The convexity of optimal transport-based waveform inversion for certain structured velocity models

Srinath Mahankali*

Stuyvesant High School
New York, NY 10282

Faculty Advisor: Dr. Yunan Yang†
Courant Institute of Mathematical Sciences
New York, NY 10012-1185

Abstract

Full–waveform inversion (FWI) is a method used to determine properties of the Earth from information on the surface. We use the squared Wasserstein distance (squared W_2 distance) as an objective function to invert for the velocity of seismic waves as a function of position in the Earth, and we discuss its convexity with respect to the velocity parameter. In one dimension, we consider constant, piecewise increasing, and linearly increasing velocity models as a function of position, and we show the convexity of the squared W_2 distance with respect to the velocity parameter on the interval from zero to the true value of the velocity parameter when the source function is a probability measure. Furthermore, we consider a two–dimensional model where velocity is linearly increasing as a function of depth and prove the convexity of the squared W_2 distance in the velocity parameter on large regions containing the true value. We discuss the convexity of the squared W_2 distance compared with the convexity of the squared L^2 norm, and we discuss the relationship between frequency and convexity of these respective distances. We also discuss multiple approaches to optimal transport for non–probability measures by first converting the wave data into probability measures.

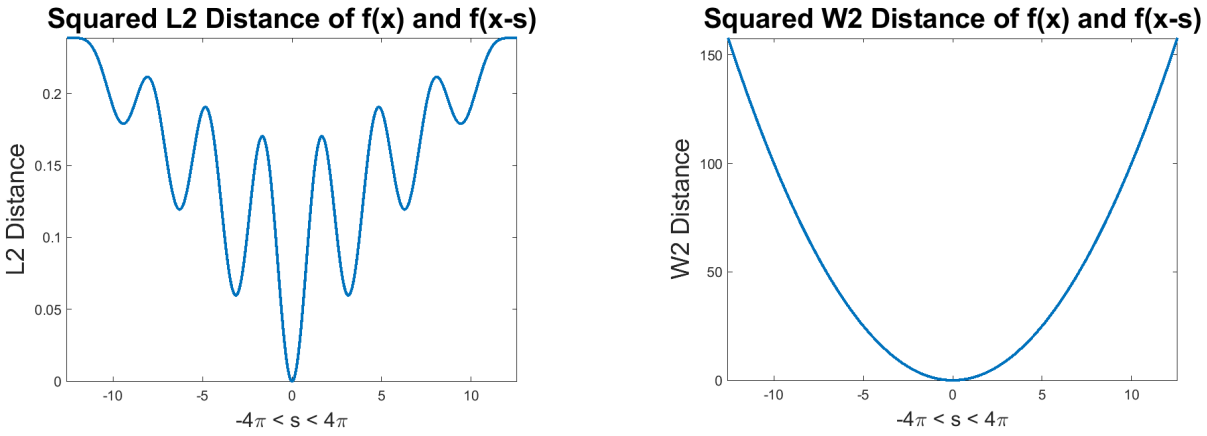
1 Introduction

The study of seismic waves has many practical applications in geology, especially in searching for natural resources such as oil or natural gas. It plays a major role in determining the type of material underground, given the position of several receivers on the surface and the amount of time it takes for the wave to rebound to the surface. The velocity of the wave (as a function of its position) is unknown, and finding the wave velocity function is equivalent to finding the underground substance. With the same wave source, different wave velocity properties produce different wave data (such as wave amplitude and travel time) measured at a given receiver. This wave data can be used to find the velocity. We use an objective function, or misfit function, which measures the "distance" between two sets of wave data. This allows us to compare the observed data with simulated data to find the true velocity function. Some examples of objective functions are the L^p distance and the pth Wasserstein distance (W_p distance) from the theory of optimal transport [20], the latter of which is the main tool of this research project.

The objective function becomes zero when the observed data and simulated data are equivalent, which occurs when we have the correct velocity model. Thus, finding the correct velocity model is an

*smahankali10@stuy.edu
†yunan.yang@nyu.edu

Copyright © SIAM
Unauthorized reproduction of this article is prohibited



(a) Squared L^2 metric as a plot in s

(b) Squared W_2 distance as a plot in s

Figure 1: Comparison between squared W_2 distance of $f(x)$ and $f(x-s)$ and squared L^2 norm of $f(x) - f(x-s)$ as plots in the shift s.

optimization problem: minimizing the objective function, which measures the error between simulated and observed wave data. Although it is necessary to have a gradient of zero to minimize the objective function, this is not enough, as it is possible to reach a saddle point or a local minimum at such a point. However, this issue is fixed if the objective function is convex, as it will have only one global minimum. Thus, it is important for the convex region near the global minimum to be as large as possible. For this reason, we investigate the convexity of the squared W_2 distance as an objective function.

The conventional choice of objective function is the least–squares (L^2) norm, used in both time [19] and frequency [15, 16] domains. If we have data from multiple receiver locations X_r (where r is the index for the receiver location), we can consider the squared L^2 norm of the difference between the predicted wave data $g(X_r, t, c)$ and the observed wave data $h(X_r, t) = g(X_r, t, c^*)$

$$\mathcal{L}(c) = \frac{1}{2} \sum_r \int_0^{\mathcal{T}} |g(X_r, t, c) - h(X_r, t)|^2 \, \mathrm{d}t, \tag{1.1}$$

from [2], where c^* is the true velocity parameter. While $\mathcal{L}(c)$ is minimized (and therefore equal to zero) exactly when $c = c^*$, algorithms for minimizing the squared L^2 norm may reach local minima instead of the global minimum when $c = c^*$, due to the nonconvexity of the squared L^2 norm as discussed in [2, 22]. In addition, its sensitivity to noise can make it an unsuitable choice of objective function [3]. Thus, we use the squared W_2 distance from [5, 22, 23] instead. If there are multiple receiver locations X_r, our final objective function will be

$$\mathcal{W}(c) = \sum_r W_2^2(g(X_r, t, c), h(X_r, t)). \tag{1.2}$$

Previous results show that the squared W_2 metric is jointly convex in translations and dilations of the data [23], suggesting that the squared W_2 metric is a suitable choice for the objective function. Taking the function

$$f(x) = \begin{cases} \frac{1}{2\pi} \sin^2(x), & -2\pi \leq x \leq 2\pi, \\ 0, & \text{otherwise.} \end{cases}$$

as an example, we compare in Figure 1 the graphs of the squared L^2 norm of $f(x) - f(x-s)$ with the squared W_2 distance between $f(x)$ and $f(x-s)$. It can be observed that the squared W_2 distance is convex in the shift s while the graph of the squared L^2 norm is not.

In this paper, we present a theoretical approach to velocity inversion using optimal transport by investigating the convexity of the squared W_2 distance as a function of the velocity parameters – this has

not been studied theoretically before. We investigate several velocity models in one dimension and show that the squared W_2 distance is a suitable objective function when inverting for the velocity parameter. In two dimensions, we consider a particular velocity model, and we show that the squared W_2 distance is a suitable objective function in the case where the source function f is nonnegative. We generalize to when the source function alternates between negative and positive values, and we show that the squared W_2 distance is a suitable objective function given certain requirements on f. These theorems suggest that the squared W_2 distance is a suitable objective function when $\inf f$ is close enough to zero. Numerical evidence suggests that the squared W_2 distance is a better objective function than the squared L^2 norm when the source function is nonnegative.

The paper is structured as follows. In Section 2, we discuss background knowledge which is used later on in this paper. First, we introduce optimal transport and the W_p distance, along with an explicit formula as well as some of its properties. Furthermore, we discuss the solution of wave equations in one dimension and introduce *ray tracing*, which is used to solve wave equations in higher dimensions. In Section 3, we first consider one–dimensional velocity models, where the velocity is either constant, piecewise constant, or linearly increasing as a function of position, and the source function is a probability measure. Then, we study a two–dimensional velocity model where the wave velocity v satisfies $v(X, z) = a + bz$ where a and b are positive constants, X is the horizontal position, and z is the depth of the wave. Initially, in Section 3.3.1, we assume that the source function f is a probability measure and that the wave amplitude is unchanged, and we show Theorem 4. In Section 3.3.2, we involve the wave amplitude and we allow the source function to alternate between positive and negative values, and we show Theorem 5. In Section 4, we compare the convexity of the squared W_2 distance with the squared L^2 norm, and we include numerical examples. We also discuss the relationship of the W_2 distance with the $\dot{\mathcal{H}}^{-1}$ norm, and we discuss multiple approaches to optimal transport for non–probability measures. We summarize this paper in Section 5 and discuss a possible direction for future research.

2 Background

We introduce optimal transport, and essential background on wave equations.

2.1 Optimal Transport

In this section, we establish the main goal of optimal transport (originally introduced by Monge [14]) and introduce the W_p distance along with some of its useful properties. Optimal transport involves probability spaces, which are nonnegative measure spaces with total measure equal to one. We discuss the convexity of the squared W_2 metric, which is a distance between two probability measures on a probability space, and introduce some of its properties.

2.1.1 General Problem

Consider two distinct probability measures μ and ν defined on the Borel sets of $\mathbb{R}^n$. The goal of optimal transport is to find a map $T : \mathbb{R}^n \to \mathbb{R}^n$, shown in Figure 2, which minimizes the total cost of mapping μ to ν according to the map T, for a given cost function $c : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ [1]. The W_p metric, based on optimal transport [24], gives the optimal transportation cost when the cost function is $c(x, y) = |x - y|^p$.

2.1.2 Computing the W_2 metric in one dimension

Let μ and ν be probability measures defined on the Borel sets of $\mathbb{R}^n$. We define the W_p distance as

$$W_p(f, g) = \left(\inf_{T \in \mathbb{M}(\mu,\nu)} \int_\Omega |x - T(x)|^p \mathrm{d}\mu \right)^{\frac{1}{p}}$$

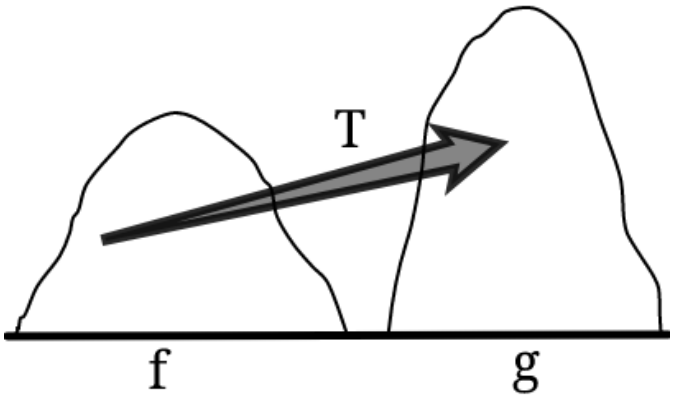


Figure 2: The two functions f and g have the same total area, and T maps all points of f to g. The goal of optimal transport is to find T such that the cost function is minimized.

where $f(x)\mathrm{d}x = \mathrm{d}\mu$, $g(y)\mathrm{d}y = \mathrm{d}\nu$, Ω is the support of μ, and $\mathbb{M}(\mu, \nu)$ is the set of all mass–preserving maps $T : \mathbb{R}^n \to \mathbb{R}^n$ which map μ to ν. The integral within the infimum is the total cost of the transport map T, so computing $W_p(f, g)$ is equivalent to minimizing the transport cost. We study the case where $p = 2$, and our objective function is $W_2^2(f, g)$.

