\mathbf{w}_{\text{MAP}}$, the corresponding estimate for the image is obtained via

$$\delta\mathbf{m}_{\text{MAP}} = g(\mathbf{z}, \mathbf{w}_{\text{MAP}}). \tag{2.10}$$

When compared with MAP estimates computed from traditional Bayesian formulations of linear inverse problems, the estimate in equation 2.9 has several important differences. The above estimate depends on the random initializations of the weights, $\mathbf{w}$ and latent variable $\mathbf{z}$, which is due to the nonlinearity introduced by the reparameterization. This renders the above minimization non-convex, i.e., its local minimum is no longer guaranteed to coincide with the global minimum. While the objective is non-convex [83], as a result of deep prior reparameterization, because $M \gg N$, first-order stochastic optimization methods [84, 85, 86] are able to minimize the objective function in equation 2.9 to small values of the residual [87, 88]. Several other challenges include increased number of iterations, establishment of a stopping criterion when maximizing equation 2.9 to prevent overfitting, and the quantification of uncertainty. Despite these challenge, we argue that invoking the deep prior outweighs the challenges since it offers a better bias-variance trade-off and requires knowledge of only a single hyperparameter. In addition, we refer to [62] for an alternative formulation, which reduces the number of iterations and therefore the number of evaluations of the computationally expensive forward modeling operator.

Conditional mean estimation

So far, the MLE and MAP estimates involved a deterministic (at least for fixed initialization of the network and latent variable) procedure maximizing the likelihood or posterior. Since we have access to the unnormalized posterior PDF, $p_{\text{post}}\left(\delta\mathbf{w} \mid \mathbf{d}\right)$ (equation 2.7), we have in principle ways to retrieve information on the statistical moments of the posterior distribution of the unknown perturbation including its mean and pointwise standard deviation. However, contrary to the two estimates discussed so far these point estimates can typically only be approximated with samples drawn from the posterior.

We obtain access to samples from the posterior, $p_{\text{post}}\left(\delta\mathbf{m} \mid \mathbf{d}\right)$ via a "push-forward" of samples from $p_{\text{post}}\left(\mathbf{w} \mid \mathbf{d}\right)$ based on the deterministic map $\delta\mathbf{m} = g(\mathbf{z}, \mathbf{w})$ for fixed $\mathbf{z}$ [89]. As a result, for any sample of the weights, $\mathbf{w}$, drawn from $p_{\text{post}}\left(\mathbf{w} \mid \mathbf{d}\right)$, we have

$$g(\mathbf{z}, \mathbf{w}) \sim p_{\text{post}}\left(\delta\mathbf{m} \mid \mathbf{d}\right). \tag{2.11}$$

Assuming access to n_w samples from the posterior, $p_{\text{post}}\left(\mathbf{w} \mid \mathbf{d}\right)$, the first moment, also known as the conditional mean, can be approximated from these samples, $\left\{\mathbf{w}_j\right\}_{j=1}^{n_{\text{w}}} \sim$

$p_{\text{post}}(\mathbf{w} \mid \mathbf{d})$, via

$$\begin{aligned}
\delta \mathbf{m}_{\text{CM}} &= \mathbb{E}_{\delta \mathbf{m} \sim p_{\text{post}}(\delta \mathbf{m} \mid \mathbf{d})} \big[\delta \mathbf{m} \big] \\
&= \mathbb{E}_{\mathbf{w} \sim p_{\text{post}}(\delta \mathbf{w} \mid \mathbf{d})} \big[g(\mathbf{z}, \mathbf{w}) \big] \\
&= \int p_{\text{post}}(\mathbf{w} \mid \mathbf{d}) g(\mathbf{z}, \mathbf{w}) \mathrm{d}\mathbf{w} \\
&\approx \frac{1}{n_{\text{w}}} \sum_{j=1}^{n_{\text{w}}} g(\mathbf{z}, \mathbf{w}_j).
\end{aligned} \tag{2.12}$$

We describe the important step of obtaining these samples from the posterior below.

Compared to the MAP estimate, the conditional mean, which corresponds to the minimum-variance estimate [90], is less prone to overfitting [91]. This was confirmed empirically for seismic imaging [60, 61]. In the experimental sections below, we will provide further evidence of advantages the conditional mean offers compared to MAP estimation.

Point-wise standard deviation estimation

In its most rudimentary form, uncertainties in the imaging step can be assessed by computing the pointwise standard deviation, which expresses the spread among the different unknown models explaining the observed data. Given samples from the posterior, this quantity can be computed via

$$\begin{aligned}
\boldsymbol{\sigma}_{\text{post}}^2 &= \mathbb{E}_{\delta \mathbf{m} \sim p_{\text{post}}(\delta \mathbf{m} \mid \mathbf{d})} \big[(\delta \mathbf{m} - \delta \mathbf{m}_{\text{CM}}) \odot (\delta \mathbf{m} - \delta \mathbf{m}_{\text{CM}}) \big] \\
&\approx \frac{1}{n_{\text{w}}} \sum_{j=1}^{n_{\text{w}}} \big(g(\mathbf{z}, \mathbf{w}_j) - \delta \mathbf{m}_{\text{CM}} \big) \odot \big(g(\mathbf{z}, \mathbf{w}_j) - \delta \mathbf{m}_{\text{CM}} \big).
\end{aligned} \tag{2.13}$$

In this expression, $\boldsymbol{\sigma}_{\text{post}}$ is the estimated pointwise standard deviation and $\odot$ represents elementwise multiplication. Again, the expectations approximated in equations 2.12 and 2.13 require samples from the posterior distribution, $p_{\text{post}}\left(\delta \mathbf{m} \mid \mathbf{d}\right)$.

Confidence intervals

As described above, the pointwise standard deviation is a quantity that summarizes the spread among the likely estimates of the unknown. Using this quantity, we can put error bars on the unknown in which case we assign probabilities (confidence) to the unknowns being in a certain interval. The interval is obtained by treating the pointwise posterior distribution as a Gaussian distribution, where the mean and standard deviation at each points are equal to the value of the conditional mean estimate and pointwise standard deviation at that point, respectively. Given a desired confidence value, e.g., 99%, sample mean $\boldsymbol{\mu}$, and sample variance $\boldsymbol{\sigma}^2$, the confidence interval is $\boldsymbol{\mu} \pm 2.576\,\boldsymbol{\sigma}$ where 99% of samples fall between the left ($\boldsymbol{\mu} - 2.576\,\boldsymbol{\sigma}$) and right ($\boldsymbol{\mu} + 2.576\,\boldsymbol{\sigma}$) tails of the Gaussian distribution [92].

2.4 Sampling from the posterior distribution

Extracting statistical information from the posterior distribution, such as the point and interval estimates introduced in the previous section, typically requires access to samples from the posterior distribution. In the following section, we first show that approximations to the point and interval estimates are instances of Monte Carlo integration, given samples from the posterior distribution. Next, we shift our attention to constructive techniques to draw these samples efficiently by introducing preconditioning and crucial strategies to select the stepsize. Finally, we describe an empirical verification of convergence of the Markov chains that we will use to verify our sampling approach.

2.4.1 Monte Carlo sampling

For most applications the posterior PDF is not directly of interest, but we need to evaluate expectations involving the posterior distribution instead. Given samples from the posterior,$\{\mathbf{w}_j\}_{j=1}^{n_{\mathrm{w}}} \sim p_{\text{post}}(\mathbf{w} \mid \mathbf{d})$, these expectations with respect to arbitrary functions can be approximated by

$$\mathbb{E}_{\mathbf{w}\sim p_{\text{post}}(\delta\mathbf{w}|\mathbf{d})}\big[f(\mathbf{w})\big] \approx \frac{1}{n_w}\sum_{j=1}^{n_w} f(\mathbf{w}_j). \tag{2.14}$$

Below we describe our proposed MCMC approach for obtaining samples from the posterior.

2.4.2 Sampling via stochastic gradient Langevin dynamics

Drawing samples from posterior distributions associated with imaging problems of high dimensionality ($M,\ N$ large) and expensive forward operators (e.g., demigration operators) is challenging [67, 7]. Among the different approaches, MCMC is a well-studied technique capable of drawing samples via a sequential random-walk procedure. This process requires evaluation of the posterior PDF at each step. The need for repeated evaluations of the forward operator, the correlation between consecutive samples [93], and the high dimensionality of the problem are the chief computational challenges for these methods. Despite these difficulties, MCMC methods have been applied successfully in imaging problems [94, 9, 95, 11, 12, 60, 61].

Aside from problems related to the required length of the Markov Chains, computing the misfit over all n_s sources in the likelihood term of the posterior PDF (equation 2.7) is problematic since this calls for many evaluations of the linearized Born scattering operator. To address this issue, we use techniques from stochastic optimization [96, 97, 98] where the gradients are evaluated for a single randomly selected source (without replacement) at each iteration. For first-order methods, this technique is known as stochastic gradient descent [SGD; 96] and widely used in the machine learning and wave-based inversion communities [16, 99, 100, 84, 85, 101, 102, 103].

While SGD bring down the computational costs, it is a stochastic optimization algorithm for finding the mode of the posterior distribution and it does not provide samples from the posterior distribution. In order to do that, we have to add a carefully calibrated noise term to the gradients. This additional noise term induces a random walk from which samples from the posterior distribution can be drawn under certain conditions [67]. Adding this noise term also avoids converge of the iterations to the MAP estimate [67]. In this chapter, we adapt an approach known as stochastic gradient Langevin dynamics [SGLD, 67], which is designed to reduce the number of necessary individual likelihood evaluations at each iteration. SGLD was originally developed for Bayesian inference on deep neural networks trained on large-scale datasets. Compared to the original formulation of Langevin dynamics [104], SGLD works on randomly selected subsets of shot data, which makes it computationally more efficient and achievable at least in 2D imaging problems. Asymptotically, SGLD provides accurate samples from the target distribution [67, 105, 106, 107, 108]—in our case, the imaging posterior distribution. It differs from variational inference [109] in that no surrogate distribution is formulated and matched to the distribution of interest.

Following the work of [67], SGLD iterations for the negative log-posterior involve at iteration k the following update for the network weights of the deep prior:

$$\begin{aligned}&\mathbf{w}_{k+1} = \mathbf{w}_k - \frac{\alpha_k}{2} \mathbf{M}_k \nabla_{\mathbf{w}} \left(\frac{n_s}{2\sigma^2} \big\| \mathbf{d}_i - \mathbf{J}(\mathbf{m}_0, \mathbf{q}_i) g(\mathbf{z}, \mathbf{w}_k) \big\|_2^2 + \frac{\lambda^2}{2} \big\| \mathbf{w}_k \big\|_2^2 \right) + \boldsymbol{\eta}_k, \\ &\boldsymbol{\eta}_k \sim \mathrm{N}(\mathbf{0}, \alpha_k \mathbf{M}_k),\end{aligned} \tag{2.15}$$

where the index, $i \subset \{1, \ldots, n_s\}$, is chosen randomly without replacement at each iteration. Once all the shots are drawn, we start all over by redrawing indices, without replacement, from $i \subset \{1, \ldots, n_s\}$. We repeat this process for K steps (see Algorithm 2), where K can be arbitrarily large. To ensure and speedup convergence, the stepsizes α_k and the adaptive preconditioning matrix $\mathbf{M}_k$ need to be chosen carefully. The additional zero-mean Gaussian noise term $\boldsymbol{\eta}_k$ with covariance matrix $\alpha_k \mathbf{M}_k$ distinguishes between the update rule in equation 2.15 and SGD optimization algorithm. It was shown by [67] that the above iterations sample from the posterior after a warmup phase, i.e., a certain number of iterations of equation 2.15. During the warmup stage, these iterations behave similarly to those of the SGD algorithm but at some point transition to the proper sampling phase [67]. Below we will comment when that transition is likely to occur.

Stepsize selection

Convergence of stochastic optimization methods such as SGD and SGLD relies on carefully designed stepsize strategies. Compared to SGD, SGLD has the additional complication of having to balance random errors due to randomly selecting shot records and the deliberate random "errors" induced by the additional Gaussian noise term, $\boldsymbol{\eta}_k$. On the one hand, the iterations in equation 2.15 need to make sufficient progress during the warmup phase so that the samples ($=$ iterations $\mathbf{w}_k$) become independent of the chain's initialization, i.e., the weights $\mathbf{w}_0$ at the start. On the other hand, after warmup the Gaussian noise term, $\boldsymbol{\eta}_k$ will start to dominate the energy of the error in the gradient caused by the stochastic approximation to the likelihood function (equation 2.5). This can be explained by the fact that the variance of error due to the stochastic gradient approximation is proportional to the square of the stepsize [96], whereas the additive noise term is drawn from a Gaussian distribution whose variance is proportional to the stepsize. Consequently, for small stepsizes, it is expected that the error in gradients will be dominated by additive noise [67], which effectively turns equation 2.15 to Langevin dynamics [104]. As a result, similar to SGD, convergence can only be guaranteed when the stepsize in equation 2.15 decreases to zero. However, this would increase the number of iterations to fully explore the posterior probability space. We avoid this situation and follow [67] who propose the following sequence of stepsizes:

$$\alpha_k = a(b+k)^{-\gamma}, \tag{2.16}$$

where $\gamma = \frac{1}{3}$ is the decay rate chosen according to [110]. The constants $a,\ b$ in this expression control the initial and final value of the stepsize. Below, we will comment how to chose these constants and how to ensure that potential posterior sampling errors [107] are avoided.

Preconditioning

In addition to selecting proper stepsizes, the converge of the iterations in equation 2.15 depends on how strongly the different weights of the deep prior are coupled to the data. Without preconditioning, i.e., $\mathbf{M}_k = \mathbf{I}$, SGLD updates all parameters with one and the same stepsize. This leads to slow convergence of the insensitive weights that are weakly coupled to the data. To avoid this situation, [68] proposed an adaptive diagonal preconditioning matrix extending the RMSprop optimization algorithm [84]. This preconditioner is deigned to speed up the initial warmup and subsequent sampling stage of the iterations in equation 2.15. To define this preconditioning matrix, let $\delta \mathbf{w}$ denote the gradient of the negative log-posterior density (equation 2.7) at the current estimate of weights $\mathbf{w}_k$, i.e.,

$$\delta \mathbf{w} = \nabla_{\mathbf{w}} \left(\frac{n_s}{2\sigma^2} \big\| \mathbf{d}_i - \mathbf{J}(\mathbf{m}_0, \mathbf{q}_i) g(\mathbf{z}, \mathbf{w}_k) \big\|_2^2 + \frac{\lambda^2}{2} \big\| \mathbf{w}_k \big\|_2^2 \right). \tag{2.17}$$

Given these gradients, define the following running pointwise sum on the pointwise square of the gradients

$$\mathbf{v}_{k+1} = \beta \mathbf{v}_k + (1 - \beta)\, \delta \mathbf{w} \odot \delta \mathbf{w} \tag{2.18}$$

where the parameter β controls the relative importance of the elementwise square of the gradient compared to the current iterate $\mathbf{v}_k$. The $\mathbf{v}_0$ is initialized as a vector with M zeros. By choosing,

$$\mathbf{M}_k = \text{diag}\left(1 \oslash \sqrt{\mathbf{v}_{k+1}}\right) \tag{2.19}$$

with $\oslash$ elementwise division, the effective stepsize for network weights with large (on average) gradients, i.e., large sensitivities, is lowered whereas weights with small (on average) gradients get updated with a larger effective stepsize. To avoid division by zero, we add a small value to the denominator of equation 2.19. By introducing the preconditioning matrix $\mathbf{M}_k$ all weights are updated similarly, which allows us to increase the stepsize. Following

[68], we set $\beta = 0.99$. In addition to leveling the playing field, for the gradients themselves the preconditioning matrix also scales the essential additive noise term so the random walk proceeds isotropically.

2.4.3 Practical verification

While there exists a well established literature on how to verify whether Markov chains produce accurate samples from the posterior distribution [see 111], these methods are typically impractical for our problem. We will adopt a more pragmatic approach to assess the accuracy of the samples drawn from our Markov chains computed with SGLD, as it will be explained in the following.

We first validate the accuracy of sampling from a single Markov chain by computing confidence intervals. These intervals are computed from posterior samples obtained via one MCMC chain using SGLD (equation 2.15). By definition, these confidence intervals provide the range within which the weights and therefore the image are expected to fall. This means that MAP estimates for the seismic image should ideally fall within these confidence intervals computed from the posterior samples. While the variability among MAP estimates is less than the variability among true posterior samples, we still find this test of practical importance. To verify this, we compute multiple MAP estimates (see equations 2.9 and 2.10) for different independent random initializations of the deep prior weights, $\mathbf{w}$. MAP estimates are obtained via stochastic optimization using the RMSprop optimization algorithm [84], which uses the same preconditioning scheme (equations 2.18 and 2.19) as SGLD. By checking whether the different MAP estimates indeed fall within the computed confidence interval, the accuracy of the samples from the posterior can at least be verified qualitatively.

Ideally, different Markov chains initialized with different weights should lead to similar statistics for samples of the posterior distribution. We verify this empirically by running chains with different independently randomly initialized weights, followed by visual inspection of the conditional mean and pointwise standard deviation derived from samples generated by the different chains. Deviations among the estimates provides us with at least an indication of areas in the image where we should be less confident on the inferred statistics.

2.4.4 The SGLD Algorithm

The different steps of generating $n_w = K/2$ samples from the posterior from K iterations of SGLD (equation 2.15) are summarized in Algorithm 2 for a given set of n_s processed shot records and their respective source signatures, $\{\mathbf{d}_i, \mathbf{q}_i\}_{i=1}^{n_s}$. Aside from shot data, Algorithm 2 requires a smooth background model, $\mathbf{m}_0$, for the squared slowness and a fixed realization for the latent variable $\mathbf{z} \sim \mathrm{N}(\mathbf{0}, \mathbf{I})$. In addition to these input vectors, SGLD requires hyperparameters to be set for the

- **Stepsize strategy.** Following [110], the decay rate parameter in equation 2.16 is set to $\gamma = \frac{1}{3}$. The stepsize constants $a,\ b$ in equation 2.16 are chosen separately for each presented numerical experiment to ensure fast convergence in the warmup phase. Specifically, we select a large enough to ensure fast convergence while making sure the initial iterations do not diverge due to large a stepsize. We selected b to be the same as the stepsize that would yeld a good convergence for the MAP estimation problem, i.e., SGLD iterations without the additive noise. This is to ensure SGLD iterations get close enough to mode(s) of the distribution toward the end. While selecting these parameters differently changes the speed of converge, the accuracy of the resulting samples is empirically verified for the chosen parameters.

- **Preconditioning.** As documented in the literature, we chose $\beta = 0.99$ in equation 2.18.

- **Noise variance.** The variance σ^2 of the noise assumed to be known.

- **Regularization parameter.** As with many inverse problems, the selection of the regularization parameter λ^{-2} is challenging. While sophisticated techniques [82] exist to estimate this parameter, we tune the regularization parameter λ^{-2} by hand to limit the imaging artifacts visually.

- **Number of iterations and warmup.** We run $10\,\mathrm{k}$ SGLD iterations (equation 2.15) in total, and we adopt the general practice of discarding the first half of the obtained samples [111].

Given the above inputs, Algorithm 2 proceeds by running K iterations during which simultaneous shot records, each made of a Gaussian weighted source aggregate, are selected, followed by calculations of the gradient (line 3), calculation of the preconditioner (lines $4-5$), stepsize (line 6), and update of the weights (line 8). After $K/2$ iterations, the updated weight also serve as samples from the posterior [111, 67].

2.5 Validating Bayesian inference

In this section, we validate our approach with synthetic examples. To mimic a realistic imaging scenario where the ground truth is known, we do this on a “quasi”-field data set, made out of noisy synthetic shot data generated from a real migrated image. After demonstrating the benefits of regularization with the deep prior, we compare the MAP estimate with the conditional mean. The latter minimizes the Bayesian risk, i.e., it minimizes in expectation the ℓ_2-norm squared difference between the true image and inverted image, given shot data [90]. We conclude by reviewing the pointwise standard deviation as a measure of uncertainty, which can be reaped from samples drawn from the posterior.

2.5.1 Problem setup

With few exceptions, synthetic models often miss realistic statistics of the spatial distribution of the seismic reflectivity. To avoid working with over simplified seismic images, we

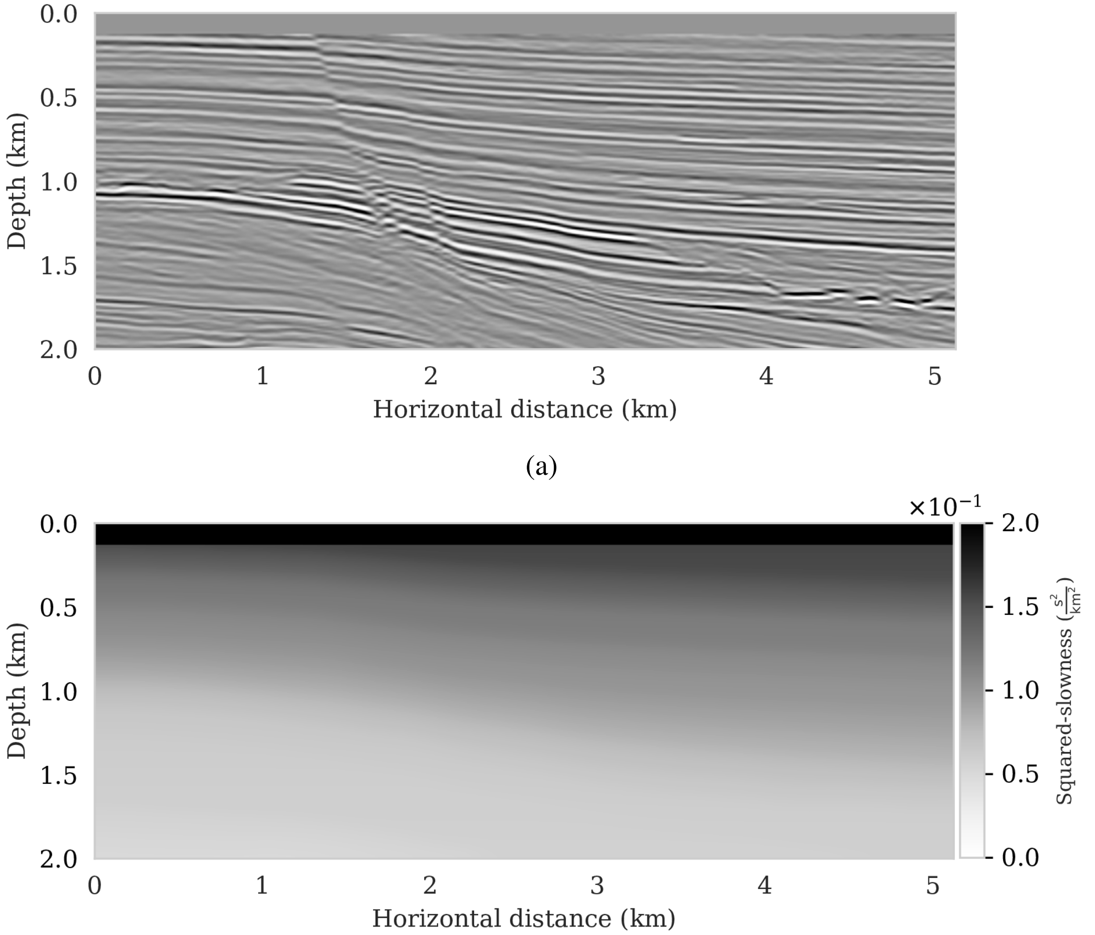
Algorithm 1 Seismic imaging posterior sampling with SGLD.

```
Input:
    $\{\mathbf{d}_i, \mathbf{q}_i\}_{i=1}^{n_s}$          // observed data and source signatures
    $\mathbf{m}_0$          // smooth background squared-slowness model
    $\mathbf{z} \sim \mathrm{N}(\mathbf{0}, \mathbf{I})$          // fixed input to the CNN
    $\lambda^{-2}$          // variance of Gaussian prior on CNN weights
    $\sigma^2$          // estimated noise variance
    $\beta$          // weighting parameter for constructing the preconditioning matrix
    $a, b$          // stepsize parameters in equation 2.16
    $K$          // maximum MCMC steps
Initialization:
    randomly initialize CNN parameters, $\mathbf{w}_0 \in \mathbb{R}^M$          [80]
    initialize vector, $\mathbf{v}_0 \in \mathbb{R}^M$ with zero
for $k = 0$ to $K - 1$ do
    randomly draw $i \subset \{1, \dots, n_s\}$          // sample without replacement
    $\delta \mathbf{w} = \nabla_{\mathbf{w}} \left( \frac{n_s}{2\sigma^2} \left\| \mathbf{d}_i - \mathbf{J}(\mathbf{m}_0, \mathbf{q}_i) g(\mathbf{z}, \mathbf{w}_k) \right\|_2^2 + \frac{\lambda^2}{2} \left\| \mathbf{w}_k \right\|_2^2 \right)$          // equation 2.17
    $\mathbf{v}_{k+1} = \beta \mathbf{v}_k + (1 - \beta)\, \delta \mathbf{w} \odot \delta \mathbf{w}$          // equation 2.18
    $\mathbf{M}_k = \mathrm{diag} \left( 1 \oslash \sqrt{\mathbf{v}_{k+1}} \right)$          // equation 2.19
    $\alpha_k = a(b + k)^{-\gamma}$          // equation 2.16
    $\boldsymbol{\eta}_k \sim \mathrm{N}(\mathbf{0}, \alpha_k \mathbf{M}_k)$          // draw noise to add to gradient
    $\mathbf{w}_{k+1} = \mathbf{w}_k - \frac{\alpha_k}{2} \mathbf{M}_k \delta \mathbf{w} + \boldsymbol{\eta}_k$          // update rule according to equation 2.15
end for
Output: $\{\mathbf{w}_k\}_{k=K/2+1}^{K}$          // samples from the posterior $p_{\text{post}}(\mathbf{w} \mid \mathbf{d})$
```

generate “quasi”-field shot data derived from a 2D subset of the real prestack Kirchhoff time migrated Parihaka-3D dataset [112, 113] released by the New Zealand government. We call our experiment “quasi” real because synthetic data is generated from migrated field data that serves as a proxy for the unknown true medium perturbations (Figure 2.1a). Due to the nature of the migration algorithm used to obtain the Parihaka dataset, the amplitudes in the extracted 2D subset are not necessarily consistent with the seismic imaging forward model presented in this chapter (equation 7.1). For this reason, we normalized the amplitudes of the extracted seismic image. Given these perturbations, shot data is generated with the linearized Born scattering operator for a made up, but realistic, smoothly varying background model $\mathbf{m}_0$ for the squared slowness (Figure 2.1b). To ensure good coverage, 205 shot records are simulated and sampled with a source spacing of $25\,\mathrm{m}$. Each shot is recorded over 1.5 seconds with 410 fixed receivers sampled at $12.5\,\mathrm{m}$ spread across full survey area. The source is a Ricker wavelet with a central frequency of $30\,\mathrm{Hz}$.

We also add a significant amount of band-limited noise to the shot data by filtering Gaussian white noise with the source wavelet. The resulting signal-to-noise ratio for all data is $-8.74\,\mathrm{dB}$, which is low. Figure 5.4 shows an example of a single noise-free (Figure 5.4a) and noisy (Figure 2.2b) shot record.

Even though our example is in 2D, the number of parameters (the weights of the deep prior network) is large (approximately 40 times larger than image dimension), which results in many SGLD iterations. In a setting where we are content with approximate Bayesian inference, i.e., where the validity of the Markov chains can be established qualitatively in the way described earlier, we found that ten thousand iterations are adequate. We adopt the general practice of discarding the first half the MCMC iterations (about 25 passes over the data) [93], which leaves five thousand iterations dedicated to posterior sampling phase [14]. The stepsize sequence is chosen according to equation 2.16 with $a,\ b$ chosen such the stepsize decreases from 10^{-2} to 5×10^{-3}.



(a)

(b)

Figure 2.1: Problem setup. (a) A 2D subset of the Parihaka dataset, considered as true model. (b) Made up smooth squared-slowness background model.

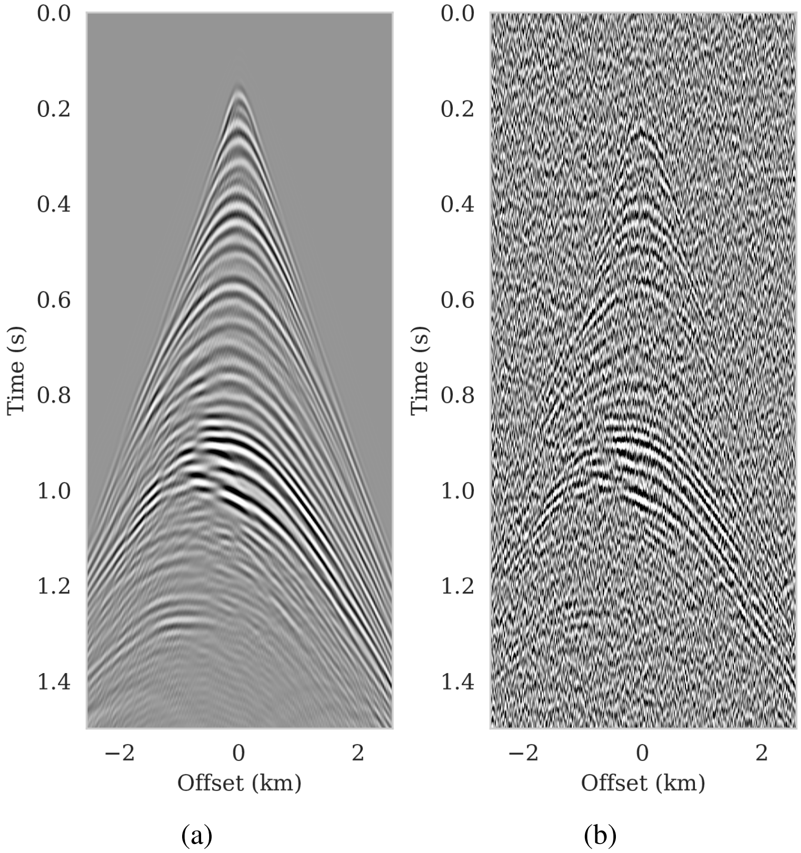


Figure 2.2: A shot record generated from an image extracted from the Parihaka dataset. (a) Noise-free linearized data. (b) Linearized data with band-limited noise.

2.5.2 Imaging with versus without the deep prior

For reference, we first compare imaging results with and without regularization. The latter is based on maximizing the likelihood (equation 2.8) whereas the former involves maximizing the posterior distribution (equation 2.9). To prevent overfitting of the MLE estimate, the number of iterations is limited to the equivalent of only four data passes (four loops over all shots). To ensure convergence, the number of data passes (or epochs) for the MAP estimate was set to 15 (about three thousand iterations). Since the ground truth is known, the optimal value for the $\lambda^{-2} = 5 \times 10^{-3}$ was found by grid search and picking the value that visually limits imaging artifacts. Results of minimizing the negative log-likelihood and negative log-posterior are included in Figures 2.3a and 2.3b, respectively. We obtained these results with the RMSprop optimization algorithm [84], which uses the same preconditioning scheme (equations 2.18 and 2.19) as SGLD that we will use later to conduct Bayesian inference. Compared to vanilla SGD with a fixed stepsize, RMSprop is an adaptive stepsize method conducive to the preconditioner introduced in equations 2.18 and 2.19. As with the SGLD updates, the gradient calculations involve a single randomly selected shot record. As expected, compared to the MAP estimate with signal-to-noise ratio (SNR) 8.79 dB, the MLE estimate (SNR 8.25 dB) lacks important details, e.g., weak reflectors in deeper sections, and exhibits strong artifacts, including imaged reflectors that are noisy and lack continuity. The latter is important since the estimated seismic image will be used to automatically track horizons.

2.5.3 Bayesian inference with deep priors

As the comparison between MLE and MAP estimates clearly showed, regularization improves the image but important issues remain. First, the use of deep priors can lead to overfitting even when a Gaussian prior on the weights is included. As reported in the literature [51, 53], stopping early can be a remedy but a stopping criterion remains elusive rendering this type of regularization less effective. Second, the uncertainty is not captured by the MAP estimation. As we will demonstrate, the ability to draw samples from the posterior remedies these issues.

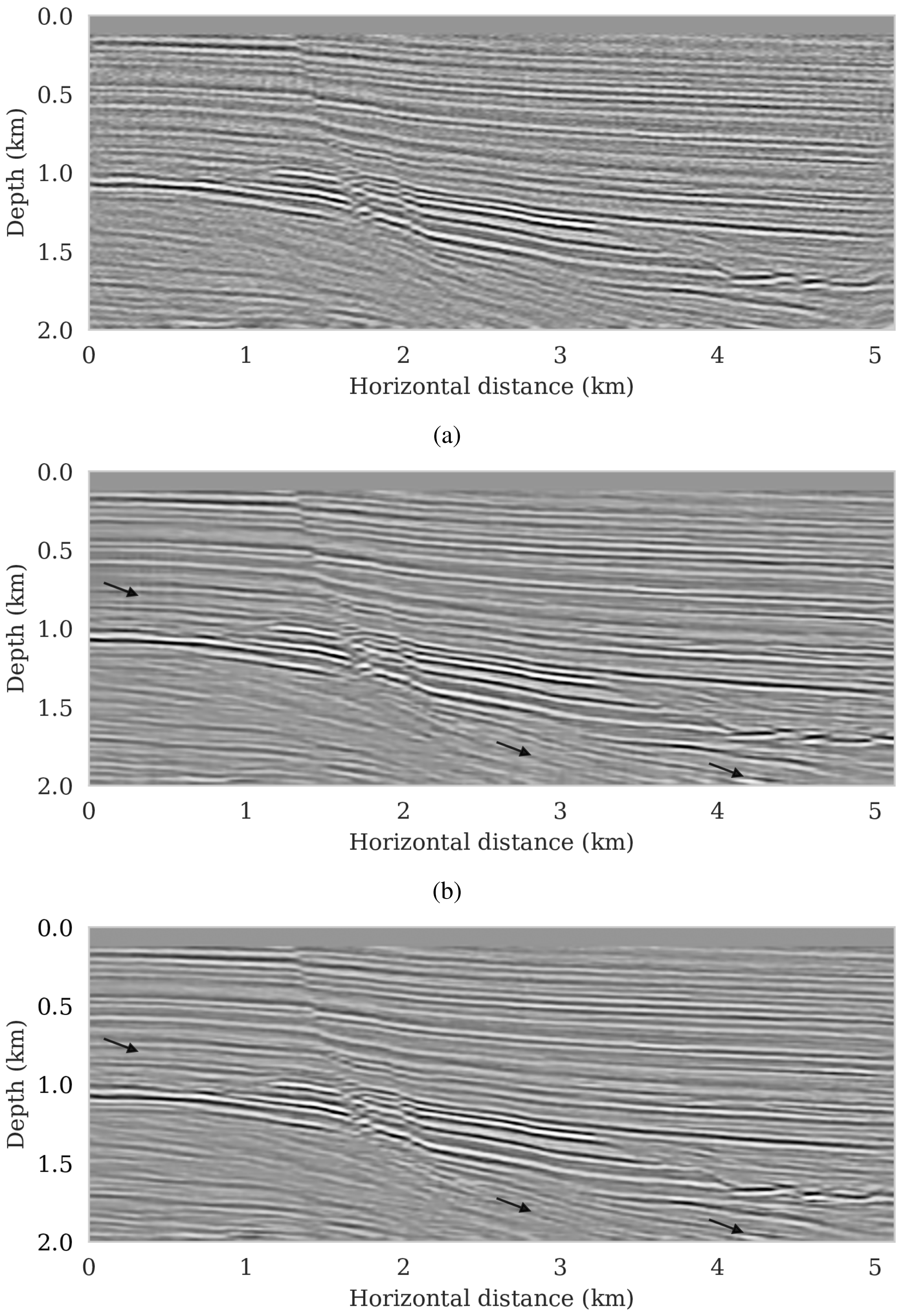
Conditional mean

As described earlier, samples from the posterior provide access to useful statistical information including approximations to moments of the distribution such as the mean. With the minor modifications proposed by [68] to the RMSprop optimization algorithm, the posterior distribution can be sampled with Algorithm 2 after a warmup phase of about 25 data passes. The resulting samples for the weights are then used, after push forward (see equation 2.11), to approximate the conditional mean, $\delta \mathbf{m}_{\text{CM}}$, by computing the sum in equation 2.12. Compared to the MAP estimate (Figures 2.3b and 2.3c), the $\delta \mathbf{m}_{\text{CM}}$ (SNR 9.66 dB) is tantamount to another significant improvement especially for weaker reflectors in the deeper part of the image and for reflectors denoted by the arrows.

While there has been a debate in the literature on the accuracy of the MAP versus conditional mean estimates in the context of regularization with handcrafted priors, such as total variation [114], we find that the conditional mean estimate negates the need to stop early and is also more robust with respect to noise.

Pointwise standard deviation and histograms

To assess variability among the different samples from the posterior, we include a plot of the pointwise standard deviation $\boldsymbol{\sigma}_{\text{post}}$ (equation 2.13) in Figure 2.4a. This quantity is a measure for uncertainty. To avoid bias by strong amplitudes in the estimated image, we also plot the stabilized division of the standard deviation by the envelope of the conditional mean in Figure 2.4b. From these plots in Figure 2.4, we observe that as expected uncertainty is large in areas with a complex geology, e.g., along the faults and along the tortuous reflectors, and in areas with relative poor illumination deep in the image and near the edges. On the other hand, the shallow areas of the image exhibit low uncertainty, which is to be expected due

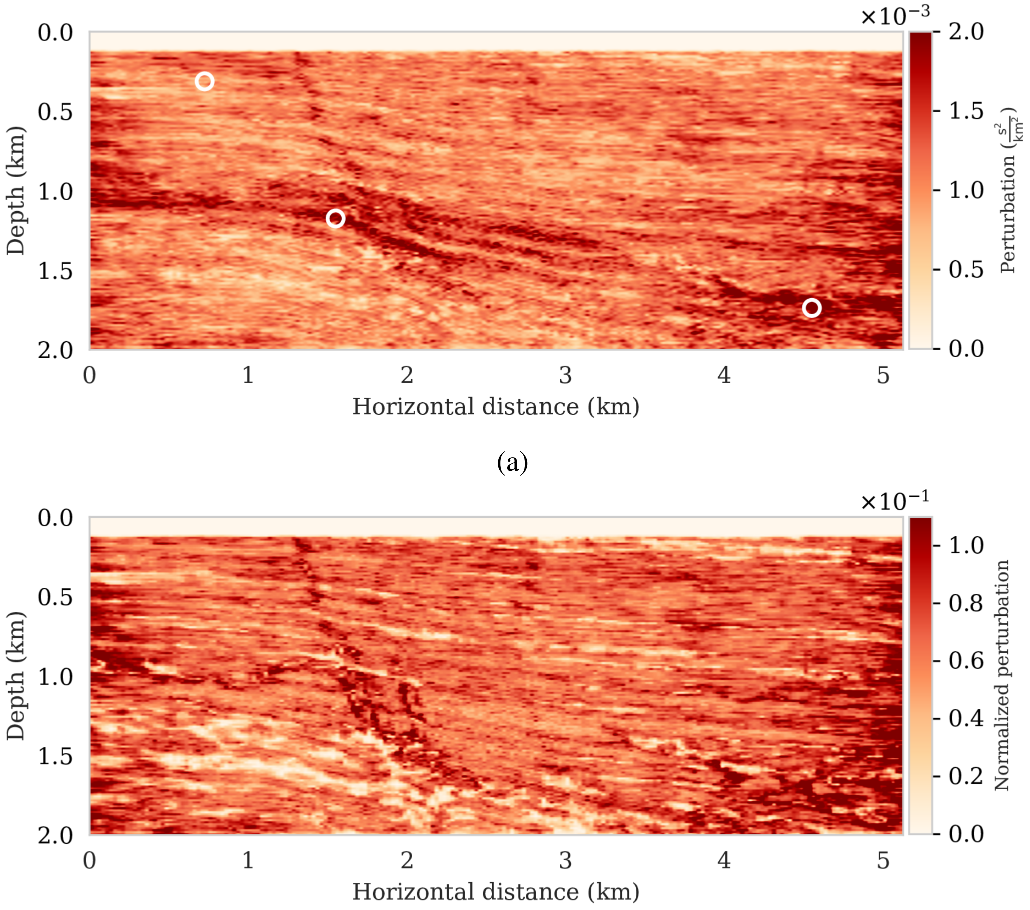


(a)

(b)

(c)

Figure 2.3: Imaging with deep priors of a 2D subset of the Parihaka dataset (a) MLE, i.e., minimizer of equation 2.4 with respect to $\delta \mathbf{m}$, with SNR $8.25\,$dB. (b) The MAP estimate, i.e., minimizer of equation 2.7, followed by a mapping onto the image space via g (equation 2.10), with SNR $8.79\,$dB. (c) The conditional (posterior) mean estimate, $\delta \mathbf{m}_{\text{CM}}$, with SNR $9.66\,$dB. All figures are displayed with the same color clipping values.



(a)

(b)

Figure 2.4: Imaging uncertainty quantification for the Parihaka example. (a) The pointwise standard deviation among samples drawn from the posterior, $\boldsymbol{\sigma}_{\text{post}}$. (b) Normalized pointwise standard deviation by the conditional mean estimate (Figure 2.3c).

to proximity to the sources and receivers.

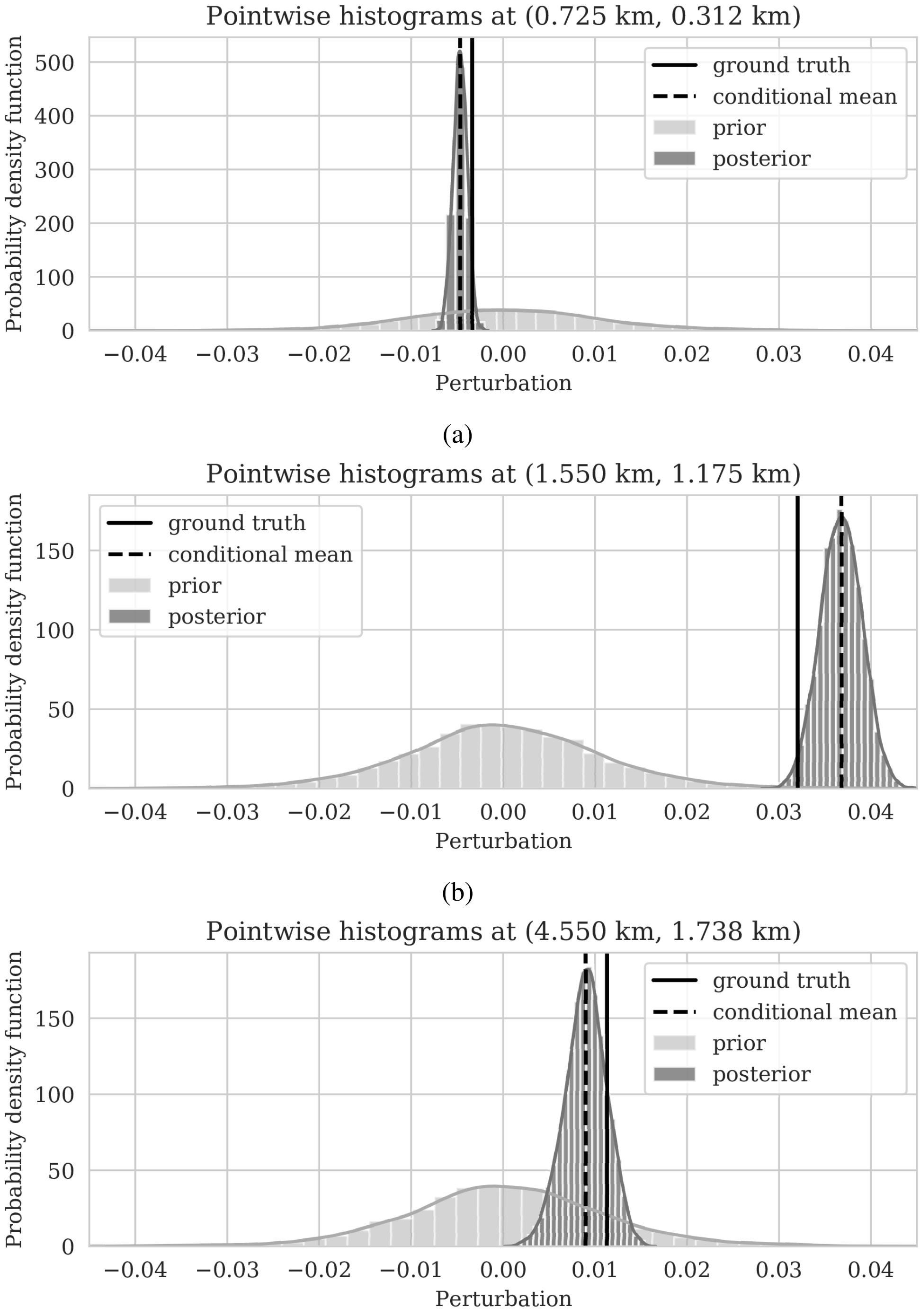
To illustrate how the posterior regularized by the deep prior is informed by the likelihood, we also calculated histograms at three locations denoted by the white circles in Figure 2.4a. Histograms from the prior are calculated by randomly sampling network weights from the prior distribution, i.e., $\mathrm{N}(\mathbf{w} \mid \mathbf{0}, \lambda^{-2}\mathbf{I})$, followed by computing the deep prior network's output for a random but fixed $\mathbf{z}$. The resulting histograms are plotted in light gray in Figure 2.5. Similarly, histograms for the posterior are computed (equation 2.11) from samples of the posterior for the weights. These are plotted in dark gray. As expected, the histograms for the posterior are considerably narrower than those of the prior, which means that the posterior is informed by the shot data. We also see that the width of the histograms increases in areas with larger variability. For comparison, we added the conditional mean estimates with dashed vertical line. When compared with the ground truth values denoted by the solid vertical lines, we observe that the ground truth falls inside of the nonzero pointwise posterior interval, which confirms the benefits of the prior.

2.5.4 Accuracy and convergence verification

Drawing samples from the posterior distribution via Markov chains can be subject to errors when the chain is not long enough [93]. Unfortunately, the required length of the chain is often infeasible in practice, certainly when the forward operators is expensive to calculate as is the case with our imaging examples. As explained earlier, we qualitative verify the accuracy of the Bayesian inversion by comparing MAP estimates with confidence intervals [92] and by running different Markov chains [9].

Consistency with empirical confidence intervals

As a first assessment of the accuracy of the MCMC sampling, we computed the relative errors of 15 MAP estimates with respect to the ground truth image obtained for a single fixed $\mathbf{z}$ but different initializations of the network weights. The decay of the relative ℓ_2-norm error for each run over $3\mathrm{k}$ iterations are plotted in Figure 2.6a and show relatively small variations from random realization to random realization. Vertical profiles of the MAP estimates at two lateral positions confirm this behavior. With few exceptions, these different MAP estimates fall well within the shaded 99% confidence intervals plotted in Figures 2.6b and 2.6c. The confidence intervals themselves were derived from samples of the posterior. Except for perhaps the deeper part of the model, we can be confident that the Bayesian inference is reasonable certainly in the light of the nonlinearity of the deep prior itself.

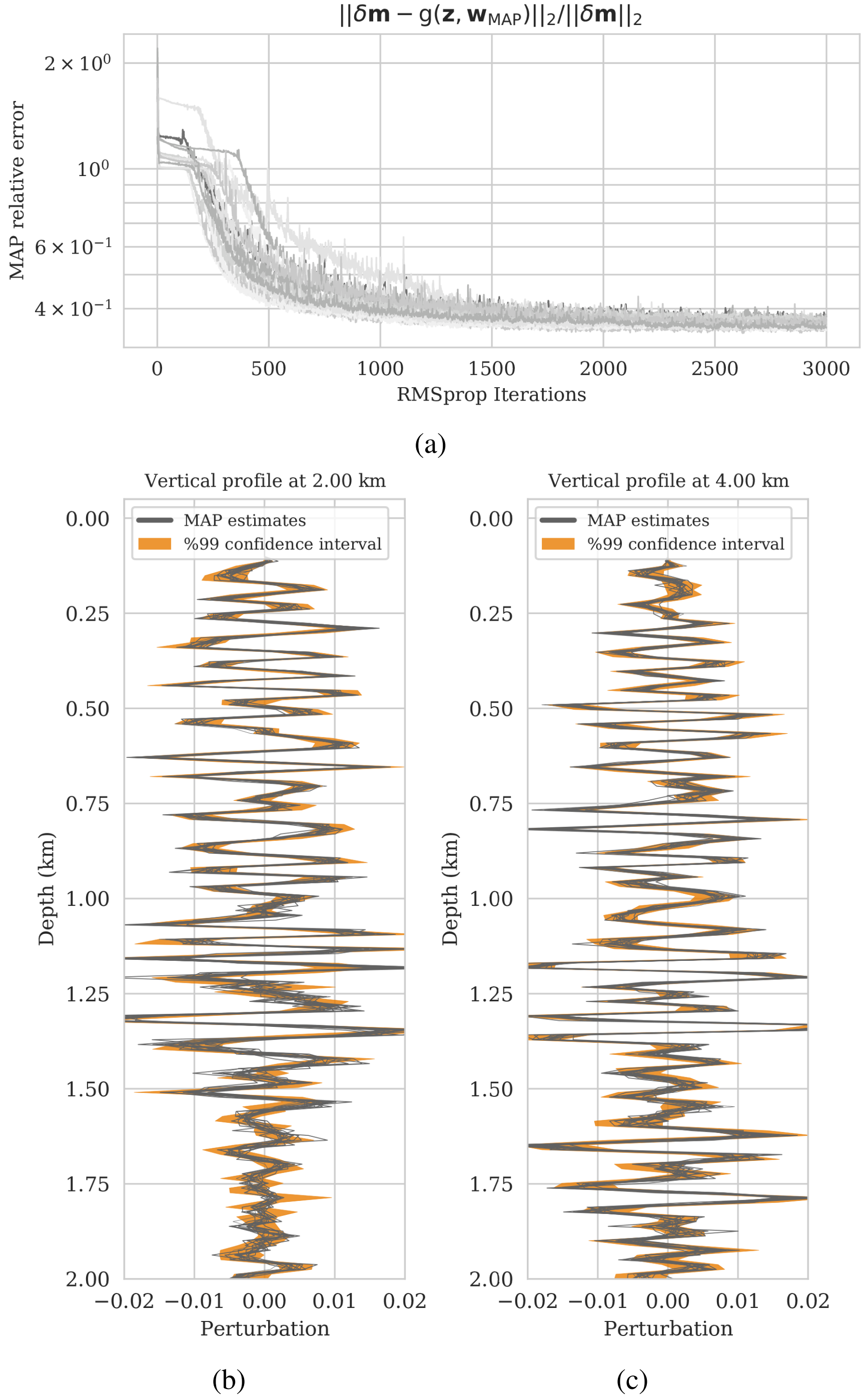


(a)

(b)

(c)

Figure 2.5: Pointwise prior (light gray) and posterior (dark gray) histograms along with the true perturbation values (solid black line) and conditional mean (dashed black line) for points located at (a) $(0.725\,\mathrm{km},\ 0.312\,\mathrm{km})$, (b) $(1.550\,\mathrm{km},\ 1.175\,\mathrm{km})$, and (c) $(4.550\,\mathrm{km},\ 1.738\,\mathrm{km})$.



(a)

(b)

(c)

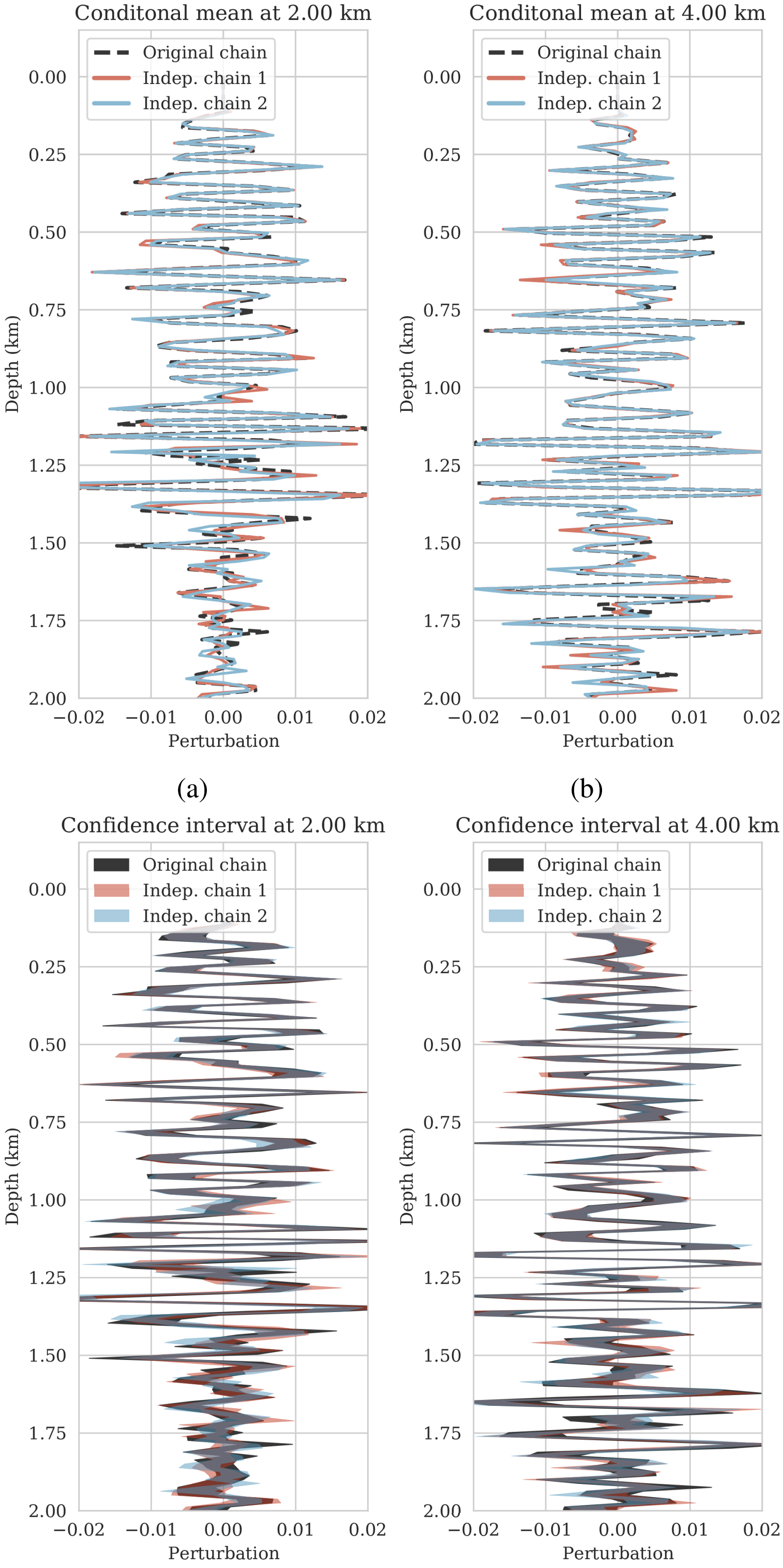
Figure 2.6: Confidence intervals empirical verification. (a) Relative error in the estimated perturbation model for 15 different initialization of the deep prior, with respect to the ground truth image. Traces of 99% confidence interval and 15 realizations of the MAP estimate, $\delta\mathbf{m}_{\mathrm{MAP}}$, at (b) $2.0\,\mathrm{km}$ and (c) $4.0\,\mathrm{km}$ horizontal location.

Chain to chain variations

To further assure our Bayesian inference is accurate, we conducted a second experiment comparing estimates for the conditional mean and confidence intervals for three different Markov chains computed from independent random initialization for the weights and $\mathbf{z}$ fixed. Since we can not afford to run the Markov chains to convergence, we expect slightly different results for the conditional mean and confidence intervals. As observed from Figure 2.7, this is indeed the case but the variations are relatively minor and confined to the deeper part of the image. This qualitative observation, in conjunction with the behavior of the MAP estimates, suggests that we can be confident that the presented Bayesian inference is reasonably accurate certainly given the task of horizon tracking at hand.

2.6 Probabilistic horizon tracking

Typically, seismic images serve as input to a decision process involving identification of certain attributes within the image preferably including an assessment of their uncertainty. With very few exceptions [21], these assessments of risk are not based on a systematic approach where errors in shot data are propagated to uncertainty in the image and subsequent tasks. To illustrate how the proposed Bayesian inference can serve to assess uncertainty on downstream tasks, we consider horizon tracking, where reflector horizons are extracted automatically from seismic images given a limited number of user specified control points. Typically, these control points are either derived from well data available in the area or from human interpretation. The horizon tracking can be deterministic, i.e., horizons are determined uniquely given a seismic image, or more realistically, it may be nondeterministic, i.e., multiple possible horizons explaining a single image. In either case, the task of delineating the stratigraphy automatically with no to little intervention by interpreters is challenging certainly in areas where the geology is complex, e.g., near faults. To resolve these complex areas high quality images including information on uncertainty are essential.



(a)

(b)

(c)

(d)

Figure 2.7: MCMC convergence diagnosis. Conditional posterior mean with three independent Markov chains at (a) $2.0\,\mathrm{km}$ and (b) $4.0\,\mathrm{km}$. Confidence intervals at (c) $2.0\,\mathrm{km}$ and (d) $4.0\,\mathrm{km}$.

To set the stage to put tasks conducted on seismic images on a firm statistical footing, we will make the assumption that these tasks are only informed by the estimated image and not by the shot data explicitly. This means when given an estimated image the task, e.g., of horizon tracking, is assumed to be statistically independent of the shot data. Formally, this conditional independence can be expressed as

$$(\mathbf{h} \perp\!\!\!\perp \mathbf{d}) \mid \delta\mathbf{m}, \tag{2.20}$$

where the random variable $\mathbf{h}$ encodes tracked horizons. The symbol $\perp\!\!\!\perp$ represents conditional independence [115] in this case between the estimated horizons and shot records, given the seismic image, i.e., the seismic images, $\delta\mathbf{m}$, obtained from the shot records, $\mathbf{d}$, contain all the needed information to predict horizons, $\mathbf{h}$. The assumed statistical independence implies that the tracked horizons, $\mathbf{h}$, can be predicted unequivocally from estimated images. Because of the independence, shot data does not bring forth additional information on the horizons. For the remainder of this chapter, we denote the task on the image by $\mathcal{H}$, which for horizon tracking implies $\mathbf{h} = \mathcal{H}(\delta\mathbf{m})$.

2.6.1 Bayesian formulation

Given the mapping from image to horizons, let $p_{\mathcal{H}}\left(\mathbf{h} \mid \delta\mathbf{m}\right)$ represent the conditional PDF of horizons given an estimate for the seismic image. This distribution is characterized by the nondeterministic behavior of $\mathcal{H}$ and assigns probabilities to horizons in the image, $\delta\mathbf{m}$. In the case where $\mathcal{H}$ represents automatic horizon tracking this mapping requires control points as an additional input. In this context, sampling from the conditional distribution is equivalent to performing automatic horizon tracking with different realizations of the control points. Alternatively, when $\mathcal{H}$ represents actions by human interpreters, samples from $p_{\mathcal{H}}\left(\mathbf{h} \mid \delta\mathbf{m}\right)$ can be thought of as horizons tracked by different individuals.

