SIAM/ASA J. UNCERTAINTY QUANTIFICATION
Vol. 9, No. 4, pp. 1499–1526

Stability of Gibbs Posteriors from the Wasserstein Loss for Bayesian Full Waveform Inversion*

Matthew M. Dunlop† and Yunan Yang†

Abstract. Recently, the Wasserstein loss function has been proven to be effective when applied to deterministic full-waveform inversion (FWI) problems. We consider the application of this loss function in Bayesian FWI so that the uncertainty can be captured in the solution. Other loss functions that are commonly used in practice are also considered for comparison. Existence and stability of the resulting Gibbs posteriors are shown on function space under weak assumptions on the prior and model. In particular, the distribution arising from the Wasserstein loss is shown to be quite stable with respect to high-frequency noise in the data. We then illustrate the difference between the resulting distributions numerically, using Laplace approximations to estimate the unknown velocity field and uncertainty associated with the estimates.

Key words. seismic inversion, Bayesian inference, Wasserstein metric, Gibbs posterior, prior distribution

AMS subject classifications. Primary, 62C10; Secondary, 86A22, 65J22, 49K40

DOI. 10.1137/20M1334218

1. Introduction.

Seismic signals generated by natural earthquakes or induced seismicity contain essential information about subsurface properties. Nowadays, vibrations of the earth can be measured on the surface more accurately and more frequently in spatial and time domains, rather than merely measuring the travel time [39]. The full wavefield could be generated by partial differential equations (PDEs) from the acoustic wave equation to 3D elastic wave modeling with attenuation. The state-of-the-art imaging technique in geophysics is the full-waveform inversion (FWI) [39], which seeks the optimal parameter by minimizing the objective function that measures the data mismatch between the recorded true data and the simulated waveforms produced by the current prediction. In the deterministic setting, it is PDE-constrained optimization.

Most current FWI studies have focused on how to efficiently and accurately solve the data-fitting problem. However, the recorded data often contain various types of noise that affect the accuracy of inversion [37]. For example, common sources of noise include the surface upon which the survey was performed, the instruments of receiving and recording, and the noise generated by the induced seismicity [28]. As a result, the resolution analysis and uncertainty quantification of the predictions are essential aspects [16, 17]. Bayesian inference provides a

*Received by the editors May 11, 2020; accepted for publication (in revised form) July 27, 2021; published electronically October 25, 2021.
https://doi.org/10.1137/20M1334218
Funding: This work was partially supported by National Science Foundation grant DMS-1913129, and by the U.S. Department of Energy Office of Science, Advanced Scientific Computing Research (ASCR), Scientific Discovery through Advanced Computing (SciDAC) program.

†Courant Institute of Mathematical Sciences, New York University, New York, NY 10012 USA (matt.dunlop@nyu.edu, yunan.yang@nyu.edu).

systematic way to quantify uncertainties for geophysical inverse problems [6, 43]. A realistic noise model is an essential a priori for Bayesian inversion of geophysical data, and the likelihood function is also partially determined by choice of the noise model. However, quantification of the noise model is nontrivial, and the common additive Gaussian assumption might not be enough to characterize the real uncertainty [27].

The shape and curvature of the likelihood surface represent information about the stability of the estimates, whose analogue in the deterministic approach of solving FWI is the objective function in PDE-constrained optimization. The oscillatory and periodic nature of waveforms leads to the main challenge of local minima, which can be significantly mitigated by a recently introduced class of objective functions from optimal transport [12, 13, 24, 42]. Attractive properties such as the convexity and insensitivity to noise have been theoretically studied in [15, 40, 41]. Since optimal transport studies probability measures, the use of the Wasserstein metrics as criteria is natural in statistical inference [3, 33], particularly in the Bayesian setting [11]. In [27], the quadratic Wasserstein metric was first used in Bayesian seismic inversion as the likelihood function. This work, together with the success of optimal-transport-based metrics in deterministic FWI, motivates us to further analyze the Wasserstein metric under the framework of Bayesian seismic inversion [21, 43]. Compared to [27], there are three new contributions to this field from our paper. First, we theoretically study the underlying noise model assumption on the data by choosing the quadratic Wasserstein metric as the likelihood function. Second, we prove the existence and stability of the corresponding posterior distributions, which are formally solutions to Bayesian inverse problems. The benefits of choosing the Wasserstein likelihood stand out from the stability estimates because of its robustness to the high-frequency noise. The noise robustness traces back to both the asymptotic connection [29] and the nonasymptotic connection [30] between the quadratic Wasserstein metric and the negative Sobolev seminorm. The conclusion in this paper is consistent with the analysis of using the quadratic Wasserstein metric in general deterministic inverse data matching problems [15]. Third, the numerical examples illustrated here consider the nonlinear inversion of the velocity parameter, which is discretized into thousands of variables. The number of unknowns here is much larger than the ones in [27].

The rest of the paper is organized as follows. We first briefly review necessary background knowledge on optimal transport, the quadratic Wasserstein metric, and the related negative Sobolev norm. As the main application of the paper, seismic waveform inversion is briefly introduced, and the major challenges are addressed. The main contribution of this paper is presented in sections 3 and 4, where Bayesian FWI is studied systematically. Starting with basics in Bayesian inversion, we then outline choices of prior distribution appropriate for seismic inversion [2, 31]. Next, we consider different likelihoods/potentials and the corresponding data models. We define four potentials ($\Phi_\cdot : X \times Y \to \mathbb{R}$) and discuss the corresponding noise models for each one of them:

$$\Phi_{L^2}(u;y) = \frac{1}{2}\int_D \|\mathcal{G}(u)(x,\cdot) - y(x,\cdot)\|^2_{L^2(T)}\,\lambda(\mathrm{d}x), \tag{1.1}$$

$$\Phi_{\dot{H}^{-1}}(u;y) = \frac{1}{2}\int_D \|\mathcal{G}(u)(x,\cdot) - y(x,\cdot)\|^2_{\dot{H}^{-1}(T)}\,\lambda(\mathrm{d}x), \tag{1.2}$$

$$\Phi_M(u;y) = \frac{1}{2}\int_D \left\| \frac{(P_\sigma \mathcal{G}(u))(x,\cdot) - (P_\sigma y)(x,\cdot)}{(P_\sigma y)(x,\cdot)} \right\|^2_{L^2(T)} \lambda(\mathrm{d}x), \tag{1.3}$$

$$\Phi_{W_2}(u;y) = \frac{1}{2}\int_D W_2\left((P_\sigma y)(x,\cdot), (P_\sigma \mathcal{G}(u))(x,\cdot)\right)^2 \lambda(\mathrm{d}x), \tag{1.4}$$

where the choice of norm on Y depends on the choice of potential; here P_σ is an operator that maps functions into probability densities. In the rest of section 3, we will rigorously define the posterior distributions (3.1) and study the existence and stability with respect to perturbations to the observed data for the different choices of likelihood. In Theorem 4.4, we establish that the posterior measures corresponding to the likelihoods discussed above are well-defined, and in Theorem 4.6, we prove that these measures are stable with respect to perturbations of the observed data measured in different norms. The theoretical analysis demonstrates the advantages of choosing Wasserstein-type likelihood. Numerical simulations are shown in section 5 to demonstrate the main findings of our study. We will present a challenging benchmark in FWI. We can tackle them using proper choices of likelihood and priors by combining the mathematical tool and the physics knowledge of the geophysical problem. Concluding remarks are offered in section 6.

2. Background.

In this section, we present some background material regarding the quadratic Wasserstein metric, the negative Sobolev norm, optimal transport, and the waveform inversion problem. To better motivate our study, we will also briefly mention the deterministic approach that is reformulated as a wave-equation-constrained optimization.

2.1. The quadratic Wasserstein distance and the negative Sobolev norm.

The Wasserstein distance comes from optimal transport, which is a classical subject in mathematical analysis that was first brought up by Monge in 1781 [25] and later expanded by Kantorovich in 1940s [22]. The core of the subject, i.e., the optimal transport problem, discusses the optimal plan that maps one probability distribution ν_1 on a measure space X into another probability distribution ν_2 on a measure space Y, intending to minimize the total transport cost of a given cost function. The transport cost function $c(x,y)$ maps pairs $(x,y) \in X \times Y$ to $\mathbb{R} \cup \{+\infty\}$, which denotes the cost of transporting one unit mass from location x to y. If $c(x,y) = |x-y|^p$ for $p \geq 1$, the optimal transport cost becomes the class of Wasserstein distance.

Definition 2.1 (the Wasserstein distance). *We denote by $\mathscr{P}_p(X)$ the set of probability measures with finite moments of order p on the measure space X. For all $p \in [1,\infty)$ and any $\nu_1, \nu_2 \in \mathscr{P}_p(X)$, the p-Wasserstein distance between ν_1 and ν_2 is defined as*

$$W_p(\nu_1,\nu_2) = \left(\inf_{T \in \mathcal{M}_{\nu_1,\nu_2}} \int_X |x - T(x)|^p \, \nu_1(\mathrm{d}x) \right)^{\frac{1}{p}}, \tag{2.1}$$

where $\mathcal{M}_{\nu_1,\nu_2} = \{T : X \to X \text{ measurable} \,|\, T^\sharp \nu_1 = \nu_2\}$[1] is the set of all measure-preserving maps that rearrange the distribution ν_1 into ν_2.

[1] Here $T^\sharp \nu_1$ denotes the pushforward of ν_1 by the map T, i.e., the measure such that $(T^\sharp \nu_1)(A) = \nu_1(T^{-1}(A))$ for all measurable $A \subseteq X$.

In this paper, we are interested in studying the case of $p = 2$. With a bit of abuse of notation, we also write $W_2(\nu_1, \nu_2) = W_2(f, g)$, where f, g are density functions of the measures ν_1 and ν_2 that are absolutely continuous with respect to the Lebesgue measure: $\mathrm{d}\nu_1 = f(x)\mathrm{d}x$, $\mathrm{d}\nu_2 = g(x)\mathrm{d}x$. The quadratic Wasserstein distance (W_2) has close connections with the negative Sobolev space $\dot{H}^{-1}$ [7, 15, 29]. We will present the most relevant results regarding this paper that clearly illustrate the close connections. In section 4, we will analyze them both as the choice of the likelihood function in Bayesian inversion.

We first introduce the *weighted* L^2, $\dot{H}^1$, and $\dot{H}^{-1}$ norms on a connected set $T \subseteq \mathbb{R}^s$. Given strictly positive probability density $f = d\nu_1$, we can define a Laplace-type linear operator $L : \dot{H}^1 \to \dot{H}^{-1}$,

$$Lh = -\Delta h + \nabla(-\log f) \cdot \nabla h = -\frac{1}{f}\nabla \cdot (f \nabla h), \tag{2.2}$$

which satisfies the fundamental integration-by-parts formula

$$\int_T (Lh_1)h_2 \,\mathrm{d}\nu_1 = \int_T h_1(Lh_2)\,\mathrm{d}\nu_1 = \int_T \nabla h_1 \cdot \nabla h_2 \,\mathrm{d}\nu_1,$$

provided h_1, h_2 have homogeneous Dirichlet or Neumann boundary conditions. Therefore, we can define the weighted Sobolev norms $\|h\|_{L^2(f)}$, $\|h\|_{\dot{H}^1(f)}$ and $\|h\|_{\dot{H}^{-1}(f)}$ for any h that satisfies $\int_T h \,\mathrm{d}\nu_1 = 0$:

$$\|h\|^2_{L^2(f)} = \int_T h^2 \,\mathrm{d}\nu_1, \quad \|h\|^2_{\dot{H}^1(f)} = \int_T |\nabla h|^2 \,\mathrm{d}\nu_1,$$

$$\|h\|^2_{\dot{H}^{-1}(f)} := \sup\left\{ \int_T h\varphi \,\mathrm{d}\nu_1 \;\middle|\; \|\varphi\|_{\dot{H}^1(f)} \le 1 \right\}^2 = \int_T h(L^{-1}h)\,\mathrm{d}\nu_1. = \int_T h(L^{-1}h) f \,\mathrm{d}x.$$

Theorem 2.2 indicates that the linearization of W_2 is *weighted* $\dot{H}^{-1}$, i.e., $W_2(f, g) \approx \|f - g\|_{\dot{H}^{-1}(f)}$ if g is an infinitesimal perturbation of f.

Theorem 2.2 (linearization of W_2 [29, 38]). *For any positive probability density functions f and $g_\varepsilon = (1 + \varepsilon h)f$ on $\mathbb{R}^d$, where $\int_T h d\nu_1 = 0$, then*

$$\|h\|^2_{\dot{H}^{-1}(f)} = \liminf_{\varepsilon \to 0} \frac{W_2(f, g_\varepsilon)^2}{\varepsilon^2}.$$

Remark 2.3. The W_2 distance and the $\dot{H}^{-1}(f)$ norm are quite different metrics if f and g are not close enough in the above linearization regime. For example, $\dot{H}^{-1}(f)$ does not have the global convexity that W_2 has with respect to signal translation and dilation [15].

If Theorem 2.2 states the asymptotic behavior of the W_2 distance (between two probability distributions that are close enough), the next theory gives a relatively global characterization between the negative Sobolev norm $\dot{H}^{-1}(T)$ and the W_2 distance. We remark that here the notation $\dot{H}^{-1}(T)$ represents the familiar $\dot{H}^{-1}$ seminorm for Lebesgue measure, i.e., $f = 1$ in (2.1), instead of the weighted $\dot{H}^{-1}$ norm. The theorem provides a *nonasymptotic* comparison in the sense that g cannot be seen as a small perturbation away from f and it still holds for f and g in any dimension.