In one dimension, it is possible to express $W_2^2(f, g)$ for probability measures f and g in a simpler way: Let F and G be the cumulative distribution functions of f and g, respectively. Then, Rachev and Rüschendorf derive the formula for the squared W_2 distance in one dimension [17] as

$$W_2^2(f, g) = \int_0^1 (F^{-1}(s) - G^{-1}(s))^2 \, \mathrm{d}s = \int_\Omega (t - G^{-1}(F(t)))^2 f(t) \, \mathrm{d}t \tag{2.1}$$

where Ω is the domain of f. This formula for $W_2^2(f, g)$ is useful when the wave data is only a function of time.

As wave data are not usually probability density functions, we can normalize a function $k(t)$ defined on $[0, \mathcal{T}]$ by replacing it with

$$\frac{k(t) + \gamma}{\int_0^{\mathcal{T}} (k(t) + \gamma) \, \mathrm{d}t}$$

for some constant $\gamma > 0$ such that $k(t) + \gamma > 0$ for all $t \in [0, \mathcal{T})$ [11].

2.1.3 Computing the W_2 Metric in Higher Dimensions

In general, there is no explicit formula to compute the W_2 metric in higher dimensions. However, certain requirements derived from the concept of cyclical monotonicity [13] make it possible to calculate the optimal map T, and therefore the W_2 metric, through numerical methods. This is shown in the following theorem of Brenier [4, 6, 22]:

Theorem 1 (Brenier's theorem)**.** *Let μ and ν be two compactly supported probability measures on $\mathbb{R}^n$. If μ is absolutely continuous with respect to the Lebesgue measure, then there is a convex function $w : \mathbb{R}^n \to \mathbb{R}$ such that the optimal map T for the cost function $c(x, y) = |x-y|^2$ is given by $T(x) = \nabla w(x)$ for μ–almost every x.*

Furthermore, if $\mu(\mathrm{d}x) = f(x)\mathrm{d}x, \nu(\mathrm{d}y) = g(y)\mathrm{d}y$, then T is differentiable μ–almost everywhere and

$$\det(\nabla T(x)) = \frac{f(x)}{g(T(x))}, \tag{2.2}$$

from the mass preserving property of T. Replacing $T(x)$ in Equation (2.2) with $\nabla w(x)$ leads to the *Monge–Ampère* equation

$$\det(D^2 w(x)) = \frac{f(x)}{g(\nabla w(x))},$$

where $D^2 w(x)$ is the Hessian matrix of w. Then, the squared W_2 distance satisfies

$$W_2^2(f,g) = \int_\Omega |x - \nabla w(x)|^2 f(x)\, \mathrm{d}x,$$

where Ω is the domain of f.

2.1.4 Properties of the Squared W_2 Distance

Results about the convexity of the squared W_2 distance with respect to changes in the data are known [9, 22, 23].

Theorem 2. *Let f and g be compactly supported probability density functions on an interval $\Omega \subset \mathbb{R}$. Then,*

$$W_2^2(f(t-s), g(t)) = W_2^2(g(t), f(t)) + s^2 + 2s \int_\Omega (x - T(x)) f(x)\, \mathrm{d}x \tag{2.3}$$

where $s \in \mathbb{R}$ and T is the optimal map from f to g. Furthermore, $W_2^2(f(t), Af(At-s))$ is convex in both A and s, for $A \in \mathbb{R}^+$.

The convexity with respect to shifts and dilations suggests that the squared W_2 distance is more suitable when inverting for wave data.

2.2 Background on Wave Equation

We introduce the partial differential equation which governs the behavior of n–dimensional waves. We also introduce d'Alembert's solution to the one–dimensional wave equation and present the *ray tracing* approach to solving higher dimensional wave equations.

2.2.1 General Wave Equation

An n–dimensional wave can be expressed as a function ψ of n position variables $x_1, x_2, \ldots, x_n$ and time t, which in general satisfies the partial differential equation

$$\frac{\partial^2 \psi}{\partial t^2} - \mathcal{C}(\mathbf{x})^2 \left(\frac{\partial^2 \psi}{\partial {x_1}^2} + \frac{\partial^2 \psi}{\partial {x_2}^2} + \cdots + \frac{\partial^2 \psi}{\partial {x_n}^2} \right) = f(\mathbf{x}, t)$$

for a variable coefficient $\mathcal{C}(\mathbf{x})$ which is a function defined on $\mathbb{R}^n$, and a source function f which is a function of both space and time. In general, the wave equation might not have an analytical solution.

2.2.2 Solution to 1D Wave Equation

In one dimension, we consider a simple case where $\mathcal{C}(\mathbf{x})$ is equal to a constant c. The partial differential equation becomes

$$\frac{\partial^2 \psi}{\partial t^2} = c^2 \frac{\partial^2 \psi}{\partial x^2},$$

but unlike its n–dimensional variant, it is possible to obtain an explicit solution as derived by d'Alembert [7]:

$$\psi(x,t;c) = \frac{j(x+ct) + j(x-ct)}{2} + \frac{1}{2c} \int_{x-ct}^{x+ct} k(s) ds \tag{2.4}$$

given the initial conditions $\psi(x,0) = j(x)$ and $\psi_t(x,0) = k(x)$.

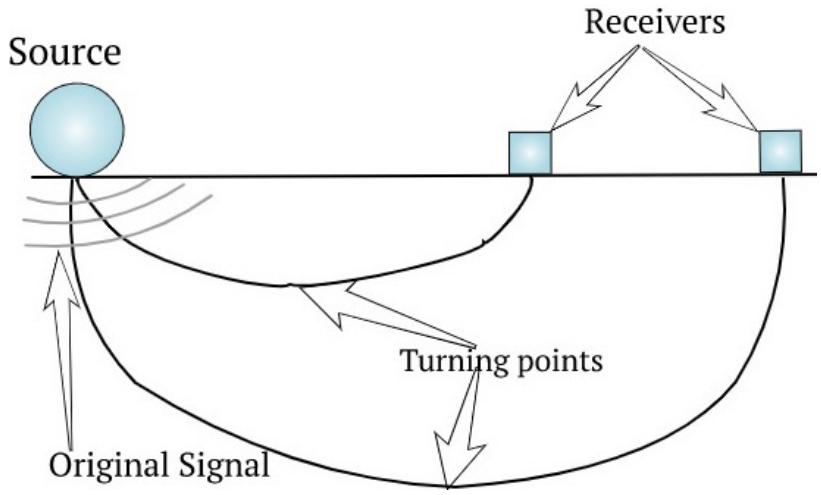


Figure 3: The method of *ray tracing* tracks the path of a ray passing through the Earth, and the paths of two rays are shown here. Each ray leaves the source location, reaches a turning point, and then returns to the surface at the indicated receiver location.

2.2.3 Solution for Higher Dimensional Wave Equation

In higher dimensions, however, a wave equation might not have an analytical solution, so we use *ray tracing* to obtain information about the wave function [18]. An example of ray tracing is shown in Figure 3, where an original signal is released, and two receivers on the surface are present to measure the wave data. We assume that the velocity v of a wave can be expressed as a function of depth z in the Earth, and the ray parameter p, or the horizontal slowness, can be expressed as $u(z)\sin\theta$ by Snell's law where $u(z) = \frac{1}{v(z)}$ and θ is the angle the ray makes with a vertical axis. The ray parameter is constant throughout the path of the ray. We also define the vertical slowness $\eta(z)$ as $\sqrt{u(z)^2 - p^2}$. The path of the ray is symmetric about a vertical line passing through a turning point at depth z_p, and $u(z_p) = p$ while $\eta(z_p) = 0$. By examining a ray passing through several layers in the Earth and eventually returning to the surface, we may calculate the horizontal distance and the traveltime of the wave as a function of the velocity. We can calculate

$$X = 2p \int_0^{z_p} \frac{dz}{\sqrt{u^2(z) - p^2}} \tag{2.5}$$

as the distance from the source to the receiver, and

$$T = 2 \int_0^{z_p} \frac{u^2(z)}{\sqrt{u^2(z) - p^2}} dz \tag{2.6}$$

as the total travel time. Using Equation (2.5), it is possible to solve for p in terms of X and the velocity parameters. It is also possible to use Equation (2.6) to solve for T as a function of X and the velocity parameters, and it is possible to calculate the predicted time T_{pred} by guessing the velocity parameter. The method of *traveltime tomography* uses the above formulas to calculate the predicted time and approaches the problem as minimizing the squared difference between the predicted time and the observed time: $(T_{pred} - T_{obs})^2$ [25].

However, this method does not work when the velocity model is not continuous [18]. In addition to using Equations (2.5) and (2.6), we make use of the wave's amplitude as well to deal with discontinuous velocity models and other issues that traveltime tomography runs into. This method is known as *full–waveform inversion* (FWI), where both the amplitude and the traveltime are used to approximate the properties of the Earth [21].

The final amplitude is asymptotically scaled by a factor of

$$A = \left(2 \int_0^{z_p} \frac{u(z)}{\sqrt{u^2(z) - p^2}} dz \right)^{-1}, \tag{2.7}$$

which is the reciprocal of the total arc length of the ray's path [18], and the ray's path is symmetric about the turning point at depth z_p. Thus, if the source function is $f(t)$, then the observed wave function is approximated by $Af(t - T_{obs})$. To invert for the velocity we use an objective function such as Equation (1.1) and Equation (1.2) to compare observed data with simulated data and minimize it using standard algorithms, and we will see that the squared W_2 metric can be a suitable choice.

3 Convexity in the Model Parameter

In this section, we study multiple velocity models in one dimension as well as a model in two dimensions, and we prove convexity of the squared W_2 distance with respect to the velocity parameter on certain domains. First, we consider one–dimensional velocity models. We begin with a model with constant velocity c and prove convexity of the squared W_2 distance with respect to c. We then consider two models with piecewise increasing velocities with respect to distance from the source and a model where velocity is linearly increasing. In every one–dimensional velocity model, we assume that the source function is nonnegative. Finally, we consider a two–dimensional model where the velocity v satisfies $v(X, z) = a + bz$ where a and b are positive constants, X is the horizontal position of the ray, and z is the current depth of the ray. In Section 3.3.1, we assume that the source function is nonnegative, which allows us to use Theorem 2 when computing the squared W_2 distance. After this, we consider a more general case where the source function is alternating in Section 3.3.2, and the predicted wave function has an amplitude which is a function of a, b, and the receiver location X. We also use the following result [12]:

Lemma 1. *Let $P : \Omega_1 \to \mathbb{R}$ and $Q : \Omega_2 \to \Omega_1$ be convex functions where $\Omega_1 \subseteq \mathbb{R}$ and $\Omega_2 \subseteq \mathbb{R}^n$ are convex sets and $n \geq 1$. Furthermore, assume P is nondecreasing. Then, $P(Q(\mathbf{x}))$ is a convex function on Ω_2.*

An example of this is when $P(x) = x^2$. Then, if Q is convex and nonnegative, we see that Q^2 is also convex on its domain.

