Deep Bayesian inference for seismic imaging with tasks

Ali Siahkoohi
School of Computational Science and Engineering,
Georgia Institute of Technology
`alisk@gatech.edu`

Gabrio Rizzuti
Georgia Institute of Technology
Now at Utrecht University
`g.rizzuti@uu.nl`

Felix J. Herrmann
School of Computational Science and Engineering,
Georgia Institute of Technology
`felix.herrmann@gatech.edu`

Abstract

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.

1 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 [1–3]. 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 [4–12], 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 [13–15], or satisfying regularity conditions related to model parameters or derivatives thereof [4, 6, 7, 9–12, 16–22]. While effective in controlled settings, handcrafted priors might introduce unwanted bias to the solution. Recent deep-learning based approaches [23–34], 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, e.g., medical imaging [35–40], 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 [41–50].

In this work, we take advantage of a novel prior recently deployed in computer vision and geophysics [51–63], 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 papers 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., 71–74] because horizon uncertainty is ultimately driven by data (through the intermediate imaging distribution), and not from label (control point) uncertainty alone.

---

[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.

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 Theory

The goal of this paper 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.

2.1 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

$$\mathbf{d}_i = \mathbf{J}(\mathbf{m}_0, \mathbf{q}_i)\delta \mathbf{m}^* + \boldsymbol{\epsilon}_i, \quad \boldsymbol{\epsilon}_i \sim \mathrm{N}(\mathbf{0}, \sigma^2 \mathbf{I}), \tag{1}$$

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

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

While this approach is in principle capable of producing high-fidelity true-amplitude images [75–77], 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 requires regularization but also calls for a statistical inference framework that allows us to draw conclusions in the presence of uncertainty.

2.2 Probabilistic imaging with Bayesian inference

To account for uncertainties in the image induced by the random noise term $\boldsymbol{\epsilon}_i$ in equation 1, we follow the seminal work of Tarantola [5] and cast our noisy imaging as a Bayesian inverse problem. Instead of calculating a single image by solving equation 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

$$p_{\text{post}}\left(\delta \mathbf{m} \mid \mathbf{d}\right) \propto p_{\text{like}}(\mathbf{d} \mid \delta \mathbf{m})\, p_{\text{prior}}(\delta \mathbf{m}). \tag{3}$$

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 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 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:

$$\begin{aligned}
-\log p_{\text{like}}\left(\mathbf{d} \mid \delta\mathbf{m}\right) &= -\sum_{i=1}^{n_s} \log p_{\text{like}}\left(\mathbf{d}_i \mid \delta\mathbf{m}\right) \\
&= \frac{1}{2\sigma^2} \sum_{i=1}^{n_s} \left\|\mathbf{d}_i - \mathbf{J}(\mathbf{m}_0, \mathbf{q}_i)\delta\mathbf{m}\right\|_2^2 + \text{const},
\end{aligned} \tag{4}$$

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 [16–19]. 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 [23, 25, 37, 48, 79]. 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 Deep priors

Following recent work by Lempitsky et al. [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:

$$\begin{aligned}
-\log p_{\text{like}}\left(\mathbf{d} \mid \mathbf{w}\right) &= -\sum_{i=1}^{n_s} \log p_{\text{like}}\left(\mathbf{d}_i \mid \mathbf{w}\right) \\
&= \frac{1}{2\sigma^2} \sum_{i=1}^{n_s} \left\|\mathbf{d}_i - \mathbf{J}(\mathbf{m}_0, \mathbf{q}_i) g(\mathbf{z}, \mathbf{w})\right\|_2^2 + \text{const},
\end{aligned} \tag{5}$$

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 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

$$p_{\text{post}}(\mathbf{w} \mid \mathbf{d}) \propto \left[\prod_{i=1}^{n_s} p_{\text{like}}(\mathbf{d}_i \mid \mathbf{w})\right] \mathrm{N}\big(\mathbf{w} \mid \mathbf{0}, \lambda^{-2}\mathbf{I}\big) \tag{6}$$

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 5, the negative log-posterior distribution becomes

$$\begin{aligned} -\log p_{\text{post}}(\mathbf{w} \mid \mathbf{d}) &= -\left[\sum_{i=1}^{n_s} \log p_{\text{like}}(\mathbf{d}_i \mid \mathbf{w})\right] - \log \mathrm{N}\big(\mathbf{w} \mid \mathbf{0}, \lambda^{-2}\mathbf{I}\big) + \text{const} \\ &= \frac{1}{2\sigma^2} \sum_{i=1}^{n_s} \big\|\mathbf{d}_i - \mathbf{J}(\mathbf{m}_0, \mathbf{q}_i) g(\mathbf{z}, \mathbf{w})\big\|_2^2 + \frac{\lambda^2}{2} \big\|\mathbf{w}\big\|_2^2 + \text{const.} \end{aligned} \tag{7}$$

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. Cheng et al. [53] provided a theoretical Bayesian perspective on deep priors, describing them as Gaussian process priors in classical Bayesian terms. Specifically, Cheng et al. [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. Cheng et al. [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, Dittmer et al. [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 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.

2.4 Estimation with Bayesian inference

Based on the expressions for the negative log-likelihood (equation 5) and posterior (equation 7), we derive expressions for different point and interval estimates [52].

2.4.1 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

$$\begin{aligned} \delta\mathbf{m}_{\text{MLE}} &= \underset{\delta\mathbf{m}}{\arg\min} -\log p_{\text{like}}(\mathbf{d} \mid \delta\mathbf{m}) \\ &= \underset{\delta\mathbf{m}}{\arg\min} \frac{1}{2\sigma^2} \sum_{i=1}^{n_s} \big\|\mathbf{d}_i - \mathbf{J}(\mathbf{m}_0, \mathbf{q}_i)\delta\mathbf{m}\big\|_2^2, \end{aligned} \tag{8}$$

where the last equality follows from equation 4. Observe that the MLE corresponds to the deterministic least-squares solution, yielded by equation 2. Unfortunately, MLE images are prone to overfitting [81, 82] that results in imaging artifacts [3].

2.4.2 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

$$\begin{aligned}\mathbf{w}_{\text{MAP}} &= \underset{\mathbf{w}}{\arg\max}\, p_{\text{post}}\left(\mathbf{w} \mid \mathbf{d}\right) \\ &= \underset{\mathbf{w}}{\arg\min}\, \frac{1}{2\sigma^2} \sum_{i=1}^{n_s} \left\| \mathbf{d}_i - \mathbf{J}(\mathbf{m}_0, \mathbf{q}_i) g(\mathbf{z}, \mathbf{w}) \right\|_2^2 + \frac{\lambda^2}{2} \left\| \mathbf{w} \right\|_2^2.\end{aligned} \tag{9}$$

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

$$\delta \mathbf{m}_{\text{MAP}} = g(\mathbf{z}, \mathbf{w}_{\text{MAP}}). \tag{10}$$

When compared with MAP estimates computed from traditional Bayesian formulations of linear inverse problems, the estimate in equation 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 [84–86] are able to minimize the objective function in equation 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 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 Siahkoohi et al. [62] for an alternative formulation, which reduces the number of iterations and therefore the number of evaluations of the computationally expensive forward modeling operator.

2.4.3 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 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

$$g(\mathbf{z}, \mathbf{w}) \sim p_{\text{post}}\left(\delta \mathbf{m} \mid \mathbf{d}\right). \tag{11}$$

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, $\{\mathbf{w}_j\}_{j=1}^{n_{\text{w}}} \sim p_{\text{post}}(\mathbf{w} \mid \mathbf{d})$, via

$$\begin{aligned}\delta \mathbf{m}_{\text{CM}} &= \mathbb{E}_{\delta \mathbf{m} \sim p_{\text{post}}(\delta \mathbf{m} \mid \mathbf{d})} \left[\delta \mathbf{m} \right] \\ &= \mathbb{E}_{\mathbf{w} \sim p_{\text{post}}(\delta \mathbf{w} \mid \mathbf{d})} \left[g(\mathbf{z}, \mathbf{w}) \right] \\ &= \int p_{\text{post}}(\mathbf{w} \mid \mathbf{d}) g(\mathbf{z}, \mathbf{w}) \mathrm{d}\mathbf{w} \\ &\approx \frac{1}{n_{\text{w}}} \sum_{j=1}^{n_{\text{w}}} g(\mathbf{z}, \mathbf{w}_j).\end{aligned} \tag{12}$$

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.

2.4.4 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

$$\begin{aligned}\boldsymbol{\sigma}_{\text{post}}^2 &= \mathbb{E}_{\delta\mathbf{m}\sim p_{\text{post}}(\delta\mathbf{m}\mid\mathbf{d})}\big[(\delta\mathbf{m}-\delta\mathbf{m}_{\text{CM}})\odot(\delta\mathbf{m}-\delta\mathbf{m}_{\text{CM}})\big] \\ &\approx \frac{1}{n_{\text{w}}}\sum_{j=1}^{n_{\text{w}}}\big(g(\mathbf{z},\mathbf{w}_j)-\delta\mathbf{m}_{\text{CM}}\big)\odot\big(g(\mathbf{z},\mathbf{w}_j)-\delta\mathbf{m}_{\text{CM}}\big).\end{aligned} \tag{13}$$

In this expression, $\boldsymbol{\sigma}_{\text{post}}$ is the estimated pointwise standard deviation and $\odot$ represents elementwise multiplication. Again, the expectations approximated in equations 12 and 13 require samples from the posterior distribution, $p_{\text{post}}\left(\delta\mathbf{m}\mid\mathbf{d}\right)$.

2.4.5 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].

3 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.

3.1 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, $\{\mathbf{w}_j\}_{j=1}^{n_{\text{w}}}\sim p_{\text{post}}(\mathbf{w}\mid\mathbf{d})$, these expectations with respect to arbitrary functions can be approximated by

$$\mathbb{E}_{\mathbf{w}\sim p_{\text{post}}(\delta\mathbf{w}\mid\mathbf{d})}\big[f(\mathbf{w})\big]\approx\frac{1}{n_w}\sum_{j=1}^{n_w}f(\mathbf{w}_j). \tag{14}$$

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

3.2 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 [7, 67]. 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 [9, 11, 12, 60, 61, 94, 95].

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 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 [96–98] 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, 84, 85, 99–103].

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 paper, 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, 105–108]—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 Welling and Teh [67], SGLD iterations for the negative log-posterior involve at iteration k the following update for the network weights of the deep prior:

$$
\begin{aligned}
&\mathbf{w}_{k+1} = \mathbf{w}_k - \frac{\alpha_k}{2}\mathbf{M}_k \nabla_{\mathbf{w}} \left(\frac{n_s}{2\sigma^2} \left\| \mathbf{d}_i - \mathbf{J}(\mathbf{m}_0, \mathbf{q}_i) g(\mathbf{z}, \mathbf{w}_k) \right\|_2^2 + \frac{\lambda^2}{2} \left\| \mathbf{w}_k \right\|_2^2 \right) + \boldsymbol{\eta}_k, \\
&\boldsymbol{\eta}_k \sim \mathrm{N}(\mathbf{0}, \alpha_k \mathbf{M}_k),
\end{aligned}
\tag{15}
$$

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 1), 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 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 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.

3.2.1 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 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 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 15 to

Langevin dynamics [104]. As a result, similar to SGD, convergence can only be guaranteed when the stepsize in equation 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 Welling and Teh [67] who propose the following sequence of stepsizes:

$$\alpha_k = a(b+k)^{-\gamma}, \tag{16}$$

where $\gamma = \frac{1}{3}$ is the decay rate chosen according to Teh et al. [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.

3.2.2 Preconditioning

In addition to selecting proper stepsizes, the converge of the iterations in equation 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, Li et al. [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 15. To define this preconditioning matrix, let $\delta\mathbf{w}$ denote the gradient of the negative log-posterior density (equation 7) at the current estimate of weights $\mathbf{w}_k$, i.e.,

$$\delta\mathbf{w} = \nabla_{\mathbf{w}} \left(\frac{n_s}{2\sigma^2} \left\| \mathbf{d}_i - \mathbf{J}(\mathbf{m}_0, \mathbf{q}_i) g(\mathbf{z}, \mathbf{w}_k) \right\|_2^2 + \frac{\lambda^2}{2} \left\| \mathbf{w}_k \right\|_2^2 \right). \tag{17}$$

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

$$\mathbf{v}_{k+1} = \beta \mathbf{v}_k + (1-\beta)\, \delta\mathbf{w} \odot \delta\mathbf{w} \tag{18}$$

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,

$$\mathbf{M}_k = \text{diag}\left(1 \oslash \sqrt{\mathbf{v}_{k+1}}\right) \tag{19}$$

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 19. By introducing the preconditioning matrix $\mathbf{M}_k$ all weights are updated similarly, which allows us to increase the stepsize. Following Li et al. [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.