Theorem 2.4 (the equivalence between $\dot{H}^{-1}$ and W_2 [30]). *Given any positive probability density functions f and g on T that are bounded from below and from above by constants a, b, where $0 < a < b < +\infty$, i.e., $a < f, g < b$, then*

$$\frac{1}{\sqrt{b}}\|f-g\|_{\dot{H}^{-1}(T)} \le W_2(f,g) \le \frac{1}{\sqrt{a}}\|f-g\|_{\dot{H}^{-1}(T)}.$$

The two theorems above shed light on the local behavior of W_2 by relating it with the well-known Sobolev norms. From Theorem 2.4, we know that if $W_2(f,g)$ is small, then the corresponding negative Sobolev space seminorm $\dot{H}^{-1}(T)$ is also small. On the other hand, Theorem 2.2 conveys the message that once W_2 drives f close enough to g, the distance becomes exactly the *weighted* $\dot{H}^{-1}$ norm.

2.2. Nonlinear seismic inversion in the deterministic setting. For a forward problem, the relationship among the input model parameter u, the forward operator $\mathcal{G}$, and the observable output data y can be expressed as

$$y = \mathcal{G}(u). \tag{2.3}$$

The explicit form of $\mathcal{G}$ can be found in subsection 4.3. The inverse problem of reconstructing u from y typically does not fulfill Hadamard's postulates of well-posedness: there might not be a solution in the strict sense, solutions might not be unique, or a solution might not depend continuously on the data. Even if $\mathcal{G}$ is linear, there is no guarantee that it is invertible. Even if $\mathcal{G}$ is invertible, the computational cost of inverting a dense matrix is prohibitive. An alternative way of formulating the inverse problem is to estimate the true model parameter u^* through the solution of an optimization problem

$$u^* = \arg\min_u J(\mathcal{G}(u), y) + R(u), \tag{2.4}$$

where J is a suitable choice of objective/loss/misfit function characterizing the difference between the data $\mathcal{G}(u)$ generated by the current (and inaccurate) model parameter u and the true observable data y. $R(u)$ here denotes the regularization term to enforce desirable properties on the solution. For example, the Tikhonov regularization $R(u) = ||\Gamma u||_2^2$ with a chosen matrix Γ is a common choice to obtain a smooth reconstruction of u.

The deterministic approach (2.4) is called FWI in exploration geophysics, which is a PDE-constrained optimization by wave equations. The forward operator $\mathcal{G}$ is highly nonlinear for FWI. The conventional objective function J is the least-squares norm (L^2) [39]. The frequency content of the data [26] constrains the accuracy of the model parameter in inversion, and high-frequency data is advantageous to achieve higher resolution in the model recovery. However, the oscillatory and periodic nature of waveforms leads to a primary challenge and results in finding only local minima when using the L^2 norm. The inversion result is also extremely sensitive to high-frequency noise in the data. Numerous works have been done on the subject to deal with these issues. One particular idea is to replace the L^2 norm with other objective functions in optimization for a wider basin of attraction and better stability. Since its first proposal [12], the W_2 distance (2.1) from optimal transport theory has been extensively studied for topics including the convexity, noise robustness, signal normalization,

and fast algorithms for computation in the geophysics community [13, 24, 40]. As a new class of objective functions for FWI, optimal-transport-based metrics are shown to be effective in dealing with local minima issues and are already used in the industry for realistic inversion.

3. The Bayesian approach.

In this section, we outline the Bayesian approach to inversion. One of the motivations for using a Bayesian approach is that there is typically uncertainty in the data, for example, from random observational noise, or from the use of a smoothing/low rank forward model. It then makes sense that uncertainty in the data should be propagated to uncertainty in the solution to the inverse problem. The Bayesian approach combines a probabilistic model for the observed data y, $\mathbb{P}(\mathrm{d}y|u)$, with a probability distribution $\mathbb{P}(\mathrm{d}u)$ representing our prior belief about the unknown u. Bayes's theorem then tells us how to construct the posterior distribution $\mathbb{P}(\mathrm{d}u|y)$ of the unknown given the data: formally, if $\mathbb{P}(\mathrm{d}y|u) = \mathbb{P}(y|u)\,\mathrm{d}y$ admits a Lebesgue density,

$$\mathbb{P}(\mathrm{d}u|y) = \frac{\mathbb{P}(y|u)\mathbb{P}(\mathrm{d}u)}{\mathbb{P}(y)}. \tag{3.1}$$

The *probability measure* $\mathbb{P}(\mathrm{d}u|y)$ is then the solution to the Bayesian inverse problem, rather than a single state as in many classical inversion approaches. This measure can be used to, for example, obtain credible bounds on the solution, or calculate the uncertainty associated with quantities of interest.

Though abstractly the Bayesian approach is quite different from classical approaches, connections between the methods exist. For example, if the prior distribution $\mathbb{P}(\mathrm{d}u) = N(0, C)$ is chosen to be Gaussian, then modes of the posterior distribution will coincide with minimizers of the classical variational problem

$$u^* = \arg\min_u \Phi(u; y) + \frac{1}{2}\langle u, C^{-1}u\rangle. \tag{3.2}$$

Here, $\Phi(u, y) = -\log \mathbb{P}(y|u)$ (the negative log-likelihood) represents the misfit between the observed data and the state u. The operator C will often be chosen as the inverse of a differential operator, so that the regularization term $\langle u, C^{-1}u\rangle$ is a Sobolev-type norm, penalizing (lack of) smoothness of the solution. Equivalently, the Tikhonov regularization in (2.4) embeds the a priori distribution of u. In particular, if the prior distribution $\mathbb{P}(\mathrm{d}u) \propto \exp(-R(u))$ and the likelihood $\mathbb{P}(y|u) \propto \exp(-J(\mathcal{G}(u), y))$, we have $\Phi(u, y) = J(\mathcal{G}(u), y)$ and the two minimization problems in (2.4) and (3.2) coincide.

Alternatively, rather than starting with a probabilistic data model $\mathbb{P}(\mathrm{d}y|u)$, we can begin with a loss function $\Phi(u; y)$ and define the "likelihood" through the relation $\mathbb{P}(\mathrm{d}y|u) = \exp(-\Phi(u; y))\,\mathbb{Q}_0(\mathrm{d}y)$ for some measure $\mathbb{Q}_0$. Note, however, that we then do not necessarily have that

$$\mathbb{P}(Y|u) = \int_Y \exp(-\Phi(u; y))\,\mathbb{Q}_0(\mathrm{d}y) = 1 \quad \text{for all} \quad u \in X, \tag{3.3}$$

and so we do not necessarily have an explicit probabilistic data model. When this is the case, the model is not strictly Bayesian; we will refer to the function $\Phi(u; y)$ as the *potential* and the probability measure $\mathbb{P}(\mathrm{d}u|y)$ *defined* by (3.1) as the *Gibbs posterior* distribution [1].

This is mainly a philosophical distinction, and the mathematical analysis of the family of measures $\mathbb{P}(\mathrm{d}u|y)$ will not be affected by whether we work with posterior or Gibbs posterior distributions.

In the next subsections, we will outline choices of prior distribution appropriate for seismic inversion. We will then consider different data models and the corresponding likelihoods. Finally, we will give the details for rigorously defining the (Gibbs) posterior distribution (3.1) and study its stability with respect to perturbations to the observed data for the different choices of likelihood.

3.1. The prior distribution. In high- and infinite-dimensional Bayesian inverse problems, where the aim is to recover a field, the prior distribution is often chosen to impose properties such as regularity and length-scale on samples. Gaussian priors are often used when continuity of the field is required, and non-Gaussian priors such as Besov and level set priors may be used when this is not desired. In this section, we will outline the details for some choices of prior that will be suitable for seismic inversion.

3.1.1. Gaussian priors. Gaussian distributions are one of the most studied and utilized classes of distributions on function spaces; we outline their definition and provide examples and some key properties. Let $D \subseteq \mathbb{R}^d$ denote a spatial domain, and let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. A random field $u : D \times \Omega \to \mathbb{R}$ is said to be a Gaussian random field if, for any finite collection of points $\{x_j\}_{j=1}^n \subseteq D$, the random vector $(u(x_1, \cdot), \ldots, u(x_n, \cdot))$ is a multivariate Gaussian random variable on $\mathbb{R}^n$. We will often drop the dependence of u on its random argument for cleaner notation.

A useful property of Gaussian random fields is that they are defined completely by their mean function $m : D \to \mathbb{R}$ and covariance function $c : D \times D \to \mathbb{R}$:

$$m(x) = \mathbb{E}(u(x)), \quad c(x, x') = \mathbb{E}(u(x) - m(x))(u(x') - m(x')), \quad x, x' \in D.$$

We will write $u \sim \mathsf{GP}(m(x), c(x, x'))$ when this is the case. When $c(x, x')$ depends only on the Euclidean distance $|x - x'|$, the Gaussian field is said to be isotropic. A common family of isotropic covariance functions used in practice is the Matérn covariance functions, defined by

$$c(x, x') = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{|x - x'|}{\ell} \right) K_\nu \left(\frac{|x - x'|}{\ell} \right)$$

for scalar parameters σ, ν, ℓ representing amplitude, regularity, and length-scale, respectively; here K_ν denotes the modified Bessel function of the second kind. Assuming a smooth mean function m, on regular enough domains, samples from a Gaussian with the covariance function will almost surely possess up to ν Sobolev and Hölder derivatives.

It is often convenient to work with the covariance operator $C : L^2(D) \to L^2(D)$, defined by $C = \mathbb{E}(u - m) \otimes (u - m)$ rather than the covariance function, in which case we will write $u \sim N(m, C)$. They are related by

$$(C\varphi)(x) = \int_D c(x, x')\varphi(x') \, \mathrm{d}x', \quad \varphi \in L^2(D).$$

Hence when the covariance function is a Green's function, the covariance operator is the inverse of the corresponding differential operator. When $D = \mathbb{R}^d$, the covariance operator

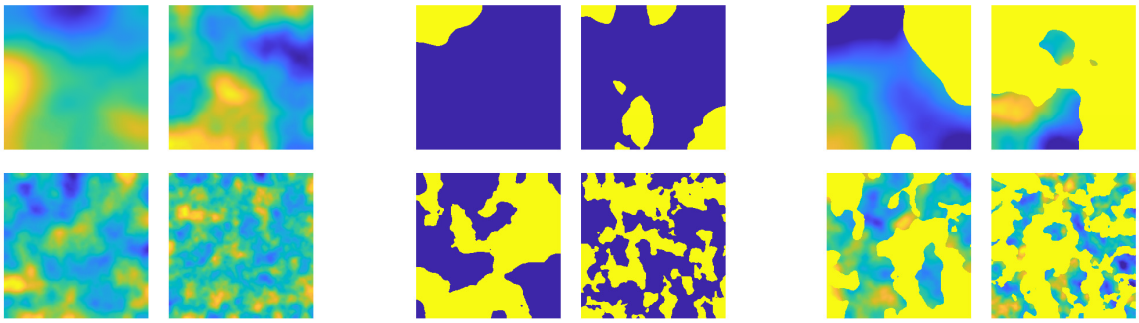

Figure 3.1. *Example of independent samples from the priors discussed in subsection 3.1. (Left) Plain Gaussian prior, (middle) plain level set prior, and (right) mixed level set prior. In all cases the underlying Gaussian random fields have Matérn covariance with the regularity and amplitude parameters fixed, and the length-scale is decreased from left to right, top to bottom within each block.*

corresponding to the Matérn covariance function above is given by

$$C = \sigma^2 \frac{\Gamma(\nu + d/2)(4\pi)^{d/2}\ell^d}{\Gamma(\nu)}(I - \ell^2\Delta)^{-\nu - d/2},$$

where Δ denotes the Laplace operator on $\mathbb{R}^d$. We can therefore generate samples $u \sim N(0, C)$ by solving the (fractional) stochastic PDE

$$(I - \ell^2\Delta)^{\nu/2 + d/4} u = \sigma \sqrt{\frac{\Gamma(\nu + d/2)(4\pi)^{d/2}\ell^d}{\Gamma(\nu)}} W,$$

where W is Gaussian white noise: $W \sim \mathsf{GP}(0, \delta(x - x'))$, or equivalently $W \sim N(0, I)$.

In Figure 3.1, the left block depicts examples of samples of Gaussian fields with Matérn covariance functions for fixed σ, ν, and ℓ decreasing from left to right, top to bottom.

3.1.2. Level-set type priors. Gaussian random fields, such as those with Matérn covariance as described above, typically have global smoothness properties associated with their samples. In some situations, such as classification problems or inference of salt models discussed later, piecewise constant or piecewise continuous samples may be desired instead. One method for producing such fields is to write them as nonlinear transformations of Gaussian fields. For example, given a Gaussian measure $\nu_0 = N(m, C)$ and scalar values $u_+, u_- \in \mathbb{R}$, one could define a prior measure by the pushforward

$$\pi_0 = F^\sharp \nu_0, \quad F(v)(x) = u_+ \mathbb{1}_{v(x) > 0} + u_- \mathbb{1}_{v(x) \leq 0}. \tag{3.4}$$

That is, π_0 is the law of the thresholded Gaussian field $F(v), v \sim \nu_0$: samples from π_0 take the values u_+, u_- almost everywhere, with interface between the values given by the level set $\{v(x) = 0\}$. Such priors have been studied previously from a nonparametric Bayesian perspective [10, 19].

Alternatively, one may desire some combination of the above and plain Gaussian priors. For example, given a product Gaussian measure $\nu_0 = N(m_1, C_1) \times N(m_2, C_2)$, one could define a mixed level set prior:

$$\pi_0 = F^\sharp \nu_0, \quad F(v, w)(x) = u_+ \mathbb{1}_{v(x)>0} + w(x)\mathbb{1}_{v(x)\leq 0}. \tag{3.5}$$

Examples of samples from the above two priors are shown in the middle and right blocks of Figure 3.1 respectively, with the length-scale of all underlying Gaussian fields decreasing from left to right, top to bottom. These examples may be generalized further to multiple interfaces, for example, by using the vector level set method [4]. Additionally, they may be generalized to allow for uncertainty in the values u_+, u_- using a hierarchical method [8]; such priors will be considered numerically in section 5.