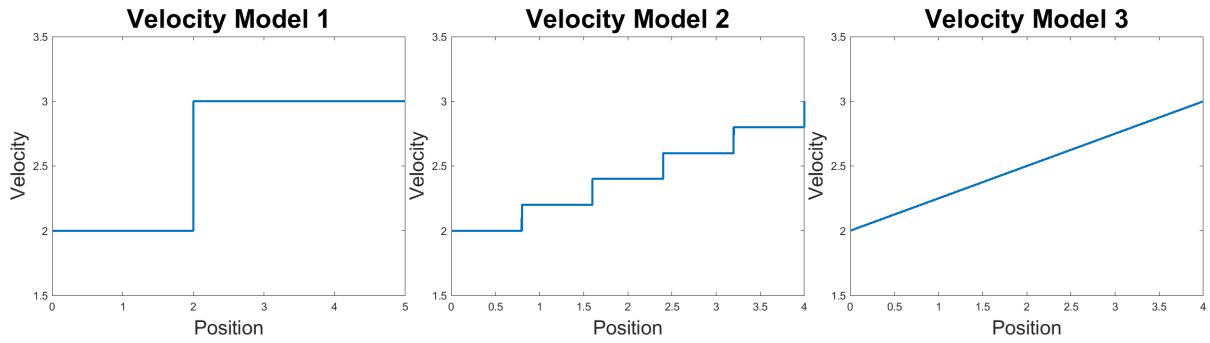
3.1 Constant Velocity in One Dimension

Let c be the constant wave velocity, and let our initial wave function be $f(t)$. For this section, we assume that $f(t)$ is a probability distribution. We assume that the final wave function at a fixed spatial location d is of the form $f(t - T(c))$, where $T(c)$ is the amount of time it takes to receive the wave signals at location d as a function of the velocity c. Since the total distance is $d, T(c) = \frac{d}{c}$, which means the predicted wave function is $f(t - \frac{d}{c})$. Letting c^* be the true value of the velocity gives us

$$W_2^2 \left(f\left(t - \frac{d}{c}\right), f\left(t - \frac{d}{c^*}\right) \right) = \left(\frac{d}{c} - \frac{d}{c^*} \right)^2,$$

as the squared W_2 distance from Theorem 2, because the integral in Equation (2.3) is zero. In addition, $\frac{d}{c} - \frac{d}{c^*}$ is a convex function of c on $(0, \infty)$, and it is nonnegative on the interval $(0, c^*]$. Hence, by Lemma 1, the squared W_2 distance is convex on $(0, c^*]$.

Remark 1. Due to the convexity of the squared W_2 distance in the interval $(0, k^*]$, choosing a small value of k as the initial guess guarantees being able to find the true velocity parameter k^* through gradient–based optimization methods.



(a) First velocity model. (b) Second velocity model. (c) Third velocity model.

Figure 4: Three types of non–constant velocity models in one dimension that we study.

3.2 Non–constant Velocity in One Dimension

When the velocity is non–constant, the d'Alembert solution Equation (2.4) does not hold anymore. Thus, we study the convexity of the squared W_2 distance for several non–constant velocity models, shown in Figure 4. We first study a model where the velocity is piecewise constant, after which we add multiple pieces. Then, we study a model where the velocity is linearly increasing as a function of position. For all of these models, we assume that the source function $f(t)$ is a probability distribution.

3.2.1 Piecewise Constant Velocity

The first scenario we study is a piecewise constant velocity model. Assume the velocity $v(x)$ satisfies $v(x) = c_1$ for $0 \leq x \leq d_1$ and a known constant c_1, and $v(x) = c_2$ for $d_1 < x \leq d_1 + d_2$, as seen in Figure 4a – we show convexity in the unknown c_2. The total travel time, as a function of c_2, is $T(c_2) = \frac{d_1}{c_1} + \frac{d_2}{c_2}$. Letting c_2^* be the true value of c_2 gives $\left(\frac{d_2}{c_2} - \frac{d_2}{c_2^*}\right)^2$ as the squared W_2 distance by Theorem 2. As the function $\frac{d_2}{c_2} - \frac{d_2}{c_2^*}$ is convex and nonnegative on $(0, c_2^*]$, the squared W_2 distance is also convex on $(0, c_2^*]$ by Lemma 1.

3.2.2 Piecewise Constant Velocity with Multiple Pieces

The second scenario we consider is the piecewise constant velocity with n pieces of equal length d, such that the velocity of the wave is

$$v(x) = \begin{cases} c_1, & x \leq d, \\ c_1 + k, & d < x \leq 2d, \\ \vdots & \vdots \\ c_1 + (n-1)k, & (n-1)d < x \leq nd. \end{cases}$$

Here, the unknown variable is k. In this case, the total travel time is

$$T(k) = \sum_{m=0}^{n-1} \frac{d}{c_1 + mk},$$

so the squared W_2 distance becomes $(T(k) - T(k^*))^2$ by Theorem 2 where k^* is the true value of the velocity parameter. We claim that the squared W_2 distance is convex in k on the interval $[0, k^*]$. To do this, we initially show that $T(k)$ is convex in k on $[0, \infty)$. Taking the second derivative, we get

$$T''(k) = \sum_{m=0}^{n-1} \frac{2dm^2}{(c_1 + mk)^3},$$

which is nonnegative for all $k \geq 0$. Hence, $T(k)$ is convex in k on $[0, \infty)$. Since $T(k)$ is strictly decreasing on $[0, \infty)$, the function $T(k) - T(k^*)$ is nonnegative (and also convex) on the interval $[0, k^*]$. Thus, the squared W_2 distance $(T(k) - T(k^*))^2$ is convex in k on the interval $[0, k^*]$ by Lemma 1.

3.2.3 Linearly Increasing Velocity

Next, we consider a linearly increasing velocity, which is of the form $c_1 + kx$ at a position x, for a known constant c_1. Here, k is unknown and k^* is the true value of the velocity parameter. Letting d be the total travel distance, we have the following integral representation for the traveltime:

$$T(k) = \int_0^d \frac{\mathrm{d}x}{c_1 + kx} = \frac{\ln(c_1 + kd) - \ln(c_1)}{k}. \tag{3.1}$$

We first prove a lemma.

Lemma 2. *$T(k)$ is convex in k on the interval $[0, \infty)$, where $T(0) = \frac{d}{c_1}$.*

Proof. We take the second derivative of T:

$$T''(k) = \frac{\mathrm{d}^2}{\mathrm{d}k^2} \int_0^d \frac{\mathrm{d}x}{c_1 + kx} = \int_0^d \frac{\mathrm{d}^2}{\mathrm{d}k^2} \left(\frac{1}{c_1 + kx} \right) \mathrm{d}x,$$

and the second equation follows by the Leibniz Integral Rule. Because $\frac{\mathrm{d}^2}{\mathrm{d}k^2} \left(\frac{1}{c_1+kx} \right) = \frac{2x^2}{(c_1+kx)^3} \geq 0$, the integrand is nonnegative. This implies the convexity of $T(k)$. □

Now we are ready to show that $(T(k) - T(k^*))^2$ is convex on the interval $[0, k^*]$.

Theorem 3. *$(T(k) - T(k^*))^2$ is convex in k on the interval $[0, k^*]$, where $T(0) = \frac{d}{c_1}$.*

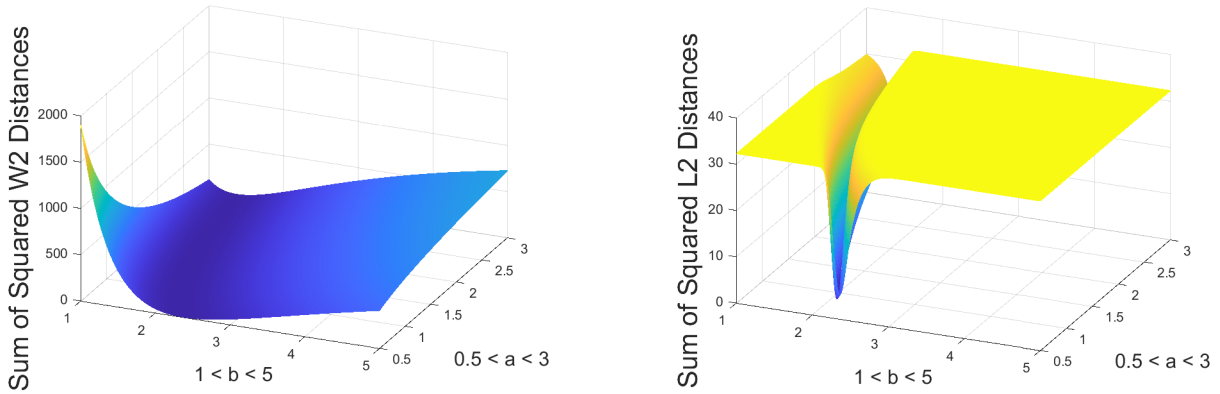
Proof. By Lemma 2, $T(k) - T(k^*)$ is convex. As the square of a nonnegative convex function is convex by Lemma 1, it is enough to determine the interval in which $T(k) - T(k^*)$ is nonnegative. From the integral representation of $T(k)$ in Equation (3.1) we see that $T(k)$ is strictly decreasing in k. Hence, $T(k) - T(k^*)$ is nonnegative on the interval $[0, k^*]$ implying the convexity of $(T(k) - T(k^*))^2$ on this interval by Lemma 1. □

3.3 A Velocity Model in Two Dimensions

Consider a velocity model in two dimensions where the predicted velocity at a point (X, z) is of the form $v(X, z) = a + bz$ where a and b are positive constants, X is the horizontal position of the ray, and z is the current depth of the ray. We analyze the travel time of a ray emanating from a single receiver which has a final distance X from the source, as well as the amplitude of the wave at a receiver. We compute the squared W_2 distance of the predicted wave function with the observed wave function (with velocity $a^* + b^*z$ at depth z) and aim to find a region in $\mathbb{R}^2$ for which this distance is convex in (a, b).