Provided samples from $p_{\mathcal{H}}\left(\mathbf{h} \mid \delta\mathbf{m}\right)$ and assuming conditional independence between

$\mathbf{h}$ and $\mathbf{d}$ given the seismic image described in equation 2.20, we can perform Bayesian inference with the posterior distribution of horizons, denoted by $p_{\text{post}}\left(\mathbf{h} \mid \mathbf{d}\right)$. Generally speaking, for any arbitrary function of horizons, f, expectations with respect to $p_{\text{post}}\left(\mathbf{h} \mid \mathbf{d}\right)$ can be computed as follows:

$$\begin{aligned}
\mathbb{E}_{\mathbf{h} \sim p_{\text{post}}(\mathbf{h} \mid \mathbf{d})}\left[f\left(\mathbf{h}\right)\right] &= \int f\left(\mathbf{h}\right) p_{\text{post}}\left(\mathbf{h} \mid \mathbf{d}\right) \mathrm{d}\mathbf{h} = \iint f\left(\mathbf{h}\right) p\left(\mathbf{h}, \delta\mathbf{m} \mid \mathbf{d}\right) \mathrm{d}\mathbf{h}\, \mathrm{d}\delta\mathbf{m} \\
&= \iint f\left(\mathbf{h}\right) p_{\mathcal{H}}\left(\mathbf{h} \mid \delta\mathbf{m}, \mathbf{d}\right) p_{\text{post}}\left(\delta\mathbf{m} \mid \mathbf{d}\right) \mathrm{d}\mathbf{h}\, \mathrm{d}\delta\mathbf{m} \\
&= \iint f\left(\mathbf{h}\right) p_{\mathcal{H}}\left(\mathbf{h} \mid \delta\mathbf{m}\right) p_{\text{post}}\left(\delta\mathbf{m} \mid \mathbf{d}\right) \mathrm{d}\mathbf{h}\, \mathrm{d}\delta\mathbf{m} \\
&= \mathbb{E}_{\delta\mathbf{m} \sim p_{\text{post}}(\delta\mathbf{m} \mid \mathbf{d})}\left[\int f\left(\mathbf{h}\right) p_{\mathcal{H}}\left(\mathbf{h} \mid \delta\mathbf{m}\right) \mathrm{d}\mathbf{h}\right] \\
&= \underbrace{\mathbb{E}_{\delta\mathbf{m} \sim p_{\text{post}}(\delta\mathbf{m} \mid \mathbf{d})} \underbrace{\mathbb{E}_{\mathbf{h} \sim p_{\mathcal{H}}(\mathbf{h} \mid \delta\mathbf{m})}\left[f\left(\mathbf{h}\right)\right].}_{\substack{\text{uncertainty in} \\ \text{horizon tracking}}}}_{\text{uncertainty in seismic imaging}}
\end{aligned} \tag{2.21}$$

The second equality in the first line of equation 2.21 follows from the law of total probability[2], the second line is obtained by applying the chain rule of PDFs[3] to the joint density $p\left(\mathbf{h}, \delta\mathbf{m} \mid \mathbf{d}\right)$, and the third line exploits the conditional independence assumption in equation 2.20. Conceptually, equation 2.21 states that we can decompose the uncertainty in horizon tracking into two parts, namely uncertainty in imaging and uncertainty in the horizon tracking task itself. Based on equation 2.21, expectations over $p_{\text{post}}\left(\mathbf{h} \mid \mathbf{d}\right)$ can be calculated via Monte Carlo integration using samples from $p_{\text{post}}\left(\mathbf{h} \mid \mathbf{d}\right)$. Thanks to the conditional independence assumption in equation 2.20, we can sample from $p_{\text{post}}\left(\mathbf{h} \mid \mathbf{d}\right)$ by sampling the imaging posterior, $p_{\text{post}}\left(\delta\mathbf{m} \mid \mathbf{d}\right)$, followed by tracking the horizons in each seismic image. This step yields an ensemble of possible horizons for each sampled image. Using samples drawn from $p_{\text{post}}\left(\mathbf{h} \mid \mathbf{d}\right)$, we approximate the expectation in equation 2.21 by the sample mean. In the following sections, we break equation 2.21 down into two

[2] $p(x) = \int_{\mathcal{Y}} p(x, y)\, \mathrm{d}y$, where $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ are two arbitrary random variables.

[3] $p(x, y) = p(x \mid y)\, p(y),\ \forall\, x \in \mathcal{X},\ y \in \mathcal{Y}$.

cases where the horizon tracker yields an unique set of horizons or multiple sets of likely horizons, given one seismic image.

Case 1*: horizons are unique given an image*

In the simplest case, where horizon tracking uniquely determines the horizons given a seismic image, the conditional PDF $p_{\mathcal{H}}\left(\mathbf{h} \mid \delta\mathbf{m}\right)$ corresponds to a delta function, i.e., we have

$$p_{\mathcal{H}}(\mathbf{h} \mid \delta\mathbf{m}) = \delta_{[\mathbf{h}=\mathcal{H}(\delta\mathbf{m})]}\left(\mathbf{h}\right), \tag{2.22}$$

where $\mathcal{H}$ represent the deterministic horizon tracking map and $\delta(\cdot)$ stands for the delta Dirac distribution. Substituting equation 2.22 into equation 2.21 yields

$$\begin{aligned} \mathbb{E}_{\mathbf{h}\sim p_{\text{post}}(\mathbf{h}|\mathbf{d})}\ \left[f\left(\mathbf{h}\right)\right] &= \mathbb{E}_{\delta\mathbf{m}\sim p_{\text{post}}(\delta\mathbf{m}|\mathbf{d})}\left[\int f\left(\mathbf{h}\right)\delta_{[\mathbf{h}=\mathcal{H}(\delta\mathbf{m})]}\left(\mathbf{h}\right)\mathrm{d}\mathbf{h}\right], \\ &= \mathbb{E}_{\delta\mathbf{m}\sim p_{\text{post}}(\delta\mathbf{m}|\mathbf{d})}\left[f\left(\mathcal{H}\left(\delta\mathbf{m}\right)\right)\right], \\ &\approx \frac{1}{n_{\text{w}}}\sum_{j=1}^{n_{\text{w}}} f\left(\mathcal{H}(\delta\mathbf{m}_j)\right), \end{aligned} \tag{2.23}$$

where $\{\delta\mathbf{m}_j\}_{j=1}^{n_{\text{w}}} \sim p_{\text{post}}(\delta\mathbf{m} \mid \mathbf{d})$ are n_{w} samples from the posterior distribution. Equation 2.23 essentially means that in case of a deterministic horizon tracker uncertainty in imaging can be translated to uncertainty in horizon tracking by simply drawing samples from the seismic imaging posterior and tracking horizons in each image. This procedure results in samples from the posterior distribution of horizons and inference of this posterior distribution is done via the equation above.

Case 2*: multiple likely horizons given an image*

The probabilistic horizon tracking approach, described in equation 2.21, also admits nondeterministic horizon trackers, e.g., automatic horizon tracking with uncertain control points, e.g., points provided by multiple human interpreters. In this case, instead of having one set of horizons for each seismic image, we have multiple realizations of horizons that each agree with a seismic image, i.e., they are samples from $p_{\mathcal{H}}(\mathbf{h} \mid \delta\mathbf{m})$. With these samples, the inner expectation in equation 2.21 can be estimated. Assuming that for each image we have n_h different realizations of tracked horizons, namely, $h_k^{(\delta\mathbf{m})} \sim p_{\mathcal{H}}(\mathbf{h} \mid \delta\mathbf{m}),\ k = 1, \cdots, n_h$, equation 2.21 becomes

$$
\begin{aligned}
\mathbb{E}_{\mathbf{h} \sim p_{\text{post}}(\mathbf{h} | \mathbf{d})} \left[f\left(\mathbf{h}\right) \right] &= \mathbb{E}_{\delta\mathbf{m} \sim p_{\text{post}}(\delta\mathbf{m} | \mathbf{d})} \mathbb{E}_{\mathbf{h} \sim p_{\mathcal{H}}(\mathbf{h} | \delta\mathbf{m})} \left[f\left(\mathbf{h}\right) \right] \\
&\approx \mathbb{E}_{\delta\mathbf{m} \sim p_{\text{post}}(\delta\mathbf{m} | \mathbf{d})} \left[\frac{1}{n_h} \sum_{k=1}^{n_h} f\left(h_k^{(\delta\mathbf{m})} \right) \right], \\
&\approx \frac{1}{n_h n_{\text{w}}} \sum_{j=1}^{n_{\text{w}}} \sum_{k=1}^{n_h} f\left(h_k^{(\delta\mathbf{m}_j)} \right)
\end{aligned}
\tag{2.24}
$$

where $h_k^{(\delta\mathbf{m}_j)}$ is the kth sample from $p_{\mathcal{H}}(\mathbf{h} \mid \delta\mathbf{m}_j)$ and $\delta\mathbf{m}_j$ is the jth sample from $p_{\text{post}}\left(\delta\mathbf{m} \mid \mathbf{d}\right)$. Because it has an extra sum over the different realizations of tracked horizons for a fixed image, equation 2.24 differs from equation 2.23. In the ensuing sections, we show how equations 2.23 and 2.24 can be used to calculate pointwise estimates for the first two moments of the posterior distribution over horizons.

Uncertainty quantification in horizon tracking

It is often beneficial to express uncertainty in the form of confidence intervals. For this purpose, equation 2.23 or 2.24 is evaluated first for $f(\mathbf{h}) = \mathbf{h}$. This yields the conditional mean estimate for the horizons denoted by

$$
\boldsymbol{\mu}_{\mathbf{h}} = \mathbb{E}_{\mathbf{h} \sim p_{\text{post}}(\mathbf{h} | \mathbf{d})} [\mathbf{h}]. \tag{2.25}
$$

Similarly, the pointwise standard deviation of the horizons can be computed by choosing $f(\mathbf{h}) = (\mathbf{h} - \boldsymbol{\mu}_{\mathbf{h}}) \odot (\mathbf{h} - \boldsymbol{\mu}_{\mathbf{h}})$. This latter point estimate can be calculated via

$$\boldsymbol{\sigma}_{\mathbf{h}}^2 = \mathbb{E}_{\mathbf{h} \sim p_{\text{post}}(\mathbf{h}|\mathbf{d})}[(\mathbf{h} - \boldsymbol{\mu}_{\mathbf{h}}) \odot (\mathbf{h} - \boldsymbol{\mu}_{\mathbf{h}})], \tag{2.26}$$

where $\boldsymbol{\sigma}_{\mathbf{h}}$ denotes the pointwise standard deviation. The 99% confidence interval for horizons is the interval $\boldsymbol{\mu}_{\mathbf{h}} \pm 2.576\, \boldsymbol{\sigma}_{\mathbf{h}}$. Contrary to most existing automatic horizon trackers, the uncertainty estimates we provide here are determined by uncertainties in the image due to noise, and possibly linearization errors, in the shot records.

2.7 Probabilistic horizon tracking—“ideal low-noise” Parihaka example

The main goal of this chapter is to derive a systematic approach to propagate uncertainty in imaging to the task at hand. For this purpose, we first consider the relatively ideal case of uncertainty due to additive random noise in the shot data. To illustrate how the presence of this noise affects the task of horizon tracking, we apply the proposed probabilistic framework to the Parihaka imaging example discussed earlier. Because the seismic shot data for this example is relatively low-frequency ($30\,$Hz source peak frequency) and the geology relatively simple, horizons are not that challenging to track. However, there is a substantial amount of noise in the shot data that we need to contend with when tracking the imaged horizons. For the latter task, we deploy the tracking approach introduced by [69], which requires the user to provide control points on the seismic horizons of interest.

To setup this tasked imaging experiment, we select 25 horizons from the conditional mean estimate (Figure 2.3c) calculated for the Parihaka seismic imaging example. Next, control points are picked for the selected horizons at various horizontal positions, separated by $1\,\mathrm{km}$. We group the control points with the same horizontal location, yielding five sets of control points. Figure 2.8 shows these five sets located at lateral positions 0.5, 1.5, 2.5, 3.5, and $4.5\,\mathrm{km}$.

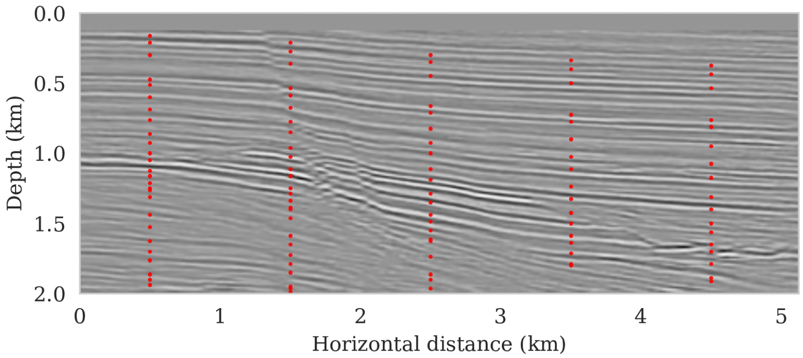


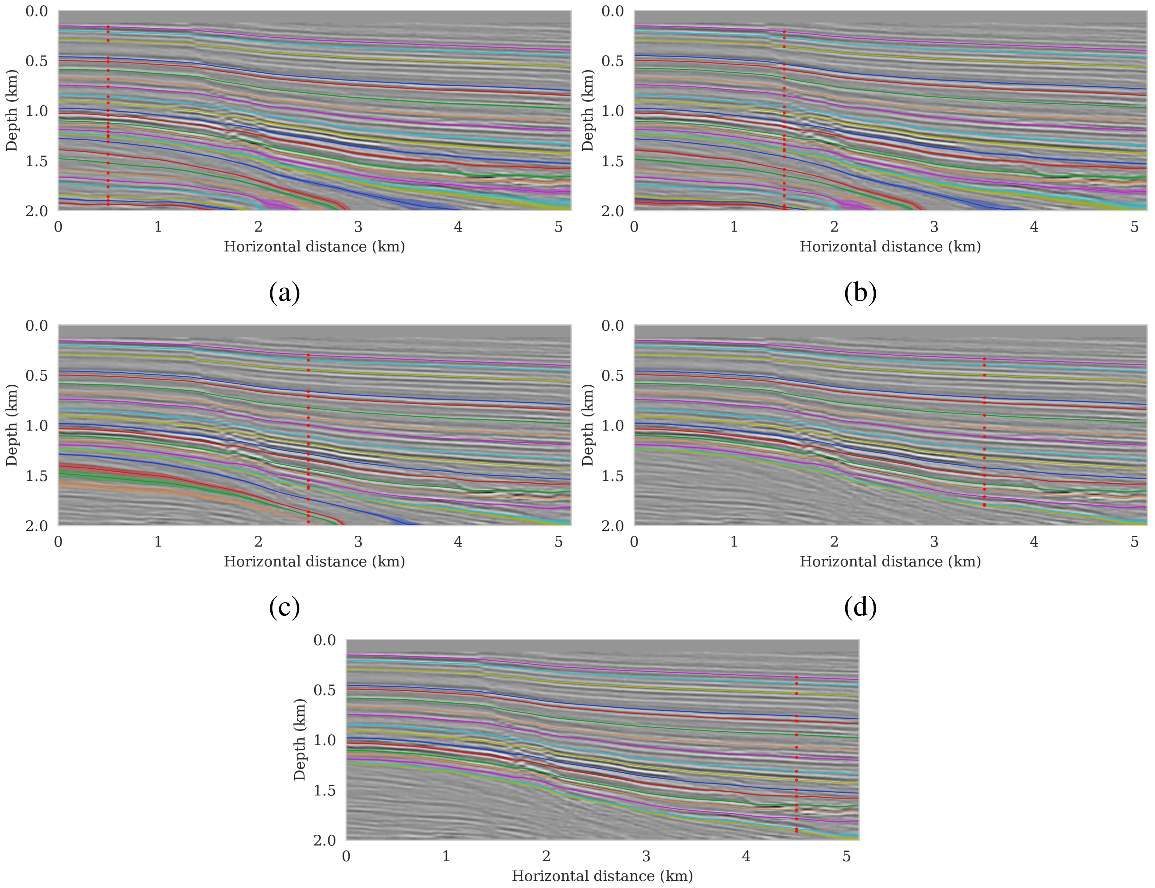
Figure 2.8: Five sets of control points identifying 25 horizons of interest.

To separate the effect of errors in the shot data and variations amongst provided control points, we consider noisy shot data first, followed by the situation where there is uncertainty due to noise in the shot data and due to variations in the control points.

2.7.1 Uncertainty due to noise in shot data (case 1)

To calculate noise-induced uncertainties in horizon tracking (case 1), we pass samples from the imaging posterior distribution, $p_{\text{post}}\left(\delta \mathbf{m} \mid \mathbf{d}\right)$, to the automatic horizon tracking software [69]. Given the five sets of selected control points (Figure 2.8), the tracker generates for each sample of the imaging posterior 25 horizons according to equation 2.23. For each set of control points, the conditional mean and 99% confidence intervals are calculated included in Figure 2.9. Each plot (Figures 2.9a – 2.9e) corresponds to tracked horizons with confidence intervals derived from different sets of control points. As expected, the results exhibit more uncertainty for horizons tracked in the deeper parts of the image and close to boundaries, which is consistent with the relative poor illumination in these areas. Moreover, uncertainty in the tracked locations increases away from the control points. This increase in uncertainty agrees with the inherent challenge of automatic horizon tracking across areas of poor illumination, faults and tortuous reflectors. This behavior is observed for each set of control points.

Aside from shot noise induced uncertainty, variations in the control point may also contribute to uncertainty in the horizon tracking. This corresponds to case 2, which we



(a)

(b)

(c)

(d)

(e)

Figure 2.9: Uncertainty in horizon tracking only due noise in the shot data (equation 2.23). Control points are located at (a) $500\,\mathrm{m}$, (b) $1500\,\mathrm{m}$, (c) $2500\,\mathrm{m}$, (d) $3500\,\mathrm{m}$, and (e) $4500\,\mathrm{m}$ horizontal location. Conditional mean estimates and 99% confidence intervals are shown in solid and shaded colors, respectively.

consider in the next section.

2.7.2 Uncertainty due to noise and uncertain control points (case 2)

The presence of noise in shot data is often not the only cause of uncertainty within the task of (automatic) horizon tracking. Human errors, or better variations in the selection of control points by interpreters, may also contribute to uncertainty. To mimic differences in selected control points, we impose a distribution over the control points. For simplicity, in this example we assume that the five sets of control points are equally likely to be accurate. This is to say, we are equally certain of the accuracy of the picked control points. This can be related to the case where we have access to wells in the seismic survey area and we are certain of the well tying procedure. Other sources of error such as uncertainty in location of the control points, multiple human interpreters, etc, can also be incorporated in the equation 2.24, but will not be considered here. Given the above assumption on the probability distribution, each realization of the seismic image gives rise to multiple equally likely tracked horizons for each of the five sets of control points.

Given these multiple tracked horizons for each image, pointwise first and second moments of $p_{\mathcal{H}} \left(\mathbf{h} \mid \delta \mathbf{m} \right)$ are calculated via equations 2.25 and 2.26 by tracking horizons in each sample from the imaging posterior. The horizon tracker yields five sets of horizons for each seismic image, each obtained using one of the five sets of control points. Results of this procedure are summarized in Figure 2.10. As in the earlier examples, we observe an increase in uncertainty with depth and at the boundaries. Contrary to small uncertainties near the control points, we now observe uncertainty everywhere along the tracked horizons, which suggests increased variability amongst horizons. The variability comes from not trusting only one set of control points, but incorporating information from all the fives sets of control points.

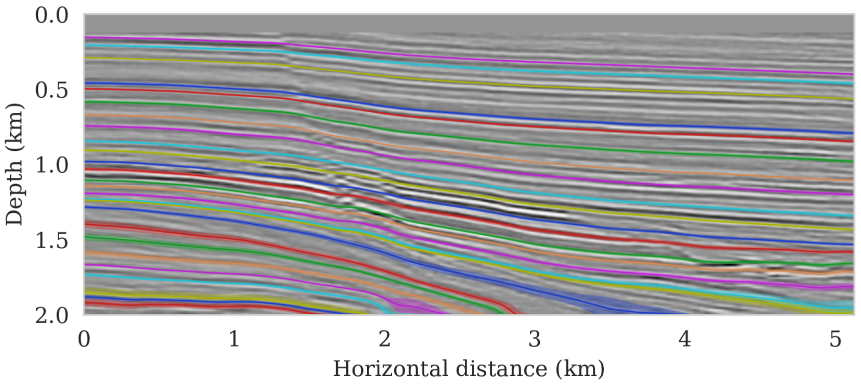


Figure 2.10: Uncertainty in horizon tracking due to a combination of uncertainties in imaging and control points (equation 2.24).

2.8 Probabilistic horizon tracking—“noisy” high-resolution Compass model example

While the seismic imaging examples considered so far were directly obtained from migrated shot data of the Parihaka dataset, the linearization errors, i.e., errors due to linearizing the wave equation with respect to the background squared-slowness model, were ignored because the quasi-real data was generated through the demigration process for a made-up background model. Aside from this simplification, the geology of the examples discussed so far was relatively simple and imaged at low resolution. To account for a more realistic setting, involving complex geology and high-resolution imaging, we will quantify imaging and horizon tracking uncertainty given high-frequency but noisy synthetic shot data. To mimic the complexities of field data, noisy shot data—generated with nonlinear forward modeling on a 2D subset of the Compass model [116]—is used as input to the proposed imaging scheme. We select the synthetic Compass model because it contains realistic heterogeneity derived from both seismic and well data collected in the North Sea [116]. Aside from a low signal-to-noise ratio of $-9.17\,\mathrm{dB}$, this example is affected by linearization errors.

As before, we first describe the problem setup, followed by a comparison between the unregularized MLE and deep-prior regularized MAP and conditional mean estimates. After presenting results on uncertainty quantification on the image, results of our probabilistic horizon tracking framework will be discussed below.

2.8.1 Problem setup

Split spread raw data, consisting of 101 shot records sampled with a source spacing of $25\,\mathrm{m}$ and a receiver sampling $12.5\,\mathrm{m}$, is generated by solving the acoustic wave equation for the 2D subset of the Compass model. To mimic broadband data, the source is a Ricker wavelet with a central frequency of $40\,\mathrm{Hz}$. Each shot is 1.6 seconds long.

To image the shot data, we first derive a kinematically correct background model via smoothing. Next, linearized data is created by subtracting shot data simulated in this smooth background model from the shot data simulated in the actual model. This data serves as input to our imaging scheme.

2.8.2 Imaging with uncertainty quantification

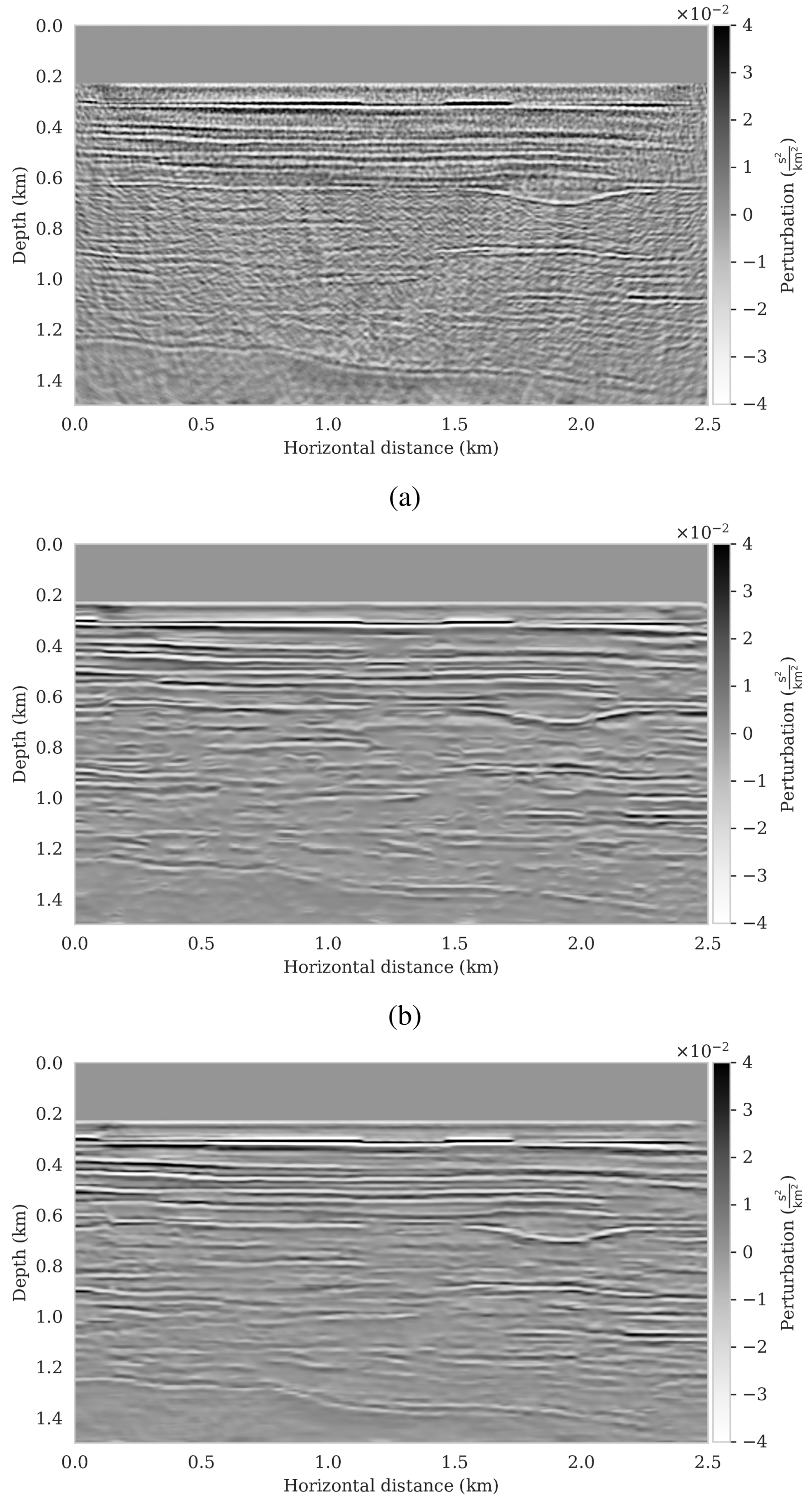
To arrive at the MLE and MAP estimates for the image, we follow the procedure outlined before. The MLE estimate is obtained by minimizing equation 2.5 with stochastic optimization limiting the number of iterations to six passes over the 101 shots. Equation 3.2 is minimized with respect to $\mathbf{w}$ with the RMSprop optimization algorithm [84] for stepsize of 10^{-3} and five thousand iterations (about 50 passes over all shots). We stopped the iterations after no further visual improvement to the image was observed. The value for the tradeoff off parameter, $\lambda^{-2} = 3 \times 10^{-5}$, was set after extensive parameter testing guided by a value that leads to the least amount of visual imaging artifacts.

Compared to the previous example, the MLE image estimate is of poor quality (SNR 2.80 dB) and contains many imaging artifacts stemming from the noise and linearization errors. The MAP estimate, on the other hand, is improved (SNR $3.91\,\mathrm{dB}$) thanks to regularization by the deep prior but it does contain unrealistic artifacts and misses details especially in the deeper parts of the image (juxtapose Figures 2.11a and 2.11b). By running Algorithm 2 for ten thousand iterations, we compute five thousand samples from the posterior distribution on the image following the stepsize strategy of equation 2.16 with $\gamma = \frac{1}{3}$ and $a,\ b$ chosen such the stepsize decreases from 5×10^{-3} to 10^{-3}. This took approximately 18 hours on a quad-core machine. The resulting estimate for the conditional mean, included in Figure 2.11c, represents a considerable visual improvement with a SNR of $4.12\,$dB. Compared to the MLE and MAP estimates, the conditional mean estimate exhibits more continuous reflectors and significantly fewer artifacts. This example confirms that images yielded by the conditional mean of inverse problems regularized by the deep prior are relatively robust to noise.

The available samples from the posterior also allows us to calculate an estimate for the pointwise standard deviation of the image. This quantity is plotted in Figure 2.12a. To avoid overprint by the strong reflectors, we also included a plot of the normalized standard deviation obtained by stabilized division by the conditional mean. Both plots for the pointwise standard deviations show a distinct correlation between difficult to image areas of complex geology, such as channels, and areas affected by relatively poor illumination near the edges and for the deep parts of the image. In the next section, we will show how the samples from the posterior inform the task of horizon tracking.

2.8.3 Horizon tracking with uncertainty quantification

Similar to the previous horizon tracking example 14 horizons are selected from the conditional mean estimate for the image (Figure 2.11c). As before, control points are picked for each horizon at the horizontal locations 62.5, 562.5, 1062.5, 1562.5, and $2062.5\,\mathrm{m}$. Control points with the same horizontal position are grouped together, yielding five sets of control points. As before, we distinguish between uncertainties related to noise and now also linearization errors in the data and uncertainty related to errors in the control parameters and in the shot data due to noise and linearization errors.

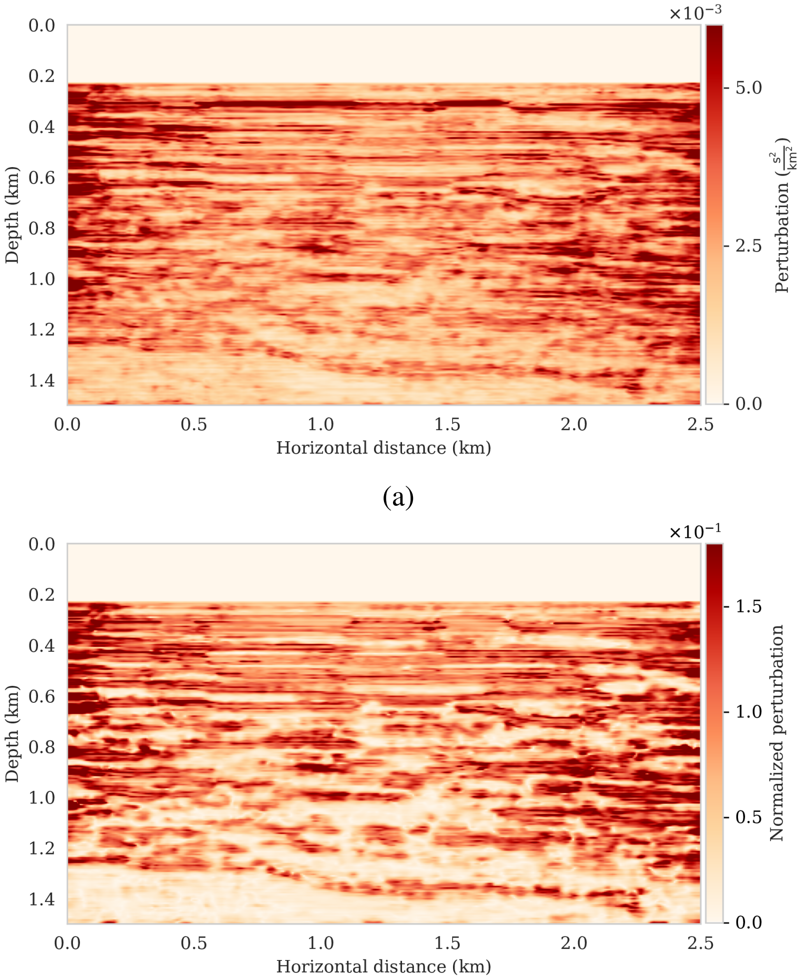


(a)

(b)

(c)

Figure 2.11: Imaging a 2D subset of the Compass model with deep priors. (a) MLE, i.e., minimizer of equation 2.4 with respect to $\delta\mathbf{m}$, with SNR $2.80\,$dB. (b) The MAP estimate, i.e., minimizer of equation 2.7, following by mapping onto the image space via g (equation 2.10), with SNR $3.91\,$dB. (c) The conditional (posterior) mean estimate, $\delta\mathbf{m}_{\text{CM}}$, with SNR $4.12\,$dB.



(a)

(b)

Figure 2.12: Imaging uncertainty quantification corresponding to the Compass model. (a) The pointwise standard deviation among samples drawn from the posterior, $\boldsymbol{\sigma}_{\text{post}}$. (b) Normalized pointwise standard deviation by the conditional mean estimate (Figure 2.11c).

Uncertainty due to noise and linearization errors in the shot data (case 1)

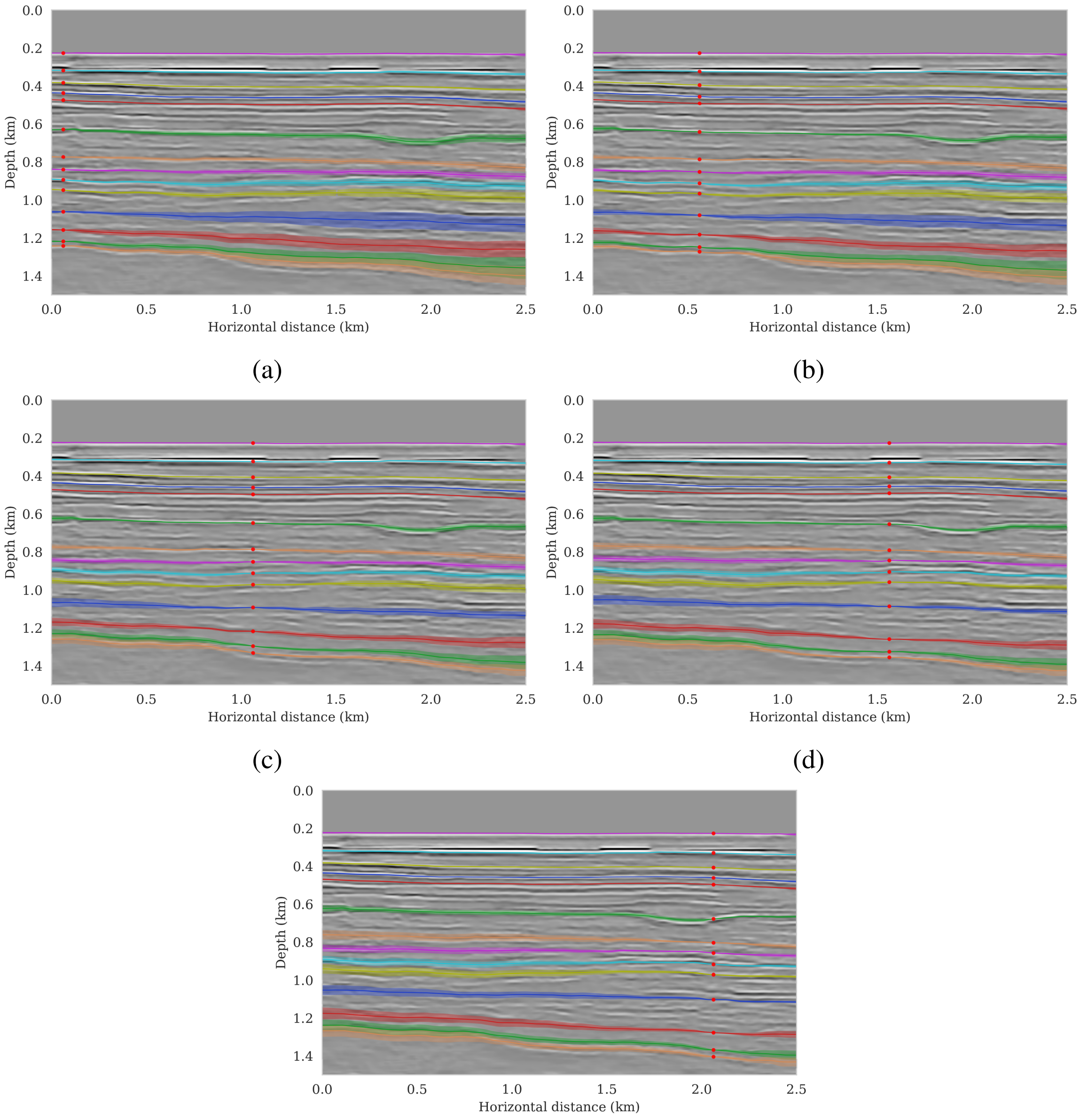
Tracked horizons plus 99% confidence intervals for the five sets of control points are included in Figures 2.13a – 2.13e. These results were computed by sampling the posterior distribution for horizon tracking, $p_{\text{post}}\left(\mathbf{h} \mid \mathbf{d}\right)$, following the procedure described above. Not unexpected, we consistently observe increases in uncertainty as we move further away from the control points and deeper into the image. This increase in the size of the confidence interval is due to the increased variability amongst the samples of the imaging posterior especially in regions that are more difficult to image, e.g., near the boundaries of the image, at the deeper parts and near regions of complex geology.

Uncertainty due to noise, linearization errors, and uncertain control points

Following the procedure described above, we also consider the effect of randomness in the horizon tracking task itself. We mimic this by imposing a distribution over the location of control points. As before, we consider the case where we are equally confident in the location of control points. The results for the conditional mean and the 99% confidence interval are presented in Figure 2.14. As expected, uncertainties increase consistently with depth, close to the boundaries, and in areas of complex geology. Compared to the previous example, these effects are more pronounced, which we argue is due to an increase in linearization errors at later times in the shot data. This increase leads to more uncertainty in deeper sections.

2.9 Discussion

The examples presented in this chapter demonstrate the beneficial regularization properties of deep priors as long as overfitting of noisy data is avoided. Unfortunately, preventing overfit is challenging in practice. To mitigate this issue, we proposed the use of conditional mean estimates rather than maximum a posteriori (MAP) estimation even though the former



(a)

(b)

(c)

(d)

(e)

Figure 2.13: Uncertainty in horizon tracking due uncertainties in imaging (equation 2.23). Control points are located at (a) $62.5\,\mathrm{m}$, (b) $562.5\,\mathrm{m}$, (c) $1062.5\,\mathrm{m}$, (d) $1562.5\,\mathrm{m}$, and (e) $2062.5\,\mathrm{m}$ horizontal location. Conditional mean estimates and 99% confidence intervals are shown in solid and shaded colors, respectively.

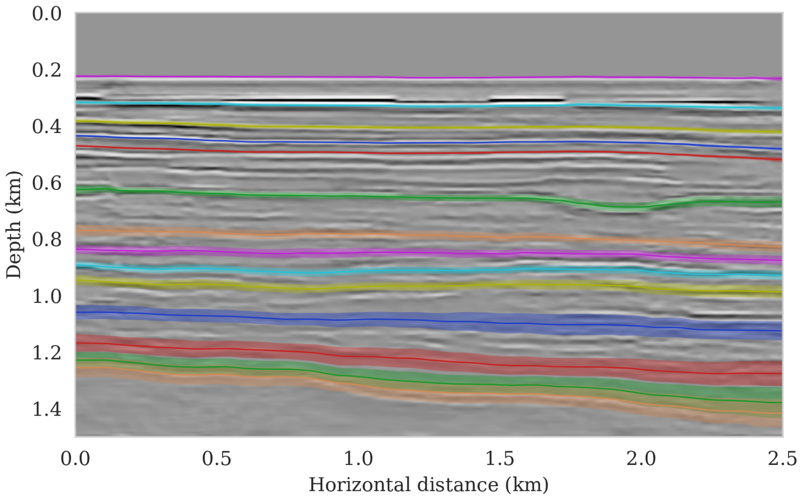


Figure 2.14: Uncertainty in horizon tracking due to a combination of uncertainties in imaging and control points (equation 2.24).

relies on sampling the posterior distribution, which is computationally expensive. Based on our experience and other studies [53, 117], estimates based on the conditional mean are comparatively robust to overfitting and yield superior results.

Even though having access to samples from the posterior has many advantages, e.g., it gives us access to the conditional mean and pointwise standard deviation estimates, its computational cost becomes typically prohibitive for large dimensional problems with expensive forward modeling operators. By using techniques from stochastic optimization, we managed to partly offset these costs by avoiding exact calculation of the multi-source data likelihood function. Similar to stochastic optimization, sometimes employed to solve wave-equation based optimization problems [16, 99, 100, 101, 102, 103], the gradient of the data misfit is calculated for artificially constructed simultaneous source experiments. This reduces the number of wave-equation solves for each gradient calculation significantly. In the context of Langevin dynamics—the theory undergirding our Markov chain Monte Carlo (MCMC) method to sample from the posterior—this stochastic approximation corresponds to the stochastic gradient Langevin dynamics (SGLD) method proposed by [67]. This approach, in combination with a preconditioning scheme and stepsize schedule, eliminates computationally prohibitive Metropolis-Hasting acceptance steps. By means of several numerical experiments and empirical accuracy measures, we established that the proposed SGLD algorithm is capable of drawing samples from the posterior with a reasonable accuracy. Aside from providing a reasonable assessment of the uncertainty, with pointwise standard deviation increasing in complex areas or in areas of relatively poor illumination, the samples from the posterior also allowed us to propagate uncertainties due to errors in the shot data all the way to the task of automatic horizon tracking or other tasks. With very few exceptions, we are not aware of this type of work on relatively large scale problems [70, 118].

While uncertainty quantification based on Bayesian inference certainly has its merits, it comes at a significant computational price even for 2D problems. These computational costs are compounded by the fact that the parameterization of seismic images in terms of a deep neural network is highly overparameterized, making it more difficult to solve the uncertainty quantification problem. Notwithstanding these challenges, the use of deep priors has several distinct advantages. First, the regularization comes from the inductive bias of the network architecture itself, which is designed to favor natural images. Second, this approach eliminates the need of having access to training data when compared to method that really prior information encoded in pretrained networks. Third, imposing a Gaussian prior on the networks weights is a common regularization strategy [65, 66]. Despite these advantages, the number of iterations needed by the SGLD algorithm remains high and prohibitive for imaging problems in 3D. For this reason, reducing the computational complexity of Bayesian inference over the weights of deep prior networks remains an activate area of research. For instance, [119] proposes to project the network weights onto a low-dimensional subspace and perform posterior inference within this reduced subspace. Unfortunately, construction of this reduced space can also be costly. For the purpose of applying this work to 3D seismic imaging, one possibility is to combine this dimensionality reduction approach with a different form of CNNs that exhibit similar inductive biases to those that we used in this work [51] but has fewer weights than the image dimensionality [120]. As a result, we may be able to exploit the inductive bias of CNNs at a lower computational cost.

Despite the fact that Bayesian inference based on MCMC methods, such as SGLD, is well-studied and widely employed, it is challenging for high-dimensional inverse problems that involve expensive forward operator. Variational (Bayesian) inference [109, 121], also known as distribution learning, can potentially overcome these challenges. In this approach, a neural network is trained to synthesize new samples from a target distribution based on a collection of training samples. Typically these samples are obtained by applying a series of learned nonlinear functions to random realizations from a canonical distribution. Early work on variational inference [122, 123, 124, 32, 125, 126, 33, 34, 48, 127] shows encouraging results, which opens enticing new perspectives on uncertainty quantification in the field of wave-equation based inversion.

Uncertainty in solving (linear) inverse problems including seismic imaging comes in two flavors. On the one hand, there is uncertainty related to noise in the input (shot) data. On the other hand, there may be modeling errors, such as linearization errors, which decrease with the accuracy of the background velocity model. We plan to study how these two types of uncertainty specifically affect the results in future work, with a certain regard for the impact of the background velocity model. Compared to the problem addressed in this chapter, this would entail multiple imaging experiments for different background velocity models and is therefore more challenging. Recent developments in full subsurface offset image volumes [128] and Fourier neural operators [129] may prove essential in addressing this problem.

2.10 Conclusions

Bayesian inference on high-dimensional inverse problems with computationally expensive forward modeling operators, has been, and continues to be a major challenge in the field of seismic imaging. Aside from obvious computational challenges, the selection of effective priors is problematic given the heterogeneity across geological scenarios and scales exhibited by elastic properties of the Earth's subsurface. To limit the possibly heavy-handed bias induced by a handcrafted prior, we propose regularization via deep priors. During this type of regularization, seismic images are restricted to the range of an untrained convolutional neural network with a fixed input, randomly initialized. Compared to conventional regularization, which tends to bias solutions towards sometimes restrictive choices made in defining these prior distributions, nonlinear deep priors derive regularizing properties from their overparameterized network architecture. The reparameterization of the seismic image by means of a deep prior leads to a Bayesian formulation where the prior is a Gaussian distribution of the weights of the network.

As long as overfitting can be avoided, regularization with deep priors is known to produce perceptually accurate results, an observation we confirmed in the context of controlled seismic imaging experiments. Unfortunately, preventing fitting the noise is difficult in practice. In addition, there is always the question how errors, e.g., bandwidth-limited noise or linearization errors, propagate to uncertainty in the image and to certain tasks to be carried out on the image, which for instance includes the task of automatic horizon tracking. To answer this question and to avoid the issue of overfitting, we propose to sample from the imaging posterior distribution and use the samples to compute the conditional mean estimate, which in our experiments exhibited more robustness to noise, and to obtain confidence intervals for the tracked horizons via our probabilistic horizon tracking framework.

Even though drawing samples from the posterior is computationally burdensome, it allows us to mitigate the imprint of overfitting while it is also conducive to a systematic framework mapping errors in shot data to uncertainty on the image and task at hand. By means of two imaging experiments derived from imaged seismic data volumes, we corroborated findings in the literature that the conditional mean estimate, i.e., the average over samples from the posterior distribution on the image, is more robust to overfitting than the maximum a posteriori estimate. The latter is the product of deterministic inversion.

Aside from improving the image quality itself with the conditional mean estimate, access to samples from the posterior also allows us to compute pointwise standard deviation on the image and confidence intervals on automatically tracked horizons.

With few exceptions as of yet, no systematic attempts have been made to account for uncertainties in the task of horizon tracking due to errors in the seismic imaging itself. These errors are caused by noise, linearization approximations, and uncertainty in the horizon tracking process itself, the latter being possibly related to differences in the selection of control points by different interpreters as part of the task of automatic horizon tracking.

To validate the proposed probabilistic tasked imaging framework, we considered realistic scenarios that are representative of two different geological settings. Our findings include: empirical verification of the accuracy of the samples from the posterior distribution; establishment of the conditional mean as a robust estimate for the image; reasonable estimates for the pointwise standard deviation on the image, showing an expected increase in variability in complex geological areas and in areas with poor illumination; and finally confidence intervals for the automatic horizon tracking given the uncertainty on the image and errors in the selection of control points guiding automatic horizon tracking.

2.11 Related material

The SGLD iterations (equation 2.15) require computing gradient of the negative-log posterior with respect to the CNN weights. This requires actions of the linearized Born scattering operator and its adjoint. For maximal numerical performance, the just-in-time Devito [130, 131] compiler was used for the wave-equation based simulations. To have access to the automatic differentiation utilities of PyTorch, we expose Devito's matrix-free implementations for the migration operator and its adjoint to PyTorch. In this way, we are able to compute the gradients required by equation 2.15 with automatic differentiation while exploiting Devito's highly optimized migration and demigration operators. For the CNN architecture, we followed [51]. For the automated horizon tracking we made use of software written by [69]. For more details on our implementation, please refer to our code on GitHub.

2.12 References

[1] G. Lambaré, J. Virieux, R. Madariaga, and S. Jin, “Iterative asymptotic inversion in the acoustic approximation,” *Geophysics*, vol. 57, no. 9, pp. 1138–1154, 1992 (page 15).

[2] G. T. Schuster, “Least-squares cross-well migration,” in *63rd Annual International Meeting, SEG*, Expanded Abstracts, 1993, pp. 110–113 (page 15).

[3] T. Nemeth, C. Wu, and G. T. Schuster, “Least-squares migration of incomplete reflection data,” *GEOPHYSICS*, vol. 64, no. 1, pp. 208–221, 1999 (pages 15, 19, 24).

[4] A. Malinverno and V. A. Briggs, “Expanded uncertainty quantification in inverse problems: Hierarchical bayes and empirical bayes,” *GEOPHYSICS*, vol. 69, no. 4, pp. 1005–1016, 2004 (page 15).

[5] A. Tarantola, *Inverse problem theory and methods for model parameter estimation*. SIAM, 2005, ISBN: 978-0-89871-572-9 (pages 15, 19).

[6] A. Malinverno and R. L. Parker, “Two ways to quantify uncertainty in geophysical inverse problems,” *GEOPHYSICS*, vol. 71, no. 3, W15–W27, 2006 (page 15).

[7] 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. eprint: http://epubs.siam.org/doi/pdf/10.1137/110845598 (pages 15, 28).

[8] A. Ray, S. Kaplan, J. Washbourne, and U. Albertin, “Low frequency full waveform seismic inversion within a tree based Bayesian framework,” *Geophysical Journal International*, vol. 212, no. 1, pp. 522–542, Oct. 2017. eprint: https://academic.oup.com/gji/article-pdf/212/1/522/21782947/ggx428.pdf (page 15).

[9] Z. Fang, C. D. Silva, R. Kuske, and F. J. Herrmann, “Uncertainty quantification for inverse problems with weak partial-differential-equation constraints,” *GEOPHYSICS*, vol. 83, no. 6, R629–R647, 2018 (pages 15, 28, 43).

[10] G. K. Stuart, S. E. Minkoff, and F. Pereira, “A two-stage Markov chain Monte Carlo method for seismic inversion and uncertainty quantification,” *GEOPHYSICS*, vol. 84, no. 6, R1003–R1020, Nov. 2019 (page 15).

[11] Z. Zhao and M. K. Sen, “A gradient based MCMC method for FWI and uncertainty analysis,” in *89th Annual International Meeting, SEG*, Expanded Abstracts, 2019, pp. 1465–1469 (pages 15, 28).

[12] M. Kotsi, A. Malcolm, and G. Ely, “Uncertainty quantification in time-lapse seismic imaging: a full-waveform approach,” *Geophysical Journal International*, vol. 222, no. 2, pp. 1245–1263, May 2020 (pages 15, 28).

[13] A. Malinverno, “Parsimonious bayesian markov chain monte carlo inversion in a nonlinear geophysical problem,” *Geophysical Journal International*, vol. 151, no. 3, pp. 675–688, 2002. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1046/j.1365-246X.2002.01847.x (page 15).

[14] D. Zhu and R. Gibson, “Seismic inversion and uncertainty quantification using transdimensional markov chain monte carlo method,” *GEOPHYSICS*, vol. 83, no. 4, R321–R334, 2018 (pages 15, 36).

[15] G. Visser, P. Guo, and E. Saygin, “Bayesian transdimensional seismic full-waveform inversion with a dipping layer parameterization,” *GEOPHYSICS*, vol. 84, no. 6, R845–R858, 2019 (page 15).

[16] T. van Leeuwen, A. Y. Aravkin, and F. Herrmann, “Seismic Waveform Inversion by Stochastic Optimization,” *International Journal of Geophysics*, vol. 2011, pp. 1–18, 2011 (pages 15, 20, 28, 63).

[17] F. J. Herrmann and X. Li, “Efficient least-squares imaging with sparsity promotion and compressive sensing,” *Geophysical Prospecting*, vol. 60, no. 4, pp. 696–712, Jul. 2012 (pages 15, 20).

[18] X. Lu, L. Han, J. Yu, and X. Chen, “L1 norm constrained migration of blended data with the fista algorithm,” *Journal of Geophysics and Engineering*, vol. 12, no. 4, pp. 620–628, 2015 (pages 15, 20).

[19] N. Tu and F. J. Herrmann, “Fast imaging with surface-related multiples by sparse inversion,” *Geophysical Journal International*, vol. 201, no. 1, pp. 304–317, Apr. 2015 (pages 15, 20).

[20] H. Zhu, S. Li, S. Fomel, G. Stadler, and O. Ghattas, “A bayesian approach to estimate uncertainty for full-waveform inversion using a priori information from depth migration,” *Geophysics*, vol. 81, no. 5, R307–R323, 2016 (page 15).

[21] G. Ely, A. Malcolm, and O. V. Poliannikov, “Assessing uncertainties in velocity models and images with a fast nonlinear uncertainty quantification method,” *GEOPHYSICS*, vol. 83, no. 2, R63–R75, 2018. eprint: https://doi.org/10.1190/geo2017-0321.1 (pages 15, 46).

[22] M. Izzatullah, R. Baptista, L. Mackey, Y. Marzouk, and D. Peter, “Bayesian seismic inversion: Measuring Langevin MCMC sample quality with kernels,” in *90th Annual International Meeting, SEG*, Expanded Abstracts, Jun. 2020, pp. 295–299 (page 15).

[23] L. Mosser, O. Dubrule, and M. Blunt, “Stochastic Seismic Waveform Inversion Using Generative Adversarial Networks as a Geological Prior,” *Mathematical Geosciences*, vol. 84, no. 1, pp. 53–79, 2019 (pages 15, 20).

[24] Z. Zhang and T. Alkhalifah, “Regularized elastic full waveform inversion using deep learning,” *GEOPHYSICS*, vol. 84, no. 5, 1SO–Z28, Sep. 2019 (page 15).

[25] Z. Fang, H. Fang, and L. Demanet, “Chapter Five - Deep generator priors for Bayesian seismic inversion,” in *Machine Learning in Geosciences*, ser. Advances in Geophysics, B. Moseley and L. Krischer, Eds., vol. 61, Elsevier, 2020, pp. 179–216 (pages 15, 20).

[26] Z. Fang, H. Fang, and L. Demanet, “Quality control of deep generator priors for statistical seismic inverse problems,” in *90th Annual International Meeting, SEG*, Expanded Abstracts, 2020, pp. 1715–1719 (page 15).

[27] Z. Liu, Y. Chen, and G. Schuster, “Deep Convolutional Neural Network and Sparse Least Squares Migration,” *Geophysics*, vol. 85, no. 4, pp. 1–57, Jul. 2020 (page 15).

[28] L. Mosser, S. Purves, and E. Naeini, “Deep Bayesian Neural Networks for Fault Identification and Uncertainty Quantification,” vol. 2020, no. 1, pp. 1–5, 2020 (page 15).

[29] H. Sun and L. Demanet, “Extrapolated full-waveform inversion with deep learning,” *Geophysics*, vol. 85, no. 3, R275–R288, 2020 (page 15).

[30] Y. Wu and Y. Lin, “InversionNet: An Efficient and Accurate Data-Driven Full Waveform Inversion,” *IEEE Transactions on Computational Imaging*, vol. 6, pp. 419–433, 2020 (page 15).

[31] V. Kazei, O. Ovcharenko, P. Plotnitskii, D. Peter, X. Zhang, and T. Alkhalifah, “Mapping full seismic waveforms to vertical velocity profiles by deep learning,” *Geophysics*, vol. 86, no. 5, pp. 1–50, 2021 (page 15).

[32] K. Kothari, A. Khorashadizadeh, M. de Hoop, and I. Dokmanić, “Trumpets: Injective Flows for Inference and Inverse Problems,” *arXiv preprint arXiv:2102.10461*, 2021 (pages 15, 65).

[33] R. Kumar, M. Kotsi, A. Siahkoohi, and A. Malcolm, “Enabling uncertainty quantification for seismic data preprocessing using normalizing flows (NF)—An interpo-

lation example," in *First International Meeting for Applied Geoscience & Energy*, Expanded Abstracts, 2021, pp. 1515–1519 (pages 15, 65).

[34] A. Siahkoohi and F. J. Herrmann, "Learning by example: fast reliability-aware seismic imaging with normalizing flows," in *First International Meeting for Applied Geoscience & Energy*, Expanded Abstracts, 2021, pp. 1580–1585 (pages 15, 65).

[35] J. Adler and O. öktem, "Deep Bayesian Inversion," *arXiv preprint arXiv:1811.05910*, 2018 (page 16).

[36] P. Putzky and M. Welling, "Invert to Learn to Invert," in *Advances in Neural Information Processing Systems*, vol. 32, Curran Associates, Inc., 2019 (page 16).

[37] M. Asim, M. Daniels, O. Leong, A. Ahmed, and P. Hand, "Invertible generative models for inverse problems: Mitigating representation error and dataset bias," in *Proceedings of the 37th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 119, PMLR, Jul. 2020, pp. 399–409 (pages 16, 20).

[38] A. Hauptmann and B. T. Cox, "Deep Learning in Photoacoustic Tomography: Current approaches and future directions," *Journal of Biomedical Optics*, vol. 25, no. 11, p. 112 903, 2020 (page 16).

[39] A. Sriram *et al.*, "End-to-End Variational Networks for Accelerated MRI Reconstruction," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2020, pp. 64–73 (page 16).

[40] S. Mukherjee, M. Carioni, O. öktem, and C.-B. Schönlieb, "End-to-end reconstruction meets data-driven regularization for inverse problems," *arXiv preprint arXiv:2106.03538*, 2021 (page 16).

[41] A. Siahkoohi, M. Louboutin, and F. J. Herrmann, "The importance of transfer learning in seismic modeling and imaging," *GEOPHYSICS*, vol. 84, no. 6, A47–A52, Nov. 2019 (page 16).

[42] H. Kaur, N. Pham, and S. Fomel, "Improving the resolution of migrated images by approximating the inverse Hessian using deep learning," *Geophysics*, vol. 85, no. 4, WA173–WA183, 2020 (page 16).

[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 (page 16).

[44] R. Rojas-Gómez, J. Yang, Y. Lin, J. Theiler, and B. Wohlberg, "Physics-consistent data-driven waveform inversion with adaptive data augmentation," *IEEE Geoscience and Remote Sensing Letters*, pp. 1–5, 2020 (page 16).

[45] H. Sun and L. Demanet, "Elastic full-waveform inversion with extrapolated low-frequency data," in *SEG Technical Program Expanded Abstracts 2020*, Society of Exploration Geophysicists, 2020, pp. 855–859 (page 16).

[46] M. Zhang, A. Siahkoohi, and F. J. Herrmann, "Transfer learning in large-scale ocean bottom seismic wavefield reconstruction," in *90th Annual International Meeting, SEG*, Expanded Abstracts, 2020, pp. 1666–1670 (page 16).

[47] R. Barbano, Z. Kereta, A. Hauptmann, S. R. Arridge, and B. Jin, "Unsupervised knowledge-transfer for learned image reconstruction," *arXiv preprint arXiv:2107.02572*, 2021 (page 16).

[48] A. Siahkoohi, G. Rizzuti, M. Louboutin, P. Witte, and F. J. Herrmann, "Preconditioned training of normalizing flows for variational inference in inverse problems," in *3rd Symposium on Advances in Approximate Bayesian Inference*, Jan. 2021 (pages 16, 20, 65).

[49] S. Qu, E. Verschuur, D. Zhang, and Y. Chen, "Training deep networks with only synthetic data: Deep-learning-based near-offset reconstruction for (closed-loop) surface-related multiple estimation on shallow-water field data," *Geophysics*, vol. 86, no. 3, A39–A43, 2021 (page 16).

[50] J.-W. Vrolijk and G. Blacquière, "Source deghosting of coarsely sampled common-receiver data using a convolutional neural network," *Geophysics*, vol. 86, no. 3, pp. V185–V196, 2021 (page 16).

[51] V. Lempitsky, A. Vedaldi, and D. Ulyanov, "Deep Image Prior," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2018, pp. 9446–9454 (pages 16, 21, 23, 39, 64, 67).

[52] S. Arridge, P. Maass, O. Öktem, and C.-B. Schönlieb, "Solving inverse problems using data-driven models," *Acta Numerica*, vol. 28, pp. 1–174, 2019 (pages 16, 23).

[53] Z. Cheng, M. Gadelha, S. Maji, and D. Sheldon, "A Bayesian Perspective on the Deep Image Prior," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2019, pp. 5443–5451 (pages 16, 22, 39, 63).

[54] M. Gadelha, R. Wang, and S. Maji, "Shape reconstruction using differentiable projections and deep priors," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 22–30 (page 16).

[55] Q. Liu, L. Fu, and M. Zhang, “Deep-seismic-prior-based reconstruction of seismic data using convolutional neural networks,” *arXiv preprint arXiv:1911.08784*, 2019 (page 16).

[56] Y. Wu and G. A. McMechan, “Parametric convolutional neural network-domain full-waveform inversion,” *GEOPHYSICS*, vol. 84, no. 6, R881–R896, 2019 (page 16).

[57] S. Dittmer, T. Kluth, P. Maass, and D. Otero Baguer, “Regularization by architecture: A deep prior approach for inverse problems,” *Journal of Mathematical Imaging and Vision*, vol. 62, no. 3, pp. 456–470, 2020 (pages 16, 22, 23).

[58] F. Kong, F. Picetti, V. Lipari, P. Bestagini, X. Tang, and S. Tubaro, “Deep Prior-Based Unsupervised Reconstruction of Irregularly Sampled Seismic Data,” *IEEE Geoscience and Remote Sensing Letters*, vol. 19, no. 7501305, 2020 (page 16).

[59] Y. Shi, X. Wu, and S. Fomel, “Deep learning parameterization for geophysical inverse problems,” in *SEG 2019 Workshop: Mathematical Geophysics: Traditional vs Learning, Beijing, China, 5-7 November 2019*, Society of Exploration Geophysicists, 2020, pp. 36–40 (page 16).

[60] A. Siahkoohi, G. Rizzuti, and F. J. Herrmann, “A deep-learning based bayesian approach to seismic imaging and uncertainty quantification,” in *82nd EAGE Conference and Exhibition*, Extended Abstracts, 2020 (pages 16, 26, 28).

[61] A. Siahkoohi, G. Rizzuti, and F. J. Herrmann, “Uncertainty quantification in imaging and automatic horizon tracking—a Bayesian deep-prior based approach,” in *90th Annual International Meeting, SEG*, Expanded Abstracts, Sep. 2020, pp. 1636–1640 (pages 16, 26, 28).

[62] A. Siahkoohi, G. Rizzuti, and F. J. Herrmann, “Weak deep priors for seismic imaging,” in *90th Annual International Meeting, SEG*, Expanded Abstracts, Sep. 2020, pp. 2998–3002 (pages 16, 25).