Remark 3.1. It may be convenient, both theoretically and numerically, to view the prior as ν_0 and compose the potential Φ with the map F. This is sometimes referred to as non-centering, and we will use this when establishing existence and well-posedness of the posterior distributions when using the level-set type priors. It can be checked, using the definition of the pushforward measure, that the stability estimates obtained are invariant under this change of parameterization.

3.2. Likelihood and loss functions. In this subsection, we outline several likelihood functions and loss functions that are the main focus of this paper. We first consider some natural likelihood functions corresponding to explicit data models, which lead to Bayesian posterior distributions. We then consider two loss functions used in variational inversion approaches, which lead to Gibbs posteriors. In particular, we discuss the noise models that correspond to these posteriors heuristically. In what follows, we assume that the data is a function defined on a spatial domain $D \subseteq \mathbb{R}^d$ equipped with a finite measure λ and on temporal domain $T \subseteq \mathbb{R}^s$ equipped with the Lebesgue measure. Note that this setup allows us to consider model setups with either discrete or continuous spatial measurements simultaneously—for example, D may be a discrete set equipped with the counting measure, or it may be a continuous set equipped with the Lebesgue or Hausdorff measure.

3.2.1. Explicit likelihood functions. We first consider the simplest case wherein the loss $\Phi(u; y) = J(\mathcal{G}(u), y)$ is given by the L^2 misfit:

$$\Phi_{L^2}(u; y) = \frac{1}{2} \int_D \|\mathcal{G}(u)(x, \cdot) - y(x, \cdot)\|^2_{L^2(T)} \, \lambda(\mathrm{d}x).$$

This loss arises as a negative log-likelihood by assuming that the data is corrupted by additive Gaussian space-time white noise:

$$y = \mathcal{G}(u) + \eta, \quad \eta \sim N(0, I).$$

Note that under this model, if $\dim(Y) = \infty$, the data y will almost surely not be valued in L^2 due to the roughness of the noise: we will have $y(x, \cdot) \in H^{-r}(T)$ almost surely for all $r > s/2$ but not $r = s/2$. To see how this leads to the above loss, observe that $y|u \sim \mathbb{Q}_u := N(\mathcal{G}(u), I)$.

Defining $\mathbb{Q}_0 = N(0, I)$, we have by the Cameron–Martin theorem that

$$\begin{aligned}\frac{\mathrm{d}\mathbb{Q}_u}{\mathrm{d}\mathbb{Q}_0}(y) &= \exp\left(-\frac{1}{2}\|\mathcal{G}(u)\|^2_{L^2(D;L^2(T))} + \langle y, \mathcal{G}(u)\rangle_{L^2(D;L^2(T))}\right) \\ &= \exp\left(-\frac{1}{2}\|\mathcal{G}(u) - y\|^2_{L^2(D;L^2(T))} + \frac{1}{2}\|y\|^2_{L^2(D;L^2(T))}\right)\end{aligned}$$

since the Cameron–Martin space of white noise is L^2. For fixed y, if $\dim(Y) < \infty$ or $y \in L^2$, we may drop the second term in the exponent when defining the loss—its appearance here ensures that the exponent is finite almost surely under the noise model.[2]

Another loss function that has been considered is the $\dot{H}^{-1}$ loss[3]

$$\Phi_{\dot{H}^{-1}}(u; y) = \frac{1}{2}\int_D \|\mathcal{G}(u)(x,\cdot) - y(x,\cdot)\|^2_{\dot{H}^{-1}(T)}\,\lambda(\mathrm{d}x).$$

This arises from a data model similar to the above, except instead of assuming that the noise is white, temporal correlations are assumed:

$$y = \mathcal{G}(u) + \eta, \quad \eta \sim N(0, \Gamma),$$

where $\Gamma = -\Delta_T$ is the negative Laplacian on the temporal variable. Similarly to the above we have $y|u \sim \mathbb{Q}_u := N(\mathcal{G}(u), \Gamma)$, and so defining $\mathbb{Q}_0 = N(0, \Gamma)$ we have that

$$\frac{\mathrm{d}\mathbb{Q}_u}{\mathrm{d}\mathbb{Q}_0}(y) = \exp\left(-\frac{1}{2}\|\mathcal{G}(u) - y\|^2_\Gamma + \frac{1}{2}\|y\|^2_\Gamma\right),$$

where $\langle\cdot,\cdot\rangle_\Gamma := \langle\cdot, \Gamma^{-1}\cdot\rangle_{L^2(D;L^2(T))} = \langle\cdot,\cdot\rangle_{L^2(D;\dot{H}^{-1}(T))}$ denotes the Cameron–Martin inner product associated with $\mathbb{Q}_0$. Note that in this case the noise will be even rougher in time than the white noise case: since the eigenvalues $\{\lambda_j\}$ of the Laplacian on T asymptotically satisfy $\lambda_j \asymp j^{2/s}$ by a Weyl-type law, we have that if $\eta \sim N(0, \Gamma)$,

$$\mathbb{E}\|\eta(x,\cdot)\|^2_{H^{-r}(T)} \asymp \sum_{j=1}^{\infty} j^{-2r/s+2/s} < \infty \quad \text{iff } r > 1 + \frac{s}{2},$$

using the Karhunen–Loève expansion of $\eta(x,\cdot)$. As in the white noise case, we may drop the second term if $\dim(Y) < \infty$ or if $y \in L^2(D; \dot{H}^{-1}(T))$.

3.2.2. Loss functions and approximate noise models. The loss functions above correspond to negative log-likelihoods from explicit data models, which will lead to Bayesian posteriors if coupled with prior distributions. However, this is not the case for general loss functions, as discussed at the beginning of section 3. Given a general loss function $\Phi(u; y)$ and

[2] In practice for the FWI problem it will be the case that $\dim(Y) < \infty$, since only a finite number of frequencies will be observed at a finite number of receivers.

[3] Note that $\|y(x,\cdot)\|_{\dot{H}^{-1}(T)}$ is well-defined only if $\int_T y(x,t)\,\mathrm{d}t = 0$. After data preprocessing, zero-frequency in time signals are removed from FWI data $y(x,\cdot)$. The mean-zero property makes $\dot{H}^{-1}$ a proper data discrepancy for comparing seismic data.

a measure $\mathbb{Q}_0$ on Y such that $\exp(-\Phi(u;y))$ is integrable with respect to $\mathbb{Q}_0$, one can define the normalized loss function $\tilde{\Phi}(u;y)$ by

$$\tilde{\Phi}(u;y) = \Phi(u;y) - \log \int_Y \exp\left(-\Phi(u;z)\right) \mathbb{Q}_0(\mathrm{d}z)$$

so that the relation (3.3) holds for $\tilde{\Phi}(u;y)$. The corresponding data-generating distribution would then be given by $\mathbb{P}(\mathrm{d}y|u) = \exp(-\tilde{\Phi}(u;y))\,\mathbb{Q}_0(\mathrm{d}y)$. However, performing this normalization may be intractable if it cannot be done analytically. Therefore, it may be preferable to work with the unnormalized loss function. In this subsection, we consider an unnormalized multiplicative noise loss function and a Wasserstein loss function. We will show that the latter could be viewed asymptotically as an unnormalized state-dependent multiplicative noise loss in the small noise limit.

In the above additive noise models, the loss function scale was dictated by the size of the noise on the observations. In the absence of an explicit data model, one needs to be careful about choosing the scale for the potential. Typically, one introduces a scalar parameter $\beta > 0$, often referred to as the inverse temperature, and works with $\beta\Phi(u;y)$ in place of $\Phi(u;y)$. The choice of this parameter is known as calibration and may either be chosen empirically [36] or treated as a hyperparameter as part of the inverse problem [43]. We do not discuss the choice of β here and assume it to be fixed.

We first introduce the Wasserstein loss function. The relevant background is presented in subsection 2.1. To evaluate the quadratic Wasserstein distance (W_2) between the reference data and the output from the forward map, we must first transform the datasets into probability densities with respect to the temporal variable. Hence, given a scalar function $\sigma : \mathbb{R} \to \mathbb{R}^+$, we define the normalization operator P_σ on functions $y : D \times T \to \mathbb{R}$ by

$$(P_\sigma y)(x,t) = \frac{1}{Z_\sigma(x)}\sigma(y(x,t)), \quad Z_\sigma(x) = \int_T \sigma(y(x,t'))\,\mathrm{d}t'. \tag{3.6}$$

Given this operator, we then define the Wasserstein loss by

$$\Phi_{W_2}(u;y) = \frac{1}{2}\int_D W_2\left((P_\sigma\mathcal{G}(u))(x,\cdot), (P_\sigma y)(x,\cdot)\right)^2 \lambda(\mathrm{d}x).$$

Note that Φ_{W_2} is not normalized in the sense described above and does not appear to correspond to a particular data model even if it were normalized. However, via linearization of the W_2 distance, we can describe an approximate data model that this loss corresponds to in the limit of small observational noise. Assume that there is $\eta : D \times T \to \mathbb{R}$ such that

$$P_\sigma y = (1+\eta)P_\sigma\mathcal{G}(u), \quad \int_T \eta(x,t)(P_\sigma\mathcal{G}(u))(x,t)\,\mathrm{d}t = 0 \quad \text{for all} \quad x \in D.$$

Then by Theorem 2.2 we may approximate for small $\|\eta(x,\cdot)\|_{\dot{H}^{-1}(P_\sigma\mathcal{G}(u))}$

$$\begin{aligned}\Phi_{W_2}(u;y) &\coloneqq \frac{1}{2}\int_D W_2\left((P_\sigma\mathcal{G}(u))(x,\cdot), (P_\sigma y)(x,\cdot)\right)^2 \lambda(\mathrm{d}x)\\ &= \frac{1}{2}\int_D W_2\Big((P_\sigma\mathcal{G}(u))(x,\cdot), \big(1+\eta(x,\cdot)\big)(P_\sigma\mathcal{G}(u))(x,\cdot)\Big)^2 \lambda(\mathrm{d}x)\end{aligned}$$

$$\approx \frac{1}{2}\int_D \|\eta(x,\cdot)\|^2_{\dot{H}^{-1}(P_\sigma\mathcal{G}(u))}\,\lambda(\mathrm{d}x) \quad \text{(by Theorem 2.2 with } \varepsilon h = \eta\text{)}$$
$$= \frac{1}{2}\int_D \left\| \frac{(P_\sigma y)(x,\cdot) - P_\sigma\mathcal{G}(u))(x,\cdot)}{(P_\sigma\mathcal{G}(u))(x,\cdot)} \right\|^2_{\dot{H}^{-1}(P_\sigma\mathcal{G}(u))} \lambda(\mathrm{d}x).$$

The above potential formally corresponds to the negative logarithm of an unnormalized Gaussian density $N(1,\mathcal{L}(u))$, for some operator $\mathcal{L}(u)$ defined below, evaluated at the ratio $P_\sigma y/P_\sigma\mathcal{G}(u)$. It suggests that the Wasserstein loss could formally be considered as asymptotically coming from the state-dependent multiplicative noise data model

$$P_\sigma y = (1+\eta)\,P_\sigma\mathcal{G}(u), \quad \eta|u \sim N(0,\mathcal{L}(u)),$$

where the operator $\mathcal{L}(u) : \mathrm{Dom}(\mathcal{L}(u)) \to L^2(D;L^2(T))$ is defined by

$$\mathcal{L}(u)\varphi = -\frac{1}{P_\sigma\mathcal{G}(u)}\nabla_T \cdot \left(P_\sigma\mathcal{G}(u)\nabla_T \left(\frac{\varphi}{P_\sigma\mathcal{G}(u)} \right) \right),$$
$$\mathrm{Dom}(\mathcal{L}(u)) = \left\{ \varphi \in L^2(D;H^2(T)) \,\middle|\, \int_T \varphi(x,t)(P_\sigma\mathcal{G}(u))(x,t)\,\mathrm{d}t = 0 \text{ for all } x \in D \right\}$$

and ∇_T denotes the gradient with respect to the temporal variable; this operator is derived from the definition of the weighted $\dot{H}^{-1}$ norm and the operator (2.2). Note that subject to appropriate boundedness and regularity of $P_\sigma\mathcal{G}(u)$, samples $\eta \sim N(1,\mathcal{L}(u))$ will have the same negative Sobolev regularity as samples $\eta \sim N(0,\Gamma)$ considered in subsection 3.2.1.

The relation of Wasserstein loss to a multiplicative noise model is interesting. The latter has been studied in the literature of Bayesian inverse problems [20] before the introduction of the Wasserstein loss for FWI [13]. One may consider the following unnormalized multiplicative noise model [18]:

$$\Phi_M(u;y) = \frac{1}{2}\int_D \left\| \frac{(P_\sigma\mathcal{G}(u))(x,\cdot) - (P_\sigma y)(x,\cdot)}{(P_\sigma y)(x,\cdot)} \right\|^2_{L^2(T)} \lambda(\mathrm{d}x),$$

which can be viewed as arising from the model

$$P_\sigma y = \eta \cdot P_\sigma\mathcal{G}(u), \quad 1/\eta \sim N(1,I).$$

Alternatively, one may formulate it by informally assuming the noise variance is proportional to the size of the observed data. In general, the data and the output of the forward map do not need to be probability densities in this model. However, for the stability of the resulting posterior, the forward map must be bounded away from zero; see [9] for a discussion. The condition can be ensured by using the same operator P_σ, and we do so in the following section for brevity. Additionally, one may prefer to use a model wherein $\eta \sim N(1,I)$ rather than its reciprocal, in which case $P_\sigma\mathcal{G}(u)$ will replace $P_\sigma y$ in the denominator; we do not provide details of the proofs for this modification but note that they are similar and slightly simpler than without this modification.

