Wasserstein metric-driven Bayesian inversion with applications to signal processing

M. Motamed* and D. Appelö†

*Department of Mathematics and Statistics, University of New Mexico, USA
†Department of Applied Mathematics, University of Colorado Boulder, USA

August 7, 2018

Abstract. We present a Bayesian framework based on a new exponential likelihood function driven by the quadratic Wasserstein metric. Compared to conventional Bayesian models based on Gaussian likelihood functions driven by the least-squares norm ($L_2$ norm), the new framework features several advantages. First, the new framework does not rely on the likelihood of the measurement noise and hence can treat complicated noise structures such as combined additive and multiplicative noise. Secondly, unlike the normal likelihood function, the Wasserstein-based exponential likelihood function does not usually generate multiple local extrema. As a result, the new framework features better convergence to correct posteriors when a Markov Chain Monte Carlo sampling algorithm is employed. Thirdly, in the particular case of signal processing problems, while a normal likelihood function measures only the amplitude differences between the observed and simulated signals, the new likelihood function can capture both the amplitude and the phase differences. We apply the new framework to a class of signal processing problems, that is, the inverse uncertainty quantification of waveforms, and demonstrate its advantages compared to Bayesian models with normal likelihood functions.

Keywords. Bayesian inversion; Wasserstein metric; measurement noise; Markov Chain Monte Carlo; Metropolis-within-Gibbs Sampling; signal processing; uncertainty quantification; waveforms

MSC. 62F15, 62P30, 65M32, 60J22, 86A15, 35L53

1 Introduction

The Euclidean distance is one of the most widely used metrics in constructing misfit functions in deterministic inversion algorithms. In a Bayesian framework (see e.g. [7, 9, 17]), this misfit function corresponds to a normal likelihood function, which is valid under the (often unrealistic) assumption that the experimental noise is additive and Gaussian. In addition to being unrealistic, a Guassian likelihood function measures only the amplitude differences between the observed and simulated signals and does not capture the phase differences.
Moreover, Gaussian likelihood functions often have multiple local extrema that may result in converging to wrong posteriors (see Section 5) when a Markov Chain Monte Carlo (MCMC) sampling algorithm [6] is used. Such disadvantages are particularly of great concern in signal processing applications, such as seismic inversion problems where the noise is not necessarily additive and the wave signals are oscillatory and cyclic.

In the present work we will propose a Bayesian framework based on an exponential likelihood function driven by the quadratic Wasserstein metric (see e.g. [18, 19]). We will compare the new framework with current Bayesian models based on normal likelihood functions when applied to a class of signal processing problems, including seismic source inversion and seismic imaging. Our main motivation comes from the recent work [20] that applies the quadratic Wasserstein metric to deterministic seismic inversion problems, where the authors demonstrate the efficiency of the quadratic Wasserstein metric compared to the $L_2$ norm in defining the misfit function. We extend the work [20] to Bayesian inversion and demonstrate applications of the new framework in the uncertainty quantification of seismic inversion problems.

The rest of the paper is organized as follows. In Section 2 we formulate the Bayesian inversion problem and discuss the likelihood and measurement noise structures. We then present the new Wasserstein metric-driven Bayesian framework in Section 3, followed by a standard MCMC algorithm in Section 4. We finally present several numerical examples related to seismic wave inversion in Section 5.

2 Bayesian inversion problem

2.1 Problem formulation

Let $g = (g_1, \ldots , g_N) \in \mathbb{R}^N$ be a vector of $N$ observed quantities, for instance, measured by an experimental device. Let further $f = (f_1, \ldots , f_N) \in \mathbb{R}^N$ be a vector of $N$ predicted quantities corresponding to $g$, computed by a forward predictive model with an input parameter vector $\theta = (\theta_1, \ldots , \theta_m) \in \Theta \subset \mathbb{R}^m$:

$$f = f(\theta) : \Theta \subset \mathbb{R}^m \rightarrow \mathbb{R}^N.$$ 

A main assumption in mathematical statistics is that the model parameter vector $\theta$, and consequently any physical quantity to be predicted by the forward model, is described by a stochastic law characterized by a probability density function (PDF). The density of $\theta$ is available to us only through the experimental measurements $g$ that are often corrupted by experimental noise and error. The goal is then to find the density of $\theta$, given the forward model $f(\theta)$ and the noisy data $g$.

In Bayesian inversion the conditional posterior probability density of the model parameter vector $\theta$ is given by the Bayes’ rule; see e.g. [7, 9, 17]:

$$\pi(\theta | g) = \frac{\pi(g | \theta) \pi(\theta)}{\int_{\Theta} \pi(g | \theta) \pi(\theta) d\theta} \propto \pi(g | \theta) \pi(\theta),$$

where $\pi(g | \theta)$ is the likelihood, and $\pi(\theta)$ is the prior density of $\theta$. The proportionality follows from the fact that the denominator (referred to as the evidence) does not depend on $\theta$ and
hence can be regarded as a constant. Assuming the observed samples \( \{g_i\}_{i=1}^N \) are independent, their joint likelihood reads

\[
L(\theta) := \pi(g|\theta) = \prod_{i=1}^N \pi(g_i|\theta),
\]

where the first equality indicates that the likelihood can be viewed as a function of only \( \theta \), referred to as the likelihood function and denoted by \( L(\theta) \).