[63] M. Tölle, M.-H. Laves, and A. Schlaefer, “A Mean-Field Variational Inference Approach to Deep Image Prior for Inverse Problems in Medical Imaging,” in *Medical Imaging with Deep Learning*, 2021 (page 16).

[64] T. M. Mitchell, “The need for biases in learning generalizations,” Department of Computer Science, Laboratory for Computer Science Research, Rutgers University, Tech. Rep. CBM-TR 5-110, 1980 (page 16).

[65] A. Krogh and J. A. Hertz, “A simple weight decay can improve generalization,” in *Advances in Neural Information Processing Systems*, 1992, pp. 950–957 (pages 16, 64).

[66] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org (pages 16, 64).

[67] M. Welling and Y. W. Teh, “Bayesian Learning via Stochastic Gradient Langevin Dynamics,” in *Proceedings of the 28th International Conference on Machine Learning*, Bellevue, Washington, USA: Omnipress, 2011, pp. 681–688, ISBN: 9781450306195 (pages 16, 28–30, 34, 63).

[68] C. Li, C. Chen, D. Carlson, and L. Carin, “Preconditioned Stochastic Gradient Langevin Dynamics for Deep Neural Networks,” in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, Phoenix, Arizona: AAAI Press, 2016, pp. 1788–1794 (pages 16, 31, 32, 40).

[69] X. Wu and S. Fomel, “Least-squares horizons with local slopes and multi-grid correlations,” *GEOPHYSICS*, vol. 83, pp. IM29–IM40, 4 2018 (pages 17, 52, 53, 68).

[70] R. Arnold and A. Curtis, “Interrogation theory,” *Geophysical Journal International*, vol. 214, no. 3, pp. 1830–1846, Jun. 2018. eprint: https://academic.oup.com/gji/article-pdf/214/3/1830/25116266/ggy248.pdf (pages 17, 64).

[71] B. Peters, J. Granek, and E. Haber, “Multiresolution neural networks for tracking seismic horizons from few training images,” *Interpretation*, vol. 7, no. 3, SE201–SE213, 2019. eprint: https://doi.org/10.1190/INT-2018-0225.1 (page 17).

[72] Z. Geng, X. Wu, Y. Shi, and S. Fomel, “Deep learning for relative geologic time and seismic horizons,” *GEOPHYSICS*, vol. 85, no. 4, WA87–WA100, Jul. 2020 (page 17).

[73] B. Peters and E. Haber, “Fully Reversible Neural Networks for Large-Scale 3D Seismic Horizon Tracking,” vol. 2020, no. 1, pp. 1–5, 2020 (page 17).

[74] Y. Shi, X. Wu, and S. Fomel, “Waveform embedding: Automatic horizon picking with unsupervised deep learning,” *GEOPHYSICS*, vol. 85, no. 4, WA67–WA76, Jul. 2020 (page 17).

[75] A. A. Valenciano, “Imaging by wave-equation inversion,” Ph.D. dissertation, Stanford University, 2008 (page 19).

[76] S. Dong *et al.*, “Least-squares reverse time migration: Towards true amplitude imaging and improving the resolution,” in *82nd Annual International Meeting, SEG*, Expanded Abstracts, 2012, pp. 1–5 (page 19).

[77] C. Zeng, S. Dong, and B. Wang, “Least-squares reverse time migration: Inversion-based imaging toward true reflectivity,” *The Leading Edge*, vol. 33, no. 9, pp. 962–968, 2014 (page 19).

[78] D. L. Donoho, “Compressed Sensing,” *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289–1306, 2006 (page 20).

[79] A. Bora, A. Jalal, E. Price, and A. G. Dimakis, “Compressed sensing using generative models,” in *Proceedings of the 34th International Conference on Machine Learning*, D. Precup and Y. W. Teh, Eds., ser. Proceedings of Machine Learning Research, vol. 70, PMLR, Jun. 2017, pp. 537–546 (page 20).

[80] X. Glorot and Y. Bengio, “Understanding the difficulty of training deep feedforward neural networks,” in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, vol. 9, PMLR, May 2010, pp. 249–256 (pages 21, 23, 35).

[81] G. Casella and R. L. Berger, *Statistical Inference*. Cengage Learning, 2002 (pages 23, 24).

[82] R. C. Aster, B. Borchers, and C. H. Thurber, *Parameter Estimation and Inverse Problems*. Elsevier, 2018 (pages 24, 34).

[83] H. Li, Z. Xu, G. Taylor, C. Studer, and T. Goldstein, “Visualizing the loss landscape of neural nets,” in *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., vol. 31, Curran Associates, Inc., 2018 (page 25).

[84] T. Tieleman and G. Hinton, *Lecture 6.5-RMSprop: Divide the gradient by a running average of its recent magnitude*, https://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf, 2012 (pages 25, 28, 31, 32, 39, 57).

[85] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014 (pages 25, 28).

[86] J. Bernstein, A. Vahdat, Y. Yue, and M.-Y. Liu, “On the distance between two neural networks and the stability of learning,” *arXiv preprint arXiv:2002.03432*, 2020 (page 25).

[87] S. S. Du, X. Zhai, B. Poczos, and A. Singh, “Gradient descent provably optimizes over-parameterized neural networks,” in *International Conference on Learning Representations*, 2019 (page 25).

[88] D. Kunin, J. Bloom, A. Goeva, and C. Seed, “Loss landscapes of regularized linear autoencoders,” in *Proceedings of the 36th International Conference on Machine Learning*, K. Chaudhuri and R. Salakhutdinov, Eds., ser. Proceedings of Machine Learning Research, vol. 97, PMLR, Sep. 2019, pp. 3560–3569 (page 25).

[89] Bogachev, V.I., *Measure Theory*. Springer-Verlag Berlin Heidelberg, 2006, ISBN: 9783540345138 (page 25).

[90] B. D. Anderson and J. B. Moore, *Optimal Filtering*. Prentice-Hall, Englewood Cliffs, NJ., 1979 (pages 26, 34).

[91] D. J. MacKay and D. J. Mac Kay, *Information theory, inference and learning algorithms*. Cambridge university press, 2003 (page 26).

[92] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning* (Springer Series in Statistics). Springer New York Inc., 2001 (pages 27, 43).

[93] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, *Bayesian Data Analysis*. CRC press, 2013 (pages 28, 36, 43).

[94] A. Curtis and A. Lomax, "Prior information, sampling distributions, and the curse of dimensionality," *Geophysics*, vol. 66, no. 2, pp. 372–378, 2001 (page 28).

[95] F. J. Herrmann, A. Siahkoohi, and G. Rizzuti, "Learned imaging with constraints and uncertainty quantification," in *Neural Information Processing Systems (NeurIPS) 2019 Deep Inverse Workshop*, Dec. 2019 (page 28).

[96] H. Robbins and S. Monro, "A stochastic approximation method," *The annals of mathematical statistics*, pp. 400–407, 1951 (pages 28, 30).

[97] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro, "Robust stochastic approximation approach to stochastic programming," *SIAM Journal on optimization*, vol. 19, no. 4, pp. 1574–1609, 2009 (page 28).

[98] C. Li, J. Huang, Z. Li, and R. Wang, "Plane-wave least-squares reverse time migration with a preconditioned stochastic conjugate gradient method," *Geophysics*, vol. 83, no. 1, S33–S46, 2018 (page 28).

[99] E. Haber, M. Chung, and F. Herrmann, "An effective method for parameter estimation with pde constraints with multiple right-hand sides," *SIAM Journal on Optimization*, vol. 22, no. 3, pp. 739–757, 2012 (pages 28, 63).

[100] F. J. Herrmann and X. Li, "Efficient least-squares imaging with sparsity promotion and compressive sensing," *Geophysical Prospecting*, vol. 60, pp. 696–712, 2012 (pages 28, 63).

[101] X. Lu, L. Han, J. Yu, and X. Chen, "L1 norm constrained migration of blended data with the FISTA algorithm," *Journal of Geophysics and Engineering*, vol. 12, pp. 620–628, 2015 (pages 28, 63).

[102] N. Tu and F. Herrmann, “Fast imaging with surface-related multiples by sparse inversion,” *Geophysical Journal International*, vol. 201, no. 1, pp. 304–417, 2015 (pages 28, 63).

[103] C. Li, J. Huang, Z. Li, and R. Wang, “Plane-wave least-squares reverse time migration with a preconditioned stochastic conjugate gradient method,” *Geophysics*, vol. 83, no. 1, S33–S46, 2018. eprint: https://doi.org/10.1190/geo2017-0339.1 (pages 28, 63).

[104] R. M. Neal, “MCMC using Hamiltonian dynamics,” *Handbook of Markov Chain Monte Carlo*, pp. 113–162, May 2011 (pages 29, 30).

[105] I. Sato and H. Nakagawa, “Approximation analysis of stochastic gradient langevin dynamics by using fokker-planck equation and ito process,” in *Proceedings of the 31st International Conference on Machine Learning*, E. P. Xing and T. Jebara, Eds., ser. Proceedings of Machine Learning Research, vol. 32, Bejing, China: PMLR, 22–24 Jun 2014, pp. 982–990 (page 29).

[106] M. Raginsky, A. Rakhlin, and M. Telgarsky, “Non-Convex Learning via Stochastic Gradient Langevin Dynamics: A Nonasymptotic Analysis,” in *Conference on Learning Theory*, PMLR, 2017, pp. 1674–1703 (page 29).

[107] N. Brosse, A. Durmus, and E. Moulines, “The promises and pitfalls of Stochastic Gradient Langevin Dynamics,” in *Advances in Neural Information Processing Systems 31*, Curran Associates, Inc., 2018, pp. 8268–8278 (pages 29, 30).

[108] W. Deng, G. Lin, and F. Liang, “A Contour Stochastic Gradient Langevin Dynamics Algorithm for Simulations of Multi-modal Distributions,” in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds., vol. 33, Curran Associates, Inc., 2020, pp. 15 725–15 736 (page 29).

[109] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul, “An Introduction to Variational Methods for Graphical Models,” *Machine Learning*, vol. 37, no. 2, pp. 183–233, 1999 (pages 29, 65).

[110] Y. W. Teh, A. H. Thiery, and S. J. Vollmer, “Consistency and fluctuations for stochastic gradient Langevin dynamics,” *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 193–225, 2016 (pages 30, 33).

[111] A. Gelman and D. B. Rubin, “Inference from Iterative Simulation Using Multiple Sequences,” *Statistical Science*, vol. 7, no. 4, pp. 457–472, 1992 (pages 32, 34).

[112] Veritas, “Parihaka 3D Marine Seismic Survey - Acquisition and Processing Report,” New Zealand Petroleum & Minerals, Wellington, Tech. Rep. New Zealand Petroleum Report 3460, 2005 (page 36).

[113] WesternGeco., “Parihaka 3D PSTM Final Processing Report,” New Zealand Petroleum & Minerals, Wellington, Tech. Rep. New Zealand Petroleum Report 4582, 2012 (page 36).

[114] M. Burger and F. Lucka, “Maximum a posteriori estimates in linear inverse problems with log-concave priors are proper Bayes estimators,” *Inverse Problems*, vol. 30, no. 11, p. 114 004, 2014 (page 40).

[115] A. P. Dawid, “Conditional Independence in Statistical Theory,” *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 41, no. 1, pp. 1–31, 1979 (page 48).

[116] 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*, Extended Abstracts, 2012 (page 56).

[117] A. G. Wilson and P. Izmailov, “Bayesian Deep Learning and a Probabilistic Perspective of Generalization,” *arXiv preprint arXiv:2002.08791*, 2020 (page 63).

[118] Y. Cho and H. Jun, “Estimation and uncertainty analysis of the co2 storage volume in the sleipner field via 4d reversible-jump markov-chain monte carlo,” *Journal of Petroleum Science and Engineering*, vol. 200, p. 108 333, 2021 (page 64).

[119] P. Izmailov, W. J. Maddox, P. Kirichenko, T. Garipov, D. Vetrov, and A. G. Wilson, “Subspace inference for Bayesian deep learning,” in *Uncertainty in Artificial Intelligence*, Proceedings of Machine Learning Research (PMLR), 2020, pp. 1169–1179 (page 64).

[120] R. Heckel and P. Hand, “Deep decoder: Concise image representations from untrained non-convolutional networks,” in *7th International Conference on Learning Representations*, OpenReview.net, 2019 (page 65).

[121] D. Rezende and S. Mohamed, “Variational inference with normalizing flows,” ser. Proceedings of Machine Learning Research, vol. 37, PMLR, Jul. 2015, pp. 1530–1538 (page 65).

[122] G. Rizzuti, A. Siahkoohi, P. A. Witte, and F. J. Herrmann, “Parameterizing uncertainty by deep invertible networks, an application to reservoir characterization,” in *90th Annual International Meeting, SEG*, Sep. 2020, pp. 1541–1545 (page 65).

[123] A. Siahkoohi, G. Rizzuti, P. A. Witte, and F. J. Herrmann, “Faster Uncertainty Quantification for Inverse Problems with Conditional Normalizing Flows,” Georgia Institute of Technology, Tech. Rep. TR-CSE-2020-2, Jul. 2020 (page 65).

[124] X. Zhang and A. Curtis, “Seismic tomography using variational inference methods,” *Journal of Geophysical Research: Solid Earth*, vol. 125, no. 4, e2019JB018589, 2020 (page 65).

[125] N. Kovachki, R. Baptista, B. Hosseini, and Y. Marzouk, *Conditional Sampling With Monotone GANs*, 2021. arXiv: 2006.06755 `[stat.ML]` (page 65).

[126] J. Kruse, G. Detommaso, R. Scheichl, and U. Köthe, “HINT: Hierarchical Invertible Neural Transport for Density Estimation and Bayesian Inference,” *Proceedings of AAAI-2021*, 2021 (page 65).

[127] X. Zhao, A. Curtis, and X. Zhang, “Bayesian seismic tomography using normalizing flows,” *Geophysical Journal International*, vol. 228, no. 1, pp. 213–239, Jul. 2021. eprint: https://academic.oup.com/gji/article-pdf/228/1/213/40348424/ggab298.pdf (page 65).

[128] M. Yang, M. Graff, R. Kumar, and F. J. Herrmann, “Low-rank representation of omnidirectional subsurface extended image volumes,” *Geophysics*, vol. 86, no. 3, S165–S183, Jan. 2021 (page 65).

[129] Z. Li *et al.*, “Fourier Neural Operator for Parametric Partial Differential Equations,” in *9th International Conference on Learning Representations*, OpenReview.net, 2021 (page 65).

[130] F. Luporini *et al.*, “Architecture and performance of devito, a system for automated stencil computation,” *CoRR*, vol. abs/1807.03032, Jul. 2018. arXiv: 1807.03032 (page 67).

[131] 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 (page 67).

CHAPTER 3
WEAK DEEP PRIORS FOR SEISMIC IMAGING

3.1 Summary

Incorporating prior knowledge on model unknowns of interest is essential when dealing with ill-posed inverse problems due to the nonuniqueness of the solution and data noise. Unfortunately, it is not trivial to fully describe our priors in a convenient and analytical way. Parameterizing the unknowns with a convolutional neural network (CNN), and assuming an uninformative Gaussian prior on its weights, leads to a variational prior on the output space that favors “natural” images and excludes noisy artifacts, as long as overfitting is prevented. This is the so-called deep-prior approach. In seismic imaging, however, evaluating the forward operator is computationally expensive, and training a randomly initialized CNN becomes infeasible. We propose, instead, a weak version of deep priors, which consists of relaxing the requirement that reflectivity models must lie in the network range, and letting the unknowns deviate from the network output according to a Gaussian distribution. Finally, we jointly solve for the reflectivity model and CNN weights. The chief advantage of this approach is that the updates for the CNN weights do not involve the modeling operator, and become relatively cheap. Our synthetic numerical experiments demonstrate that the weak deep prior is more robust with respect to noise than conventional least-squares imaging approaches, with roughly twice the computational cost of reverse-time migration, which is the affordable computational budget in large-scale imaging problems.

3.2 Introduction

Linearized seismic imaging involves an inconsistent, ill-conditioned linear inverse problem due to presence of shadow zones and complex structures in the subsurface, coherent linearization errors, and noisy data. Due to nonuniqueness, using prior information as regularization is essential. This particular choice is crucial because it typically affects the final result. Conventional methods mostly rely on handcrafted and unrealistic priors, such as a Gaussian or Laplace distributed model parameters (in the physical or in a transform domain). These simplifying assumptions, while being practical, negatively bias the outcome of the inversion.

Recent proposals [1, 2, 3, 4, 5, 6, 7] make use of convolutional neural networks (CNN) as a prior. Specifically, [7] reparameterize the unknown reflectivity model by a CNN and impose a Gaussian prior on its weights. These authors show that the combination of the functional form of a CNN and a Gaussian prior on its weights is a suitable prior for seismic imaging. However, since every update to CNN weights requires the action of the forward operator and its adjoint, tuning randomly initialized CNN weights need many stochastic optimization steps. In seismic imaging, computing the action of the forward operator—i.e., linearized Born scattering operator, and its adjoint is computationally expensive, which might limit the application of deep priors.

We propose the *weak deep prior*, a computationally convenient formulation that relaxes deep priors. Instead of reparameterizing the unknowns with CNNs, we let the unknown reflectivity to be distributed according to a Gaussian distribution centered at the CNN network output. Next, we jointly solve for the reflectivity model and CNN weights. This formulation decouples the forward operator with the CNN, allowing for fast and forward-operator free updates of CNN weights, while partially keeping the advantages of the deep prior. The proposed formulation additionally allows for imposing handcrafted or physical hard constraints on the unknowns, which is often not feasible when imposing deep priors [8].

In general, numerous efforts involve the incorporation of ideas from deep learning in seismic processing and inversion [9, 10, 11, 12, 13, 14, 15]. Deep prior itself have been utilized by [4] to perform seismic data reconstruction. [5] propose to pretrain a randomly initialized CNN before reparameterizing the velocity model in the context of Full-Waveform Inversion. [6] use the deep priors in the context of denoising. Finally, [7] proposes a deep-prior based Bayesian framework for seismic imaging and perform uncertainty quantification.

Our work is organized as follows. We first introduce the original concept of deep prior and how it can be integrated in seismic imaging. Next, we develop the weak deep prior framework and the associated optimization problem. We conclude by showcasing the proposed method using a synthetic example involving a 2D portion of a real migrated image of the 3D Parihaka dataset [16, 17] in the presence of strong noise.

3.3 Seismic imaging

Seismic imaging is the problem of estimating the short-wavelength structure of the Earth's subsurface, denoted by $\delta\mathbf{m}$ given data recorded at the surface, $\delta\mathbf{d}_i, \ i = 1, 2, \cdots, N$, where N is the number of shot records. Besides observed data, this inverse problem requires a smooth background squared-slowness model, $\mathbf{m}_0$, and estimated source signatures, $\mathbf{q}_i$. When noise in the data can be approximated by a zero-mean Gaussian random variable, ℓ_2-norm data discrepancy defines the likelihood function [18]. Assuming the noise covariance is $\sigma^2\mathbf{I}$, we can write the negative log-likelihood of the observed data as follows:

$$\begin{aligned} -\log p_{\text{like}}\left(\{\delta\mathbf{d}_i\}_{i=1}^{N} \,|\delta\mathbf{m}\right) &= -\sum_{i=1}^{N} \log p_{\text{like}}\left(\delta\mathbf{d}_i|\delta\mathbf{m}\right) \\ &= \frac{1}{2\sigma^2}\sum_{i=1}^{N} \|\delta\mathbf{d}_i - \mathbf{J}(\mathbf{m}_0, \mathbf{q}_i)\delta\mathbf{m}\|_2^2 \quad + \underbrace{\text{const}}_{\text{Ind. of } \delta\mathbf{m}} . \end{aligned} \tag{3.1}$$

In these expressions, p_{like} denotes the likelihood probability density function, and $\mathbf{J}$ is the linearized Born scattering operator. The maximum likelihood estimate (MLE), denoted by $\widehat{\delta\mathbf{m}}_{\text{MLE}}$, is obtained by minimizing the negative-log likelihood defined in Equation 3.1 with respect to $\delta\mathbf{m}$. Notoriously, MLE estimators tend to produce imaging artifacts. To address this issue, we discuss a special kind of prior based on neural networks: the so-called deep priors.

3.4 Imaging with deep priors

Parameterizing the unknown variables with a CNN, with a fixed input, has shown promising results in inverse problems [1, 2, 3, 4, 5, 6, 7]. In this approach, weights and biases are Gaussian random variables and they are tuned to fit the observed data. The success of this approach hinges on the special structure of the CNN, which tends to favor noise-free looking images. Despite this feature, it should be noted that a stopping criteria is still essential to avoid overfitting the noise in observed data. Notwithstanding this challenge, we propose to parameterize the unknown reflectivity model by a CNN—i.e., $\delta\mathbf{m} = g(\mathbf{z}, \mathbf{w})$, where $\mathbf{z} \sim \mathrm{N}(\mathbf{0}, \mathbf{I})$ is the fixed input to the CNN and $\mathbf{w}$ denotes the unknown CNN weights. Imposing a Gaussian prior on $\mathbf{w}$ with covariance matrix $\lambda^{-2}\mathbf{I}$ allows us to formulate the negative log-posterior distribution for $\mathbf{w}$ as follows:

$$\begin{aligned} & p_{\text{post}} \left(\mathbf{w} \middle| \left\{ \delta\mathbf{d}_i \right\}_{i=1}^{N} \right) \propto \left[\prod_{i=1}^{N} p_{\text{like}} \left(\delta\mathbf{d}_i | \mathbf{w} \right) \right] p_w \left(\mathbf{w} \right), \\ & \text{where} \quad p_w \left(\mathbf{w} \right) = \mathrm{N}(\mathbf{w} | \mathbf{0}, \lambda^{-2}\mathbf{I}). \end{aligned} \tag{3.2}$$

In the equation above, p_w and p_{post} denote the prior and posterior probability density functions, respectively. The maximum a posteriori estimator (MAP), denoted by $\widehat{\mathbf{w}}_{\text{deep}}$, is obtained by maximizing Equation 3.2 with respect to $\mathbf{w}$.

As stated before, there are two challenges in employing deep priors in seismic imaging. The first challenge is finding a stopping criteria while maximizing the posterior in Equation 3.2 to prevent noise overfit. [7] propose to perform stochastic gradient Langevin dynamics [SGLD, 19] steps to obtain samples from this posterior distribution. Using these samples, these authors approximate the conditional mean estimator, which prevents overfitting and at the same time, yields a seismic image that has less imaging artifacts compared to the MAP estimator. However, sampling the posterior is a challenging feat in and of itself and is outside of the scope of this discussion. Another challenge associated with deep-prior based imaging is the number of iterations needed to optimize the CNN weights. Unless the CNN is pretrained, its weights are initialized randomly, hence, solving for $\mathbf{w}$ requires many iterations involving the seismic modeling operator and its adjoint and may not be computationally practical. Unfortunately, unlike other imaging modalities, such as medical imaging, we generally do not have access to detailed information on the subsurface. This limits the scope of the pretraining phase, which in turn might adversely bias the outcome of the inversion, and contradicts the premises of this work. In the next section, we introduce our proposed method and discuss how to address the computational challenges associated with optimizing the CNN's randomly initialized weights, while keeping the advantages of the deep-prior based imaging.

3.5 Imaging with weak deep prior

The deep-prior based imaging problem can equivalently be casted as the following constrained optimization problem:

$$\begin{aligned} &\mathop{\arg\min}_{\delta\mathbf{m},\, \mathbf{w}} \frac{1}{2\sigma^2} \left[\sum_{i=1}^{N} \| \delta\mathbf{d}_i - \mathbf{J}(\mathbf{m}_0, \mathbf{q}_i)\delta\mathbf{m} \|_2^2 \right] + \frac{\lambda^2}{2} \|\mathbf{w}\|_2^2 \\ &\text{subject to} \quad \delta\mathbf{m} = g(\mathbf{z}, \mathbf{w}), \end{aligned} \tag{3.3}$$

where we restrict the feasible model to the output of $g(\mathbf{z}, \mathbf{w})$. To address the computational challenge associated with deep-prior based imaging, we propose to relax the constraint in problem 3.3 and let $\delta\mathbf{m}$ be a random variable distributed according to a Gaussian distribution centered at $g(\mathbf{z}, \mathbf{w})$ with covariance matrix $\gamma^{-2}\mathbf{I}$. We denote the defined prior on $\delta\mathbf{m}$ as the *weak* deep prior. By decoupling the forward operator and the CNN weights, observed data becomes conditionally independent from $\mathbf{w}$, given $\delta\mathbf{m}$. We can write the joint posterior distribution for $(\delta\mathbf{m}, \mathbf{w})$ using the defined prior as follows:

$$
\begin{aligned}
& p_{\text{post}}\left(\delta\mathbf{m}, \mathbf{w} \,|\, \left\{\delta\mathbf{d}_i\right\}_{i=1}^{N}\right) \\
& \quad \propto \left[\prod_{i=1}^{N} p_{\text{like}}\left(\delta\mathbf{d}_i | \delta\mathbf{m}\right)\right] p_{\text{weak}}\left(\delta\mathbf{m} | \mathbf{w}\right) p_w(\mathbf{w}), \\
& \text{where} \quad p_{\text{weak}}\left(\delta\mathbf{m} | \mathbf{w}\right) = \mathrm{N}(\delta\mathbf{m} | g(\mathbf{z}, \mathbf{w}), \gamma^{-2}\mathbf{I}).
\end{aligned}
\tag{3.4}
$$

In Equation 3.4, $p_{\text{weak}}\left(\delta\mathbf{m}|\mathbf{w}\right)$ denotes the weak deep prior, which is equivalent to a Gaussian distribution centered at $g(\mathbf{z}, \mathbf{w})$ with covariance matrix $\gamma^{-2}\mathbf{I}$. γ is a hyperparameter that needs to be tuned. We solve the imaging with weak deep prior problem by minimizing the negative log-posterior defined in Equation 3.4 as follows:

$$
\begin{aligned}
\widehat{\delta\mathbf{m}}_{\text{weak}}, \widehat{\mathbf{w}}_{\text{weak}} = \underset{\delta\mathbf{m},\, \mathbf{w}}{\arg\min} \; & \Bigg[\frac{1}{2\sigma^2} \sum_{i=1}^{N} \left\| \delta\mathbf{d}_i - \mathbf{J}(\mathbf{m}_0, \mathbf{q}_i)\delta\mathbf{m} \right\|_2^2 \\
& + \frac{\gamma^2}{2} \left\| \delta\mathbf{m} - g(\mathbf{z}, \mathbf{w}) \right\|_2^2 + \frac{\lambda^2}{2} \left\| \mathbf{w} \right\|_2^2 \Bigg]
\end{aligned}
\tag{3.5}
$$

where $\widehat{\delta\mathbf{m}}_{\text{weak}}$ and $\widehat{\mathbf{w}}_{\text{weak}}$ are the obtained reflectivity and CNN weights by solving the imaging with weak deep prior problem. We consider $\widehat{\delta\mathbf{m}}_{\text{weak}}$ as the final estimate in this approach. When $\gamma \to \infty$, the solution to problem 3.5 is the same as the solution to problem 3.3.

In formulation above, updating the parameters $\mathbf{w}$ does not involve the action of the forward operator, hence, weights of the CNN can be quickly and independently updated. Moreover, the optimization problem 3.5 offers flexibility to impose any intersection of physical or handcrafted hard constraints, $\mathcal{C}$, by limiting the search space to $\delta\mathbf{m} \in \mathcal{C}$ while minimizing the objective with respect to $\delta\mathbf{m}$, using standard constrained optimization techniques [20]. In a similar fashion, [8] use a Total-Variation constraint in the context of seismic imaging to jointly solve the imaging problem and train a generative model capable of directly sampling the posterior using the Expectation-Maximization method. As the main contribution of this work, we choose not to utilize hard constraints and focus on the computational aspect of the weak deep prior.

3.6 Algorithm and implementation details

To limit the computational cost—i.e., number of wave-equation solves, we use stochastic optimization algorithms to solve the optimization problems 3.3 and 3.5. We approximate the negative-log likelihood term (see Equation 3.1) using a single simultaneous source, made of a Gaussian weighted source aggregate. While we could use stochastic gradient descent algorithm [SGD, 21], we avoid it because of several challenges associated with it. For example, even though SGD's "noisy" (approximate) gradient is an unbiased estimate of true gradient, its variance is proportional to square of the step size. Therefore, choosing the step size is a trade-off between convergence speed and accuracy. Additionally, SGD updates different components of the unknown with the same step size—i.e., no preconditioning, which is not desirable when the objective has varying sensitivity with respect to different components of the unknowns. Various stochastic optimization algorithms to some extent address these issues by diagonally weighting the gradient by the norm of the past gradients [22] or the (weighted) mean of past squared gradients [23]. We use Adagrad [22] with step size 2×10^{-3} to update $\delta \mathbf{m}$ while estimating $\widehat{\delta \mathbf{m}}_{\text{MLE}}$, and when solving optimization problem 3.5. To update $\mathbf{w}$, either in optimization problem 3.3 or 3.5, we use RMSprop [23] with step size 10^{-3}. We set the step sizes by extensive hyper-parameter tuning. In Algorithm 2, which summarizes our proposed approach, `Adagrad` and `RMSprop` are optimization subroutines that given the objective value and the step size, provide an update for $\delta \mathbf{m}$ and $\mathbf{w}$, respectively.

As mentioned before, the weak deep prior allows for fast updates of the CNN weights (see the inner loop in lines 5 – 8 of Algorithm 2). However, choosing the number of updates for $\mathbf{w}$ per each $\delta \mathbf{m}$ update is a trade-off between reducing computational cost (many $\mathbf{w}$ updates) and preserving the the deep prior advantages (maintained by employing

Algorithm 2 Seismic imaging with weak deep prior.

```
Input:
    $\mathbf{z} \sim \mathrm{N}(\mathbf{0}, \mathbf{I})$: fixed input to the CNN
    $\lambda,\ \gamma$: trade-off parameters
    $\sigma^2$: estimated noise variance
    $T$: stochastic optimization steps for $\delta\mathbf{m}$
    $K$: inner loop stochastic optimization steps for $\mathbf{w}$
    $\eta$, $\tau$: step sizes to update $\delta\mathbf{m}$ and $\mathbf{w}$, respectively
    $\{\delta\mathbf{d}_i, \delta\mathbf{q}_i\}_{i=1}^N$: observed data and source signatures
    $\mathbf{m}_0$: smooth background squared-slowness model
    Adagrad: Adagrad algorithm to update $\delta\mathbf{m}$
    RMSprop: RMSprop algorithm to update $\mathbf{w}$
Initialization:
    Randomly initialize CNN parameters, $\mathbf{w}$
    $\delta\mathbf{m} = \mathbf{0}$
for $t = 1$ to $T$ do
    Randomly sample $(\delta\mathbf{d}, \mathbf{q})$ from $\{\delta\mathbf{d}_i, \mathbf{q}_i\}_{i=1}^N$
    $\mathcal{L}(\delta\mathbf{m}) = \frac{N}{2\sigma^2}\|\delta\mathbf{d} - \mathbf{J}(\mathbf{m}_0, \mathbf{q})\delta\mathbf{m}\|_2^2 + \frac{\gamma^2}{2}\|\delta\mathbf{m} - g(\mathbf{z}, \mathbf{w})\|_2^2$
    $\delta\mathbf{m} \leftarrow$ Adagrad $(\mathcal{L}(\delta\mathbf{m}), \eta)$
    for $k = 1$ to $K$ do
        $\mathcal{L}(\mathbf{w}) = \frac{\gamma^2}{2}\|\delta\mathbf{m} - g(\mathbf{z}, \mathbf{w})\|_2^2 + \frac{\lambda^2}{2}\|\mathbf{w}\|_2^2$
        $\mathbf{w} \leftarrow$ RMSprop $(\mathcal{L}(\mathbf{w}), \tau)$
    end for
end for
Output: $\delta\mathbf{m}$
```

several $\mathbf{w}$ updates). In the extreme case, if we update $\mathbf{w}$ once per $\delta\mathbf{m}$ update, there is no computational gain compared to the deep-prior based approach. On the other hand, if we solve for $\mathbf{w}$ after each update to $\delta\mathbf{m}$—i.e., $\|\delta\mathbf{m} - g(\mathbf{z}, \mathbf{w})\|_2^2 \simeq 0$, the CNN has almost no effect in the next update for $\delta\mathbf{m}$. To strike a balance between the number of updates to $\delta\mathbf{m}$ and $\mathbf{w}$, we choose to alternatingly take one gradient step for $\delta\mathbf{m}$ and ten gradient steps for $\mathbf{w}$.

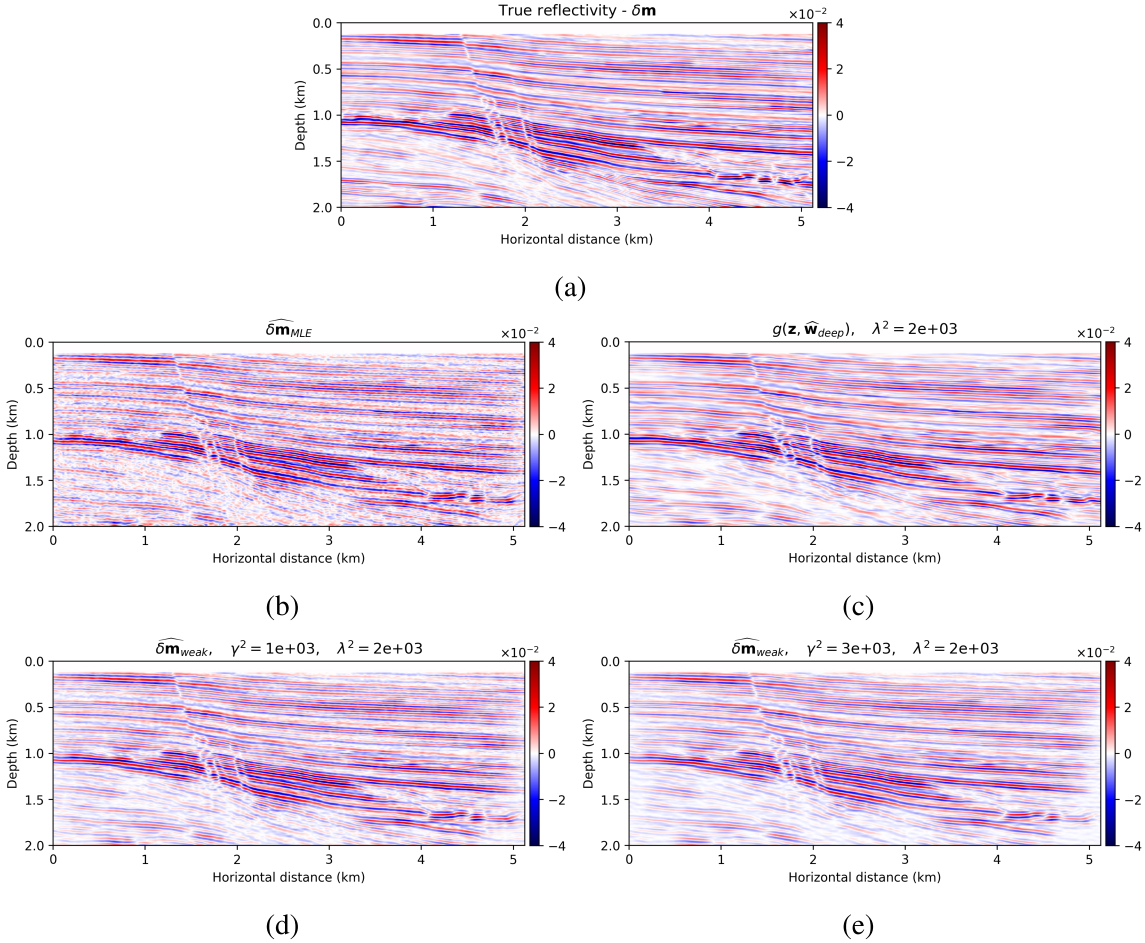
We use Devito [24, 25] to compute matrix-free actions of the linearized Born scattering operator and its adjoint. By integrating these operators into PyTorch, we are able to solve the optimization problems 3.3 and 3.5 with automatic differentiation. We follow [1] for the CNN architecture. We provide more details regarding to our implementation on GitHub.

3.7 Numerical experiments

We compare the seismic images obtained by solving problems 3.3 and 3.5, when applied to a “quasi” real field data example consisting of a 2D portion of the Kirchoff time migrated 3D Parihaka dataset (see Figure 3.1a). These imaging results are set as the ground truth for the experiment here discussed. Synthetic data is obtained by applying the linearized Born scattering operator to this “true” reflectivity image. The dataset includes 205 shot records sampled with a source spacing of $25\,\mathrm{m}$ and 1.5 seconds recording time. There are 410 fixed receivers sampled at $12.5\mathrm{m}$ spread across the survey area. The source is a Ricker wavelet with a central frequency of $30\,\mathrm{Hz}$. To demonstrate the regularization effect of our method, we add a significant amount of noise to the shot records, yielding a low signal-to-noise ratio of the “observed” data of $-18.01\,\mathrm{dB}$. To limit the computational costs, we mix the shot records according to normally distributed source encodings. By conducting extensive parameter tuning, we set $\lambda^2 = 2 \times 10^3$ (Equations 3.3 and 3.5) throughout all experiments. We also set $\sigma^2 = 0.01$ (Equations 3.1, 3.3, and 3.5), which is equal to the variance of the measurement noise. To provide evidence regarding to the computational feasibility of the weak deep prior formulation, we fix the number of passes over the dataset—i.e., we use roughly twice the computational cost of reverse-time migration (computational budget for large-scale least-squares imaging) However, as mentioned before, the deep prior formulation requires more iterations to generate a reasonable image. We use 15 passes over the dataset to solve problem 3.3—i.e., to compute $\widehat{\mathbf{w}}_{\text{deep}}$. Note that taking one gradient step for $\delta\mathbf{m}$ takes roughly 60 times more time than one update of $\mathbf{w}$, without GPU acceleration. Therefore, we neglect the CNN weights update times in our comparisons.

The imaging results are included in Figure 7.4.2. Figure 3.1a indicates the reflectivity that we have used to generate linearized data. Figure 3.1b is the MLE image—i.e., conventional least-squares reverse-time migration, $\widehat{\delta\mathbf{m}}_{\text{MLE}}$, obtained by minimizing Equation 3.1 with Adagrad for two passes over the dataset. Figures 3.1c shows the the deep-prior based image, $g(\mathbf{z}, \widehat{\mathbf{w}}_{\text{deep}})$, computed by running RMSprop for 15 passes over the dataset. Figures 3.1d and 3.1e show the obtained results using the proposed method by solving problem 3.5 for two passes over the dataset using values $\gamma = 10^3$ and $\gamma = 3 \times 10^3$, respectively.

We make the following observations. As expected, Figure 3.1b contains imaging artifacts since no prior regularization is in effect. Although the deep prior has been successful in generating a realistic result (Figures 3.3), computing the solution required 15 passes over the source experiments, which is practically not attainable for larger problems. The solution to the weak deep prior imaging problem, with $\gamma = 10^3$ (Equation 3.5), generates a seismic image with considerably less artifacts compared to MLE (compare Figures 3.1b and 3.1d), using the same number of wave-equation solves. By comparing Figure 3.1d with the image obtained by deep prior based imaging (Figure 3.1c), we observe that the proposed method is able to provide the benefits of deep prior while remaining computationally feasible. However, Figure 3.1d is slightly less smooth compared to the true reflectivity (Figure 3.1a) and deep prior based recovery (Figure 3.1d). We can increase the the deep prior penalty by increasing γ to 3×10^3 (Figure 3.1e) to get an image with less artifacts compared to Figure 3.1d. The cost of heavier penalty is amplitude underestimation compared to the



(a)

(b)

(c)

(d)

(e)

Figure 3.1: Imaging with the proposed method. a) True model. b) $\widehat{\delta\mathbf{m}}_{\text{MLE}}$. c) $g(\mathbf{z}, \widehat{\mathbf{w}}_{\text{deep}})$. d, e) $\widehat{\delta\mathbf{m}}_{\text{weak}}$, with $\gamma = 10^3$ and 3×10^3, respectively.

deep-prior result (Figure 3.1c).

3.8 Conclusions

The proposed method is an alternative to classical constrained optimization, where hand-crafted regularization is considered instead. While practical and ubiquitous, the latter approach is based on heavy-handed assumptions, which inevitably leaves a strong imprint on the final result. Conversely, constraints by deep priors only requires an uninformative Gaussian prior on the network weights. While deep priors have been recently proven successful for many imaging problems, a naive implementation for seismic imaging, which involves lengthy wave-equation solvers, leads to a computationally expensive scheme. By relaxing the deep prior, we decouple model and network updates when optimizing, hence a relatively cheap training phase. As verified by our numerical experiment, we are still able to resolve the imaging artifacts present in conventional least-squares imaging when data is contaminated by strong noise. Compared to reverse-time migration, the deep weak prior approach requires twice its computational cost, an affordable computational budget in large-scale imaging problems.

3.9 References

[1] V. Lempitsky, A. Vedaldi, and D. Ulyanov, “Deep Image Prior,” in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2018, pp. 9446–9454 (pages 82, 84, 89).

[2] Z. Cheng, M. Gadelha, S. Maji, and D. Sheldon, “A Bayesian Perspective on the Deep Image Prior,” in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2019, pp. 5443–5451 (pages 82, 84).

[3] M. Gadelha, R. Wang, and S. Maji, “Shape reconstruction using differentiable projections and deep priors,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 22–30 (pages 82, 84).

[4] Q. Liu, L. Fu, and M. Zhang, “Deep-seismic-prior-based reconstruction of seismic data using convolutional neural networks,” *arXiv preprint arXiv:1911.08784*, 2019 (pages 82, 84).

[5] Y. Wu and G. A. McMechan, “Parametric convolutional neural network-domain full-waveform inversion,” *GEOPHYSICS*, vol. 84, no. 6, R881–R896, 2019 (pages 82, 84).

[6] Y. Shi, X. Wu, and S. Fomel, “Deep learning parameterization for geophysical inverse problems,” in *SEG 2019 Workshop: Mathematical Geophysics: Traditional vs Learning, Beijing, China, 5-7 November 2019*, Society of Exploration Geophysicists, 2020, pp. 36–40 (pages 82–84).

[7] A. Siahkoohi, G. Rizzuti, and F. J. Herrmann, “A deep-learning based bayesian approach to seismic imaging and uncertainty quantification,” in *82nd EAGE Conference and Exhibition*, Extended Abstracts, 2020 (pages 82–84).

[8] F. J. Herrmann, A. Siahkoohi, and G. Rizzuti, “Learned imaging with constraints and uncertainty quantification,” in *Neural Information Processing Systems (NeurIPS) 2019 Deep Inverse Workshop*, Dec. 2019 (pages 82, 86).

[9] O. Ovcharenko, V. Kazei, M. Kalita, D. Peter, and T. A. Alkhalifah, “Deep learning for low-frequency extrapolation from multi-offset seismic data,” *GEOPHYSICS*, vol. 84, no. 6, R989–R1001, Nov. 2019 (page 82).

[10] G. Rizzuti, A. Siahkoohi, and F. J. Herrmann, “Learned iterative solvers for the Helmholtz equation,” *81st EAGE Conference and Exhibition 2019*, 2019 (page 82).

[11] A. Siahkoohi, R. Kumar, and F. J. Herrmann, “Deep-learning based ocean bottom seismic wavefield recovery,” in *SEG Technical Program Expanded Abstracts 2019*, Aug. 2019, pp. 2232–2237 (page 82).

[12] A. Siahkoohi, M. Louboutin, and F. J. Herrmann, “The importance of transfer learning in seismic modeling and imaging,” *GEOPHYSICS*, vol. 84, no. 6, A47–A52, Nov. 2019 (page 82).

[13] A. Siahkoohi, D. J. Verschuur, and F. J. Herrmann, “Surface-related multiple elimination with deep learning,” in *SEG Technical Program Expanded Abstracts 2019*, Aug. 2019, pp. 4629–4634 (page 82).

[14] H. Sun and L. Demanet, “Extrapolated full waveform inversion with convolutional neural networks,” in *SEG Technical Program Expanded Abstracts 2019*, 2019 (page 82).

[15] Z. Zhang and T. Alkhalifah, “Regularized elastic full waveform inversion using deep learning,” *GEOPHYSICS*, vol. 84, no. 5, 1SO–Z28, Sep. 2019 (page 82).

[16] Veritas, “Parihaka 3D Marine Seismic Survey - Acquisition and Processing Report,” New Zealand Petroleum & Minerals, Wellington, Tech. Rep. New Zealand Petroleum Report 3460, 2005 (page 83).

[17] WesternGeco., “Parihaka 3D PSTM Final Processing Report,” New Zealand Petroleum & Minerals, Wellington, Tech. Rep. New Zealand Petroleum Report 4582, 2012 (page 83).

[18] A. Tarantola, *Inverse problem theory and methods for model parameter estimation*. SIAM, 2005, ISBN: 978-0-89871-572-9 (page 83).

[19] M. Welling and Y. W. Teh, “Bayesian Learning via Stochastic Gradient Langevin Dynamics,” in *Proceedings of the 28th International Conference on Machine Learning*, Bellevue, Washington, USA: Omnipress, 2011, pp. 681–688, ISBN: 9781450306195 (page 84).

[20] B. Peters, B. R. Smithyman, and F. J. Herrmann, “Projection methods and applications for seismic nonlinear inverse problems with multiple constraints,” *GEOPHYSICS*, vol. 84, no. 2, R251–R269, 2019 (page 86).

[21] H. E. Robbins, “A stochastic approximation method,” 2007 (page 87).

[22] J. Duchi, E. Hazan, and Y. Singer, “Adaptive subgradient methods for online learning and stochastic optimization,” *Journal of Machine Learning Research*, vol. 12, no. null, pp. 2121–2159, Jul. 2011 (page 87).

[23] T. Tieleman and G. Hinton, *Lecture 6.5-RMSprop: Divide the gradient by a running average of its recent magnitude*, https://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf, 2012 (page 87).

[24] F. Luporini *et al.*, “Architecture and performance of devito, a system for automated stencil computation,” *CoRR*, vol. abs/1807.03032, Jul. 2018. arXiv: 1807.03032 (page 89).

[25] 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 (page 89).

CHAPTER 4
PRECONDITIONED TRAINING OF NORMALIZING FLOWS FOR VARIATIONAL INFERENCE IN INVERSE PROBLEMS

4.1 Summary

Obtaining samples from the posterior distribution of inverse problems with expensive forward operators is challenging especially when the unknowns involve the strongly heterogeneous Earth. To meet these challenges, we propose a preconditioning scheme involving a conditional normalizing flow (NF) capable of sampling from a low-fidelity posterior distribution directly. This conditional NF is used to speed up the training of the high-fidelity objective involving minimization of the Kullback-Leibler divergence between the predicted and the desired high-fidelity posterior density for indirect measurements at hand. To minimize costs associated with the forward operator, we initialize the high-fidelity NF with the weights of the pretrained low-fidelity NF, which is trained beforehand on available model and data pairs. Our numerical experiments, including a 2D toy and a seismic compressed sensing example, demonstrate that thanks to the preconditioning considerable speed-ups are achievable compared to training NFs from scratch.

4.2 Introduction

Our aim is to perform approximate Bayesian inference for inverse problems characterized by computationally expensive forward operators, $F : \mathcal{X} \to \mathcal{Y}$, with a data likelihood, $\pi_{\text{like}}(\boldsymbol{y} \mid \boldsymbol{x})$:

$$\boldsymbol{y} = F(\boldsymbol{x}) + \boldsymbol{\epsilon}, \tag{4.1}$$

where $\boldsymbol{x} \in \mathcal{X}$ is the unknown model, $\boldsymbol{y} \in \mathcal{Y}$ the observed data, and $\boldsymbol{\epsilon} \sim \mathrm{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$ the measurement noise. Given a prior density, $\pi_{\text{prior}}(\boldsymbol{x})$, variational inference [VI, 1] based on normalizing flows [NFs, 2] can be used where the Kullback-Leibler (KL) divergence is minimized between the predicted and the target—i.e., *high-fidelity*, posterior density $\pi_{\text{post}}(\boldsymbol{x} \mid \boldsymbol{y})$ [3, 4, 5, 6, 7]:

$$\min_{\theta} \mathbb{E}_{\boldsymbol{z}\sim\pi_z(\boldsymbol{z})}\left[\frac{1}{2\sigma^2}\left\|F\big(T_\theta(\boldsymbol{z})\big)-\boldsymbol{y}\right\|_2^2-\log\pi_{\text{prior}}\big(T_\theta(\boldsymbol{z})\big)-\log\Big|\det\nabla_z T_\theta(\boldsymbol{z})\Big|\right]. \quad (4.2)$$

In the above expression, $T_\theta : \mathcal{Z}_x \to \mathcal{X}$ denotes a NF with parameters $\boldsymbol{\theta}$ and a Gaussian latent variable $\boldsymbol{z} \in \mathcal{Z}_x$. The above objective consists of the data likelihood term, regularization on the output of the NF, and a log-determinant term that is related to the entropy of the NF output. The last term is necessarily to prevent the output of the NF from collapsing on the maximum a posteriori estimate. For details regarding the derivation of the objective in Equation (4.2), we refer to Appendix A. During training, we replace the expectation by Monte-Carlo averaging using mini-batches of $\boldsymbol{z}$. After training, samples from the approximated posterior, $\pi_\theta(\boldsymbol{x} \mid \boldsymbol{y}) \approx \pi_{\text{post}}(\boldsymbol{x} \mid \boldsymbol{y})$, can be drawn by evaluating $T_\theta(\boldsymbol{z})$ for $\boldsymbol{z} \sim \pi_z(\boldsymbol{z})$ [4]. It is important to note that Equation (4.2) trains a NF specific to the observed data $\boldsymbol{y}$. While the above VI formulation in principle allows us to train a NF to generate samples from the posterior given a single observation $\boldsymbol{y}$, this variational estimate requires access to a prior density, and the training calls for repeated evaluations of the forward operator, F, as well as the adjoint of its Jacobian, $\nabla F^\top$. As in multi-fidelity Markov chain Monte Carlo (MCMC) sampling [8], the costs associated with the forward operator may become prohibitive even though VI-based methods are known to have computational advantages over MCMC [9].

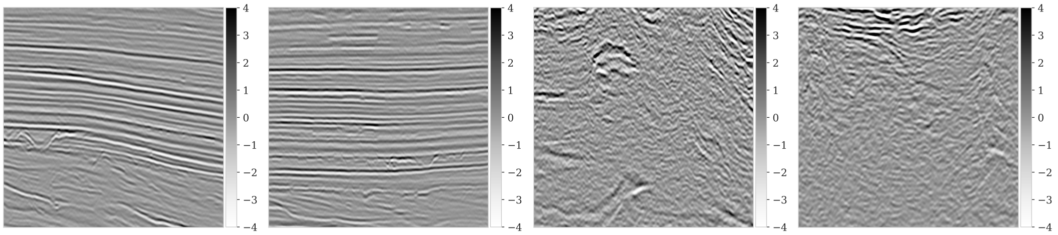
Aside from the above computational considerations, reliance on having access to a prior may be problematic especially when dealing with images of the Earth's subsurface, which are the result of complex geological processes that do not lend themselves to be easily captured by hand-crafted priors. Under these circumstances, data-driven priors—or even better data-driven posteriors obtained by training over model and data pairs sampled from the joint distribution, $\widehat{\pi}_{y,x}(\boldsymbol{y}, \boldsymbol{x})$—are preferable. More specifically, we follow [4], [10], and [11], and formulate the objective function in terms of a block-triangular conditional NF, $G_{\phi} : \mathcal{Y} \times \mathcal{X} \to \mathcal{Z}_y \times \mathcal{Z}_x$, with latent space $\mathcal{Z}_y \times \mathcal{Z}_x$:

$$
\begin{aligned}
&\min_{\phi} \mathbb{E}_{\boldsymbol{y},\boldsymbol{x} \sim \widehat{\pi}_{y,x}(\boldsymbol{y},\boldsymbol{x})} \left[\frac{1}{2} \left\| G_{\phi}(\boldsymbol{y}, \boldsymbol{x}) \right\|^2 - \log \left| \det \nabla_{y,x} G_{\phi}(\boldsymbol{y}, \boldsymbol{x}) \right| \right], \\
&\text{where} \quad G_{\phi}(\boldsymbol{y}, \boldsymbol{x}) = \begin{bmatrix} G_{\phi_y}(\boldsymbol{y}) \\ G_{\phi_x}(\boldsymbol{y}, \boldsymbol{x}) \end{bmatrix}, \ \boldsymbol{\phi} = \left\{ \boldsymbol{\phi}_y, \boldsymbol{\phi}_x \right\}.
\end{aligned}
\tag{4.3}
$$

Thanks to the block-triangular structure of G_{ϕ}, samples of the approximated posterior, $\pi_{\phi}(\boldsymbol{x} \mid \boldsymbol{y}) \approx \pi_{\text{post}}(\boldsymbol{x} \mid \boldsymbol{y})$ can be drawn by evaluating $G_{\phi_x}^{-1}(G_{\phi_y}(\boldsymbol{y}), \boldsymbol{z})$ for $\boldsymbol{z} \sim \pi_z(\boldsymbol{z})$ [12]. Unlike the objective in Equation (4.2), training G_{ϕ} does not involve multiple evaluations of F and $\nabla F^{\top}$, nor does it require specifying a prior density. However, its success during inference heavily relies on having access to training pairs from the joint distribution, $\boldsymbol{y}, \boldsymbol{x} \sim \widehat{\pi}_{y,x}(\boldsymbol{y}, \boldsymbol{x})$. Unfortunately, unlike medical imaging, where data is abundant and variability among patients is relatively limited, samples from the joint distribution are unavailable in geophysical applications. Attempts have been made to address this lack of training pairs including the generation of simplified artificial geological models [13], but these approaches cannot capture the true heterogeneity exhibited by the Earth's subsurface. This is illustrated in Figure 4.1, which shows several true seismic image patches drawn from the Parihaka dataset. Even though samples are drawn from a single data set, they illustrate significant differences between shallow (Figures 4.1a and 4.1b) and deeper (Figures 4.1c and 4.1d) sections.

To meet the challenges of computational cost, heterogeneity and lack of access to training pairs, we propose a preconditioning scheme where the two described VI methods are combined to:

1. take maximum advantage of available samples from the joint distribution $\widehat{\pi}_{y,x}(\boldsymbol{y}, \boldsymbol{x})$, to pretrain G_{ϕ} by minimizing Equation (4.3). We only incur these costs once, by



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

Figure 4.1: Subsurface images estimated from a real seismic survey, indicating strong heterogeneity between (a), (b) shallow and (c), (d) deep parts of the same survey area.

training this NF beforehand. As these samples typically come from a different (neighboring) region, they are considered as low-fidelity;

2. exploit the invertibility of $G_{\phi_x}(\boldsymbol{y}, \,\cdot\,)$, which gives us access to a low-fidelity posterior density, $\pi_\phi(\boldsymbol{x} \mid \boldsymbol{y})$. For a given $\boldsymbol{y}$, this trained (conditional) prior can be used in Equation (4.2);

3. initialize T_θ with weights from the pretrained $G_{\phi_x}^{-1}$. This initialization can be considered as an instance of transfer learning [14], and we expect a considerable speed-up when solving Equation (4.2). This is important since it involves inverting F, which is computationally expensive.

4.3 Related work

In the context of variational inference for inverse problems with expensive forward operators, [15] train a generative model to sample from the posterior distribution, given indirect measurements of the unknown model. This approach is based on an Expectation Maximization technique, which infers the latent representation directly instead of using an inference encoding model. While that approach allows for inclusion of hand-crafted priors, capturing the posterior is not fully developed. Like [10], we also use a block-triangular map between the joint model and data distribution and their respective latent spaces to train a network to generate samples from the conditional posterior. By imposing an additional monotonicity constraint, these authors train a generative adversarial network [GAN, 16] to directly sample from the posterior distribution. To allow for scalability to large scale problems, we work with NFs instead, because they allow for more memory efficient training [17, 18, 19, 20]. Our contribution essentially corresponds to a reformulation of [21] and [8]. In that work, transport-based maps are used as non-Gaussian proposal distributions during MCMC sampling. As part of the MCMC, this transport map is then tuned to match the target density, which improves the efficiency of the sampling. [8] extend this approach by proposing a preconditioned MCMC sampling technique where a transport-map trained to sample from a low-fidelity posterior distribution is used as a preconditioner. This idea of multi-fidelity preconditioned MCMC inspired our work where we setup a VI objective instead. We argue that this formulation can be faster and may be easier to scale to large-scale Bayesian inference problems [9].

Finally, there is a conceptional connection between our work and previous contributions on amortized variational inference [22], including an iterative refinement step [23, 24, 25, 26]. Although similar in spirit, our approach is different from these attempts because we adapt the weights of our conditional generative model to account for the inference errors instead of correcting the inaccurate latent representation of the out-of-distribution data.

4.4 Multi-fidelity preconditioning scheme

For an observation $\boldsymbol{y}$, we define a NF $T_{\phi_x} : \mathcal{Z}_x \to \mathcal{X}$ as

$$T_{\phi_x}(\boldsymbol{z}) := G_{\phi_x}^{-1}(G_{\phi_y}(\boldsymbol{y}), \boldsymbol{z}), \tag{4.4}$$

where we obtain $\boldsymbol{\phi} = \{\boldsymbol{\phi}_y, \boldsymbol{\phi}_x\}$ by training G_ϕ through minimizing the objective function in Equation (4.3). To train G_ϕ, we use available low-fidelity training pairs $\boldsymbol{y}, \boldsymbol{x} \sim \widehat{\pi}_{y,x}(\boldsymbol{y}, \boldsymbol{x})$. We perform this training phase beforehand, similar to the pretraining phase during transfer learning [14]. Thanks to the invertibility of G_ϕ, it provides an expression for the posterior. We refer to this posterior as low-fidelity because the network is trained with often scarce and out-of-distribution training pairs. Because the Earth's heterogeneity does not lend itself to be easily captured by hand-crafted priors, we argue that this NF can still serve as a (conditional) prior in Equation (4.2):

$$\pi_{\text{prior}}(\boldsymbol{x}) := \pi_{\boldsymbol{z}}\big(G_{\phi_x}(\boldsymbol{y}, \boldsymbol{x})\big) \, \Big| \det \nabla_x G_{\phi_x}(\boldsymbol{y}, \boldsymbol{x}) \Big|. \tag{4.5}$$

To train the high-fidelity NF given observed data $\boldsymbol{y}$, we minimize the KL divergence between the predicted and the high-fidelity posterior density, $\pi_{\text{post}}(\boldsymbol{x} \mid \boldsymbol{y})$ [3, 4]

$$\min_{\phi_x} \, \mathbb{E}_{\boldsymbol{z} \sim \pi_z(\boldsymbol{z})} \left[\frac{1}{2\sigma^2} \left\| F\big(T_{\phi_x}(\boldsymbol{z})\big) - \boldsymbol{y} \right\|_2^2 - \log \pi_{\text{prior}}\big(T_{\phi_x}(\boldsymbol{z})\big) - \log \Big| \det \nabla_z T_{\phi_x}(\boldsymbol{z}) \Big| \right], \tag{4.6}$$

where the prior density of Equation (4.5) is used. Notice that this minimization problem differs from the one stated in Equation (4.2). Here, the optimization involves "fine-tuning" the low-fidelity network parameters ϕ_x introduced in Equation (4.3). Moreover, this low-fidelity network is also used as a prior. While other choices exist for the latter—e.g., it could be replaced or combined with a hand-crafted prior in the form of constraints [27] or by a separately trained data-driven prior [13], using the low-fidelity posterior as a prior (cf. Equation (4.5)) has certain distinct advantages. First, it removes the need for training a separate data-driven prior model. Second, use of the low-fidelity posterior may be more informative [28] than its unconditional counterpart because it is conditioned by the observed data $\boldsymbol{y}$. In addition, our multi-fidelity approach has strong connections with online variational Bayes [29] where data arrives sequentially and previous posterior approximates are used as priors for subsequent approximations.

In summary, the problem in Equation (4.6) can be interpreted as an instance of transfer learning [14] for conditional NFs. This formulation is particularly useful for inverse problems with expensive forward operators, where access to high fidelity training samples, i.e. samples from the target distribution, is limited. In the next section, we present two numerical experiments designed to show the speed-up and accuracy gained with our proposed multi-fidelity formulation.