3.3. The (Gibbs) posterior distribution. In this section, we establish both that the (Gibbs) posterior measures corresponding to the likelihoods/potentials discussed above are well-defined (existence), and that these measures are stable with respect to perturbations of the observed data (well-posedness). The existence is established using the theory/assumptions of [35], which is a generalization of the result in [6] that provides an abstract statement of Bayes's theorem at the level of measures. Given a prior measure π_0 on a space X and a potential Φ of the form considered in the previous subsection, we show that the measure π_Φ^y defined by

$$\pi_\Phi^y(\mathrm{d}u) = \frac{1}{Z_\Phi(y)} \exp(-\Phi(u;y))\, \pi_0(\mathrm{d}u), \quad Z_\Phi(y) = \int_X \exp(-\Phi(u;y))\, \pi_0(\mathrm{d}u),$$

defines a Radon probability measure on X. The definition of π_Φ^y is essentially a restatement of Bayes's theorem as given in (3.1).

Stability of the posterior will be established with respect to the *Hellinger* distance d_{H} on probability measures:

$$d_{\mathrm{H}}(\pi, \pi')^2 = \int_X \left(\sqrt{\frac{\mathrm{d}\pi}{\mathrm{d}\nu_0}(u)} - \sqrt{\frac{\mathrm{d}\pi'}{\mathrm{d}\nu_0}(u)} \right)^2 \nu_0(\mathrm{d}u),$$

where the measure ν_0 is such that both π, π' are absolutely continuous with respect to ν_0.[4] That is, we show if a perturbation is made to the observed data, then the Hellinger distance between the resulting posteriors is bounded above by some norm of the perturbation. We, in particular, show that for the Wasserstein and $\dot{H}^{-1}$ choices of likelihood, if the observed data is perturbed, then the Hellinger distance between the corresponding posteriors is bounded above by the $\dot{H}^{-1}$ norm of this perturbation. This is in contrast to the choice of L^2 likelihood, in which it is only bounded by the L^2 norm of the perturbation. This essentially shows that the Wasserstein and $\dot{H}^{-1}$ likelihoods lead to posterior distributions, which are *more robust* with respect to high-frequency noise on the data.

A useful property of the Hellinger distance is that it allows us to bound expectations of quantities of interest:

$$\|\mathbb{E}^{\pi}(f) - \mathbb{E}^{\pi'}(f)\|_S \leq C(f) d_{\mathrm{H}}(\pi, \pi')$$

for any $f \in L^2(X, \pi; S) \cap L^2(X, \pi'; S)$. For example, the choice $f(u) = u$ provides us with stability of the posterior means with respect to the data.

4. Existence and well-posedness of the (Gibbs) posterior. In this section, we provide assumptions that lead to the existence and well-posedness of the (Gibbs) posteriors arising from the likelihoods and priors in the previous section. We then show the applicability of the theory to the full-waveform inversion problem under a large class of prior distributions.

[4] This definition can be seen to be independent of the choice of ν_0. Such a measure ν_0 may always be found in practice, for example, $\nu_0 = \frac{1}{2}(\pi + \pi')$; however, in our setup we may always take it to be the prior measure $\nu_0 = \pi_0$.

4.1. Notation, assumptions, and supporting lemmas. In what follows $(X, \|\cdot\|_X)$ will denote a separable Banach space. $(D, \mathcal{D}, \lambda)$ will denote a finite measure space representing the spatial domain, and $T \subset \mathbb{R}^p$ will denote a compact subset of Euclidean space, representing the temporal domain. Any L^p space over D will be with respect to λ, and any L^p space over T will be with respect to the Lebesgue measure. We define the following subsets of $L^2(D; L^2(T))$:

$$Y_0 = \left\{ y \in L^\infty(D; L^\infty(T)) \,\middle|\, \int_T y(x,t)\,\mathrm{d}t = 0 \text{ for all } x \in D \right\},$$
$$Y_{1,\varepsilon} = \left\{ \rho \in L^\infty(D; L^\infty(T)) \,\middle|\, \int_T \rho(x,t)\,\mathrm{d}t = 1 \text{ for all } x \in D,\ \varepsilon \le \rho \le \varepsilon^{-1} \right\}, \quad \varepsilon > 0.$$

Thus, Y_0 is a set of bounded functions with no zero-frequency component, and $Y_{1,\varepsilon}$ is a set of probability densities bounded above and below by positive constants. On any set $Y_{1,\varepsilon}$, we have the following equivalence of metrics, which follows directly from Theorem 2.4.

Lemma 4.1. *Let $\varepsilon > 0$; then for any $\rho, \rho' \in Y_{1,\varepsilon}$,*

$$\varepsilon \|\rho - \rho'\|^2_{L^2(D;\dot{H}^{-1}(T))} \le \int_D W_2(\rho(x,\cdot), \rho'(x,\cdot))^2 \,\lambda(\mathrm{d}x) \le \varepsilon^{-1} \|\rho - \rho'\|^2_{L^2(D;\dot{H}^{-1}(T))}.$$

Given a map $\sigma : \mathbb{R} \to \mathbb{R}_+$ we recall the normalization operator P_σ (3.6). Under appropriate assumptions on σ, we have the following result.

Lemma 4.2. *Let $\sigma : \mathbb{R} \to \mathbb{R}_+$ be locally Lipschitz, and given $r > 0$ let $y, y' \in Y_0$ with $\|y\|_{L^\infty(D;L^\infty(T))}, \|y'\|_{L^\infty(D;L^\infty(T))} < r$. Then there exists $\varepsilon(r) > 0$ such that $P_\sigma y, P_\sigma y' \in Y_{1,\varepsilon(r)}$, and $L_\sigma(r)$ such that*

$$\|P_\sigma y - P_\sigma y'\|_{L^2(D;L^2(T))} \le L_\sigma(r) \|y - y'\|_{L^2(D;L^2(T))},$$
$$\|P_\sigma y - P_\sigma y'\|_{L^2(D;\dot{H}^{-1}(T))} \le L_\sigma(r) \|y - y'\|_{L^2(D;\dot{H}^{-1}(T))}.$$

Proof. The map $\sigma : \mathbb{R} \to \mathbb{R}_+$ is locally Lipschitz and hence locally bounded. It follows that since $|y|, |y'| < r$ almost everywhere, there exists $k(r) > 0$ such that $k(r) \le \sigma(y(x,t)), \sigma(y'(x,t)) \le k(r)^{-1}$ for almost all $(x,t) \in D \times T$, and so $P_\sigma y, P_\sigma y' \in Y_{1,\varepsilon(r)}$ for some $\varepsilon(r) > 0$ by construction of the map P_σ.

With $Z = L^2(T)$ or $Z = \dot{H}^{-1}(T)$, we have that for all $x \in D$,

$$\|(P_\sigma y)(x,\cdot) - (P_\sigma y')(x,\cdot)\|_Z = \sup\left\{ \int_T |(P_\sigma y)(x,t) - (P_\sigma y')(x,t)| \varphi(t)\,\mathrm{d}t \,\middle|\, \|\varphi\|_{Z^*} \le 1 \right\}.$$

Again using the boundedness $|y|, |y'| < r$, we have $\int_T \sigma(y(x,t))\,\mathrm{d}t, \int_T \sigma(y'(x,t))\,\mathrm{d}t \ge |T| k(r) > 0$, and so for all $x \in D$

$$\begin{aligned}
\|(P_\sigma y)(x,\cdot) - (P_\sigma y')(x,\cdot)\|_Z &\le \frac{1}{|T|k(r)} \sup\left\{ \int_T |\sigma(y(x,t)) - \sigma(y'(x,t))| \varphi(t)\,\mathrm{d}t \,\middle|\, \|\varphi\|_{Z^*} \le 1 \right\} \\
&\le \frac{L(r)}{|T|k(r)} \sup\left\{ \int_T |y(x,t) - y'(x,t)| \varphi(t)\,\mathrm{d}t \,\middle|\, \|\varphi\|_{Z^*} \le 1 \right\} \\
&= \frac{L(r)}{|T|k(r)} \|y - y'\|_Z,
\end{aligned}$$

where $L(r)$ is the local Lipschitz coefficient of σ. Squaring and integrating over D gives the result. ∎

In the theory that follows we will make use of the following assumptions.

Assumptions 4.3. The data y, forward map $\mathcal{G}$, and prior π_0 satisfy the following:

(i) $y \in Y_0$;

(ii) $\mathcal{G} : X \to Y_0$ is continuous with respect to the $L^2(D; L^2(T))$ norm π_0-almost surely; and

(iii) there exists an increasing function $R_{\mathcal{G}} : \mathbb{R}_+ \to \mathbb{R}_+$ such that $\|\mathcal{G}(u)\|_{L^\infty(D;L^\infty(T))} \leq R_{\mathcal{G}}(\|u\|_X)$ for all $u \in X$.

4.2. Existence and well-posedness. We recall the four loss functions $\Phi_\cdot : X \times Y \to \mathbb{R}$ introduced in section 3 and defined by (1.1)–(1.4). In the cases of Φ_{L^2} and Φ_M we equip the data space $Y = Y_0$ with the $L^2(D; L^2(T))$ norm, and for $\Phi_{H^{-1}}$ and Φ_{W_2} we equip it with the $L^2(D; \dot{H}^{-1}(T))$ norm. We may then establish existence and well-posedness of the corresponding (Gibbs) posterior distributions.

Theorem 4.4 (existence). *Let π_0 be a Borel probability measure on X, and let Assumptions 4.3 hold. Then for any choice $\Phi \in \{\Phi_{L^2}, \Phi_{H^{-1}}, \Phi_M, \Phi_{W_2}\}$,*

$$Z_\Phi(y) = \int_X \exp(-\Phi(u; y))\, \pi_0(\mathrm{d}u)$$

is strictly positive and finite, and

$$\pi_\Phi^y(\mathrm{d}u) := \frac{1}{Z_\Phi(y)} \exp\left(-\Phi(u; y)\right)\, \pi_0(\mathrm{d}u)$$

defines a Radon probability measure on X.

Proof. We show that the assumptions of Theorem 4.3 in [35] are satisfied for each choice of Φ. Note that in [35], the data space is assumed to be a quasi-Banach space, and so in particular complete; however, studying the proofs therein, one finds that this completeness property is never used—an incomplete norm structure on the data space suffices to obtain the same results. This is important for our applications, as we equip Y with different incomplete norms. We also note that continuity of the likelihood in the data component is not required for the proof of existence, either by examining the proof in [35] or using Theorem 2.4 in [23].

We must establish the following properties for each choice of Φ:

(i) $\Phi(\cdot; y)$ is measurable for each $y \in Y$.

(ii) For each $r > 0$ there exists $M_{0,r} \in \mathbb{R}$ such that for each $u \in X$, $y \in Y$ with $\|u\|_X, \|y\|_Y < r$, $|\Phi(u; y)| \leq M_{0,r}$.

(iii) For each $r > 0$ there exists a measurable $M_{1,r} : \mathbb{R}_+ \to \mathbb{R}$ such that for each $u \in X$, $y \in Y$ with $\|y\|_Y < r$, $\Phi(u; y) \geq M_{1,r}(\|u\|_X)$, and

$$\int_X \exp(-M_{1,r}(\|u\|_X))\, \pi_0(\mathrm{d}u) < \infty.$$

The latter is trivially satisfied for any probability measure π_0 by all choices of Φ considered with $M_{1,r}(\|u\|) \equiv 0$. We verify the other two properties for each Φ in turn.

$\boxed{\Phi_{L^2}}$ (i) The map $\mathcal{G}$ is assumed to be π_0-a.s. continuous from X into Y_0 equipped with the $L^2(D;L^2(T))$ topology, and the $L^2(D;L^2(T))$ norm is continuous with respect to its own topology, so measurability follows.

(ii) Fix $r>0$ and choose $u\in X$, $y\in Y$ with $\|u\|_X, \|y\|_Y < r$. Then by assumption $\|\mathcal{G}(u)\|^2_{L^2(D;L^2(T))} \le \lambda(D)|T|\|\mathcal{G}(u)\|_{L^\infty(D;L^\infty(T))} \le \lambda(D)|T|R_{\mathcal{G}}(r)^2$. We may then bound

$$\begin{aligned}|\Phi_{L^2}(u;y)| &\le \frac{1}{2}(\|\mathcal{G}(u)\|_{L^2(D;L^2(T))} + \|y\|_{L^2(D;L^2(T))})^2\\ &\le \frac{1}{2}\lambda(D)|T|(\|\mathcal{G}(u)\|_{L^\infty(D;L^\infty(T))} + \|y\|_{L^\infty(D;L^\infty(T))})^2\\ &\le \frac{1}{2}\lambda(D)|T|(R_{\mathcal{G}}(r)+r)^2 =: M_{0,r}.\end{aligned}$$

$\boxed{\Phi_{H^{-1}}}$ (i) The map $\mathcal{G}$ is assumed to be π_0-a.s. continuous from X into Y_0 equipped with the $L^2(D;L^2(T))$ topology, and the embedding of $(Y_0, \|\cdot\|_{L^2(D;\dot{L}^2(T))})$ into $(Y_0, \|\cdot\|_{L^2(D;\dot{H}^{-1}(T))})$ is continuous. Since the $L^2(D;\dot{H}^{-1}(T))$ norm is continuous with respect to its own topology, measurability follows.