3.3 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 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 9 and 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 18 and 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.

3.4 The SGLD Algorithm

The different steps of generating $n_w = K/2$ samples from the posterior from K iterations of SGLD (equation 15) are summarized in Algorithm 1 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 1 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 Teh et al. [110], the decay rate parameter in equation 16 is set to $\gamma = \frac{1}{3}$. The stepsize constants $a,\ b$ in equation 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 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 15) in total, and we adopt the general practice of discarding the first half of the obtained samples [111].

Given the above inputs, Algorithm 1 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 [67, 111].

4 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.

4.1 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 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 1a). Due to the nature of the migration algorithm used to obtain the

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 16
  $K$   // maximum MCMC steps
Initialization:
  randomly initialize CNN parameters, $\mathbf{w}_0 \in \mathbb{R}^M$   Glorot and Bengio [80]
  initialize vector, $\mathbf{v}_0 \in \mathbb{R}^M$ with zero
for $k = 0$ to $K - 1$ do
  randomly draw $i \subset \{1, \ldots, n_s\}$   // 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 17
  $\mathbf{v}_{k+1} = \beta \mathbf{v}_k + (1 - \beta)\, \delta\mathbf{w} \odot \delta\mathbf{w}$   // equation 18
  $\mathbf{M}_k = \operatorname{diag}\left(1 \oslash \sqrt{\mathbf{v}_{k+1}}\right)$   // equation 19
  $\alpha_k = a(b + k)^{-\gamma}$   // equation 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 15
end for
Output: $\{\mathbf{w}_k\}_{k=K/2+1}^{K}$   // samples from the posterior $p_{\text{post}}(\mathbf{w} \mid \mathbf{d})$
```

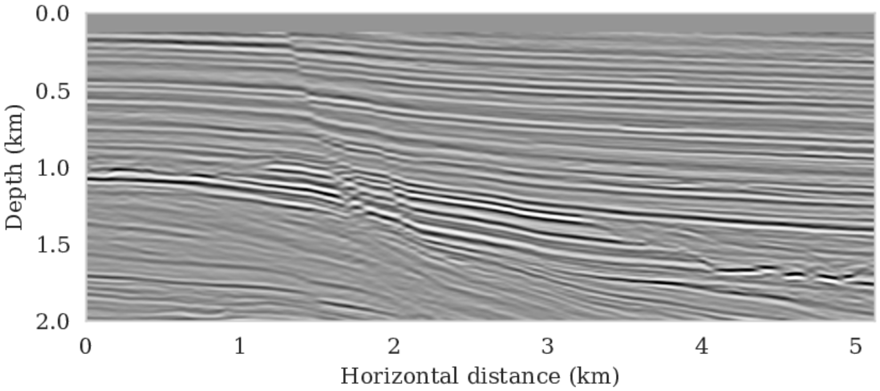
Parihaka dataset, the amplitudes in the extracted 2D subset are not necessarily consistent with the seismic imaging forward model presented in this paper (equation 1). For this reason, we normalized the amplitudes of the extracted seismic image. Given these perturbations, shot data is generated with the linearized Born scattering operator for a made up, but realistic, smoothly varying background model $\mathbf{m}_0$ for the squared slowness (Figure 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 2 shows an example of a single noise-free (Figure 2a) and noisy (Figure 2b) shot record.

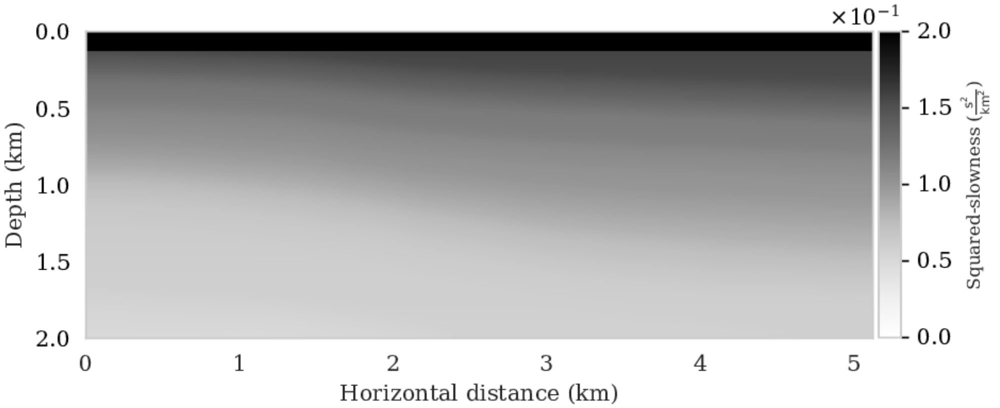
Even though our example is in 2D, the number of parameters (the weights of the deep prior network) is large (approximately 40 times larger than image dimension), which results in many SGLD iterations. In a setting where we are content with approximate Bayesian inference, i.e., where the validity of the 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 16 with $a,\ b$ chosen such the stepsize decreases from 10^{-2} to 5×10^{-3}.

4.2 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 8) whereas the former involves maximizing the posterior distribution (equation 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



(a)



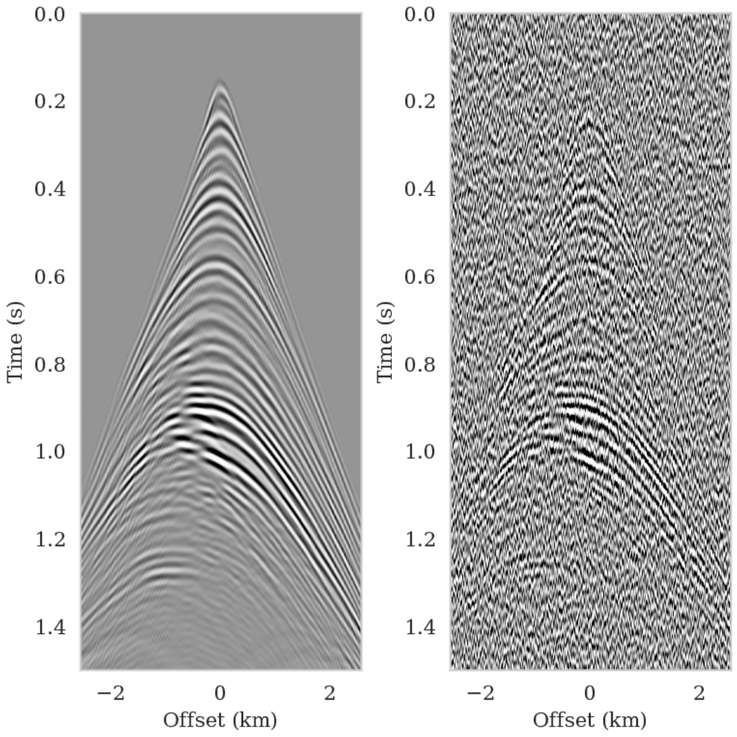
(b)

Figure 1: Problem setup. (a) A 2D subset of the Parihaka dataset, considered as true model. (b) Made up smooth squared-slowness background model.

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 3a and 3b, respectively. We obtained these results with the RMSprop optimization algorithm [84], which uses the same preconditioning scheme (equations 18 and 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 18 and 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) $8.79\,\text{dB}$, the MLE estimate (SNR $8.25\,\text{dB}$) 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.

4.3 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.



(a) (b)

Figure 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.

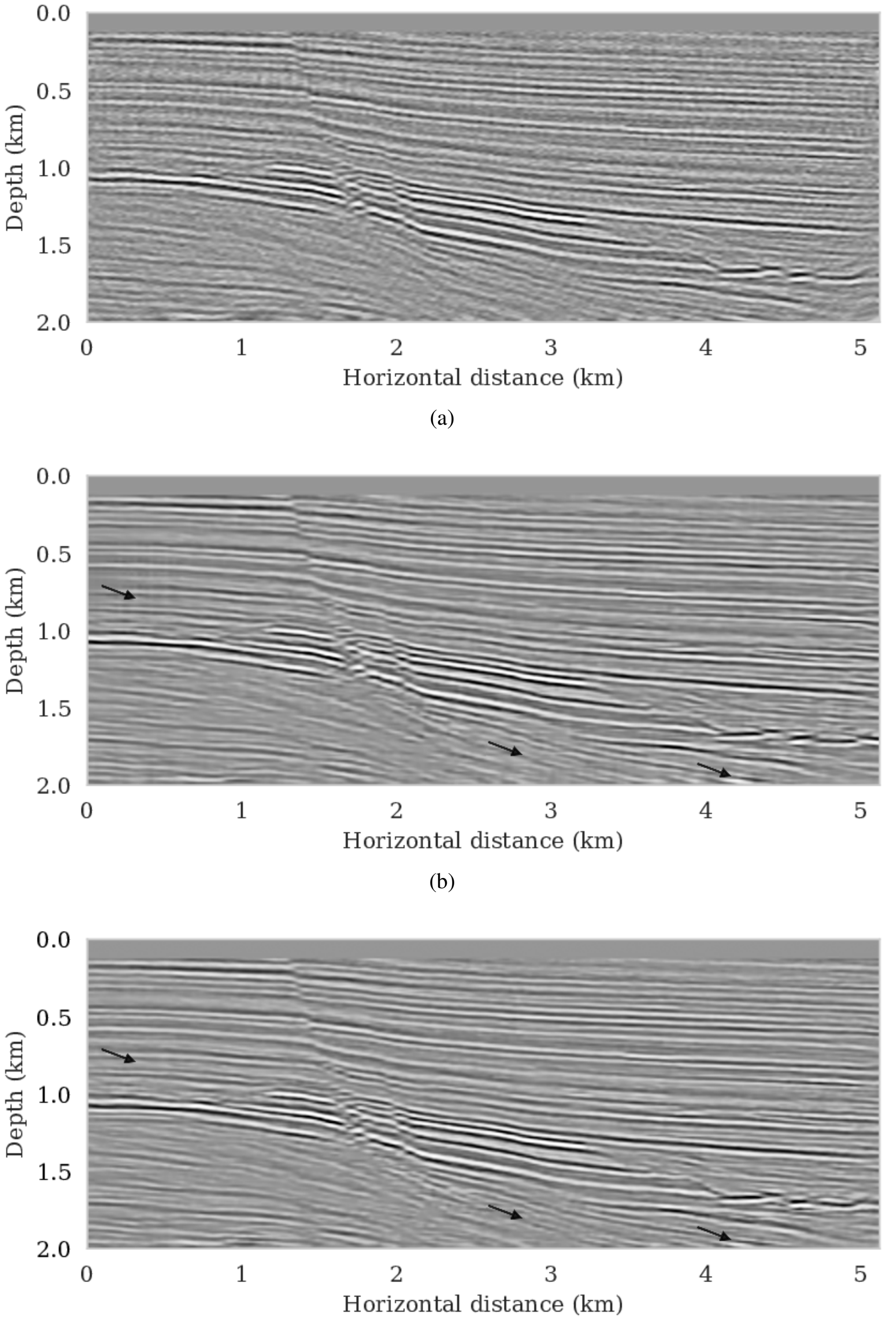
4.3.1 Conditional mean

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 Li et al. [68] to the RMSprop optimization algorithm, the posterior distribution can be sampled with Algorithm 1 after a warmup phase of about 25 data passes. The resulting samples for the weights are then used, after push forward (see equation 11), to approximate the conditional mean, $\delta\mathbf{m}_{\text{CM}}$, by computing the sum in equation 12. Compared to the MAP estimate (Figures 3b and 3c), the $\delta\mathbf{m}_{\text{CM}}$ (SNR $9.66\,\text{dB}$) 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.

4.3.2 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 13) in Figure 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 4b. From these plots in Figure 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 to proximity to the sources and receivers.

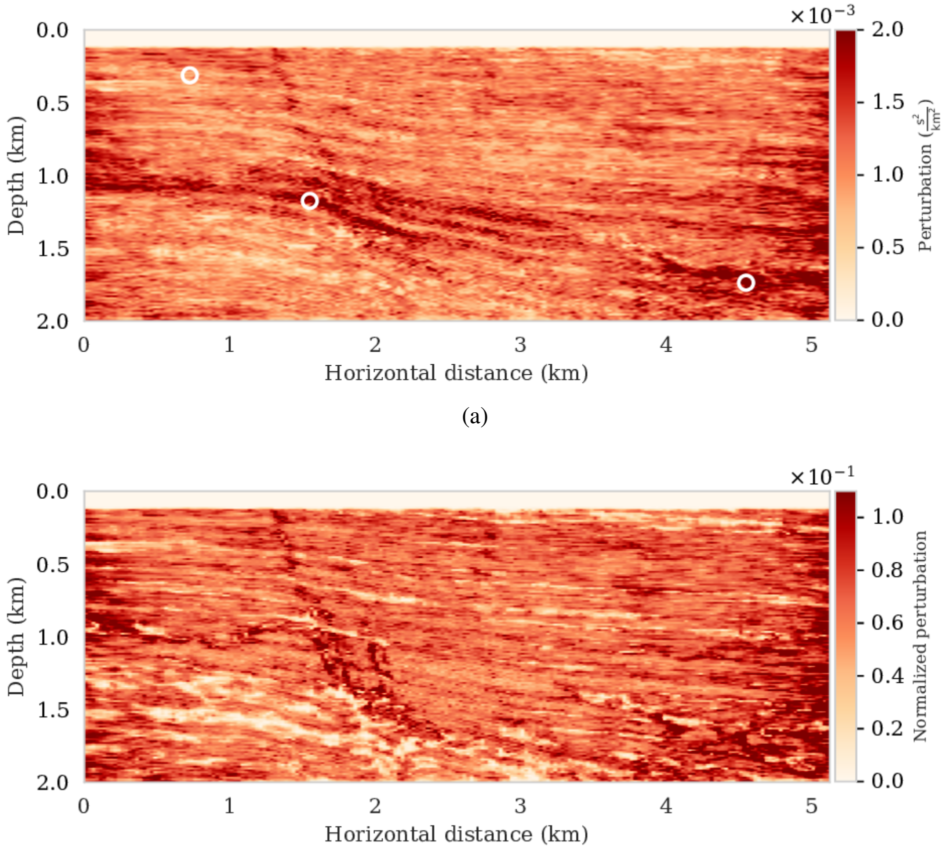


(a)

(b)

(c)

Figure 3: Imaging with deep priors of a 2D subset of the Parihaka dataset (a) MLE, i.e., minimizer of equation 4 with respect to $\delta\mathbf{m}$, with SNR $8.25\,\text{dB}$. (b) The MAP estimate, i.e., minimizer of equation 7, followed by a mapping onto the image space via g (equation 10), with SNR $8.79\,\text{dB}$. (c) The conditional (posterior) mean estimate, $\delta\mathbf{m}_{\text{CM}}$, with SNR $9.66\,\text{dB}$. All figures are displayed with the same color clipping values.



(a)

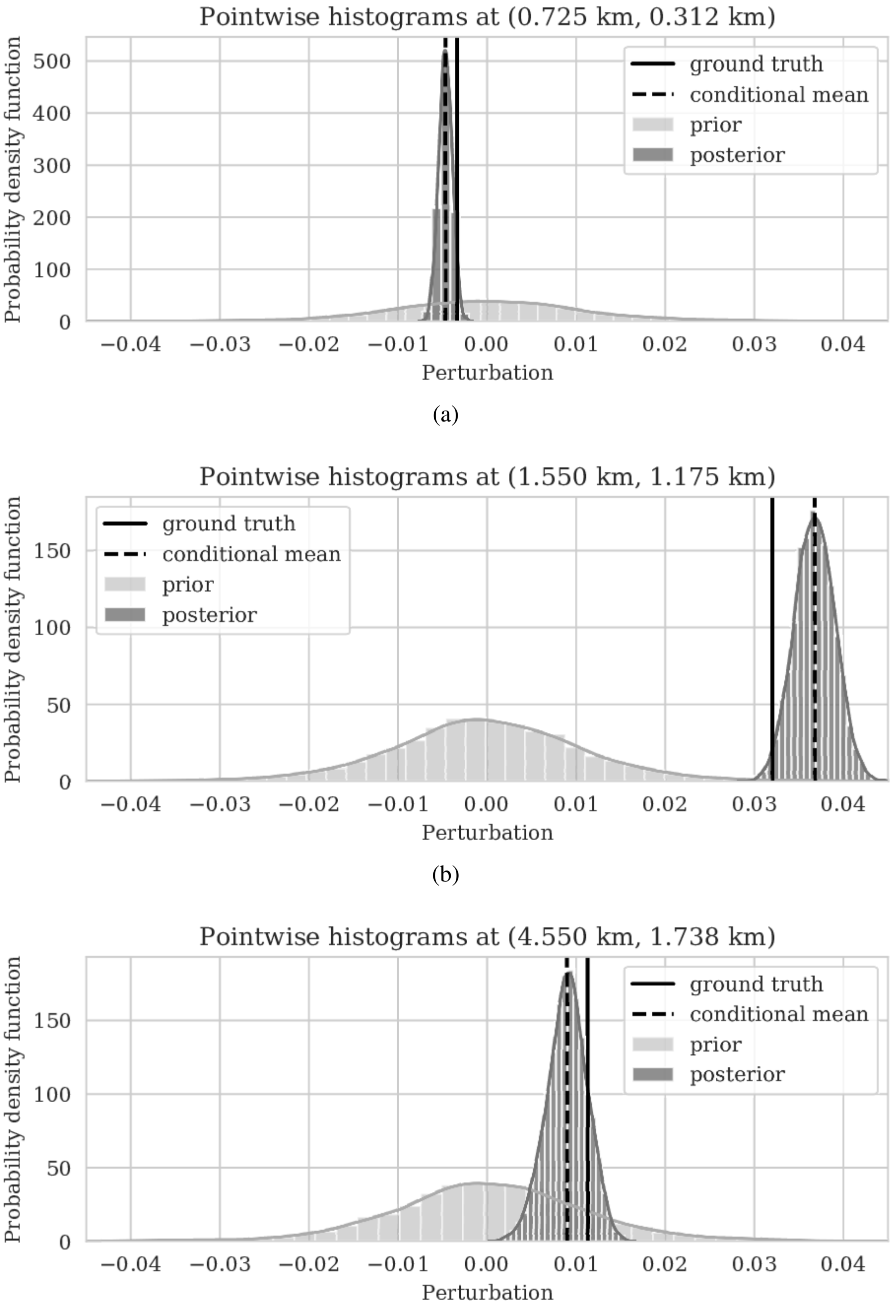
(b)

Figure 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 3c).

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 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 5. Similarly, histograms for the posterior are computed (equation 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.

4.4 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].



(a)

(b)

(c)

Figure 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) $(4.550\,\mathrm{km},\ 1.738\,\mathrm{km})$.

4.4.1 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 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 6b and 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.

4.4.2 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 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.

5 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.

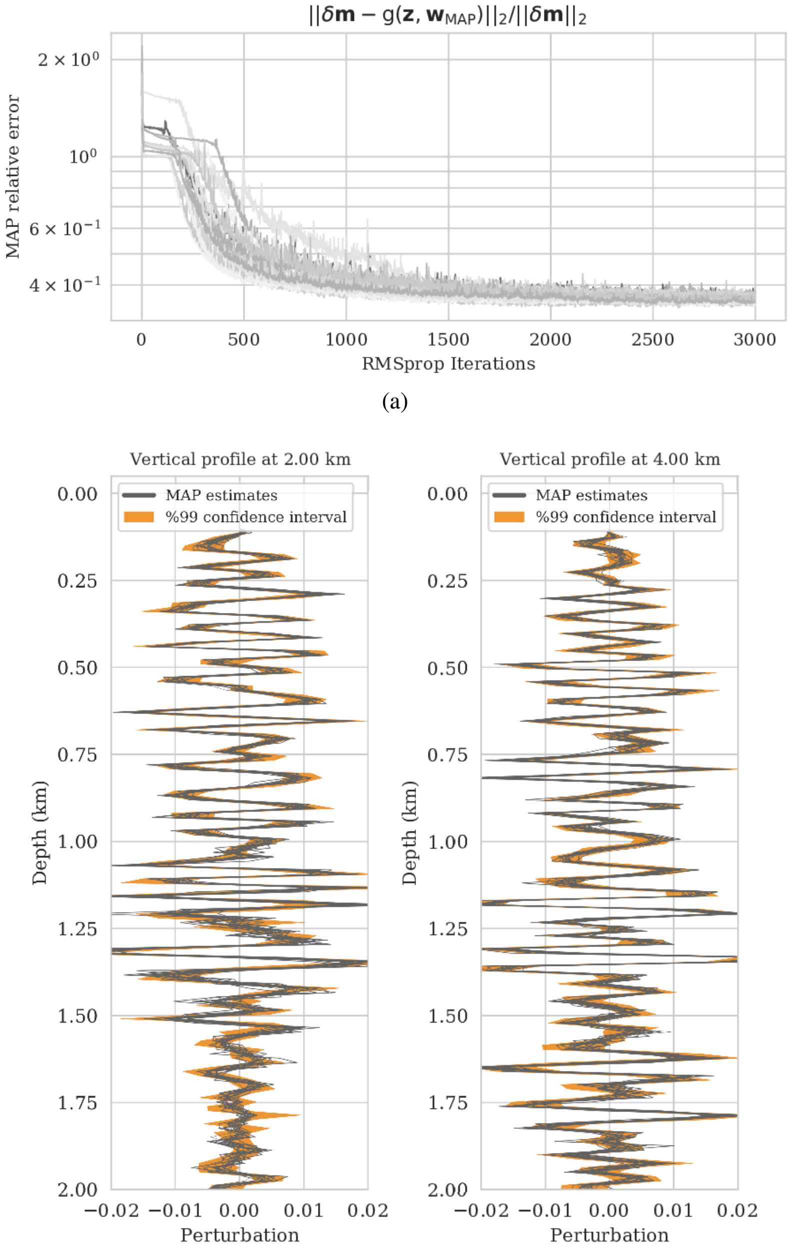
To set the stage to put tasks conducted on seismic images on a firm statistical footing, we will make the assumption that these tasks are only informed by the estimated image and not by the shot data explicitly. This means when given an estimated image the task, e.g., of horizon tracking, is assumed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

$$(\mathbf{h} \perp\!\!\!\perp \mathbf{d}) \mid \delta\mathbf{m}, \tag{20}$$