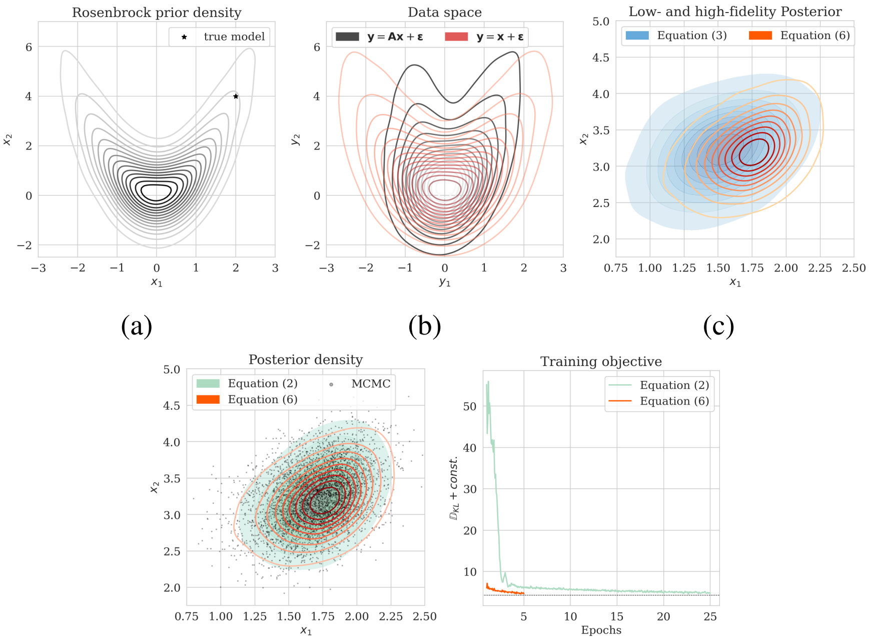
4.5 Numerical experiments

We present two synthetic examples aimed at verifying the anticipated speed-up and increase in accuracy of the predicted posterior density via our multi-fidelity preconditioning scheme. The first example is a two-dimensional problem where the posterior density can be accurately and cheaply sampled via MCMC. The second example demonstrates the effect of the preconditioning scheme in a seismic compressed sensing [30, 31] problem. Details regarding training hyperparameters and the NF architectures are included in Appendix B. Code to reproduce our results are made available on GitHub. Our implementation relies on InvertibleNetworks.jl [32], a recently-developed memory-efficient framework for training invertible networks in the Julia programming language.

4.5.1 2D toy example

To illustrate, the advantages of working with our multi-fidelity scheme, we consider the 2D Rosenbrock distribution, $\pi_{\text{prior}}(\boldsymbol{x}) \propto \exp\left(-\frac{1}{2}x_1^2 - (x_2 - x_1^2)^2\right)$, plotted in Figure 4.2a. High-fidelity data $\boldsymbol{y} \in \mathbb{R}^2$ are generated via $\boldsymbol{y} = \boldsymbol{A}\,\boldsymbol{x} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}, 0.4^2\boldsymbol{I})$ and $\boldsymbol{A} \in \mathbb{R}^{2\times 2}$ is a forward operator. To control the discrepancy between the low- and high-fidelity samples, we set $\boldsymbol{A}$ equal to $\bar{\boldsymbol{A}}\ /\ \rho(\bar{\boldsymbol{A}})$, where $\rho(\cdot)$ is the spectral radius of $\bar{\boldsymbol{A}} = \boldsymbol{\Gamma} + \gamma\boldsymbol{I}$, $\boldsymbol{\Gamma} \in \mathbb{R}^{2\times 2}$ has independent and normally distributed entries, and $\gamma = 3$. By choosing smaller values for γ, we make $\boldsymbol{A}$ more dissimilar to the identity matrix, therefore increasing the discrepancy between the low- and high-fidelity posterior.

Figure 4.2b depicts the low- (purple) and high-fidelity (red) data densities. The dark star represents the unknown model. Low-fidelity data samples are generated with the identity as the forward operator. During the pretraining phase conducted beforehand, we minimize the objective function in Equation (4.3) for 25 epochs.



(a) (b) (c)

(d) (e)

Figure 4.2: (a) Prior, (b) low- and high-fidelity data, (c) low- (blue) and high-fidelity (orange) approximated posterior densities, and (d) approximated posterior densities via MCMC (dark circles), and objectives in Equations (4.2) in green and (4.6) in orange. (e) Objective value during training via Equations (4.2) in green and (4.6) in orange.

The pretrained low-fidelity posterior is subsequently used to precondition the minimization of (4.6) given observed data $\boldsymbol{y}$. The resulting low- and high-fidelity estimates for the posterior as plotted in Figure 4.2c differ significantly. In Figure (4.2d), the accuracy of the proposed method is verified by comparing the approximated high-fidelity posterior density (orange contours) with the approximation (in green) obtained by minimizing the objective of Equation (4.2). The overlap between the orange contours and the green shaded area confirms consistency between the two methods. To assess the accuracy of the estimated densities themselves, we also include samples from the posterior (dark circles) obtained via stochastic gradient Langevin dynamics [33], an MCMC sampling technique. As expected, the estimated posterior densities with and without the preconditioning scheme are in agreement with the MCMC samples.

Finally, to illustrate the performance our multi-fidelity scheme, we consider the convergence plot in Figure 4.2e where the objective values of Equations (4.2) and (4.6) are compared. As explained in Appendix A, the values of the objective functions correspond to the KL divergence (plus a constant) between the posterior given by Equation (4.2) and the posterior distribution obtained by our multi-fidelity approach (Equation (4.6)). As expected, the multi-fidelity objective converges much faster because of the "warm start". In addition, the updates of T_{ϕ_x} via Equation (4.6) succeed in bringing down the KL divergence within only five epochs (see orange curve), whereas it takes 25 epochs via the objective in Equation (4.2) to reach approximately the same KL divergence. This pattern holds for smaller values of γ too as indicated in Table 4.1. According to Table 4.1, the improvements by our multi-fidelity method become more pronounced if we decrease the γ. This behavior is to be expected since the samples used for pretraining are more and more out of distribution in that case. We refer to Appendix C for additional figures for different values of γ.

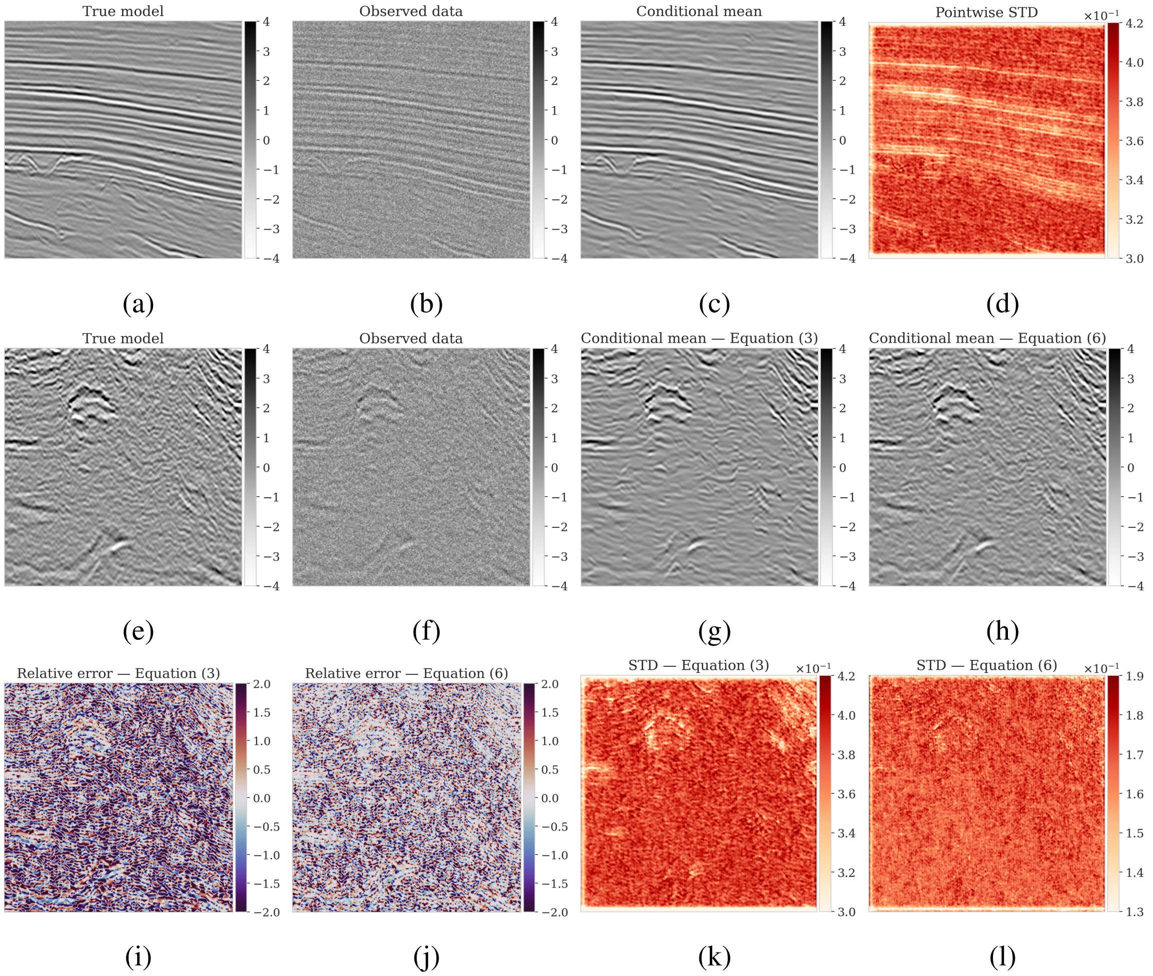
Table 4.1: The KL divergence (plus some constant, see Appendix A) between different estimated posterior densities and high-fidelity posterior distribution for different values of γ. Second column corresponds to the low-fidelity posterior estimate obtained via Equation (4.3), third column relates to the posterior estimate via Equation (4.2), and finally, the last column shows the same quantity for the posterior estimate via the multi-fidelity preconditioned scheme.

γ	Low-fidelity	Without preconditioning	With preconditioning
3	6.13	4.88	4.66
2	8.43	5.60	5.26
1	8.84	6.26	6.51
0	14.73	8.41	8.45

4.5.2 Seismic compressed sensing example

This experiment is designed to show challenges with geophysical inverse problems due to the Earth's strong heterogeneity. We consider the inversion of noisy indirect measurements of image patches $\boldsymbol{x} \in \mathbb{R}^{256 \times 256}$ sampled from deeper parts of the Parihaka seismic dataset. The observed measurements are given by $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{\epsilon}$ where $\boldsymbol{\epsilon} \sim \mathrm{N}(\boldsymbol{0}, 0.2^2\boldsymbol{I})$. For simplicity, we chose $\boldsymbol{A} = \boldsymbol{M}^T\boldsymbol{M}$ with $\boldsymbol{M}$ a compressing sensing matrix with $66.66\,\%$ subsampling. The measurement vector $\boldsymbol{y}$ corresponds to a pseudo-recovered model contaminated with noise.

To mimic a realistic situation in practice, we change the likelihood distribution by reducing the standard deviation of the noise to 0.01 in combination with using image patches sampled from the shallow part of the Parihaka dataset. As we have seen in Figure 4.1, these patches differ in texture. Given pairs $\boldsymbol{y}, \boldsymbol{x} \sim \widehat{\pi}_{y,x}(\boldsymbol{y}, \boldsymbol{x})$, we pretrain our network by minimizing Equation (4.3). Figures 4.3a and 4.3b contain a pair not used during pretraining. Estimates for the conditional mean and standard deviation obtained by drawing 1000 samples from the pretrained conditional NF for the noisy indirect measurement (Figure 4.3b) are included in Figures 4.3c and 4.3d. Both estimates exhibit the expected behavior because the examples in Figure 4.3a and 4.3b are within the distribution. As anticipated, this observation no longer holds if we apply this pretrained network to indirect data depicted in



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

(e) (f) (g) (h)

(i) (j) (k) (l)

Figure 4.3: Seismic compressed sensing. First row indicates the performance of the pretrained network on data not used during pretraining. Second row is compares the pretraining-based recovery with the accelerated scheme result. Last row compares the recovery errors and pointwise STDs for the two recoveries.

Figure 4.3f, which is sampled from the deeper part. However, these results are significantly improved when the pretrained network is fine-tuned by minimizing Equation (4.6). After fine tuning, the fine details in the image are recovered (compare Figures 4.3g and 4.3h). This improvement is confirmed by the relative errors plotted in Figures 4.3i and 4.3j, as well as by the reduced standard deviation (compare Figures 4.3k and 4.3l).

4.6 Conclusions

Inverse problems in fields such as seismology are challenging for several reasons. The forward operators are complex and expensive to evaluate numerically while the Earth is highly heterogeneous. To handle this situation and to quantify uncertainty, we propose a preconditioned scheme for training normalizing flows for Bayesian inference. The proposed scheme is designed to take full advantage of having access to training pairs drawn from a joint distribution, which for the reasons stated above is close but not equal to the actual joint distribution. We use these samples to train a normalizing flow via likelihood maximization leveraging the normalizing property. We use this pretrained low-fidelity estimate for the posterior as a prior and preconditioner for the actual variational inference on the observed data, which minimizes the Kullback-Leibler divergence between the predicted and the desired posterior density. By means of a series of examples, we demonstrate that our preconditioned scheme leads to considerable speed-ups compared to training a normalizing flow from scratch.

4.7 References

[1] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul, "An Introduction to Variational Methods for Graphical Models," *Machine Learning*, vol. 37, no. 2, pp. 183–233, 1999 (page 96).

[2] D. Rezende and S. Mohamed, "Variational inference with normalizing flows," ser. Proceedings of Machine Learning Research, vol. 37, PMLR, Jul. 2015, pp. 1530–1538 (page 97).

[3] Q. Liu and D. Wang, "Stein Variational Gradient Descent: A General Purpose Bayesian Inference Algorithm," in *Advances in Neural Information Processing Systems*, vol. 29, Curran Associates, Inc., 2016, pp. 2378–2386 (pages 97, 101).

[4] J. Kruse, G. Detommaso, R. Scheichl, and U. Köthe, *HINT: Hierarchical Invertible Neural Transport for Density Estimation and Bayesian Inference*, 2019 (pages 97, 98, 101).

[5] G. Rizzuti, A. Siahkoohi, P. A. Witte, and F. J. Herrmann, "Parameterizing uncertainty by deep invertible networks, an application to reservoir characterization," in *90th Annual International Meeting, SEG*, Sep. 2020, pp. 1541–1545 (page 97).

[6] A. Siahkoohi, G. Rizzuti, P. A. Witte, and F. J. Herrmann, "Faster Uncertainty Quantification for Inverse Problems with Conditional Normalizing Flows," Georgia Institute of Technology, Tech. Rep. TR-CSE-2020-2, Jul. 2020 (page 97).

[7] H. Sun and K. L. Bouman, "Deep probabilistic imaging: Uncertainty quantification and multi-modal solution characterization for computational imaging," *arXiv preprint arXiv:2010.14462*, 2020 (page 97).

[8] B. Peherstorfer and Y. Marzouk, "A transport-based multifidelity preconditioner for Markov chain Monte Carlo," *Advances in Computational Mathematics*, vol. 45, no. 5-6, pp. 2321–2348, 2019 (pages 97, 100).

[9] 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 (pages 97, 100).

[10] N. Kovachki, R. Baptista, B. Hosseini, and Y. Marzouk, "Conditional Sampling With Monotone GANs," *arXiv preprint arXiv:2006.06755*, 2020 (pages 98, 99).

[11] R. Baptista, O. Zahm, and Y. Marzouk, “An adaptive transport framework for joint and conditional density estimation,” *arXiv preprint arXiv:2009.10303*, 2020 (page 98).

[12] Y. Marzouk, T. Moselhy, M. Parno, and A. Spantini, “Sampling via measure transport: An introduction,” *Handbook of uncertainty quantification*, pp. 1–41, 2016 (page 98).

[13] L. Mosser, O. Dubrule, and M. Blunt, “Stochastic Seismic Waveform Inversion Using Generative Adversarial Networks as a Geological Prior,” *Mathematical Geosciences*, vol. 84, no. 1, pp. 53–79, 2019 (pages 98, 101).

[14] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, “How transferable are features in deep neural networks?” In *Proceedings of the 27th International Conference on Neural Information Processing Systems*, ser. NIPS’14, Montreal, Canada, 2014, pp. 3320–3328 (pages 99–101).

[15] F. J. Herrmann, A. Siahkoohi, and G. Rizzuti, “Learned imaging with constraints and uncertainty quantification,” in *Neural Information Processing Systems (NeurIPS) 2019 Deep Inverse Workshop*, Dec. 2019 (page 99).

[16] I. Goodfellow *et al.*, “Generative Adversarial Nets,” in *Proceedings of the 27th International Conference on Neural Information Processing Systems*, ser. NIPS’14, Montreal, Canada, 2014, pp. 2672–2680. eprint: http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf (page 100).

[17] S. C. v. Leemput, J. Teuwen, B. v. Ginneken, and R. Manniesing, “MemCNN: A Python/PyTorch package for creating memory-efficient invertible neural networks,” *Journal of Open Source Software*, vol. 4, no. 39, p. 1576, Jul. 30, 2019 (page 100).

[18] P. Putzky and M. Welling, *Invert to Learn to Invert*, 2019 (page 100).

[19] B. Peters, E. Haber, and K. Lensink, “Fully reversible neural networks for large-scale surface and sub-surface characterization via remote sensing,” *arXiv preprint arXiv:2003.07474*, 2020 (page 100).

[20] B. Peters and E. Haber, “Fully reversible neural networks for large-scale 3d seismic horizon tracking,” in *82nd EAGE Annual Conference & Exhibition*, European Association of Geoscientists & Engineers, 2020, pp. 1–5 (page 100).

[21] M. D. Parno and Y. M. Marzouk, “Transport Map Accelerated Markov Chain Monte Carlo,” *SIAM/ASA Journal on Uncertainty Quantification*, vol. 6, no. 2, pp. 645–682, 2018 (page 100).

[22] S. Gershman and N. Goodman, “Amortized Inference in Probabilistic Reasoning,” in *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 36, 2014, pp. 517–522 (page 100).

[23] D. Hjelm, R. R. Salakhutdinov, K. Cho, N. Jojic, V. Calhoun, and J. Chung, “Iterative Refinement of the Approximate Posterior for Directed Belief Networks,” in *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, Eds., vol. 29, Curran Associates, Inc., 2016, pp. 4691–4699 (page 100).

[24] R. Krishnan, D. Liang, and M. Hoffman, “On the challenges of learning with inference networks on sparse, high-dimensional data,” in *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, A. Storkey and F. Perez-Cruz, Eds., ser. Proceedings of Machine Learning Research, vol. 84, PMLR, Sep. 2018, pp. 143–151 (page 100).

[25] Y. Kim, S. Wiseman, A. Miller, D. Sontag, and A. Rush, “Semi-amortized variational autoencoders,” in *Proceedings of the 35th International Conference on Machine Learning*, J. Dy and A. Krause, Eds., ser. Proceedings of Machine Learning Research, vol. 80, PMLR, Oct. 2018, pp. 2678–2687 (page 100).

[26] J. Marino, Y. Yue, and S. Mandt, “Iterative amortized inference,” *arXiv preprint arXiv:1807.09356*, 2018 (page 100).

[27] B. Peters, B. R. Smithyman, and F. J. Herrmann, “Projection methods and applications for seismic nonlinear inverse problems with multiple constraints,” *GEOPHYSICS*, vol. 84, no. 2, R251–R269, 2019 (page 101).

[28] Y. Yang and S. Soatto, “Conditional Prior Networks for Optical Flow,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 271–287 (page 101).

[29] C. Zeno, I. Golan, E. Hoffer, and D. Soudry, “Task Agnostic Continual Learning Using Online Variational Bayes,” *arXiv preprint arXiv:1803.10123*, 2018 (page 101).

[30] E. J. Candes, J. K. Romberg, and T. Tao, “Stable signal recovery from incomplete and inaccurate measurements,” *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, vol. 59, no. 8, pp. 1207–1223, 2006 (page 102).

[31] D. L. Donoho, “Compressed Sensing,” *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289–1306, 2006 (page 102).

[32] P. Witte, G. Rizzuti, M. Louboutin, A. Siahkoohi, and F. Herrmann, *InvertibleNetworks.jl: A Julia framework for invertible neural networks*, version v1.1.0, Nov. 2020 (page 102).

[33] M. Welling and Y. W. Teh, “Bayesian Learning via Stochastic Gradient Langevin Dynamics,” in *Proceedings of the 28th International Conference on Machine Learning*, Bellevue, Washington, USA: Omnipress, 2011, pp. 681–688, ISBN: 9781450306195 (page 104).

CHAPTER 5
RELIABLE AMORTIZED VARIATIONAL INFERENCE WITH PHYSICS-BASED LATENT DISTRIBUTION CORRECTION

5.1 Summary

Bayesian inference for high-dimensional inverse problems is challenged by the computational costs associated with the forward operator during posterior sampling, as well as the selection of an appropriate prior distribution that encodes our prior knowledge about the unknown. Amortized variational inference addresses these challenges where a deep neural network is trained to approximate the posterior distribution over existing pairs of model and data. When fed previously unseen data and normally distributed latent samples as input, the pretrained deep neural network—in our case a conditional normalizing flow—provides posterior samples associated with the input data virtually for free. However, the accuracy of this approach solely relies on the availability of high-fidelity training data, which seldom exists in geophysical inverse problems because of the highly heterogeneous structure of the Earth. In addition, the ability of amortized variational inference to approximate the posterior distribution for a previously unseen data hinges on the data being drawn from the training data distribution. As such, we offer a solution that increases the resilience of amortized variational inference when faced with data distribution shifts, e.g., changes in the forward model or prior distribution. Our proposed method involves learning a physics-based correction to the conditional normalizing flow latent distribution to provide a more accurate approximation to the posterior distribution for the observed data at hand, which might be drawn from a slightly shifted data distribution. To accomplish this, instead of feeding standard Gaussian latent samples to the pretrained conditional normalizing flow, we parameterize the latent distribution by a Gaussian distribution with an unknown mean

and diagonal covariance. These unknown quantities are then estimated through solving a variational inference objective that involves minimizing the Kullback-Leibler divergence between the corrected posterior distribution estimate and the true posterior distribution. By means of a realistic seismic imaging example, we show that our correction step improves the robustness of amortized variational inference with respect to a certain class of data distribution shifts, e.g., changes in number of source experiments and noise variance as well as shifts in the prior distribution. While generic and applicable to other inverse problems, our proposed latent distribution correction when applied to least squares imaging provides a seismic image with limited artifacts and an assessment of its uncertainty with approximately the same cost as five reverse-time migrations, which might be affordable in large-scale problems.

5.2 Introduction

Inverse problems involve the estimation of an unknown quantity based on noisy indirect observations. The problem is typically solved by minimizing the difference between observed and predicted data, where predicted data can be computed by modeling the underlying data generation process through a forward operator. Due to the presence of noise in the data, forward modeling errors, and the inherent nullspace of the forward operator, minimization of the data misfit alone negatively impacts the quality of the obtained solution [1]. Casting inverse problems into a probabilistic Bayesian framework allows for a more comprehensive description of their solution, where instead of finding one single solution, a distribution of solutions to the inverse problem—known as the posterior distribution—is obtained whose samples are consistent with the observed data [2]. The posterior distribution can be sampled to extract statistical information that allows for quantification of uncertainty, i.e., assessing the variability among the possible solutions to the inverse problem.

Uncertainty qualification and Bayesian inference in inverse problems often require high-dimensional posterior distribution sampling, for instance through the use of Markov chain Monte Carlo [MCMC, 3, 4, 5, 6]. Because of their sequential nature, MCMC sampling methods require a large number of sampling steps to perform accurate Bayesian inference [7], which reduces their applicability to large-scale problems due to the costs associated with the forward operator [8, 9, 4, 5, 10, 11, 12, 13, 14]. As an alternative, variational inference methods [15, 16, 17, 18, 19, 20, 21, 22, 23] approximate the posterior distribution with a surrogate and easy-to-sample distribution. By means of this approximation, sampling is turned into an optimization problem, in which the parameters of the surrogate distribution are tuned in order to minimize the divergence between the surrogate and posterior distributions. This surrogate distribution is then used for conducting Bayesian inference. While variational inference methods may have computational advantages over MCMC methods in high-dimensional inverse problems [24, 25], the resulting approximation to the posterior distribution is typically non-amortized, i.e., it is specific to the observed data used in solving the variational inference optimization problem. Thus, the variational inference optimization problem must be solved again for every new set of observations. Solving this optimization problem may require numerous iterations [19, 20], which may not be feasible in inverse problems with computationally costly forward operators, such as seismic imaging.

On the other hand, amortized variational inference [26, 27, 28, 29, 30, 31, 32, 33, 34, 35] reduces Bayesian inference computational costs by incurring an up-front optimization cost for finding a surrogate conditional distribution, typically parameterized by deep neural networks [31], that approximate the posterior distribution across a family of observed data instead of being specific to a single observed dataset. This optimization problem involves maximization of the probability density function (PDF) of the surrogate conditional distribution over existing pairs model and data [28]. Following optimization, samples from the posterior distribution for previously unseen data may be obtained by sampling the surrogate conditional distribution, which does not require further optimization or MCMC sampling. While drastically reducing the cost of Bayesian inference, amortized variational inference can only be used for inverse problems where a dataset of model and data pairs is available that sufficiently captures the underlying joint distribution. In reality, such an assumption is rarely true in geophysical applications due to the Earth's strong heterogeneity across geological scenarios and our lack of access to its interior [29, 36, 37]. Additionally, the accuracy of Bayesian inference with data-driven amortized variational inference methods degrades as the distribution of the data shifts with respect to pretraining data [38]. Among these shifts are changes in the distribution of noise, the number of observed data in multi-source inverse problems, and the distribution of unknowns, in other words, the prior distribution.

In this work, we leverage amortized variational inference to accelerate Bayesian inference while building resilience against data distribution shifts through a data-specific physics-based latent distribution correction method. During this process, the latent distribution of a normalizing-flow-based surrogate conditional distribution [31] is adapted to minimize the Kullback-Leibler (KL) divergence between the predicted and true posterior distributions. The invertibility of the conditional normalizing flow—a family of invertible neural networks [39]—guarantees the existence of an adapted latent distribution [40] that when "pushed forward" by the conditional normalization flow matches the posterior distribution. During pretraining, the conditional normalizing flow learns to Gaussianize the input model and data joint samples [31], resulting in a standard Gaussian latent distribution. As a result, for slightly shifted data distributions, the conditional normalization flow can provide samples from the the posterior distribution given an "approximately Gaussian" latent distribution as input [41, 42]. Motivated by this, and to limit the costs of the latent distribution correction step, we learn a simple diagonal (elementwise) scaling and shift to the latent distribution through a physic-based objective that minimizes the KL divergence between the predicted and true posterior distributions. As with amortized variational inference, after latent distribution correction, we gain cheap access to corrected posterior samples. Besides offering computational advantages, our proposed method implicitly learns

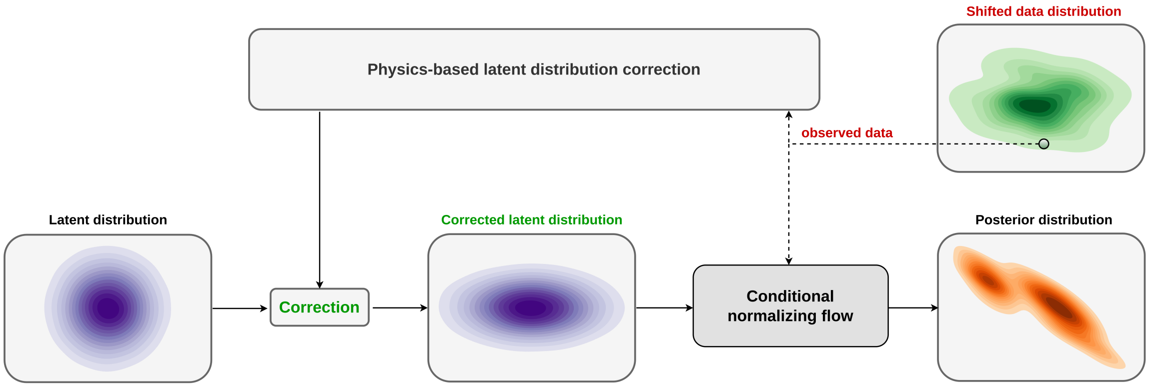

Shifted data distribution
Physics-based latent distribution correction
observed data
Latent distribution
Corrected latent distribution
Posterior distribution
Correction
Conditional normalizing flow

Figure 5.1: Schematic representation of our proposed method. We modify the standard Gaussian latent distribution of a pretrained conditional normalizing flow through a computationally cheap diagonal physics-based correction procedure to mitigate the errors due to data distribution shifts. Upon correction, the new latent samples result in corrected posterior samples when fed into the pretrained conditional normalizing flow.

the prior distribution during conditional normalizing flow pretraining. As advocated in the literature [43, 40] learned priors have the potential to better describe the prior information when compared to generic handcrafted priors that are chosen purely for their simplicity and applicability. A schematic representation of our proposed method is shown in Figure 5.1.

5.2.1 Related work

In the context of variational inference for inverse problems, [19], [44], [45], [46], and [47] proposed a non-amortized variational inference approach to approximate posterior distributions through the use of normalizing flows. These methods do not require training data, however they require choosing a prior distribution and repeated computationally expensive evaluation of the forward operator and the adjoint of its Jacobian. Therefore, the proposed methods may prove computationally expensive when applied to inverse problems involving computationally expensive forward operators. To speed up the convergence of non-amortized variational inference, [29] introduces a normalizing-flow-based nonlinear preconditioning scheme. In this approach, a pretrained conditional normalizing flow capable of providing a low-fidelity approximation to the posterior distribution is used to warm-start the variational inference optimization procedure. In a related work, [48] partially address challenges associated with non-amortized variational inference by learning a normalizing-flow-based prior distribution in a learned low-dimensional space via an injective network. Additionally to learning a prior, this approach also allowed non-amortized variational inference in a lower dimensional space, which could potentially have computational benefits.

Alternatively, amortized variational inference was applied by [49], [31], [32], [30], and [33] to further reduce the computational costs associated with Bayesian inference. These methods learn an implicit prior distribution from training data and provide posterior samples for previously unseen data for a negligible cost due to the low cost of forward evaluation of neural networks. The success of such techniques hinges on having access to high-quality training data, including pairs of model and data that sufficiently capture the underlying model and data joint distribution. To address this limitation, [29] take amortized variational inference a step further by proposing a two-stage multifidelity approach where during the first stage a conditional normalizing flow is trained in the context of amortized variational inference. To account for any potential shift in data distribution, the weights of this pretrained conditional normalizing flow are then further finetuned during an optional second stage of physics-based variational inference, which is customized for the specific imaging problem at hand. While limiting the risk of errors caused by shifts in the distribution of data, the second physics-based stage can be computationally expensive due to the high dimensionality of the weight space of conditional normalizing flows. Our work differs from the proposed method in [29] in that we learn to correct the latent distribution of the conditional normalizing flow, which typically has a much smaller dimensionality (approximately $\times 90$ in our case) than the dimension of the conditional normalizing flow weight space.

The work we present is principally motivated by [40], which demonstrates that normalizing flows—due to their invertibility—can mitigate biases caused by shifts in the data distribution. This is achieved by reparameterizing the unknown by a pretrained normalizing flow with fixed weights while optimizing over the latent variable in order to fit the data. The reparameterization together with a Gaussian prior on the latent variable act as a regularization while the invertibility ensures the existence of a latent variable that fits the data. [40] exploit this property using a normalizing flow that is pretrained to capture the prior distribution associated with an inverse problem. By computing the maximum-a-posterior estimate in the latent space, [40], as well as [23] and [34], limit biases originating from data distribution shifts while utilizing the prior knowledge of the normalizing flow. We extend this method by obtaining an approximation to the full posterior distribution of an inverse problem instead of a point estimate, e.g., maximum-a-posteriori.

Our work is also closely related to the non-amortized variational inference techniques presented by [50] and [48], in which the latent distribution of a normalizing flow is altered in order to perform Bayesian inference. In contrast to our approach, these methods employ a pretrained normalizing flow that approximates the prior distribution. As a result, it is necessary to significantly alter the latent distribution in order to adapt the pretrained normalizing flow to sample from the posterior distribution. In response, [50] and [48] train a second normalizing flow aimed at learning a latent distribution that approximates the posterior distribution after passing through the pretrained normalizing flow. Our study, however, utilizes a conditional normalizing flow, which, before any corrections are applied, already approximates the posterior distribution. We argue that our approach requires a simpler correction in the latent space to mitigate biases caused by shifts in the data distribution. This is crucial when dealing with large-scale inverse problems with computationally expensive forward operators.

5.2.2 Main contributions

The main contribution of our work involves a variational inference formulation for solving probabilistic Bayesian inverse problems that leverages the benefits of data-driven learned posteriors whilst being informed by physics and data. The advantages of this formulation include

- Enhancing the solution quality of inverse problems by implicitly learning the prior distribution from the data;
- Reliably reducing the cost of uncertainty quantification and Bayesian inference; and
- Providing safeguards against data distribution shifts.

5.2.3 Outline

In the sections below, we first formulate multi-source inverse problems mathematically and cast them within a Bayesian framework. We then describe variational inference and examine how existing model and data pairs can be used to obtain an approximation to the posterior distribution that is amortized, i.e., the approximation holds over a distribution of data rather than a specific set of observations. We showcase amortized variational inference on a high-dimensional seismic imaging example in a controlled setting where we assume observed data during inference is drawn from the same distribution as training seismic data. As means to mitigate potential errors due to data distribution shifts, we introduce our proposed correction approach to amortized variational inference, which exploits the advantages of learned posteriors while reducing potential errors induced by certain data distribution shifts. Two realistic seismic imaging examples are presented, in which the distribution of the data is shifted by altering the forward model and the prior distribution. These numerical experiments are intended to demonstrate the ability of the proposed latent distribution correction method to correct for errors caused by shifts in the distribution of data. Finally, we verify our proposed Bayesian inference method by conducting posterior contraction experiments.

5.3 Theory

Our purpose is to present a technique for using deep neural networks to accelerate Bayesian inference for ill-posed inverse problems while ensuring that the inference is robust with respect to data distribution shifts through the use of physics. We begin with an introduction to Bayesian inverse problems and discuss variational inference [15] as a probabilistic framework for solving Bayesian inverse problems.

5.3.1 Inverse problems

We are concerned with estimating an unknown multidimensional quantity $\mathbf{x}^* \in \mathcal{X}$, often referred to as the unknown model, given N noisy and indirect observed data (e.g, shot records in seismic imaging) $\mathbf{y} = \{\mathbf{y}_i\}_{i=1}^N$ with $\mathbf{y}_i \in \mathcal{Y}$. Here $\mathcal{X}$ and $\mathcal{X}$ denote the space of unknown models and data, respectively. The physical underlying data generation process is assumed to be encoded in forward modeling operators, $\mathcal{F}_i : \mathcal{X} \to \mathcal{Y}$, which relates the unknown model to the observed data via the forward model

$$\mathbf{y}_i = \mathcal{F}_i(\mathbf{x}^*) + \boldsymbol{\epsilon}_i, \quad i = 1, \ldots, N. \tag{5.1}$$

In the above expression, $\boldsymbol{\epsilon}_i$ is a vector of measurement noise, which might also include errors in the forward modeling operator. Solving ill-posed inverse problems is challenged by noise in the observed data, potential errors in the forward modeling operator, and the intrinsic nontrivial nullspace of the forward operator [1]. These challenges can lead to non-unique solutions where different estimates of the unknown model may fit the observed data equally well. Under such conditions, the use of a single model estimate ignores the intrinsic variability within inverse problem solutions, which increases the risk of overfitting the data. Therefore, not only does the process of estimating $\mathbf{x}^*$ from$\mathbf{y}$ require regularization, but it also calls for a statistical inference framework that allows us to characterize the variability among the solutions by quantifying the solution uncertainty [2].

5.3.2 Bayesian inference for solving inverse problems

To systematically quantify the uncertainty, we cast the inverse problem into a Bayesian framework [2]. In this framework, instead of having a single estimate of the unknown, the solution is characterized by a probability distribution over the solution space $\mathcal{X}$ that is conditioned on data, namely the posterior distribution. This conditional distribution, denoted by $p_{\text{post}}(\mathbf{x} \mid \mathbf{y})$, can according to the Bayes' rule be written as follows:

$$p_{\text{post}}(\mathbf{x} \mid \mathbf{y}) = \frac{p_{\text{like}}(\mathbf{y} \mid \mathbf{x})\, p_{\text{prior}}(\mathbf{x})}{p_{\text{data}}(\mathbf{y})}, \tag{5.2}$$

which equivalently can be expressed as

$$\begin{aligned} -\log p_{\text{post}}(\mathbf{x} \mid \mathbf{y}) &= -\sum_{i=1}^{N} \log p_{\text{like}}(\mathbf{y}_i \mid \mathbf{x}) - \log p_{\text{prior}}(\mathbf{x}) + \log p_{\text{data}}(\mathbf{y}) \\ &= \frac{1}{2\sigma^2} \sum_{i=1}^{N} \left\| \mathbf{y}_i - \mathcal{F}_i(\mathbf{x}) \right\|_2^2 - \log p_{\text{prior}}(\mathbf{x}) + \text{const.} \end{aligned} \tag{5.3}$$

In equations 5.2 and 5.3, the likelihood function $p_{\text{like}}(\mathbf{y} \mid \mathbf{x})$ quantifies how well the predicted data fits the observed data given the PDF of the noise distribution. For simplicity, we assume the distribution of the noise is a zero-mean Gaussian distribution with covariance $\sigma^2\mathbf{I}$ but other choices can be incorporated. The prior distribution $p_{\text{prior}}(\mathbf{x})$ encodes prior beliefs on the unknown quantity, which can also be interpreted as a regularizer for the inverse problem. Finally, $p_{\text{data}}(\mathbf{y})$ denotes the data PDF, which is a normalization constant that is independent of $\mathbf{x}$.

Acquiring statistical information regarding the posterior distribution requires access to samples from the posterior distribution. Sampling the posterior distribution, commonly achieved via MCMC [3] or variational inference techniques [15], is computationally costly in high-dimensional inverse problems due to the costs associated with many needed evaluations of the forward operator [51, 2, 52, 4, 24, 53, 5, 54, 11, 12, 25]. For multi-source inverse problem the costs are especially high as evaluating the likelihood function involves N forward operator evaluations (equation 5.3). Stochastic gradient Langevin dynamics [SGLD; 9, 55, 6] alleviates the need to evaluate the likelihood for all the N forward operators by allowing for stochastic approximations to the likelihood, i.e., evaluating the likelihood over randomly selected indices $i \in \{1, \ldots, N\}$. While SGLD can provably provide accurate posterior samples with more favorable computational costs [9], due to the sequential nature of MCMC methods, SGLD still requires numerous iterations to fully traverse the probability space [7], which is computationally challenging in large-scale multi-source inverse problems. In the next section, we introduce variational inference as an alternative Bayesian inference method that has the potential to scale better than MCMC-methods in inverse problems with costly forward operators [24, 25].

5.3.3 Variational inference

As an alternative to MCMC-based methods, variational inference methods [15] reduce the problem of sampling from the posterior distribution $p_{\text{post}}(\mathbf{x} \mid \mathbf{y})$ to an optimization problem. The optimization problem involves approximating the posterior PDF via the PDF of a tractable surrogate distribution $p_{\phi}(\mathbf{x})$ with parameters $\boldsymbol{\phi}$ by minimizing a divergence (read "distance") between $p_{\phi}(\mathbf{x})$ and $p_{\text{post}}(\mathbf{x} \mid \mathbf{y})$ with respect to surrogate distribution parameters $\boldsymbol{\phi}$. This optimization problem can be solved approximately, which allows for trading off computational cost for accuracy [15]. After optimization, we gain access to samples from the posterior distribution by sampling $p_{\phi}(\mathbf{x})$ instead, which does not involve forward operator evaluations.

Due to its simplicity and connections to the maximum likelihood principle [56], we formulate variational inference via the Kullback-Leibler (KL) divergence. The KL divergence can be explained as the cross-entropy of $p_{\text{post}}(\mathbf{x} \mid \mathbf{y})$ relative to $p_{\phi}(\mathbf{x})$ minus the entropy of $p_{\phi}(\mathbf{x})$. This definition describes the reverse KL divergence, denoted by $\mathbb{KL}\left(p_{\phi}(\mathbf{x}) \,||\, p_{\text{post}}(\mathbf{x} \mid \mathbf{y})\right)$, which is not equal to the forward KL divergence, $\mathbb{KL}\left(p_{\text{post}}(\mathbf{x} \mid \mathbf{y}) \,||\, p_{\phi}(\mathbf{x})\right)$.

This non-symmetry in KL divergence leads to different computational and approximation properties during variational inference, which we describe in detail in the following sections. We will first describe the reverse KL divergence, followed by the forward KL divergence. Finally, we will describe normalizing flows as a way of parameterized surrogate distributions to facilitate variational inference.

Non-amortized variational inference

The reverse KL divergence is the common choice for formulating variational inference [19, 44, 45, 46, 47] in which the the physically-informed posterior density guides the optimization over $\boldsymbol{\phi}$. The reverse KL divergence can be mathematically stated as

$$\mathbb{KL}\left(p_{\phi}(\mathbf{x}) \;||\; p_{\text{post}}(\mathbf{x} \mid \mathbf{y}_{\text{obs}})\right) = \mathbb{E}_{\mathbf{x}\sim p_{\phi}(\mathbf{x})}\big[-\log p_{\text{post}}(\mathbf{x} \mid \mathbf{y}_{\text{obs}}) + \log p_{\phi}(\mathbf{x})\big], \tag{5.4}$$

where $\mathbf{y}_{\text{obs}} \sim p_{\text{data}}(\mathbf{y})$ refers to a specific single observed data. $\mathbf{x}$ in the right hand side of the expression in equation 5.4 is a random variable obtained by sampling the surrogate distribution $p_{\phi}(\mathbf{x})$, over which we evaluate the expectation. Variational inference using the reverse KL divergence involves minimizing equation 5.4 with respect to $\boldsymbol{\phi}$ during which the logarithm of the posterior PDF is approximated by the logarithm of the surrogate PDF, when evaluated over samples from the surrogate distribution. By expanding the negative-log posterior density via Bayes' rule (equation 5.3), we write the non-amortized variational inference optimization problem as

$$\boldsymbol{\phi}^{*} = \underset{\phi}{\arg\min}\, \mathbb{E}_{\mathbf{x}\sim p_{\phi}(\mathbf{x})}\left[\frac{1}{2\sigma^2} \sum_{i=1}^{N} \big\|\mathbf{y}_{\text{obs},i} - \mathcal{F}_i(\mathbf{x})\big\|_2^2 - \log p_{\text{prior}}\left(\mathbf{x}\right) + \log p_{\phi}(\mathbf{x}) \right]. \tag{5.5}$$

The expectation in the above equation is approximated with a sample mean over samples drawn from $p_{\phi}(\mathbf{x})$. The optimization problem in equation 5.5 can be solved using stochastic gradient descent and its variants [57, 58, 59, 60] where at each iteration the objective function is evaluated over a batch of samples drawn from $p_{\phi}(\mathbf{x})$ and randomly selected (without replacement) indices $i \in \{1, \dots, N\}$. To solve this optimization problem, there are two considerations to take into account. First consideration involves the tractable computation of the surrogate PDF and its gradient with respect to $\boldsymbol{\phi}$. As described in the following sections, normalizing flows [17], which are a family of specially designed invertible neural networks [39], facilitate the computation of these quantities via the change-of-variable formula in probability distributions [61]. The second consideration involves differentiating (with respect to $\boldsymbol{\phi}$) the expectation (sample mean) operation in equation 5.5. Evaluating this expectation requires sampling from the surrogate distribution $p_{\phi}(\mathbf{x})$, which depends $\boldsymbol{\phi}$. Differentiating through the sampling procedure from the surrogate distribution $p_{\phi}(\mathbf{x})$ can be facilitated through the reparameterization trick [62]. In this approach sampling from $p_{\phi}(\mathbf{x})$ is interpreted as passing latent samples $\mathbf{z} \in \mathcal{Z}$ from a simple base distribution, such as standard Gaussian distribution, through a parametric function parameterized by $\boldsymbol{\phi}$ [62]. With this interpretation, the expectation over $p_{\phi}(\mathbf{x})$ can be computed over the latent distribution instead, which does not depend on $\boldsymbol{\phi}$, followed by a mapping of latent samples through the parametric function. This process enables computing the gradient of the expression in equation 5.5 with respect to $\boldsymbol{\phi}$ [62].

Following optimization, $p_{\phi^*}(\mathbf{x})$ provides unlimited samples from the posterior distribution—virtually for free. While there are indications that this approach can be computationally favorable compared to MCMC sampling methods [24, 25], each iteration during optimization problem 5.5 involves evaluating the forward operator and the adjoint of its Jacobian, which can be computationally costly depending on N and the number of iterations required to solve 5.5. In addition, and more importantly, this approach is non-amortized—i.e., the resulting surrogate distribution $p_{\phi^*}(\mathbf{x})$ approximates the posterior distribution for the specific data $\mathbf{y}_{\text{obs}}$ that is used to solve optimization problem 5.5. This necessitates the optimization problem to be solved again for a new instance of the inverse problem with different data. In the next section, we introduce an amortized variational inference approach that addresses these limitations.

Amortized variational inference

Similarly to reverse KL divergence, forward KL divergence involves calculating the difference between the logarithms of the surrogate PDF and the posterior PDF. In contrast to reverse KL divergence, however, to compute the forward KL divergence the PDFs are evaluated over samples from the posterior distribution rather than the surrogate distribution samples (see equation 5.4). The forward KL divergence can be written as follows

$$\mathbb{KL}\left(p_{\text{post}}(\mathbf{x} \mid \mathbf{y}) \mid\mid p_{\phi}(\mathbf{x})\right) = \mathbb{E}_{\mathbf{x} \sim p_{\text{post}}(\mathbf{x} \mid \mathbf{y})}\big[-\log p_{\phi}(\mathbf{x}) + \log p_{\text{post}}(\mathbf{x} \mid \mathbf{y})\big]. \tag{5.6}$$

Following the expression above, it is infeasible to evaluate the forward KL divergence in inverse problems as it requires access to samples from the posterior distribution—the samples that we are ultimately after and do not have access to. However, the average (over data) forward KL divergence can be computed using available model and data pairs in the form of samples from the joint distribution $p(\mathbf{x}, \mathbf{y})$. This involves integrating (marginalizing) the forward KL divergence over existing data $\mathbf{y} \sim p_{\text{data}}(\mathbf{y})$:

$$\begin{aligned}
&\mathbb{E}_{\mathbf{y} \sim p_{\text{data}}(\mathbf{y})}\Big[\mathbb{KL}\left(p_{\text{post}}(\mathbf{x} \mid \mathbf{y}) \mid\mid p_{\phi}(\mathbf{x})\right)\Big] \\
&= \mathbb{E}_{\mathbf{y} \sim p_{\text{data}}(\mathbf{y})}\mathbb{E}_{\mathbf{x} \sim p_{\text{post}}(\mathbf{x} \mid \mathbf{y})}\Big[-\log p_{\phi}(\mathbf{x} \mid \mathbf{y}) + \underbrace{\log p_{\text{post}}(\mathbf{x} \mid \mathbf{y})}_{\text{constant w.r.t. } \phi}\Big] \\
&= \iint \underbrace{p_{\text{data}}(\mathbf{y}) p_{\text{post}}(\mathbf{x} \mid \mathbf{y})}_{=p(\mathbf{x},\mathbf{y})}\Big[-\log p_{\phi}(\mathbf{x} \mid \mathbf{y})\Big]\mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{y} + \text{const} \\
&= \mathbb{E}_{(\mathbf{x},\mathbf{y}) \sim p(\mathbf{x},\mathbf{y})}\big[-\log p_{\phi}(\mathbf{x} \mid \mathbf{y})\big] + \text{const}.
\end{aligned} \tag{5.7}$$

In the above expression $p_{\phi}(\mathbf{x} \mid \mathbf{y})$ represents a surrogate conditional distribution that approximates the posterior distribution for any data $\mathbf{y} \sim p_{\text{data}}(\mathbf{y})$. The third line in equation 5.7 is the result of applying the chain rule of PDFs[1]. By minimizing the average KL divergence we obtain the following amortized variational inference objective:

[1] $p(x, y) = p(x \mid y)\, p(y),\ \forall\, x \in \mathcal{X},\ y \in \mathcal{Y}$.

$$
\begin{aligned}
\boldsymbol{\phi}^{*} &= \underset{\phi}{\arg\min}\ \mathbb{E}_{\mathbf{y}\sim p_{\text{data}}(\mathbf{y})}\big[\mathbb{KL}\left(p_{\text{post}}(\mathbf{x}\mid\mathbf{y})\ ||\ p_{\phi}(\mathbf{x}\mid\mathbf{y})\right)\big] \\
&= \underset{\phi}{\arg\min}\ \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim p(\mathbf{x},\mathbf{y})}\big[-\log p_{\phi}(\mathbf{x}\mid\mathbf{y})\big].
\end{aligned} \tag{5.8}
$$

The above optimization problem represent a supervised learning framework to for obtaining fully-learned posteriors using existing pairs of model and data. The expectation is approximated with a sample mean over available model and data joint samples. Note that this method does not impose any explicit assumption on the noise distribution (see equation 5.3), and the information about the forward model is implicitly encoded in the model and data pairs. As a result, this formulation is an instance of likelihood-free simulation-based inference methods [63, 64] that allows us to approximate the posterior distribution for previously unseen data as,

$$
p_{\phi^{*}}(\mathbf{x}\mid\mathbf{y}_{\text{obs}}) \approx p_{\text{post}}(\mathbf{x}\mid\mathbf{y}_{\text{obs}}), \quad \forall\,\mathbf{y}_{\text{obs}}\sim p_{\text{data}}(\mathbf{y}). \tag{5.9}
$$

Equation 5.9 holds for previously unseen data drawn from $p_{\text{data}}(\mathbf{y})$ provided that the optimization problem 5.8 is solved accurately [31, 38], i.e., $\mathbb{E}_{\mathbf{y}\sim p_{\text{data}}(\mathbf{y})}\big[\mathbb{KL}\left(p_{\text{post}}(\mathbf{x}\mid\mathbf{y})\ ||\ p_{\phi^{*}}(\mathbf{x}\mid\mathbf{y})\right)\big] = 0$. Following an one-time upfront cost of training, equation 5.9 can be used to samples the posterior distribution with no additional forward operator evaluations. While computationally cheap, the accuracy of the amortized variational inference approach in equation 5.8 is directly linked to the quality and quantity of model and data pairs used during optimization [63]. This raises questions regarding the reliability of this approach in domains that sufficiently capturing the underlying joint model and data distribution is challenging, e.g., in geophysical applications due to the Earth's strong heterogeneity across geological scenarios and our lack of access to its interior [29, 36, 37]. To increases the resilience of amortized variational inference when faced with data distribution shifts, e.g., changes in the forward model or prior distribution, we propose a latent distribution correction to physically inform the inference. Before describing our proposed physics-based latent distribution correction approach, we introduce conditional normalizing flows [31] to parameterize the surrogate conditional distribution for amortized variational inference.

5.3.4 Conditional normalizing flows for amortized variational inference

To limit the computational cost of amortized variational inference, both during optimization and inference, it is imperative that the surrogate conditional distribution be able to: (1) approximate complex distributions, i.e., it should have a high representation power, which is required to represent possibly multi-modal distributions; (2) support cheap density estimation, which involves computing the density $p_\phi(\mathbf{x} \mid \mathbf{y})$ for given $\mathbf{x}$ and $\mathbf{y}$; and (3) permit fast sampling from $p_\phi(\mathbf{x} \mid \mathbf{y})$ for cheap posterior sampling during inference. These characteristics are provided by conditional normalizing flows [31], which are a family of invertible neural networks [39] that are capable of approximating complex conditional distributions [65, 66].

A conditional normalizing flows—in the context of amortized variational inference—aims to map input samples $\mathbf{z}$ from a latent standard multivariate Gaussian distribution $\mathrm{N}(\mathbf{z} \mid \mathbf{0}, \mathbf{I})$ to samples from the posterior distribution given the observed data $\mathbf{y} \sim p_{\text{data}}(\mathbf{y})$ as an additional input. This nonlinear mapping can formally be stated as $f_\phi^{-1}(\,\cdot\,; \mathbf{y}) : \mathcal{Z} \to \mathcal{X}$, with $f_\phi^{-1}(\mathbf{z}; \mathbf{y})$ being the inverse of the conditional normalizing flow with respect to its first argument. Due to the low computational cost of evaluating invertible neural networks in reverse [39], using conditional normalizing flows as a surrogate conditional distribution $p_\phi(\mathbf{x} \mid \mathbf{y})$ allows for extremely fast sampling from $p_\phi(\mathbf{x} \mid \mathbf{y})$. In addition to low-cost sampling, the invertibility of conditional normalizing flows permits straightforward and cheap estimation of the density $p_\phi(\mathbf{x} \mid \mathbf{y})$. This allows for tractable amortized variational inference via equation 5.8 through the following change-of-variable formula in probability distributions [61],

$$p_\phi(\mathbf{x} \mid \mathbf{y}) = \mathrm{N}\left(f_\phi(\mathbf{x}; \mathbf{y}) \,\middle|\, \mathbf{0}, \mathbf{I}\right) \left|\det \nabla_{\mathbf{x}} f_\phi(\mathbf{x}; \mathbf{y})\right|, \quad \forall\, \mathbf{x}, \mathbf{y} \sim p(\mathbf{x}, \mathbf{y}). \tag{5.10}$$

In the above formula, $\mathrm{N}\left(f_\phi(\mathbf{x};\mathbf{y})\,\middle|\,\mathbf{0},\mathbf{I}\right)$ represents the PDF for a multivariate standard Gaussian distribution evaluated at $f_\phi(\mathbf{x};\mathbf{y})$. Thanks to the special design of invertible neural networks [39], density estimation via equation 5.10 is cheap since evaluating the conditional normalizing flow and the determinant of its Jacobian $\det \nabla_{\mathbf{x}} f_\phi(\mathbf{x};\mathbf{y})$ are almost free of cost. Given the expression for $p_\phi(\mathbf{x} \mid \mathbf{y})$ in equation 5.10, we derive the following training objective for amrotized conditional normalizing flows:

$$\begin{aligned}\boldsymbol{\phi}^* &= \underset{\phi}{\arg\min}\ \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim p(\mathbf{x},\mathbf{y})}\big[-\log p_\phi(\mathbf{x} \mid \mathbf{y})\big] \\ &= \underset{\phi}{\arg\min}\ \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim p(\mathbf{x},\mathbf{y})}\left[\frac{1}{2}\left\|f_\phi(\mathbf{x};\mathbf{y})\right\|_2^2 - \log\left|\det \nabla_{\mathbf{x}} f_\phi(\mathbf{x};\mathbf{y})\right|\right].\end{aligned} \tag{5.11}$$

In the above objective, the ℓ_2-norm follows from a standard Gaussian distribution assumption on the latent variable, i.e., the output of the normalizing flow. The second term quantifies the relative change of density volume [papamakarios2021] and can be interpreted as an entropy regularization of $p_\phi(\mathbf{x} \mid \mathbf{y})$, which prevents the conditional normalizing flow from converging to solutions, e.g., $f_\phi(\mathbf{x};\mathbf{y}) := \mathbf{0}$. Due to the particular design of invertible networks [39, 31], computing the gradient of $\det \nabla_{\mathbf{x}} f_\phi(\mathbf{x};\mathbf{y})$ has a negligible extra cost. Figure 5.2 illustrates the pretraining phase as a schematic.

After training, given a previously unseen observed data $\mathbf{y}_{\text{obs}} \sim p_{\text{data}}(\mathbf{y})$ we sample from the posterior distribution using the inverse of the conditional normalizing flow. We achieve this by by feeding latent samples $\mathbf{z} \sim \mathrm{N}(\mathbf{z} \mid \mathbf{0},\mathbf{I})$ to the conditional normalizing flow's inverse $f_\phi^{-1}(\mathbf{z};\mathbf{y}_{\text{obs}})$ while conditioning on the observed data $\mathbf{y}_{\text{obs}}$,

$$f_\phi^{-1}(\mathbf{z};\mathbf{y}_{\text{obs}}) \sim p_{\text{post}}(\mathbf{x} \mid \mathbf{y}_{\text{obs}}), \quad \mathbf{z} \sim \mathrm{N}(\mathbf{z} \mid \mathbf{0},\mathbf{I}). \tag{5.12}$$

As the process above does not involve forward operator evaluations, sampling with pretrained conditional normalizing flows is fast once an upfront cost of amortized variational inference is incurred. In the next section, we apply the above amortized variational inference to a seismic imaging example in a controlled setting in which we assume no data

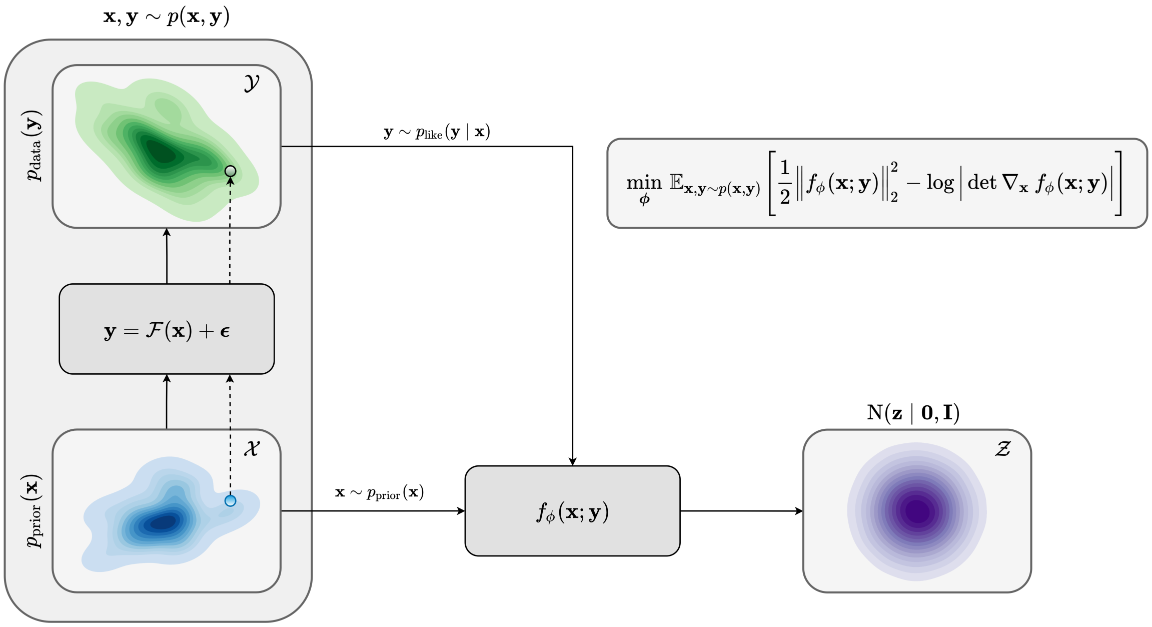

$\mathbf{x}, \mathbf{y} \sim p(\mathbf{x}, \mathbf{y})$
$\mathcal{Y}$
$p_{\text{data}}(\mathbf{y})$
$\mathbf{y} \sim p_{\text{like}}(\mathbf{y} \mid \mathbf{x})$
$\min_{\phi} \mathbb{E}_{\mathbf{x},\mathbf{y} \sim p(\mathbf{x},\mathbf{y})} \left[\frac{1}{2} \left\| f_\phi(\mathbf{x}; \mathbf{y}) \right\|_2^2 - \log \left| \det \nabla_{\mathbf{x}} f_\phi(\mathbf{x}; \mathbf{y}) \right| \right]$
$\mathbf{y} = \mathcal{F}(\mathbf{x}) + \boldsymbol{\epsilon}$
$\mathrm{N}(\mathbf{z} \mid \mathbf{0}, \mathbf{I})$
$\mathcal{X}$
$\mathcal{Z}$
$p_{\text{prior}}(\mathbf{x})$
$\mathbf{x} \sim p_{\text{prior}}(\mathbf{x})$
$f_\phi(\mathbf{x}; \mathbf{y})$

Figure 5.2: A schematic representation of pretraining conditional normalizing flows in the context of amortized variational inference. During pretraining, joint model and data joint samples $\mathbf{x}, \mathbf{y} \sim p(\mathbf{x}, \mathbf{y})$ from the training dataset and are fed to the conditional normalizing flow. The training objective (equation 5.11) enforces the conditional normalizing flow to Gaussianize its input.

distribution shifts during inference

5.4 Validating amortized variational inference

The objective of this example is to apply amortized variational inference to the high-dimensional seismic imaging problem. We show that a relatively good pretrained conditional normalizing flow within the context of amortized variational inference can be used to provide approximate posterior samples for previously unseen seismic data that is drawn from the same distribution as training seismic data. We begin by introducing seismic imaging and describe challenges with Bayesian inference in this problem.