### 2.2 Likelihood and noise structure

A major step in Bayesian analysis is the selection of the likelihood function. This is usually done based on the structure of the measurement noise. The most common choice is the Gaussian likelihood, built based on the assumption that the measurement noise \( \{\varepsilon_i\}_{i=1}^N \) in the \( N \) measured quantities \( \{g_i\}_{i=1}^N \) is additive and normally distributed with zero mean and a standard deviation \( \sigma \):

\[
g_i = f_i(\theta) + \varepsilon_i, \quad \varepsilon_i \sim \text{Normal}(0, \sigma), \quad i = 1, \ldots, N.
\]

In this case the likelihood function reads

\[
L_{\text{norm}}(\theta) = \pi_{\text{norm}}(g|\theta) = \frac{1}{(2\pi)^{N/2} \sigma^N} \exp\left(\frac{-1}{2\sigma^2} \sum_{i=1}^N (g_i - f_i(\theta))^2\right).
\]

One critical problem with this setting is that the Gaussian additive noise assumption may not be realistic. For instance, in ultrasound and laser imaging with \( g \in \mathbb{R}^N \) being an \( N \)-pixel image, the noise may be multiplicative. A practical example is a speckle noise (see e.g. [8]) with mean one and variance \( 1/s \):

\[
g_i = \varepsilon_i f_i(\theta), \quad \varepsilon_i \sim \text{Gamma}(s, 1/s), \quad i = 1, \ldots, N.
\]

Here Gamma\((a, b)\) is a Gamma distribution with a shape parameter \( a > 0 \) and a rate parameter \( b > 0 \). In some applications, the noise structure may be even more complicated. For instance the measurement noise may be both additive and multiplicative, taking the form

\[
g_i = \varepsilon_i^{(1)} f_i(\theta) + \varepsilon_i^{(2)}, \quad i = 1, \ldots, N. \tag{1}
\]

Such complicated noise structures may not be easily treated by Bayesian frameworks that rely on the likelihood of the noise, even when a non-Gaussian likelihood function is employed. It is to be noted that the noise structure is not the only problem with Gaussian likelihoods. Even if the noise structure is additive and can be modeled by a Gaussian likelihood function, the likelihood may have multiple local extrema, resulting in wrong posteriors when applying a MCMC sampling method; see Section 5. Moreover, in signal processing problems, a Gaussian likelihood function that is based on the \( L_2 \) norm measures only the amplitude differences between the observed and predicted signals and does not capture the phase differences. This latter problem is indeed a crucial limitation for instance in: a) seismic inversion with moderately to highly oscillatory seismic waves, and b) MRI noise reduction where the retainment of the image fine features is desired.
In what follows we present a Bayesian framework based on an exponential likelihood function driven by a quadratic Wasserstein metric. Unlike conventional Bayesian analysis, this framework does not rely on the likelihood of the measurement noise and hence can treat complicated noise structures as in [1]. Moreover, since the Wasserstein metric has better optimization and fitting properties than the least-squares norm (see e.g. [20] and the references therein), the new framework may feature better convergence properties than the standard frameworks based on Gaussian likelihoods.

3 Wasserstein metric-driven Bayesian framework

The Wasserstein metric is a distance function defined between two probability distributions. It corresponds to the minimum “cost” of turning one distribution into the other. The metric has been applied to various fields, including image processing [15], computer vision [10], Stochastic programming, [5], and seismic imaging [20]. We refer to [18, 19] for a complete introduction to Wasserstein distances.

Consider two discrete-time signals \( f, g \in \mathbb{R}^N \) given at a discrete set of time level \( \{t_i\}_{i=1}^N \). In order to compute the discrete quadratic Wasserstein metric for the two signals, we follow [20] and construct two discrete CDFs as follows:

- Shift the signals to ensure positivity: select a constant \( c \) so that \( f_i + c > 0 \) and \( g_i + c > 0 \), for all \( i = 1, \ldots, N \).

- Rescale the shifted signals so that they share a common unit total mass:
  \[
  \tilde{f} = \frac{f + c}{\langle f + c \rangle}, \quad \tilde{g} = \frac{g + c}{\langle g + c \rangle},
  \]
  where \( \langle f \rangle = \sum_{i=1}^N f_i \) is the total mass of the signal \( f \).

- Find the discrete CDFs \( F = (F_1, \ldots, F_N) \) and \( G = (G_1, \ldots, G_N) \) with the components
  \[
  F_i = \sum_{j=1}^i \tilde{f}_j, \quad G_i = \sum_{j=1}^i \tilde{g}_j, \quad i = 1, \ldots, N,
  \]
  where \( \tilde{f}_j \) and \( \tilde{g}_j \) are the \( j \)-th components of \( \tilde{f} \) and \( \tilde{g} \) in [2], respectively.

The discrete quadratic Wasserstein distance between the two signals then reads

\[
\lVert (f, g) \rVert_{W^2} = \sum_{i=1}^N |t_i - T_i|^2 \tilde{f}_i, \quad T = G^{-1} \circ F.
\]

We note that \( T = G^{-1} \circ F \) is the “optimal” map from the density \( \tilde{f} \) to the density \( \tilde{g} \). Intuitively, if we view the two densities as two piles of dirt with the same unit mass, the map would correspond to the minimum “cost” of turning one pile into the other, i.e. the minimum amount of dirt that needs to be moved times the distance it has to be moved.
This corresponds to the horizontal distance between the two discrete CDFs \( F \) and \( G \). For computing the metric \( d_W \) in (4), the map \( T \) is computed by interpolation.

We then consider an exponential likelihood function

\[
L_{\text{exp}}(\theta) = \pi_{\text{exp}}(g|\theta) = s^N \exp(-s d_W(f(\theta), g)),
\]

where \( s > 0 \) is a rate parameter and will be considered as a hyperparameter in the MCMC sampling algorithm.