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 paper, we denote the task on the image by $\mathcal{H}$, which for horizon tracking implies $\mathbf{h} = \mathcal{H}(\delta\mathbf{m})$.

5.1 Bayesian formulation

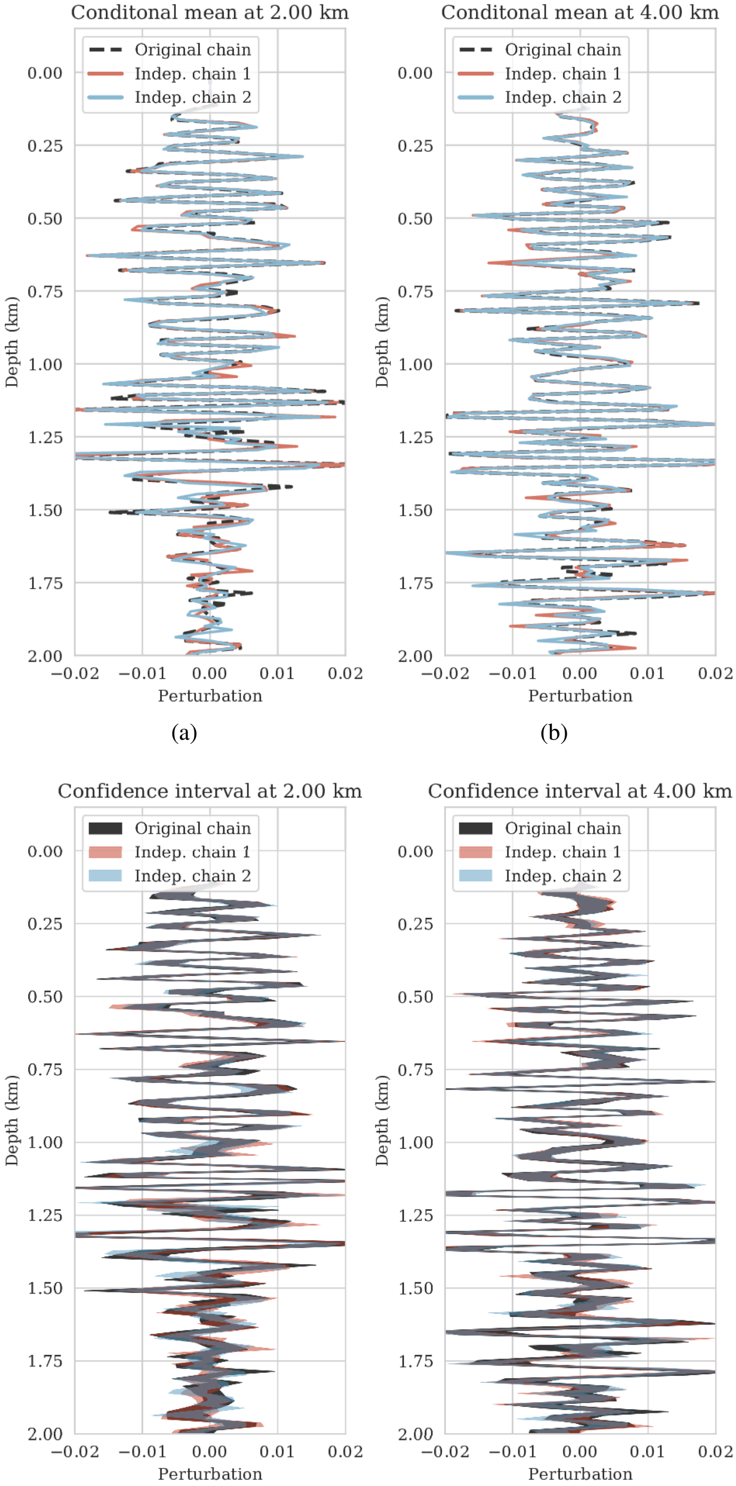
Given the mapping from image to horizons, let $p_{\mathcal{H}}\left(\mathbf{h} \mid \delta\mathbf{m}\right)$ represent the conditional PDF of horizons given an estimate for the seismic image. This distribution is characterized by the nondeterministic



(a)

(b) (c)

Figure 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, $\delta \mathbf{m}_{\text{MAP}}$, at (b) $2.0\,\text{km}$ and (c) $4.0\,\text{km}$ horizontal location.



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

Figure 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) $4.0\,\mathrm{km}$.

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 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:

$$\begin{aligned}
\mathbb{E}_{\mathbf{h}\sim p_{\text{post}}(\mathbf{h}\mid\mathbf{d})}\left[f\left(\mathbf{h}\right)\right] &= \int f\left(\mathbf{h}\right) p_{\text{post}}\left(\mathbf{h} \mid \mathbf{d}\right) \mathrm{d}\mathbf{h} = \iint f\left(\mathbf{h}\right) p\left(\mathbf{h}, \delta\mathbf{m} \mid \mathbf{d}\right) \mathrm{d}\mathbf{h}\, \mathrm{d}\delta\mathbf{m} \\
&= \iint f\left(\mathbf{h}\right) p_{\mathcal{H}}\left(\mathbf{h} \mid \delta\mathbf{m}, \mathbf{d}\right) p_{\text{post}}\left(\delta\mathbf{m} \mid \mathbf{d}\right) \mathrm{d}\mathbf{h}\, \mathrm{d}\delta\mathbf{m} \\
&= \iint f\left(\mathbf{h}\right) p_{\mathcal{H}}\left(\mathbf{h} \mid \delta\mathbf{m}\right) p_{\text{post}}\left(\delta\mathbf{m} \mid \mathbf{d}\right) \mathrm{d}\mathbf{h}\, \mathrm{d}\delta\mathbf{m} \\
&= \mathbb{E}_{\delta\mathbf{m}\sim p_{\text{post}}(\delta\mathbf{m}\mid\mathbf{d})}\left[\int f\left(\mathbf{h}\right) p_{\mathcal{H}}\left(\mathbf{h} \mid \delta\mathbf{m}\right) \mathrm{d}\mathbf{h}\right] \\
&= \underbrace{\mathbb{E}_{\delta\mathbf{m}\sim p_{\text{post}}(\delta\mathbf{m}\mid\mathbf{d})} \underbrace{\mathbb{E}_{\mathbf{h}\sim p_{\mathcal{H}}(\mathbf{h}\mid\delta\mathbf{m})}\left[f\left(\mathbf{h}\right)\right]}_{\substack{\text{uncertainty in} \\ \text{horizon tracking}}}.}_{\text{uncertainty in seismic imaging}}
\end{aligned} \tag{21}$$

The second equality in the first line of equation 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 20. Conceptually, equation 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 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 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 21 by the sample mean. In the following sections, we break equation 21 down into two cases where the horizon tracker yields an unique set of horizons or multiple sets of likely horizons, given one seismic image.

5.1.1 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

$$p_{\mathcal{H}}(\mathbf{h} \mid \delta\mathbf{m}) = \delta_{[\mathbf{h}=\mathcal{H}(\delta\mathbf{m})]}\left(\mathbf{h}\right), \tag{22}$$

where $\mathcal{H}$ represent the deterministic horizon tracking map and $\delta(\cdot)$ stands for the delta Dirac distribution. Substituting equation 22 into equation 21 yields

$$\begin{aligned}
\mathbb{E}_{\mathbf{h}\sim p_{\text{post}}(\mathbf{h}\mid\mathbf{d})}\left[f\left(\mathbf{h}\right)\right] &= \mathbb{E}_{\delta\mathbf{m}\sim p_{\text{post}}(\delta\mathbf{m}\mid\mathbf{d})}\left[\int f\left(\mathbf{h}\right) \delta_{[\mathbf{h}=\mathcal{H}(\delta\mathbf{m})]}\left(\mathbf{h}\right) \mathrm{d}\mathbf{h}\right], \\
&= \mathbb{E}_{\delta\mathbf{m}\sim p_{\text{post}}(\delta\mathbf{m}\mid\mathbf{d})}\left[f\left(\mathcal{H}\left(\delta\mathbf{m}\right)\right)\right], \\
&\approx \frac{1}{n_{\text{w}}} \sum_{j=1}^{n_{\text{w}}} f\left(\mathcal{H}(\delta\mathbf{m}_j)\right),
\end{aligned} \tag{23}$$

where $\left\{\delta\mathbf{m}_j\right\}_{j=1}^{n_{\text{w}}} \sim p_{\text{post}}(\delta\mathbf{m} \mid \mathbf{d})$ are n_{w} samples from the posterior distribution. Equation 23 essentially means that in case of a deterministic horizon tracker uncertainty in imaging can be

---

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

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.

5.1.2 Case 2: multiple likely horizons given an image

The probabilistic horizon tracking approach, described in equation 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 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 21 becomes

$$\begin{aligned}\mathbb{E}_{\mathbf{h}\sim p_{\text{post}}(\mathbf{h}\mid\mathbf{d})}\ [f\left(\mathbf{h}\right)] &= \mathbb{E}_{\delta\mathbf{m}\sim p_{\text{post}}(\delta\mathbf{m}\mid\mathbf{d})}\mathbb{E}_{\mathbf{h}\sim p_{\mathcal{H}}(\mathbf{h}\mid\delta\mathbf{m})}\left[f\left(\mathbf{h}\right)\right] \\ &\approx \mathbb{E}_{\delta\mathbf{m}\sim p_{\text{post}}(\delta\mathbf{m}\mid\mathbf{d})}\left[\frac{1}{n_h}\sum_{k=1}^{n_h} f\left(h_k^{(\delta\mathbf{m})}\right)\right], \\ &\approx \frac{1}{n_h n_{\text{w}}}\sum_{j=1}^{n_{\text{w}}}\sum_{k=1}^{n_h} f\left(h_k^{(\delta\mathbf{m}_j)}\right)\end{aligned} \tag{24}$$

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 24 differs from equation 23. In the ensuing sections, we show how equations 23 and 24 can be used to calculate pointwise estimates for the first two moments of the posterior distribution over horizons.

5.1.3 Uncertainty quantification in horizon tracking

It is often beneficial to express uncertainty in the form of confidence intervals. For this purpose, equation 23 or 24 is evaluated first for $f(\mathbf{h}) = \mathbf{h}$. This yields the conditional mean estimate for the horizons denoted by

$$\boldsymbol{\mu}_{\mathbf{h}} = \mathbb{E}_{\mathbf{h}\sim p_{\text{post}}(\mathbf{h}\mid\mathbf{d})}[\mathbf{h}]. \tag{25}$$

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

$$\boldsymbol{\sigma}_{\mathbf{h}}^2 = \mathbb{E}_{\mathbf{h}\sim p_{\text{post}}(\mathbf{h}\mid\mathbf{d})}[(\mathbf{h} - \boldsymbol{\mu}_{\mathbf{h}}) \odot (\mathbf{h} - \boldsymbol{\mu}_{\mathbf{h}})], \tag{26}$$

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.

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

The main goal of this paper 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 Wu and Fomel [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 3c) calculated for the Parihaka seismic imaging example. Next, control points are picked for the selected horizons at various horizontal positions, separated by $1\,\text{km}$. We group the control points

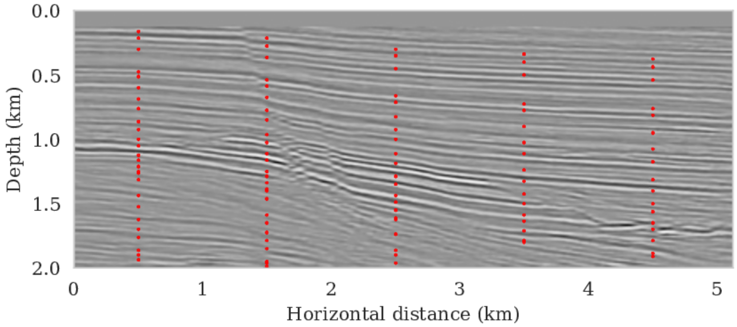


Figure 8: Five sets of control points identifying 25 horizons of interest.

with the same horizontal location, yielding five sets of control points. Figure 8 shows these five sets located at lateral positions 0.5, 1.5, 2.5, 3.5, and $4.5\,\mathrm{km}$.

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.