5.4.1 Seismic imaging

We are concerned with constructing an image of the Earth's subsurface using indirect surface measurements that record the Earth's response to synthetic sources being fired on the surface. The nonlinear relationship between these measurements, known as shot records, and the squared-slowness model of the Earth's subsurface is governed by the wave equation. By linearizing this nonlinear relation, seismic imaging aims to estimates the short-wavelength component of the Earth's subsurface squared-slowness model. In its simplest acoustic form, the linearization with respect to the slowness model—around a known, smooth background squared slowness model $\mathbf{m}_0$—leads to the following linear forward problem:

$$\mathbf{d}_i = \mathbf{J}(\mathbf{m}_0, \mathbf{q}_i)\delta\mathbf{m}^* + \boldsymbol{\epsilon}_i, \quad i = 1, \ldots, N. \tag{5.13}$$

We invert the above forward model to estimate the ground truth seismic image $\delta\mathbf{m}^*$ from N processed (linearized) shot records $\{\mathbf{d}_i\}_{i=1}^{N}$ where $\mathbf{J}(\mathbf{m}_0, \mathbf{q}_i)$ represents the linearized Born scattering operator [67]. This operator is parameterized by the source signature $\mathbf{q}_i$ and the smooth background squared-slowness model $\mathbf{m}_0$. Noise is denoted by $\boldsymbol{\epsilon}_i$, and represents measurement noise and linearization errors. While amortized variational inference does not require knowing the distribution of the noise, to simulate pairs of data and model for simplicity we assume the noise distribution is a zero-centered Gaussian distribution with known covariance $\sigma^2\mathbf{I}$. Due to the presence of shadow zones and noisy finite-aperture shot data, seismic imaging corresponds to solving an inconsistent and ill-conditioned linear inverse problem [68, 69, 70]. To avoid the risk of overfitting the data and to quantify uncertainty, we cast the seismic imaging problem into a Bayesian inverse problem [2].

To address the challenge of Bayesian inference in this high-dimensional inverse problem, we adhere to our amortized variational inference framework. Within this approach, for an one-time upfront cost of training a conditional normalizing flow, we get access to posterior samples for previously unseen observed data that are drawn from the same distribution as the distribution of training seismic data. This includes data acquired in areas of the Earth with similar geologies, e.g. in neighboring surveys. In addition, in our framework no explicit prior distribution needs to chosen as the conditional normalizing flow learns the prior distribution during pretraining. The implicitly learned prior distribution by the conditional normalizing flow minimizes the risk of negatively biasing the outcome of Bayesian inference by using overly simplistic priors. In the next section, we describe the setup for our amortized variational inference for seismic imaging.

Acquisition geometry

To mimic the complexity of real seismic images, we propose a “quasi”-real data example in which we generate synthetic data by applying the linearized Born scattering operator to 4750 2D sections with size $3075\,\mathrm{m} \times 5120\,\mathrm{m}$ extracted from the shallow section of the Kirchhoff migrated Parihaka-3D field dataset [71, 72]. We consider a $12.5\,\mathrm{m}$ vertical and $20\,\mathrm{m}$ horizontal grid spacing, and we augment an artificial $125\,\mathrm{m}$ water column on top of these images. We parameterize the linearized Born scattering operator via a fictitious background squared-slowness model, derived from the Kirchhoff migrated images. To ensure good coverage, we simulate 102 shot records with a source spacing of $50\,\mathrm{m}$. Each shot is recorded for two seconds with 204 fixed receivers sampled at $25\,\mathrm{m}$ spread on top of the model. The source is a Ricker wavelet with a central frequency of $30\,\mathrm{Hz}$. To mimic a more realistic imaging scenario, we add band-limited noise to the shot records, where the noise is obtained by filtering white noise with the source wavelet (Figure 5.4b).

Training configuration

Casting seismic imaging into amortized variational inference, as described in this chapter, is hampered by the high-dimensionality of the data due to the multi-source nature of this inverse problem. To avoid computational complexities associated with directly using N shot records as input to the conditional normalizing flow, we choose to condition the conditional normalizing flow on the reverse-time migrated image, which can be estimated by applying the adjoint of the linearized Born scattering operator to the shot records,

$$\delta\mathbf{m}_{\text{RTM}} = \sum_{i=1}^{N} \mathbf{J}(\mathbf{m}_0, \mathbf{q}_i)^\top \mathbf{d}_i. \tag{5.14}$$

While $\mathbf{d}_i$ in the above expression is defined according to the linearized forward model in equation 7.1, which does not involve linearization errors, our method can handle observed data simulated from wave-equation based nonlinear forward modeling. Conditioning on the reverse-time migrated image and not on the shot records directly may result in learning an approximation to the true posterior distribution [73]. While technique from statistics involving learned summary functions [28, 38] can reduce the dimensionality of the observed data, we propose to limit the Bayesian inference bias induced by conditioning on the reverse-time migrated image via our physics-based latent variable correction approach. We leave utilizing summary functions in the context of seismic imaging Bayesian inference to future work.

To create training pairs, $(\delta\mathbf{m}^{(i)}, \delta\mathbf{m}_{\text{RTM}}^{(i)}),\ i = 1, \ldots, 4750$, we first simulate (see Figure 5.2) noisy seismic data according to the above-mentioned acquisition design for all extracted seismic images $\delta\mathbf{m}^{(i)}$ from shallow sections of the imaged Parihaka dataset. Next, we compute $\delta\mathbf{m}_{\text{RTM}}^{(i)}$ by applying reverse-time-migration to the observed data for each image $\delta\mathbf{m}^{(i)}$. We train a conditional normalizing flow on the resulting pairs $(\delta\mathbf{m}^{(i)}, \delta\mathbf{m}_{\text{RTM}}^{(i)}),\ i = 1, \ldots, 4750$ according to the objective function in equation 5.11 with the Adam stochastic optimization method [60] with a batchsize of 16 for one thousand passes over the training dataset (epochs). We use an initial stepsize of 10^{-4} and decrease it after each epoch until reaching the final stepsize of 10^{-6}. To monitor overfitting, we evaluate the objective function at the end of every epoch over random subsets of the validation set, consisting of 530 seismic images extracted from the shallow sections of the imaged Parihaka dataset and the associated reverse-time migrated images. As illustrated in Figure 5.3, the training and validation objective values exhibit a decreasing trend, which suggests no overfitting. We stopped the training after one thousand epochs due to a slowdown in the decrease of the training and validation objective values.

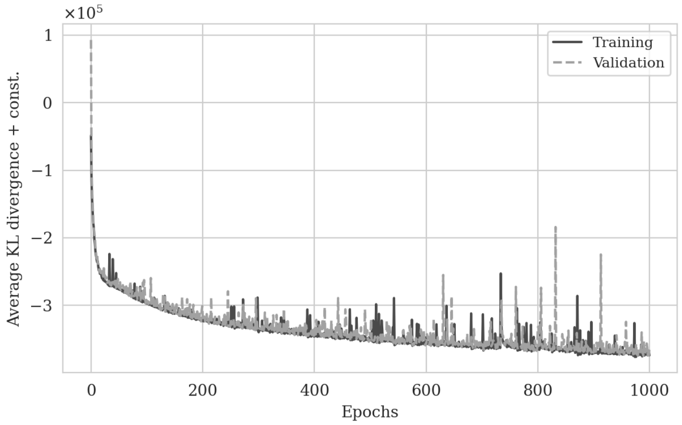


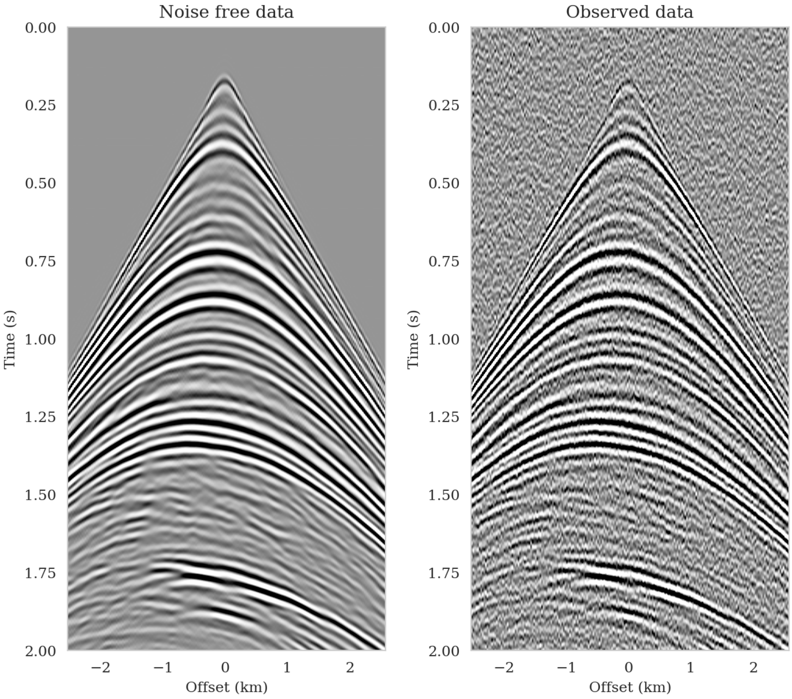
Figure 5.3: Training and validation objective values as a function of epochs. The validation objective value is computed over randomly selected batched of the validation set at the end of each epoch.

5.4.2 Results and observations

Following training, the pretrained conditional normalizing flow is able to produce samples from the posterior distribution for seismic data not used in training. To demonstrate this, we simulate seismic data for a previously unseen perturbation model using the forward model 7.1 with the same noise variance. Figure 5.4 shows an example of a single noise-free (Figure 5.4a) and noisy (Figure 5.4b) shot record for one of 102 sources.

We perform reverse-time migration to obtain the necessary input for the conditional normalizing flow to obtain posterior samples. We show the ground-truth seismic image (to be estimated) and the resulting reverse-time migrated image in Figures 5.5a and 7.2, respectively. Clearly, the reverse-time migrated image has grossly wrong amplitudes, and more importantly, due to limited-aperture shot data, the edges of the image are not well illuminated.

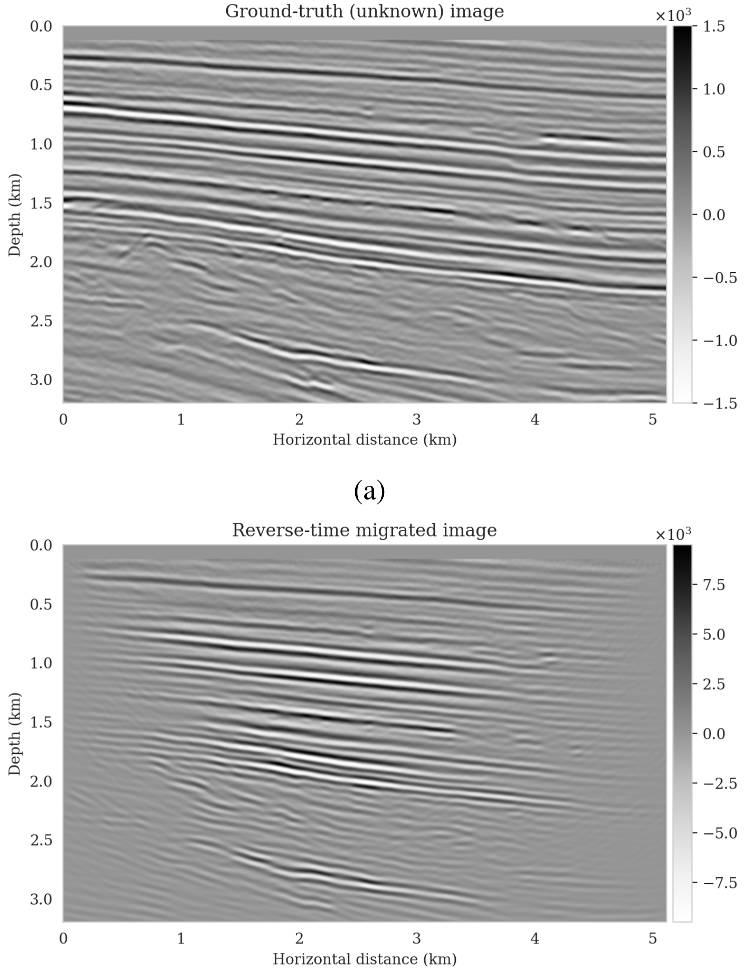
We obtain one thousand posterior samples by providing the reverse-time migrated image and latent samples drawn from the standard Gaussian distribution to the pretrained conditional normalizing flow (equation 5.9). This process is fast as it does not require any forward operator evaluations. To illustrate the variability among the posterior samples, we show six of them in Figure 5.6. As shown in Figure 5.6, these image samples have ampli-



(a)

(b)

Figure 5.4: A shot record generated from an image extracted from the Parihaka dataset. (a) Noise-free linearized data. (b) Linearized data with bandwidth-limited noise.



(a)

(b)

Figure 5.5: Amortized variational inference testing phase setup. (a) High-fidelity ground-truth image. (b) Reverse-time migrated image with SNR $-12.17\,\mathrm{dB}$.

tudes in the same range as the ground-truth image and better predict the reflectors at the edges of the image compared to the reverse-time migration image (Figure 7.2). In addition, the posterior samples indicate improve imaging in deep regions, which is typically more difficult due to the placement of the sources and receiver near the surface.

Samples from the posterior provide access to useful statistical information including approximations to moments of the distribution such as the mean and pointwise standard deviation (Figure 5.7). We compute the mean of the posterior samples to obtain the conditional mean estimate, i.e., the expected value of the posterior distribution. This estimate is depicted in Figure 5.7a. From Figure 5.7a, we observe that the overall amplitudes are well recovered by the conditional mean estimate, which includes partially recovered reflectors in badly illuminated areas close to the boundaries. Although the reconstructions are not perfect, they significantly improve upon the reverse-time migrations estimate. We did not observe a significant increase in the signal-to-noise ratio (SNR) of the conditional mean estimate when more than one thousand samples from the posterior are drawn. We use the one thousand samples to also estimate the pointwise standard deviation (Figure 5.7b), which serves as an assessment of the uncertainty. To avoid bias from strong amplitudes in the estimated image, we also plot the stabilized division of the standard deviation by the envelope of the conditional mean in Figure 5.7c. As expected, the pointwise standard deviation in Figures 5.7b and 5.7c indicate that we have the most uncertainty in areas of complex geology—e.g., near channels and tortuous reflectors, and in areas with a relatively poor illumination (deep and close to boundaries). The areas with large uncertainty align well with difficult-to-image parts of the model.

After incurring an upfront cost of training the conditional normalizing flow, the computational cost of sampling the posterior distribution is low as it does not involve any forward operator evaluations. However, the accuracy of the presented results is directly linked to the availability of high-quality training data that fully represent the joint distribution for model and data. Due to our lack of access to the subsurface of the Earth, obtaining high-quality

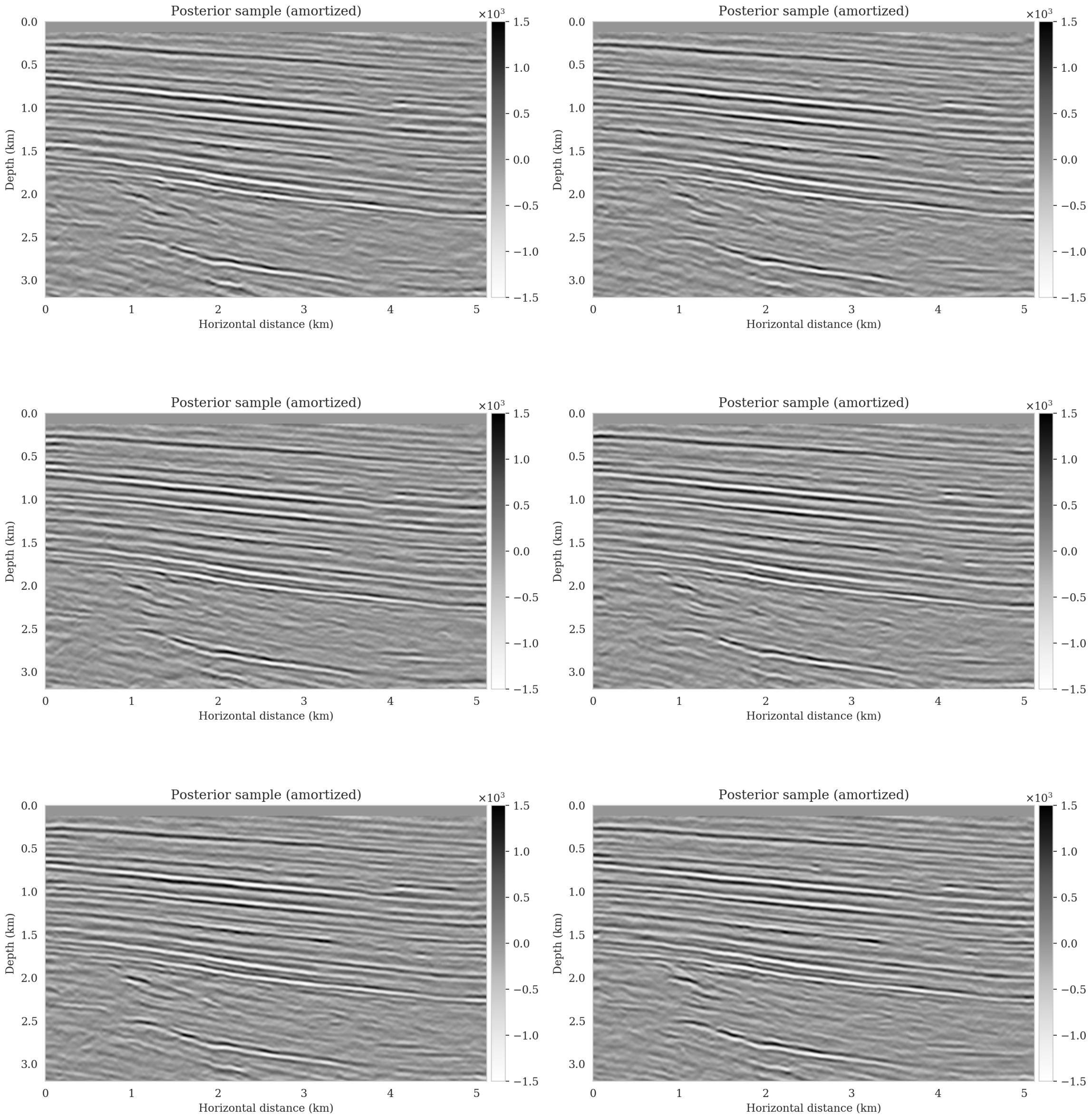
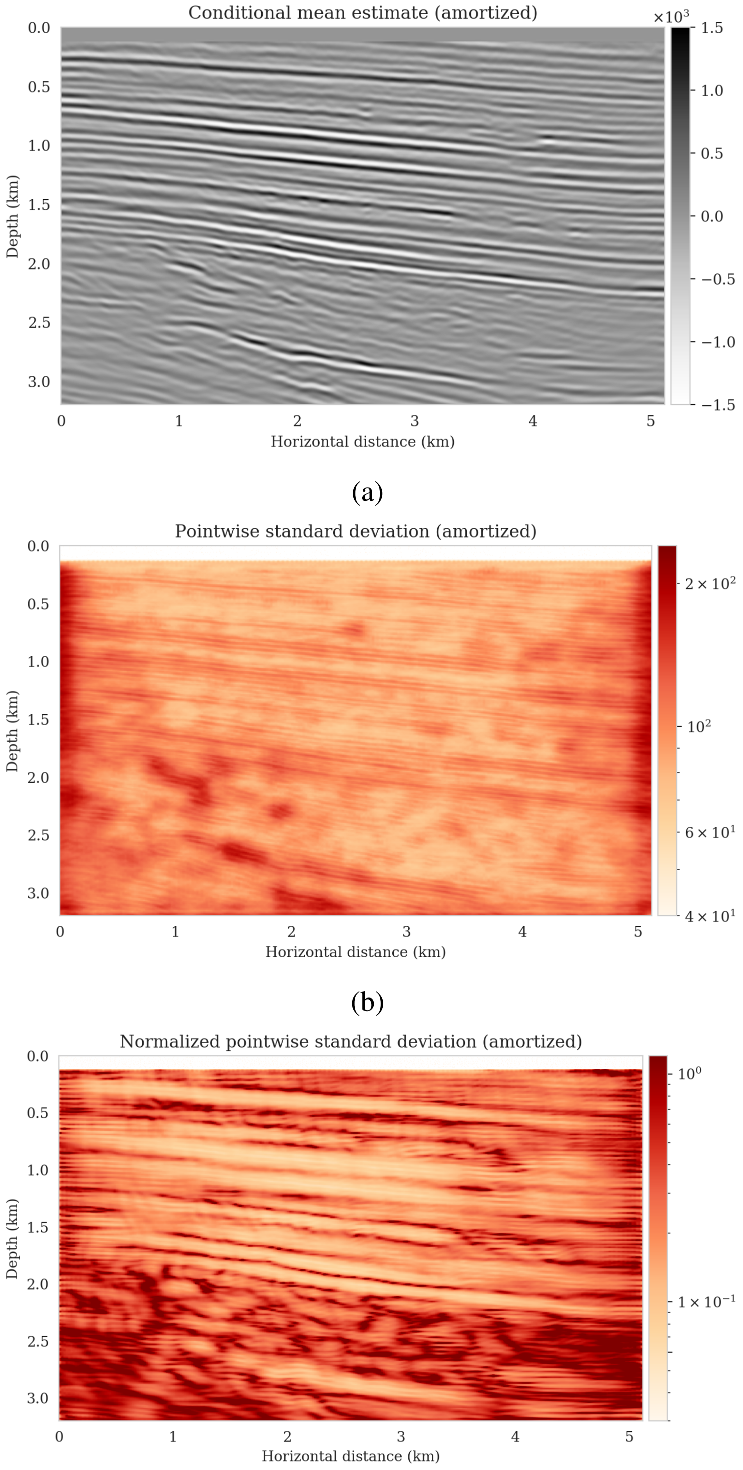


Figure 5.6: Samples drawn from posterior distribution using the pretrained conditional normalizing flow via equation 5.9 with SNRs ranging from $8.08\,\mathrm{dB}$ to $8.92\,\mathrm{dB}$.



(a)

(b)

(c)

Figure 5.7: Amortized variational inference results. (a) The conditional (posterior) mean estimate with SNR $9.44\,\mathrm{dB}$. (b) The pointwise standard deviation among samples drawn from the posterior. (c) Normalized pointwise standard deviation by the conditional mean estimate (Figure 5.7a).

training data is challenging when dealing with geophysical inverse problems. To address this issue, we propose to supplement amortized variational inference with a physics-based latent distribution correction technique that increases the reliability of this approach when dealing with moderate shifts in the data distribution during inference.

5.5 A physics-based treatment to data distribution shifts during inference

For accurate Bayesian inference in the context of amortized variational inference, the surrogate conditional distribution $p_{\phi}(\mathbf{x} \mid \mathbf{y})$ must yield a zero amortized variational inference objective value (equation 5.8). Achieving this objective is challenging due to lack of access to model and data pairs that sufficiently captures the underlying joint distribution in equation 5.8. Additionally, due to potential shifts to the joint distribution during inference, i.e., shifts in the prior distribution or the forward (likelihood) model (equation 5.1), the conditional normalizing flow can no longer reliably provide samples from the posterior distribution due to lack of generalization. Under such conditions feeding latent samples drawn from a standard Gaussian distribution to the conditional normalizing flow may lead to posterior sampling errors. To quantify the posterior distribution approximation error and to propose our correction method, we will use the invariance of the KL divergence to differentiable and invertible mappings [74]. This property allows use to relate the errors that the conditional normalizing flow makes in approximating the posterior distribution to the errors that it makes in Gaussianizing the input model and data pairs.

5.5.1 KL divergence invariance relation

The errors that the pretrained conditional normalizing flows makes in approximating the posterior distribution can be formally quantified using the invariance of the KL divergence under diffeomorphism mappings [74]. Using this relation, we relate the posterior distribution approximation errors to the errors that the conditional normalizing flow makes in turning its inputs to latent samples drawn from a standard multivariate Gaussian distribution. Specifically, for observed data $\mathbf{y}_{\text{obs}}$ drawn from a shifted data distribution $\widehat{p}_{\text{data}}(\mathbf{y}) \neq p_{\text{data}}(\mathbf{y})$, the invariance relation states

$$\mathbb{KL}\ (p_\phi(\mathbf{z} \mid \mathbf{y}_{\text{obs}}) \ ||\ \mathrm{N}(\mathbf{z} \mid \mathbf{0}, \mathbf{I})) = \mathbb{KL}\ (p_{\text{post}}(\mathbf{x} \mid \mathbf{y}_{\text{obs}}) \ ||\ p_\phi(\mathbf{x} \mid \mathbf{y}_{\text{obs}})) > 0. \tag{5.15}$$

In this expression, $p_\phi(\mathbf{z} \mid \mathbf{y}_{\text{obs}})$ representing the distribution of conditional normalizing flow output $\mathbf{z} = f_\phi(\mathbf{x}; \mathbf{y}_{\text{obs}})$ for inputs $\mathbf{x} \sim p(\mathbf{x} \mid \mathbf{y}_{\text{obs}})$. We refer to this distribution as the shifted latent distribution as it is the result of a data distribution shift translated through the conditional normalizing flow to the latent space. Equation 5.15 states that the conditional normalizing flow fails to accurately Gaussianize the input models $\mathbf{x} \sim p(\mathbf{x} \mid \mathbf{y}_{\text{obs}})$ for the given data $\mathbf{y}_{\text{obs}}$. Failure to take into account the mismatch between the shifted latent distribution $p_\phi(\mathbf{z} \mid \mathbf{y}_{\text{obs}})$ and $\mathrm{N}(\mathbf{z} \mid \mathbf{0}, \mathbf{I})$ leads to posterior sampling errors as the KL divergence between the predicted and true posterior distributions is nonzero (equation 5.15). In other words, feeding latent samples drawn from a standard Gaussian distribution to $f_\phi^{-1}(\mathbf{z}; \mathbf{y}_{\text{obs}})$ produces samples from $p_\phi(\mathbf{x} \mid \mathbf{y}_{\text{obs}})$, which does not accurately approximate the true posterior distribution under the assumption of data distribution shift. On the other hand, with the same reasoning via the KL divergence invariance relation, feeding samples from the shifted latent distribution $p_\phi(\mathbf{z} \mid \mathbf{y}_{\text{obs}})$ to the conditional normalizing flow yields accurate posterior samples. However, obtaining samples from $p_\phi(\mathbf{z} \mid \mathbf{y}_{\text{obs}})$ is not trivial as we do not have a closed-form expression for its density. In the next section, we introduce a physics-based approximation to the shifted-latent distribution.

5.5.2 Physics-based latent distribution correction

Ideally, performing accurate posterior sampling via the pretrained conditional normalizing flow—in the presence of data distribution shifts—requires passing samples from the shifted latent distribution $p_\phi(\mathbf{z} \mid \mathbf{y}_{\text{obs}})$ to $f_\phi^{-1}(\mathbf{z}; \mathbf{y}_{\text{obs}})$. Unfortunately, accurately sampling $p_\phi(\mathbf{z} \mid \mathbf{y}_{\text{obs}})$ requires access to the true posterior distribution, which we are ultimately after and do not have access to. Alternatively, we propose to quantify $p_\phi(\mathbf{z} \mid \mathbf{y}_{\text{obs}})$ using Bayes' rule,

$$p_\phi(\mathbf{z} \mid \mathbf{y}_{\text{obs}}) = \frac{p_{\text{like}}(\mathbf{y}_{\text{obs}} \mid \mathbf{z})\, p_{\text{prior}}(\mathbf{z})}{\widehat{p}_{\text{data}}(\mathbf{y}_{\text{obs}})}, \tag{5.16}$$

where the physics-informed likelihood function $p_{\text{like}}(\mathbf{y}_{\text{obs}} \mid \mathbf{z})$ and the prior distribution $p_{\text{prior}}(\mathbf{z})$ over the latent variable are defined as

$$\begin{aligned} -\log p_\phi(\mathbf{z} \mid \mathbf{y}_{\text{obs}}) &= -\sum_{i=1}^{N} \log p_{\text{like}}(\mathbf{y}_{\text{obs},i} \mid \mathbf{z}) - \log p_{\text{prior}}(\mathbf{z}) + \log \widehat{p}_{\text{data}}(\mathbf{y}_{\text{obs}}) \\ &:= \frac{1}{2\sigma^2} \sum_{i=1}^{N} \left\| \mathbf{y}_{\text{obs},i} - \mathcal{F}_i \circ f_\phi^{-1}(\mathbf{z}; \mathbf{y}_{\text{obs}}) \right\|_2^2 + \frac{1}{2} \left\| \mathbf{z} \right\|_2^2 + \text{const.} \end{aligned} \tag{5.17}$$

In the above expression, the physics-informed likelihood function $p_{\text{like}}(\mathbf{y}_{\text{obs}} \mid \mathbf{z})$ follows from the forward model in equation 5.1 with a Gaussian assumption on the noise with mean zero and covariance matrix $\sigma^2\mathbf{I}$, and the prior distribution $p_{\text{prior}}(\mathbf{z})$ is chosen as a standard Gaussian distribution with mean zero and covariance matrix $\mathbf{I}$. The choice of the likelihood function ensures physics and data fidelity by giving more importance to latent variables that once passed through the pretrained conditional normalizing flow and the forward operator provide smaller data misfits while the prior distribution $p_{\text{prior}}(\mathbf{z})$ injects our prior beliefs about the latent variable, which is by design chosen to be distributed according to a standard Gaussian distribution.

Due to our choice of the likelihood function and prior distribution above, the effective prior distribution over the unknown $\mathbf{x}$ is in fact a conditional prior characterized by the pretrained conditional normalizing flow [40]. As observed by [75] and [34], using a conditional prior may be more informative than its unconditional counterpart because it is conditioned by the observed data $\mathbf{y}_{\text{obs}}$. Our approach can be also viewed as an instance of online variational Bayes [76] where data arrives sequentially and previous posterior approximates are used as priors for subsequent approximations.

In the next section, we improve the available amortized approximation to the posterior distribution by relaxing the standard Gaussian distribution assumption of the conditional normalizing flow latent distribution.

Gaussian relaxation of the latent distribution

By definition, feeding samples from $p_\phi(\mathbf{z} \mid \mathbf{y}_{\text{obs}})$ to the pretrained amortized conditional normalizing flows provides samples from the posterior distribution (see discussion beneath equation 5.15). To maintain the low computational cost of sampling with amortized variational inference, it is imperative that $p_\phi(\mathbf{z} \mid \mathbf{y}_{\text{obs}})$ is sampled as cheaply as possible. To this end, we exploit the fact that conditional normalizing flows in the context of amortized variational inference are trained to Gaussianize the input model random variable (equation 5.11). This suggests that the shifted latent distribution $p_\phi(\mathbf{z} \mid \mathbf{y}_{\text{obs}})$ will be close to a standard Gaussian distribution for a certain class of data distribution shifts. We exploit this property and approximate the shifted latent distribution $p_\phi(\mathbf{z} \mid \mathbf{y}_{\text{obs}})$ via a Gaussian distribution with an unknown mean and diagonal covariance matrix,

$$p_\phi(\mathbf{z} \mid \mathbf{y}_{\text{obs}}) \approx \mathrm{N}\big(\mathbf{z} \mid \boldsymbol{\mu}, \mathrm{diag}(\mathbf{s})^2\big), \quad \mathbf{z} \in \mathcal{Z}. \tag{5.18}$$

In the above expression, the vector $\boldsymbol{\mu}$ corresponds to the mean and the vector $\mathrm{diag}(\mathbf{s})^2$ represents a diagonal covariance matrix with diagonal entries s $\odot$ s (with the symbol $\odot$ denoting elementwise multiplication) that need to be determined. We estimate these quantities by minimizing the reverse KL divergence between the relaxed Gaussian latent distribution $\mathrm{N}\big(\mathbf{z} \mid \boldsymbol{\mu}, \mathrm{diag}(\mathbf{s})^2\big)$ and the shifted latent distribution $p_\phi(\mathbf{z} \mid \mathbf{y}_{\text{obs}})$. According to the variational inference objective function associated with the reverse KL divergence in equation 5.5, this correction can be achieved by solving the following optimization problem,

$$
\begin{aligned}
\boldsymbol{\mu}^*, \mathbf{s}^* &= \operatorname*{arg\,min}_{\boldsymbol{\mu}, \mathbf{s}} \mathbb{KL} \left(\mathrm{N}\big(\mathbf{z} \mid \boldsymbol{\mu}, \mathrm{diag}(\mathbf{s})^2\big) \mid\mid p_{\phi}(\mathbf{z} \mid \mathbf{y}_{\text{obs}}) \right) \\
&= \operatorname*{arg\,min}_{\boldsymbol{\mu}, \mathbf{s}} \mathbb{E}_{\mathbf{z} \sim \mathrm{N}(\mathbf{z} \mid \mathbf{0}, \mathbf{I})} \Bigg[\frac{1}{2\sigma^2} \sum_{i=1}^{N} \big\| \mathbf{y}_{\text{obs},i} - \mathcal{F}_i \circ f_{\phi}\big(\mathbf{s} \odot \mathbf{z} + \boldsymbol{\mu}; \mathbf{y}_{\text{obs}}\big) \big\|_2^2 \\
&\qquad\qquad + \frac{1}{2} \big\| \mathbf{s} \odot \mathbf{z} + \boldsymbol{\mu} \big\|_2^2 - \log \Big| \det \mathrm{diag}(\mathbf{s}) \Big| \Bigg].
\end{aligned}
\tag{5.19}
$$

We solve optimization problem 5.19 with the Adam optimizer where we select random batches of latent variable variables $\mathbf{z} \sim \mathrm{N}(\mathbf{z} \mid \mathbf{0}, \mathbf{I})$ and data indices. We initialize the optimization problem 5.19 by $\boldsymbol{\mu} = \mathbf{0}$ and $\mathrm{diag}(\mathbf{s})^2 = \mathbf{I}$. This initialization acts as a warm-start and an implicit regularization [40] since $f_{\phi}^{-1}(\mathbf{z}; \mathbf{y}_{\text{obs}})$ for standard Gaussian distributed latent samples $\mathbf{z}$ provides approximate samples from the posterior distribution—thanks to amortization over different observed data $\mathbf{y}$. As a result, we expect the optimization problem 5.19 to be solved relatively cheaply. Additionally, the imposed standard Gaussian distribution prior on $\mathbf{s} \odot \mathbf{z} + \boldsymbol{\mu}$ regularizes inversion for the corrections since $\mathbb{KL} \left(p_{\phi}(\mathbf{z} \mid \mathbf{y}_{\text{obs}}) \mid\mid \mathrm{N}(\mathbf{z} \mid \mathbf{0}, \mathbf{I}) \right)$ is minimized during amortized variational inference (equation 5.15). To relax the (conditional) prior imposed by the pretrained conditional normalizing flow, instead of a standard Gaussian prior, a Gaussian prior with a larger variance can be imposed on the corrected latent variable. Conditional normalizing flows' inherent invertibility allows the normalizing flow to represent any solution $\mathbf{x} \in \mathcal{X}$ in the solution space. This has the additional benefit of limiting the adverse affects of imperfect pretraining of f_{ϕ} in domains where access to high-fidelity training data is limited, which is often the case in practice. Figure 5.8 summarizes our proposed method latent distribution correction method.

Inference with corrected latent distribution

Once the optimization problem 5.19 is solved with respect to $\boldsymbol{\mu}$ and $\mathbf{s}$, we obtain corrected posterior samples by passing samples from the corrected latent distribution $\mathrm{N}\big(\mathbf{z} \mid$

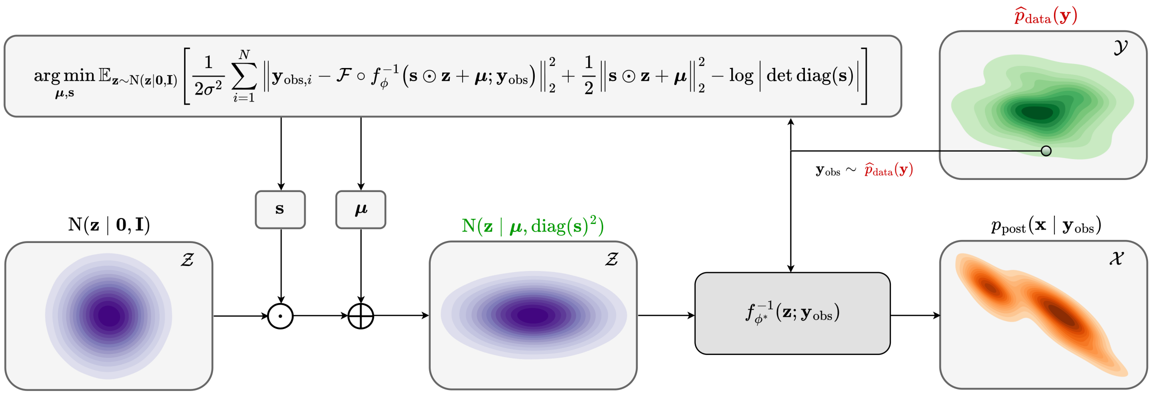

$\widehat{p}_{\mathrm{data}}(\mathbf{y})$
$\mathcal{Y}$
$\underset{\boldsymbol{\mu},\mathbf{s}}{\arg\min}\,\mathbb{E}_{\mathbf{z}\sim\mathrm{N}(\mathbf{z}\mid\mathbf{0},\mathbf{I})}\left[\frac{1}{2\sigma^2}\sum_{i=1}^{N}\left\|\mathbf{y}_{\mathrm{obs},i}-\mathcal{F}\circ f_{\phi}^{-1}(\mathbf{s}\odot\mathbf{z}+\boldsymbol{\mu};\mathbf{y}_{\mathrm{obs}})\right\|_2^2+\frac{1}{2}\left\|\mathbf{s}\odot\mathbf{z}+\boldsymbol{\mu}\right\|_2^2-\log\left|\det\mathrm{diag}(\mathbf{s})\right|\right]$
$\mathbf{y}_{\mathrm{obs}}\sim\widehat{p}_{\mathrm{data}}(\mathbf{y})$
$\mathbf{s}$
$\boldsymbol{\mu}$
$\mathrm{N}(\mathbf{z}\mid\mathbf{0},\mathbf{I})$
$\mathcal{Z}$
$\mathrm{N}(\mathbf{z}\mid\boldsymbol{\mu},\mathrm{diag}(\mathbf{s})^2)$
$\mathcal{Z}$
$f_{\phi^*}^{-1}(\mathbf{z};\mathbf{y}_{\mathrm{obs}})$
$p_{\mathrm{post}}(\mathbf{x}\mid\mathbf{y}_{\mathrm{obs}})$
$\mathcal{X}$

Figure 5.8: A schematic representation of out proposed method. When dealing with nonzero amortized variational inference objective value (equation 5.7) or in presence of data distribution shifts during inference, we adapt the latent distribution of the pretrained conditional normalizing flow via a diagonal physics-based correction. After the correction, the new latent samples result in corrected posterior samples when fed to the pretrained normalizing flow.

$\boldsymbol{\mu}^*,\mathrm{diag}(\mathbf{s}^*)^2\big)\approx p_\phi(\mathbf{z}\mid\mathbf{y}_{\mathrm{obs}})$ to the conditional normalizing flow,

$$\mathbf{x}=f_\phi^{-1}(\mathbf{s}^*\odot\mathbf{z}+\boldsymbol{\mu}^*;\mathbf{y}_{\mathrm{obs}}),\quad \mathbf{z}\sim\mathrm{N}(\mathbf{z}\mid\mathbf{0},\mathbf{I}). \tag{5.20}$$

These corrected posterior samples are implicitly regularized by the reparameterization with the pretrained conditional normalizing flow and the standard Gaussian distribution prior on $\mathbf{z}$ [40, 34]. Next section applies this physics-based correction to a seismic imaging example, in which we use the pretrained conditional normalizing flow from the earlier example.

5.6 Latent distribution correction applied to seismic imaging

The purpose of our proposed latent distribution correction approach is to accelerate Bayesian inference while maintaining fidelity to a specific observed dataset and physics. While this method is generic and can be applied to a variety of inverse problems, it is particularly relevant when solving geophysical inverse problems, where the unknown quantity is high dimensional, the forward operator is computationally costly to evaluate, and there is a lack of access to high-quality training data that represents the true heterogeneity of the Earth's subsurface. Therefore, we apply this approach to seismic imaging to utilize the advantages of generative models for solving inverse problems, including fast conditional sampling and learned prior distributions, while limiting the negative bias induced by shifts in data distributions.

The results are presented for two cases. The first case involves introducing a series of changes in the distribution of observed data, for example changing the number of sources and the noise levels. This is followed by correcting for the error in predictions made by the pretrained conditional normalizing flow using our proposed method. In the second case, in addition to the shifts in the distribution of the observed data (forward model), we also introduce a shift in the prior distribution. We accomplish this by selecting a ground-truth image from a deeper section of the Parihaka dataset that has different image characteristics than the training images, such as tortuous reflectors and more complex geological features. In both cases, we expect the outcome of the Bayesian algorithm to improve following the correction of latent distributions described above. We provide qualitative and quantitative evaluations of the Bayesian inference results.

5.6.1 Shift in the forward model

In the following example, we introduce shifts in the distribution of observed data—compared to the pretraining phase—by changing the forward model. The shift involves reducing the number of sources (N in equation 5.1) by a factor of two to four, while adding band-limited noise with 1.5 to three times larger standard deviation (σ in equation 5.3). We will demonstrate the potential pitfalls of relying solely on the pretrained conditional normalizing flows in circumstances where the distribution of observed data has shifted. With the use of our latent distribution correction, we will demonstrate that we are able to correct for errors that are made by the pretrained conditional normalizing flow as a result of changes to data distribution.

Following the description of the problem setup, we will also provide comparisons between the conditional mean estimation quality before and after the latent distribution correction step. Before moving on to our results relating to uncertainty quantification on the image, we demonstrate the importance of the correction step by visualizing the improvements in fitting the observed data. Lastly, we perform a series of experiments to verify our Bayesian inference results.

Problem setup

To induce shifts in the data distribution, we reduce the number of sources and increase the standard deviation of the added band-limited noise. Consequently, we have reduced the amount of data (due to having fewer source experiments) and decreased the SNR of each shot record (due to being contaminated with stronger noise). As a consequence, seismic imaging becomes more challenging, i.e., more difficult to estimate the ground truth image, and it is expected that the uncertainty associated with the problem will also increase.

We use the same ground-truth image as in the previous example (Figure 5.5a), while experimenting with 25, 51, 102 sources and adding band-limited noise that has 1.5, 2.0, 2.5, and 3.0 times larger standard deviation than the pretraining setup. For each combination of source number and noise level (12 combinations in total), we compute the reverse-time migrated image corresponding to that combination. Next, we perform latent distribution corrections for each of the 12 seismic imaging instances. All latent distribution correction optimization problems (equation 5.19) are solved using the Adam optimization algorithm [60] for five passes over the shot records (epochs). We did not observe a significant decrease in the objective function after five epochs. The objective function is evaluated each iteration by drawing a single latent sample from the standard Gaussian distribution and randomly selecting (without replacement) a data index $i \in \{1, \ldots, 25\}$. We use a stepsize of 10^{-1} and decrease it by a factor of 0.9 at the end of every two epochs.

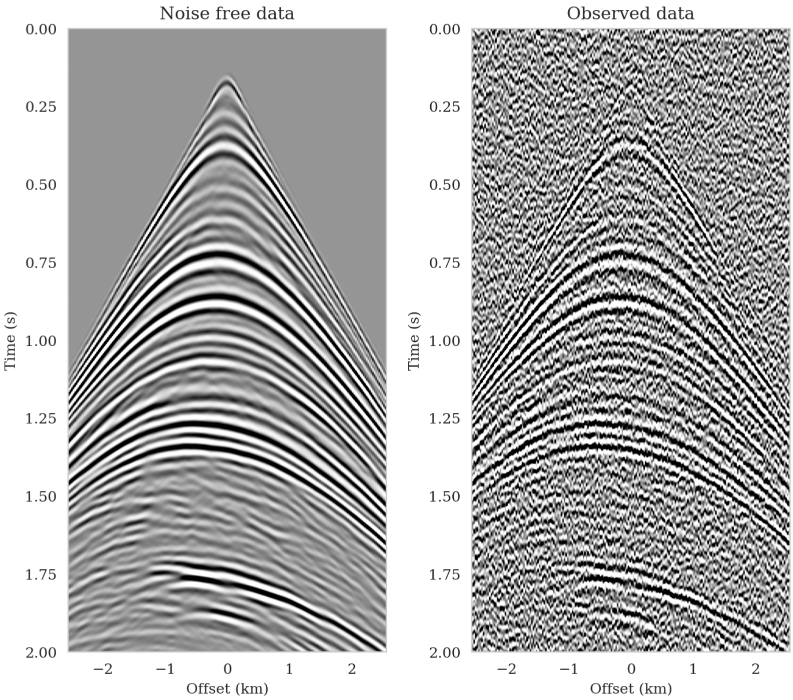
After solving the optimization problem 5.19 for the different seismic imaging instances, we obtain corrected posterior samples for each instance. The next section provides a detailed discussion of the latent distribution correction that was applied to one such instance that had a significant shift in data distribution.

Improved Bayesian inference via latent distribution correction

The aim of this section is to demonstrate how latent distribution correction can be used to mitigate errors induced by data distribution shifts. Specifically, we present the results for the case where the number of sources is reduced by a factor of four ($N = 25$) as compared to the pretrained data generation setup. The 25 sources are spread periodically over the survey area with a source sampling of approximately 200 meters. Moreover, we contaminate the resulting shot records with band-limited noise with an increased standard deviation of 2.5 times when compared to the pretraining phase. The overall SNR for the data thus becomes $-2.78\,\mathrm{dB}$, which is $7.95\,\mathrm{dB}$ lower that the SNR of the observed data during pretraining (Figure 5.4b). Figure 5.9 shows one of the ‘$\texttt{25}$ shot records.

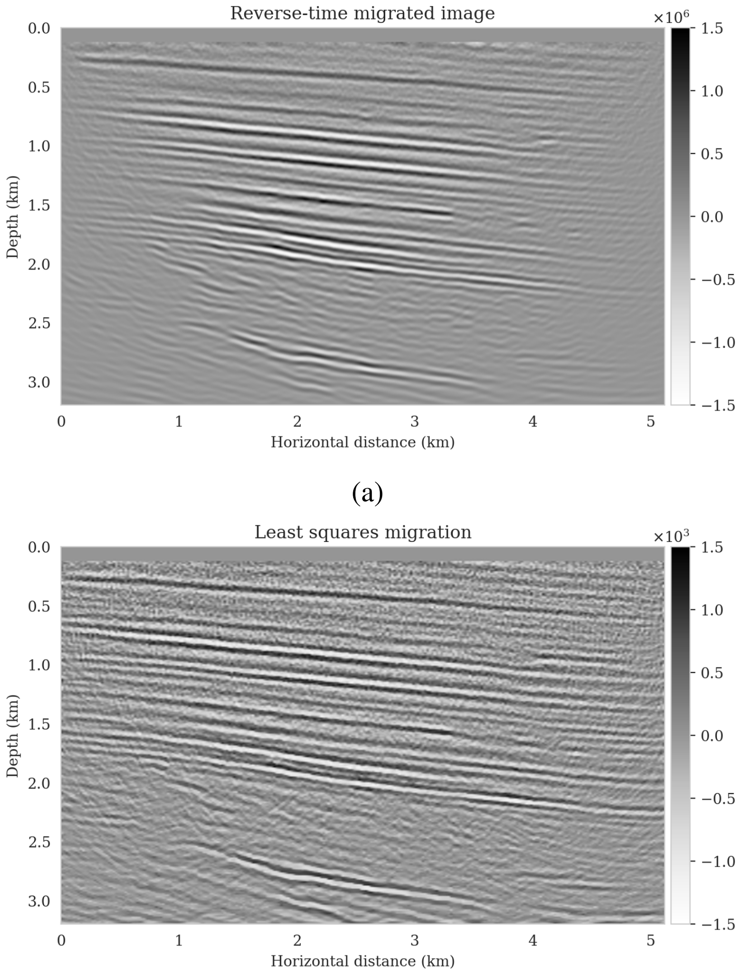
Utilizing the above observed dataset, we compute the reverse-time migrated image as an input to our pretrained conditional normalizing flow (Figure 5.10a). In contrast to the reverse-time migrated image shown in Figure 5.4b, this migrated image is, as expected, noisier, and it displays visible near-source imaging artifacts as a result of coarse source sampling. Additionally, we compute the least-squares migrated seismic image that is obtained by minimizing the negative-log likelihood (see the likelihood term in equation 5.3). This image, shown in Figure 5.10b, was constructed by fitting the data without incorporating any prior information. It is evident from this image that there are strong artifacts caused by noise in the data, underscoring the importance of incorporating prior knowledge into solving seismic imaging.

Improvements in posterior samples and conditional mean estimate To obtain amortized (uncorrected) posterior samples, we feed the reverse-time migrated image (Figure 5.10a) and latent samples drawn from the standard Gaussian distribution to the pretrained normal-



(a) (b)

Figure 5.9: A shot record from the shifted data distribution. (a) Noise-free linearized data (same as Figure 5.4a). (b) Noisy linearized data with 2.5 larger band-limited noise standard deviation (SNR $-2.78\,\mathrm{dB}$).



(a)

(b)

Figure 5.10: Latent distribution correction experiment setup. (a) Reverse-time migrated image corresponding to the shifted forward model with SNR $-8.22\,\mathrm{dB}$. (b) Least squares imaging, which is equivalent to the minimizing $\sum_{i=1}^{N} \left\| \mathbf{d}_i - \mathbf{J}(\mathbf{m}_0, \mathbf{q}_i)\delta\mathbf{m} \right\|_2^2$ with respect to $\delta\mathbf{m}$ with no regularization. The SNR for this estimate is $6.90\,\mathrm{dB}$.

izing flow. These samples, which are shown in the left column of Figure 5.11, contain artifacts near the top of the image. These artifacts are related to the near-source reverse-time migrated image artifacts (Figure 5.10a). Since the reverse-time migrated images used during pretraining do not contain near-source imaging artifacts—due to fine source sampling—the pretrained normalizing flow fails to eliminate them. Further, the uncorrected posterior samples do not accurately predict reflectors as they approach the boundaries and deeper sections of the image.

To illustrate the improved posterior sample quality following latent distribution correction, we feed latent samples drawn from the corrected latent distribution to the pretrained normalizing flow (right column of Figure 5.11). Comparing the left and right columns in Figure 5.11 indicates an improvement in the quality of samples from the posterior distribution, which can be attributed to the attenuation of near-top artifacts and an improvement in the image quality close to the boundary and deeper reflectors in the image. Moreover, the SNR values of the posterior samples after correction are approximately $3\,\mathrm{dB}$ higher, which represents a significant improvement.

To compute the conditional mean estimate, we simulate one thousand posterior samples before and after latent distribution correction. As with the posterior samples before correction, drawing samples after correction is very cheap once the correction is done as it only requires evaluating the conditional normalizing flow over the corrected latent samples. Figures 5.12a and 5.12b show conditional mean estimates before and after latent distribution correction, respectively. The conditional mean estimate before correction reveals similar artifacts as the posterior samples before correction, in particular, near-top imaging artifacts due to coarse sources sampling and less illumination of reflectors located closer to the boundary and deeper portions of the image. The importance of our proposed latent distribution correction can be observed by juxtaposing the conditional mean estimate before (Figure 5.12a) and after correction (Figure 5.12b). The conditional mean estimate obtained after latent distribution correction eliminates the aforementioned inaccuracies and

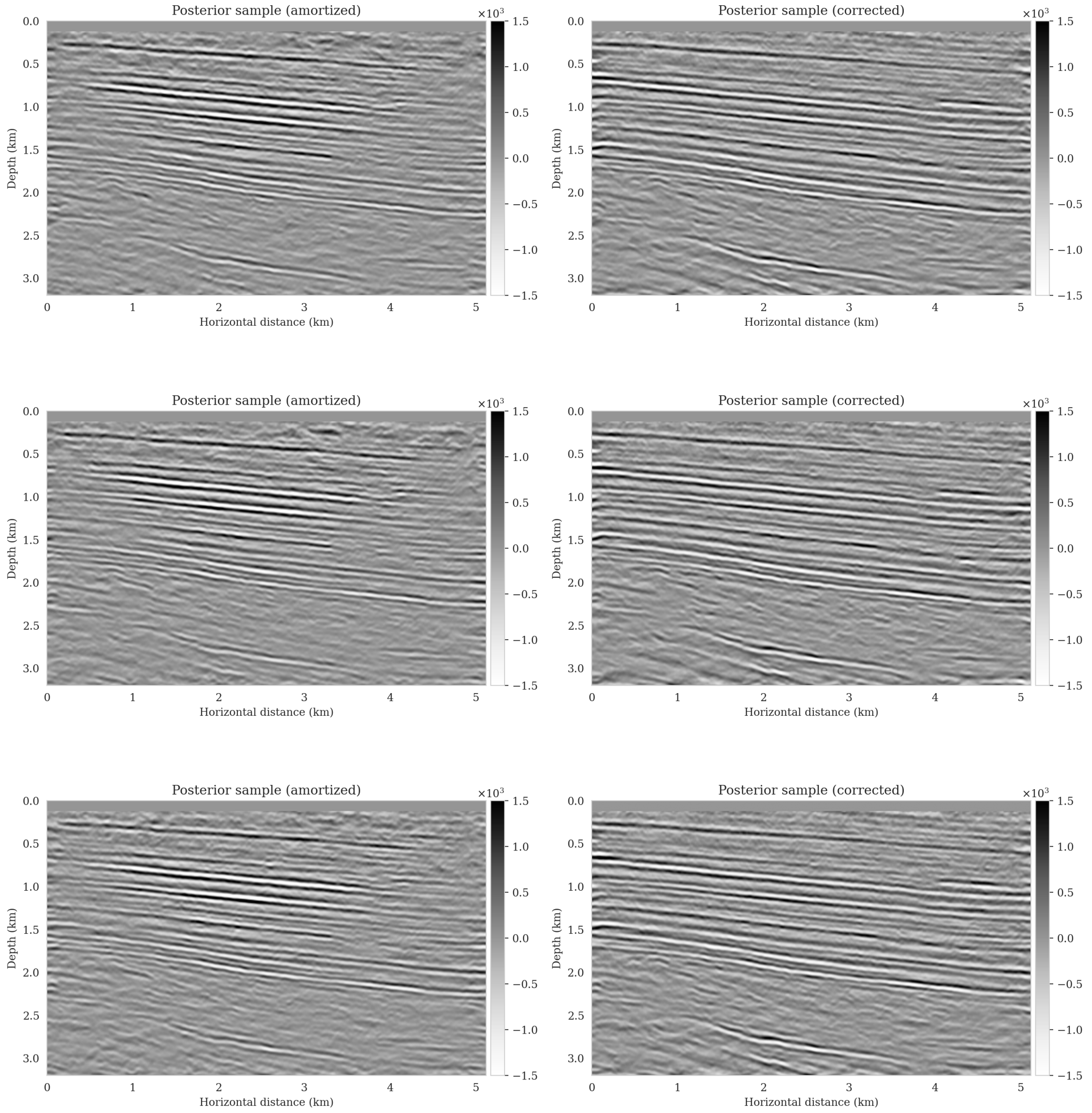
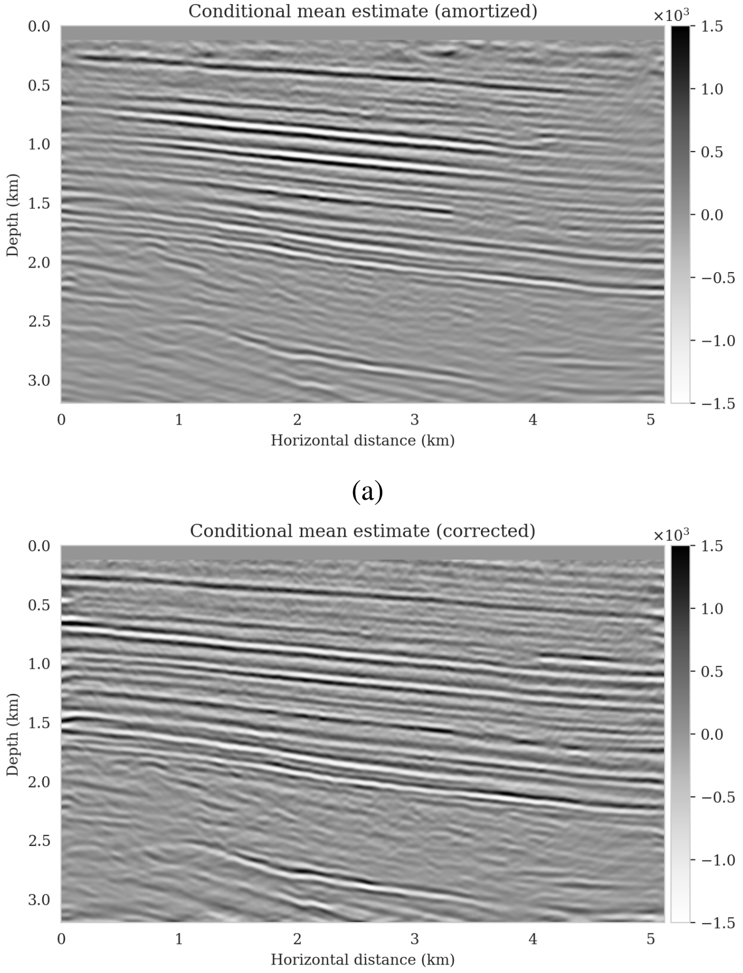


Figure 5.11: Samples from the posterior distribution (left) without latent distribution correction with SNRs ranging from $4.57\,\mathrm{dB}$ to $5.21\,\mathrm{dB}$; and (right) after latent distribution correction with SNRs ranging from $7.80\,\mathrm{dB}$ to $8.53\,\mathrm{dB}$.

enhances the quality of the image by approximately $4\,\mathrm{dB}$. We gain similar improvements in SNR compared to the least-squares migrated image (Figure 5.10b) with virtually the same cost, i.e., five passes over the shot records. This is significant improvement in SNR also is complimented by access to information regarding the uncertainty of the image.

Data-space quality control As the latent distribution correction step involves finding latent samples that are better suited to fit the data (equation 5.19), we can expect an improvement in fitting the observed data after correction. Predicted data is obtained by applying the forward operator to the conditional mean estimates, before and after latent distribution correction. Figures 5.13a and 5.13b show the predicted shot records before and after correction, respectively. In spite of the fact that both predicted data appear to be similar to ideal noise-free data (Figure 5.9a), the data residual associated with the conditional mean without correction reveals several coherent events that contain valuable information about the unknown seismic image. The latent distribution correction allows us to fit these coherent events as indicated by the data residual associated with the corrected conditional mean estimate.