It is to be noted that the Wasserstein metric can be computed in any dimension through the (numerical) solution of an Monge–Ampère equation, see e.g. [2]. For dimensions larger than one, where there are more than one signal, this may be a costly procedure. We therefore follow the work by Yang et. al. [20] and use a “trace-by-trace” approach, summing up one dimensional quadratic Wasserstein distances; see Section 5 where there are more than one signal collected by several receivers.

### 4 Numerical algorithm

We employ a Metropolis-Hastings-within-Gibbs sampling strategy (see e.g. [7]) within the proposed Bayesian framework. The algorithm, summerized in Algorithm 1, iteratively generates a sequence of samples, forming a Markov chain, whose distribution approaches the target distribution of parameters in the limit. A candidate sample is produced based on the current sample value and is later accepted or rejected with some probability \( \alpha \); see below.

The algorithm consists of two interactive parts: a Gibbs sampler and a Metropolis-Hastings sampler. The Gibbs sampler is used to sample the hyperparameter \( s > 0 \) with a fixed parameter vector \( \theta \), and the Metropolis-Hastings sampler is used to sample \( \theta \) with a fixed hyperparameter \( s \).

**Gibbs sampler.** The Gibbs sampler employs the proposed exponential likelihood (5) and a Gamma prior on its rate parameter \( s \), i.e. \( s \sim \text{Gamma}(a,b) \), where \( a > 0 \) and \( b > 0 \) are the shape and rate parameters of the prior. Given a fixed \( \theta \) we will then have

\[
\pi(s|\theta, g) \propto \pi_{\text{exp}}(g|\theta, s) \pi_{\text{prior}}(s) \propto s^N e^{-s d_W} s^{a-1} e^{-b s} = s^{a+N-1} e^{-s(b+d_W)},
\]

where \( d_W = d_W(f(\theta), g) \) is the Wasserstein metric given by (4). Hence we obtain a Gamma posterior on \( s \):

\[
s \sim \text{Gamma}(a^*, b^*), \quad a^* = a + N, \quad b^* = b + d_W.
\]

This will be used to generate a new sample of \( s \), given a fixed \( \theta \).

**Metropolis-Hastings sampler.** The Metropolis-Hastings sampler employs the proposed exponential likelihood (5) with a fixed rate parameter \( s \) obtained by the Gibbs sampler. The posterior from which a sequence of samples are to be generated reads

\[
\pi(\theta|s, g) \propto \pi_{\text{exp}}(g|\theta, s) \pi_{\text{prior}}(\theta).
\]

Given a sample value \( \theta^{(i)} \), a new sample \( \theta^{(i+1)} \) is generated as follows. A candidate sample, say \( \tilde{\theta} \), is first generated by a proposal distribution \( q(\tilde{\theta}|\theta^{(i)}, \theta) \) from the current sample \( \theta^{(i)} \).
This candidate sample is then accepted or rejected with probability (see e.g. [3])

$$\alpha = \frac{\pi(\tilde{\theta}|s, g) q(\theta^{(i)}|\tilde{\theta})}{\pi(\theta^{(i)}|s, g) q(\theta, \theta^{(i)})} = \frac{\pi_{\text{exp}}(g|\tilde{\theta}, s) \pi_{\text{prior}}(\tilde{\theta}) q(\theta^{(i)}, \tilde{\theta})}{\pi_{\text{exp}}(g|\theta^{(i)}, s) \pi_{\text{prior}}(\theta^{(i)}) q(\theta, \theta^{(i)})}.$$ 

It is to be noted that the choice of prior would depend on the problem at hand and expert’s opinion. For instance, the prior may be informative with uniform or other types of distributions, or it may be non-informative, i.e. $\pi_{\text{prior}}(\theta) = 1$. We may even consider hierarchical priors that include other hyper-priors. For the sake of simplicity and without imposing any restriction, in the numerical examples in Section 5 we consider uniform priors on $\theta$,

$$\pi_{\text{prior}}(\theta) = \prod_{i=1}^{m} \pi_{\text{prior}}(\theta_i), \quad \theta_i \sim \text{Unif}(\alpha_i, \beta_i), \quad (6)$$

where the bounds $\{(\alpha_i, \beta_i)\}_{i=1}^{m}$ are to be selected depending on the problem at hand. We further use a Gaussian random walk proposal to generate a new sample $\tilde{\theta}$ from a current sample $\theta^{(i)}$:

$$\tilde{\theta} \sim \text{Normal}(\theta^{(i)}, \Sigma), \quad (7)$$

where $\Sigma \in \mathbb{R}^{m \times m}$ is a covariance matrix to be selected so that the proposal distribution is neither too wide nor too narrow. The former would result in an acceptance rate close to zero, and the chain would rarely move to a different sample. The latter would result in an acceptance rate close to one, however, the generated samples would not cover the whole support of posterior. We note that since such a proposal is symmetric, i.e. $q(\theta^{(i)}, \tilde{\theta}) = q(\tilde{\theta}, \theta^{(i)})$, the ratio $\alpha$ simplifies to $\alpha = \pi(\tilde{\theta}|s, g)/\pi(\theta^{(i)}|s, g)$. This ratio was used in the original version of the Metropolis algorithm [12].

5 Application to seismic inversion

In this section, we apply the proposed Bayesian framework to the inverse uncertainty quantification of wave propagation problems. More specifically, we consider two problems: source inversion and material inversion. We present four numerical examples with complicated noise structures of the form (1). We demonstrate the applicability and performance of the proposed framework in comparison with the conventional Bayesian framework based on Gaussian likelihoods.