(ii) Choose such $u\in X$, $y\in Y$. Then we obtain a similar bound $M_{0,r}$ as for Φ_{L^2}, using that

$$\|\mathcal{G}(u)\|^2_{L^2(D;\dot{H}^{-1}(T))} \le C_p(T)\|\mathcal{G}(u)\|^2_{L^2(D;L^2(T))} \le C_p(T)\lambda(D)|T|R_{\mathcal{G}}(r)^2,$$

where $C_p(T)$ is the Poincaré constant of the domain T.

$\boxed{\Phi_M}$ (i) By assumption $\mathcal{G}$ is π_0-a.s. continuous into Y_0 equipped with the $L^2(D;L^2(T))$ norm. By Lemma 4.2 the map P_σ is continuous, as is the mapping $\rho \mapsto (\rho - P_\sigma y)/P_\sigma y$ since $P_\sigma y > 0$ for all $y\in Y_0$. Measurability then follows by continuity of the $L^2(D;L^2(D))$ norm.

(ii) Choose such $u\in X$, $y\in Y$. Then $\|\mathcal{G}(u)\|_{L^\infty(D;L^\infty(T))} \le R_{\mathcal{G}}(r)$; without loss of generality assume $R_{\mathcal{G}}(r)\ge r$. By Lemma 4.2, there exists $\varepsilon(r)>0$ such that $P_\sigma(y)\ge\varepsilon(r)$, and so

$$\begin{aligned}|\Phi_M(u;y)| &\le \frac{1}{2}\varepsilon(r)^{-2}\|P_\sigma\mathcal{G}(u) - P_\sigma y\|^2_{L^2(D;L^2(T))}\\ &\le \frac{1}{2}\varepsilon(r)^{-2}L_\sigma(R_{\mathcal{G}}(r))^2\|\mathcal{G}(u)-y\|^2_{L^2(D;L^2(T))}.\end{aligned}$$

The bound now follows from the bound for Φ_{L^2}.

$\boxed{\Phi_W}$ (i) This follows from Lemma 4.1 and the above arguments for $\Phi_{\dot{H}^{-1}}$.

(ii) Choose such $u\in X$, $y\in Y$. Then $\|\mathcal{G}(u)\|_{L^\infty(D;L^\infty(T))} \le R_{\mathcal{G}}(r)$; again without loss of generality assume $R_{\mathcal{G}}(r)\ge r$. Using Lemma 4.2, there exist $\varepsilon(r)>0, \tilde{\varepsilon}(r)>0$ such that $P_\sigma y \in Y_{1,\varepsilon(r)}, P_\sigma\mathcal{G}(u)\in Y_{1,\tilde{\varepsilon}(r)}$. We may therefore use Lemmas 4.1 and 4.2 again to bound

$$\begin{aligned}|\Phi_{W_2}(u;y)| &\le \frac{1}{2}(\varepsilon(r)\wedge\tilde{\varepsilon}(r))^{-1}\|P_\sigma\mathcal{G}(u) - P_\sigma y\|^2_{L^2(D;\dot{H}^{-1}(T))}\\ &\le \frac{1}{2}(\varepsilon(r)\wedge\tilde{\varepsilon}(r))^{-1}L_\sigma(R_{\mathcal{G}}(r))^2\|\mathcal{G}(u)-y\|^2_{L^2(D;\dot{H}^{-1}(T))}\end{aligned}$$

$$\leq \frac{1}{2}(\varepsilon(r) \wedge \tilde{\varepsilon}(r))^{-1} L_\sigma(R_{\mathcal{G}}(r))^2 \lambda(D)|T| \|\mathcal{G}(u) - y\|^2_{L^\infty(D;L^\infty(T))}$$
$$\leq \frac{1}{2}(\varepsilon(r) \wedge \tilde{\varepsilon}(r))^{-1} L_\sigma(R_{\mathcal{G}}(r))^2 \lambda(D)|T|(R_{\mathcal{G}}(r) + r)^2 =: M_{0,r},$$

where $a \wedge b$ denotes $\min\{a, b\}$. ∎

The following analysis will demonstrate the regularity of the posterior with respect to perturbations of the data. First, we assert the following lemma regarding the local Lipschitz properties of the potentials that are considered here.

Lemma 4.5. *Let Assumption 4.3(iii) hold and choose* $\Phi \in \{\Phi_{L^2}, \Phi_{H^{-1}}, \Phi_M, \Phi_{W_2}\}$. *For each* $r > 0$ *there exists* $M_{2,r} : \mathbb{R}_+ \to \mathbb{R}_+$ *such that for each* $u \in X$, $y, y' \in Y$ *with* $\|y\|_{L^\infty(D;L^\infty(T))}, \|y'\|_{L^\infty(D;L^\infty(T))} < r$,

$$|\Phi(u; y) - \Phi(u; y')| \leq M_{2,r}(\|u\|_X)\|y - y'\|_Y.$$

Proof. We verify for each choice of Φ in turn; given such a Φ, fix $r > 0$, and choose $u \in X$, $y, y' \in Y$ with $\|y\|_{L^\infty(D;L^\infty(T))}, \|y'\|_{L^\infty(D;L^\infty(T))} < r$.

$\boxed{\Phi_{L^2}}$ We have that $\|\cdot\|_Y = \|\cdot\|_{L^2(D;L^2(T))}$. We may calculate, using the assumption,

$$\begin{aligned}
|\Phi_{L^2}(u; y) - \Phi_{L^2}(u; y')| &= \left|\langle \mathcal{G}(u), y'\rangle_Y - \langle \mathcal{G}(u), y\rangle_Y + \frac{1}{2}\|y\|_Y^2 - \frac{1}{2}\|y'\|_Y^2\right| \\
&\leq |\langle \mathcal{G}(u), y - y'\rangle_Y| + \frac{1}{2}\left|\|y\|_Y^2 - \|y'\|_Y^2\right| \\
&\leq \|\mathcal{G}(u)\|_Y\|y - y'\|_Y + \frac{1}{2}(\|y\|_Y + \|y'\|_Y)\|y - y'\|_Y \\
&\leq \lambda(D)^{1/2}|T|^{1/2}(R_{\mathcal{G}}(\|u\|_X) + r)\|y - y'\|_Y,
\end{aligned}$$

and so we may take $M_{2,r}(\|u\|_X) = \lambda(D)^{1/2}|T|^{1/2}(R_{\mathcal{G}}(\|u\|_X) + r)$.

$\boxed{\Phi_{H^{-1}}}$ We have that $\|\cdot\|_Y = \|\cdot\|_{L^2(D;\dot{H}^{-1}(T))}$. We may calculate similarly to the L^2 case above to deduce that we may take

$$M_{2,r}(\|u\|_X) = C_p(T)^{1/2}\lambda(D)^{1/2}|T|^{1/2}(R_{\mathcal{G}}(\|u\|_X) + r),$$

where $C_p(T)$ is the Poincaré constant of the domain T.

$\boxed{\Phi_M}$ We have that $\|\cdot\|_Y = \|\cdot\|_{L^2(D;L^2(T))}$. Using Lemma 4.2, we have that $P_\sigma y, P_\sigma y' \in Y_{1,\varepsilon(r)}$. Using the assumption, we may then calculate, again using Lemma 4.2,

$$\begin{aligned}
|\Phi_M(u; y) - \Phi_M(u; y')| &= \frac{1}{2}\left|\left\|\frac{P_\sigma\mathcal{G}(u) - P_\sigma y}{P_\sigma y}\right\|_Y^2 - \left\|\frac{P_\sigma\mathcal{G}(u) - P_\sigma y'}{P_\sigma y'}\right\|_Y^2\right| \\
&= \frac{1}{2}\left|\left\|\frac{P_\sigma\mathcal{G}(u)}{P_\sigma y}\right\|_Y^2 - \left\|\frac{P_\sigma\mathcal{G}(u)}{P_\sigma y'}\right\|_Y^2 + 2\left\langle\frac{P_\sigma\mathcal{G}(u)}{P_\sigma y'} - \frac{P_\sigma\mathcal{G}(u)}{P_\sigma y}, 1\right\rangle_Y\right| \\
&\leq \left(\left\|\frac{P_\sigma\mathcal{G}(u)}{P_\sigma y}\right\|_Y + \left\|\frac{P_\sigma\mathcal{G}(u)}{P_\sigma y'}\right\|_Y + \|1\|_Y\right)\left\|\frac{P_\sigma\mathcal{G}(u)}{P_\sigma y} - \frac{P_\sigma\mathcal{G}(u)}{P_\sigma y'}\right\|_Y
\end{aligned}$$

$$\leq \left(2\varepsilon(r)^{-1}\left\|P_\sigma\mathcal{G}(u)\right\|_Y + \lambda(D)^{1/2}|T|^{1/2}\right) \times \varepsilon(r)^{-2}\|P_\sigma\mathcal{G}(u)\|_Y \left\|P_\sigma y - P_\sigma y'\right\|_Y$$
$$\leq \left(2\varepsilon(r)^{-1}\varepsilon(R_{\mathcal{G}}(\|u\|_X))^{-1} + \lambda(D)^{1/2}|T|^{1/2}\right) \times \varepsilon(r)^{-2}\varepsilon(R_{\mathcal{G}}(\|u\|_X))^{-1}L_\sigma(r)\left\|y - y'\right\|_Y,$$

which provides $M_{2,r}(\|u\|_X)$.

$\boxed{\Phi_{W_2}}$ We have that $\|\cdot\|_Y = \|\cdot\|_{L^2(D;\dot{H}^{-1}(T))}$. By Lemma 4.2, there exist $\varepsilon(r), \tilde{\varepsilon}(\|u\|_X) > 0$ such that $P_\sigma y, P_\sigma y' \in Y_{1,\varepsilon(r)}$, $P_\sigma\mathcal{G}(u) \in Y_{1,\tilde{\varepsilon}(\|u\|_X)}$. Using the (reverse) triangle inequality for $W_2(\cdot,\cdot)$, Cauchy–Schwarz, Lemmas 4.1 and 4.2, and the assumption, we may calculate

$$\begin{aligned}
&|\Phi_{W_2}(u;y) - \Phi_{W_2}(u;y')| \\
&\leq \frac{1}{2}\int_D \left|W_2((P_\sigma\mathcal{G}(u))(x,\cdot),(P_\sigma y)(x,\cdot))^2 - W_2((P_\sigma\mathcal{G}(u))(x,\cdot),(P_\sigma y')(x,\cdot))^2\right| \lambda(\mathrm{d}x) \\
&\leq \frac{1}{2}\int_D \left|W_2((P_\sigma\mathcal{G}(u))(x,\cdot),(P_\sigma y)(x,\cdot)) + W_2((P_\sigma\mathcal{G}(u))(x,\cdot),(P_\sigma y')(x,\cdot))\right| \\
&\qquad \times W_2((P_\sigma y)(x,\cdot),(P_\sigma y')(x,\cdot))\,\lambda(\mathrm{d}x) \\
&\leq (\tilde{\varepsilon}(\|u\|_X) \wedge \varepsilon(r))^{-1/2}\varepsilon(r)^{-1/2}(\|P_\sigma\mathcal{G}(u) - P_\sigma y\|_Y + \|P_\sigma\mathcal{G}(u) - P_\sigma y'\|_Y) \\
&\qquad \times \|P_\sigma y - P_\sigma y'\|_Y \\
&\leq 2(\tilde{\varepsilon}(\|u\|_X) \wedge \varepsilon(r))^{-1/2}\varepsilon(r)^{-1/2}C_p(T)^{1/2}\lambda(D)^{1/2} \\
&\qquad \times (\tilde{\varepsilon}(\|u\|_X)^{-1} + \varepsilon(r)^{-1})L_\sigma(r)\|y - y'\|_Y,
\end{aligned}$$

which provides $M_{2,r}(\|u\|_X)$. ∎

The following now follows from Lemma 4.5 and Theorem 4.4 in [35].

Theorem 4.6 (well-posedness). *Let π_0 be a Borel probability measure on X, and let Assumptions 4.3 hold. Choose $\Phi \in \{\Phi_{L^2}, \Phi_{H^{-1}}, \Phi_M, \Phi_{W_2}\}$, let $r > 0$ and assume*

$$S_{\Phi,r} := \int_X M_{2,r}(\|u\|_X)^2\,\pi_0(\mathrm{d}u) < \infty,$$

where $M_{2,r}(\cdot)$ is the corresponding function from Lemma 4.5. Then there exists $C_\Phi(r) > 0$ such that for all $y, y' \in Y$ with $\|y\|_{L^\infty(D;L^\infty(T))}, \|y'\|_{L^\infty(D;L^\infty(T))} < r$,

$$d_{\mathrm{H}}(\pi_\Phi^y, \pi_\Phi^{y'}) \leq C_\Phi(r)\|y - y'\|_Y.$$

Remark 4.7. We address that $\|\cdot\|_Y = \|\cdot\|_{L^2(D;\dot{H}^{-1}(T))}$ for $\Phi_{H^{-1}}$ and Φ_{W_2} while $\|\cdot\|_Y = \|\cdot\|_{L^2(D;L^2(T))}$ for Φ_{L^2} and Φ_M from the proof of Lemma 4.5. In particular, in combination with Theorem 4.6, we have the following two stability results for W_2-based posterior and L^2-based posterior:

$$d_{\mathrm{H}}(\pi_{\Phi_{W_2}}^y, \pi_{\Phi_{W_2}}^{y'}) \leq C_{W_2}(r)\|y - y'\|_{L^2(D;\dot{\boldsymbol{H}}^{-1}(\boldsymbol{T}))},$$

$$d_{\mathrm{H}}(\pi^{y}_{\Phi_{L^2}}, \pi^{y'}_{\Phi_{L^2}}) \leq C_{L^2}(r)\|y - y'\|_{L^2(D;\boldsymbol{L^2(T)})},$$

where d_{H} represents the Hellinger distance.