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

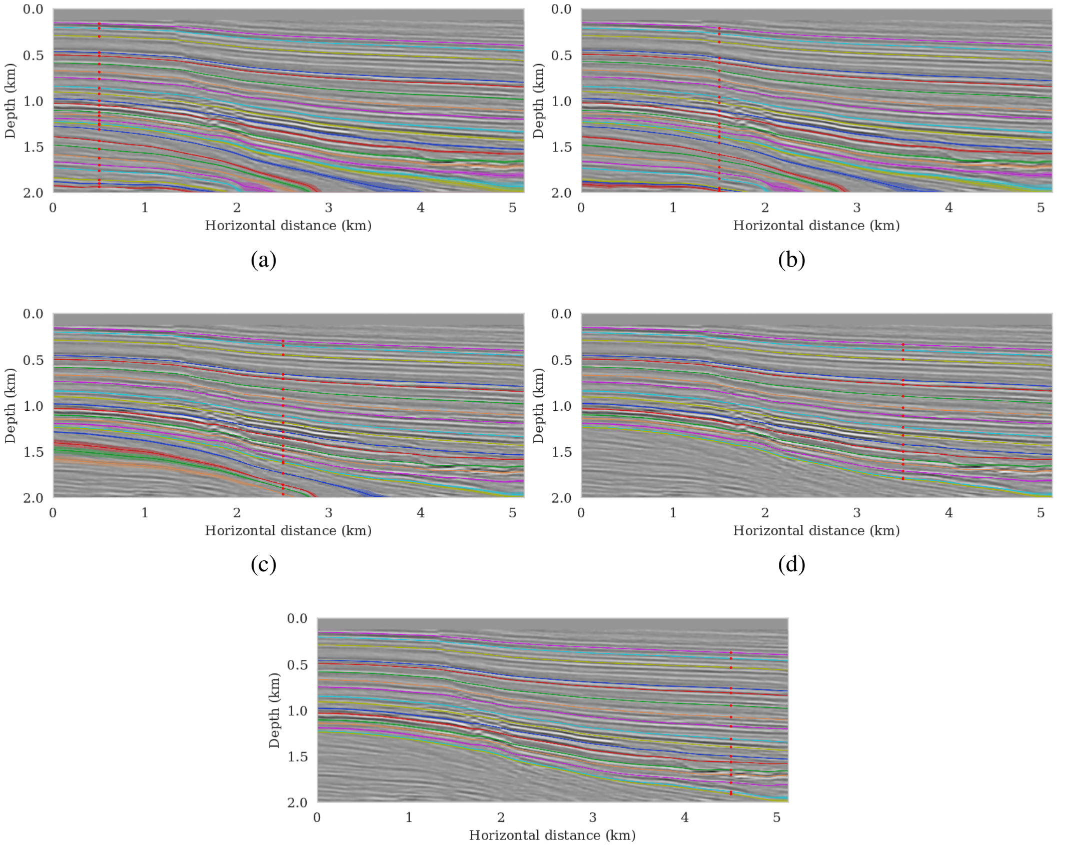
To calculate noise-induced uncertainties in horizon tracking (case 1), we pass samples from the imaging posterior distribution, $p_{\text{post}}\left(\delta\mathbf{m} \mid \mathbf{d}\right)$, to the automatic horizon tracking software [69]. Given the five sets of selected control points (Figure 8), the tracker generates for each sample of the imaging posterior 25 horizons according to equation 23. For each set of control points, the conditional mean and 99% confidence intervals are calculated included in Figure 9. Each plot (Figures 9a – 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 consider in the next section.

6.2 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 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 25 and 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 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



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

Figure 9: Uncertainty in horizon tracking only due noise in the shot data (equation 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.

variability comes from not trusting only one set of control points, but incorporating information from all the fives sets of control points.

7 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

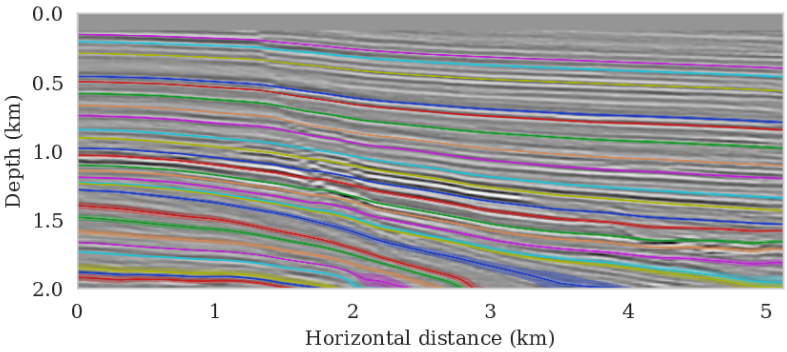


Figure 10: Uncertainty in horizon tracking due to a combination of uncertainties in imaging and control points (equation 24).

uncertainty quantification on the image, results of our probabilistic horizon tracking framework will be discussed below.

7.1 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.

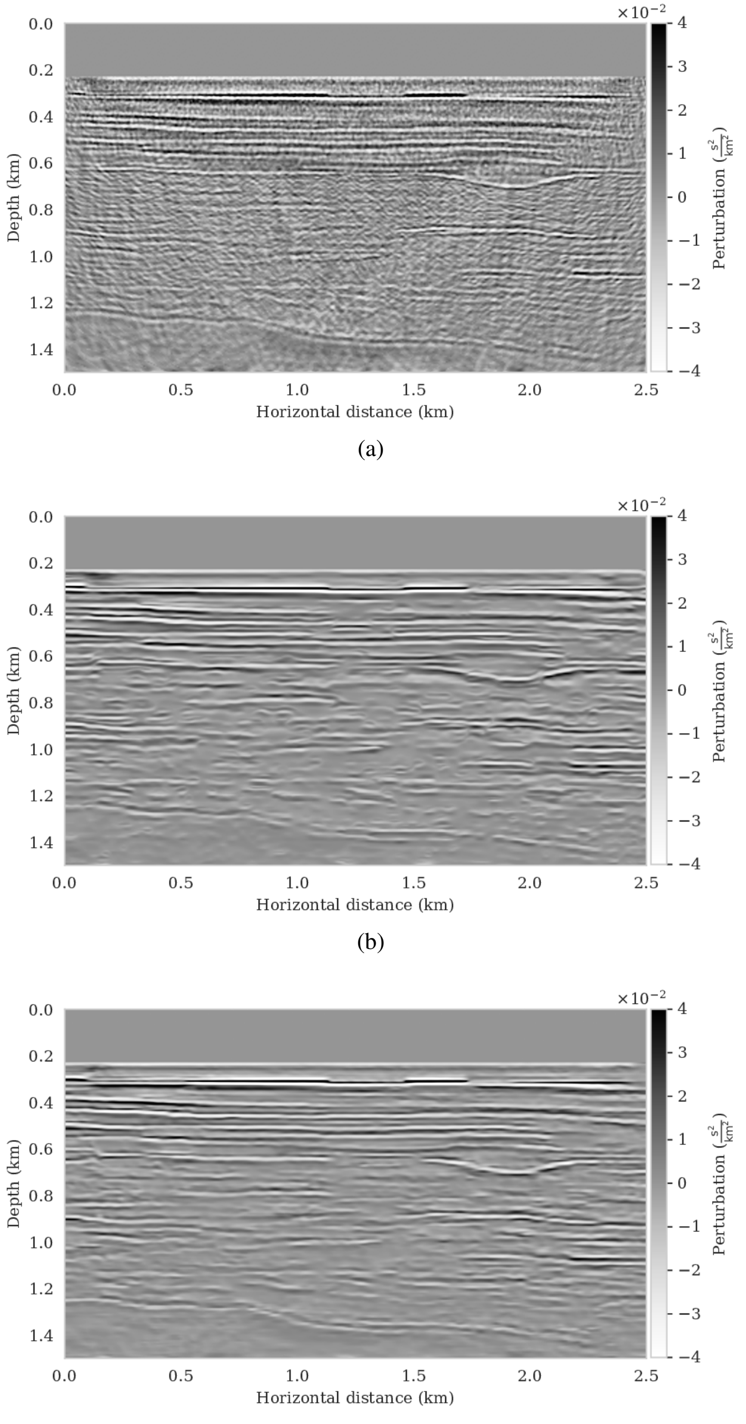
To image the shot data, we first derive a kinematically correct background model via smoothing. Next, linearized data is created by subtracting shot data simulated in this smooth background model from the shot data simulated in the actual model. This data serves as input to our imaging scheme.

7.2 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 5 with stochastic optimization limiting the number of iterations to six passes over the 101 shots. Equation 6 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 $2.80\,\mathrm{dB}$) 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 11a and 11b). By running Algorithm 1 for ten thousand iterations, we compute five thousand samples from the posterior distribution on the image following the stepsize strategy of equation 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 11c, represents a considerable visual improvement with a SNR of $4.12\,\mathrm{dB}$. 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 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.

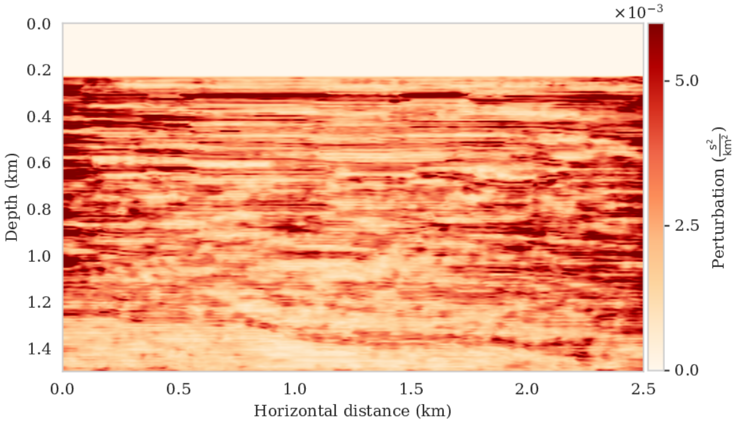


(a)

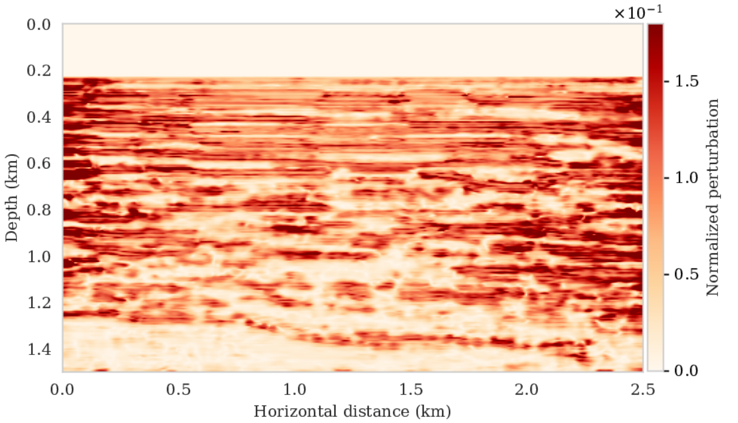
(b)

(c)

Figure 11: Imaging a 2D subset of the Compass model with deep priors. (a) MLE, i.e., minimizer of equation 4 with respect to $\delta \mathbf{m}$, with SNR $2.80\,\mathrm{dB}$. (b) The MAP estimate, i.e., minimizer of equation 7, following by mapping onto the image space via g (equation 10), with SNR $3.91\,\mathrm{dB}$. (c) The conditional (posterior) mean estimate, $\delta \mathbf{m}_{\mathrm{CM}}$, with SNR $4.12\,\mathrm{dB}$.



(a)



(b)

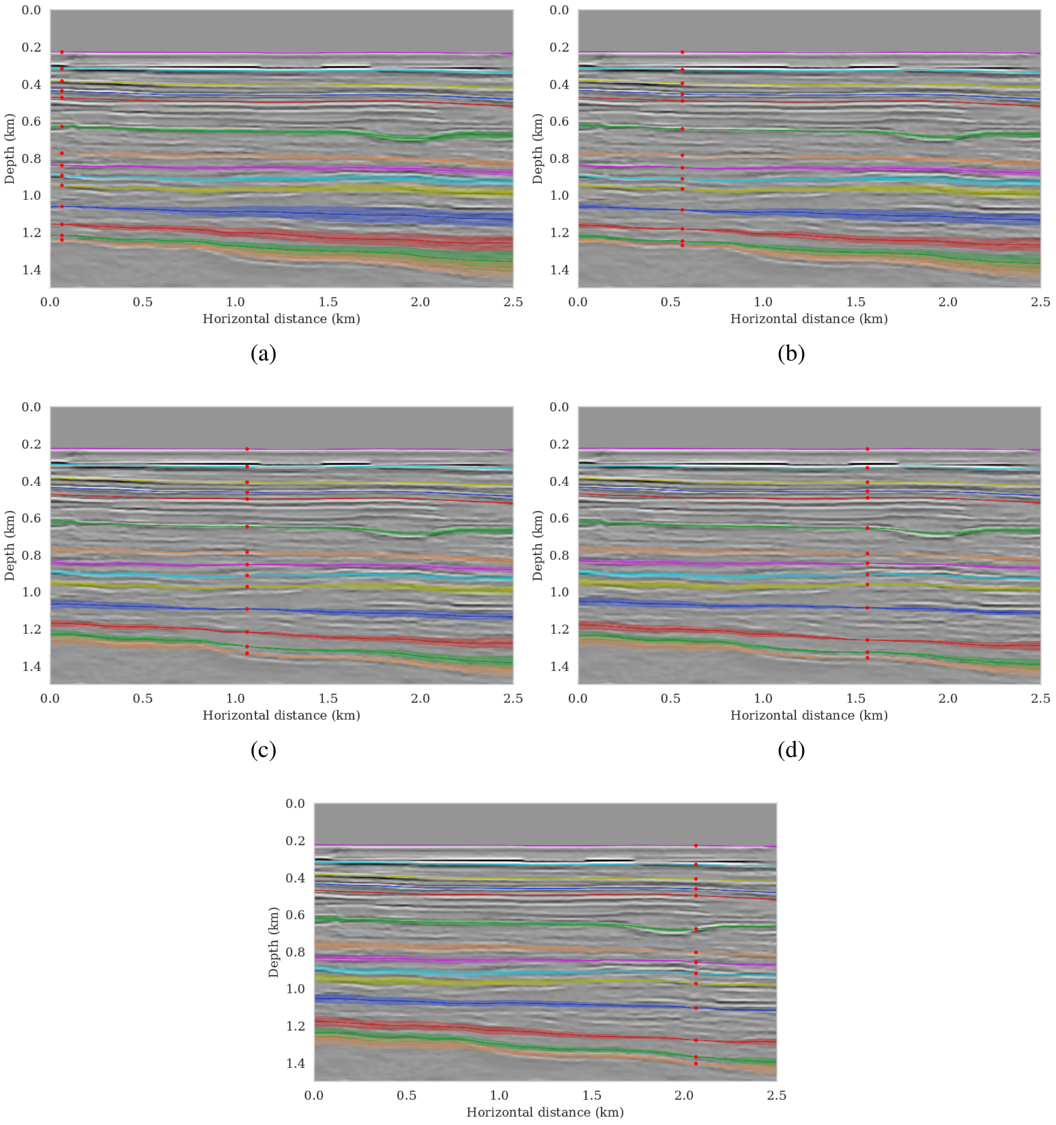
Figure 12: Imaging uncertainty quantification corresponding to the Compass model. (a) The pointwise standard deviation among samples drawn from the posterior, $\boldsymbol{\sigma}_{\text{post}}$. (b) Normalized pointwise standard deviation by the conditional mean estimate (Figure 11c).

7.3 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 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.

7.3.1 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 13a – 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.



(a) (b)

(c) (d)

(e)

Figure 13: Uncertainty in horizon tracking due uncertainties in imaging (equation 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.

7.3.2 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 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.

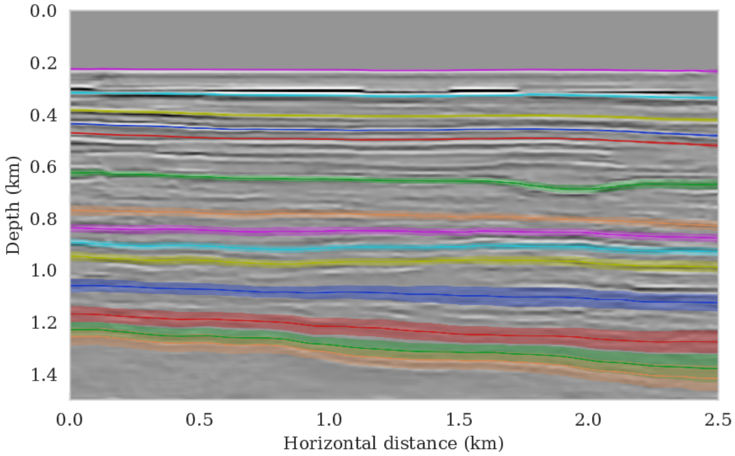


Figure 14: Uncertainty in horizon tracking due to a combination of uncertainties in imaging and control points (equation 24).

8 Discussion

The examples presented in this paper 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 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, 99–103], 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 Welling and Teh [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, Izmailov et al. [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 [32–34, 48, 122–127] 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 paper, 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.

9 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.

10 Related material

The SGLD iterations (equation 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 15 with automatic differentiation while exploiting Devito's highly optimized migration and demigration operators. For the CNN architecture, we followed Lempitsky et al. [51]. For the automated horizon tracking we made use of software written by Wu and Fomel [69]. For more details on our implementation, please refer to our code on GitHub.

11 Acknowledgments

This research was carried out with the support of Georgia Research Alliance and partners of the ML4Seismic Center.

12 Appendix A

12.1 Deep-prior network architecture

The CNN architecture proposed by Lempitsky et al. [51] is a variation of the widely used U-net architecture, as described by Ronneberger et al. [132]. 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 [51] 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 15 illustrates the exact deep prior architecture used in the Parihaka example in this paper, which closely follows the architecture advocated by Lempitsky et al. [51]. 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 15. The left most shaded block in Figure 15 is the $m \times n \times c$ input, which in the case of deep priors is a fixed random array, with m, n 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.

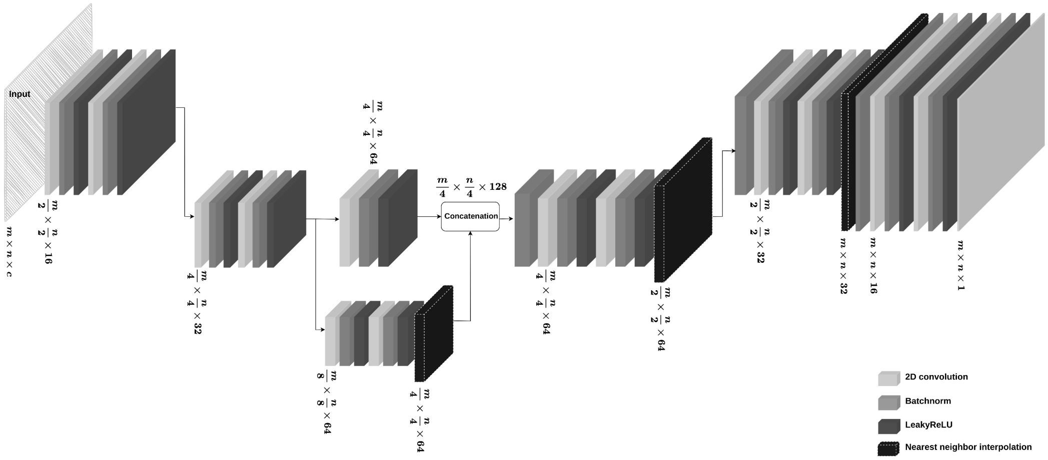


Figure 15: The architecture used for the Parihaka example.

The Compass example uses the same architecture as Figure 15, 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.

References

[1] Gilles Lambaré, Jean Virieux, Raul Madariaga, and Side Jin. Iterative asymptotic inversion in the acoustic approximation. *Geophysics*, 57(9):1138–1154, 1992.

[2] Gerard T Schuster. Least-squares cross-well migration. In *63rd Annual International Meeting, SEG*, pages 110–113. Expanded Abstracts, 1993. doi: 10.1190/1.1822308.

[3] Tamas Nemeth, Chengjun Wu, and Gerard T Schuster. Least-squares migration of incomplete reflection data. *GEOPHYSICS*, 64(1):208–221, 1999. doi: 10.1190/1.1444517.