5.1 A one-dimensional source inversion problem

The first example concerns a simple one-dimensional source inversion problem where an initial wave pulse, represented by two unknown parameters, propagates with a constant speed.

Problem formulation. Consider the Cauchy problem for the 1D wave equation

$$u_{tt}(t, x) - u_{xx}(t, x) = 0, \quad t \in [0, T], \quad x \in \mathbb{R}, \quad (8)$$

$$u(0, x) = h(x; \theta), \quad u_t(0, x) = 0. \quad (9)$$
Algorithm 1 Metropolis-Hastings-within-Gibbs Sampling in the Wasserstein-Bayesian framework

1. **Initialization**: Select an arbitrary point \((\theta^{(0)}, s^{(0)})\), and set \(i = 0\).

2. **Gibbs sampler**: Generate \(s^{(i+1)}\) from the posterior \(\pi(s|\theta^{(i)}, g)\) with Gamma distribution,
   \[
   s^{(i+1)} \sim \text{Gamma}(a^{*}, b^{*}), \quad a^{*} = a + N, \quad b^{*} = b + d_{W}(f(\theta^{(i)}), g).
   \]

3. **Metropolis-Hastings sampler**: Generate \(\theta^{(i+1)}\) as follows:
   - Sample a candidate \(\tilde{\theta}\) from a proposal distribution \(q(\theta^{(i)}, \tilde{\theta})\).
   - Compute the ratio
     \[
     \alpha(\theta^{(i)}, \tilde{\theta}) = \frac{\pi_{\text{exp}}(g|\tilde{\theta}, s^{(i+1)}) \pi_{\text{prior}}(\tilde{\theta}) q(\theta^{(i)}, \tilde{\theta})}{\pi_{\text{exp}}(g|\theta^{(i)}, s^{(i+1)}) \pi_{\text{prior}}(\theta^{(i)}) q(\theta^{(i)}, \theta^{(i)})},
     \]
     with the likelihood \(\pi_{\text{exp}}\) and the prior \(\pi_{\text{prior}}\) given in (5) and (6), respectively.
   - Set
     \[
     \theta^{(i+1)} = \begin{cases} \tilde{\theta} & \text{if } \text{Unif}(0, 1) \leq \alpha(\theta^{(i)}, \tilde{\theta}), \\ \theta^{(i)} & \text{otherwise}. \end{cases}
     \]

4. **Iteration**: Increment \(i\) by 1 and go to step 2.

The initial data

\[
    h(x; \theta) = a \left( e^{-100(x-x_0-0.5)^2} + e^{-100(x-x_0)^2} + e^{-100(x-x_0+0.5)^2} \right), \quad \theta = (x_0, a),
\]

acts as a source term generating an initial wave pulse. This model problem has two parameters: the initial location \(x_0\) and the amplitude \(a\) of the wave pulse collected in a parameter vector denoted by \(\theta\). The solution to this simple problem is given by d’Alembert’s formula

\[
    u(t, x; \theta) = \frac{1}{2} h(x-t; \theta) + \frac{1}{2} h(x+t; \theta). \tag{10}
\]

**Synthetic data.** Consider a uniform array of \(N_r = 7\) receivers located at

\[
    x_1 = -3, \quad x_2 = -2, \quad x_3 = -1, \quad x_4 = 0, \quad x_5 = 1, \quad x_6 = 2, \quad x_7 = 3.
\]

Each receiver located at \(x_r\), with \(r = 1, \ldots, N_r\), records a noisy discrete-time signal \(g(t_k, x_r)\) over the time interval \([0, T]\) at \(N\) discrete time levels \(t_k = (k - 1) \Delta t\), with \(\Delta t = T/(N - 1)\) and \(k = 1, \ldots, N\). Let \(f(t_k, x_r; \theta)\) denote the corresponding simulated signal for a given \(\theta\), computed by (10). We select a fixed parameter vector

\[
    \theta^* = (0, 5),
\]
and generate synthetic data as follows. We first compute $f(t_k, x_r; \theta^*)$ by (10) and then generate $g(t_k, x_r)$ by polluting $f(t_k, x_r; \theta^*)$ with a noise of the form (1): we multiply $f$ by i.i.d. Gamma noise and then add an i.i.d uniform noise to the product, that is,

$$
g(t_k, x_r) = \varepsilon_{kr}^{(1)} f(t_k, x_r; \theta^*) + \varepsilon_{kr}^{(2)}, \quad \varepsilon_{kr}^{(1)} \sim \text{Gamma}(60, 60), \quad \varepsilon_{kr}^{(2)} \sim \text{Unif}(-0.25, 0.25).
$$

Let $g_r \in \mathbb{R}^N$ denote the discrete-time signal recorded at $x_r$, $r = 1, \ldots, N_r$. We have a set of $N_r = 7$ such signals, shown in Figure 1 with $T = 5$ and $N = 101$. Similarly, we let $f_r(\theta) \in \mathbb{R}^N$ be the $r$-th simulated signal, consisting of $f(t_k, x_r; \theta)$ at all $N$ discrete time levels $\{t_k\}_{k=1}^N$.