The main difference between these two stability estimates lies in the functional spaces for the noise $y - y'$ defined on the domain $T \subseteq \mathbb{R}^s$. As an L^2-based Sobolev seminorm, $\dot{H}^{-1}$ can be considered as a weighted L^2 norm with a weighting factor being asymptotically $|\boldsymbol{\xi}|^{-1}$, where $\boldsymbol{\xi}$ is the frequency. Thus, the high-frequency content of the input is imposed with a much smaller weight compared to the low-frequency content of the input. In another word, $\|y - y'\|_{L^2(D;\boldsymbol{\dot{H}^{-1}(T)})}$ is much smaller than $\|y - y'\|_{L^2(D;\boldsymbol{L^2(T)})}$ given high-frequency noise $y - y'$. These two different estimates consequently lead to a stronger stability result for W_2-based posterior and $\dot{H}^{-1}$-based posterior in the regime of high-frequency noises.

Theorem 4.6 is another theoretical derivation that demonstrates the better stability of the W_2 metric, but in a Bayesian inference framework in the context of solving Bayesian inverse problems. The result is well aligned with the analysis for the deterministic approach for inverse data matching problems [15]. The asymptotic and nonasymptotic connections between the W_2 metric and the $\dot{H}^{-1}$ seminorm play the central role in achieving such estimates as weaker norms smooth out the input.

Remark 4.8. The assumption that $S_{\Phi,r} < \infty$ is key to applying this result, which from the definitions of $M_{2,r}(\cdot)$ in Lemma 4.5 will depend on the growth rate of the forward map $\mathcal{G}$ and the structure of the density mapping P_σ, as well as the prior π_0. In the main application considered in this article, FWI, $\mathcal{G}$ will be uniformly bounded, and thus ensuring this condition holds involves balancing the growth/decay rates of the map σ with the tails of the prior. The Gaussian and level set priors discussed in subsection 3.1 all have at least squared exponential moments, and so relatively weak conditions are required on σ to ensure the required integrability.

4.3. Application to FWI. We show that the above theory is applicable when the forward model is taken to be that described in subsection 2.2, with the choices of priors described in subsection 3.1. Specifically, we define the forward map $\mathcal{G} : X \to Y$ as follows.

Denote by $\mathbb{R}^d_U = \mathbb{R}^{d-1} \times \mathbb{R}_+$ the upper half space, and by $T \subset (0, \infty)$ the temporal domain with $|T| < \infty$. Let $\{s_j\}_{j=1}^{N_s}$ be a collection of point sources, where $s_j(x,t) = w(t)\delta(x - x_j)$, $x_j \in \mathbb{R}^d_U$, and the time-dependent function $w(t) \in H^1(T) \subset L^\infty(T)$. Let the space for the state parameter $X = C^0(\mathbb{R}^d_U; \mathbb{R}^k)$. We then define a transport function $F : X \to L^\infty(\mathbb{R}^d_U; \mathbb{R}_+)$ that maps the state parameter u to the physical parameter m (for Bayesian inversion). We will specify the choice of F later in the numerical examples. Given $u \in X$, we define the slowness $m : \mathbb{R}^d_U \to (v_{\max}^{-2}, v_{\min}^{-2}) \subseteq \mathbb{R}_+$ by $m = F(u)$, where the constants $v_{\min}$ and $v_{\max}$ are physical lower and upper bounds for the wave velocity. For each j, let $v_j \in L^2(\mathbb{R}^d_U; L^2(T))$ solve the following wave equation:

$$\begin{cases} m(x)\frac{\partial^2 v_j}{\partial t^2}(x,t) - \Delta v_j(x,t) = s_j(x,t), & (x,t) \in \mathbb{R}^d_U \times T, \\ v_j(x,0) = 0,\ \frac{\partial v_j}{\partial t}(x,0) = 0, & x \in \mathbb{R}^d_U, \\ \nabla_x v_j(x,t) \cdot \mathbf{n} = 0, & x \in \partial\mathbb{R}^d_U. \end{cases} \tag{4.1}$$

Let $D_0 \subset \mathbb{R}^d_U$ be a compact subset denoting the receiver locations; we assume D_0 is either

finite or has positive Lebesgue measure, and we equip it with either the counting measure or Lebesgue measure respectively.

Given a bounded linear observation operator $\mathcal{O} : L^2(\mathbb{R}^d_U; L^2(T)) \to L^2(D_0; L^2(T))$ we write $v_j^D = \mathcal{O} \circ v_j$ as the observable part of the solution.[5] Additionally, we assume the observable part of the solution to be essentially bounded and thus in $L^\infty(D_0; L^\infty(T))$. We concatenate the v_j^D for the N_s different source terms, writing the total amount of observable data $v^D \in Y_0 = L^\infty(D; L^\infty(T))$, where $D = D_0 \times \{1, \dots, N_s\}$. The forward mapping $\mathcal{G}$ is then defined as the mapping $u \mapsto v^D$.

We first note that though the forward problem is defined on an infinite spatial domain, we are only interested in inversion on a compact subset—numerically the forward problem will be restricted to this subset using perfectly matched layer absorbing boundary conditions to mimic the true physics [14]. The state parameter and slowness function will hence be assumed to be defined only on this compact subset rather than all of $\mathbb{R}^d_U$. We verify that Assumptions 4.3 are satisfied by the above with this assumption in place:

- Assumption 4.3(i) will hold, assuming that the data is preprocessed to remove the zero-frequency component.
- The continuous dependence of the map $\mathcal{G}$ in the L^2 norm with respect to the wave slowness $m(x) = v^{-2}(x)$ was addressed in [34, Theorem 2.8.3]. Assuming the transport map $F : X \to L^\infty$ is continuous almost surely under the prior π_0 with respect to the L^2 norm, the required π_0-a.s. continuity of the map $\mathcal{G}$ with respect to the state parameter u, i.e., Assumption 4.3(ii), will follow. If the map F is defined pointwise so that $F(u)(x) = f(u(x))$ for some continuous bounded $f : \mathbb{R}^k \to \mathbb{R}_+$, the required continuity will follow for any choice of prior on X. In the case of a level set prior (3.4), F is continuous into L^2 at $u \in X$ if the set $\{x \,|\, u(x) = 0\}$ has zero Lebesgue measure: this can be ensured to hold π_0-a.s. by choosing the measure ν_0 in (3.4) to be supported on a space of C^1 functions, for example, a sufficiently smooth Gaussian as discussed in subsection 3.1.1 [19]. The mixed level set prior (3.5) leads to the required π_0-a.s. continuity with an analogous setup.
- Since we have zero initial conditions and the source pulse $w(t)$ has a finite H^1 norm (and thus a finite L^∞ norm), we assume that the L^∞ energy estimates for the wave equation (4.1) guarantee that $\mathcal{G}(u) < C$, where C depends only on $\|w\|_{H^1(T)}, v_{\min}, v_{\max}$ and is independent of the model parameter u. As discussed in Remark 4.8, since the bound on $\mathcal{G}$ is global, the condition that $S_{\Phi,r} < \infty$ in Theorem 4.6 is true under weak conditions on the prior measure π_0 and density mapping P_σ—in particular, the choices of prior outlined in subsection 3.1 and any σ with up to exponential growth/decay.

For this particular FWI problem, the temporal domain T in our theoretical framework becomes a 1D domain. Therefore, one can efficiently calculate the four potentials, Φ_{L^2}, $\Phi_{H^{-1}}$, Φ_M, Φ_{W_2} as defined in (1.1)–(1.4). In particular, the calculation of the W_2 potential, Φ_{W_2}, is equivalent to the so-called trace-by-trace approach in [42], for which there is an explicit solution to the corresponding 1D optimal transport problem.

[5] The observation operator $\mathcal{O}$ will typically be the restriction to the subset $D_0 \times T$, though depending on the choice of D_0 may require mollification in space to ensure the operator is bounded.

5. Numerical examples. We illustrate the difference between the resulting distributions corresponding to the different potentials numerically for the FWI problem described in subsection 4.3. First, we present a simple constant-velocity example to illustrate the different structure among several likelihood functions and their corresponding (Gibbs) posterior distributions using a Gaussian prior. We then consider the recovery of a continuous velocity field using a Gaussian prior. Laplace approximations are first made to the posteriors, which correspond to second-order Taylor approximations of the posteriors around their modes. We look at the effect of perturbation of the data by noise uncorrelated in time on these Laplace approximations; we may compute various metrics between probability measures explicitly in the case of Gaussian. Moreover, we consider a mixed level set prior for the recovery of a salt model, where the velocity within the salt is unknown a priori. We again use Laplace approximations to provide approximations to the posteriors efficiently.

Throughout this section we fix the transformation map $F : X \to L^\infty$ in subsection 4.3 to be given by $F(u)(x) = \alpha_- \tanh(u(x)) + \alpha_+$ and $\alpha_\pm = v_{\min}^{-2} \pm v_{\max}^{-2}$. Hence for any $u(x) \in \mathbb{R}$, the velocity $v(x) = 1/\sqrt{m(x)} = 1/\sqrt{F(u(x))} \in (v_{\min}, v_{\max})$. The hyperbolic tangent function is chosen as a smooth monotone mapping from the unbounded range of Gaussian processes into a positive bounded interval to match the physical meaning of the unknown parameter within the hyperbolic equation.

For all numerical tests, the wave equation is solved on a rectangular domain and the boundaries are imposed with the absorbing boundary conditions [14] to approximate the unbounded domain and reduce reflections from the upper surface. The acoustic wave equation (4.1) is solved numerically by a second-order finite difference scheme.

5.1. Simple illustration of the posterior distributions. We first illustrate the difference in the potential functions by plotting a slice of the unnormalized likelihood functions and their posteriors in the direction of constant wave speed. The reference velocity is 3 km/s discretized on a rectangle domain 1.5 km by 2 km with a spatial spacing of 10 m. The source pulse is a superposition of three Ricker wavelets of different phase shifts, centered at 20 Hz. We equally distribute 8 point sources at $z = 50$ m and 200 receivers at $z = 1.4$ km. The observed data y is recorded for 1.5 s, containing transmissions only. Since these potentials have different scaling, we rescale them with their values evaluated at $v(x) \approx 1.395$ (i.e., $u(x) = 0$) for illustration purposes. We impose the same Gaussian prior with mean $v(x) \approx 2.63$ ($u(x) = -1$) for all three potentials, as shown in Figure 5.1.

The shape of the unnormalized likelihood functions highlight one significant advantage of the W_2 metric for FWI—mitigating cycle-skipping issues. The unimodal feature in both the W_2 likelihood function and its posterior corresponds to the desirable convexity property in the deterministic FWI [42] that heavily replies on local optimization algorithms. In turn, it facilitates sampling the posterior for Bayesian inversion. On the other hand, the multimodal feature in the plots for $\dot{H}^{-1}$ and L^2 are unsurprising as they compare signals locally, failing to capture the large phase changes in the waveforms. As expected, the shapes improve in the posteriors once the prior information is imposed, regularizing the likelihood function. We remark that the unimodality of W_2 is unlikely to hold for general coefficient function $v(x)$, but its multimodality is expected to be much less severe than $\dot{H}^{-1}$ and L^2 based on abundance numerical evidence in deterministic FWI [42].

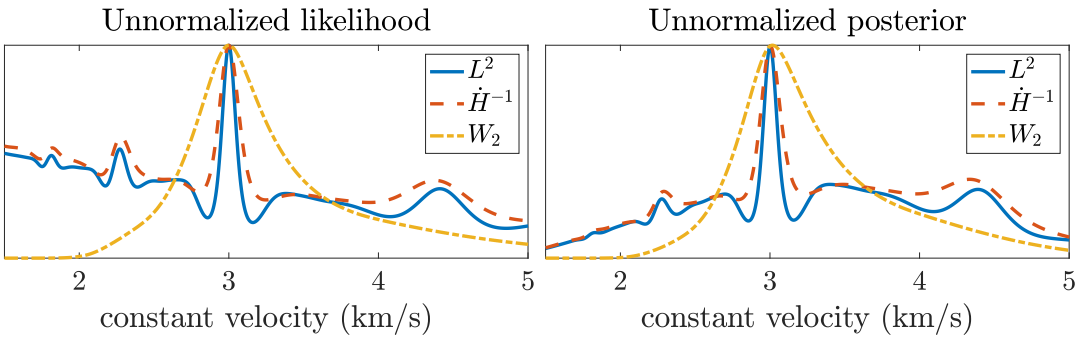

Figure 5.1. *(Left) An illustration of the unnormalized likelihood functions for potentials* $\Phi_{L^2}(u;y)$, $\Phi_{\dot{H}^{-1}}(u;y)$, $\Phi_{W_2}(u;y)$ *by restricting to constant wave speed. (Right) The corresponding unnormalized posterior distributions with the same Gaussian prior.*

5.2. Continuous velocity field. In this subsection we consider recovery of a continuous field $m(x) = F(u(x))$. We place a Gaussian prior on u and compare the distributions arising from each choice of loss function. Figure 5.2 shows the true velocity field $v = 1/\sqrt{m}$ (0.93 km in depth and 2.31 km in width) along with $N_s = 6$ sources (Ricker wavelet centered at 15 Hz) and 77 receivers equally distributed in the additional water layer (thickness 0.6 km) at $z = 150$ m. The discretization of the wave equations is 30 m in space and 3 ms in time. The total recording time is 1.8 s. The state parameter $u = F^{-1}(m)$ is also shown.