[4] Alberto Malinverno and Victoria A Briggs. Expanded uncertainty quantification in inverse problems: Hierarchical bayes and empirical bayes. *GEOPHYSICS*, 69(4):1005–1016, 2004. doi: 10.1190/1.1778243.

[5] Albert Tarantola. *Inverse problem theory and methods for model parameter estimation*. SIAM, 2005. ISBN 978-0-89871-572-9. doi: 10.1137/1.9780898717921.

[6] Alberto Malinverno and Robert L Parker. Two ways to quantify uncertainty in geophysical inverse problems. *GEOPHYSICS*, 71(3):W15–W27, 2006. doi: 10.1190/1.2194516.

[7] James Martin, Lucas C. Wilcox, Carsten Burstedde, and OMAR Ghattas. A Stochastic Newton MCMC Method for Large-scale Statistical Inverse Problems with Application to Seismic Inversion. *SIAM Journal on Scientific Computing*, 34(3):A1460–A1487, 2012. URL http://epubs.siam.org/doi/abs/10.1137/110845598.

[8] Anandaroop Ray, Sam Kaplan, John Washbourne, and Uwe Albertin. Low frequency full waveform seismic inversion within a tree based Bayesian framework. *Geophysical Journal International*, 212(1):522–542, 10 2017. ISSN 0956-540X. doi: 10.1093/gji/ggx428. URL https://doi.org/10.1093/gji/ggx428.

[9] Zhilong Fang, Curt Da Silva, Rachel Kuske, and Felix J. Herrmann. Uncertainty quantification for inverse problems with weak partial-differential-equation constraints. *GEOPHYSICS*, 83(6): R629–R647, 2018. doi: 10.1190/geo2017-0824.1.

[10] Georgia K Stuart, Susan E Minkoff, and Felipe Pereira. A two-stage Markov chain Monte Carlo method for seismic inversion and uncertainty quantification. *GEOPHYSICS*, 84(6): R1003–R1020, 11 2019. doi: 10.1190/geo2018-0893.1.

[11] Zeyu Zhao and Mrinal K Sen. A gradient based MCMC method for FWI and uncertainty analysis. In *89th Annual International Meeting, SEG*, pages 1465–1469. Expanded Abstracts, 2019. doi: 10.1190/segam2019-3216560.1.

[12] M Kotsi, A Malcolm, and G Ely. Uncertainty quantification in time-lapse seismic imaging: a full-waveform approach. *Geophysical Journal International*, 222(2):1245–1263, 05 2020. ISSN 0956-540X. doi: 10.1093/gji/ggaa245.

[13] Alberto Malinverno. Parsimonious bayesian markov chain monte carlo inversion in a nonlinear geophysical problem. *Geophysical Journal International*, 151(3):675–688, 2002. doi: 10.1046/j.1365-246X.2002.01847.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1046/j.1365-246X.2002.01847.x.

[14] Dehan Zhu and Richard Gibson. Seismic inversion and uncertainty quantification using transdimensional markov chain monte carlo method. *GEOPHYSICS*, 83(4):R321–R334, 2018. doi: 10.1190/geo2016-0594.1.

[15] Gerhard Visser, Peng Guo, and Erdinc Saygin. Bayesian transdimensional seismic full-waveform inversion with a dipping layer parameterization. *GEOPHYSICS*, 84(6):R845–R858, 2019. doi: 10.1190/geo2018-0785.1.

[16] Tristan van Leeuwen, Aleksandr Y. Aravkin, and F. Herrmann. Seismic Waveform Inversion by Stochastic Optimization. *International Journal of Geophysics*, 2011:1–18, 2011.

[17] Felix J. Herrmann and Xiang Li. Efficient least-squares imaging with sparsity promotion and compressive sensing. *Geophysical Prospecting*, 60(4):696–712, 07 2012. doi: 10.1111/j.1365-2478.2011.01041.x. URL https://slim.gatech.edu/Publications/Public/Journals/GeophysicalProspecting/2012/herrmann11GPelsqIm/herrmann11GPelsqIm.pdf.

[18] Xinting Lu, Liguo Han, Jianglong Yu, and Xue Chen. L1 norm constrained migration of blended data with the fista algorithm. *Journal of Geophysics and Engineering*, 12(4):620–628, 2015.

[19] Ning Tu and Felix J. Herrmann. Fast imaging with surface-related multiples by sparse inversion. *Geophysical Journal International*, 201(1):304–317, 04 2015. URL https://slim.gatech.edu/Publications/Public/Journals/GeophysicalJournalInternational/2014/tu2014fis/tu2014fis.pdf.

[20] Hejun Zhu, Siwei Li, Sergey Fomel, Georg Stadler, and Omar Ghattas. A bayesian approach to estimate uncertainty for full-waveform inversion using a priori information from depth migration. *Geophysics*, 81(5):R307–R323, 2016. doi: 10.1190/geo2015-0641.1.

[21] Gregory Ely, Alison Malcolm, and Oleg V. Poliannikov. Assessing uncertainties in velocity models and images with a fast nonlinear uncertainty quantification method. *GEOPHYSICS*, 83(2):R63–R75, 2018. doi: 10.1190/geo2017-0321.1. URL https://doi.org/10.1190/geo2017-0321.1.

[22] Muhammad Izzatullah, Ricardo Baptista, Lester Mackey, Youssef Marzouk, and Daniel Peter. Bayesian seismic inversion: Measuring Langevin MCMC sample quality with kernels. In *90th Annual International Meeting, SEG*, pages 295–299. Expanded Abstracts, 06 2020. doi: 10.1190/segam2020-3422419.1.

[23] Lukas Mosser, Olivier Dubrule, and M Blunt. Stochastic Seismic Waveform Inversion Using Generative Adversarial Networks as a Geological Prior. *Mathematical Geosciences*, 84(1): 53–79, 2019. doi: 10.1007/s11004-019-09832-6.

[24] Zhen-Dong Zhang and Tariq Alkhalifah. Regularized elastic full waveform inversion using deep learning. *GEOPHYSICS*, 84(5):1SO–Z28, 9 2019. doi: 10.1190/geo2018-0685.1.

[25] Zhilong Fang, Hongjian Fang, and Laurent Demanet. Chapter Five - Deep generator priors for Bayesian seismic inversion. In Ben Moseley and Lion Krischer, editors, *Machine Learning in Geosciences*, volume 61 of *Advances in Geophysics*, pages 179–216. Elsevier, 2020. doi: https://doi.org/10.1016/bs.agph.2020.07.002. URL https://www.sciencedirect.com/science/article/pii/S0065268720300029.

[26] Zhilong Fang, Hongjian Fang, and Laurent Demanet. Quality control of deep generator priors for statistical seismic inverse problems. In *90th Annual International Meeting, SEG*, pages 1715–1719. Expanded Abstracts, 2020. doi: 10.1190/segam2020-3428013.1.

[27] Zhaolun Liu, Yuqing Chen, and Gerard Schuster. Deep Convolutional Neural Network and Sparse Least Squares Migration. *Geophysics*, 85(4):1–57, 7 2020. doi: 10.1190/geo2019-0412.1.

[28] L. Mosser, S. Purves, and E.Z. Naeini. Deep Bayesian Neural Networks for Fault Identification and Uncertainty Quantification. 2020(1):1–5, 2020. ISSN 2214-4609. doi: https://doi.org/10.3997/2214-4609.202032036. URL https://www.earthdoc.org/content/papers/10.3997/2214-4609.202032036.

[29] Hongyu Sun and Laurent Demanet. Extrapolated full-waveform inversion with deep learning. *Geophysics*, 85(3):R275–R288, 2020. doi: 10.1190/geo2019-0195.1.

[30] Y. Wu and Y. Lin. InversionNet: An Efficient and Accurate Data-Driven Full Waveform Inversion. *IEEE Transactions on Computational Imaging*, 6:419–433, 2020.

[31] Vladimir Kazei, Oleg Ovcharenko, Pavel Plotnitskii, Daniel Peter, Xiangliang Zhang, and Tariq Alkhalifah. Mapping full seismic waveforms to vertical velocity profiles by deep learning. *Geophysics*, 86(5):1–50, 2021. doi: 10.1190/geo2019-0473.1.

[32] Konik Kothari, AmirEhsan Khorashadizadeh, Maarten de Hoop, and Ivan Dokmanić. Trumpets: Injective Flows for Inference and Inverse Problems. *arXiv preprint arXiv:2102.10461*, 2021. URL https://arxiv.org/abs/2102.10461.