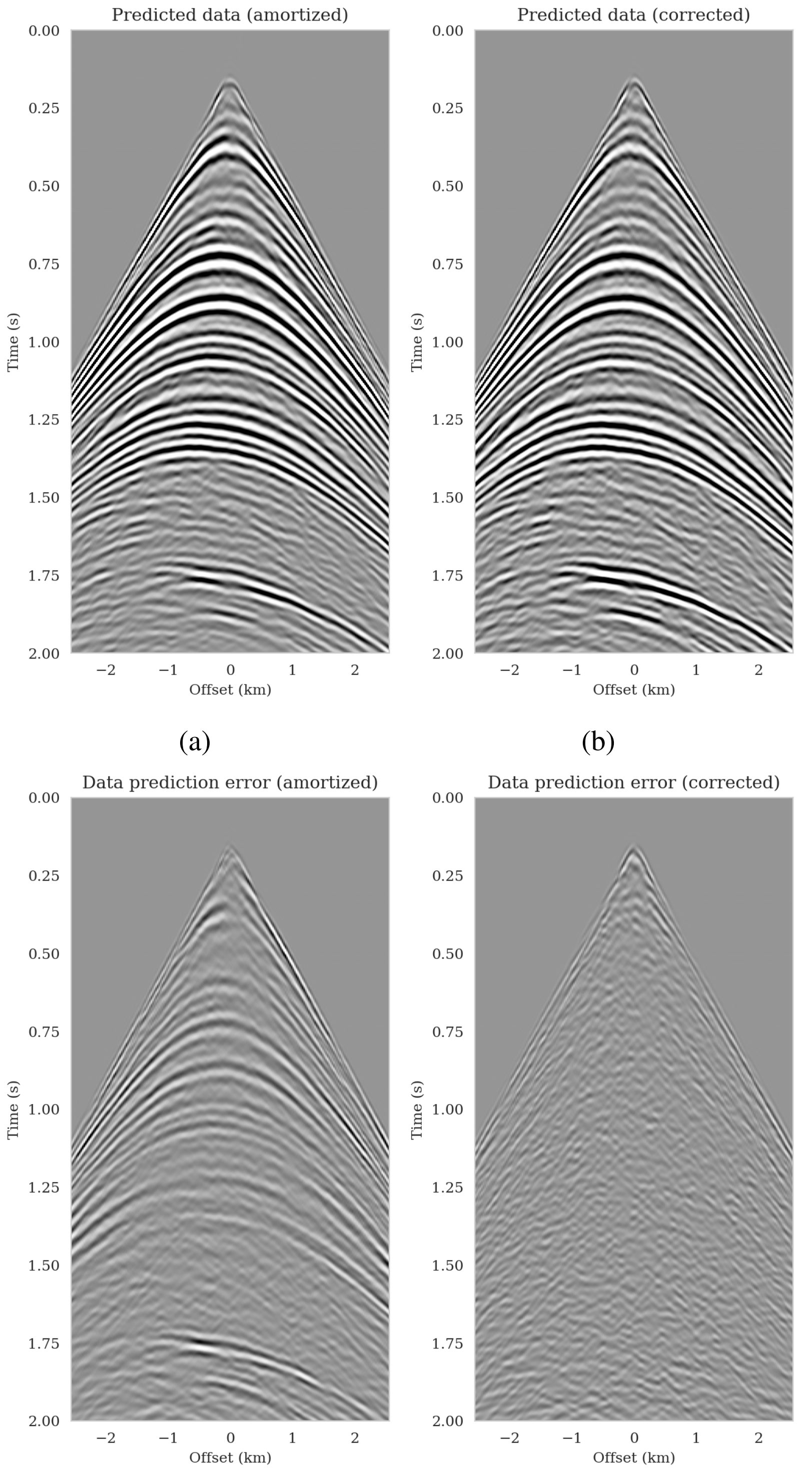
Uncertainty quantification—pointwise standard deviation and histograms We exploit cheap access to corrected samples from the posterior in order to extract information regarding uncertainty in the image estimates. Figure 5.14a displays the pointwise standard deviation among the one thousand corrected posterior samples. The overprint by the strong reflectors can be reduced by normalizing the standard deviation using a stabilized division by the conditional mean (Figure 5.14b). The pointwise standard deviation plots indicate high uncertainty in areas near the boundaries of the image and in the deep parts of the image where illumination is relatively poor. This observation is more evident in Figure 5.15, which displays three vertical profiles as 99% confidence intervals (orange colored shading) illustrating the expected increasing trend of uncertainty with depth. We additionally observe that the ground truth (dashed black) falls within the confidence intervals for most of



(a)

(b)

Figure 5.12: Improvements in conditional mean estimate due to latent distribution correction. (a) The conditional (posterior) mean estimate using the pretrained conditional normalizing flow without correction (SNR $6.29\,\mathrm{dB}$). (b) The conditional mean estimate after latent distribution correction (SNR $10.36\,\mathrm{dB}$).



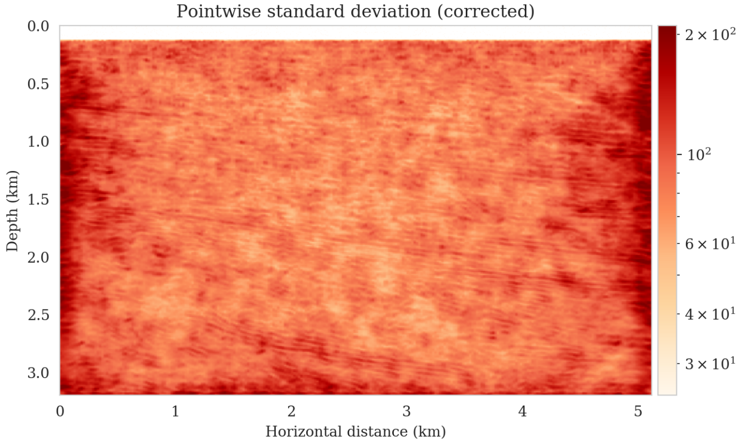
(a)

(b)

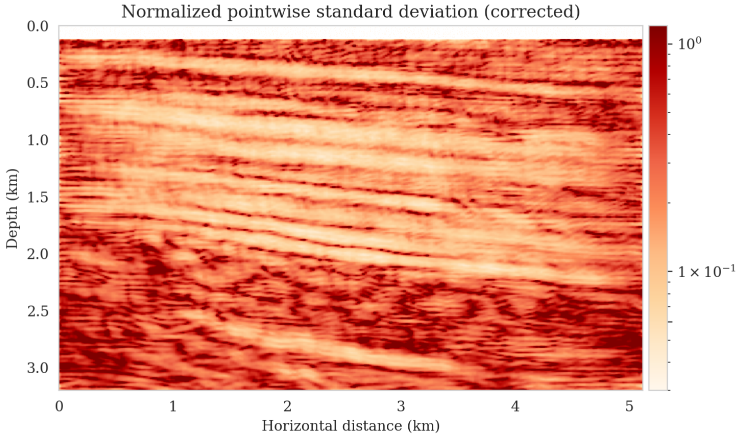
(c)

(d)

Figure 5.13: Quality control in data space. Data is simulated by applying the forward operator to the conditional mean estimate (a) before (SNR $11.62\,\mathrm{dB}$); and (b) after latent distribution correction (SNR $16.57\,\mathrm{dB}$). (c) Prediction errors associated with Figure 5.13a. (d) Prediction errors associated with Figure 5.13b (after latent distribution correction).



(a)

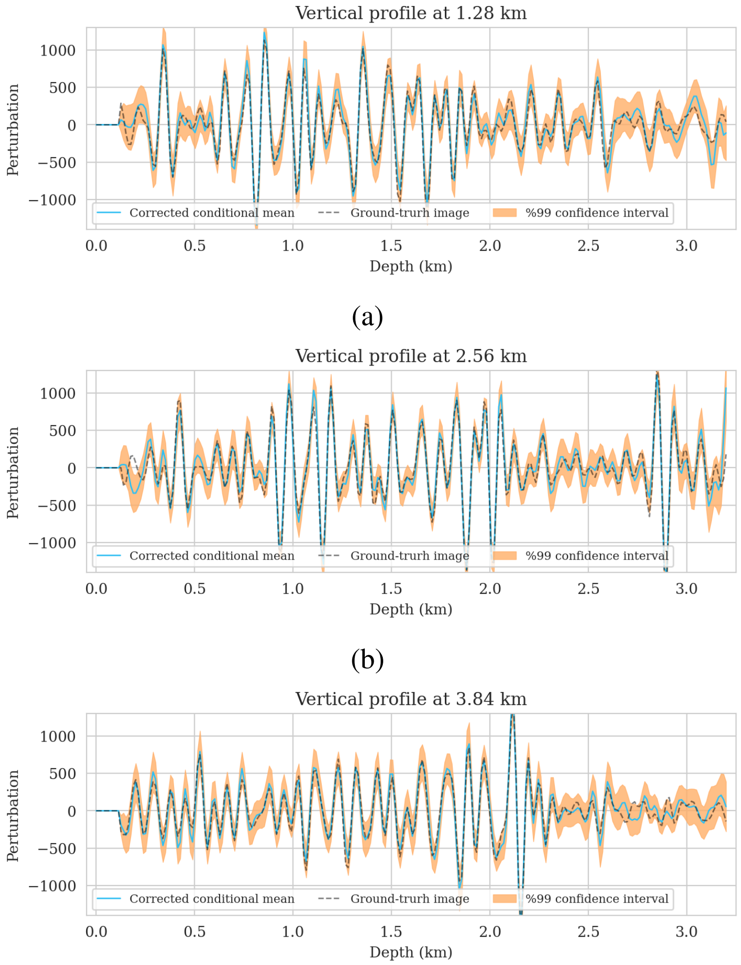


(b)

Figure 5.14: Uncertainty quantification with latent distribution correction. (a) The pointwise standard deviation among samples drawn from the posterior after latent distribution correction. (b) Normalized pointwise standard deviation by the conditional mean estimate (Figure 5.12b).

the areas.

To demonstrate how the corrected posterior is informed by the observed data, we calculated histograms at three locations in Figure 5.14a. Prior histograms are calculated by feeding latent samples drawn from the standard Gaussian distribution to the pretrained conditional normalizing flow without using data conditioning (see [31] and [30] for more information). These samples in the image spaces are indicative of samples from the prior distribution implicitly learned by the conditional normalizing flow during pretraining. The resulting prior histograms are shown in Figure 5.16. Corresponding histograms are also ob-



(a)

(b)

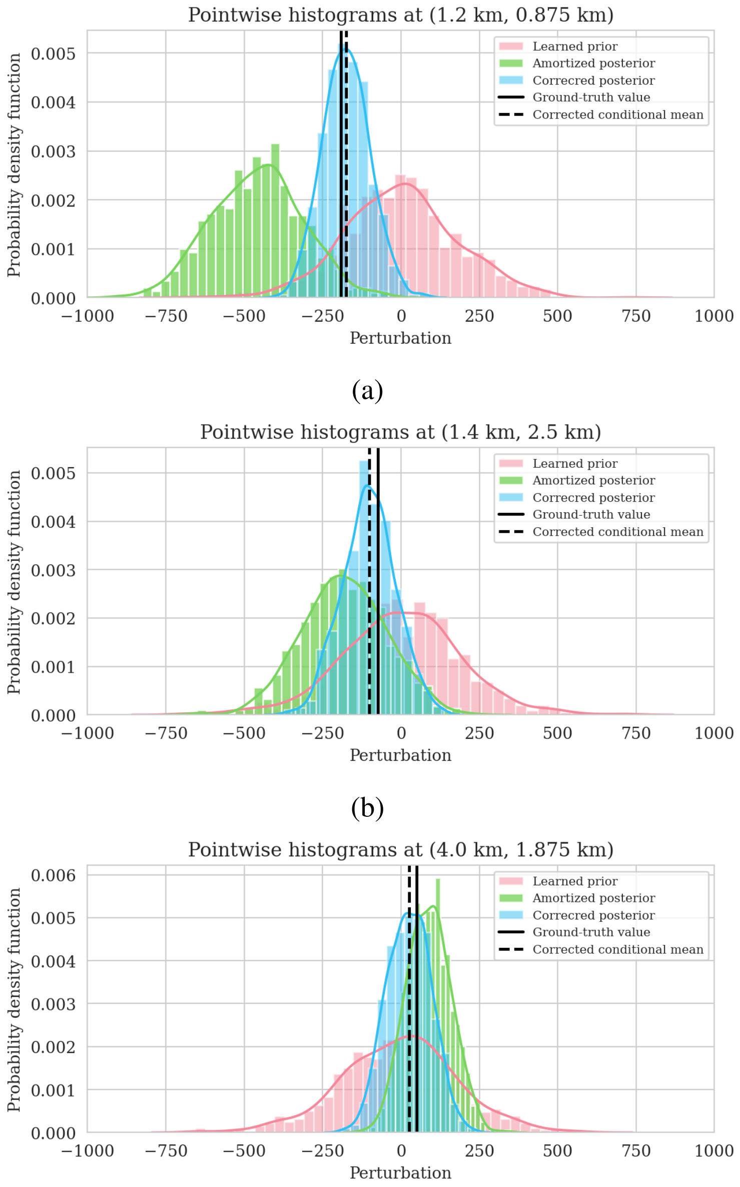
(c)

Figure 5.15: Confidence intervals for three vertical profiles. Traces of 99% confidence interval (shaded orange color), corrected conditional mean (solid blue), and ground truth (dashed black) at (a) $1.28\,\mathrm{km}$, (b) $2.56\,\mathrm{km}$, and (c) $3.84\,\mathrm{km}$ horizontal location.

tained for the uncorrected amortized posterior distribution (equation 5.12). As mentioned before, the uncorrected posterior distribution serves as an implicit conditional prior for the subsequent step of correction of the latent distribution. The green histograms in Figure 5.16 represent the uncorrected amortized posterior distribution. A similar procedure is followed to obtain histograms after latent distribution correction (blue histograms). As expected, the histograms of the posterior distribution are considerably narrower than those of the learned prior, which indicates that the posterior is further informed by the specific observed dataset and physics. As a means of evaluating the effect of latent distribution correction, we provide a vertical solid line showing the ground truth value's location. All three corrected posterior histograms for each location are shifted towards the ground truth, and their (conditional) mean plotted with the dashed vertical line indicates improved recovery of the ground truth. Compared to the amortized uncorrected histograms, the corrected histograms in Figures 5.16a – 5.16b are further contracted, suggesting that the latent distribution correction step has further informed the inference by the data.

Bayesian inference verification

While we investigated the accuracy of the conditional mean estimate after correction, we do not have access to the underlying true posterior distribution to verify our proposed posterior sampling method. This is partly due to our learned prior and the implicit conditional prior used in latent variable correction, which make traditional MCMC-based comparisons challenging. To further validate our Bayesian inference procedure, we conduct a series of experiments in which we investigate the effect of gradual increase in the number of sources (N) and reduction of the noise level. As the number of sources increases and the noise level decreases, we expect to see an increase in seismic image quality and a decrease in uncertainty.



(a)

(b)

(c)

Figure 5.16: Pointwise prior (red), uncorrected amortized posterior (green), and latent distribution corrected posterior (blue) histograms along with the true perturbation values (solid black line) and the corrected conditional mean (dashed black line) for points located at (a) $(1.2\,\mathrm{km},\ 0.875\,\mathrm{km})$, (b) $(1.4\,\mathrm{km},\ 2.5\,\mathrm{km})$, and (c) $(4.0\,\mathrm{km},\ 1.875\,\mathrm{km})$.

Estimation accuracy The accuracy of the Bayesian parameter estimation method is directly affected by the amount of data that has been collected [2]. That is to say with more observed data (larger N), we should be able to obtain a more accurate seismic image estimate. The same principle allows us to stack out noise when increasing the fold in seismic data acquisition. In order to assess whether our Bayesian inference approach has this property, we repeat the latent distribution correction process while varying the number of sources (using $N = 25, 51, 102$ sources) and the amount of band-limited noise (standard deviations $1.5, 2.0, 2.5$, and 3.0 times greater than the noise standard deviation during pretraining). Each of the 12 instances of latent distribution correction problems is treated similarly with respect to the number of passes made over the shot records and other optimization parameters. For each of the 12 combinations of source numbers and noise levels, we calculate the corrected conditional mean estimate and plot the SNRs as a function of the noise's standard deviation in Figure 5.17.

For a fixed number of sources (25, 51, and 102 sources shown with red, green, and blue colors respectively), we plot the corrected conditional means SNR as a function of the noise standard deviation. There is a clear increase in SNR trend as we decrease the noise level. In the same way, for each fixed noise level, the SNR increases with the number of sources. This verifies the our Bayesian inference method yields a more accurate estimate of the conditional mean for larger number of sources and smaller noise levels.

Bayesian posterior contraction An alternative Bayesian inference verification method involves analyzing the Bayesian posterior contraction, that is, the decrease of uncertainty with more data. To examine whether or not our Bayesian inference method possesses this property, we visually inspect the resulting pointwise standard deviation plots in Figure 5.18 for the 12 possible combinations of source numbers and noise levels. Each row corresponds to the pointwise standard deviation plot for a fixed noise standard deviation (σ), where the number of sources (N) decreases from left to right. In each column, we maintain

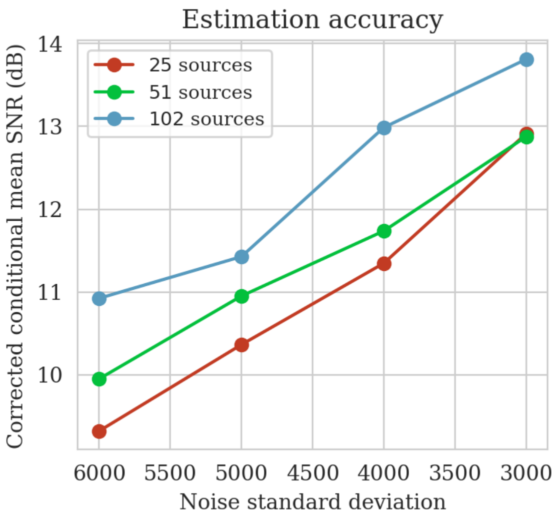


Figure 5.17: Estimation accuracy as a function of number of sources and noise levels. Colors correspond to different source numbers.

the number of sources and we plot the pointwise standard deviation as we increase the noise standard deviation from top to bottom. There is a consistent increase in standard deviation values as we move from the top-left to the bottom-right corner. In other words, the posterior contract (shrinks) when we have more data (more numbers of sources, Figure 5.18 from right to left) and when we have less noise (Figure 5.18 from bottom to top), which effectively means more data.

Figure 5.19 offers an alternate method of visualizing posterior contraction, displaying box plots of the standard deviation values for each of the 12 images in Figure 5.18. In each of the three box plots, the vertical axis corresponds to the noise standard deviation, and the horizontal axis represents the possible values in the posterior pointwise standard deviation plots. The box indicates the values that are between the first and third quartiles (where half of the possible values fall) and the line in the middle indicates the median value. Figures 5.19a to 5.19c show box plots for experiments with 25, 51, and 102 sources, respectively, with each box plot color reflecting a particular noise level. In each of the Figures 5.19a to 5.19c, we observe a decrease in the range of posterior standard deviation values, including median and quantiles, as we lower the noise level from left to right. Sim-

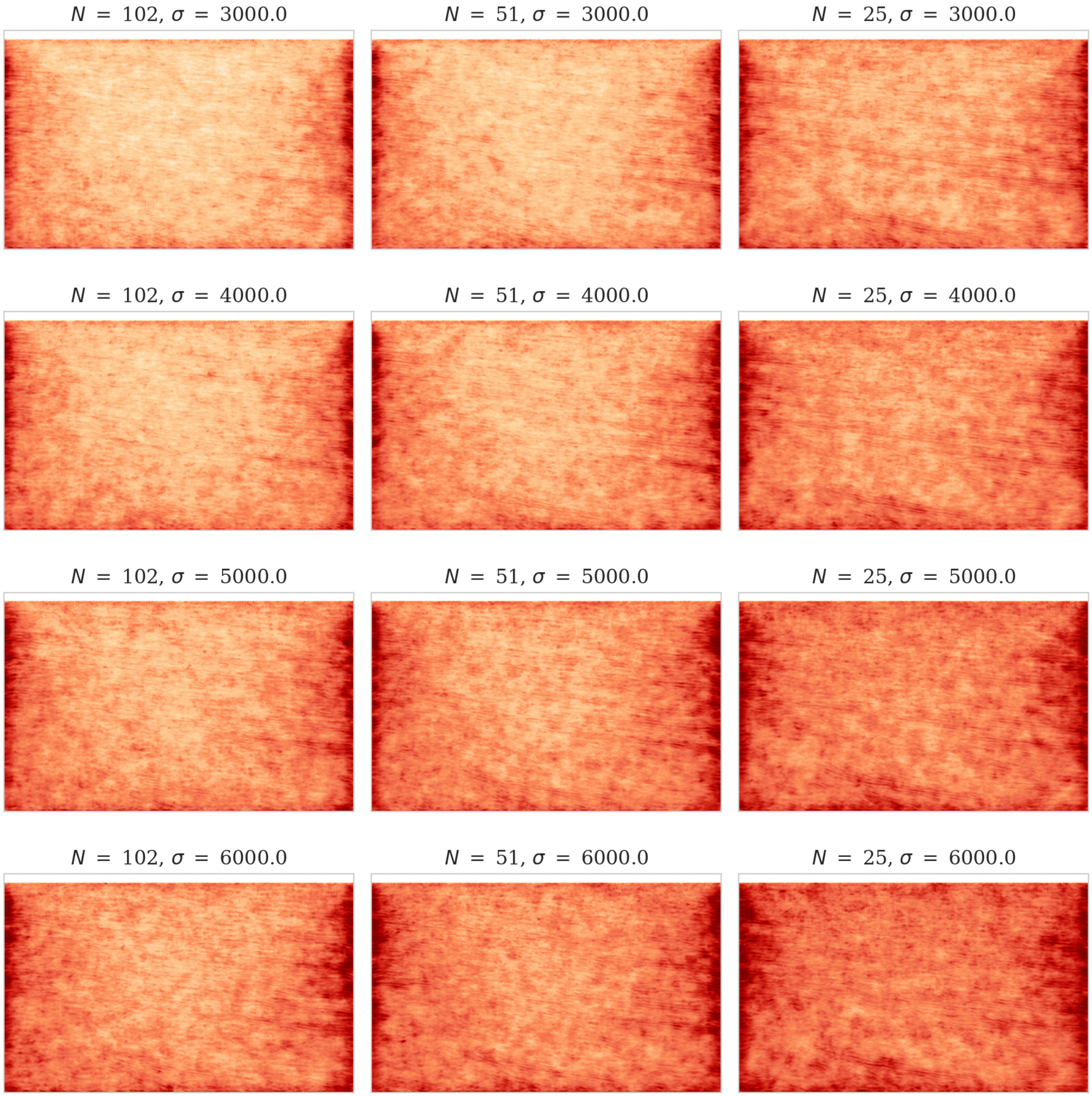
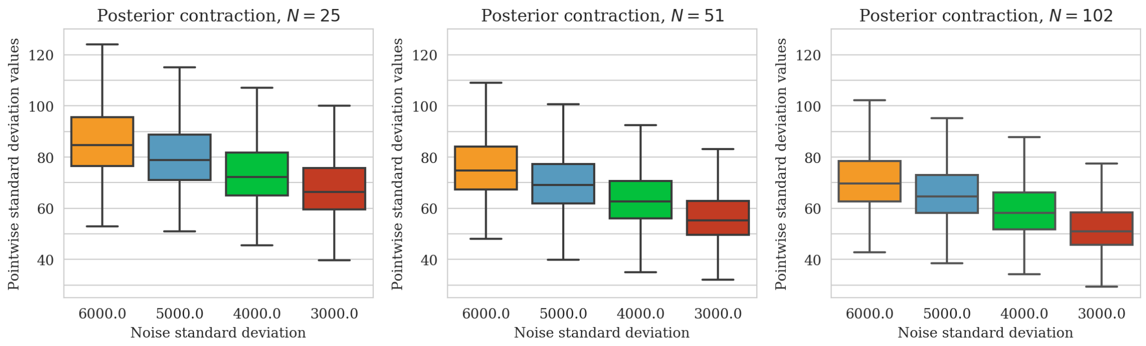


Figure 5.18: Bayesian posterior contraction: visual inspection. Pointwise standard deviations for varying number of sources (decreasing left to right) and noise variances (increases top to bottom).



(a) (b) (c)

Figure 5.19: Box plots of pointwise standard deviation values as a function of noise level for number of sources (a) $N = 25$, (b) $N = 51$, and (c) $N = 102$.

ilarly, for the same noise levels, that is, box plots of the same color, the standard deviations decrease from Figure 5.19a to Figure 5.19c (increasing number of sources). The observed trends in Figures 5.18 and 5.19 verify that our Bayesian inference method exhibits the Bayesian posterior contraction property.

5.6.2 Shift in the forward model and prior distribution

This example is intended to demonstrate how our latent distribution correction method can perform Bayesian inference when the unknown ground truth image has properties that differ from the seismic images in the pretraining dataset. This expectation is based on the invertible nature of conditional normalizing flows, which allows them to represent any image in the image space [40].

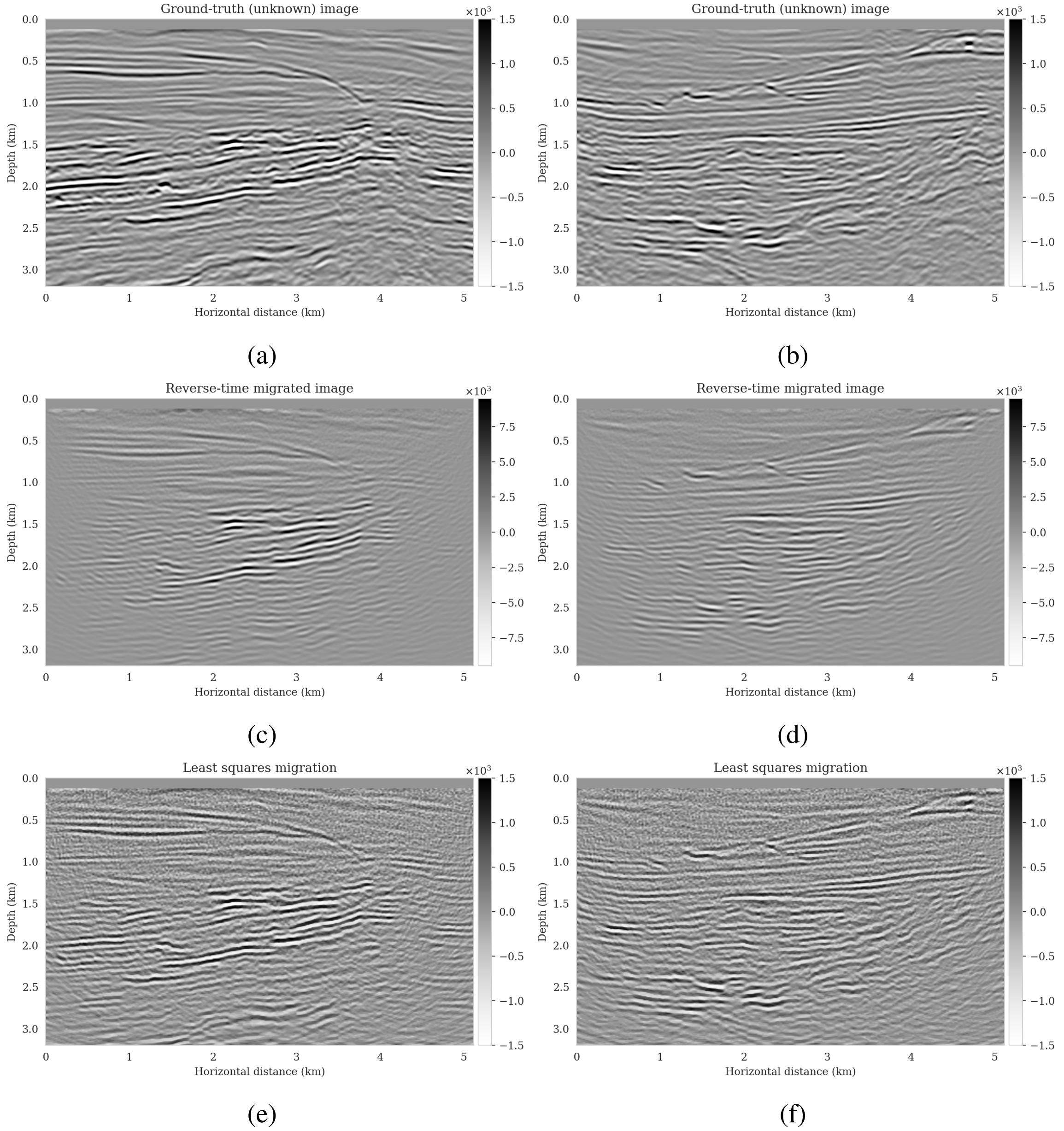
As a means to mimic this scenario, unlike images used in pretraining, we have extracted two 2D seismic images from deeper sections of the Parihaka dataset. In these sections, there are fewer continuous reflectors and more complex geological features. Figures 5.20a and 5.20b show two images with drastically different geological features when compared to Figure 5.5a. Following the pretraining data acquisition setup, we add a water column on top of these images to reduce the near-source artifacts. Similar to the previous experiment involving forward model shifts, we use only 25 sources with 200 meters sampling distance and add noise with a 2.5 fold larger standard deviation than during the pretraining phase. With this setup, the noisy observed data associated with experiments involving seismic images in Figures 5.20a and 5.20b have a SNR of $-1.56\,\mathrm{dB}$ and $-2.41\,\mathrm{dB}$, respectively.

As part of our analysis, we compute the reverse-time (second row in Figure 5.10) and least-squares (third row in Figure 5.10) migrated images for these two ground-truth images, where the former images serve as inputs to the pretrained conditional normalizing flows. Similar to the previous example, the reverse-time migrated images contain the near-source artifacts and are contaminated by the input noise. Moreover, the least-squares migrated images highlight the importance of including prior information in this imaging problem, since this image contains strong noise-related artifacts, which might impact downstream tasks, such as horizon tracking.

Conditional mean and pointwise standard deviations

To obtain samples from the posterior distribution, we feed the reverse-time migrated images (Figures 5.20c and 5.20d) into the pretrained conditional normalizing flow. The posterior is sampled using either standard Gaussian distributions or corrected latent samples. It is apparent, once more, that the uncorrected conditional mean estimates are significantly contaminated by artifacts in the near-source region (Figures 5.21a and 5.21b). Another type of noticeable error in these predicted images includes lower amplitudes in deeper and closer to boundary reflectors. A comparison of the conditional means estimates before (Figures 5.21a and 5.21b) and after (Figures 5.21c and 5.21d) latent distribution correction indicates attenuation of near-source artifacts as well as an improvement in reflector illumination near the boundary and deeper sections where images are more difficult to capture. Following latent distribution correction, the conditional mean estimate SNR is improved by three to four decibels for both images.

Careful inspection at the boundaries in the corrected conditional mean estimates reveals some nonrealistic events near the boundaries. The plots of pointwise standard deviations



Ground-truth (unknown) image

Ground-truth (unknown) image

Reverse-time migrated image

Reverse-time migrated image

Least squares migration

Least squares migration

(a)

(b)

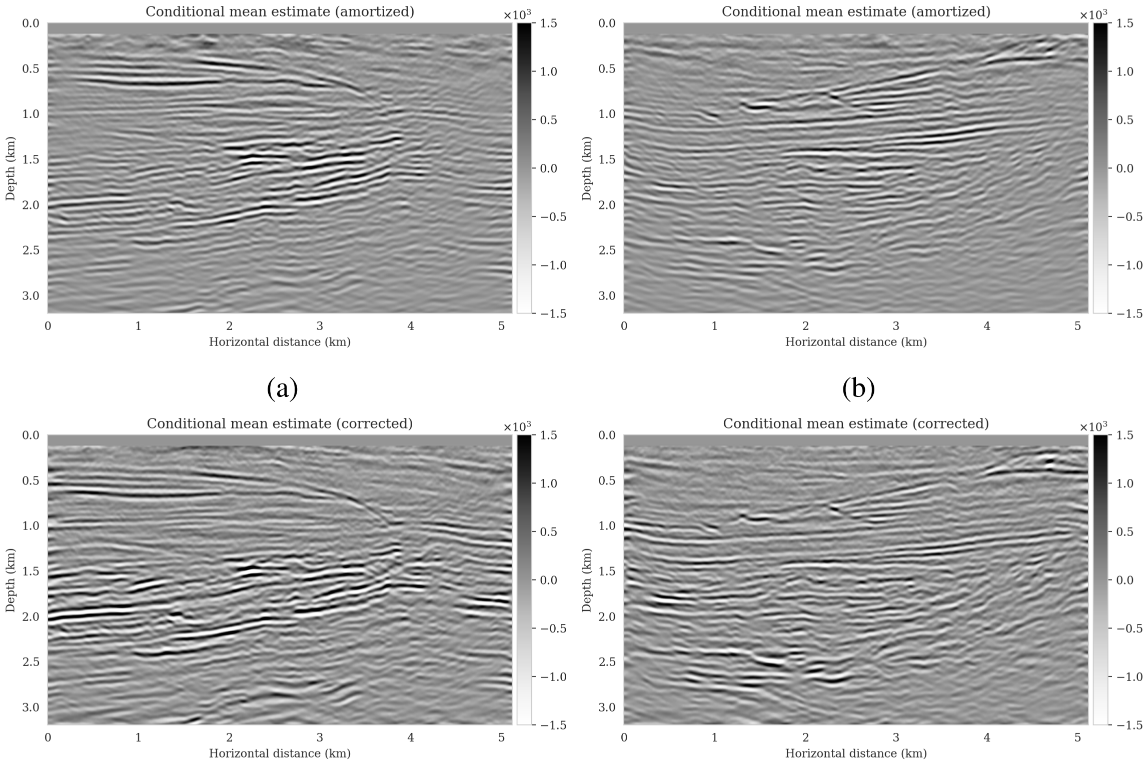
(c)

(d)

(e)

(f)

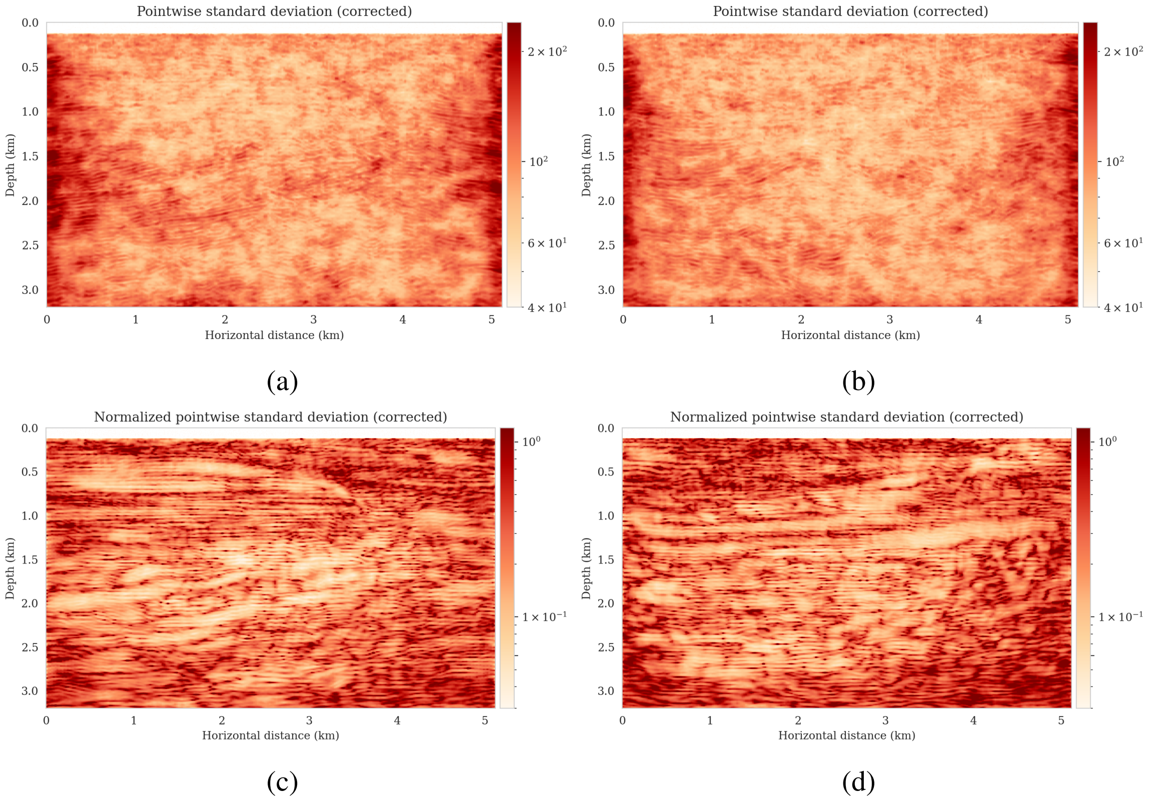
Figure 5.20: Setup for experiments involving shifts in the prior distribution. (a) and (b) Two high-fidelity ground-truth images from deeper sections of the Parihaka dataset. (c) and (d) Reverse-time migrated image corresponding ground-truth images in Figures 5.20a (SNR $-8.02\,\mathrm{dB}$) and 5.20b (SNR $-8.49\,\mathrm{dB}$), respectively. (d) Least squares imaging results (no regularization) corresponding ground-truth images in Figures 5.20a (SNR $4.94\,\mathrm{dB}$) and 5.20b (SNR $5.59\,\mathrm{dB}$), respectively.



(a) (b)

(c) (d)

Figure 5.21: The improvement in conditional mean estimate due to latent distribution correction. (a) and (b) Amortized uncorrected conditional (posterior) mean estimates with SNRs 5.47\,\mathrm{dB} and $6.17\,\mathrm{dB}$, respectively. (c) and (d) Conditional (posterior) mean estimates after latent distribution correction with SNRs $9.40\,\mathrm{dB}$ and $9.11\,\mathrm{dB}$, respectively.



(a)

(b)

(c)

(d)

Figure 5.22: Uncertainty quantification with latent distribution correction. (a) and (b) The pointwise standard deviation estimates among samples drawn from the posterior after latent distribution correction. (b) Pointwise standard deviations normalized by the envelop of conditional mean estimates.

(Figures 5.22a and 5.22b) associated with corrected conditional means, however, clearly indicate that there is uncertainty for these events. This illustrates the importance of uncertainty quantification and not relying on a single estimate when addressing ill-posed inverse problems. To diminish the imprint of strong reflectors in the pointwise standard deviations plots, we also display these images when normalized with respect to the envelopes of the conditional mean estimates in Figures 5.22c and 5.22d.

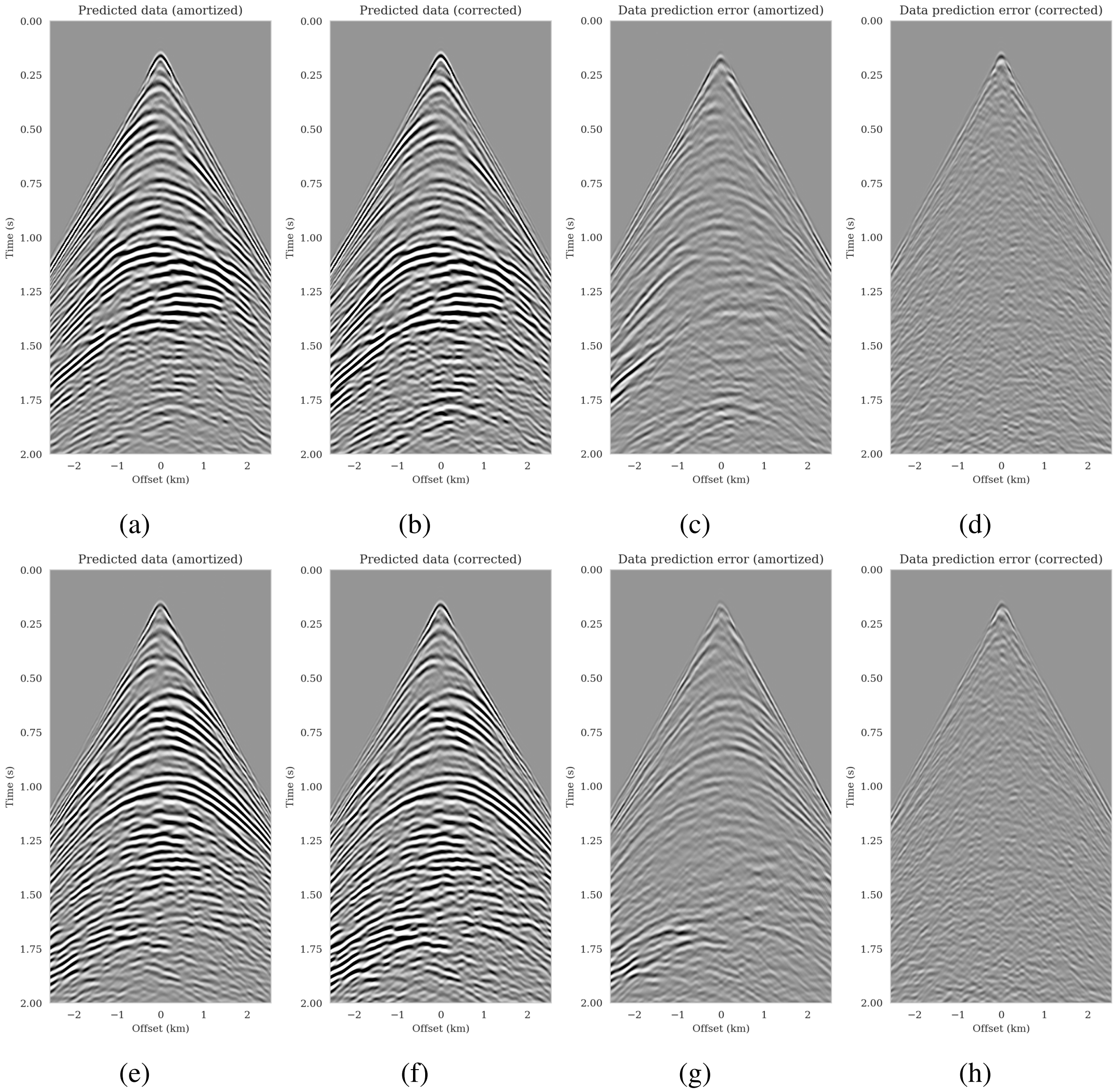
Data residuals

As before, we confirm that the latent distribution correction improves the fit of the data. Figure 5.23 shows the predicted data as well as the data residuals for all conditional mean estimates. The predicted data are obtained by applying the forward operator to the conditional mean estimates, both before and after latent distribution correction. The predicted data before and after correction are presented in the first and second columns, respectively. The corresponding data residuals before and after correction, which are computed by subtracting the predicted data from ideal noise-free data, can be seen in the third and last columns of Figure 5.23. Evidently, the latent distribution correction stage has resulted in a better fit with the observed data, as coherent data events show up in Figures 5.23c and 5.23g, but are attenuated in the corrected residual plots (Figures 5.23d and 5.23h).

5.7 Discussion

The examples presented demonstrate that deep neural networks trained in the context of amortized variational inference can facilitate solving inverse problems in two ways: (1) incorporating prior knowledge gained through pretraining; and (2) accelerating Bayesian inference and uncertainty quantification. Despite the fact that amortized variational inference is capable of sampling the posterior distribution without requiring forward operator evaluations during inference, the extent to which it is considered reliable is dependent upon the availability of high quality training data. We demonstrate this limitation of amortized variational inference via a seismic imaging example where we alter the number of sources and variance of noise in comparison to the pretraining phase. The obtained posterior samples via amortized variational inference had unusual near-source artifacts, which we can be attributed to lack of these artifacts in the training reverse-time migrated images.

As part of our efforts to extend the application of these models to domains with limited access to high-quality training data, we developed a physics-based variational inference formulation over the latent space of a pretrained conditional normalizing flow that mitigates some of the posterior sampling errors induced by data distribution shifts. In this approach, a diagonal correction to the latent distribution is learned that ensures that the conditional normalizing flow output distribution better matches the desired posterior distribution. Based on observations and other research [40, 34], we found that normalizing flows, because of



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

(e) (f) (g) (h)

Figure 5.23: Quality control in data space. The first and second row correspond to experiments involving Figures 5.20a and 5.20b, respectively. (a) and (e) Predicted data before correction with SNRs $9.13\,\mathrm{dB}$ and $9.02\,\mathrm{dB}$, respectively. (b) and (f) Predicted data after latent distribution correction with SNRs $15.27\,\mathrm{dB}$ and $14.52\,\mathrm{dB}$, respectively. (c) and (g) Data residuals (before correction) associated with Figures 5.23a and 5.23e, respectively. (d) and (h) Data residuals (after correction) associated with Figures 5.23b and 5.23f, respectively.

their invertibility, are capable of partially mitigating errors related to changes in data distributions when they are used to solve inverse problems. We leave the delineation of which data distribution shifts can be handled by our diagonal correction approach to future work. Needless to say, by adhering to more complex transformations in the latent space, e.g., via a neural network, a wide range of data distribution shifts can be handled but it would potentially require a more computationally costly correction step.

Latent distribution correction requires several passes over the data (five in our example), which includes evaluation of the forward operator and the adjoint of its Jacobian. Due to the amortized nature of our approach, these costs are significantly lower than those incurred by other non-amortized variational inference methods [19, 44, 45, 46, 47]. By amortizing over the data, the pretrained conditional normalizing flow provides posterior samples for previously unseen data (drawn from the same distribution as training data) without the need for latent distribution correction. In presence of moderate data distribution shifts, the learned posterior is adjusted to new observed data via the latent distribution correction step, which could be considered an instance of transfer learning [77]. In contrast, existing methods that perform variational inference in the latent space [50, 48] require a more substantial modification to the latent distribution, as the pretrained model in these methods originally provides samples from the prior distribution. For large-scale inverse problems, such as seismic imaging, where solving a partial differential equation is required to evaluate the forward operator, reduced computational costs of Bayesian inference are particularly important. Our approach reduces the computational costs associated with Bayesian inference in such problems while also complementing the inversion with a learned prior. In order to quantify the extent to which a diagonal correction to the latent distribution mitigates errors resulting from data distribution shifts, more research is required.

As far as seismic imaging is concerned, uncertainty can be attributed to two main sources [78, 79, 80]: (1) data errors, including measurement noise; (2) modeling errors, including linearization errors, which diminish with the accuracy of the background velocity model. The scope of our research is focused on the first source of uncertainty, and we will explore how variational inference models can be used to capture errors in background models in future work. In contrast to the problem highlighted in this chapter, capturing the uncertainty caused by errors in the background model would require generating training data involving imaging experiments for a variety of plausible background velocity models, which would be computationally intensive. Recent developments in Fourier neural operators [81, 82] may prove to be useful in addressing this problem.

5.8 Conclusions

In high-dimensional inverse problems with computationally expensive forward modeling operators, Bayesian inference is challenging due to the cost of sampling the high-dimensional posterior distribution. The computational costs associated with the forward operator often limit the applicability of sampling the posterior distribution with traditional Markov chain Monte Carlo and variational inference methods. Added to the computational challenges are the difficulties associated with selecting a prior distribution that encodes our prior knowledge while not negatively biasing the outcome of Bayesian inference. Amortized variational inference addresses these challenges by incurring an offline initial training cost for a deep neural network that can approximate the posterior distribution for previously unseen observed data, distributed according to the training data distribution. When high-quality training data is readily available, and as long as there are no shifts in the data distribution, the pretrained network is capable of providing samples from the posterior distribution for previously unknown data virtually free of additional costs.

Unfortunately, in certain domains, such as geophysical inverse problems, where the structure of the Earth's subsurface is unknown, it can be challenging to obtain a training dataset, e.g., a collection of images of the subsurface, which statistically captures the strong heterogeneity exhibited by the Earth's subsurface. Furthermore, changes to the data generation process could negatively influence the quality of Bayesian inferences with amortized variational inferences due to generalization errors associated with neural networks. To address these challenges while exploiting the computational benefits of amortized variational inference, we proposed a data-specific, physics-based, and computationally cheap correction to the latent distribution of a conditional normalizing flow, pretrained to via an amortized variational inference objective. This correction involves solving a variational inference problem in the latent space of the pretrained conditional normalizing flow where we obtain a diagonal correction to the latent distribution such that the predicted posterior distribution more closely matches the desired posterior distribution.

Using a seismic imaging example, we demonstrate that the proposed latent distribution correction, at a cost of five reverse-time migrations, can be used to mitigate the effects of data distribution shifts, which includes changes in the forward model as well as the prior distribution. Our evaluation indicated improvements in seismic image quality, comparable to least squares imaging, after the latent distribution correction step, as well as estimate on on the uncertainty of the image. We presented the pointwise standard deviation as a measure of uncertainty in the image, which indicated an increase in variability in complex geological areas and poorly illuminated areas. This approach will enable computationally feasible uncertainty quantification in large-scale inverse problems, which otherwise would be computationally expensive to achieve.

5.9 Related material

The latent distribution correction optimization problem (equation 5.19) involves computing gradients of the composition of the forward operator and the pretrained conditional normalizing flow with respect to the latent variable. Computing this gradient requires actions of the forward operator and the adjoint of its Jacobian. In our numerical experiments, these operations involved solving wave equations. For maximal numerical performance, we use JUDI [83] to construct wave-equation solvers, which utilizes the just-in-time Devito [84, 85] compiler for the wave-equation based simulations. The invertible network architectures are implemented using InvertibleNetworks.jl [86], a memory-efficient framework for training invertible nets in Julia programming language. For more details on our implementation, please refer to our code on GitHub.

5.10 References

[1] R. C. Aster, B. Borchers, and C. H. Thurber, *Parameter Estimation and Inverse Problems*. Elsevier, 2018 (pages 113, 120).

[2] A. Tarantola, *Inverse problem theory and methods for model parameter estimation*. SIAM, 2005, ISBN: 978-0-89871-572-9 (pages 113, 120, 121, 130, 159).

[3] C. Robert and G. Casella, *Monte Carlo statistical methods*. Springer-Verlag, 2004 (pages 114, 121).

[4] 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. eprint: http://epubs.siam.org/doi/pdf/10.1137/110845598 (pages 114, 121).

[5] Z. Fang, C. D. Silva, R. Kuske, and F. J. Herrmann, "Uncertainty quantification for inverse problems with weak partial-differential-equation constraints," *GEOPHYSICS*, vol. 83, no. 6, R629–R647, 2018 (pages 114, 121).

[6] A. Siahkoohi, G. Rizzuti, and F. J. Herrmann, "Deep Bayesian inference for seismic imaging with tasks," *Geophysics*, vol. 87, no. 5, Jun. 2022 (pages 114, 122).

[7] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, *Bayesian Data Analysis*. CRC press, 2013 (pages 114, 122).

[8] A. Curtis and A. Lomax, "Prior information, sampling distributions, and the curse of dimensionality," *Geophysics*, vol. 66, no. 2, pp. 372–378, 2001 (page 114).

[9] M. Welling and Y. W. Teh, "Bayesian Learning via Stochastic Gradient Langevin Dynamics," in *Proceedings of the 28th International Conference on Machine Learning*, Bellevue, Washington, USA: Omnipress, 2011, pp. 681–688, ISBN: 9781450306195 (pages 114, 122).

[10] F. J. Herrmann, A. Siahkoohi, and G. Rizzuti, "Learned imaging with constraints and uncertainty quantification," in *Neural Information Processing Systems (NeurIPS) 2019 Deep Inverse Workshop*, Dec. 2019 (page 114).

[11] Z. Zhao and M. K. Sen, "A gradient based MCMC method for FWI and uncertainty analysis," in *89th Annual International Meeting, SEG*, Expanded Abstracts, 2019, pp. 1465–1469 (pages 114, 121).

[12] M. Kotsi, A. Malcolm, and G. Ely, “Uncertainty quantification in time-lapse seismic imaging: a full-waveform approach,” *Geophysical Journal International*, vol. 222, no. 2, pp. 1245–1263, May 2020 (pages 114, 121).

[13] A. Siahkoohi, G. Rizzuti, and F. J. Herrmann, “A deep-learning based bayesian approach to seismic imaging and uncertainty quantification,” in *82nd EAGE Conference and Exhibition*, Extended Abstracts, 2020 (page 114).

[14] A. Siahkoohi, G. Rizzuti, and F. J. Herrmann, “Uncertainty quantification in imaging and automatic horizon tracking—a Bayesian deep-prior based approach,” in *90th Annual International Meeting, SEG*, Expanded Abstracts, Sep. 2020, pp. 1636–1640 (page 114).

[15] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul, “An Introduction to Variational Methods for Graphical Models,” *Machine Learning*, vol. 37, no. 2, pp. 183–233, 1999 (pages 114, 120–122).

[16] M. J. Wainwright and M. I. Jordan, *Graphical models, exponential families, and variational inference*. Now Publishers Inc, 2008 (page 114).

[17] D. Rezende and S. Mohamed, “Variational inference with normalizing flows,” ser. Proceedings of Machine Learning Research, vol. 37, PMLR, Jul. 2015, pp. 1530–1538 (pages 114, 124).

[18] Q. Liu and D. Wang, “Stein Variational Gradient Descent: A General Purpose Bayesian Inference Algorithm,” in *Advances in Neural Information Processing Systems*, vol. 29, Curran Associates, Inc., 2016, pp. 2378–2386 (page 114).

[19] G. Rizzuti, A. Siahkoohi, P. A. Witte, and F. J. Herrmann, “Parameterizing uncertainty by deep invertible networks, an application to reservoir characterization,” in *90th Annual International Meeting, SEG*, Sep. 2020, pp. 1541–1545 (pages 114, 116, 123, 169).

[20] X. Zhang and A. Curtis, “Seismic tomography using variational inference methods,” *Journal of Geophysical Research: Solid Earth*, vol. 125, no. 4, e2019JB018589, 2020 (page 114).

[21] M. Tölle, M.-H. Laves, and A. Schlaefer, “A Mean-Field Variational Inference Approach to Deep Image Prior for Inverse Problems in Medical Imaging,” in *Medical Imaging with Deep Learning*, 2021 (page 114).

[22] D. Li, H. Denli, C. MacDonald, K. Basler-Reeder, A. Baumstein, and J. Daves, “Multiparameter geophysical reservoir characterization augmented by generative networks,” in *First International Meeting for Applied Geoscience & Energy*, Society of Exploration Geophysicists, 2021, pp. 1364–1368 (page 114).

[23] D. Li and H. Denli, "Traversing within the gaussian typical set: Differentiable gaussianization layers for inverse problems augmented by normalizing flows," *arXiv preprint arXiv:2112.03860*, 2021 (pages 114, 118).

[24] 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 (pages 114, 121, 122, 124).

[25] X. Zhang, M. A. Nawaz, X. Zhao, and A. Curtis, "An introduction to variational inference in geophysical inverse problems," in *Inversion of Geophysical Data*, Elsevier, 2021, pp. 73–140 (pages 114, 121, 122, 124).

[26] Y. Kim, S. Wiseman, A. Miller, D. Sontag, and A. Rush, "Semi-amortized variational autoencoders," in *Proceedings of the 35th International Conference on Machine Learning*, J. Dy and A. Krause, Eds., ser. Proceedings of Machine Learning Research, vol. 80, PMLR, Oct. 2018, pp. 2678–2687 (page 114).

[27] R. Baptista, O. Zahm, and Y. Marzouk, "An adaptive transport framework for joint and conditional density estimation," *arXiv preprint arXiv:2009.10303*, 2020 (page 114).

[28] 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 (pages 114, 132).

[29] A. Siahkoohi, G. Rizzuti, M. Louboutin, P. Witte, and F. J. Herrmann, "Preconditioned training of normalizing flows for variational inference in inverse problems," in *3rd Symposium on Advances in Approximate Bayesian Inference*, Jan. 2021 (pages 114–117, 126).

[30] A. Siahkoohi and F. J. Herrmann, "Learning by example: fast reliability-aware seismic imaging with normalizing flows," in *First International Meeting for Applied Geoscience & Energy*, Expanded Abstracts, 2021, pp. 1580–1585 (pages 114, 117, 155).

[31] J. Kruse, G. Detommaso, R. Scheichl, and U. Köthe, "HINT: Hierarchical Invertible Neural Transport for Density Estimation and Bayesian Inference," *Proceedings of AAAI-2021*, 2021 (pages 114, 115, 117, 126–128, 155).

[32] N. Kovachki, R. Baptista, B. Hosseini, and Y. Marzouk, *Conditional Sampling With Monotone GANs*, 2021. arXiv: 2006.06755 `[stat.ML]` (pages 114, 117).

[33] A. Khorashadizadeh, K. Kothari, L. Salsi, A. A. Harandi, M. de Hoop, and I. Dokmanić, "Conditional Injective Flows for Bayesian Imaging," *arXiv preprint arXiv:2204.07664*, 2022 (pages 114, 117).

[34] 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 (pages 114, 118, 141, 144, 167).

[35] A. Taghvaei and B. Hosseini, "An optimal transport formulation of bayes' law for nonlinear filtering algorithms," *arXiv preprint arXiv:2203.11869*, 2022 (page 114).

[36] J. Sun, K. A. Innanen, and C. Huang, "Physics-guided deep learning for seismic inversion with hybrid training and uncertainty analysis," *Geophysics*, vol. 86, no. 3, R303–R317, 2021 (pages 115, 126).

[37] P. Jin, X. Zhang, Y. Chen, S. X. Huang, Z. Liu, and Y. Lin, "Unsupervised learning of full-waveform inversion: Connecting CNN and partial differential equation in a loop," in *International Conference on Learning Representations*, 2022 (pages 115, 126).

[38] M. Schmitt, P.-C. Bürkner, U. Köthe, and S. T. Radev, "Bayesflow can reliably detect model misspecification and posterior errors in amortized bayesian inference," *arXiv preprint arXiv:2112.08866*, 2021 (pages 115, 126, 132).

[39] L. Dinh, J. Sohl-Dickstein, and S. Bengio, *Density estimation using Real NVP*, 2016 (pages 115, 124, 127, 128).

[40] M. Asim, M. Daniels, O. Leong, A. Ahmed, and P. Hand, "Invertible generative models for inverse problems: Mitigating representation error and dataset bias," in *Proceedings of the 37th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 119, PMLR, Jul. 2020, pp. 399–409 (pages 115–118, 141, 143, 144, 162, 167).

[41] M. D. Parno and Y. M. Marzouk, "Transport Map Accelerated Markov Chain Monte Carlo," *SIAM/ASA Journal on Uncertainty Quantification*, vol. 6, no. 2, pp. 645–682, 2018 (page 115).

[42] B. Peherstorfer and Y. Marzouk, "A transport-based multifidelity preconditioner for Markov chain Monte Carlo," *Advances in Computational Mathematics*, vol. 45, no. 5-6, pp. 2321–2348, 2019 (page 115).

[43] A. Bora, A. Jalal, E. Price, and A. G. Dimakis, "Compressed sensing using generative models," in *Proceedings of the 34th International Conference on Machine Learning*, D. Precup and Y. W. Teh, Eds., ser. Proceedings of Machine Learning Research, vol. 70, PMLR, Jun. 2017, pp. 537–546 (page 116).

[44] A. Andrle, N. Farchmin, P. Hagemann, S. Heidenreich, V. Soltwisch, and G. Steidl, "Invertible Neural Networks Versus MCMC for Posterior Reconstruction in Graz-

ing Incidence X-Ray Fluorescence," in *SSVM*, Springer, 2021, pp. 528–539 (pages 116, 123, 169).

[45] X. Zhao, A. Curtis, and X. Zhang, "Bayesian seismic tomography using normalizing flows," *Geophysical Journal International*, vol. 228, no. 1, pp. 213–239, Jul. 2021. eprint: https://academic.oup.com/gji/article-pdf/228/1/213/40348424/ggab298.pdf (pages 116, 123, 169).

[46] X. Zhang and A. Curtis, "Bayesian geophysical inversion using invertible neural networks," *Journal of Geophysical Research: Solid Earth*, vol. 126, no. 7, e2021JB022320, 2021 (pages 116, 123, 169).

[47] X. Zhao, A. Curtis, and X. Zhang, "Interrogating subsurface structures using probabilistic tomography: An example assessing the volume of irish sea basins," *Journal of Geophysical Research: Solid Earth*, vol. 127, no. 4, e2022JB024098, 2022 (pages 116, 123, 169).

[48] K. Kothari, A. Khorashadizadeh, M. de Hoop, and I. Dokmanić, "Trumpets: Injective Flows for Inference and Inverse Problems," *arXiv preprint arXiv:2102.10461*, 2021 (pages 117, 118, 169).

[49] J. Adler and O. öktem, "Deep Bayesian Inversion," *arXiv preprint arXiv:1811.05910*, 2018 (page 117).

[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 (pages 118, 169).

[51] A. Malinverno and V. A. Briggs, "Expanded uncertainty quantification in inverse problems: Hierarchical bayes and empirical bayes," *GEOPHYSICS*, vol. 69, no. 4, pp. 1005–1016, 2004 (page 121).

[52] A. Malinverno and R. L. Parker, "Two ways to quantify uncertainty in geophysical inverse problems," *GEOPHYSICS*, vol. 71, no. 3, W15–W27, 2006 (page 121).

[53] A. Ray, S. Kaplan, J. Washbourne, and U. Albertin, "Low frequency full waveform seismic inversion within a tree based Bayesian framework," *Geophysical Journal International*, vol. 212, no. 1, pp. 522–542, Oct. 2017. eprint: https://academic.oup.com/gji/article-pdf/212/1/522/21782947/ggx428.pdf (page 121).

[54] G. K. Stuart, S. E. Minkoff, and F. Pereira, "A two-stage Markov chain Monte Carlo method for seismic inversion and uncertainty quantification," *GEOPHYSICS*, vol. 84, no. 6, R1003–R1020, Nov. 2019 (page 121).