We chose a relatively flat Gaussian prior on u with mean m_0 given by a smoothed version of the true field, and covariance function given by the Matérn kernel with parameters $\sigma = 0.7$, $\nu = 3$, and $\ell = 0.05$. This choice of σ ensures that the prior on the velocity at each point is approximately uniformly distributed in the range $(v_{\min}, v_{\max})$. The prior mean and standard deviation are shown in Figure 5.3. The loss functions are all normalized such that $\Phi(m_0) = 1$, then inverse temperature parameter $\beta = 10^7$ is fixed—that is, the potentials $\Phi(u)$ are replaced by $\beta\Phi(u)/\Phi(m_0)$. In practice β will be found, for example, hierarchically [27], but we fix it here for a more direct comparison between the distributions. Note, however, that due to the difference in nature of the likelihood functions, the distributions will not be directly comparable quantitatively.

We denote by $\mathcal{L}^y_\Phi$ the Laplace approximation to π^y_Φ, that is, $\mathcal{L}^y_\Phi = N(m_\Phi, \mathcal{C}_\Phi)$, where

$$m_\Phi = \operatorname*{arg\,min}_{u\in X} \Phi(u;y) + \frac{1}{2}\langle u, \mathcal{C}_0^{-1}u\rangle, \quad \mathcal{C}_\Phi^{-1} = \nabla_u^2\Phi(u_\Phi;y) + \mathcal{C}_0^{-1}.$$

In finite dimensions, this approximation arises by performing a second-order Taylor approximation of the negative logarithm of the posterior Lebesgue density—note that when the potential Φ is quadratic, such as arises when the forward map is linear and the noise is additive Gaussian, this approximation would be exact. Additionally, in the limit of large data or small noise, the posterior and Laplace approximations coincide [32]. Numerically we calculate the maximum a posteriori (MAP) estimates m_Φ using the L-BFGS algorithm, initialized at a smoothed version of the true field to somewhat avoid multimodality issues when using the L^2 loss. We approximate the covariance $\mathcal{C}_\Phi$ using the methodology described in [20, section 4.2].

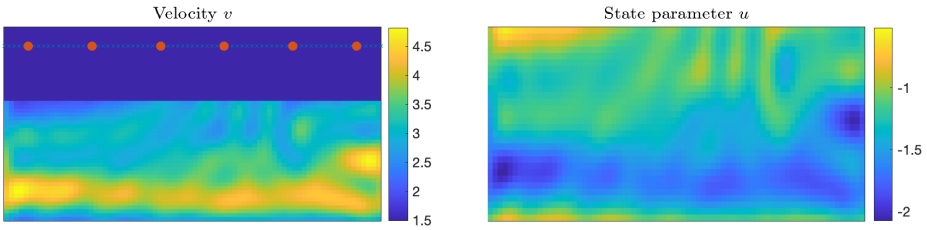

Figure 5.2. *(Left) The true continuous velocity field v we aim to infer in subsection 5.2, the location of the six sources $\{s_j\}$ and the set D_0 on which the solution is measured at each time. (Right) The state parameter $u = F^{-1}(1/v^2)$, restricted to the domain below the water level.*

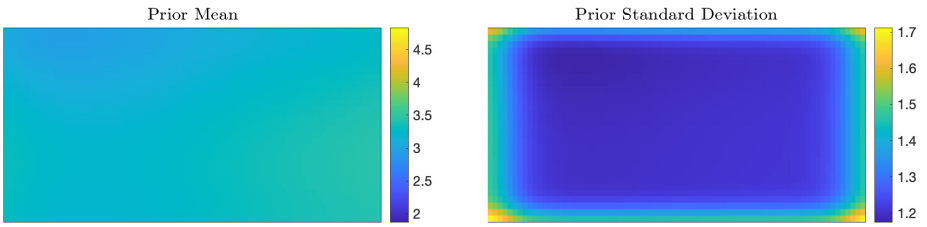

Figure 5.3. *The pushforward of the choice of prior mean $m_0(x)$ and standard deviation $\sqrt{c_0(x,x)}$ on u to $v = 1/\sqrt{F(u)}$, used in subsection 5.2.*

Remark 5.1. We could instead choose to probe the posterior distribution directly using Markov chain Monte Carlo (MCMC) methods—though the dimension of the field we are inferring is high, dimension robust methods are available [5]. However, the statistical performance of such methods is highly dependent on how informative the data is relative to the prior. In the setups we consider the data is highly informative, and so generating a large number of uncorrelated samples using MCMC would likely be computationally prohibitive. Additionally, in the cases of the L^2 and $\dot{H}^{-1}$ likelihoods the posterior is likely to be highly multimodal, as illustrated in the previous subsection, which could lead to additional issues with exploring the full posterior; this would be less of an issue with the W_2 likelihood, however.

We consider both clean observations $y = \mathcal{G}(u)$ and noisy observations $y' = \mathcal{G}(u) + \eta$, where η has the form

$$\eta(t,x) = \left(1 + \frac{y(t,x)}{\|y\|_{L^\infty}}\right)\eta_0(t)$$

and η_0 is temporal white noise; this form of noise is similar to that considered in [41] and does not lie in L^2 in the continuous time limit. The resulting signal-to-noise ratio is 14.6 dB.

In Figures 5.4 and 5.5 we show the results of the inversion with the clean and noisy data, respectively, for the three potentials Φ_{L^2}, $\Phi_{\dot{H}^{-1}}$, and Φ_{W_2}.[6] For the clean data, both the L^2 and W_2 likelihoods lead to good reconstruction in the mean, though the standard deviation fields are different, with the W_2 likelihood leading to higher uncertainty in the regions where

[6]We do not consider the potential Φ_M in this section as it is very sensitive to the choice of mapping $\mathbb{P}_\sigma$; the potential Φ_{W_2} is also sensitive to this choice. However, the previous study allows for it to be chosen intuitively.

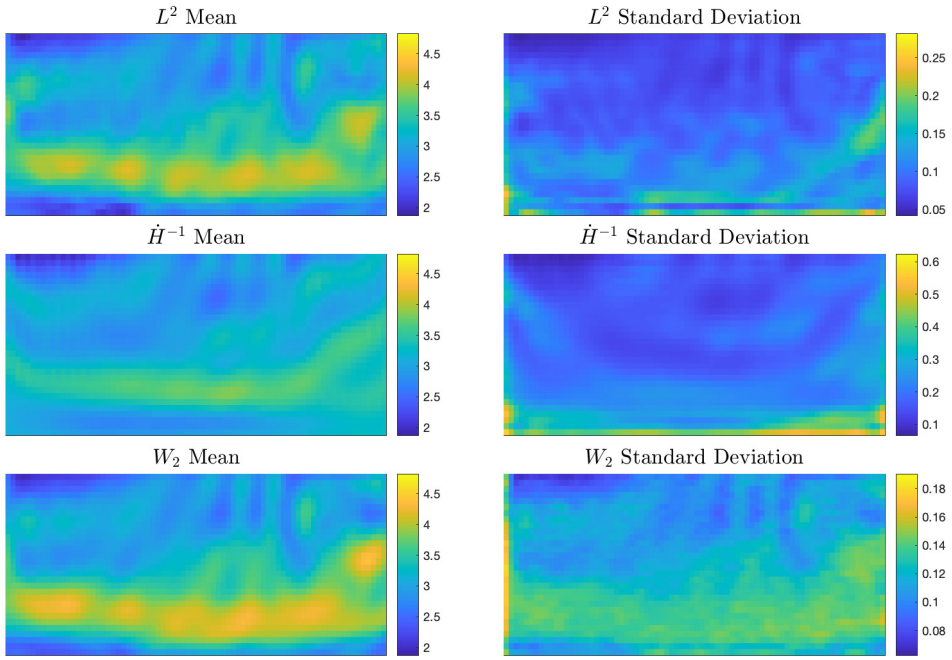

Figure 5.4. *The means (left) and standard deviations (right) of the Laplace approximations arising using each of the loss functions $\Phi_{L^2}, \Phi_{\dot{H}^{-1}}$, and Φ_{W_2} with clean data y in subsection 5.2.*

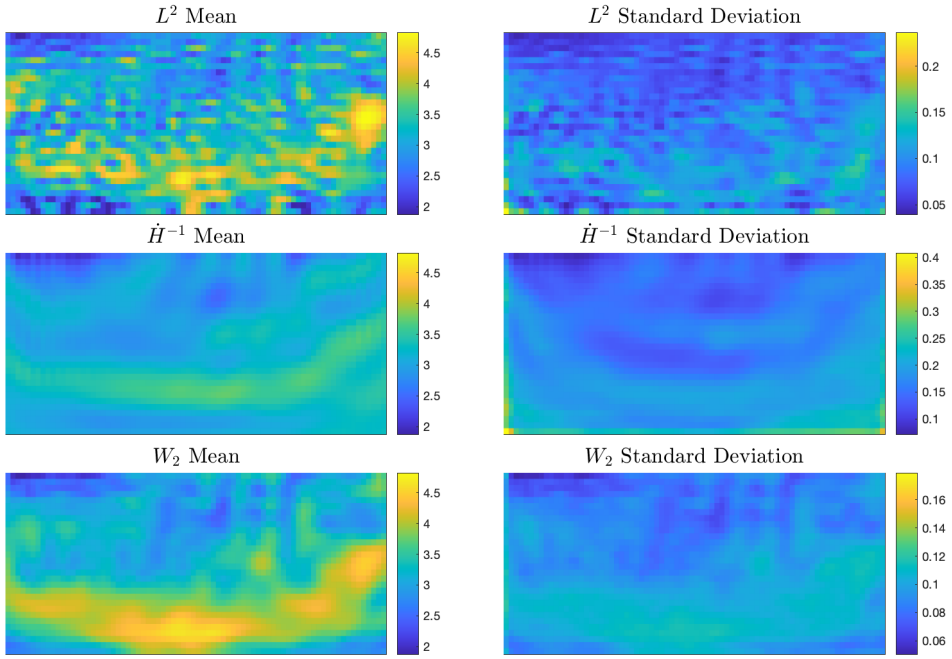

Figure 5.5. *The means (left) and standard deviations (right) of the Laplace approximations arising using each of the loss functions $\Phi_{L^2}, \Phi_{\dot{H}^{-1}}$, and Φ_{W_2} with noisy data y' in subsection 5.2.*

Table 5.1

Wasserstein distances between the Laplace approximations for the Gibbs posteriors considered in subsection 5.2 when perturbed by noise uncorrelated in time.

	Φ_{L^2}	$\Phi_{\dot{H}^{-1}}$	Φ_{W_2}
$d_{\text{Wass}}(\mathcal{L}_\Phi^y, \mathcal{L}_\Phi^{y'})$	9.34	2.83	6.60

the velocity is higher. The $\dot{H}^{-1}$ reconstruction is less accurate in the mean but has higher uncertainty to account for this. For the noisy reconstruction, there is a stark difference for the L^2 reconstruction, with a poor quality noisy estimate obtained. The W_2 reconstruction is only slightly worse than that with the clean data, and again uncertainty is highest in regions with the highest velocity. The $\dot{H}^{-1}$ reconstruction is essentially unchanged from the clean data, though the average uncertainty has decreased slightly.

These sensitivities can be quantified using the Wasserstein distance between the Laplace approximations, which has an analytic form. We do not consider the Hellinger distance here, as considered in theory—the resulting Laplace approximations are close to being singular, and so the distances between the approximations are close to the maximum value of 1, whereas the Wasserstein distance is finite even for singular measures. In Table 5.1 we show $d_{\text{Wass}}(\mathcal{L}_\Phi^y, \mathcal{L}_\Phi^{y'})$ for each of the potentials Φ considered, for the fixed noisy realization of the data y' used in Figure 5.5. This illustrates the balance needed between stability and accuracy: $\Phi_{\dot{H}^{-1}}$ leads to a more stable but less accurate posterior than Φ_{W_2}, and conversely Φ_{L^2} leads to a similarly accurate but much less stable posterior than Φ_{W_2}.

5.3. Salt model. We now consider the recovery of a constant region, corresponding to salt, over a continuous background velocity. The true velocity field (1.95 km in width and 1.11 km in depth) is shown in Figure 5.6 along with the source and receiver locations; we increase the number of sources to $N_s = 11$ (Ricker wavelet centered at 10 Hz) in this example, while the number of receivers is 65. They are equally distributed in the additional water layer (thickness 300 m) at $z = 150$ m. The discretization of the forward wave equation is 30 m in space and 2 ms in time. The total recording time for the wave propagation is 2 s. We consider both a Gaussian prior as in the previous subsection and the mixed level set method outlined in subsection 3.1, choosing the Gaussian field underlying the level set function to have a Matérn covariance prior with parameters $\sigma = 10$, $\nu = 2$, and $\ell = 0.25$, and fixing the continuous background field. The salt velocity is assumed unknown in the mixed level set model, and a relatively flat prior of $N(3, 4^2)$ is placed on it; for reference, the true value is 4.79. Numerically the indicator functions in the level set prior (3.5) are replaced with smoothed versions to allow for the computation of gradients. The potentials are normalized as in the previous subsection, and we choose a lower inverse temperature parameter $\beta = 10^4$ to allow for more influence from the prior. Additionally, we do not corrupt the observations with any noise in this example.