[33] Rajiv Kumar, Maria Kotsi, Ali Siahkoohi, and Alison Malcolm. Enabling uncertainty quantification for seismic data preprocessing using normalizing flows (NF)—An interpolation example. In *First International Meeting for Applied Geoscience & Energy*, pages 1515–1519. Expanded Abstracts, 2021. doi: 10.1190/segam2021-3583705.1.

[34] Ali Siahkoohi and Felix J Herrmann. Learning by example: fast reliability-aware seismic imaging with normalizing flows. In *First International Meeting for Applied Geoscience & Energy*, pages 1580–1585. Expanded Abstracts, 2021. doi: 10.1190/segam2021-3581836.1.

[35] Jonas Adler and Ozan Öktem. Deep Bayesian Inversion. *arXiv preprint arXiv:1811.05910*, 2018. URL https://arxiv.org/abs/1811.05910.

[36] Patrick Putzky and Max Welling. Invert to Learn to Invert. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/ac1dd209cbcc5e5d1c6e28598e8cbbe8-Paper.pdf.

[37] Muhammad Asim, Max Daniels, Oscar Leong, Ali Ahmed, and Paul Hand. Invertible generative models for inverse problems: mitigating representation error and dataset bias. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 399–409. PMLR, 13–18 Jul 2020. URL http://proceedings.mlr.press/v119/asim20a.html.

[38] Andreas Hauptmann and Ben T Cox. Deep Learning in Photoacoustic Tomography: Current approaches and future directions. *Journal of Biomedical Optics*, 25(11):112903, 2020.

[39] Anuroop Sriram, Jure Zbontar, Tullie Murrell, Aaron Defazio, C Lawrence Zitnick, Nafissa Yakubova, Florian Knoll, and Patricia Johnson. End-to-End Variational Networks for Accelerated MRI Reconstruction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 64–73. Springer, 2020.

[40] Subhadip Mukherjee, Marcello Carioni, Ozan Öktem, and Carola-Bibiane Schönlieb. End-to-end reconstruction meets data-driven regularization for inverse problems. *arXiv preprint arXiv:2106.03538*, 2021. URL https://arxiv.org/abs/2106.03538.

[41] Ali Siahkoohi, Mathias Louboutin, and Felix J. Herrmann. The importance of transfer learning in seismic modeling and imaging. *GEOPHYSICS*, 84(6):A47–A52, 11 2019. doi: 10.1190/geo2019-0056.1.

[42] Harpreet Kaur, Nam Pham, and Sergey Fomel. Improving the resolution of migrated images by approximating the inverse Hessian using deep learning. *Geophysics*, 85(4):WA173–WA183, 2020. doi: 10.1190/geo2019-0315.1.

[43] Gregory Ongie, Ajil Jalal, Christopher A. Metzler, Richard G. Baraniuk, Alexandros G. Dimakis, and Rebecca Willett. Deep learning techniques for inverse problems in imaging. *IEEE Journal on Selected Areas in Information Theory*, 1(1):39–56, 2020. doi: 10.1109/JSAIT.2020.2991563.

[44] Renán Rojas-Gómez, Jihyun Yang, Youzuo Lin, James Theiler, and Brendt Wohlberg. Physics-consistent data-driven waveform inversion with adaptive data augmentation. *IEEE Geoscience and Remote Sensing Letters*, pages 1–5, 2020. doi: 10.1109/LGRS.2020.3022021.

[45] Hongyu Sun and Laurent Demanet. Elastic full-waveform inversion with extrapolated low-frequency data. In *SEG Technical Program Expanded Abstracts 2020*, pages 855–859. Society of Exploration Geophysicists, 2020. doi: 10.1190/segam2020-3428087.1.

[46] Mi Zhang, Ali Siahkoohi, and Felix J. Herrmann. Transfer learning in large-scale ocean bottom seismic wavefield reconstruction. In *90th Annual International Meeting, SEG*, pages 1666–1670. Expanded Abstracts, 2020. doi: 10.1190/segam2020-3427882.1.

[47] Riccardo Barbano, Zeljko Kereta, Andreas Hauptmann, Simon R Arridge, and Bangti Jin. Unsupervised knowledge-transfer for learned image reconstruction. *arXiv preprint arXiv:2107.02572*, 2021. URL https://arxiv.org/abs/2107.02572.

[48] Ali Siahkoohi, 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*, 1 2021. URL https://openreview.net/pdf?id=P9m1sMaNQ8T.

[49] Shan Qu, Eric Verschuur, Dong Zhang, and Yangkang 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*, 86(3):A39–A43, 2021. doi: 10.1190/geo2020-0723.1.

[50] Jan-Willem Vrolijk and Gerrit Blacquière. Source deghosting of coarsely sampled common-receiver data using a convolutional neural network. *Geophysics*, 86(3):V185–V196, 2021. doi: 10.1190/geo2020-0186.1.

[51] V. Lempitsky, A. Vedaldi, and D. Ulyanov. Deep Image Prior. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9446–9454, June 2018. doi: 10.1109/CVPR.2018.00984.

[52] Simon Arridge, Peter Maass, Ozan Öktem, and Carola-Bibiane Schönlieb. Solving inverse problems using data-driven models. *Acta Numerica*, 28:1–174, 2019. doi: 10.1017/S0962492919000059.

[53] Zezhou Cheng, Matheus Gadelha, Subhransu Maji, and Daniel Sheldon. A Bayesian Perspective on the Deep Image Prior. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5443–5451, June 2019.

[54] Matheus Gadelha, Rui Wang, and Subhransu Maji. Shape reconstruction using differentiable projections and deep priors. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 22–30, 2019.

[55] Qun Liu, Lihua Fu, and Meng Zhang. Deep-seismic-prior-based reconstruction of seismic data using convolutional neural networks. *arXiv preprint arXiv:1911.08784*, 2019.

[56] Yulang Wu and George A McMechan. Parametric convolutional neural network-domain full-waveform inversion. *GEOPHYSICS*, 84(6):R881–R896, 2019. doi: 10.1190/geo2018-0224.1.

[57] Sören Dittmer, Tobias Kluth, Peter Maass, and Daniel Otero Baguer. Regularization by architecture: A deep prior approach for inverse problems. *Journal of Mathematical Imaging and Vision*, 62(3):456–470, 2020.

[58] Fantong Kong, Francesco Picetti, Vincenzo Lipari, Paolo Bestagini, Xiaoming Tang, and Stefano Tubaro. Deep Prior-Based Unsupervised Reconstruction of Irregularly Sampled Seismic Data. *IEEE Geoscience and Remote Sensing Letters*, 19(7501305), 2020. doi: 10.1109/LGRS.2020.3044455.

[59] Yunzhi Shi, Xinming Wu, and Sergey Fomel. Deep learning parameterization for geophysical inverse problems. In *SEG 2019 Workshop: Mathematical Geophysics: Traditional vs Learning, Beijing, China, 5-7 November 2019*, pages 36–40. Society of Exploration Geophysicists, 2020. doi: 10.1190/iwmg2019_09.1.

[60] Ali Siahkoohi, Gabrio Rizzuti, and Felix J. Herrmann. A deep-learning based bayesian approach to seismic imaging and uncertainty quantification. In *82nd EAGE Conference and Exhibition*. Extended Abstracts, 2020. doi: 10.3997/2214-4609.202010770.

[61] Ali Siahkoohi, Gabrio Rizzuti, and Felix J. Herrmann. Uncertainty quantification in imaging and automatic horizon tracking—a Bayesian deep-prior based approach. In *90th Annual International Meeting, SEG*, pages 1636–1640. Expanded Abstracts, 9 2020. doi: 10.1190/segam2020-3417560.1.

[62] Ali Siahkoohi, Gabrio Rizzuti, and Felix J. Herrmann. Weak deep priors for seismic imaging. In *90th Annual International Meeting, SEG*, pages 2998–3002. Expanded Abstracts, 9 2020. doi: 10.1190/segam2020-3417568.1.

[63] Malte Tölle, Max-Heinrich Laves, and Alexander Schlaefer. A Mean-Field Variational Inference Approach to Deep Image Prior for Inverse Problems in Medical Imaging. In *Medical Imaging with Deep Learning*, 2021.