[55] C. Li, C. Chen, D. Carlson, and L. Carin, “Preconditioned Stochastic Gradient Langevin Dynamics for Deep Neural Networks,” in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, Phoenix, Arizona: AAAI Press, 2016, pp. 1788–1794 (page 122).

[56] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer-Verlag New York, 2006 (page 122).

[57] H. Robbins and S. Monro, “A stochastic approximation method,” *The annals of mathematical statistics*, pp. 400–407, 1951 (page 123).

[58] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro, “Robust stochastic approximation approach to stochastic programming,” *SIAM Journal on optimization*, vol. 19, no. 4, pp. 1574–1609, 2009 (page 123).

[59] T. Tieleman and G. Hinton, *Lecture 6.5-RMSprop: Divide the gradient by a running average of its recent magnitude*, https://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf, 2012 (page 123).

[60] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014 (pages 123, 132, 146).

[61] C. Villani, *Optimal Transport: Old and New*. Springer-Verlag Berlin Heidelberg, 2009, vol. 338 (pages 124, 127).

[62] D. P. Kingma and M. Welling, “Auto-Encoding Variational Bayes,” in *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014 (page 124).

[63] 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 (page 126).

[64] A. Lavin *et al.*, “Simulation intelligence: Towards a new generation of scientific methods,” *arXiv preprint arXiv:2112.03235*, 2021 (page 126).

[65] T. Teshima, I. Ishikawa, K. Tojo, K. Oono, M. Ikeda, and M. Sugiyama, “Coupling-based invertible neural networks are universal diffeomorphism approximators,” in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33, Curran Associates, Inc., 2020, pp. 3362–3373 (page 127).

[66] I. Ishikawa, T. Teshima, K. Tojo, K. Oono, M. Ikeda, and M. Sugiyama, “Universal approximation property of invertible neural networks,” *arXiv preprint arXiv:2204.07415*, 2022 (page 127).

[67] J. E. Gubernatis, E. Domany, J. A. Krumhansl, and M. Huberman, “The Born approximation in the theory of the scattering of elastic waves by flaws,” *Journal of Applied Physics*, vol. 48, no. 7, pp. 2812–2819, 1977 (page 130).

[68] G. Lambaré, J. Virieux, R. Madariaga, and S. Jin, “Iterative asymptotic inversion in the acoustic approximation,” *Geophysics*, vol. 57, no. 9, pp. 1138–1154, 1992 (page 130).

[69] G. T. Schuster, “Least-squares cross-well migration,” in *63rd Annual International Meeting, SEG*, Expanded Abstracts, 1993, pp. 110–113 (page 130).

[70] T. Nemeth, C. Wu, and G. T. Schuster, “Least-squares migration of incomplete reflection data,” *GEOPHYSICS*, vol. 64, no. 1, pp. 208–221, 1999 (page 130).

[71] Veritas, “Parihaka 3D Marine Seismic Survey - Acquisition and Processing Report,” New Zealand Petroleum & Minerals, Wellington, Tech. Rep. New Zealand Petroleum Report 3460, 2005 (page 131).

[72] WesternGeco., “Parihaka 3D PSTM Final Processing Report,” New Zealand Petroleum & Minerals, Wellington, Tech. Rep. New Zealand Petroleum Report 4582, 2012 (page 131).

[73] 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, May 2022 (page 132).

[74] R. Liu, A. Chakrabarti, T. Samanta, J. K. Ghosh, and M. Ghosh, “On divergence measures leading to jeffreys and other reference priors,” *Bayesian Analysis*, vol. 9, no. 2, pp. 331–370, 2014 (page 139).

[75] Y. Yang and S. Soatto, “Conditional Prior Networks for Optical Flow,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 271–287 (page 141).

[76] C. Zeno, I. Golan, E. Hoffer, and D. Soudry, “Task Agnostic Continual Learning Using Online Variational Bayes,” *arXiv preprint arXiv:1803.10123*, 2018 (page 141).

[77] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, “How transferable are features in deep neural networks?” In *Proceedings of the 27th International Conference on Neural Information Processing Systems*, ser. NIPS’14, Montreal, Canada, 2014, pp. 3320–3328 (page 169).

[78] P. Thore, A. Shtuka, M. Lecour, T. Ait-Ettajer, and R. Cognot, “Structural uncertainties: Determination, management, and applications,” *Geophysics*, vol. 67, no. 3, pp. 840–852, 2002 (page 169).

[79] K. Osypov *et al.*, "Model-uncertainty quantification in seismic tomography: Method and applications," *Geophysical Prospecting*, vol. 61, no. 6-Challenges of Seismic Imaging and Inversion Devoted to Goldin, pp. 1114–1134, 2013 (page 169).

[80] G. Ely, A. Malcolm, and O. V. Poliannikov, "Assessing uncertainties in velocity models and images with a fast nonlinear uncertainty quantification method," *GEOPHYSICS*, vol. 83, no. 2, R63–R75, 2018. eprint: https://doi.org/10.1190/geo2017-0321.1 (page 169).

[81] Z. Li *et al.*, "Fourier Neural Operator for Parametric Partial Differential Equations," in *9th International Conference on Learning Representations*, OpenReview.net, 2021 (page 170).

[82] A. Siahkoohi, M. Louboutin, and F. J. Herrmann, "Velocity continuation with Fourier neural operators for accelerated uncertainty quantification," in *2nd International Meeting for Applied Geoscience & Energy*, 2022 (page 170).

[83] 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. eprint: https://doi.org/10.1190/geo2018-0174.1 (page 171).

[84] F. Luporini *et al.*, "Architecture and performance of devito, a system for automated stencil computation," *CoRR*, vol. abs/1807.03032, Jul. 2018. arXiv: 1807.03032 (page 171).

[85] 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 (page 171).

[86] P. Witte, G. Rizzuti, M. Louboutin, A. Siahkoohi, and F. Herrmann, *InvertibleNetworks.jl: A Julia framework for invertible neural networks*, version v1.1.0, Nov. 2020 (page 172).

CHAPTER 6
THE IMPORTANCE OF TRANSFER LEARNING IN SEISMIC MODELING AND IMAGING

6.1 Summary

Accurate forward modeling is essential for solving inverse problems in exploration seismology. Unfortunately, it is often not possible to afford being physically or numerically accurate. To overcome this conundrum, we make use of raw and processed data from nearby surveys. We propose to use this data, consisting of shot records or velocity models, to pre-train a neural network to correct for the effects of, for instance, the free surface or numerical dispersion, both of which can be considered as proxies for incomplete or inaccurate physics. Given this pre-trained neural network, we apply transfer learning to finetune this pre-trained neural network so it performs well on its task of mapping low-cost, but low-fidelity, solutions to high-fidelity solutions for the current survey. As long as we can limit ourselves during finetuning to using only a small fraction of high-fidelity data, we gain processing the current survey while using information from nearby surveys. We demonstrate this principle by removing surface-related multiples and ghosts from shot records and the effects of numerical dispersion from migrated images and wave simulations.

6.2 Introduction

In a perfect world, the wave equation should be able to inform us how to image seismic field data. Often we are met with data behaving erratically or with imaging problems that are too computationally demanding. For these reasons, lots of time and effort has been devoted to wave-equation based imaging technology. We propose to map low-cost low-fidelity solutions to high-fidelity ones whether this involves the free surface or inaccurate numerical wave simulations. Our goal is to do this by using information that is available from nearby surveys.

Several attempts have been made to incorporate machine-learning techniques into seismic processing, imaging, and inversion. Applications range from missing-trace interpolation [1, 2] and low-frequency interpolation [3] to the use of a pre-trained neural network as a prior for full-waveform inversion [4]. [5] demonstrate the possible application of deep learning in seismic data processing. We extend ideas in [5] by utilizing transfer learning to achieve more accurate results by utilizing nearby surveys.

Compared to other imaging modalities, such as medical imaging, we generally do not have access to the ground truth whether this concerns ideal field data or detailed information on the subsurface. This lack of information forces us to fundamentally rethink how to apply machine learning. We combine established data-processing flows with transfer learning [6]. This allows us to map computationally cheap low-fidelity solutions to high-fidelity ones with a pre-trained Convolutional Neural Network (CNN). We train this CNN with data from a survey that is close so that it shares main geological features. Because of its proximity, we assume that this nearby survey and the current survey have similar statistical properties and thus we can use the nearby survey to pre-train our network. Since this assumption does not hold completely, we use transfer learning to finetune this pre-trained network with a small percentage of additional training data obtained from the current survey. This allows us to replace computationally expensive processing steps with cheaper learned ones, when processing the current survey.

This chapter is organized as follows. First, we describe three approaches for data, gradient, and simulation conditioning designed to deal with cheap low-fidelity simulations. Next, we explain our transfer learning technique, followed by an introduction to Generative Adversarial Networks [GANs, 7], which we use to map low-fidelity solutions to high-fidelity ones. After describing the used CNN architecture, we demonstrate the capabilities of our approach and the importance of transfer learning on the removal of the effects of the free surface and of a poorly discretized wave equation.

6.3 Theory

When solving inverse problems, we implicitly make assumptions on the data. For instance, during seismic imaging we often assume the data to be generated without a free surface, which is not the case in practice. We also assume our simulations as part of imaging free of numerical dispersion, an assumption that may not always be computationally affordable. If these assumptions are violated, wave-equation based inversions will suffer because our simulations will be inconsistent with the observed data. Below we discuss our machine-learning approach to data/gradient/simulation conditioning and how transfer learning and GANs factor into our approach.

6.3.1 Data/gradient/simulation conditioning

Let $\mathbf{d}_{\text{observed}}$ be the observed data that we try to fit with forward modeling $F(\mathbf{m})$ parameterized by the model parameters $\mathbf{m}$. In the ideal case, $\mathbf{d}_{\text{observed}}$ is noise free and physically accurate, so F exactly describes the observed data when $\mathbf{m} = \mathbf{m}^*$ with $\mathbf{m}^*$ the ground-truth model parameters. In that ideal situation, we have $\mathbf{d}_{\text{observed}} = F(\mathbf{m}^*)$.

As stated before, we usually do not have access to the "true" forward modeling operator, or if we do, its numerical evaluation can be too expensive. Let $\underline{F}(\mathbf{m})$ denote our low-fidelity forward operator. Given true model parameters $\mathbf{m} = \mathbf{m}^*$, this approximation does not explain the observed data, i.e., $\mathbf{d}_{\text{observed}} \neq \underline{F}(\mathbf{m}^*)$ and even in the absence of noise, inverting for $\mathbf{m}$ from observed data will not lead to $\mathbf{m}^*$.

To overcome this problem, we propose three different types of conditionings, namely data conditioning, where we take out unmodeled components from the observed data, gradient, and simulation conditioning, where we remove artifacts due to low-fidelity simulations that suffer from numerical dispersion. Because of recent success of CNNs in "image-to-image" mappings [8, 9], we propose to condition with CNNs, which we train on pairs of low- and high-fidelity simulations. Our examples with numerical dispersion serve as proxies for situations where the numerical simulation for the physics may be inadequate. However, we do not claim that our approach is superior to recent work by [10] on removing the effects of temporal dispersion.

Let $\mathcal{G}_{\theta_d}$, $\mathcal{G}_{\theta_g}$, and $\mathcal{G}_{\theta_A}$ be the CNNs used to condition the low-fidelity observed data, gradients, and wavefield simulations, parameterized by θ_d, θ_g, and θ_A, respectively. Mathematically, these conditionings take the following form:

$$\mathbf{d}_{\text{conditioned}} = \mathcal{G}_{\theta_d}\left(\mathbf{d}_{\text{observed}}\right), \tag{6.1}$$

$$\mathbf{g}_{\text{conditioned}} = \mathcal{G}_{\theta_g}\left(\underline{J}^{\top}(\delta\mathbf{d})\right), \text{ and} \tag{6.2}$$

$$\mathbf{u}_{\text{conditioned}} = \mathcal{G}_{\theta_A}\left(\underline{A}^{-1}\mathbf{q}\right) \tag{6.3}$$

where $\mathbf{d}_{\text{observed}}$ represents observed data including the effects of the free surface. The symbol $\underline{J}$ denotes the low-fidelity Jacobian (linearized Born scattering operator), which acts on $\delta\mathbf{d}$ the data residual—i.e., observed data after removal of the direct wave. The matrix $\underline{A}$ corresponds to the low-fidelity discretized wave equation, and $\mathbf{q}$ to the source. In our examples, $\mathbf{d}_{\text{conditioned}}$, $\mathbf{g}_{\text{conditioned}}$, and $\mathbf{u}_{\text{conditioned}}$ represent shot records without free-surface effects, single-shot reverse-time migrations, and wavefield snapshots after numerical-dispersion removal, respectively.

The above operations are key to iterative wave-equation based inversion where observed data are matched with simulated data. Because each operation involves the wave equation, simulations either miss important physics or are too expensive to evaluate accurately. The key idea now is to condition by training CNNs with pairs of low- and high-fidelity simulations. With these trained CNNs, we map low-fidelity data, gradients, and simulations to high-fidelity ones. We assume that these low- and high-fidelity pairs are available from nearby surveys to pre-train our network.

6.3.2 Transfer learning

Since the Earth subsurface is unknown, finding suitable training sets is often a challenge in the geosciences. We avoid this problem by working with low- and high-fidelity solutions that are both assumed available, e.g., from nearby surveys, albeit the latter often at a higher cost. Generally speaking, we gain if the costs of training and applying the CNNs to the current survey are smaller than the gains we make by avoiding expensive processing, such as the removal of surface-related multiples or conducting accurate high-fidelity simulations. Obviously, this approach can become challenging because of dissimilarities that may exist between the current and nearby surveys. Instead on relying on CNNs to generalize, which may call for massive amounts of training or may not be feasible at all, we rely on transfer learning [6] to handle the fact that the probability distributions for the nearby and current surveys may not be the same.

During transfer learning, the weights of a pre-trained neural network from the nearby surveys are finetuned to work with data from the current survey. Since transfer learning can be done with a relatively small fraction ($\approx 5\%$) of data from the pertinent survey, this can lead to an economically viable workflow since this type of data can often be made available, e.g., by applying more expensive conventional processing on a small fraction of the data of the current survey. Following [11], we avoid overfitting by only finetuning the deeper task-dependent CNN layers during transfer learning. We demonstrate the feasibility of our approach by means of a series of synthetic examples, based on numerical solutions of the wave equation. First, we discuss how we train our networks.

6.3.3 Training objective

Our main goal is to find nonlinear mappings for our three conditioning problems. For this purpose, we use GANs [7]. This generative approach is capable of training CNNs designed to conduct complex mappings [8, 9]. GANs derive their success from an adversarial training procedure with two CNNs, the generator, which trains to do the mapping, and a discriminator, which trains to distinguish between mapped low-fidelity simulations and true-high-fidelity simulations. We train by minimizing the following two stochastic expectations:

$$\begin{aligned} &\min_{\theta} \underset{\mathbf{x}\sim p_X(\mathbf{x}),\, \mathbf{y}\sim p_Y(\mathbf{y})}{\mathbb{E}} \left[\left(1 - \mathcal{D}_{\phi}\left(\mathcal{G}_{\theta}(\mathbf{x})\right)\right)^2 + \lambda \left\|\mathcal{G}_{\theta}(\mathbf{x}) - \mathbf{y}\right\|_1 \right], \\ &\min_{\phi} \underset{\mathbf{x}\sim p_X(\mathbf{x}),\, \mathbf{y}\sim p_Y(\mathbf{y})}{\mathbb{E}} \left[\left(\mathcal{D}_{\phi}\left(\mathcal{G}_{\theta}(\mathbf{x})\right)\right)^2 + \left(1 - \mathcal{D}_{\phi}\left(\mathbf{y}\right)\right)^2 \right]. \end{aligned} \tag{6.4}$$

The expectations are computed with respect to pairs $\{\mathbf{x},\, \mathbf{y}\}$ of low- and high-fidelity images drawn from the probability distributions $p_X(\mathbf{x})$ and $p_Y(\mathbf{y})$. By alternating the above minimizations, we simultaneously train the generator $\mathcal{G}_{\theta}$ to create a low-to-high fidelity map and the discriminator $\mathcal{D}_{\phi}$ to discern the low-to-high fidelity map from a true high-fidelity one. Following [9], we include an additional ℓ_1-norm misfit term weighted by λ. This term is designed to ensure that each realization of $\mathcal{G}_{\theta}(\mathbf{x})$ maps to a particular $\mathbf{y}$—i.e., $\mathbf{x} \mapsto \mathbf{y}$ rather than solely fooling the discriminator.

We minimize the above objectives with a variant of Stochastic Gradient Descent known as the Adam optimizer [12] with momentum parameter $\beta = 0.9$ and a linearly decaying step size with initial value $\mu = 2 \times 10^{-4}$ for both the generator and discriminator networks. During each iteration of the Adam optimizer, the expectations in the above expressions are approximated by evaluations of the objective value and gradient for a single randomly selected training pair. These pairs are selected without replacement.

6.3.4 CNN architecture

"Image-to-image" mappings typically call for a hierarchical neural network consisting of encoder-decoder neural networks [9, 8]. Since the low- and high-fidelity data share a lot of information, we use CNNs that contain skip connections that exploit similarities, which allows for faster convergence during training [13]. We use the exact architecture provided by [8], which includes Residual Blocks, the main building block of ResNets, introduced by [13], for the generator $\mathcal{G}_\theta$. For the discriminator $\mathcal{D}_\phi$, we utilize the "PatchGAN" architecture, exactly as it was introduced by [9]. We designed and implemented our deep architectures in TensorFlow[1]. To carry out our wave-equation simulations with finite differences, we use Devito[2] [14, 15].

6.4 Numerical experiments

Wave-equation based imaging relies on the assumption that observed data and simulations are consistent, which is often not the case. To handle this situation, we demonstrate how CNNs can be used to remove the effects of the free surface and numerical dispersion. Our data conditioning partly replaces the numerically expensive process of SRME [16] and deghosting by the action of a transfer-trained CNN that removes the imprint of the free surface. In the second and third examples, we correct migrated shot records and wave simulations by CNNs trained to correct numerical dispersion.

6.4.1 Surface-related multiple elimination and deghosting

While wave-equation based techniques for removing the free surface are applied routinely, these techniques are computationally expensive and often require dense sampling. Here, we follow a different approach where we train a CNN to carry out the joint task of surface-related multiple removal and source-receiver side deghosting.

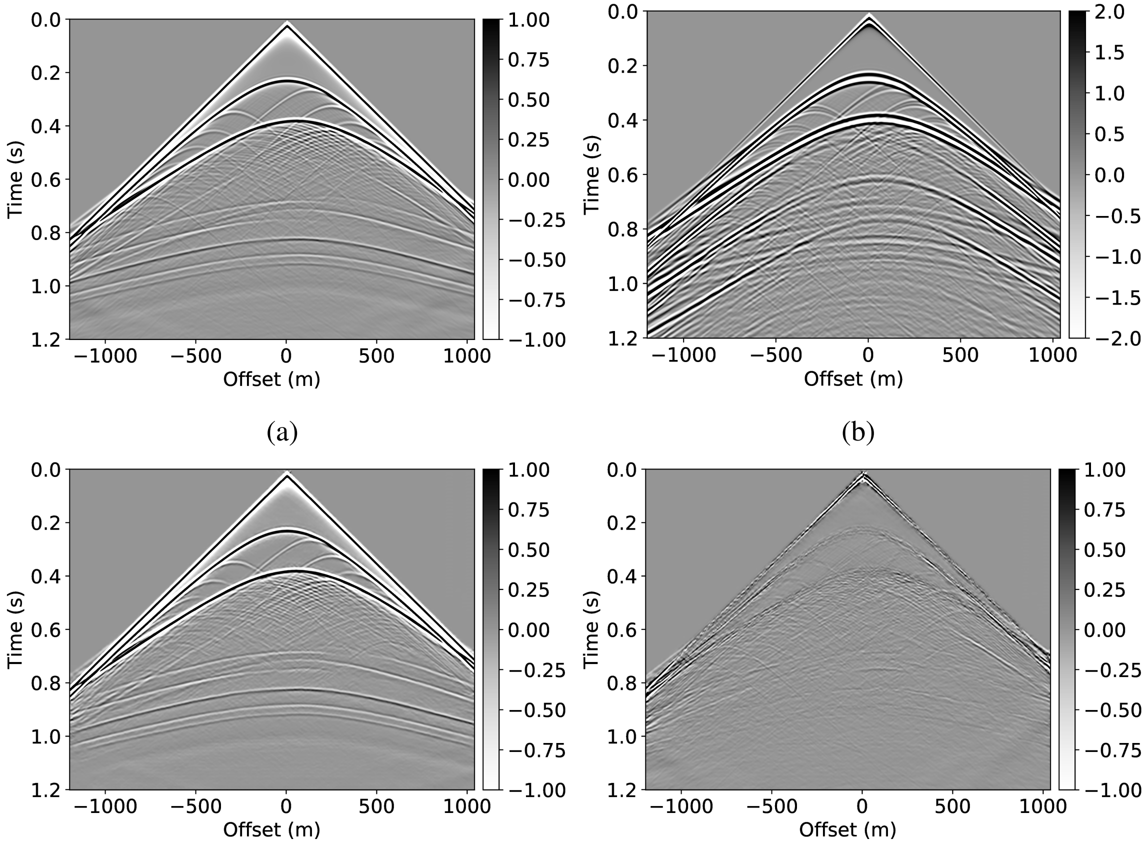
Specifically, we consider the situation where we have access to pairs of shot records from a neighboring survey before and after removal of the free-surface effects through standard processing. These pairs of unprocessed and processed data allow us to pre-train our CNN. To mimic this realistic scenario, we simulate pairs of shot records with and without a free surface for nearby velocity models that differ by 25% in water depth from the velocity model that we use for testing—i.e., we increase and decrease the water depth by $50\ \mathrm{m}$ to simulate training pairs used to pre-train our network. We make this choice for

[1] https://www.tensorflow.org/

[2] https://www.devitoproject.org/

demonstration purposes only, other more realistic choices are possible. We simulate our marine surveys with a Ricker wavelet centered at 50 Hz and with sources and receivers towed at a depth of 9 m and 14 m, respectively. Considering the ± 50 m difference in water depth between training and testing velocity models, shot records simulated on training velocity models are cycle-skipped with respect to the testing shot records. This explains the need for transfer training.

We finetune our CNN on only 5% of unprocessed/processed shot records from the current survey—i.e., we only process a fraction of the data conventionally. We numerically simulate this scenario by generating shot records in the true model (correct water depth) with and without a free surface. While this additional training round requires extra processing, we argue that we actually reduce the computational costs by as much as 95% if we ignore the time it takes to apply our neural network and the time it takes to pre-train our network on training pairs from nearby surveys. In Figure 6.1, we present our result for a shot record not seen during training. We obtained this result by applying Equation 6.1 with a CNN trained by minimizing objectives 6.4 for $\lambda = 1000$ with 100 passes through 802 shot records from the nearby surveys, followed by a round of transfer learning involving 100 passes over only 21 shot records of the current survey. We find the value for λ after extensive parameter testing. During the selection of the λ-value, we make sure that the output of the discriminator averages in the end to $\frac{1}{2}$ (by choosing λ not too large). We also make certain that the training pairs map one-to-one (by choosing λ large enough). As reported in the literature by [17], training results typically do not vary by much when varying this parameter. Our results confirm the ability of a (pre-)trained neural network to remove both surface-related multiples and ghosts (cf. Figures 6.1c and 6.1d). We carry out this training by using data from nearby surveys followed by transfer training on a small number of data pairs from the current survey. Finally, we did extensive testing to make sure we are not overfitting by monitoring the loss function over independently randomly selected shots from the unprocessed 95% of the shot records.



(a)

(b)

(c)

(d)

Figure 6.1: Joint removal of surface-related multiples and ghost. a) Modeled shot record without free surface. b) Modeled shot record with free surface. c) Removal of the free-surface effects. d) Difference between a) and c).

6.4.2 Numerical dispersion removal

To demonstrate how CNNs handle incomplete and/or inaccurate physics, we consider two examples where we use a poor discretization (only second order) for the Laplacian. Because of this choice, our low-fidelity wave simulations are numerically dispersed and we aim to remove the effects of this dispersion by properly trained CNNs. For this purpose, we first train a CNN on pairs of low- and high-fidelity single-shot reverse-time migrations from a background velocity model that is obtained from a "nearby" survey. We perform high-fidelity simulations by using a more expensive 20^{th}-order stencil. Secondly, we correct wave simulations themselves with a neural network that is trained on a family of related "nearby" velocity models. Both examples are not meant to claim possible speedups of finite-difference calculations. Instead, they are intended to demonstrate how CNNs can make non-trivial corrections such as the removal of numerical dispersion.

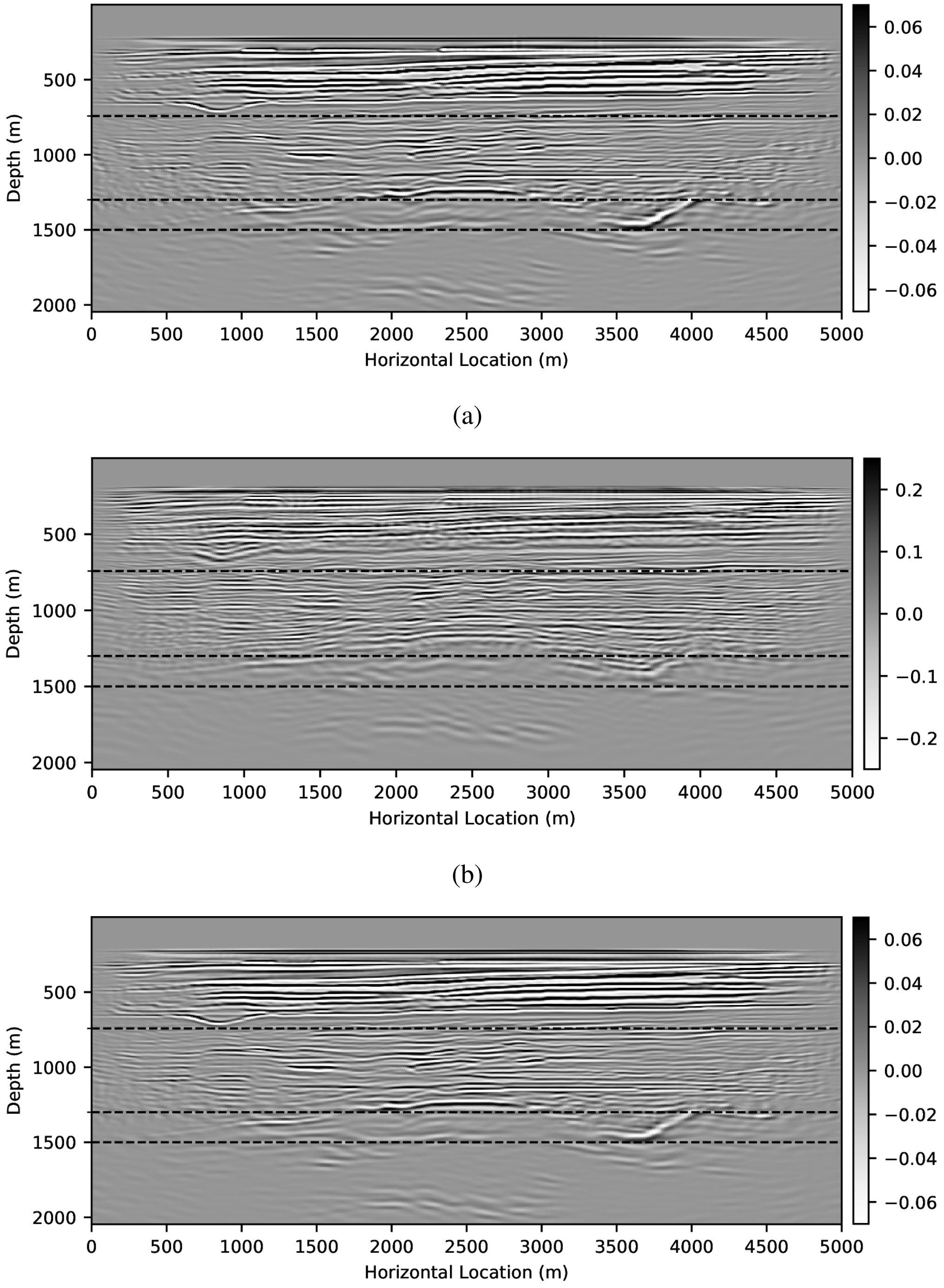
As an example of gradient conditioning, we pre-train a CNN by minimizing objectives 6.4 with $\lambda = 100$ for 100 passes over 804 pairs of low- and high-fidelity single-shot reverse-time migrations simulated for four "nearby" surveys defined by four different vertical 2D slices taken from the 3D BG Compass velocity model. As before, we find the value for λ via extensive parameter testing. Because these 2D slices are different from the current velocity model, we need to transfer train. We do this by carrying out an additional training round via 20 passes over only 11 low- and high-fidelity gradient pairs, simulated on one in every 20 shot locations, out of 201 available shot locations. This means we only perform high-fidelity migration for 5% of the shots records in the current velocity model. Given the few transfer training pairs, we prevent overfitting during transfer learning by only finetuning the deeper task-dependent CNN layers. After this two-stage training procedure, we apply corrections for each shot separately to numerically dispersed gradients computed for the current model. Figure 6.2 depicts the image we obtain by summing over 95% of the corrected gradients and over 5% of the high-fidelity gradients. As expected, there is a bias in the image quality at the high-fidelity shot locations. While the low-fidelity migrations contain large errors in positioning and amplitudes, our transfer-trained CNN corrects these errors, see for instance along the dotted lines in Figure 6.2, horizontal locations $800\ \mathrm{m}$, $2500\ \mathrm{m}$, and $3700\ \mathrm{m}$.

In the final experiment, we remove numerical dispersion from low-fidelity wave simulations. We follow [3] and crop five pieces of the Marmousi model so that their size is reduced to 10% of the original model. To make the size the same as the original model, we interpolate the selected subsets. We generate from these models low- and high-fidelity wavefield pairs for various time-steps and source positions.

We train a CNN by minimizing objectives 6.4 for $5 \times 401 \times 11 = 22055$ low- and high-fidelity wavefield snapshots simulated on these five training velocity models at 401 source locations and 11 randomly selected time snapshots. Since the cropped velocity structures are less complex than the original velocity model, an additional round of transfer learning is required. We finetune the CNN by training on 1500 (5%) low- and high-fidelity snapshot pairs. For training, we use $\lambda = 100$ in objectives 6.4 and we made 4.5 passes through the full data set (i.e., we touch all shots four times and half of them five times) followed by 11 passes during transfer learning. Figure 6.3 depicts the results of the wavefield corrections. While the low-fidelity wave simulations show a heavy imprint of numerical dispersion, e.g., strong phase and amplitude distortions, the result depicted in Figure 6.3c, demonstrates that the CNN has mostly corrected these errors. As we stated before, we do not claim to improve on recent work by [10] with this example. We only want to demonstrate that CNNs can be used for this purpose.

6.5 Conclusions

We showed that pre-trained convolutional neural networks (CNN), trained via an adversarial objective function introduced by generative adversarial networks, can conduct complex tasks including removal of the free-surface effects and mitigation of the effects of numerical dispersion during reverse-time migration and wave simulations. We were able to achieve

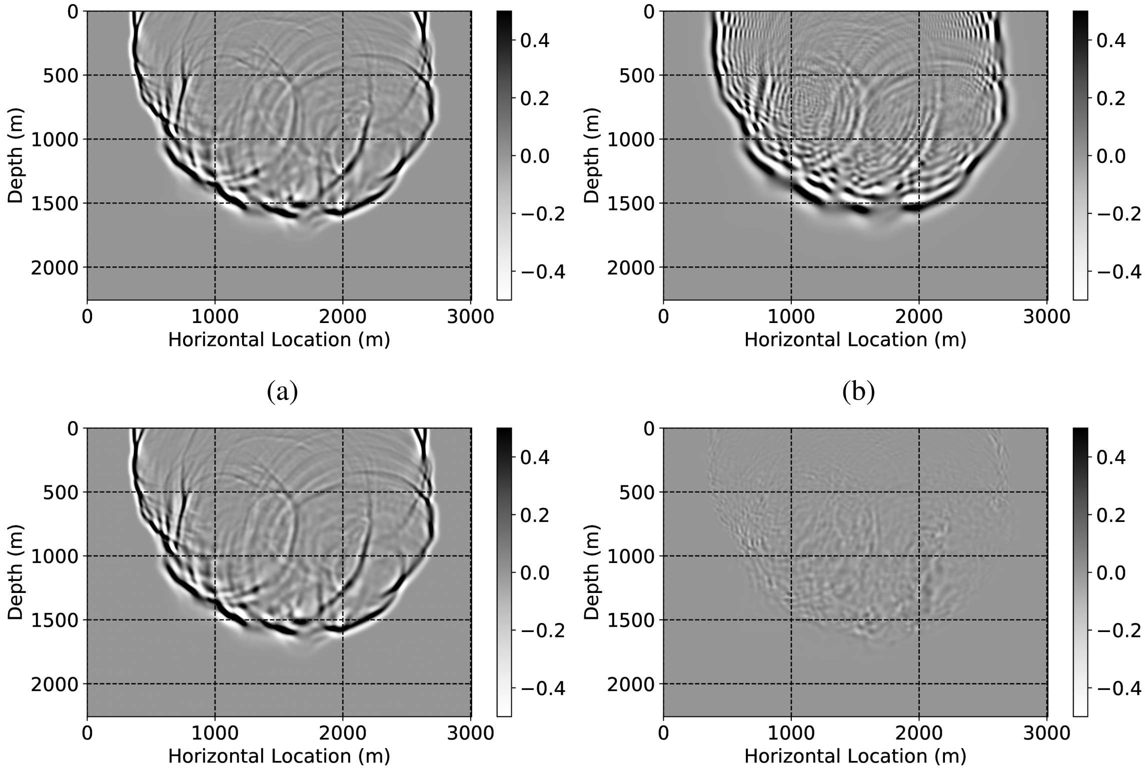


(a)

(b)

(c)

Figure 6.2: Removal of numerical dispersion from migrated shot records. a) Image obtained with high-fidelity migrated shot records. b) Image obtained with low-fidelity migrated shot records. c) Migrated image obtained from corrected low-fidelity shot records.



(a)

(b)

(c)

(d)

Figure 6.3: Removal of numerical dispersion from wavefields. a) High-fidelity simulated wavefield. b) Low-fidelity simulated wavefield. c) Corrected wavefield by the transfer-trained CNN. d) Difference between a) and c).

this by exposing our CNNs to only a small percentage of training data pertinent to the task at hand. Our experiments are exclusively based on numerical simulations. They demonstrate that as long as we are able to pre-train the neural network, e.g., by using data from a neighboring survey or by wave simulations in related velocity models, we can get good performance after finetuning this network with only a few low- and high-fidelity pairs pertinent to the current model. We argue that this may lead to future improvements in efficiency where computationally expensive (e.g., wave-equation driven) processing can partly be replaced by a potentially numerically more efficient neural network. Even though seismic data processing and imaging may differ significantly from location to location, seismic waves and Earth models do share information allowing us to pre-train neural networks by priming them for transfer learning. By following this approach, we rely less on generalizability of our CNNs and more on the availability of information from nearby surveys.

So far, our results are limited to two-dimensional image-to-image mappings, applied to numerically modeled data, and the challenge will be to scale these mappings to higher dimensions. Key in this development will be the ability of these neural networks to generalize sufficiently so that the cost of transfer learning remains small enough.

6.6 Acknowledgments

The authors thank Charles Jones for providing us with the BG Compass velocity model. We also thank Xiaowei Hu for his open-access repository[3] on GitHub. Our software implementation built on this work.

[3] https://github.com/xhujoy/CycleGAN-tensorflow

6.7 References

[1] S. Mandelli, F. Borra, V. Lipari, P. Bestagini, A. Sarti, and S. Tubaro, “Seismic data interpolation through convolutional autoencoder,” *SEG Technical Program Expanded Abstracts 2018*, pp. 4101–4105, 2018. eprint: https://library.seg.org/doi/pdf/10.1190/segam2018-2995428.1 (page 182).

[2] A. Siahkoohi, R. Kumar, and F. Herrmann, “Seismic Data Reconstruction with Generative Adversarial Networks,” *80th EAGE Conference and Exhibition 2018*, Nov. 2018 (page 182).

[3] H. Sun and L. Demanet, “Low-frequency extrapolation with deep learning,” *SEG Technical Program Expanded Abstracts 2018*, pp. 2011–2015, 2018. eprint: https://library.seg.org/doi/pdf/10.1190/segam2018-2997928.1 (pages 182, 191).

[4] L. Mosser, O. Dubrule, and M. Blunt, “Stochastic Seismic Waveform Inversion Using Generative Adversarial Networks as a Geological Prior,” *Mathematical Geosciences*, vol. 84, no. 1, pp. 53–79, 2019 (page 182).

[5] A. Siahkoohi, M. Louboutin, R. Kumar, and F. J. Herrmann, “Deep-convolutional neural networks in prestack seismic: Two exploratory examples,” *SEG Technical Program Expanded Abstracts 2018*, pp. 2196–2200, 2018. eprint: https://library.seg.org/doi/pdf/10.1190/segam2018-2998599.1 (page 182).

[6] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, “How transferable are features in deep neural networks?” In *Proceedings of the 27th International Conference on Neural Information Processing Systems*, ser. NIPS’14, Montreal, Canada, 2014, pp. 3320–3328 (pages 182, 185).

[7] I. Goodfellow *et al.*, “Generative Adversarial Nets,” in *Proceedings of the 27th International Conference on Neural Information Processing Systems*, ser. NIPS’14, Montreal, Canada, 2014, pp. 2672–2680. eprint: http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf (pages 182, 185).

[8] J. Johnson, A. Alahi, and L. Fei-Fei, “Perceptual Losses for Real-Time Style Transfer and Super-Resolution,” in *Computer Vision – European Conference on Computer Vision (ECCV) 2016*, Springer International Publishing, 2016, pp. 694–711. eprint: https://link.springer.com/chapter/10.1007%2F978-3-319-46475-6_43 (pages 183, 185, 186).

[9] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-Image Translation with Conditional Adversarial Networks,” in *The IEEE Conference on Computer Vi-*

sion and Pattern Recognition (CVPR), Jul. 2017, pp. 5967–5976. eprint: https://ieeexplore.ieee.org/document/8100115 (pages 183, 185–187).

[10] L. Amundsen and Ø. Pedersen, "Elimination of temporal dispersion from the finite-difference solutions of wave equations in elastic and anelastic models," *GEOPHYSICS*, vol. 84, no. 2, T47–T58, 2019. eprint: https://doi.org/10.1190/geo2018-0281.1 (pages 184, 191).

[11] M. Long, Y. Cao, J. Wang, and M. I. Jordan, "Learning transferable features with deep adaptation networks," in *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*, ser. ICML'15, Lille, France: JMLR.org, 2015, pp. 97–105 (page 185).

[12] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *CoRR*, vol. abs/1412.6980, 2014. arXiv: 1412.6980 (page 186).

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, pp. 770–778. eprint: https://ieeexplore.ieee.org/document/7780459 (pages 186, 187).

[14] 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 (page 187).

[15] F. Luporini *et al.*, "Architecture and performance of devito, a system for automated stencil computation," *CoRR*, vol. abs/1807.03032, Jul. 2018. arXiv: 1807.03032 (page 187).

[16] D. J. Verschuur, A. Berkhout, and C. Wapenaar, "Adaptive surface-related multiple elimination," *GEOPHYSICS*, vol. 57, no. 9, pp. 1166–1177, 1992. eprint: http://dx.doi.org/10.1190/1.1443330 (page 187).

[17] T. Hu, Z. Han, A. Shrivastava, and M. Zwicker, "Render4completion: Synthesizing multi-view depth maps for 3d shape completion," *arXiv preprint arXiv:1904.08366*, 2019 (page 188).

CHAPTER 7
VELOCITY CONTINUATION WITH FOURIER NEURAL OPERATORS FOR ACCELERATED UNCERTAINTY QUANTIFICATION

7.1 Summary

Seismic imaging is an ill-posed inverse problem that is challenged by noisy data and modeling inaccuracies—due to errors in the background squared-slowness model. Uncertainty quantification is essential for determining how variability in the background models affects seismic imaging. Due to the costs associated with the forward Born modeling operator as well as the high dimensionality of seismic images, quantification of uncertainty is computationally expensive. As such, the main contribution of this work is a survey-specific Fourier neural operator surrogate to velocity continuation that maps seismic images associated with one background model to another virtually for free. While being trained with only 200 background and seismic image pairs, this surrogate is able to accurately predict seismic images associated with new background models, thus accelerating seismic imaging uncertainty quantification. We support our method with a realistic data example in which we quantify seismic imaging uncertainties using a Fourier neural operator surrogate, illustrating how variations in background models affect the position of reflectors in a seismic image.

7.2 Introduction

Seismic imaging involves estimating the short-wavelength component of the Earth's subsurface squared-slowness model—known as the seismic image—given shot records and an estimation of the smooth background squared-slowness model. This linearized imaging problem is challenged by the computationally expensive forward operator as well as presence of measurements noise, linearization errors, modeling errors, and the nontrivial nullspace of the linearized forward Born modeling operator [1, 2, 3]. These challenges highlight the importance of uncertainty quantification (UQ) in seismic imaging, where instead of finding one seismic image estimate, a distribution of seismic images is obtained that explains the observed data [4], consequently reducing the risk of data overfit and enabling UQ [5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15].

The seismic imaging uncertainty can be attributed to two main sources [16, 17, 18]: (1) errors in the data, which include measurement and linearization errors; and (2) modeling errors, which include errors in the estimation of the background squared-slowness model. In this chapter, we focus on uncertainties with respect to the background model as it has the main contributing factor to imaging uncertainty due to its effect on reflector positioning [19, 20, 18]. To this end, we assume there are no measurement or linearization errors and consider the map from shot records to the seismic image to be represented by the deterministic reverse-time migration (RTM) algorithm. In this setup, quantifying the uncertainty in seismic imaging—due to errors in the background model—involves computing numerous RTMs with all the background model posterior samples [8, 9]. Due to the high-dimensionality of seismic images, the number of seismic imaging posterior samples required to obtain accurate estimations of the imaging posterior moments is large, making UQ with respect to the background model computationally expensive for 2D and unfeasible for large 3D problems. These costs can be alleviated via velocity continuation methods [21, 22, 23, 24, 25] that map seismic images associated with one background model to another, without directly solving the imaging problem. Velocity continuation methods are designed to be frugal compared to computing a new RTM for each new background model, which makes them suitable for large-scale seismic imaging uncertainty quantification [19, 20].

While several velocity continuation methods have been proposed [21, 22, 23, 24, 25], they typically involve solving partial differential equations (PDEs), which can be still costly in the context of seismic imaging UQ. To address this challenge, we propose a neural network surrogate for velocity continuation that is capable of mapping seismic images associated with one background model to another with negligible computational cost. Motivated by the success of Fourier neural operators [FNOs, 26] in approximating the solution operator of PDEs [27, 28, 29], we chose them as the architecture for our neural network surrogate. Due to our main interest in accelerating velocity continuation in the context of UQ, we train a survey-specific FNO that acts as a surrogate for velocity continuation for the specific survey at hand. This choice, while not offering generalizations across different survey areas in the Earth, can speed up seismic imaging UQ for the survey at hand, which involves computing seismic images associated with many background model posterior samples. We show that the FNO can be trained using 200 pairs of background and seismic image pairs, making the survey-specific training procedure computationally viable. To scale this method to industry size problems, transfer learning [30] can further reduce the upfront costs of training the FNO. After training, the FNO can be used to obtain samples from the imaging posterior almost free of cost.

In the next section, we describe seismic imaging by introducing the forward Born modeling operator, through which the seismic image relates to the background model. Next, we define velocity continuation, followed by describing our proposed FNO-based approach for accelerating seismic imaging UQ. Finally, we evaluate the performance of the trained neural operator on a realistic dataset, and we demonstrate how the imaging uncertainty affects the positioning of the reflectors in the seismic image.

7.3 Theory

We introduce a deep-network surrogate for velocity continuation in order to enable faster quantification of uncertainty—due to errors in the background velocity model—in seismic imaging. We begin with an introduction to seismic imaging and the linearized forward model associated with it.

7.3.1 Seismic imaging

The inverse problem that we tackle involves the process of estimating the short-wavelength component of the Earth's unknown subsurface squared-slowness model given measurements recorded at the surface. This problem, also known as seismic imaging, can be formulated as a linear inverse problem by linearizing the nonlinear relationship between shot records and the squared-slowness model, governed by the wave-equation. In its simplest acoustic form, the linearization with respect to the slowness model—around a background squared slowness model $\mathbf{m}_0$—leads to a linear inverse problem for estimating the ground truth seismic image $\delta\mathbf{m}^*$ with the following forward model,

$$\delta\mathbf{d}_i = \mathbf{J}(\mathbf{m}_0, \mathbf{q}_i)\delta\mathbf{m}^*. \tag{7.1}$$

In the above expression, $\delta\mathbf{d} = \{\delta\mathbf{d}_i\}_{i=1}^{n_s}$ are n_s linearized shot records and $\mathbf{J}(\mathbf{m}_0, \mathbf{q}_i)$ represents the linearized Born scattering operator. This operator is parameterized by the source signature $\mathbf{q}_i$ and the background squared-slowness model $\mathbf{m}_0$, which is typically estimated in the previous inversion steps [8, 9]. In this work, we focus on potential inaccuracies in the background model, the main source of variability and uncertainty in seismic imaging [19, 20, 18]. Therefore, we assume there are not measurement and linearization errors, which leads to a deterministic mapping from $\{\delta\mathbf{d}_i\}_{i=1}^{n_s}$ to $\delta\mathbf{m}$ for a given background model. We use RTM for this deterministic map, which involves applying the adjoint linearized Born scattering operator to the linearized shot records for all source experiments,

$$\delta\mathbf{m}_{\text{RTM}} = \sum_{i=1}^{n_s} \mathbf{J}(\mathbf{m}_0, \mathbf{q}_i)^{\top}\delta\mathbf{d}_i. \tag{7.2}$$

Since this map is deterministic, the uncertainty in the background model can be translated into uncertainty in seismic imaging by evaluating Equation 7.2 for the background models sampled from the posterior distribution $p(\mathbf{m}_0 \mid \mathbf{d})$, where $\mathbf{d}$ represents the shot records with no linearization (the field data) [8, 9]. Bayesian inference involving high-dimensional posterior distributions, for example seismic imaging and full-waveform inversion, requires many samples from the posterior distribution for accurate Monte Carlo approximation of high-dimensional integrals [31]. As a result, the mapping in Equation 7.2 must be evaluated over numerous background model samples from $p(\mathbf{m}_0 \mid \mathbf{d})$, which is computationally expensive due to costs of applying $\mathbf{J}(\mathbf{m}_0, \mathbf{q}_i)^\top$, $i = 1, \dots, n_s$. To address this computational challenge, our proposed method involves a FNO-based velocity continuation approach that is capable of mapping seismic images associated with one background model to another, effectively replacing a costly demingration-migration. We describe this approach in the next section.

7.3.2 Velocity continuation

At its core, velocity continuation is a process which alters a seismic image based on changes in the background model [21, 22, 24, 25]. Using this technique, an initial seismic image associated with an initial background model $\mathbf{m}_{\text{init}}$ is altered to approximate the seismic image associated with a target background model $\mathbf{m}_{\text{target}}$ without computing Equation 7.2. This process can be interpreted as a map [23]:

$$\mathcal{T}_{(\mathbf{m}_{\text{init}}, \mathbf{m}_{\text{target}})} : \delta\mathcal{M} \to \delta\mathcal{M}, \tag{7.3}$$

where $\delta\mathcal{M}$ denotes the space of seismic images, and the velocity continuation map is parameterized by $\mathbf{m}_{\text{init}}$ and $\mathbf{m}_{\text{target}}$. We take advantage of recent advances in deep learning to train a surrogate model for velocity continuation that approximates $\mathcal{T}_{(\mathbf{m}_{\text{init}}, \mathbf{m}_{\text{target}})} = \mathbf{J}(\mathbf{m}_{\text{target}}, \mathbf{q}_i)^\top (\mathbf{J}(\mathbf{m}_{\text{init}}, \mathbf{q}_i)^\top)^\dagger$, which can be evaluated virtually for free instead of solving four PDEs. Through this approach, we are able to significantly accelerate velocity continuation, trading wave-equation solves for a simple neural network inference. This is of considerable importance in the context of seismic imaging UQ. Before demonstrating the benefits of our approach, we introduce FNOs in the context of velocity continuation.

7.3.3 Fourier neural operators for velocity continuation

In light of their success in learning mesh-free solution operators to PDEs [27, 28, 29], we choose FNOs [26] as a surrogate model for velocity continuation—a process that can be interpreted as a double linearized PDE solve. The main components of FNOs are the Fourier layers, which involve a Fourier transform over the spatial dimensions of their input, followed by a learned pointwise multiplication and an inverse Fourier transform. These layers act as long-kernel convolutional layer, akin to pseudo-spectral methods, which explains the representation power of FNOs [29]. To train a FNO as a surrogate model for velocity continuation, we define it as a map

$$\mathcal{G}_{\mathbf{w}} : \mathcal{M} \times \delta\mathcal{M} \to \delta\mathcal{M}, \tag{7.4}$$

where $\mathcal{G}_{\mathbf{w}}$ denotes the FNO with weights $\mathbf{w}$, and $\mathcal{M}$ is the space of background models. In words, we design $\mathcal{G}_{\mathbf{w}}$ as a neural network that takes as input the target background model and the initial seismic image, and outputs the target seismic image. This choice is analogous to the structure of the velocity continuation map $\mathcal{T}_{(\mathbf{m}_{\text{init}},\mathbf{m}_{\text{target}})}$ (cf. Equation 7.3), except that we fix the initial background model $\mathbf{m}_{\text{init}}$ to an arbitrary background model posterior sample, hence making the dependence of $\mathcal{G}_{\mathbf{w}}$ to $\mathbf{m}_{\text{init}}$ implicit. With this choice of inputs and output for $\mathcal{G}_{\mathbf{w}}$, the FNO's task is to perturb the input initial seismic image according to the provided target background model in order to predict the target seismic image.

Due to our interest in accelerating velocity continuation in the context of UQ, we train a survey-specific FNO that acts as a surrogate for velocity continuation for the specific survey at hand. This choice, while not offering generalizations across different survey areas in the Earth, can speed up seismic imaging UQ for the survey at hand, which involves computing seismic images associated with many posterior background model samples $\mathbf{m}_0 \sim p(\mathbf{m}_0 \mid \mathbf{d})$. By training the FNO on a small number of these background model and seismic image pairs, i.e., approximately 200, we can accelerate the velocity continuation process for the rest of the background model posterior samples while limiting the risk of introducing generalization errors due to the strong heterogeneity of Earth. To achieve this, we construct a set of N training input-output pairs in the form of

$$\left\{ \left(\left(\mathbf{m}_0^{(i)}, \delta \mathbf{m}_{\text{init}} \right), \delta \mathbf{m}_{\text{RTM}}^{(i)} \right) \,\middle|\, i = 1, \ldots, N \right\}, \tag{7.5}$$

where $(\mathbf{m}_0^{(i)}, \delta \mathbf{m}_{\text{init}})$ is the input target background and initial seismic image training pair, and $\delta \mathbf{m}_{\text{RTM}}^{(i)}$ is the associated target seismic image. Training involves minimizing the squared ℓ_2-norm of the difference between the FNO output and the target seismic image with respect to FNO weights,

$$\mathbf{w}^* = \underset{\mathbf{w}}{\arg\min} \ \frac{1}{N} \sum_{i=1} \| \mathcal{G}_{\mathbf{w}}(\mathbf{m}_0^{(i)}, \delta \mathbf{m}_{\text{init}}) - \delta \mathbf{m}_{\text{RTM}}^{(i)} \|_2^2. \tag{7.6}$$

We solve optimization problem 7.6 with the Adam stochastic optimization algorithm [32]. After training, the trained FNO approximates the velocity continuation map for a fixed homogeneous initial background model, i.e.,

$$\mathcal{G}_{\mathbf{w}^*}(\mathbf{m}_{\text{target}}, \delta \mathbf{m}_{\text{init}}) \approx \mathcal{T}_{(\mathbf{m}_{\text{init}}, \mathbf{m}_{\text{target}})}(\delta \mathbf{m}_{\text{init}}) = \delta \mathbf{m}_{\text{target}}, \tag{7.7}$$

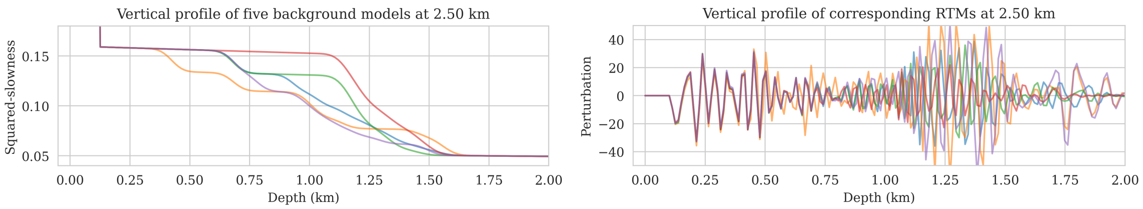
where $\mathbf{m}_{\text{init}}$ is fixed to an arbitrary background model posterior sample and $\delta \mathbf{m}_{\text{target}}$ denotes the seismic image associated with $\delta \mathbf{m}_{\text{init}}$. Given background model samples $\mathbf{m}_0 \sim p(\mathbf{m}_0 \mid \mathbf{d})$ as input, the FNO outputs samples from the imaging posterior, $p(\mathbf{m}_0 \mid \delta \mathbf{d})$. This process accelerates seismic imaging uncertainty quantification as no further RTMs (Equation 7.2) need to be computed. Training the FNO with as little as 200 training pairs makes the survey-specific training procedure computationally viable. To further accelerate the process, training of the FNO can be started as the same time as the background model posterior sampling phase, using the already collected posterior samples as training data. In the next section, we show the results of approximating the velocity continuation map with a FNO via a quasi-real seismic experiment.

7.4 Numerical experiments

The purpose of the presented numerical experiments here is to demonstrate the ability to approximate the velocity continuation map (Equation 7.3) using a FNO. We further show how this trained FNO can be used for UQ by showing the effect of imaging uncertainty on the positioning of the reflectors. We begin by describing the training setup, including the seismic acquisition geometry.

7.4.1 Acquisition geometry and training configuration

Our examples involve imaging a 2D subset of the Parihaka [33, 34] prestrack Kirchhoff migration field dataset. We use this 2D section to create linearized data according to the linear forward model in Equation 7.1, where we consider no measurement and linearization errors. The model is discretized on a $12.5\,\mathrm{m}$ vertical and $20\,\mathrm{m}$ horizontal grid, and the data is acquired with 102 equally spaced sources and 204 fixed receivers. We use a Ricker wavelet with a central frequency of $30\,\mathrm{Hz}$ as the source signature. For presentation purposes, we augment a $125\,\mathrm{m}$ water column on top of these models to limit the near source imaging artifacts. We create background models by assuming access to an oracle, which provides geologically consistent background models with the 2D section. Given 200 background models, we migrate the simulated seismic data via Equation 7.2 to obtain the corresponding seismic images. Figures 7.1a and 7.1b show the vertical profile of five randomly selected background and seismic images at $2.50\,\mathrm{km}$, respectively. These images, displaying strong amplitude and phase differences, highlight the importance of quantifying the uncertainty in imaging when dealing with errors in the background model. To reduce the costs associated



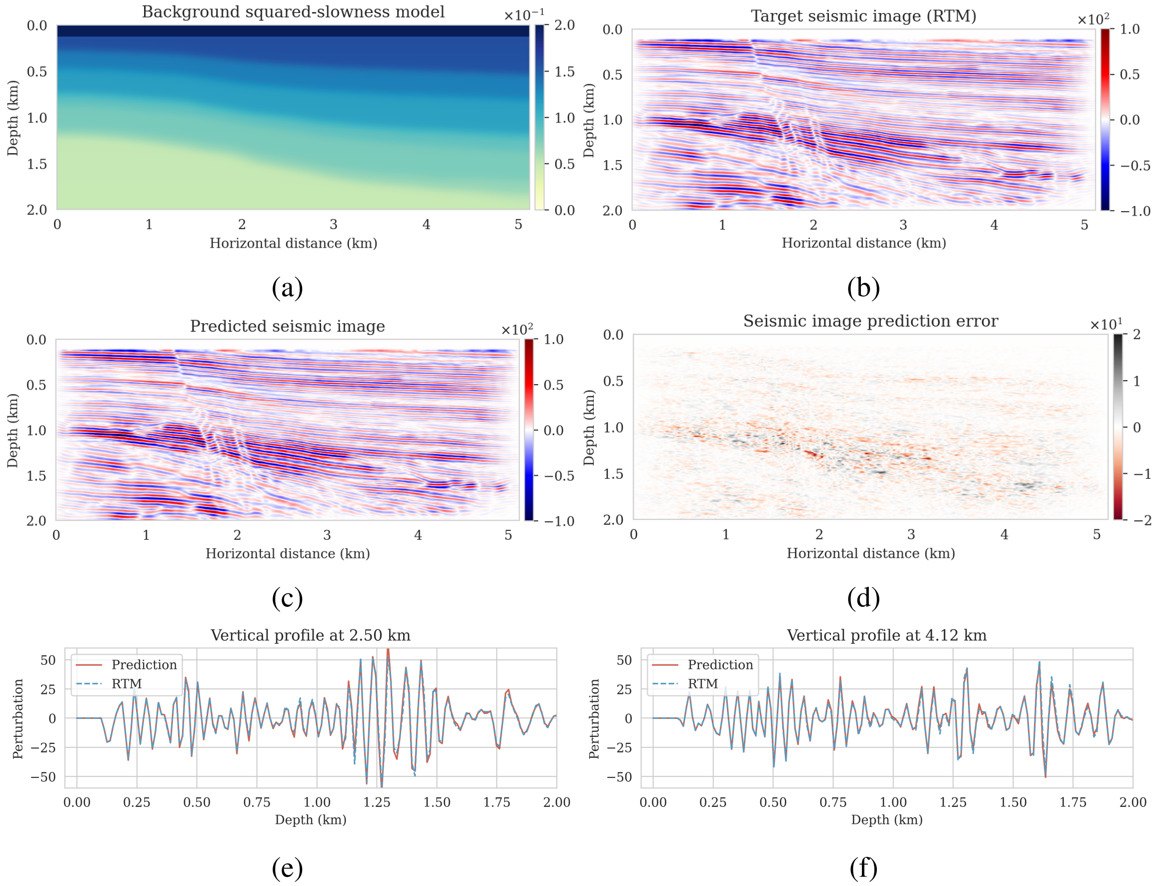
(a) (b)

Figure 7.1: Variation among five (a) background; and (b) seismic images (RTMs), plotted as a vertical profile at $2.50\,\mathrm{km}$.

with UQ, we train the FNO surrogate using the 200 training pairs (cf. Equation 7.5) for 500 epochs with the Adam optimizer [32].

7.4.2 Results

To evaluate the accuracy of the trained FNO in predicting seismic images, we compare its output to the target seismic image, i.e., the RTM image obtained via Equation 7.2 when using the target background model (Figure 7.2a) to parameterize the Born scattering forward operator. This is conducted over the testing dataset, which is derived using the same procedure as the training dataset using the background model creating oracle. Figures 7.2b and 7.2c show the target and predicted seismic images, respectively, and Figures 7.2d includes the difference between them. We observe that the network has accurately predicted the target image, where errors are mostly due to amplitude differences. The accuracy of the prediction in terms of phase and amplitude can be further confirmed by focusing on two vertical profiles, at $2.50\,\mathrm{km}$ (Figure 7.2e) and $4.125\,\mathrm{km}$ (Figure 7.2f) horizontal locations, which include regions with fault and torturous reflectors. We use the trained FNO for UQ by providing testing background models and predicting the associated seismic images. We visualize the obtained uncertainties by showing the variability in the location of reflectors, determined via an automatic horizon tracking software [35]. We pass the seismic images predicted by the FNO to the horizon tracker for 25 selected horizons. As a result, we obtain multiple instances of each horizons, from which we compute pointwise mean and standard



(a)

(b)

(c)

(d)

(e)

(f)

Figure 7.2: Velocity continuation with FNOs. Target (a) background; and (b) seismic images. (c) Predicted seismic image with the FNO. (d) Difference between target and prediction. (e) and (f) Vertical profile comparisons between target (dashed blue) and predicted (red) seismic images.

deviations. Figure 7.3 indicates the result where the solid lines correspond to the mean among different instances of each horizons and shaded areas indicate the mean plus and minus the pointwise standard deviation. These shaded areas indicate uncertainties in the location of the reflectors, which are due to the variability in the background models. As expected, we find a general increase of uncertainty with depth. We also observe that the areas of high uncertainty are correlated with areas of poor illumination, faults and tortuous reflectors.