In Figure 5.7 we show the output using the potentials Φ_{L^2}, Φ_{W_2} when using a Gaussian prior, as well as when using Φ_{W_2} with the mixed level set prior. When a Gaussian prior is used, the W_2 reconstruction recovers more of the salt than the L^2 reconstruction, though the uncertainty for both is similar. On the other hand, using the level set method, much more of

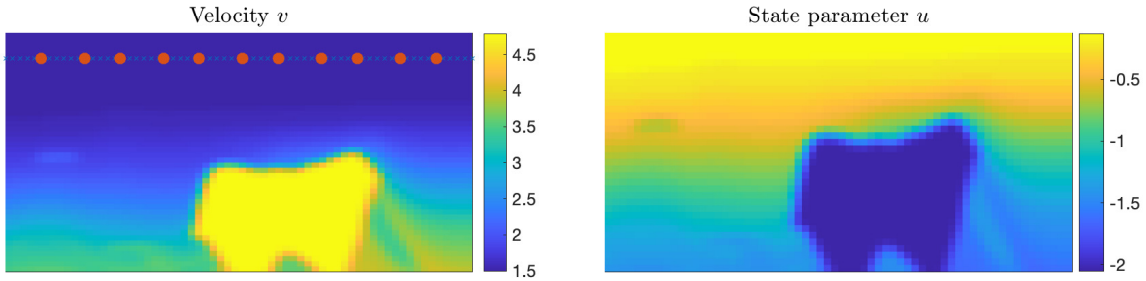

Figure 5.6. *(Left) The true salt velocity field v we aim to infer in subsection 5.3, the location of the eleven sources $\{s_j\}$, and the set D_0 on which the solution is measured at each time. (Right) The state parameter $u = F^{-1}(1/v^2)$, restricted to the domain below the water level.*

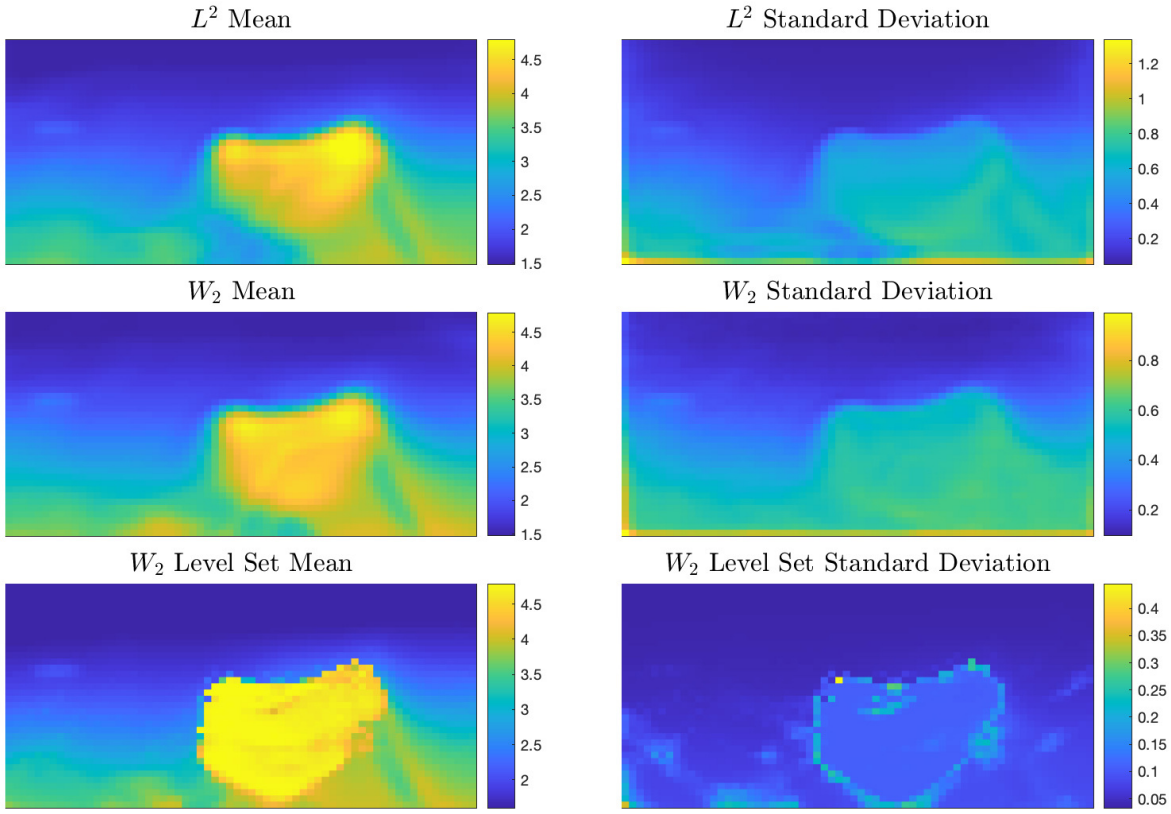

Figure 5.7. *The means (left) and standard deviations (right) of the Laplace approximations arising using each of the loss functions $\Phi_{L^2}, \Phi_{\dot{H}^{-1}}$, and Φ_{W_2} with clean data y in subsection 5.3.*

the salt is recovered, and the true value of the salt velocity is recovered almost exactly. The standard deviation illustrates the uncertainty in the boundary location, as well as small areas within the salt where the reconstruction is less accurate.

6. Conclusion. In this paper, we have analyzed the Bayesian properties of four potentials that originate from the objective functions for deterministic full-waveform inversion. We have demonstrated that two of the potentials are equipped with explicit noise models, while the other two can be seen as approximately having state-dependent multiplicative noise that leads to well-defined and stable Gibbs posteriors. As a main component of the paper, our stability results show that posteriors that are based on the Wasserstein loss and the negative Sobolev norm are more stable with respect to high-frequency noise on the data. Numerical inversions under a Bayesian setting also illustrate the robustness and advantages of choosing the Wasserstein-based potentials for seismic inversion.

Acknowledgments. We thank Prof. Georg Stadler and Prof. Andrew Stuart for constructive discussions.

REFERENCES

[1] P. Alquier, J. Ridgway, and N. Chopin, *On the properties of variational approximations of Gibbs posteriors*, J. Mach. Learn. Res., 17 (2016), pp. 8374–8414.
[2] A. Asnaashari, R. Brossier, S. Garambois, F. Audebert, P. Thore, and J. Virieux, *Regularized seismic full waveform inversion with prior model information*, Geophys., 78 (2013), pp. R25–R36.
[3] E. Bernton, P. E. Jacob, M. Gerber, and C. P. Robert, *Inference in Generative Models Using the Wasserstein Distance*, preprint, https://arxiv.org/abs/1701.05146v1, 2017.
[4] A. L. Bertozzi, X. Luo, A. M. Stuart, and K. C. Zygalakis, *Uncertainty quantification in graph-based classification of high dimensional data*, SIAM/ASA J. Uncertain. Quantif., 6 (2018), pp. 568–595, https://doi.org/10.1137/17M1134214.
[5] A. Beskos, M. Girolami, S. Lan, P. E. Farrell, and A. M. Stuart, *Geometric MCMC for infinite-dimensional inverse problems*, J. Comput. Phys., 335 (2017), pp. 327–351.
[6] M. Dashti and A. M. Stuart, *The Bayesian approach to inverse problems*, in Handbook of Uncertainty Quantification, Springer, 20167, pp. 311–428.
[7] J. Dolbeault, B. Nazaret, and G. Savaré, *A new class of transport distances between measures*, Calc. Var. Partial Differential Equations, 34 (2009), pp. 193–231.
[8] O. R. A. Dunbar, M. M. Dunlop, C. M. Elliott, V. H. Hoang, and A. M. Stuart, *Reconciling Bayesian and perimeter regularization for binary inversion*, SIAM J. Sci. Comput., 42 (2020), A1984–A2013, https://doi.org/10.1137/18M1179559.
[9] M. M. Dunlop, *Multiplicative Noise in Bayesian Inverse Problems: Well-Posedness and Consistency of MAP Estimators*, preprint, https://arxiv.org/abs/1910.14632, 2019.
[10] M. M. Dunlop, M. A. Iglesias, and A. M. Stuart, *Hierarchical Bayesian level set inversion*, Stat. Comput., 27 (2017), pp. 1555–1584.
[11] T. A. El Moselhy and Y. M. Marzouk, *Bayesian inference with optimal maps*, J. Comput. Phys., 231 (2012), pp. 7815–7850.
[12] B. Engquist and B. D. Froese, *Application of the Wasserstein metric to seismic signals*, Commun. Math. Sci., 12 (2014), pp. 979–988.
[13] B. Engquist, B. D. Froese, and Y. Yang, *Optimal transport for seismic full waveform inversion*, Commun. Math. Sci., 14 (2016), pp. 2309–2330.
[14] B. Engquist and A. Majda, *Absorbing boundary conditions for numerical simulation of waves*, Proc. Natl. Acad. Sci. USA, 74 (1977), pp. 1765–1766.
[15] B. Engquist, K. Ren, and Y. Yang, *The quadratic Wasserstein metric for inverse data matching*, Inverse Problems, 36 (2020), 055001.
[16] A. Fichtner and J. Trampert, *Resolution analysis in full waveform inversion*, Geophys. J. Internat., 187 (2011), pp. 1604–1624.
[17] W. P. Gouveia and J. A. Scales, *Bayesian seismic waveform inversion: Parameter estimation and uncertainty analysis*, J. Geophys. Res. Solid Earth, 103 (1998), pp. 2759–2779.

[18] M. Iglesias, M. Park, and M. V. Tretyakov, *Bayesian inversion in resin transfer molding*, Inverse Problems, 34 (2018), 105002.
[19] M. A. Iglesias, Y. Lu, and A. M. Stuart, *A Bayesian level set method for geometric inverse problems*, Interfaces Free Boundaries, 18 (2016), pp. 181–217.
[20] T. Isaac, N. Petra, G. Stadler, and O. Ghattas, *Scalable and efficient algorithms for the propagation of uncertainty from data through inference to prediction for large-scale problems, with application to flow of the Antarctic ice sheet*, J. Comput. Phys., 296 (2015), pp. 348–368.
[21] M. Izzatullah, T. van Leeuwen, and D. Peter, *Bayesian uncertainty estimation for full waveform inversion: A numerical study*, in SEG Technical Program Expanded Abstracts 2019, Society of Exploration Geophysicists, 2019, pp. 1685–1689.
[22] L. V. Kantorovich, *On the translocation of masses*, J. Math. Sci. Dokl. Akad. Nauk SSSR, 133 (2006), pp. 8–227.
[23] J. Latz, *On the well-posedness of Bayesian inverse problems*, SIAM/ASA J. Uncertain. Quantif., 8 (2020), pp. 451–482, https://doi.org/10.1137/19M1247176.
[24] L. Métivier, R. Brossier, Q. Mérigot, E. Oudet, and J. Virieux, *Measuring the misfit between seismograms using an optimal transport distance: Application to full waveform inversion*, Geophys. J. Internat., 205 (2016), pp. 345–377.
[25] G. Monge, *Mémoire sur la théorie des déblais et de remblais. histoire de l'académie royale des sciences de paris*, avec les Mémoires de Mathématique et de Physique pour la même année, 1781, pp. 666–704.
[26] P. Mora, *Inversion = migration + tomography*, Geophys., 54 (1989), pp. 1575–1586.
[27] M. Motamed and D. Appelo, *Wasserstein metric-driven Bayesian inversion with applications to signal processing*, Int. J. Uncertain. Quantif., 9 (2019), pp. 394–414.
[28] V. Oropeza and M. Sacchi, *Simultaneous seismic data denoising and reconstruction via multichannel singular spectrum analysis*, Geophys., 76 (2011), pp. V25–V32.
[29] F. Otto and C. Villani, *Generalization of an inequality by Talagrand and links with the logarithmic Sobolev inequality*, J. Funct. Anal., 173 (2000), pp. 361–400.
[30] R. Peyre, *Comparison between W_2 distance and $\dot{H}^{-1}$ norm, and localization of Wasserstein distance*, ESAIM Control Optim. Calc. Var., 24 (2018), pp. 1489–1501.
[31] J. A. Scales and L. Tenorio, *Prior information and uncertainty in inverse problems*, Geophys., 66 (2001), pp. 389–397.
[32] C. Schillings, B. Sprungk, and P. Wacker, *On the convergence of the Laplace approximation and noise-level-robustness of Laplace-based Monte Carlo methods for Bayesian inverse problems*, Numer. Math., 145 (2020), pp. 915–971.
[33] M. Sommerfeld and A. Munk, *Inference for empirical Wasserstein distances on finite spaces*, J. R. Stat. Soc. Ser. B. Stat. Methodol., 80 (2018), pp. 219–238.
[34] C. C. Stolk, *On the Modeling and Inversion of Seismic Data*, Ph.D. thesis, Universiteit Utrecht, 2000.
[35] T. J. Sullivan, *Well-posed Bayesian inverse problems and heavy-tailed stable quasi-Banach space priors*, Inverse Probl. Imaging, 11 (2017), pp. 857–874.
[36] N. Syring and R. Martin, *Calibrating general posterior credible regions*, Biometrika, 106 (2018), pp. 479–486.
[37] A. Tarantola, *Inverse Problem Theory and Methods for Model Parameter Estimation*, SIAM, 2005, https://doi.org/10.1137/1.9780898717921.
[38] C. Villani, *Topics in Optimal Transportation*, Grad. Stud. Math. 58, American Mathematical Society, Providence, RI, 2003.
[39] J. Virieux and S. Operto, *An overview of full-waveform inversion in exploration geophysics*, Geophys., 74 (2009), pp. WCC1–WCC26.
[40] Y. Yang, *Analysis and Application of Optimal Transport for Challenging Seismic Inverse Problems*, preprint, https://arxiv.org/abs/1902.01226, 2019.
[41] Y. Yang and B. Engquist, *Analysis of optimal transport and related misfit functions in full-waveform inversion*, Geophys., 83 (2018), pp. A7–A12.
[42] Y. Yang, B. Engquist, J. Sun, and B. F. Hamfeldt, *Application of optimal transport and the quadratic Wasserstein metric to full-waveform inversion*, Geophys., 83 (2018), pp. R43–R62.
[43] H. Zhu, S. Li, S. Fomel, G. Stadler, and O. Ghattas, *A Bayesian approach to estimate uncertainty for full-waveform inversion using a priori information from depth migration*, Geophys., 81 (2016), pp. R307–R323.