![Figure 1: Observed noisy signals at seven receiver locations over time.](image)

**Computations.** The goal is to employ Bayesian inversion and compute the conditional posterior of the model parameters $\theta$. We follow Algorithm 1 with the following choices:

- **Likelihood**: $\pi_{exp}(g|\theta) = s^N \exp(-s \sum_{r=1}^{N_r} d_W(f_r(\theta), g_r))$.
- **Priors**: $\theta_1 \sim \text{Unif}(-3, 3)$, $\theta_2 \sim \text{Unif}(2, 8)$, $s \sim \text{Gamma}(1, 0.1)$.
- **Proposal**: A Gaussian random walk [7] with covariance $\Sigma = \text{diag}(0.005, 0.005)$.
- **Initial point**: $\theta^{(0)} = (0.6, 3)$, $s^{(0)} = 70$. 

8
We also consider the standard normal likelihood based on the Euclidean distance:

$$
\pi_{\text{norm}}(\mathbf{g}|\mathbf{\theta}) = s^{N/2} \exp\left(-\frac{s}{2} \sum_{r=1}^{N_r} d_2(\mathbf{f}_r(\mathbf{\theta}), \mathbf{g}_r)\right), \quad s = \sigma^{-2},
$$

where $d_2(\mathbf{f}_r(\mathbf{\theta}), \mathbf{g}_r) = \sum_{k=1}^{N} |g(t_k, x_r) - f(t_k, x_r; \mathbf{\theta})|^2$. We note that in this case $s = \sigma^{-2}$ will be the precision parameter.

We run the algorithm with $M_0 = 25000$ iterations and remove the first $M_b = 5000$ samples, known as the burn-in period. We also use a thinning period of $M_t = 4$, that is, we keep every 4th samples and discard the rest. This would give a total of $M = 5000$ Markov chain samples. Figure 2 shows the posterior histograms and the trace plots of the Markov chain samples generated by the algorithm with both (a) the proposed likelihood and (b) the Gaussian likelihood. We observe that the samples obtained by the standard Gaussian likelihood converge to wrong distributions, while the Markov chain samples obtained by the proposed exponential likelihood function are centered around the true, noise-free parameters.

The wrong convergence of the standard Bayesian approach, as shown in Figure 2, is due to the existence of multiple local maxima in the Gaussian likelihood function. Consequently, the Markov chain samples may get trapped in a local maximum and hence miss the global maximum. To illustrate this, in Figure 3 we plot the log-likelihood functions $\log L_{\text{exp}}(\theta_1, 5)$ and $\log L_{\text{norm}}(\theta_1, 5)$ versus $\theta_1 \in [-3, 3]$. We observe that the Gaussian likelihood has multiple local maxima, while the proposed exponential likelihood has a more-convex shape, making it easy for MCMC to infer its global maximum.

5.2 A two-dimensional seismic source inversion problem

This example concerns the inference of a seismic source in a layered isotropic elastic material in two dimensions. The source models a simplified earthquake, represented by four unknown parameters.

**Problem formulation.** We consider the initial-boundary value problem (IBVP) for the elastodynamic wave equation,

$$
\begin{align*}
\nu(x) u_{tt}(t, x) - \nabla \cdot \sigma(u(t, x)) &= h(t, x; \mathbf{\theta}) \quad &\text{in } [0, T] \times D, \\
u(0, x) &= 0, \quad u_t(0, x) &= 0 \quad &\text{on } \{t = 0\} \times D, \\
\sigma(u(t, x)) \cdot \hat{n} &= 0 \quad &\text{on } [0, T] \times \partial D_0, \\
\mathbf{u}_t(t, x) &= B(x) \mathbf{\sigma}(u(t, x)) \cdot \hat{n} \quad &\text{on } [0, T] \times \partial D_1.
\end{align*}
$$

The solution $\mathbf{u} = (u_1, u_2) \uparrow$ represents the displacement field, $t \in [0, T]$ is the time, $x = (x_1, x_2) \in D$ is the location, and $\mathbf{\sigma}$ is the stress tensor given by

$$
\mathbf{\sigma}(\mathbf{u}) = \lambda(x) \nabla \cdot \mathbf{u} I + \mu(x) (\nabla \mathbf{u} + (\nabla \mathbf{u})^\top),
$$

where $I$ is the identity matrix. The computational domain is a box $D = [-10000, 10000] \times [-15000, 0]$, consisting of two layers: the top layer $D_I$ extends over $-1000 \leq x_2 \leq 0$, and
Figure 2: Posterior histograms and the trace plots of the Markov chain samples in two cases: (a) with the proposed Wasserstein-driven exponential likelihood, and (b) with the standard Gaussian likelihood. Clearly, the samples obtained by the Gaussian likelihood function do not converge to the correct distribution, while the samples obtained by the proposed likelihood are accumulated close to the true, noise-free parameters.
Figure 3: The log-likelihood functions $\log L_{\text{exp}}(\theta_1, 5)$ and $\log L_{\text{norm}}(\theta_1, 5)$ versus $\theta_1 \in [-3, 3]$.

The bottom layer $D_{II}$ is given by $x_2 \leq -1000$. The material properties are characterized by the density $\nu$ and the Lamé parameters, $\lambda$ and $\mu$. The system (11) admits longitudinal (or pressure) and transverse (or shear) waves, which, in the case of constant density, propagate at the wave speeds

$$c_p = \sqrt{(2\mu + \lambda)/\nu}, \quad c_s = \sqrt{\mu/\nu},$$

respectively. The material density and velocities are assumed to be known, given by

$$\nu(x) = \begin{cases} 2600 & x \in D_I, \\ 2700 & x \in D_{II} \end{cases}, \quad c_p(x) = \begin{cases} 4000 & x \in D_I, \\ 6000 & x \in D_{II} \end{cases}, \quad c_s(x) = \begin{cases} 2000 & x \in D_I, \\ 3464 & x \in D_{II} \end{cases}.$$

We impose a homogeneous Neumann (stress-free) boundary condition (11c) on the free surface $\partial D_0 = \{x | x_2 = 0\}$, and the first-order Clayton-Engquist non-reflecting boundary conditions [4] on the other three boundaries $\partial D_1 = \partial D \setminus \partial D_0$. Here, $\hat{n}$ is the outward unit normal to the boundary, and $B$ is a given matrix related to the non-reflecting boundary condition.