In this example, we use Devito [36, 37] for the wave-equation based simulations. We based our PyTorch FNO implementation on the original implementation. The code to reproduce our results are made available on GitHub.

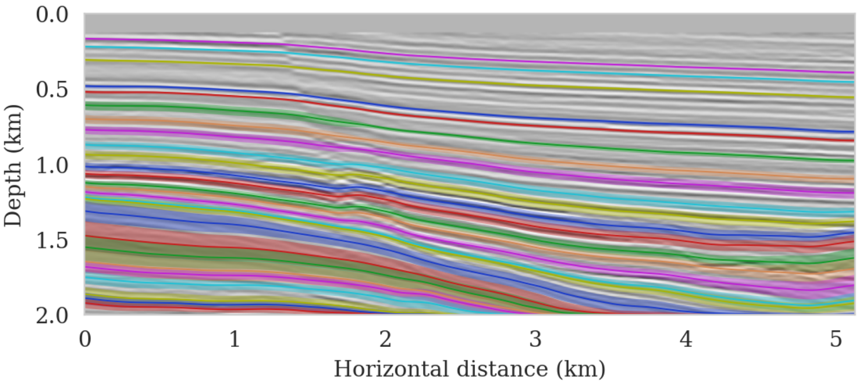


Figure 7.3: Uncertainty in the tracked horizons.

7.5 Conclusions and discussion

Quantifying the uncertainty in seismic imaging due to errors in the background model involves solving many seismic imaging problems that vary in the parameterization of the background model of the forward operator. To reduce the computational cost of this process—mainly due to the computational costs of the forward operator—we proposed to train a survey-specific Fourier neural operator surrogate that mimics velocity continuation. This surrogate model maps seismic images associated with one background model to another virtually for free, which has the benefit of accelerating uncertainty quantification. We showed that this surrogate model can be trained with as few as 200 training pairs while still providing a good seismic image prediction accuracy. Further research is required in training a reliable global surrogate, being able to generalize across other survey areas and more realistic physics.

7.6 References

[1] G. Lambaré, J. Virieux, R. Madariaga, and S. Jin, "Iterative asymptotic inversion in the acoustic approximation," *Geophysics*, vol. 57, no. 9, pp. 1138–1154, 1992 (page 198).

[2] G. T. Schuster, "Least-squares cross-well migration," in *63rd Annual International Meeting, SEG*, Expanded Abstracts, 1993, pp. 110–113 (page 198).

[3] T. Nemeth, C. Wu, and G. T. Schuster, "Least-squares migration of incomplete reflection data," *GEOPHYSICS*, vol. 64, no. 1, pp. 208–221, 1999 (page 198).

[4] A. Tarantola, *Inverse problem theory and methods for model parameter estimation*. SIAM, 2005, ISBN: 978-0-89871-572-9 (page 198).

[5] A. Malinverno and R. L. Parker, "Two ways to quantify uncertainty in geophysical inverse problems," *GEOPHYSICS*, vol. 71, no. 3, W15–W27, 2006 (page 198).

[6] 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. eprint: http://epubs.siam.org/doi/pdf/10.1137/110845598 (page 198).

[7] A. Ray, S. Kaplan, J. Washbourne, and U. Albertin, "Low frequency full waveform seismic inversion within a tree based Bayesian framework," *Geophysical Journal International*, vol. 212, no. 1, pp. 522–542, Oct. 2017. eprint: https://academic.oup.com/gji/article-pdf/212/1/522/21782947/ggx428.pdf (page 198).

[8] H. Zhu, S. Li, S. Fomel, G. Stadler, and O. Ghattas, "A bayesian approach to estimate uncertainty for full-waveform inversion using a priori information from depth migration," *Geophysics*, vol. 81, no. 5, R307–R323, 2016 (pages 198, 200, 201).

[9] Z. Fang, C. D. Silva, R. Kuske, and F. J. Herrmann, "Uncertainty quantification for inverse problems with weak partial-differential-equation constraints," *GEOPHYSICS*, vol. 83, no. 6, R629–R647, 2018 (pages 198, 200, 201).

[10] G. K. Stuart, S. E. Minkoff, and F. Pereira, "A two-stage Markov chain Monte Carlo method for seismic inversion and uncertainty quantification," *GEOPHYSICS*, vol. 84, no. 6, R1003–R1020, Nov. 2019 (page 198).

[11] F. J. Herrmann, A. Siahkoohi, and G. Rizzuti, "Learned imaging with constraints and uncertainty quantification," in *Neural Information Processing Systems (NeurIPS) 2019 Deep Inverse Workshop*, Dec. 2019 (page 198).

[12] A. Siahkoohi, G. Rizzuti, and F. J. Herrmann, "A deep-learning based bayesian approach to seismic imaging and uncertainty quantification," in *82nd EAGE Conference and Exhibition*, Extended Abstracts, 2020 (page 198).

[13] A. Siahkoohi, G. Rizzuti, and F. J. Herrmann, "Uncertainty quantification in imaging and automatic horizon tracking—a Bayesian deep-prior based approach," in *90th Annual International Meeting, SEG*, Expanded Abstracts, Sep. 2020, pp. 1636–1640 (page 198).

[14] A. Siahkoohi and F. J. Herrmann, "Learning by example: fast reliability-aware seismic imaging with normalizing flows," in *First International Meeting for Applied Geoscience & Energy*, Expanded Abstracts, 2021, pp. 1580–1585 (page 198).

[15] A. Siahkoohi, G. Rizzuti, and F. J. Herrmann, "Deep bayesian inference for seismic imaging with tasks," *arXiv preprint arXiv:2110.04825*, 2021 (page 198).

[16] P. Thore, A. Shtuka, M. Lecour, T. Ait-Ettajer, and R. Cognot, "Structural uncertainties: Determination, management, and applications," *Geophysics*, vol. 67, no. 3, pp. 840–852, 2002 (page 198).

[17] K. Osypov *et al.*, "Model-uncertainty quantification in seismic tomography: Method and applications," *Geophysical Prospecting*, vol. 61, no. 6-Challenges of Seismic Imaging and Inversion Devoted to Goldin, pp. 1114–1134, 2013 (page 198).

[18] G. Ely, A. Malcolm, and O. V. Poliannikov, "Assessing uncertainties in velocity models and images with a fast nonlinear uncertainty quantification method," *GEOPHYSICS*, vol. 83, no. 2, R63–R75, 2018. eprint: https://doi.org/10.1190/geo2017-0321.1 (pages 198, 200).

[19] S. Fomel and E. Landa, "Structural uncertainty of time-migrated seismic images," *Journal of Applied Geophysics*, vol. 101, pp. 27–30, 2014 (pages 198, 200).

[20] O. V. Poliannikov and A. E. Malcolm, "The effect of velocity uncertainty on migrated reflectors: Improvements from relative-depth imaging," *Geophysics*, vol. 81, no. 1, S21–S29, 2016 (pages 198, 200).

[21] S. Fomel, "Time-migration velocity analysis by velocity continuation," *Geophysics*, vol. 68, no. 5, pp. 1662–1672, 2003 (pages 198, 201).

[22] S. Fomel, "Velocity continuation and the anatomy of residual prestack time migration," *Geophysics*, vol. 68, no. 5, pp. 1650–1661, 2003 (pages 198, 201).

[23] A. A. Duchkov and M. V. De Hoop, "Velocity continuation in the downward continuation approach to seismic imaging," *Geophysical Journal International*, vol. 176, no. 3, pp. 909–924, 2009 (pages 198, 201).

[24] T. van Leeuwen and F. J. Herrmann, "Wave-equation Extended Images-Computation and Velocity Continuation," in *74th EAGE Conference and Exhibition incorporating EUROPEC 2012*, European Association of Geoscientists & Engineers, 2012, cp–293 (pages 198, 201).

[25] M. Yang, M. Graff, R. Kumar, and F. J. Herrmann, "Low-rank representation of omnidirectional subsurface extended image volumes," *Geophysics*, vol. 86, no. 3, S165–S183, 2021 (pages 198, 201).

[26] Z. Li *et al.*, "Fourier Neural Operator for Parametric Partial Differential Equations," in *9th International Conference on Learning Representations*, OpenReview.net, 2021 (pages 199, 202).

[27] J. Q. Toledo-Marín, G. Fox, J. P. Sluka, and J. A. Glazier, "Deep learning approaches to surrogates for solving the diffusion equation for mechanistic real-world simulations," *Frontiers in Physiology*, vol. 12, 2021 (pages 199, 202).

[28] T. Konuk and J. Shragge, "Physics-guided deep learning using fourier neural operators for solving the acoustic vti wave equation," in *82nd EAGE Annual Conference & Exhibition*, European Association of Geoscientists & Engineers, vol. 2021, 2021, pp. 1–5 (pages 199, 202).

[29] N. Kovachki, S. Lanthaler, and S. Mishra, "On universal approximation and error bounds for fourier neural operators," *Journal of Machine Learning Research*, vol. 22, Art–No, 2021 (pages 199, 202).

[30] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" In *Proceedings of the 27th International Conference on Neural Information Processing Systems*, ser. NIPS'14, Montreal, Canada, 2014, pp. 3320–3328 (page 199).

[31] A. Gelman and D. B. Rubin, "Inference from Iterative Simulation Using Multiple Sequences," *Statistical Science*, vol. 7, no. 4, pp. 457–472, 1992 (page 201).

[32] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014 (pages 203, 205).

[33] Veritas, "Parihaka 3D Marine Seismic Survey - Acquisition and Processing Report," New Zealand Petroleum & Minerals, Wellington, Tech. Rep. New Zealand Petroleum Report 3460, 2005 (page 204).

[34] WesternGeco., "Parihaka 3D PSTM Final Processing Report," New Zealand Petroleum & Minerals, Wellington, Tech. Rep. New Zealand Petroleum Report 4582, 2012 (page 204).

[35] X. Wu and S. Fomel, “Least-squares horizons with local slopes and multi-grid correlations,” *GEOPHYSICS*, vol. 83, pp. IM29–IM40, 4 2018 (page 205).

[36] F. Luporini *et al.*, “Architecture and performance of devito, a system for automated stencil computation,” *CoRR*, vol. abs/1807.03032, Jul. 2018. arXiv: 1807.03032 (page 206).

[37] 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 (page 206).

CHAPTER 8
CONCLUSIONS

In summary, this thesis has contributed several methods based on the recent advancements in machine learning to address the challenges associated with solving large-scale inverse problems. These challenges are: (1) choosing a prior distribution that captures our prior knowledge about the unknown, while preventing unwanted biases due to overly simplistic priors; and (2) the computational costs associated with solving inverse problems, in particular costs of sampling the posterior distribution during Bayesian inference. I summarize the findings and conclusions that were presented throughout this thesis in regard to these research questions.

8.1 Deep priors for imaging and uncertainty quantification

Bayesian inference on high-dimensional inverse problems with computationally expensive forward modeling operators, has been, and continues to be a major challenge in the field of seismic imaging. Aside from obvious computational challenges, the selection of effective priors is problematic given the heterogeneity across geological scenarios and scales exhibited by elastic properties of the Earth's subsurface. To limit the possibly heavy-handed bias induced by a handcrafted prior, in chapter 2, I proposed regularization via deep priors. I presented examples demonstrating the beneficial regularization properties of deep priors and proposed the use of conditional mean estimates rather than maximum a posteriori (MAP) estimation due to its robustness to overfitting. To partly offset the costs associated with Bayesian inference, I used stochastic gradient Langevin dynamics (SGLD) to avoiding exact calculation of the multi-source data likelihood function. Aside from providing a reasonable assessment of the uncertainty, with pointwise standard deviation increasing in complex areas or in areas of relatively poor illumination, I also used the samples from the posterior to propagate uncertainties due to errors in the shot data all the way to the task of automatic horizon tracking or other tasks.

While deep priors provide a favorable regularization property, it comes at a significant computational price due to the large number of optimization steps required to update the randomly initialized weights. These computational costs are compounded by the fact that the parameterization of seismic images in terms of a deep neural network is highly overparameterized. To partially reduce the costs of employing deep priors, chapter 3 relaxes the deep priors by decoupling seismic image and network weights updates during optimization, hence reducing the costs of solving the inverse problem. Via numerical experiments I verified that this relaxation is still able to resolve the imaging artifacts present in conventional least-squares imaging when data is contaminated by strong noise.

8.2 Low-cost and reliable Bayesian inference with amortized variational inference

Chapters 4 and 5 take a variational inference approach to reduce the computational costs of Bayesian inference in large-scale inverse problems. These methods are fundamentally different than the Markov chain Monte Carlo approach proposed in chapter 2 and have the potential to reduce posterior sampling costs by incurring an up-front cost of pretraining a conditional normalizing flow that fully learns the posterior distribution. These methods are designed to take full advantage of having access to training pairs drawn from a joint distribution, which in some domains such as geophysical inverse problems, is close but not equal to the actual joint distribution of model and data. The pretrained network then can be used preconditioning a physics-based variational inference formulation that involves approximating the target posterior distribution, after which, we gain cheap access to samples from the posterior distribution.

Chapter 4 improves the data-driven posterior approximation via the pretrained conditional flow by finetuning its weights such that its output better approximates the target posterior distribution in KL divergence sense. By means of a series of examples, I demonstrated that this preconditioned scheme leads to considerable speed-ups compared to training a normalizing flow from scratch. On the other hand, chapter 5 proposes a correction to the conditional normalizing flow latent distribution such that the predicted posterior distribution more closely matches to the desired posterior distribution. Using a seismic imaging example, I demonstrated that the proposed latent distribution correction can be used to mitigate the errors due to data distribution shifts, which includes changes in the forward model as well as the prior distribution. Observations indicated improvements in seismic image quality after the latent distribution correction step, as well as reasonable estimates of pointwise standard deviations on the image, suggesting an increase in variability in complex geological areas and poorly illuminated areas.

8.3 Mitigating forward modeling errors

Chapter 6 proposed a data-driven approach to mitigate numerical dispersion artifacts in wave-equation simulations that are due to course finite-difference discretizations. I demonstrates that as long as we are able to pre-train the neural network, e.g., by using data from a neighboring survey or by wave simulations in related velocity models, we can get good performance after finetuning this network with only a few low- and high-fidelity pairs pertinent to the current model. This may lead to improvements in efficiency where computationally expensive (e.g., wave-equation driven) processing can partly be replaced by a potentially numerically more efficient neural network.

Chapter 7 reduces the computational cost of seismic imaging uncertainty quantification due to errors in the wave-equation squared-slowness parameterization. This source of uncertainty is fundamentally different than the uncertainty quantified in chapters 2, 4 and 5 in which we assumed access to accurate background squared-slowness models. Quantifying the uncertainty due to errors in the background model requires solving numerous seismic imaging problems, which is computationally burdensome. To reduce the computational costs, I proposed to train a survey-specific Fourier neural operator surrogate that mimics velocity continuation—a mapping that translates seismic images associated with one background model to another. I showed that this surrogate model can be trained with as few as 200 training pairs while still providing a good seismic image prediction accuracy.

8.4 Current limitations and future directions

- In the context of Bayesian inference with deep priors, the number of iterations needed by the SGLD algorithm remains high and prohibitive for imaging problems involving 3D seismic images. These costs can be potentially reduced by projecting the network weights onto a low-dimensional subspace and performing posterior inference within this reduced subspace [1]. Alternatively, a Gaussian variational inference approximation to the the deep prior network weights might be a promising alternative.

- In the context of reliable data-driven variational inference for large-scale inverse problems, quantifying the extent to which finetuning the pretrained conditional normalizing flow weights and the diagonal correction to the latent distribution can mitigates errors resulting from data distribution shifts. A theoretical analysis for this approach would be an interesting future research direction.

- The neural surrogate for velocity continuation is specific to the survey that is trained on. To further reduce the computational cost of Bayesian inference, the surrogate model can be trained on a family of survey areas and later can be adapted to the specific survey of interest with a fewer training pairs.

8.5 References

[1] P. Izmailov, W. J. Maddox, P. Kirichenko, T. Garipov, D. Vetrov, and A. G. Wilson, "Subspace inference for Bayesian deep learning," in *Uncertainty in Artificial Intelligence*, Proceedings of Machine Learning Research (PMLR), 2020, pp. 1169–1179 (page 215).

Appendices

APPENDIX A
DEEP-PRIOR NETWORK ARCHITECTURE

The CNN architecture proposed by [1] is a variation of the widely used U-net architecture, as described by [2]. The U-net architecture is composed of an encoder and a decoder module, where information, i.e., intermediate values, from the encoder module is also passed to the decoder through skip connections.

The following are the major differences between the deep prior architecture [1] and the U-net architecture. Contrary to U-net, the convolutional layers in the decoding module of the deep prior architecture do not increase the dimensionality of the intermediate values. Rather, dimensionality-preserving convolutional layers, that is, stride-one convolutions, are augmented with user-defined interpolation schemes to achieve upsampling. This enables the degree of smoothness in the image space to be controlled by the interpolation kernel. Another difference worth noting is the way intermediate values from the encoding phase are incorporated into the decoding module through skip connections. In the deep prior architecture, the intermediate values in the encoding phase are passed through an additional convolutional layer before being fed into the decoder module.

Figure A.1 illustrates the exact deep prior architecture used in the Parihaka example in this paper, which closely follows the architecture advocated by [1]. The blocks in the figure represent the intermediate values in the network, where the color indicates the operation that produced these values. For instance, the lightest gray blocks are intermediate values obtained from applying two-dimensional convolutions, whereas the darkest gray block with dashed white edges represents the result of a user-selected interpolation method, which is in this instance nearest neighbor interpolation. For further information regarding the rest of the colors, please refer to the legend in Figure A.1. The left most shaded block in Figure A.1 is the $m \times n \times c$ input, which in the case of deep priors is a fixed random array, with m, n

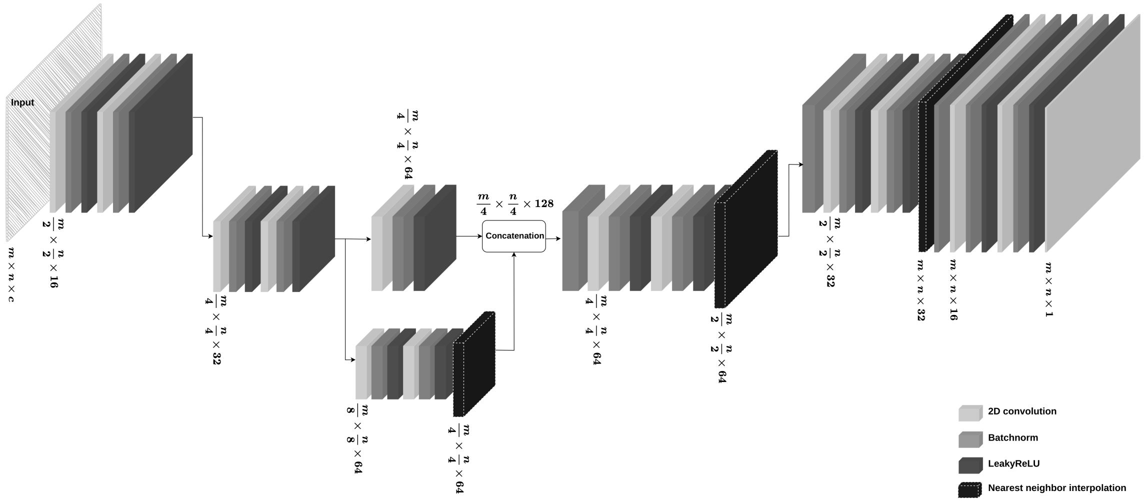

Input
$m \times n \times c$
$\frac{m}{2} \times \frac{n}{2} \times 16$
$\frac{m}{4} \times \frac{n}{4} \times 32$
$\frac{m}{4} \times \frac{n}{4} \times 64$
$\frac{m}{4} \times \frac{n}{4} \times 128$
Concatenation
$\frac{m}{8} \times \frac{n}{8} \times 64$
$\frac{m}{4} \times \frac{n}{4} \times 64$
$\frac{m}{4} \times \frac{n}{4} \times 64$
$\frac{m}{2} \times \frac{n}{2} \times 64$
$\frac{m}{2} \times \frac{n}{2} \times 32$
$m \times n \times 32$
$m \times n \times 16$
$m \times n \times 1$
2D convolution
Batchnorm
LeakyReLU
Nearest neighbor interpolation

Figure A.1: The architecture used for the Parihaka example.

indicating its horizontal and vertical dimensions, and c which represents the number of channels. If a layer alters the dimension of an intermediate value, the new dimensions are inserted adjacent to the representative block. For instance, the leftmost convolutional layer that takes in the input noise changes its dimensionality to $\frac{m}{2} \times \frac{n}{2} \times 16$. All two-dimensional convolutional layers use 5×5 kernels. The convolutional layers that reduce the horizontal and vertical dimensions have a stride of two while all the other convolutional layers have a stride of one. Lastly, we apply a fixed scaler to the output of the network such that the range of the output values for the CNN with fixed input and randomly drawn weights $\mathbf{w} \sim \mathrm{N}\big(\mathbf{w} \mid \mathbf{0}, \lambda^{-2}\mathbf{I}\big)$ covers the a priori known range of amplitudes in the unknown perturbation model.

The Compass example uses the same architecture as Figure A.1, except that it includes an additional downsampling layer in the encoder and an additional upsampling layer in the decoder. We found the addition of these layers to be helpful since the frequency content of the model in the Compass example was higher and the dimensions larger.

APPENDIX B
MATHEMATICAL DERIVATIONS

Let $f : \mathcal{Z} \to \mathcal{X}$ be a bijective transformation that maps a random variable $\boldsymbol{z} \sim \pi_z(\boldsymbol{z})$ to $\boldsymbol{x} \sim \pi_x(\boldsymbol{x})$. We can write the change of variable formula [3] that relates probability density functions π_z and π_x in the following manner:

$$\pi_x(\boldsymbol{x}) = \pi_z(\boldsymbol{z}) \left| \det \nabla_x f^{-1}(\boldsymbol{x}) \right|, \quad f(\boldsymbol{z}) = \boldsymbol{x}, \quad \boldsymbol{x} \in \mathcal{X}. \tag{B.1}$$

This relation serves as the basis for the objective functions used throughout this paper.

B.1 Derivation of the physics-based variational inference objective

In equation 4.2, we train a bijective transformation, denoted by $T_\theta : \mathcal{Z}_x \to \mathcal{X}$, that maps the latent distribution $\pi_{z_x}(\boldsymbol{z}_x)$ to the high-fidelity posterior density $\pi_{\text{post}}(\boldsymbol{x} \mid \boldsymbol{y})$. We optimize the parameters of T_θ by minimizing the KL divergence between the push-forward density [4], denoted by $\pi_\theta(\cdot \mid \boldsymbol{y}) := T_\sharp\, \pi_{z_x}$, and the posterior density:

$$\begin{aligned} &\mathop{\arg\min}_\theta \, \mathbb{D}_{\text{KL}} \left(\pi_\theta(\cdot \mid \boldsymbol{y}) \, || \, \pi_{\text{post}}(\cdot \mid \boldsymbol{y}) \right) \\ &= \mathop{\arg\min}_\theta \, \mathbb{E}_{\boldsymbol{x} \sim \pi_\theta(\boldsymbol{x} \mid \boldsymbol{y})} \Big[- \log \pi_{\text{post}}(\boldsymbol{x} \mid \boldsymbol{y}) + \log \pi_\theta(\boldsymbol{x} \mid \boldsymbol{y}) \Big]. \end{aligned} \tag{B.2}$$

In the above expression, we can rewrite the expectation with respect to $\pi_\theta(\boldsymbol{x} \mid \boldsymbol{y})$ as the expectation with respect to the latent distribution, followed by a mapping via T_θ—i.e.,

$$\mathop{\arg\min}_\theta \, \mathbb{E}_{\boldsymbol{z}_x \sim \pi_{z_x}(\boldsymbol{z})} \Big[- \log \pi_{\text{post}}(T_\theta(\boldsymbol{z}_x) \mid \boldsymbol{y}) + \log \pi_\theta(T_\theta(\boldsymbol{z}_x) \mid \boldsymbol{y}) \Big]. \tag{B.3}$$

The last term in the expectation above can be further simplified via the change of variable formula in equation B.1. If $\boldsymbol{x} = T_\theta(\boldsymbol{z}_x)$, then:

$$\pi_\theta(\boldsymbol{x} \mid \boldsymbol{y}) = \pi_{z_x}(\boldsymbol{z}_x) \left| \det \nabla_x T_\theta^{-1}(\boldsymbol{x}) \right| = \pi_{z_x}(\boldsymbol{z}_x) \left| \det \nabla_{z_x} T_\theta(\boldsymbol{z}_x) \right|^{-1}. \tag{B.4}$$

The last equality in equation B.4 holds due to the invertibility of T_θ and the differentiability of its inverse (inverse function theorem). By combining equations B.3 and B.4, we arrive at the following objective function for training T_θ:

$$\arg\min_\theta \mathbb{E}_{\boldsymbol{z}_x \sim \pi_{z_x}(\boldsymbol{z})} \Big[- \log \pi_{\text{post}}(T_\theta(\boldsymbol{z}_x) \mid \boldsymbol{y}) + \log \pi_{z_x}(\boldsymbol{z}_x) - \log \Big| \det \nabla_{z_x} T_\theta(\boldsymbol{z}_x) \Big| \Big]. \tag{B.5}$$

Finally, by ignoring the $\log \pi_{z_x}(\boldsymbol{z}_x)$ term, which is constant with respect to $\boldsymbol{\theta}$, using Bayes' rule, and inserting our data likelihood model from equation 4.1, we derive equation 4.2:

$$\min_\theta \mathbb{E}_{\,\boldsymbol{z}_x \sim \pi_{z_x}(\boldsymbol{z}_x)} \left[\frac{1}{2\sigma^2} \left\| F\big(T_\theta(\boldsymbol{z}_x)\big) - \boldsymbol{y} \right\|_2^2 - \log \pi_{\text{prior}}\big(T_\theta(\boldsymbol{z}_x)\big) - \log \Big| \det \nabla_{z_x} T_\theta(\boldsymbol{z}_x) \Big| \right]. \tag{B.6}$$

Next, based on this equation, we derive the objective function used in the pretraining phase.

B.2 Derivation of the data-driven variational inference objective

The derivation of objective in equation 4.3 follows directly from the change of variable formula in equation B.1, applied to a bijective map, $G_\phi^{-1} : \mathcal{Z}_y \times \mathcal{Z}_x \to \mathcal{Y} \times \mathcal{X}$, where $\mathcal{Z}_y$

and $\mathcal{Z}_y$ are Gaussian latent spaces. That is to say:

$$\widehat{\pi}_{y,x}(\boldsymbol{y},\boldsymbol{x}) = \pi_{z_y,z_x}(\boldsymbol{z}_y,\boldsymbol{z}_x)\left|\det \nabla_{y,x} G_\phi(\boldsymbol{y},\boldsymbol{x})\right|, \quad G_\phi(\boldsymbol{y},\boldsymbol{x}) = \begin{bmatrix} \boldsymbol{z}_y \\ \boldsymbol{z}_x \end{bmatrix}. \tag{B.7}$$

Given (low-fidelity) training pairs, $\boldsymbol{y},\boldsymbol{x} \sim \widehat{\pi}_{y,x}(\boldsymbol{y},\boldsymbol{x})$, the maximum likelihood estimate of $\boldsymbol{\phi}$ is obtained via the following objective:

$$\begin{aligned}
&\arg\max_{\phi} \mathbb{E}_{\boldsymbol{y},\boldsymbol{x}\sim\widehat{\pi}_{y,x}(\boldsymbol{y},\boldsymbol{x})} \left[\log \widehat{\pi}_{y,x}(\boldsymbol{y},\boldsymbol{x})\right] \\
&= \arg\min_{\phi} \mathbb{E}_{\boldsymbol{y},\boldsymbol{x}\sim\widehat{\pi}_{y,x}(\boldsymbol{y},\boldsymbol{x})} \left[-\log \pi_{z_y,z_x}(\boldsymbol{z}_y,\boldsymbol{z}_x) - \log\left|\det \nabla_{y,x}\, G_\phi(\boldsymbol{y},\boldsymbol{x})\right|\right] \\
&= \arg\min_{\phi} \mathbb{E}_{\boldsymbol{y},\boldsymbol{x}\sim\widehat{\pi}_{y,x}(\boldsymbol{y},\boldsymbol{x})} \left[\frac{1}{2}\left\|G_\phi(\boldsymbol{y},\boldsymbol{x})\right\|^2 - \log\left|\det \nabla_{y,x}\, G_\phi(\boldsymbol{y},\boldsymbol{x})\right|\right],
\end{aligned} \tag{B.8}$$

that is the objective function in equation 4.3. The NF trained via the objective function, given samples from the latent distribution, draws samples from the low-fidelity joint distribution, $\widehat{\pi}_{y,x}$.

By construction, G_ϕ is a block-triangular map—i.e.,

$$G_\phi(\boldsymbol{y},\boldsymbol{x}) = \begin{bmatrix} G_{\phi_y}(\boldsymbol{y}) \\ G_{\phi_x}(\boldsymbol{y},\boldsymbol{x}) \end{bmatrix}, \ \boldsymbol{\phi} = \{\boldsymbol{\phi}_y, \boldsymbol{\phi}_x\}. \tag{B.9}$$

[5] showed that after solving the optimization problem in equation 4.3, G_ϕ approximates the well-known triangular Knothe-Rosenblat map [6]. As shown in [7], the triangular structure and G_ϕ's invertibility yields the following property,

$$\left(G_{\phi_x}^{-1}(G_{\phi_y}(\boldsymbol{y}),\cdot\,)\right)_\sharp \pi_{z_x} = \widehat{\pi}_{\text{post}}(\,\cdot \mid \boldsymbol{y}), \tag{B.10}$$

where $\widehat{\pi}_{\text{post}}$ denotes the low-fidelity posterior probability density function. The expression above means we can get access to low-fidelity posterior distribution samples by simply evaluating $G_{\phi_x}^{-1}(G_{\phi_y}(\boldsymbol{y}),\boldsymbol{z}_x)$ for $\boldsymbol{z}_x \sim \pi_{z_x}(\boldsymbol{z}_x)$ for a given observed data $\boldsymbol{y}$.

APPENDIX C
TRAINING DETAILS AND NETWORK ARCHITECTURES FOR EXAMPLES IN CHAPTER 4

For our network architecture, we adapt the recursive coupling blocks proposed by [5], which use invertible coupling layers from [8] in a hierarchical way. In other words, we recursively divide the incoming state variables and apply an affine coupling layer. The final architecture is obtained by composing several of these hierarchical coupling blocks. The hierarchical structure leads to dense triangular Jacobian, which is essential in representation power of NFs [5].

For all examples in this paper, we use the hierarchical coupling blocks as described in [5]. The affine coupling layers within each hierarchal block contain a residual block as described in [9]. Each residual block has the following dimensions: 64 input, 128 hidden, and 64 output channels, except for the first and last coupling layer where we have 4 input and output channels, respectively. We use the Wavelet transform and its transpose before feeding seismic images into the network and after the last layer of the NFs.

Below, we describe the network architectures and training details regarding the two numerical experiments described in the paper. Throughout the experiments, we use the Adam optimization algorithm [10].

C.1 2D toy example

We use 8 hierarchal coupling blocks, as described above for both G_{ϕ_x} and G_{ϕ_y} (equation 4.3). As a result, due to our proposed method in equation 4.4, we choose the same architecture for T_{ϕ_x} (equation 4.2).

For pretraining G_θ according to equation 4.3, we use 5000 low-fidelity joint training pairs, $\boldsymbol{y}, \boldsymbol{x} \sim \widehat{\pi}_{y,x}(\boldsymbol{y}, \boldsymbol{x})$. We minimize equation 4.3 for 25 epochs with a batch size of 64

and a (starting) learning rate of 0.001. We decrease the learning rate each epoch by a factor of 0.9.

For the preconditioned step—i.e., solving equation 4.6, we use 1000 latent training samples. We train for 5 epochs with a batch size of 64 and a learning rate 0.001. Finally, as a comparison, we solve the objective in equation 4.6 for a randomly initialized NF with the same 1000 latent training samples for 25 epochs. We decrease the learning rate each epoch by 0.9.

C.2 Seismic compressed sensing

We use 12 hierarchal coupling blocks, as described above for both G_{ϕ_x}, G_{ϕ_y}, and we use the same architecture for T_{ϕ_x} as G_{ϕ_x}.

For pretraining G_θ according to equation 4.3, we use 5282 low-fidelity joint training pairs, $\boldsymbol{y}, \boldsymbol{x} \sim \widehat{\pi}_{y,x}(\boldsymbol{y}, \boldsymbol{x})$. We minimize equation 4.3 for 50 epochs with a batch size of 4 and a starting learning rate of 0.001. Once again, we decrease the learning rate each epoch by 0.9.

For the preconditioned step—i.e., solving equation 4.6, we use 1000 latent training samples. We train for 10 epochs with a batch size 4 and a learning rate of 0.0001, where we decay the step by 0.9 in every 5th epoch.

APPENDIX D

2D TOY EXAMPLE FROM CHAPTER 4—MORE RESULTS

Here we show the effect γ on our proposed method in the 2D toy experiment. By choosing smaller values for γ, we make $\bar{\boldsymbol{A}} \,/\, \rho(\bar{\boldsymbol{A}})$ with $\bar{\boldsymbol{A}} = \boldsymbol{\Gamma} + \gamma \boldsymbol{I}$ less close to the identity matrix, hence enhancing the discrepancy between the low- and high-fidelity posterior. The first row of Figure D.1 shows the low- (purple) and high-fidelity (red) data densities for decreasing values of γ from 2 down to 0. The second row depicts the predicted posterior densities via the preconditioned scheme (orange contours) and the low-fidelity posterior in green along with MCMC samples (dark circles). The third row compares the preconditioned posterior densities to samples obtained via the low-fidelity pretrained NF—i.e., equation 4.3. Finally, the last row shows the objective function values during training with (orange) and without (green) preconditioning.

We observe that by decreasing γ from 2 to 0, the low-fidelity posterior approximations become worse. As a result, the objective function for the preconditioned approach (orange) at the beginning start from a higher value, indicating more mismatch between low- and high-fidelity posterior densities. Finally, our preconditioning method consistently improves upon low-fidelity posterior by training for 5 epochs.

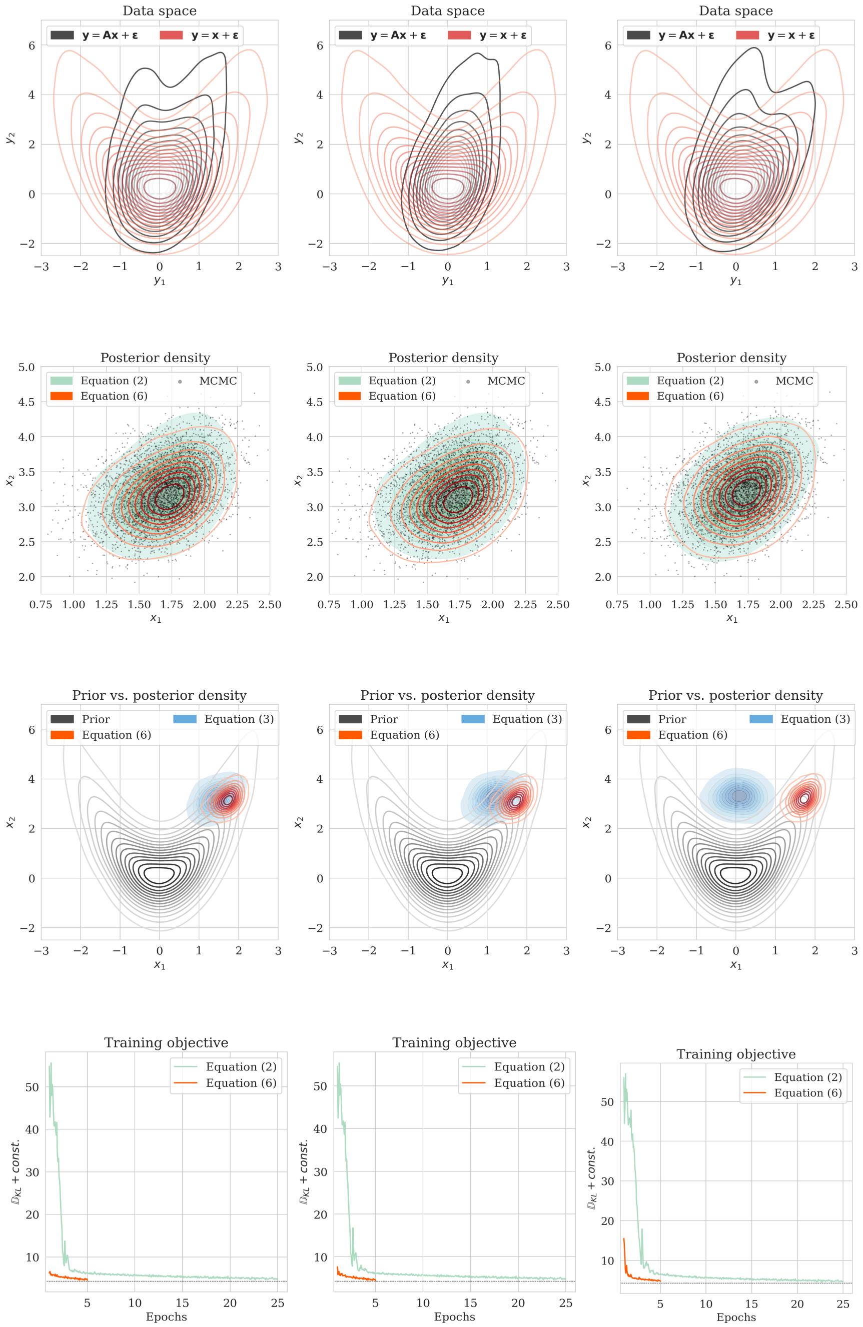


Figure D.1: 2D toy example with decreasing values of $\gamma = 2, 1, 0$ from first to last column, respectively. First row: Low- and high-fidelity data. Second row: approximated posterior densities via MCMC (dark circles), and objectives in equations 4.2 and 4.6. Third row: overlaid prior density with predicted posterior densities via objectives in equations 4.3 and 4.6. Last row: training objective values during training via equations 4.2 and 4.6.

APPENDIX E

SEISMIC COMPRESSED SENSING FROM CHAPTER 4—MORE RESULTS

Here we show more examples to verify the pretraining phase obtained via solving the objective in equation 4.2. Each row in Figure E.1 corresponds to a different testing image. The first column shows the different true seismic images used to create low-fidelity compressive sensing data, depicted in the second column. The third and last columns correspond to the conditional mean and pointwise STD estimates, respectively. Clearly, the pretrained network is able to successfully recover the true image, and consistently indicates more uncertainty in areas with low-amplitude events.

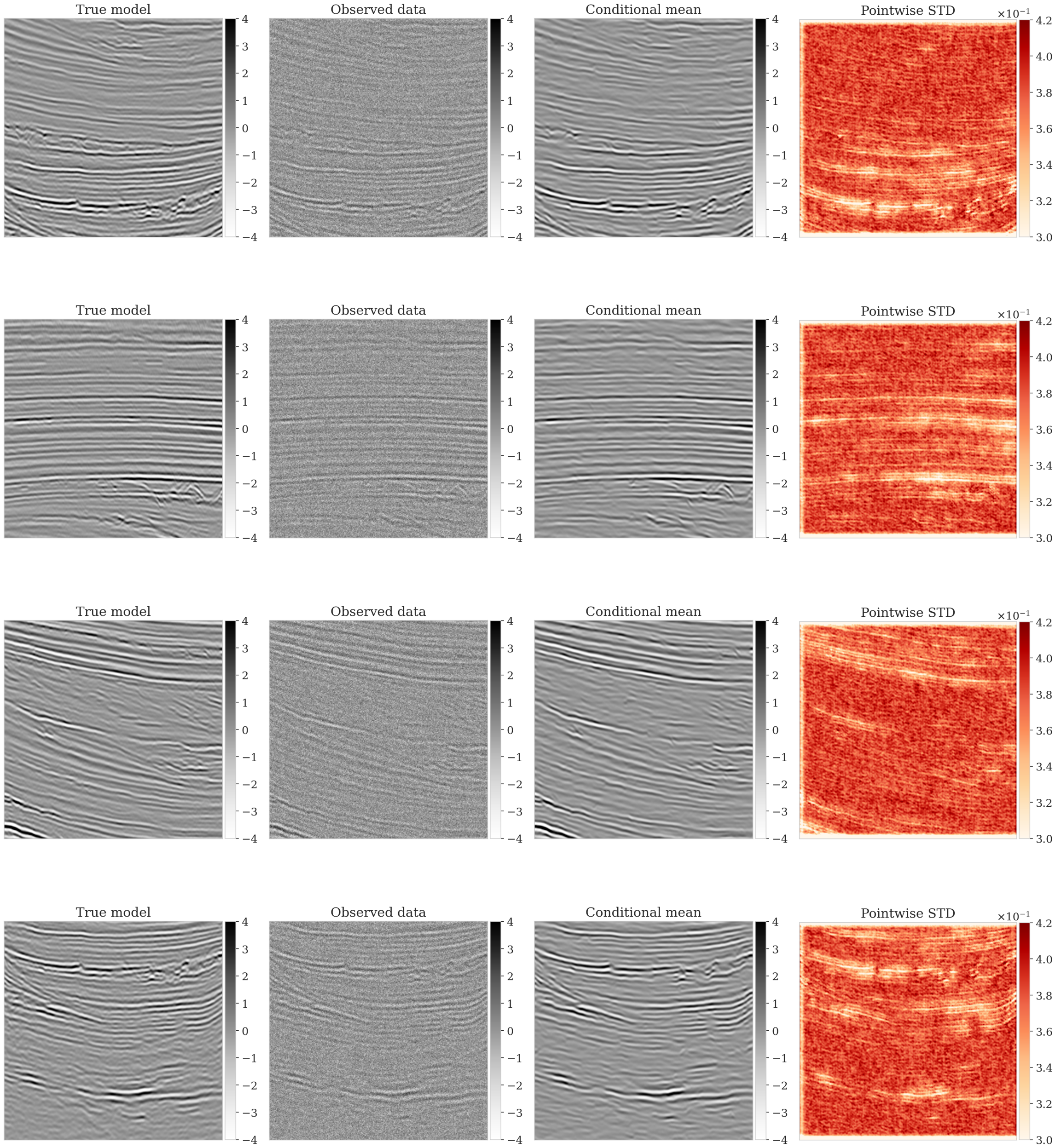


Figure E.1: Seismic compressed sensing for four different low-fidelity images. First column: true seismic images. Second column: low-fidelity observed data. Third and last columns: conditional mean and pointwise STD estimates obtained by drawing 1000 samples from the pretrained conditional NF.

APPENDIX F
PERMISSIONS TO USE COPYRIGHTED MATERIAL

F.1 Chapter 2

The content of chapter 2 was published as a technical article in *Geophysics* under the title "Deep Bayesian inference for seismic imaging with tasks":

- Siahkoohi, Ali, Gabrio Rizzuti, and Felix J. Herrmann. "Deep Bayesian inference for seismic imaging with tasks". In: *Geophysics* 87.5 (June 2022). doi: 10.1190/geo2021-0666.1.
- Copyright © 2022 Society of Exploration Geophysicists.
- The author retains the right to reuse all or part of the work in a thesis or dissertation, as stated in the Copyright Agreement.

Manuscript Number: GEO-2021-0666.R2

GEOPHYSICS Transfer of Copyright Agreement

Agreement must be signed by lead or corresponding author and uploaded to ScholarOne Manuscripts before manuscript receives final acceptance.

Article title: Deep Bayesian inference for seismic imaging with tasks

Names of all authors: Ali Siahkoohi, Gabrio Rizzuti, and Felix J. Herrmann

For and in consideration of the potential publication of the article listed above (the "Work") by the Society of Exploration Geophysicists ("SEG"), I, the undersigned, as lead and/or corresponding author and/or rightsholders and acting on behalf of all authors and/or owners of copyright and other rights in the Work ("Authors"), hereby transfer, assign, and convey all right, title, interest, and worldwide copyright in the Work to SEG, effective if and when the Work is accepted for publication, subject only to limitations expressed in this Transfer of Copyright (the "Agreement"). I warrant that I am authorized and empowered to represent all Authors with respect to the Agreement, which is executed jointly and severally by Authors if copyright transfer is from multiple individuals or entities. The Work includes the article and all material to be published within and with the article in any and all media, including but not limited to tables, figures, graphs, source code, movies, and other multimedia. Responses to discussions of the Work, errata, and other similar material directly related to the Work that may arise subsequent to publication also are considered part of the Work. In the event the Work incorporates copyrighted material of others, Authors warrant that all required permissions or releases have been secured. Authors agree to indemnify, defend, and hold SEG, its directors, officers, employees, and agents harmless against any claims to the contrary.

Authors shall retain the following royalty-free rights:

1. All proprietary rights in the Work that are not transferred to SEG in the Agreement, including the right to any patentable subject matter that may be contained in the Work
2. The right to reproduce and distribute part or all of the Work, including figures, drawings, tables, and abstracts of the Work, with proper attribution and copyright acknowledgment, in connection with Authors' teaching and technical collaborations
3. The right to make oral presentation of the same or similar information as that contained in the Work provided acknowledgment is made of SEG copyright ownership and publication status
4. The right to post a final accepted version of the manuscript or the final SEG-formatted version on Authors' personal Web sites (not including social-network sites used for article sharing such as ResearchGate), employers' Web sites, or in institutional repositories operated and controlled exclusively by Authors' employers provided that (a) the GEOPHYSICS-accepted or GEOPHYSICS-published version is presented without modification unless modification is noted and fully described; (b) SEG copyright notice and a full citation appear with the paper; (c) a live link to the version of record in the SEG Library using a Digital Object Identifier (DOI) permalink is provided; (d) the posting is noncommercial in nature, and the paper is made available to users without charge; (e) notice is provided that use is subject to SEG terms of use and conditions; and (f) a posting of the Work in an institutional repository neither carries nor is implied as carrying a license in conflict with SEG copyright and terms of use
5. The right to prepare and hold copyright in derivative publications based on the Work provided that the derivative work is published subsequent to the official date of the Work's publication by SEG and the Work is cited
6. The right to reuse all or part of the Work in a thesis or dissertation that the Author writes and is required to submit to satisfy criteria of degree-granting institutions, with full citation of the Work, its copyright status, and a DOI permalink to the version of record
7. The nonexclusive right, after publication by GEOPHYSICS, to republish and distribute print versions of the Work or excerpts therefrom without obtaining permission from SEG, provided that (a) the paper is not republished in a journal, book, or collection of conference abstracts or proceedings; and (b) no fee is charged for the printed versions. Permission must be obtained from SEG for other republication of the Work.

SEG may republish the Work or portions thereof in any future SEG publication or compilation in any form and in any language, and SEG retains exclusive right to license third parties to do the same.

This Agreement entitles Authors (or in the case of a Work Made for Hire, the employer) to retain all rights not transferred, assigned, or conveyed to SEG in the Agreement. Authors confirm that the Work has not been published previously elsewhere, nor is it under consideration by any other publisher, that the Work does not infringe any copyright or invade any right of privacy or publicity, and that Authors have the full power to enter into this Agreement and to make the grants contained herein. Authors warrant that they have complied and will comply with Ethical Guidelines for SEG Publications and that they agree to the provisions of these guidelines and all other SEG publications policies applicable to authors. This Agreement shall be binding on Authors' heirs, executors, administrators, and assigns, and shall be construed in accordance with the laws of the State of Oklahoma, United States of America.

IN WITNESS WHEREOF, I have executed this Transfer of Copyright on this 13th day of June, 20 22.

Ali Siahkoohi

Name of author (print or type)

DocuSigned by:
Ali Siahkoohi
A87710CC46614A7...

Signature

Company for which work was performed (if applicable)

Authorized by/Title

SIGN HERE IF U.S. GOVERNMENT EMPLOYED ALL AUTHORS WHEN WORK WAS PREPARED.

I certify that the article named above was prepared solely by a U.S. government employee(s) as part of his/her (their) official duties and therefore legally cannot be copyrighted. Authors agree to all other terms of this agreement.

Name (print or type) Date

Signature

F.2 Chapter 3

The content of chapter 3 was published as a conference proceeding at *SEG Technical Program Expanded Abstracts 2020* under the title “Weak deep priors for seismic imaging”:

- Siahkoohi, Ali, Gabrio Rizzuti, and Felix J. Herrmann. “Weak deep priors for seismic imaging”. In: 90th Annual International Meeting. Society of Exploration Geophysicists. Expanded Abstracts, Sept. 2020, pp. 2998–3002. doi: 10.1190/segam2020-3417568.1.

- Copyright © 2020 Society of Exploration Geophysicists

- As stated at https://library.seg.org/page/policies/open-access, “Authors may reuse all or part of their papers published with SEG in a thesis or dissertation that authors write and are required to submit to satisfy criteria of degree-granting institutions.”

F.3 Chapter 4

The content of chapter 4 was presented at the 3rd Symposium on Advances in Approximate Bayesian Inference under the title “Preconditioned training of normalizing flows for variational inference in inverse problems”:

- Siahkoohi, Ali, Gabrio Rizzuti, Mathias Louboutin, Philipp Witte, and Felix J. Herrmann. “Preconditioned training of normalizing flows for variational inference in inverse problems”. In: 3rd Symposium on Advances in Approximate Bayesian Inference. Jan. 2021.

- Openly accessible at https://openreview.net/pdf?id=P9m1sMaNQ8T.

- No transfer of copyright has been signed and the copyright remains with the author.

F.4 Chapter 5

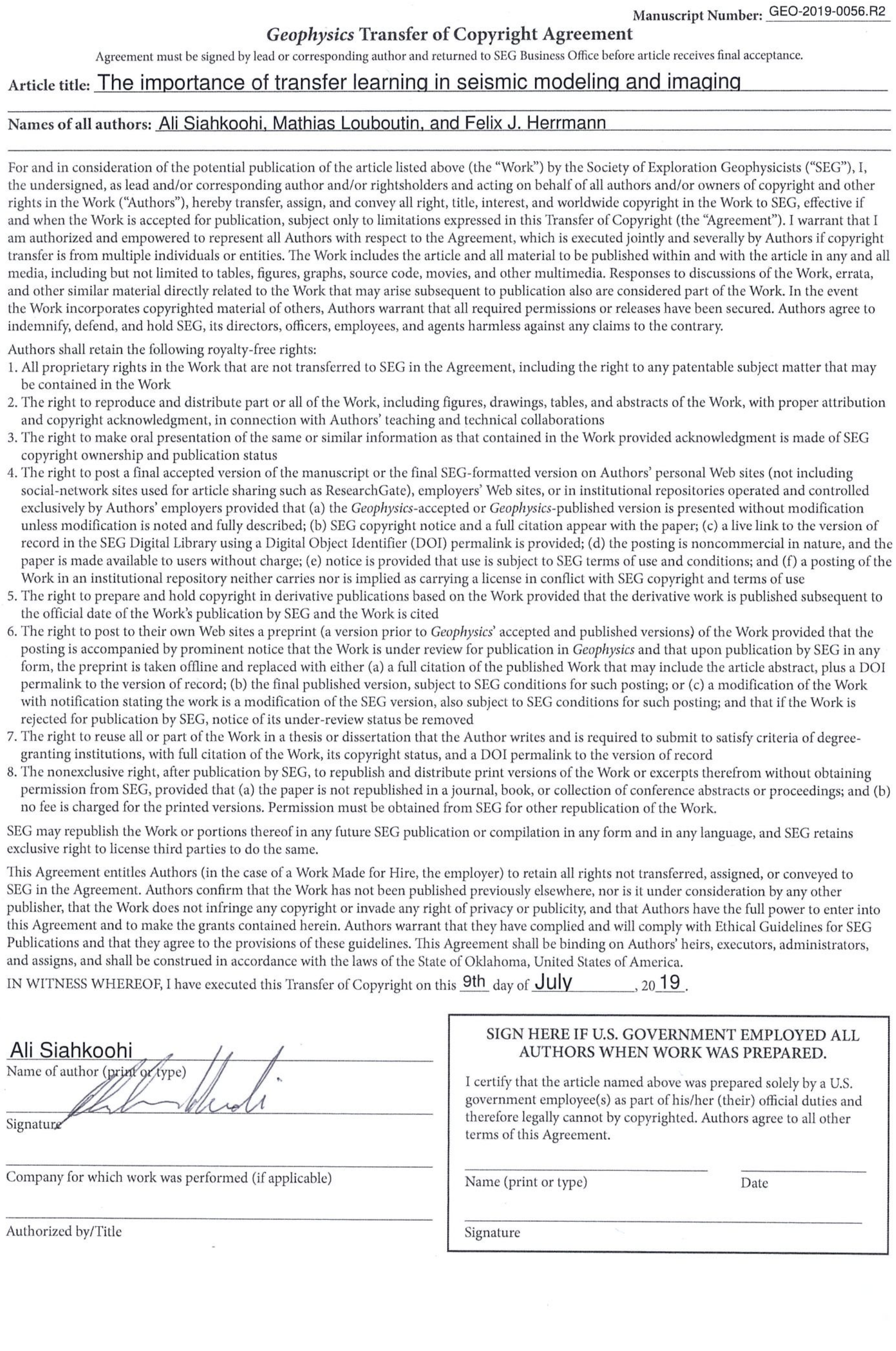
The content of chapter 5 will be submitted in a modified form to *Geophysics* in July 2022 under the title "Reliable amortized variational inference with physics-based latent distribution correction":

- Siahkoohi, Ali, Gabrio Rizzuti, Rafael Orozco, and Felix J. Herrmann. "Reliable amortized variational inference via physics-based latent distribution correction". To be submitted to Geophysics (July 2022).
- No transfer of copyright has been signed and the copyright remains with the author.

F.5 Chapter 6

The content of chapter 6 was published as a technical article in *Geophysics* under the title "The importance of transfer learning in seismic modeling and imaging":

- Siahkoohi, Ali, Mathias Louboutin, and Felix J. Herrmann. "The importance of transfer learning in seismic modeling and imaging". In: Geophysics 84.6 (Nov. 2019), A47–A52. doi: 10.1190/geo2019-0056.1.
- Copyright © 2019 Society of Exploration Geophysicists.
- The author retains the right to reuse all or part of the work in a thesis or dissertation, as stated in the Copyright Agreement.



Manuscript Number: GEO-2019-0056.R2

Geophysics Transfer of Copyright Agreement

Agreement must be signed by lead or corresponding author and returned to SEG Business Office before article receives final acceptance.

Article title: The importance of transfer learning in seismic modeling and imaging

Names of all authors: Ali Siahkoohi, Mathias Louboutin, and Felix J. Herrmann

For and in consideration of the potential publication of the article listed above (the "Work") by the Society of Exploration Geophysicists ("SEG"), I, the undersigned, as lead and/or corresponding author and/or rightsholders and acting on behalf of all authors and/or owners of copyright and other rights in the Work ("Authors"), hereby transfer, assign, and convey all right, title, interest, and worldwide copyright in the Work to SEG, effective if and when the Work is accepted for publication, subject only to limitations expressed in this Transfer of Copyright (the "Agreement"). I warrant that I am authorized and empowered to represent all Authors with respect to the Agreement, which is executed jointly and severally by Authors if copyright transfer is from multiple individuals or entities. The Work includes the article and all material to be published within and with the article in any and all media, including but not limited to tables, figures, graphs, source code, movies, and other multimedia. Responses to discussions of the Work, errata, and other similar material directly related to the Work that may arise subsequent to publication also are considered part of the Work. In the event the Work incorporates copyrighted material of others, Authors warrant that all required permissions or releases have been secured. Authors agree to indemnify, defend, and hold SEG, its directors, officers, employees, and agents harmless against any claims to the contrary.

Authors shall retain the following royalty-free rights:

1. All proprietary rights in the Work that are not transferred to SEG in the Agreement, including the right to any patentable subject matter that may be contained in the Work
2. The right to reproduce and distribute part or all of the Work, including figures, drawings, tables, and abstracts of the Work, with proper attribution and copyright acknowledgment, in connection with Authors' teaching and technical collaborations
3. The right to make oral presentation of the same or similar information as that contained in the Work provided acknowledgment is made of SEG copyright ownership and publication status
4. The right to post a final accepted version of the manuscript or the final SEG-formatted version on Authors' personal Web sites (not including social-network sites used for article sharing such as ResearchGate), employers' Web sites, or in institutional repositories operated and controlled exclusively by Authors' employers provided that (a) the *Geophysics*-accepted or *Geophysics*-published version is presented without modification unless modification is noted and fully described; (b) SEG copyright notice and a full citation appear with the paper; (c) a live link to the version of record in the SEG Digital Library using a Digital Object Identifier (DOI) permalink is provided; (d) the posting is noncommercial in nature, and the paper is made available to users without charge; (e) notice is provided that use is subject to SEG terms of use and conditions; and (f) a posting of the Work in an institutional repository neither carries nor is implied as carrying a license in conflict with SEG copyright and terms of use
5. The right to prepare and hold copyright in derivative publications based on the Work provided that the derivative work is published subsequent to the official date of the Work's publication by SEG and the Work is cited
6. The right to post to their own Web sites a preprint (a version prior to *Geophysics*' accepted and published versions) of the Work provided that the posting is accompanied by prominent notice that the Work is under review for publication in *Geophysics* and that upon publication by SEG in any form, the preprint is taken offline and replaced with either (a) a full citation of the published Work that may include the article abstract, plus a DOI permalink to the version of record; (b) the final published version, subject to SEG conditions for such posting; or (c) a modification of the Work with notification stating the work is a modification of the SEG version, also subject to SEG conditions for such posting; and that if the Work is rejected for publication by SEG, notice of its under-review status be removed
7. The right to reuse all or part of the Work in a thesis or dissertation that the Author writes and is required to submit to satisfy criteria of degree-granting institutions, with full citation of the Work, its copyright status, and a DOI permalink to the version of record
8. The nonexclusive right, after publication by SEG, to republish and distribute print versions of the Work or excerpts therefrom without obtaining permission from SEG, provided that (a) the paper is not republished in a journal, book, or collection of conference abstracts or proceedings; and (b) no fee is charged for the printed versions. Permission must be obtained from SEG for other republication of the Work.

SEG may republish the Work or portions thereof in any future SEG publication or compilation in any form and in any language, and SEG retains exclusive right to license third parties to do the same.

This Agreement entitles Authors (in the case of a Work Made for Hire, the employer) to retain all rights not transferred, assigned, or conveyed to SEG in the Agreement. Authors confirm that the Work has not been published previously elsewhere, nor is it under consideration by any other publisher, that the Work does not infringe any copyright or invade any right of privacy or publicity, and that Authors have the full power to enter into this Agreement and to make the grants contained herein. Authors warrant that they have complied and will comply with Ethical Guidelines for SEG Publications and that they agree to the provisions of these guidelines. This Agreement shall be binding on Authors' heirs, executors, administrators, and assigns, and shall be construed in accordance with the laws of the State of Oklahoma, United States of America.

IN WITNESS WHEREOF, I have executed this Transfer of Copyright on this 9th day of July, 2019.

Ali Siahkoohi
Name of author (print or type)

Signature

Company for which work was performed (if applicable)

Authorized by/Title

SIGN HERE IF U.S. GOVERNMENT EMPLOYED ALL AUTHORS WHEN WORK WAS PREPARED.

I certify that the article named above was prepared solely by a U.S. government employee(s) as part of his/her (their) official duties and therefore legally cannot by copyrighted. Authors agree to all other terms of this Agreement.

Name (print or type)　　Date

Signature

F.6 Chapter 7

The content of chapter 7 was published as a conference proceeding at *SEG Technical Program Expanded Abstracts 2022* under the title “Velocity continuation with Fourier neural operators for accelerated uncertainty quantification”:

- Siahkoohi, Ali, Mathias Louboutin, and Felix J. Herrmann. “Velocity continuation with Fourier neural operators for accelerated uncertainty quantification”. In: 2nd International Meeting for Applied Geoscience & Energy. Society of Exploration Geophysicists. June 2022.
- Copyright © 2022 Society of Exploration Geophysicists
- As stated at https://library.seg.org/page/policies/open-access, “Authors may reuse all or part of their papers published with SEG in a thesis or dissertation that authors write and are required to submit to satisfy criteria of degree-granting institutions.”