3.3.1 Constant Amplitude

While the source function and the observed wave function always have the same amplitude in one dimension, for higher dimensions this is not the case. With a source function $f(t)$, the observed wave data is of the form $Af(t-T)$ for constants A and T because there are no reflections when the velocity is a continuous function of depth. However, we analyze the convexity with the assumption that $A = 1$. We treat the observed wave data as $f(t - T(X, a^*, b^*))$ where $T(X, a, b) = T_{pred}$ is the predicted traveltime expressed as a function of a, b, and X and $T(X, a^*, b^*) = T_{obs}$ is the observed traveltime of the wave data. X is the receiver location, or the distance from the source to the receiver. Furthermore, we assume that $f(t)$ is a probability distribution with compact support in the interval $[p_1, p_2] \subset [0, \mathcal{T}]$, where $\mathcal{T} > T_{obs} + p_2$.



(a) Sum of squared W_2 distance

(b) Sum of squared L^2 norm

Figure 5: Here, $f(t) = e^{-50(t-5)^2}$ is the source function and $(a^*, b^*) = (1, 2)$ is the true value of the velocity parameter. As f is not a probability density function, the wave data, which is of the form $g(X, t, a, b) = f(t - T(X, a, b))$, is normalized according to the formula $\widetilde{g}(X, t, a, b) = \dfrac{g(X, t, a, b)}{\int_0^{\mathcal{T}} g(X, t, a, b)\ \mathrm{d}t}$ where $\mathcal{T} = 50$. The objective function $\mathcal{W}(a, b) = \displaystyle\sum_{X_r=10}^{100} W_2^2(\widetilde{g}(t - T(X_r, a, b)), \widetilde{g}(t - T(X_r, a^*, b^*)))$, where X_r is the location of receiver r, is compared with $\mathcal{L}(a, b) = \displaystyle\sum_{X_r=10}^{100} \int_0^{\mathcal{T}} |f(t - T(X_r, a, b)) - f(t - T(X_r, a^*, b^*))|^2\ \mathrm{d}t$.

When the source function is a probability distribution, this method is equivalent to traveltime tomography because the squared W_2 distance is equal to $(T_{pred} - T_{obs})^2$ by Theorem 2.

Even when the observed wave data is not a probability distribution, we may normalize the data to reduce it to this case. For example, if we normalize a nonnegative function $k(t)$ of the form $Af(t-T)$ with compact support in the interval $[0, \mathcal{T}]$ using the formula $\widetilde{k}(t) = \frac{k(t)}{\int_0^{\mathcal{T}} k(t)\ \mathrm{d}t}$, where $\widetilde{k}(t)$ is the normalized wave data, then the amplitude of the normalized data remains constant. Thus, the squared W_2 distance becomes $(T_{pred} - T_{obs})^2$. Using this expression for the squared W_2 distance, we can determine its convexity in a large region containing (a^*, b^*) and points arbitrarily close to the origin.

We first explicitly compute $T(X, a, b)$. Letting the ray parameter be p, we find that $X = \frac{2\eta_1}{bu_1 p}$ from Equation (2.5) where $u_1 = \frac{1}{a}$ is the initial slowness, and $\eta_1 = \sqrt{u_1^2 - p^2}$ is the initial vertical slowness. We can solve for p and η_1 in terms of X, a, and b as

$$p = \frac{2}{\sqrt{b^2X^2 + 4a^2}}, \qquad \eta_1 = \frac{bX}{a\sqrt{b^2X^2 + 4a^2}}.$$

From these formulas as well as Equation (2.6), we can solve for the time:

$$T(X, a, b) = \frac{2}{b}\left(\ln\left(\frac{\sqrt{b^2X^2 + 4a^2} + bX}{2a}\right)\right).$$

Then, the squared W_2 distance becomes $(T(X, a, b) - T(X, a^*, b^*))^2$ where $v_1(X, z) = a^* + b^*z$ is the true velocity function. We first claim that $T(X, a, b)$ is convex subject to a restriction on $\frac{bX}{2a}$.

Lemma 3. *Let $S_0 > 0$ be the largest root of the equation $S^2\phi(S) + 2\phi(S) - S^2 = 0$, where*

$$\phi(S) = 2\left(1 + \frac{1}{S^2}\right)^{\frac{3}{2}} \ln(\sqrt{S^2 + 1} + S) - \frac{2}{S^2} - 3.$$

Then, the traveltime $T(X, a, b)$ is jointly convex in (a, b) whenever $\frac{bX}{2a} \geq S_0$.

Proof. Let $y_1 = \sqrt{b^2X^2 + 4a^2}$ and $S = \frac{bX}{2a}$. We can compute the first order derivatives of $T(X, a, b)$ as

$$\frac{\partial T}{\partial a} = \frac{-2X}{ay_1} \text{ and } \frac{\partial T}{\partial b} = \frac{2X}{by_1} - \frac{2\ln\left(\sqrt{S^2+1}+S\right)}{b^2}.$$

The Hessian matrix of $T(X, a, b)$, where we treat X as a constant, becomes $\frac{2X}{y_1^3}\mathbf{H}$, where

$$\mathbf{H} = \begin{bmatrix} 8+4S^2 & 2SX \\ 2SX & X^2\left(\frac{2(S^2+1)^{\frac{3}{2}}}{S^3}\ln(\sqrt{S^2+1}+S) - \frac{2}{S^2} - 3\right) \end{bmatrix}.$$

To prove the convexity of $T(X, a, b)$, it is enough to show that the Hessian matrix of $T(X, a, b)$ is positive semidefinite, which is equivalent to showing that $\mathbf{H}$ is positive semidefinite. Using Sylvester's Criterion, because $8+4S^2$ is always positive, we see that $\mathbf{H}$ is positive semidefinite exactly when $\det \mathbf{H} \geq 0$. Letting

$$\phi(S) = 2(1+\frac{1}{S^2})^{\frac{3}{2}}\ln(\sqrt{S^2+1}+S) - \frac{2}{S^2} - 3,$$

we see that

$$\det \mathbf{H} = 8X^2\phi(S) + 4S^2X^2(\phi(S)-1) = 4X^2(S^2\phi(S) + 2\phi(S) - S^2),$$

so it is enough to show that $S^2\phi(S) + 2\phi(S) - S^2 \geq 0$ for all $S \geq S_0$. Since S_0 is the largest root of $S^2\phi(S) + 2\phi(S) - S^2 = 0$, it is enough to show this inequality for all sufficiently large S. Observe that $\phi(S) \geq \ln(S)$ for all sufficiently large S, implying that $S^2\phi(S) + 2\phi(S) - S^2 \geq S^2(\ln(S) - 1) + 2\ln(S)$, which is at least 0 for all sufficiently large S. Since $S = \frac{bX}{2a}$, this proves the lemma. □

We are now ready to prove the convexity of $(T(X, a, b) - T(X, a^*, b^*))^2$ over a certain region U.

Theorem 4. *Let a^* and b^* be positive constants, and let*

$$U := \{(a, b) \in \mathbb{R}^2 : a, b > 0, \frac{bX}{2a} \geq S_0, T(X, a, b) \geq T(X, a^*, b^*)\},$$

where $X > 0$ is a fixed constant. Then U is nonempty and $(T(X, a, b) - T(X, a^*, b^*))^2$ is jointly convex in $(a, b) \in U$.

Proof. First, as a decreases, both $\frac{bX}{2a}$ and $T(X, a, b)$ increase and approach infinity, implying that U is nonempty. From Lemma 3 we see that the function $T(X, a, b) - T(X, a^*, b^*)$ is jointly convex in $(a, b) \in U$. In addition, as $T(X, a, b) \geq T(X, a^*, b^*)$ on U we have that $T(X, a, b) - T(X, a^*, b^*)$ is also nonnegative on U. Thus, by Lemma 1, the squared W_2 distance, which is $(T(X, a, b) - T(X, a^*, b^*))^2$, is jointly convex in $(a, b) \in U$. □

Remark 2. To ensure that $(a^*, b^*) \in U$, it is enough to require that $\frac{b^*X}{2a^*} \geq S_0$ because $T(X, a^*, b^*) - T(X, a^*, b^*)$ is equal to 0. Thus, choosing large values of X will ensure that the squared W_2 distance is convex in (a, b) in a region containing (a^*, b^*). To find a suitable initial guess, we may choose a point (a, b) such that $\frac{bX}{2a} \geq S_0$ and scale it by a sufficiently small constant in order to satisfy the condition $T(X, a, b) \geq T(X, a^*, b^*)$.

However, this function is not suitable as an objective function because $(T(X, a, b) - T(X, a^*, b^*))^2$ may equal 0, its minimum, even when a is not equal to a^* or b is not equal to b^* – there is not enough information to find a^* and b^* through one receiver alone. To fix this, we add multiple receiver locations X_r.

Fact 1. *The equation $T(X, a_1, b_1) - T(X, a_2, b_2) = 0$ has at most one solution in $X > 0$ where $(a_1, b_1) \neq (a_2, b_2)$ are fixed ordered pairs of positive real numbers.*

Proof. Assume for the sake of contradiction that the equation $T(X, a_1, b_1) = T(X, a_2, b_2)$ has two positive solutions. Since $T(0, a_1, b_1) = T(0, a_2, b_2) = 0$, there are at least three nonnegative solutions in X to the equation $T(X, a_1, b_1) = T(X, a_2, b_2)$. Therefore, the equation $T'(X, a_1, b_1) = T'(X, a_2, b_2)$, or equivalently,

$$\frac{2}{\sqrt{b_1^2 X^2 + 4a_1^2}} = \frac{2}{\sqrt{b_2^2 X^2 + 4a_2^2}},$$

has at least two positive solutions in X by the Mean Value Theorem. Thus, there are two positive solutions to the equation $(b_1^2 - b_2^2)X^2 + (4a_1^2 - 4a_2^2) = 0$, a contradiction. □

This implies that the value of $T(X_r, a, b)$ at two different receiver locations X_r uniquely determines the pair (a, b). We use the objective function $\sum_r (T(X_r, a, b) - T(X_r, a^*, b^*))^2$, where the sum is over the receivers r, and the X_r are the receiver locations. We plot this objective function in Figure 5a, where $a^* = 1$ and $b^* = 2$ and the X_r range from 10 to 100 inclusive. We compare it to the sum of the squared L^2 norm where the set of receiver locations is the same and the source function is $f(t) = e^{-50(t-5)^2}$. The L^2–based objective function is mostly flat with a sharp incline close to the true velocity parameter $(1, 2)$. It is clearly nonconvex, as shown in Figure 5b. On the other hand, the W_2–based objective function (which does not depend on the source function, as long as it is nonnegative) appears convex in (a, b).