The function $h$ represents the seismic source. We consider the case of a point moment tensor source that models a simplified earthquake,

$$h(t, x; \theta) = S(t) M \nabla \delta(x - x_s), \quad (13)$$

located at $x_s = (0, z_s)$, where $\nabla \delta$ is the gradient of the Dirac distribution. The source time function $S(t)$ and the moment tensor are assumed to be

$$S(t) = \frac{\omega_s}{\sqrt{2\pi}} e^{-\omega_s^2(t-t_s)^2/2}, \quad M = m_s \begin{pmatrix} 10^{10} & 0.8 \times 10^{10} \\ 0.8 \times 10^{10} & 10^{10} \end{pmatrix}.$$

parametrized by the frequency $\omega_s$, the center time $t_s$, and a constant $m_s$. Hence, the source parameter vector $\theta$ consists of four parameters,

$$\theta = (z_s, t_s, \omega_s, m_s).$$
Synthetic data. We place a uniform array of \( N_r = 90 \) receivers located at
\[ x_r = (x_{1,r}, x_{2,r}) = (-8900 + 200(r-1), 0), \quad r = 1, 2, \ldots, N_r. \]

Figure 4 shows the computational domain and the configuration of the problem. Each receiver located at \( x_r, \) with \( r = 1, \ldots, N_r, \) records a noisy vector-valued discrete-time signal \( \tilde{g}(t_k, x_r) = (g_1(t_k, x_r), g_2(t_k, x_r)) \) over the time interval \([0, T]\) at \( N \) discrete time levels \( t_k = (k-1) \Delta t, \) with \( \Delta t = T/(N-1) \) and \( k = 1, \ldots, N. \) Moreover, let \( \tilde{f}(t_k, x_r; \theta) = (f_1(t_k, x_r; \theta), f_2(t_k, x_r; \theta)) \) denote the corresponding simulated signal for a given \( \theta \) and computed by numerically solving (11). We employ a second-order accurate finite difference scheme, proposed in [14], with grid size \( \Delta x_1 = \Delta x_2 = 100 \) and time step \( \Delta t = 0.0125. \) We set the final time \( T = 8 \) with \( N = 641 \) time steps. We also employ the technique proposed in [16] for regularized approximations of the Dirac distribution and its gradient to achieve point-wise convergence of the solution away from the singular source. We refer to [11] for details of the employed forward solver.

We then choose a fixed parameter vector \( \theta^* = (-2000, 1, 4, 10^4), \)

and generate synthetic data as follows. We first numerically compute \( \tilde{f}(t_k, x_r; \theta^*) \) and then generate \( \tilde{g}(t_k, x_r) \) by polluting \( \tilde{f}(t_k, x_r; \theta^*) \) with a noise of the form (1):
\[ g_{1,2}(t_k, x_r) = \varepsilon_{kr}^{(1)} f_{1,2}(t_k, x_r; \theta^*) + \varepsilon_{kr}^{(2)}, \quad \varepsilon_{kr}^{(1)} \sim \text{Gamma}(500, 500), \quad \varepsilon_{kr}^{(2)} \sim \text{Unif}(-\frac{1}{80}, \frac{1}{80}). \]

Let \( g_r \in \mathbb{R}^N \) denote the discrete-time signal recorded at the receiver location \( x_r, \) with \( r = 1, \ldots, N_r. \) We note that since there are two (a horizontal and a vertical) components of the solution, we have a set of \( 2N_r = 180 \) such signals. Similarly, we let \( f_r(\theta) \in \mathbb{R}^N \) be the \( r \)-th simulated signal, consisting of \( f_1(t_k, x_r; \theta) \) or \( f_2(t_k, x_r; \theta) \) at all \( N \) discrete time levels \( \{t_k\}_{k=1}^N. \)

Computations. The goal is to employ Bayesian inversion and compute the conditional posterior of the model parameters \( \theta. \) We follow Algorithm 1 with the following choices:

1. Initialize the prior distribution of the model parameters \( \theta. \)
2. Generate synthetic data \( g_r \) for \( r = 1, \ldots, N_r. \)
3. For each \( t_k \), compute the forward model \( f_r(\theta) \) for \( r = 1, \ldots, N_r. \)
4. Update the posterior distribution of \( \theta \) using Bayes' theorem.
5. Repeat steps 2-4 until convergence.

Let \( \theta^\text{post} \) be the mode of the posterior distribution of the model parameters. We can then use \( \theta^\text{post} \) as a solution to the inverse problem.
• **Likelihood**: \[ \pi_{\exp}(\mathbf{g}|\mathbf{\theta}) = s^N \exp(-s \sum_{r=1}^{2N} d_W(f_r(\mathbf{\theta}), \mathbf{g}_r)). \]

• **Priors**: \[ \theta_1 \sim \text{Unif}(-2200, -1800), \quad \theta_2 \sim \text{Unif}(0.5, 1.5), \quad \theta_3 \sim \text{Unif}(2, 6), \quad \theta_4 \sim \text{Unif}(9600, 10400), \quad s \sim \text{Gamma}(1, 0.1). \]

• **Proposal**: A Gaussian random walk \([7]\) with covariance \(\Sigma\), which is initially chosen to be diagonal and then updated based on the first 1000 MCMC samples.