[64] Tom M Mitchell. The need for biases in learning generalizations. Technical Report CBM-TR 5-110, Department of Computer Science, Laboratory for Computer Science Research, Rutgers University, 1980. URL http://www.cs.cmu.edu/~tom/pubs/NeedForBias_1980.pdf.

[65] Anders Krogh and John A Hertz. A simple weight decay can improve generalization. In *Advances in Neural Information Processing Systems*, pages 950–957, 1992.

[66] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

[67] Max Welling and Yee Whye Teh. Bayesian Learning via Stochastic Gradient Langevin Dynamics. In *Proceedings of the 28th International Conference on Machine Learning*, ICML'11, pages 681–688, Madison, WI, USA, 2011. Omnipress. ISBN 9781450306195. doi: 10.5555/3104482.3104568. URL https://dl.acm.org/doi/abs/10.5555/3104482.3104568.

[68] Chunyuan Li, Changyou Chen, David Carlson, and Lawrence Carin. Preconditioned Stochastic Gradient Langevin Dynamics for Deep Neural Networks. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, pages 1788–1794. AAAI Press, 2016. doi: 10.5555/3016100.3016149. URL https://dl.acm.org/doi/abs/10.5555/3016100.3016149.

[69] Xinming Wu and Sergey Fomel. Least-squares horizons with local slopes and multi-grid correlations. *GEOPHYSICS*, 83:IM29–IM40, 2018. doi: 10.1190/geo2017-0830.1. URL https://doi.org/10.1190/geo2017-0830.1.

[70] Richard Arnold and Andrew Curtis. Interrogation theory. *Geophysical Journal International*, 214(3):1830–1846, 06 2018. ISSN 0956-540X. doi: 10.1093/gji/ggy248. URL https://doi.org/10.1093/gji/ggy248.

[71] Bas Peters, Justin Granek, and Eldad Haber. Multiresolution neural networks for tracking seismic horizons from few training images. *Interpretation*, 7(3):SE201–SE213, 2019. doi: 10.1190/INT-2018-0225.1. URL https://doi.org/10.1190/INT-2018-0225.1.

[72] Zhicheng Geng, Xinming Wu, Yunzhi Shi, and Sergey Fomel. Deep learning for relative geologic time and seismic horizons. *GEOPHYSICS*, 85(4):WA87–WA100, 7 2020. doi: 10.1190/geo2019-0252.1.

[73] B. Peters and E. Haber. Fully Reversible Neural Networks for Large-Scale 3D Seismic Horizon Tracking. 2020(1):1–5, 2020. ISSN 2214-4609. doi: https://doi.org/10.3997/2214-4609.202012019. URL https://www.earthdoc.org/content/papers/10.3997/2214-4609.202012019.

[74] Yunzhi Shi, Xinming Wu, and Sergey Fomel. Waveform embedding: Automatic horizon picking with unsupervised deep learning. *GEOPHYSICS*, 85(4):WA67–WA76, 7 2020. doi: 10.1190/geo2019-0438.1.

[75] Alejandro A Valenciano. *Imaging by wave-equation inversion*. PhD thesis, Stanford University, 2008.

[76] S Dong, J Cai, M Guo, S Suh, Z Zhang, B Wang, and e Z Li. Least-squares reverse time migration: towards true amplitude imaging and improving the resolution. In *82nd Annual International Meeting, SEG*, pages 1–5. Expanded Abstracts, 2012. doi: 10.1190/segam2012-1488.1.

[77] Chong Zeng, Shuqian Dong, and Bin Wang. Least-squares reverse time migration: Inversion-based imaging toward true reflectivity. *The Leading Edge*, 33(9):962–968, 2014.

[78] David L Donoho. Compressed Sensing. *IEEE Transactions on Information Theory*, 52(4): 1289–1306, 2006.

[79] Ashish Bora, Ajil Jalal, Eric Price, and Alexandros G. Dimakis. Compressed sensing using generative models. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 537–546. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/bora17a.html.

[80] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 249–256. PMLR, May 2010. URL https://proceedings.mlr.press/v9/glorot10a.html.

[81] George Casella and Roger L Berger. *Statistical Inference*. Cengage Learning, 2002.

[82] Richard C Aster, Brian Borchers, and Clifford H Thurber. *Parameter Estimation and Inverse Problems*. Elsevier, 2018. doi: 10.1016/C2015-0-02458-3.

[83] Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/a41b3bb3e6b050b6c9067c67f663b915-Paper.pdf.

[84] Tijmen Tieleman and Geoffrey 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.

[85] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. URL https://arxiv.org/pdf/1412.6980.pdf.

[86] Jeremy Bernstein, Arash Vahdat, Yisong Yue, and Ming-Yu Liu. On the distance between two neural networks and the stability of learning. *arXiv preprint arXiv:2002.03432*, 2020. URL https://arxiv.org/pdf/2002.03432.pdf.

[87] Simon S. Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=S1eK3i09YQ.

[88] Daniel Kunin, Jonathan Bloom, Aleksandrina Goeva, and Cotton Seed. Loss landscapes of regularized linear autoencoders. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3560–3569. PMLR, 09–15 Jun 2019. URL http://proceedings.mlr.press/v97/kunin19a.html.

[89] Bogachev, V.I. *Measure Theory*. Springer-Verlag Berlin Heidelberg, 2006. ISBN 9783540345138. doi: 10.1007/978-3-540-34514-5. URL https://books.google.com/books?id=4vmOjgEACAAJ.

[90] Brian D.O. Anderson and John B. Moore. *Optimal Filtering*. Prentice-Hall, Englewood Cliffs, NJ., 1979.

[91] David JC MacKay and David JC Mac Kay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

[92] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., 2001.

[93] Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian Data Analysis*. CRC press, 2013. doi: 10.1201/9780429258480. URL http://www.stat.columbia.edu/~gelman/book/BDA3.pdf.

[94] Andrew Curtis and Anthony Lomax. Prior information, sampling distributions, and the curse of dimensionality. *Geophysics*, 66(2):372–378, 2001.

[95] Felix J. Herrmann, Ali Siahkoohi, and Gabrio Rizzuti. Learned imaging with constraints and uncertainty quantification. In *Neural Information Processing Systems (NeurIPS) 2019 Deep Inverse Workshop*, 12 2019. URL https://arxiv.org/pdf/1909.06473.pdf.

[96] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

[97] Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4): 1574–1609, 2009.

[98] Chuang Li, Jianping Huang, Zhenchun Li, and Rongrong Wang. Plane-wave least-squares reverse time migration with a preconditioned stochastic conjugate gradient method. *Geophysics*, 83(1):S33–S46, 2018.

[99] Eldad Haber, Matthias Chung, and Felix Herrmann. An effective method for parameter estimation with pde constraints with multiple right-hand sides. *SIAM Journal on Optimization*, 22(3):739–757, 2012.

[100] Felix J. Herrmann and Xiang Li. Efficient least-squares imaging with sparsity promotion and compressive sensing. *Geophysical Prospecting*, 60:696–712, 2012. doi: 10.1111/j.1365-2478.2011.01041.x.

[101] Xinting Lu, Liguo Han, Jianglong Yu, and Xue Chen. L1 norm constrained migration of blended data with the FISTA algorithm. *Journal of Geophysics and Engineering*, 12:620–628, 2015. doi: 10.1088/1742-2132/12/4/620.

[102] Ning Tu and F.J. Herrmann. Fast imaging with surface-related multiples by sparse inversion. *Geophysical Journal International*, 201(1):304–417, 2015. doi: 10.1093/gji/ggv020.