3.3.2 Varying Amplitude

In general, the amplitude of the wave equation solution is non–constant. The predicted wave function at receiver X can be considered to be of the form

$$g(t, a, b) = A(X, a, b) f(t - T(X, a, b)),$$

where $A(X, a, b)$ is the amplitude as a function of the receiver location X and the velocity parameters a, b. The observed wave function at receiver X can be considered of the form $g(t, a^*, b^*)$ where (a^*, b^*) is the true velocity parameter. Using Equation (2.7), we obtain

$$A(X, a, b) = \frac{b}{\sqrt{b^2 X^2 + 4a^2}\left(\frac{\pi}{2} - \arcsin\left(\frac{2a}{\sqrt{b^2 X^2 + 4a^2}}\right)\right)}.$$

Since b and X are positive, we may simplify this expression to get

$$A(X, a, b) = \frac{1}{X\sqrt{1 + \left(\frac{2a}{bX}\right)^2}\left(\frac{\pi}{2} - \arcsin\left(\frac{1}{\sqrt{\left(\frac{bX}{2a}\right)^2 + 1}}\right)\right)} = \frac{1}{RX\sqrt{1 + \left(\frac{1}{S}\right)^2}} \tag{3.2}$$

where

$$S = \frac{bX}{2a} \text{ and } R = \frac{\pi}{2} - \arcsin\left(\frac{1}{\sqrt{S^2 + 1}}\right).$$

While the amplitude function in Equation (3.2) is not fully accurate, it is still a good approximation if the velocity model is continuous.

We will only consider points (a, b) with $a > \frac{a^*}{100}$ and $\frac{b}{a} \geq \frac{b^*}{a^*}$, which is a positive constant independent of X. This also implies that $b > \frac{b^*}{100}$. We treat X as a sufficiently large constant, which causes $S = \frac{bX}{2a} \geq \frac{b^* X}{2a^*}$ to be large as well. In particular, X is taken to be large enough so that $\frac{b^* X}{2a^*} > S_0$. For large S, the amplitude is approximately $\frac{2}{\pi X}$. Taking the derivative of the amplitude with respect to S gives

$$\frac{\partial}{\partial S} A(X, S) = \frac{1}{X}\left(\frac{1}{\left(\frac{1}{S^2} + 1\right)^{\frac{3}{2}} S^3 R} - \frac{S}{\left(\frac{1}{S^2} + 1\right)^{\frac{3}{2}} S^3 R^2}\right) = \mathcal{O}\left(\frac{1}{S^2}\right).$$

Thus,

$$\left|\frac{\partial}{\partial a}A(X,a,b)\right| = \left|\frac{\partial}{\partial a}S\cdot\frac{\partial}{\partial S}A(X,S)\right| = \left|\frac{-S}{a}\cdot\frac{\partial}{\partial S}A(X,S)\right| = \mathcal{O}\left(\frac{1}{S}\right)$$

because $a > \frac{a^*}{100}$. Similarly, because $b > \frac{b^*}{100}$,

$$\left|\frac{\partial}{\partial b}A(X,a,b)\right| = \left|\frac{\partial}{\partial b}S\cdot\frac{\partial}{\partial S}A(X,S)\right| = \left|\frac{S}{b}\cdot\frac{\partial}{\partial S}A(X,S)\right| = \mathcal{O}\left(\frac{1}{S}\right).$$

Therefore, it is reasonable to assume the amplitude remains unchanged for large values of X, and we let A be this amplitude. We approximate a predicted wave function $g(t,a,b) = A(X,a,b)f(t-T(X,a,b))$ where X is a fixed large constant by the new function $g_1(t,a,b) = Af(t-T(X,a,b))$ so it suffices to approximate the observed wave function $h(t) = g(t,a^*,b^*)$ by the wave function $h_1(t) = g_1(t,a^*,b^*)$.

As the W_2 distance is only defined when both of its inputs have total mass 1, we normalize all wave data of the form $k(t)$ using the formula

$$\widetilde{k}(t) = \frac{k(t)+\gamma}{\int_0^{\mathcal{T}}(k(t)+\gamma)\mathrm{d}t}$$

on the interval $[0,\mathcal{T}]$ and 0 everywhere else. Here, γ is a positive constant such that $k(t)+\gamma > 0$ for all t.

Eventually, we compute the squared W_2 distance between $\widetilde{g}_1(t,a,b)$ and $\widetilde{h}_1(t)$. We have

$$\widetilde{g}_1(t,a,b) = \frac{Af(t-T_{pred})+\gamma}{AI_0+\gamma\mathcal{T}} \text{ and } \widetilde{h}_1(t) = \widetilde{g}_1(t,a^*,b^*) = \frac{Af(t-T_{obs})+\gamma}{AI_0+\gamma\mathcal{T}}, \tag{3.3}$$

where $\int_0^{\mathcal{T}} f(t)\,\mathrm{d}t = I_0$, $T_{pred} = T(X,a,b)$, and $T_{obs} = T(X,a^*,b^*)$. We also assume that γ is large enough to ensure that $\widetilde{g}_1(t,a,b)$ and $\widetilde{h}_1(t)$ are strictly positive with total mass 1.

Then, we let

$$G(t,a,b) = \int_0^t \widetilde{g}_1(y,a,b)\,\mathrm{d}y \text{ and } H(t) = \int_0^t \widetilde{h}_1(y)\,\mathrm{d}y = G(t,a^*,b^*).$$

Since $\widetilde{g}_1(t,a,b)$ and $\widetilde{h}_1(t)$ are both probability distributions, we may compute the squared W_2 distance between them using Equation (2.1) to get

$$W_2^2(\widetilde{g}_1,\widetilde{h}_1) = \int_0^1 (G^{-1}(s,a,b)-H^{-1}(s))^2\,\mathrm{d}s$$

where $G^{-1}(s,a,b):[0,1]\to[0,\mathcal{T}]$ is the unique function satisfying

$$G(G^{-1}(s,a,b),a,b) = s \text{ and } G^{-1}(G(t,a,b),a,b) = t$$

for all $s\in[0,1]$ and $t\in[0,\mathcal{T}]$.

Observe that $\widetilde{g}_1 = \widetilde{g}_1(t,a,b)$ and $G = G(t,a,b)$ can be expressed as functions of time t and the velocity parameters a,b, and $G^{-1}(s) = G^{-1}(s,a,b)$ can be expressed as a function of s and a,b. It is also possible to express $\widetilde{g}_1 = \widetilde{g}_1(t,T_{pred})$ as a function of t and the predicted traveltime T_{pred} using Equation (3.3). Thus, we can alternatively express $G = G(t,T_{pred})$ as a function of t and T_{pred}, and $G^{-1}(s) = G^{-1}(s,T_{pred})$ as a function of s and T_{pred}. In the proofs of the following claims, we sometimes omit the variables a,b, and T_{pred} in the arguments of $\widetilde{g}_1, G$, and G^{-1}.

We compute $\int_0^1 (G^{-1}(s))^2\,\mathrm{d}s$ to simplify $W_2^2(\widetilde{g}_1,\widetilde{h}_1)$.

Lemma 4. *Let $I_0 = \int_0^{\mathcal{T}} f(t)\,\mathrm{d}t$, $I_1 = \int_0^{\mathcal{T}} tf(t)\,\mathrm{d}t$, and $I_2 = \int_0^{\mathcal{T}} t^2 f(t)\,\mathrm{d}t$. Then, if $T_{obs}\le T_{pred}\le \mathcal{T}-p_2$,*

$$\int_0^1 (G^{-1}(s,T_{pred}))^2\,ds = \frac{1}{I_0+\frac{\gamma\mathcal{T}}{A}}\left(\frac{\gamma\mathcal{T}^3}{3A}+I_2+2T_{pred}I_1+T_{pred}^2I_0\right).$$

Proof. From the substitution $s = G(t)$, we can compute $\int_0^1 (G^{-1}(s))^2 \, \mathrm{d}s$ as

$$\int_0^1 (G^{-1}(s))^2 \, \mathrm{d}s = \int_0^{\mathcal{T}} t^2 \widetilde{g}_1(t) \, \mathrm{d}t = \int_0^{\mathcal{T}} t^2 \frac{f(t - T_{pred}) + \frac{\gamma}{A}}{I_0 + \frac{\gamma \mathcal{T}}{A}} \, \mathrm{d}t.$$

We may write the integral as

$$\int_0^{\mathcal{T}} t^2 \frac{f(t - T_{pred}) + \frac{\gamma}{A}}{I_0 + \frac{\gamma \mathcal{T}}{A}} \, \mathrm{d}t = \frac{1}{I_0 + \frac{\gamma \mathcal{T}}{A}} \left(\frac{\gamma \mathcal{T}^3}{3A} + \int_0^{\mathcal{T}} t^2 f(t - T_{pred}) \, \mathrm{d}t \right).$$

However, note that $\int_0^{\mathcal{T}} t^2 f(t - T_{pred}) \, \mathrm{d}t = \int_{-T_{pred}}^{\mathcal{T} - T_{pred}} (t + T_{pred})^2 f(t) \, \mathrm{d}t$. As f has compact support $[p_1, p_2] \subset [0, \mathcal{T}]$ and $T_{pred} \leq \mathcal{T} - p_2$, we may write the integral as $\int_0^{\mathcal{T}} (t + T_{pred})^2 f(t) \, \mathrm{d}t$. This becomes $I_2 + 2T_{pred} I_1 + T_{pred}^2 I_0$, where $I_2 = \int_0^{\mathcal{T}} t^2 f(t) \, \mathrm{d}t, I_1 = \int_0^{\mathcal{T}} t f(t) \, \mathrm{d}t$, and $I_0 = \int_0^{\mathcal{T}} f(t) \, \mathrm{d}t$, which proves the lemma. □

Now, observe that the squared W_2 distance may be expressed as a function of $T_{pred} = T(X, a, b)$, because $G^{-1}(s, T_{pred})$ is a function of both s and T_{pred}. Using this, we may compute the first and second derivatives of $W_2^2(\widetilde{g}_1, \widetilde{h}_1)$ with respect to T_{pred}. To compute the first derivative of $W_2^2(\widetilde{g}_1, \widetilde{h}_1)$ with respect to T_{pred}, it is enough to find the first derivative of G^{-1} with respect to T_{pred}.