• **Initial point**: \(\mathbf{\theta}^{(0)} = (-2100, 0.7, 3, 9800), \quad s^{(0)} = 2000.\)

We run the algorithm with \(M_0 = 10000\) iterations and remove the first \(M_b = 2000\) samples. We also use a thinning period of \(M_t = 4\). This would give a total of \(M = 2000\) Markov chain samples. Figure 5 shows the posterior histograms and the trace plots of the Markov chain samples generated by the algorithm.

![Figure 5: Posterior histograms and the trace plots of the Markov chain samples computed by the proposed Wasserstein-driven Bayesian technique.](image)

It is to be noted that similar to the first numerical example, we observed wrong convergence of the samples using the standard Gaussian likelihood function.

### 5.3 A two-dimensional material inversion problem

This example concerns the inference of wave speed in a two-dimensional spatial domain, consisting of ten non-overlapping sub-domains, each with a different unknown wave speed.

**Problem formulation.** Consider the initial-boundary value problem for the acoustic wave equation in a two-dimensional spatial domain \(D = [-1, 1] \times [-2, 0]::

\[
\begin{align*}
\nabla_x \cdot (a^2(x; \mathbf{\theta}) \nabla_x u(t, x)) &= F_s(t, x), & t \in [0, T], & x \in D, \\
u(t, x) &= 0, & u_t(t, x) &= 0, & t = 0, & x \in D \\
u(t, x) &= 0, & u(t, x) &= 0, & t \in [0, T], & x \in \partial D.
\end{align*}
\]

(14)
The source $F_s$ is a Ricker wavelet applied on a square $D_s$ of size $0.08 \times 0.08$ centered at a given point $(x_{1,s}, x_{2,s})$:

$$F_s(t, \mathbf{x}) = \begin{cases} 
10^3 (1 - 20 (t - 0.1)^2) \exp(-10 (t - 0.1)^2), & \mathbf{x} \in D_s, \\
0, & \mathbf{x} \not\in D_s.
\end{cases}$$

The wave speed $a$ is assumed to be described by ten (unknown) parameters $\theta = (\theta_1, \ldots, \theta_{10})$.

The computational domain $D$ is split into ten non-overlapping sub-domains of size $0.4 \times 1$, each with a constant velocity $a = \theta_i$, with $i = 1, \ldots, 10$; see Figure 6.

**Synthetic data.** The acquisition geometry consists of $N_s = 5$ sources $F_s$ evenly centered at depth $x_2 = -0.2$ along the horizontal line $x_1 \in [-0.8, 0.8]$, and $N_r = 201$ evenly spaced receivers at the surface ($x_2 = 0$) with a spacing of 0.02 and the left ($x_1 = -1$) and right ($x_1 = 1$) boundaries with a spacing of 0.04. Figure 6 shows the computational domain and the configuration of the problem.

![Figure 6: A schematic representation of the computational domain, the wave speed structure, and the arrays of sources (red ×) and the uniform array of receivers (black ▽).](image)

Each receiver located at $\mathbf{x}_r$, with $r = 1, \ldots, N_r$, records a (noisy) discrete-time signal $g_s(t_k, \mathbf{x}_r)$, due to a single source $F_s$, with $s = 1, \ldots, N_s$, and over the time interval $[0, T]$ at $N$ discrete time levels $t_k = (k - 1) \Delta t$, with $\Delta t = T/(N - 1)$ and $k = 1, \ldots, N$. Moreover, let $f_s(t_k, \mathbf{x}_r; \theta)$ denote the corresponding simulated signal for a given $\theta$, obtained by employing a finite difference scheme based on the second-order central difference discretization of (14); see [13] for details of the deterministic solver. We use a uniform grid with spatial grid-lengths $\Delta x_1 = \Delta x_2 = 0.02$ and time step $\Delta t = 0.005$. We set the final time $T = 4$ with $N = 801$ time steps. We choose a fixed parameter vector

$$\theta^* = (3, 2, 3.5, 2.5, 4, 3, 4.5, 3.5, 5, 4),$$

14
and generate synthetic data by first computing \( f_s(t_k, x_r; \theta^*) \) and then polluting it with a noise of the form (1):

\[
g_s(t_k, x_r) = \varepsilon^{(1)}_{skr} f_s(t_k, x_r; \theta^*) + \varepsilon^{(2)}_{skr}, \quad \varepsilon^{(1)}_{skr} \sim \text{Gamma}(1000, 1000), \quad \varepsilon^{(2)}_{skr} \sim \text{Unif}(-0.05, 0.05).
\]

Let \( g_{sr} \in \mathbb{R}^N \) and \( f_{sr}(\theta) \in \mathbb{R}^N \) denote the vectors of recorded and simulated discrete-time signals due to a single source \( F_s \), \( s = 1, \ldots, N_s \), at the receiver location \( x_r, r = 1, \ldots, N_r \). Note that since there are \( N_s = 5 \) single sources, we have in total \( N_s N_r = 1005 \) such signals.