[103] Chuang Li, Jianping Huang, Zhenchun Li, and Rongrong Wang. Plane-wave least-squares reverse time migration with a preconditioned stochastic conjugate gradient method. *Geophysics*, 83(1):S33–S46, 2018. doi: 10.1190/geo2017-0339.1. URL https://doi.org/10.1190/geo2017-0339.1.

[104] Radford M. Neal. MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, pages 113–162, May 2011. doi: 10.1201/b10905. URL http://dx.doi.org/10.1201/b10905.

[105] Issei Sato and Hiroshi Nakagawa. Approximation analysis of stochastic gradient langevin dynamics by using fokker-planck equation and ito process. In Eric P. Xing and Tony Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 982–990, Bejing, China, 22–24 Jun 2014. PMLR. URL https://proceedings.mlr.press/v32/satoa14.html.

[106] Maxim Raginsky, Alexander Rakhlin, and Matus Telgarsky. Non-Convex Learning via Stochastic Gradient Langevin Dynamics: A Nonasymptotic Analysis. In *Conference on Learning Theory*, pages 1674–1703. PMLR, 2017.

[107] Nicolas Brosse, Alain Durmus, and Eric Moulines. The promises and pitfalls of Stochastic Gradient Langevin Dynamics. In *Advances in Neural Information Processing Systems 31*, pages 8268–8278. Curran Associates, Inc., 2018. URL http://papers.nips.cc/paper/8048-the-promises-and-pitfalls-of-stochastic-gradient-langevin-dynamics.pdf.

[108] Wei Deng, Guang Lin, and Faming Liang. A Contour Stochastic Gradient Langevin Dynamics Algorithm for Simulations of Multi-modal Distributions. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 15725–15736. Curran Associates, Inc., 2020.

[109] Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. An Introduction to Variational Methods for Graphical Models. *Machine Learning*, 37(2):183–233, 1999. doi: 10.1023/A:1007665907178.

[110] Yee Whye Teh, Alexandre H Thiery, and Sebastian J Vollmer. Consistency and fluctuations for stochastic gradient Langevin dynamics. *The Journal of Machine Learning Research*, 17(1): 193–225, 2016.

[111] Andrew Gelman and Donald B. Rubin. Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science*, 7(4):457–472, 1992. ISSN 08834237. URL http://www.jstor.org/stable/2246093.

[112] Veritas. Parihaka 3D Marine Seismic Survey - Acquisition and Processing Report. Technical Report New Zealand Petroleum Report 3460, New Zealand Petroleum & Minerals, Wellington, 2005.

[113] WesternGeco. Parihaka 3D PSTM Final Processing Report. Technical Report New Zealand Petroleum Report 4582, New Zealand Petroleum & Minerals, Wellington, 2012.

[114] Martin Burger and Felix Lucka. Maximum a posteriori estimates in linear inverse problems with log-concave priors are proper Bayes estimators. *Inverse Problems*, 30(11):114004, 2014. doi: 10.1088/0266-5611/30/11/114004.

[115] A. P. Dawid. Conditional Independence in Statistical Theory. *Journal of the Royal Statistical Society. Series B (Methodological)*, 41(1):1–31, 1979. ISSN 00359246. URL http://www.jstor.org/stable/2984718.

[116] CE Jones, JA Edgar, JI 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. doi: 10.3997/2214-4609.20148575.

[117] Andrew Gordon Wilson and Pavel Izmailov. Bayesian Deep Learning and a Probabilistic Perspective of Generalization. *arXiv preprint arXiv:2002.08791*, 2020.

[118] Yongchae Cho and Hyunggu 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*, 200:108333, 2021. ISSN 0920-4105. doi: https://doi.org/10.1016/j.petrol.2020.108333. URL https://www.sciencedirect.com/science/article/pii/S0920410520313875.

[119] Pavel Izmailov, Wesley J Maddox, Polina Kirichenko, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Subspace inference for Bayesian deep learning. In *Uncertainty in Artificial Intelligence*, pages 1169–1179. Proceedings of Machine Learning Research (PMLR), 2020.

[120] Reinhard Heckel and Paul Hand. Deep decoder: Concise image representations from untrained non-convolutional networks. In *7th International Conference on Learning Representations*. OpenReview.net, 2019. URL https://openreview.net/forum?id=rylV-2C9KQ.

[121] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. volume 37 of *Proceedings of Machine Learning Research*, pages 1530–1538. PMLR, 07–09 Jul 2015. URL http://proceedings.mlr.press/v37/rezende15.html.

[122] Gabrio Rizzuti, Ali Siahkoohi, Philipp A. Witte, and Felix J. Herrmann. Parameterizing uncertainty by deep invertible networks, an application to reservoir characterization. In *90th Annual International Meeting, SEG*, pages 1541–1545, 09 2020. doi: 10.1190/segam2020-3428150.1.

[123] Ali Siahkoohi, Gabrio Rizzuti, Philipp A. Witte, and Felix J. Herrmann. Faster Uncertainty Quantification for Inverse Problems with Conditional Normalizing Flows. Technical Report TR-CSE-2020-2, Georgia Institute of Technology, 07 2020. URL http://arxiv.org/abs/2007.07985.

[124] Xin Zhang and Andrew Curtis. Seismic tomography using variational inference methods. *Journal of Geophysical Research: Solid Earth*, 125(4):e2019JB018589, 2020.

[125] Nikola Kovachki, Ricardo Baptista, Bamdad Hosseini, and Youssef Marzouk. Conditional Sampling With Monotone GANs, 2021.

[126] Jakob Kruse, Gianluca Detommaso, Robert Scheichl, and Ullrich Köthe. HINT: Hierarchical Invertible Neural Transport for Density Estimation and Bayesian Inference. *Proceedings of AAAI-2021*, 2021. URL https://arxiv.org/pdf/1905.10687.pdf.

[127] Xuebin Zhao, Andrew Curtis, and Xin Zhang. Bayesian seismic tomography using normalizing flows. *Geophysical Journal International*, 228(1):213–239, 07 2021. ISSN 0956-540X. doi: 10.1093/gji/ggab298. URL https://doi.org/10.1093/gji/ggab298.

[128] Mengmeng Yang, Marie Graff, Rajiv Kumar, and Felix J. Herrmann. Low-rank representation of omnidirectional subsurface extended image volumes. *Geophysics*, 86(3):S165–S183, 1 2021. doi: 10.1190/geo2020-0152.1. URL https://slim.gatech.edu/Publications/Public/Journals/Geophysics/2021/yang2020lrpo/Paper_final.html.

[129] Zongyi Li, Nikola Borislavov Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew M. Stuart, and Anima Anandkumar. Fourier Neural Operator for Parametric Partial Differential Equations. In *9th International Conference on Learning Representations*. OpenReview.net, 2021. URL https://openreview.net/forum?id=c8P9NQVtmnO.

[130] F. Luporini, M. Lange, M. Louboutin, N. Kukreja, J. Hückelheim, C. Yount, P. Witte, P. H. J. Kelly, F. J. Herrmann, and G. J. Gorman. Architecture and performance of devito, a system for automated stencil computation. *CoRR*, abs/1807.03032, jul 2018. URL http://arxiv.org/abs/1807.03032.

[131] M. Louboutin, M. Lange, F. Luporini, N. Kukreja, P. A. Witte, F. J. Herrmann, P. Velesko, and G. J. Gorman. Devito (v3.1.0): an embedded domain-specific language for finite differences and geophysical exploration. *Geoscientific Model Development*, 12(3):1165–1187, 2019. doi: 10.5194/gmd-12-1165-2019. URL https://www.geosci-model-dev.net/12/1165/2019/.

[132] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.