Lemma 5. *The first derivative of $G^{-1}(s, T_{pred})$ with respect to T_{pred} is*

$$\frac{\partial}{\partial T_{pred}} G^{-1}(s, T_{pred}) = \frac{f(G^{-1}(s, T_{pred}) - T_{pred})}{f(G^{-1}(s, T_{pred}) - T_{pred}) + \frac{\gamma}{A}} = 1 - \frac{\frac{\gamma}{A}}{f(G^{-1}(s, T_{pred}) - T_{pred}) + \frac{\gamma}{A}}. \tag{3.4}$$

Proof. Because $\int_0^{G^{-1}(s)} \widetilde{g}_1(t) \, \mathrm{d}t = G(G^{-1}(s)) = s$, we may apply the Leibniz integral rule to see that

$$\begin{aligned} 0 = \frac{\partial}{\partial T_{pred}} s &= \frac{\partial}{\partial T_{pred}} \int_0^{G^{-1}(s)} \widetilde{g}_1(t) \, \mathrm{d}t \\ &= \widetilde{g}_1(G^{-1}(s)) \frac{\partial}{\partial T_{pred}} G^{-1}(s) + \int_0^{G^{-1}(s)} \frac{\partial}{\partial T_{pred}} \widetilde{g}_1(t) \, \mathrm{d}t. \end{aligned}$$

Thus, we have that

$$\frac{\partial}{\partial T_{pred}} G^{-1}(s) = \frac{-\int_0^{G^{-1}(s)} \frac{\partial}{\partial T_{pred}} \widetilde{g}_1(t) \, \mathrm{d}t}{\widetilde{g}_1(G^{-1}(s))} = \frac{-\left(I_0 + \frac{\gamma \mathcal{T}}{A}\right) \int_0^{G^{-1}(s)} \frac{\partial}{\partial T_{pred}} \widetilde{g}_1(t) \, \mathrm{d}t}{f(G^{-1}(s) - T_{pred}) + \frac{\gamma}{A}}.$$

From Equation (3.3), we see that

$$-\int_0^{G^{-1}(s)} \frac{\partial}{\partial T_{pred}} \widetilde{g}_1(t) \, \mathrm{d}t = \frac{\int_0^{G^{-1}(s)} f'(t - T_{pred}) \, \mathrm{d}t}{I_0 + \frac{\gamma \mathcal{T}}{A}} = \frac{f(G^{-1}(s) - T_{pred})}{I_0 + \frac{\gamma \mathcal{T}}{A}},$$

where we use the fact that $f(-T_{pred}) = 0$. Thus, Equation (3.4) holds. □

Now, we use this to compute the first derivative of the squared W_2 distance with respect to T_{pred}.

Lemma 6. *Suppose $T_{obs} \leq T_{pred} \leq \mathcal{T} - p_2$. The first derivative of $W_2^2(\widetilde{g}_1(t, T_{pred}), \widetilde{h}_1(t))$ with respect to T_{pred} is*

$$\frac{2}{I_0 + \frac{\gamma \mathcal{T}}{A}} \left(\left(I_0 - \frac{\gamma \mathcal{T}}{A} \right) (T_{pred} - T_{obs}) + \frac{\gamma}{A} \int_0^{\mathcal{T}} \int_{T_{obs}}^{T_{pred}} \frac{\frac{\gamma}{A}}{f(G^{-1}(H(t), y) - y) + \frac{\gamma}{A}} \, dy \, dt \right).$$

Proof. Because $\frac{\partial}{\partial T_{pred}}(H^{-1}(s)) = 0$,

$$\frac{\partial}{\partial T_{pred}} W_2^2(\widetilde{g}_1, \widetilde{h}_1) = \frac{\partial}{\partial T_{pred}} \int_0^1 (G^{-1}(s))^2 \, \mathrm{d}s - 2\int_0^1 H^{-1}(s) \frac{\partial}{\partial T_{pred}}(G^{-1}(s)) \, \mathrm{d}s,$$

using the Leibniz integral rule. By Lemma 4 and Lemma 5 we simplify

$$\frac{\partial}{\partial T_{pred}} W_2^2(\widetilde{g}_1, \widetilde{h}_1) = \frac{2T_{pred}I_0 + 2I_1}{I_0 + \frac{\gamma \mathcal{T}}{A}} - 2\int_0^1 H^{-1}(s) - \frac{\frac{\gamma}{A} H^{-1}(s)}{f(G^{-1}(s) - T_{pred}) + \frac{\gamma}{A}} \, \mathrm{d}s.$$

From the change of variables $s = H(t)$, we see that $\int_0^1 H^{-1}(s) \, \mathrm{d}s = \int_0^{\mathcal{T}} t\widetilde{h}_1(t) \, \mathrm{d}t$, which can be computed as

$$\begin{aligned}\int_0^{\mathcal{T}} \frac{tf(t - T_{obs}) + \frac{\gamma t}{A}}{I_0 + \frac{\gamma \mathcal{T}}{A}} \, \mathrm{d}t &= \frac{1}{I_0 + \frac{\gamma \mathcal{T}}{A}} \left(\frac{\gamma \mathcal{T}^2}{2A} + \int_{-T_{obs}}^{\mathcal{T} - T_{obs}} (t + T_{obs}) f(t) \, \mathrm{d}t \right) \\ &= \frac{\frac{\gamma \mathcal{T}^2}{2A} + I_1 + T_{obs} I_0}{I_0 + \frac{\gamma \mathcal{T}}{A}}.\end{aligned}$$

This means

$$\frac{\partial}{\partial T_{pred}} W_2^2(\widetilde{g}_1, \widetilde{h}_1) = \frac{2}{I_0 + \frac{\gamma \mathcal{T}}{A}} \left(I_0(T_{pred} - T_{obs}) + \frac{\gamma}{A} \int_0^1 \frac{H^{-1}(s)}{\widetilde{g}_1(G^{-1}(s))} \, \mathrm{d}s - \frac{\gamma \mathcal{T}^2}{2A} \right).$$

Furthermore, observe that $\int_0^1 \frac{H^{-1}(s)}{\widetilde{g}_1(G^{-1}(s))} \, \mathrm{d}s = \int_0^{\mathcal{T}} H^{-1}(G(t)) \, \mathrm{d}t$ from the substitution $s = G(t)$. Because $H^{-1}(G(t))$ is the inverse function of $G^{-1}(H(t))$, we have that $\int_0^{\mathcal{T}} H^{-1}(G(t)) \, \mathrm{d}t = \mathcal{T}^2 - \int_0^{\mathcal{T}} G^{-1}(H(t)) \, \mathrm{d}t$. Thus,

$$\frac{\partial}{\partial T_{pred}} W_2^2(\widetilde{g}_1, \widetilde{h}_1) = \frac{2}{I_0 + \frac{\gamma \mathcal{T}}{A}} \left(I_0(T_{pred} - T_{obs}) + \frac{\gamma \mathcal{T}^2}{2A} - \frac{\gamma}{A} \int_0^{\mathcal{T}} G^{-1}(H(t)) \, \mathrm{d}t \right). \tag{3.5}$$

Observe that $G^{-1}(H(t), T_{obs}) = t$. From Equation (3.4), we have that

$$\begin{aligned}G^{-1}(H(t), T_{pred}) &= t + \int_{T_{obs}}^{T_{pred}} \frac{\partial}{\partial y} G^{-1}(H(t), y) \, \mathrm{d}y \\ &= t + \int_{T_{obs}}^{T_{pred}} \left(1 - \frac{\frac{\gamma}{A}}{f(G^{-1}(H(t), y) - y) + \frac{\gamma}{A}} \right) \mathrm{d}y.\end{aligned}$$

Substituting this expression for $G^{-1}(H(t))$ into Equation (3.5) gives

$$\frac{2}{I_0 + \frac{\gamma \mathcal{T}}{A}} \left(\left(I_0 - \frac{\gamma \mathcal{T}}{A} \right) (T_{pred} - T_{obs}) + \frac{\gamma}{A} \int_0^{\mathcal{T}} \int_{T_{obs}}^{T_{pred}} \frac{\frac{\gamma}{A}}{f(G^{-1}(H(t), y) - y) + \frac{\gamma}{A}} \, \mathrm{d}y \, \mathrm{d}t \right)$$

as the value of $\frac{\partial}{\partial T_{pred}} W_2^2(\widetilde{g}_1, \widetilde{h}_1)$. □

Expressing $\frac{\partial}{\partial T_{pred}} W_2^2(\widetilde{g}_1, \widetilde{h}_1)$ in this form allows us to prove the convexity of $W_2^2(\widetilde{g}_1, \widetilde{h}_1)$ with respect to a, b subject to a restriction on the source function f. Here, we assume that f reaches both positive and negative values.

Theorem 5. *Let $q = \frac{\frac{\gamma}{A}}{\sup f + \frac{\gamma}{A}}$ and suppose that $I_0 \geq (1 - q)\frac{\gamma \mathcal{T}}{A}$. Let a^* and b^* be positive constants such that $\frac{b^* X}{2a^*} \geq S_0$ and let*

$$V = \{(a, b) \in \mathbb{R}^2 : a > \frac{a^*}{100}, \frac{bX}{2a} \geq \frac{b^* X}{2a^*}, T(X, a^*, b^*) \leq T(X, a, b) \leq \mathcal{T} - p_2\}.$$

Then, V is nonempty and $W_2^2(\widetilde{g}_1(t, a, b), \widetilde{h}_1(t))$ is jointly convex in $(a, b) \in V$.

Before beginning the proof of Theorem 5, we remark that the function $f(t) = (t-15)^2 - A$ with domain $[0, \mathcal{T}]$, where $\mathcal{T} = 30, \gamma = 2A$, and $0 < A \leq 73$, satisfies the condition $I_0 \geq (1-q)\frac{\gamma\mathcal{T}}{A}$ and reaches both positive and negative values. To see this, observe that $I_0 = 2250 - 30A \geq 60$ while $(1-q)\frac{\gamma\mathcal{T}}{A} = \frac{60 \sup f}{\sup f + 2} < 60$. Hence, our assumption that $I_0 \geq (1-q)\frac{\gamma\mathcal{T}}{A}$ is a reasonable assumption to make.

Proof. First, the point (a^*, b^*) is clearly in V, so V is nonempty. Now, we prove that $W_2^2(\widetilde{g}_1(t, a, b), \widetilde{h}_1(t))$ is jointly convex in $(a, b) \in V$. Because $\frac{b^* X}{2a^*} \geq S_0$, V is a subset of U, which is defined in Theorem 4. This implies that $T(X, a, b)$ is jointly convex in a, b over the region V by Lemma 3. Thus, it is enough to show that $W_2^2(\widetilde{g}_1, \widetilde{h}_1)$ is convex and nondecreasing in T_{pred} in the interval $[T_{obs}, \mathcal{T} - p_2]$. We first claim that $W_2^2(\widetilde{g}_1, \widetilde{h}_1)$ is nondecreasing in T_{pred} in this interval. By Lemma 6, we see that

$$\begin{aligned}
\frac{\partial}{\partial T_{pred}} W_2^2(\widetilde{g}_1, \widetilde{h}_1) &\geq \frac{2}{I_0 + \frac{\gamma\mathcal{T}}{A}} \left(\left(I_0 - \frac{\gamma\mathcal{T}}{A} \right) (T_{pred} - T_{obs}) + \frac{\gamma}{A} \int_0^{\mathcal{T}} \int_{T_{obs}}^{T_{pred}} q \, \mathrm{d}y \, \mathrm{d}t \right) \\
&= \frac{2}{I_0 + \frac{\gamma\mathcal{T}}{A}} \left(I_0 + (q-1)\frac{\gamma\mathcal{T}}{A} \right) (T_{pred} - T_{obs})
\end{aligned}$$