**Computations.** The goal is to employ Bayesian inversion and compute the conditional posterior of the model parameter vector \( \theta = (\theta_1, \ldots, \theta_{10}) \). We follow Algorithm 1 with the following choices:

- **Likelihood:** \( \pi_{\text{exp}}(g | \theta) = s^N \exp(-s \sum_{s=1}^{N_s} \sum_{r=1}^{N_r} d_W(f_{sr}(\theta), g_{sr})) \).
- **Priors:** \( \theta_j \sim \text{Unif}(1, 6), \quad j = 1, \ldots, 10, \quad s \sim \text{Gamma}(1, 0.1) \).
- **Proposal:** A Gaussian random walk (7) with covariance \( \Sigma \), which is initially chosen to be diagonal and then updated based on the first 1000 MCMC samples.
- **Initial point:** \( \theta_j^{(0)} = 3.8, \quad j = 1, \ldots, 10, \quad s^{(0)} = 5000 \).

We run the algorithm with \( M_0 = 80000 \) iterations and remove the first \( M_b = 65000 \) samples. We also use a thinning period of \( M_t = 3 \). This would give a total of \( M = 5000 \) Markov chain samples. Figure 7 shows the posterior histograms of the Markov chain samples generated by the proposed algorithm.

Figure 8 shows a map of the original noise-free parameters (a), the mean map (b) and the standard deviation map (c) obtained by the generated MCMC samples of the parameter posteriors.

6 Conclusion

We have presented a robust Bayesian framework based on a new exponential likelihood function driven by the quadratic Wasserstein metric. We demonstrated that this framework is able to treat complicated noise structures and presented several numerical examples subject to combined additive and multiplicative noise. As in the deterministic setting we observe that the convexity properties of the Wasserstein metric translates into better convergence, here towards the correct posteriors.

It is to be noted that in this work we exclusively used a “shift and rescale” approach to convert signals into positive probabilities. Although this procedure was shown to produce good results, there are many other alternatives, e.g. taking the exponential of the signal, squaring the signal etc., that could prove to be more suitable for certain problems.

As the framework presented here is particularly suitable for wave propagation problems we plan to apply it to geoacoustic inversion of water and seafloor parameters in shallow water range dependent environments. There the material properties vary smoothly and consequently the forward problems can be efficiently simulated with our arbitrary order wave solvers [1].
Figure 7: Posterior histograms of the Markov chain samples generated by the proposed Wasserstein-driven Bayesian algorithm.

Figure 8: Map of the original noise-free parameters (a), the mean map (b) and the standard deviation map (c) obtained by the generated MCMC samples of the parameter posteriors.

Acknowledgements. The authors are indebted to Dr. Gabriel Huerta at Sandia National Laboratories for his guidance and stimulating discussions.
References

[1] D. Appelö, T. Hagstrom, and A. Vargas. Hermite methods for the scalar wave equation. SISC (in revision), 2018.

[2] J.-D. Benamou, B. D. Froese, and A. M. Oberman. Numerical solution of the Optimal Transportation problem using the Monge-Ampère equation. Journal of Computational Physics, 260:107–126, 2014.

[3] S. Chib and E. Greenberg. Understanding the Metropolis-Hastings algorithm. The American Statistician, 49:327–335, 1995.

[4] R. Clayton and B. Engquist. Absorbing boundary conditions for acoustic and elastic wave equations. Bull. Seismol. Soc. Amer., 67:1529–1540, 1977.

[5] P. Mohajerin Esfahani and D. Kuhn. Data-driven distributionally robust optimization using the Wasserstein metric: performance guarantees and tractable reformulations. Mathematical Programming, 2017.

[6] D. Gamerman and H. F. Lopes. Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference. Chapman and Hall/CRC, 2006.

[7] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. Bayesian data analysis. Chapman and Hall/CRC, 2004.

[8] J. Goodman. Speckle Phenomena in Optics: Theory and Applications. Roberts and Company Publishers, 2007.

[9] J. Kaipo and E. Somersalo. Statistical and Computational Inverse Problems. Springer, New York, 2005.

[10] S. Kolouri, Y. Zou, and G. K. Rohde. Sliced Wasserstein kernels for probability distributions. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 5258–5267, June 2016.

[11] Q. Long, M. Motamed, and R. Tempone. Fast Bayesian optimal experimental design for seismic source inversion. Computer Methods in Applied Mechanics and Engineering, 291:123–145, 2015.

[12] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. The Journal of Chemical Physics, 21:1087–1092, 1953.

[13] M. Motamed, F. Nobile, and R. Tempone. A stochastic collocation method for the second order wave equation with a discontinuous random speed. Numer. Math., 123:493–536, 2013.
[14] S. Nilsson, N. A. Petersson, B. Sjögreen, and H.-O. Kreiss. Stable difference approximations for the elastic wave equation in second order formulation. *SIAM J. Numer. Anal.*, 45:1902–1936, 2007.

[15] N. Papadakis. Optimal transport for image processing. Habilitation, Signal and Image Processing, Université de Bordeaux, 2015. HAL Id: tel-01246096.

[16] B. Sjögreen and N. A. Petersson. Source estimation by full wave form inversion. *Journal of Scientific Computing*, 59:247–276, 2014.

[17] A. M. Stuart. Inverse problems: A Bayesian perspective. *Acta Numerica*, 19:451–559, 2010.

[18] C. Villani. *Topics in Optimal Transportation*, volume 58 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, Rhode Island, 2003.

[19] C. Villani. *Optimal Transport: Old and New*, volume 338 of *Comprehensive Studies in Mathematics*. Springer-Verlag, Berlin, 2009.

[20] Y. Yang, B. Engquist, J. Sun, and B. F. Hamfeldt. Application of optimal transport and the quadratic Wasserstein metric to full-waveform inversion. *Geophysics*, 83(1):R43–R62, 2018.