because $T_{pred} \geq T_{obs}$. Furthermore, because $I_0 \geq (1-q)\frac{\gamma\mathcal{T}}{A}$, every factor is nonnegative, implying that $\frac{\partial}{\partial T_{pred}} W_2^2(\widetilde{g}_1, \widetilde{h}_1)$ is nonnegative. Next, we claim that $W_2^2(\widetilde{g}_1, \widetilde{h}_1)$ is convex in T_{pred} in the interval $[T_{obs}, \mathcal{T} - p_2]$. Using Equation (3.5) and Lemma 5, we see that

$$\begin{aligned}
\frac{\partial^2}{\partial T_{pred}^2} W_2^2(\widetilde{g}_1, \widetilde{h}_1) &= \frac{\partial}{\partial T_{pred}} \left(\frac{2}{I_0 + \frac{\gamma\mathcal{T}}{A}} \left(I_0 (T_{pred} - T_{obs}) + \frac{\gamma\mathcal{T}^2}{2A} - \frac{\gamma}{A} \int_0^{\mathcal{T}} G^{-1}(H(t)) \, \mathrm{d}t \right) \right) \\
&= \frac{2}{I_0 + \frac{\gamma\mathcal{T}}{A}} \left(I_0 - \frac{\gamma}{A} \int_0^{\mathcal{T}} \frac{\partial}{\partial T_{pred}} G^{-1}(H(t)) \, \mathrm{d}t \right) \\
&= \frac{2}{I_0 + \frac{\gamma\mathcal{T}}{A}} \left(I_0 - \frac{\gamma}{A} \int_0^{\mathcal{T}} 1 - \frac{\frac{\gamma}{A}}{f(G^{-1}(H(t)) - T_{pred}) + \frac{\gamma}{A}} \, \mathrm{d}t \right) \\
&\geq \frac{2}{I_0 + \frac{\gamma\mathcal{T}}{A}} \left(I_0 + (q-1)\frac{\gamma\mathcal{T}}{A} \right).
\end{aligned}$$

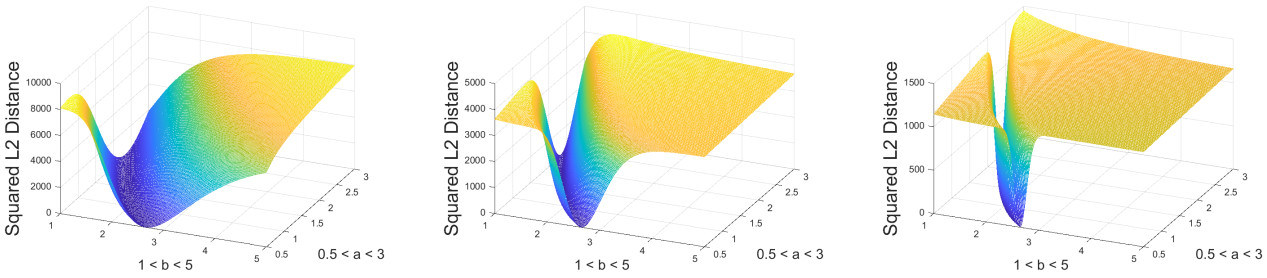
Since $I_0 \geq (1-q)\frac{\gamma\mathcal{T}}{A}$, we see that $\frac{\partial^2}{\partial T_{pred}^2} W_2^2(\widetilde{g}_1, \widetilde{h}_1)$ is nonnegative in the interval $[T_{obs}, \mathcal{T} - p_2]$. Thus, $W_2^2(\widetilde{g}_1, \widetilde{h}_1)$ is convex and nondecreasing in T_{pred} in the interval $[T_{obs}, \mathcal{T} - p_2]$. Since T_{pred} is jointly convex in $(a, b) \in V$, $W_2^2(\widetilde{g}_1, \widetilde{h}_1)$ is also jointly convex in $(a, b) \in V$ by Lemma 1. □

Remark 3. Through the Mean Value Theorem, Theorem 5 can be reformulated as a result with a condition involving an upper bound on $\frac{\mathrm{d}f}{\mathrm{d}t}$. This reformulation illustrates how the convexity of $W_2^2(\widetilde{g}_1, \widetilde{h}_1)$ is influenced by the frequency of the source function.

Remark 4. The requirements needed to apply Lemma 1 are not satisfied for $T_{pred} \in (0, T_{obs}]$. To see this, observe that the squared W_2 distance is nonnegative everywhere and 0 when $T_{pred} = T_{obs}$. Thus, it is impossible for the squared W_2 distance to be nondecreasing in $T_{pred} \in (0, T_{obs}]$, although it may be convex in this interval.

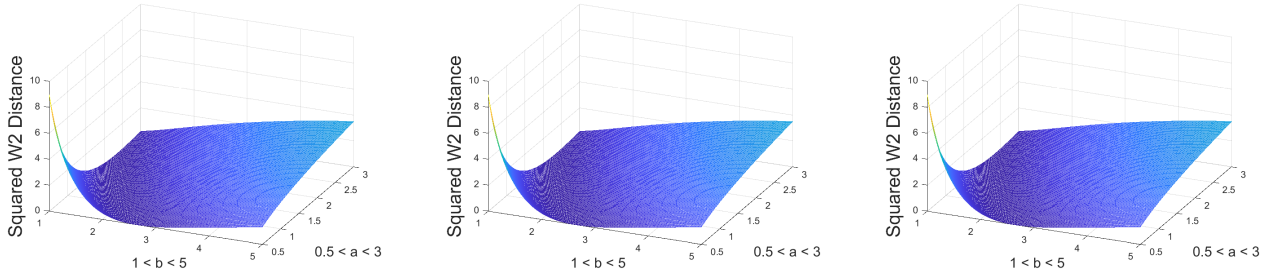
4 Numerical Results

We continue with the velocity model studied in Section 3.3. We compute both the L^2 and W_2 distance and compare the convexity of the two objective functions in (a, b), as shown in Figures 6 and 7. The source wave function is of the form $f(t) = e^{-\alpha(t-5)^2}$ and we consider $\alpha \in \{2, 10, 100\}$. Here, $(a^*, b^*) = (1, 2)$ and



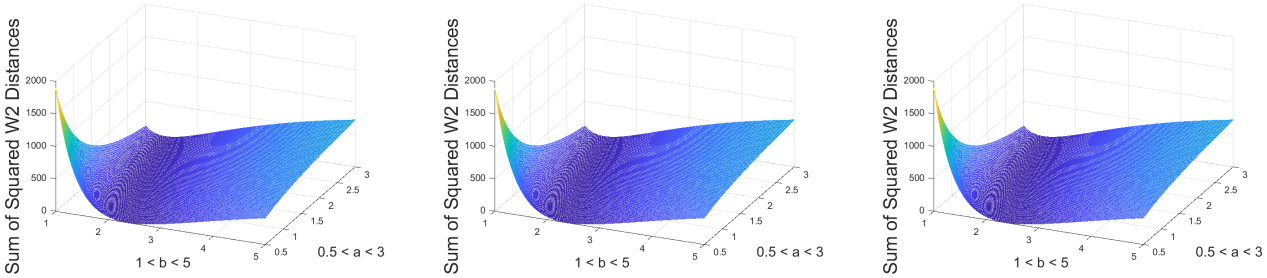
(a) $\alpha = 2$ (b) $\alpha = 10$ (c) $\alpha = 100$

Figure 6: Plots of $||g - h||_2^2$, where there is only one receiver at $X = 10$.



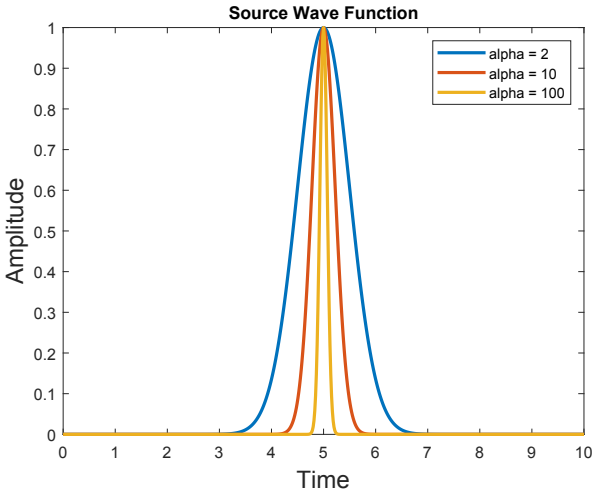
(a) $\alpha = 2$ (b) $\alpha = 10$ (c) $\alpha = 100$

Figure 7: Plots of $W_2^2(\widetilde{g}(X,t), \widetilde{h}(X,t))$, where there is only one receiver at $X = 10$.

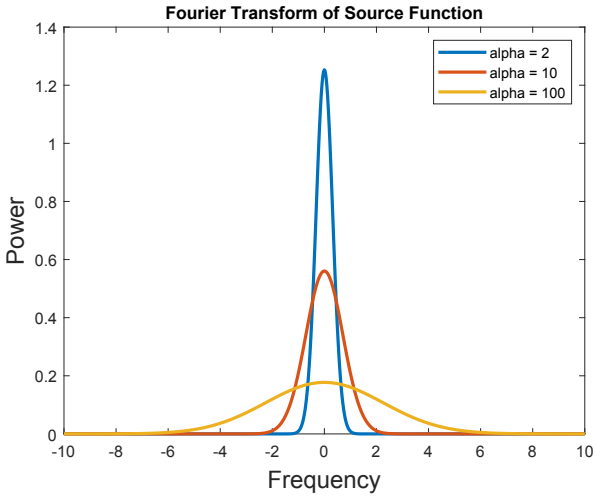


(a) $\alpha = 2$ (b) $\alpha = 10$ (c) $\alpha = 100$

Figure 8: Plots of $\sum_r W_2^2(\widetilde{g}(X_r,t), \widetilde{h}(X_r,t))$, where the receiver locations X_r range from 10 to 100, inclusive.



(a) The source functions are of the form $e^{-\alpha(t-5)^2}$.



(b) The Fourier transforms of the source functions.

Figure 9: The Fourier transform of the source function depends on the value of α. As α increases, the average absolute value of the frequency also increases.

$\gamma = 10^{-6}$. We can consider the observed data to be of the form $A(X,a,b)f(t-T(X,a,b))$ because the velocity is a continuous function of the depth. To ensure that the wave data is compactly supported, we use the time range $[0,50]$ to compute the squared W_2 distance.

While the squared W_2 distance appears to be mostly convex for $\alpha \in \{2,10,100\}$, the squared L^2 norm is certainly nonconvex. The plot of the squared L^2 norm has large flat regions with a steep incline closer to where the L^2 norm is minimized, as shown in Figures 6a to 6c. Although the squared L^2 norm is nonconvex, by decreasing the value of α we increase the size of the convex region around (a^*,b^*) of the squared L^2 norm. As the graph of the source function becomes sharper (Figure 9a), so does the graph of the squared L^2 norm. The squared W_2 distance, on the other hand, is relatively flat throughout the entire domain and does not have a steep incline closer to the minimum, as shown in Figures 7a to 7c. Thus, the squared W_2 distance should be convex on a much larger region containing the minimum. The squared W_2 distance is also highly insensitive to the choice of source function, and this suggests that the squared W_2 distance can be used to solve various seismic inversion problems, in contrast with the squared L^2 norm.

In addition, we plot the sum of the squared W_2 distance taken over multiple receiver locations in Figure 8. The summation of the squared W_2 distance over multiple receiver locations is highly convex regardless of α, as seen in Figures 8a to 8c. Furthermore, the summation of the squared W_2 distance over multiple receiver locations is also very close to the summation of $(T_{pred}-T_{obs})^2$ over multiple receiver locations, and appears to be convex in (a,b) for (a,b) closer to the origin. Thus, as an initial guess for (a,b), it appears to be better to choose points (a,b) which are very close to the origin. This is equivalent to choosing points (a,b) such that the predicted travel time is large.

4.1 Frequency Analysis

We observe that the changes in the squared L^2 or W_2 distances between $g(t,a,b)$ and $h(t)$ depends solely on the value of α, which in turn affects the frequency of the source function. The graph of the source function in the frequency domain can be derived by taking a Fourier transform. Letting

$$\widehat{f}(k) = \int_{-\infty}^{\infty} e^{-2\pi ikt} f(t)\,\mathrm{d}t = \int_{-\infty}^{\infty} e^{-2\pi ikt} e^{-\alpha(t-5)^2}\,\mathrm{d}t,$$

we get that $\widehat{f}(k) = \sqrt{\frac{\pi}{\alpha}} e^{-\frac{(\pi k)^2}{\alpha} - 10\pi i k}$, and the power at a frequency k is given by the magnitude, which is $|\widehat{f}(k)| = \sqrt{\frac{\pi}{\alpha}} e^{-\frac{(\pi k)^2}{\alpha}}$. The graph of $|\widehat{f}(k)|$ is centered at $k = 0$ regardless of the value of α, but as the value of α increases, the plot of $|\widehat{f}(k)|$ becomes wider, as shown in Figure 9b. In other words, if $|k_1| > |k_2|$ where k_1 and k_2 are fixed frequencies, then $\frac{|\widehat{f}(k_1)|}{|\widehat{f}(k_2)|}$ increases as α increases. Thus, as the value of α increases, the average absolute value of the frequency of the source function also increases. For large values of α, the plot of the squared L^2 norm also has the properties of higher frequency data, as the plot is very sharp close to $(a^*, b^*) = (1, 2)$. The high–frequency data seen in Figure 6c is explained by the squared L^2 norm weighting low–frequency and high–frequency terms equally, by the Plancherel Theorem. On the other hand, the plot of the squared W_2 distance is virtually unchanged as α increases, suggesting that the frequency of the the squared W_2 distance is highly insensitive to the frequency of the source function.

The relationship between the W_2 distance between two functions g and h and a weighted $\dot{\mathcal{H}}^{-1}$ distance between them helps provide an explanation for these observations regarding frequency [10, 20]. We define the space $\dot{\mathcal{H}}^1(\mathbb{R}^d)$ through the seminorm

$$||f||^2_{\dot{\mathcal{H}}^1(\mathbb{R}^d)} = \int_{\mathbb{R}^d} |k|^2 |\widehat{f}(k)|^2 \, \mathrm{d}k$$

and the space $\dot{\mathcal{H}}^{-1}(\mathbb{R}^d)$ is defined as the dual of $\dot{\mathcal{H}}^1(\mathbb{R}^d)$ through the norm

$$||f||_{\dot{\mathcal{H}}^{-1}(\mathbb{R}^d)} = \sup\{|\langle F, f\rangle_{L^2}| : ||F||_{\dot{\mathcal{H}}^1} \leq 1\}.$$

It is known [20] that the W_2 distance is asymptotically equivalent to the $\dot{\mathcal{H}}^{-1}$ norm, which weights terms of lower frequency over terms with higher frequency. Specifically, if μ is a probability measure and $\mathrm{d}\pi$ is an infinitesimal perturbation with zero total mass, then $W_2(\mu, \mu + \mathrm{d}\pi) = ||\mathrm{d}\pi||_{\dot{\mathcal{H}}^{-1}_{(\mathrm{d}\mu)}} + o(\mathrm{d}\pi)$ [10]. While the objective functions in Figures 7 and 8 are not globally convex, the relationship between the squared W_2 distance and the squared $\dot{\mathcal{H}}^{-1}$ metric offers an explanation for the smoothness of the plots in these figures, which display properties of low–frequency data.

4.2 Optimal Transport for Non–probability Measures

In general, the wave data tends to alternate between positive and negative values, and the total integral of the observed or predicted wave function does not have to be 1. Thus, we cannot immediately use the squared W_2 distance as our objective function, because it is only defined on probability distributions. The current approach to normalizing the wave data requires two steps: first, transform the wave data to a nonnegative function, and second, divide by the total mass [8]. This ensures that the normalized wave data satisfies the positivity and total mass requirements. Although there are several possible ways to complete the first step, the known methods of doing this have their own drawbacks.

If the source function f is positive, the first step becomes unnecessary. The requirement in Theorem 4 is not very strong, suggesting that when f is positive, the squared W_2 distance is suitable as an objective function. However, this method does not generalize well to source functions that alternate between negative and positive values. In this case, we complete the first step by initially replacing an alternating function k by $k + \gamma$, where $\inf k + \gamma > 0$. Then, we divide by the total mass of $k + \gamma$ in the interval $[0, \mathcal{T}]$. This method of normalization takes into account the wave amplitude as well. However, the convexity of the squared W_2 distance, in this case, is not as general as with the previous method. Further restrictions on the source f are necessary, as shown by the requirements in Theorem 5. This suggests that the squared W_2 distance is suitable as an objective function when the normalization constant γ is sufficiently close to 0, or equivalently, when $\inf f$ is sufficiently close to 0.

5 Conclusions

In this paper, we study the convexity of full–waveform inversion using the squared W_2 distance as an objective function with respect to the velocity model parameter. We show that the squared W_2 distance is a suitable objective function for multiple velocity models when the received signal is nonnegative. Next, we show that the squared W_2 distance is suitable in some cases, in a two–dimensional velocity model where the received signal alternates between positive and negative values. We review the smoothing property of the squared W_2 distance by its relation to the squared $\dot{\mathcal{H}}^{-1}$ distance, and contrast this with the sharpness of the squared L^2 norm, which is very sensitive to high–frequency signals. We also discuss the drawbacks of the normalization methods used in this paper. A natural direction for future research is to generalize the W_2 distance to compare functions alternating between positive and negative values.

6 Acknowledgements

Firstly, the author would like to thank Dr. Yunan Yang for her mentorship and guidance during this project. The author thanks Dr. Tanya Khovanova and Boya Song for proofreading this paper and for providing feedback. Finally, the author is thankful to the PRIMES–USA program for making this research project possible. This work is supported in part by the National Science Foundation through grant DMS–1913129.

References

[1] Luigi Ambrosio and Nicola Gigli. A user's guide to optimal transport. In *Modelling and optimisation of flows on networks*, pages 1–155. Springer, 2013.

[2] Hyoungsu Baek, Henri Calandra, and Laurent Demanet. Velocity estimation via registration-guided least-squares inversion. *Geophysics*, 79(2):R79–R89, 2014.

[3] Ebru Bozdağ, Jeannot Trampert, and Jeroen Tromp. Misfit functions for full waveform inversion based on instantaneous phase and envelope measurements. *Geophysical Journal International*, 185(2):845–870, 2011.

[4] Yann Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on pure and applied mathematics*, 44(4):375–417, 1991.

[5] Jing Chen, Yifan Chen, Hao Wu, and Dinghui Yang. The quadratic Wasserstein metric for earthquake location. *Journal of Computational Physics*, 373:188–209, 2018.

[6] Guido De Philippis and Alessio Figalli. The Monge–Ampère equation and its link to optimal transportation. *Bulletin of the American Mathematical Society*, 51(4):527–580, 2014.

[7] Laurent Demanet. Waves and imaging class notes-18.325, 2016.

[8] Björn Engquist and Brittany D. Froese. Application of the Wasserstein metric to seismic signals. *arXiv preprint arXiv:1311.4581*, 2013.

[9] Björn Engquist, Brittany D. Froese, and Yunan Yang. Optimal transport for seismic full waveform inversion. *arXiv preprint arXiv:1602.01540*, 2016.

[10] Björn Engquist, Kui Ren, and Yunan Yang. The quadratic Wasserstein metric for inverse data matching. *Inverse Problems*, 36(5):055001, 2020.

[11] Björn Engquist and Yunan Yang. Seismic imaging and optimal transport. *arXiv preprint arXiv:1808.04801*, 2018.

[12] Ayman Hourieh. The composition of two convex functions is convex. Mathematics Stack Exchange. URL:https://math.stackexchange.com/q/287725 (version: 2013-01-26).

[13] Martin Knott and Cyril S. Smith. On the optimal mapping of distributions. *Journal of Optimization Theory and Applications*, 43(1):39–49, 1984.

[14] Gaspard Monge. Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences de Paris*, 1781.

[15] R. Gerhard Pratt. Inverse theory applied to multi-source cross-hole tomography.: Part 2: Elastic wave-equation method1. *Geophysical Prospecting*, 38(3):311–329, 1990.

[16] R. Gerhard Pratt and Michael H. Worthington. Inverse theory applied to multi-source cross-hole tomography. part 1: Acoustic wave-equation method 1. *Geophysical prospecting*, 38(3):287–310, 1990.

[17] Svetlozar T. Rachev and Ludger Rüschendorf. *Mass Transportation Problems: Volume I: Theory*, volume 1. Springer Science & Business Media, 1998.

[18] Peter M. Shearer. *Introduction to seismology*. Cambridge university press, 2019.

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

[20] Cédric Villani. *Topics in optimal transportation*. American Mathematical Soc., 2003.

[21] Jean Virieux and Stéphane Operto. An overview of full-waveform inversion in exploration geophysics. *Geophysics*, 74(6):WCC1–WCC26, 2009.

[22] Yunan Yang. *Optimal transport for seismic inverse problems*. PhD thesis, The University of Texas at Austin, 2018.

[23] Yunan Yang. Analysis and application of optimal transport for challenging seismic inverse problems. *arXiv preprint arXiv:1902.01226*, 2019.

[24] Yunan Yang and Björn Engquist. Analysis of optimal transport and related misfit functions in full-waveform inversion. *Geophysics*, 83(1):A7–A12, 2018.

[25] Colin A. Zelt. Traveltime tomography using controlled-source seismic data. *Encyclopedia of solid earth geophysics*, 2:1453–1473, 2011.