Fall 11-15-2022

Convexity of Regularized Optimal Transport Dissimilarity Measures for Signed Signals

Christian P. Fowler

Christian Fowler

Candidate

Mathematics

Department

This thesis is approved, and it is acceptable in quality and form for publication:

Approved by the Thesis Committee:

Dr. Mohammad Motamed , Chairperson

Dr. Stephen Lau

Dr. Gabriel Huerta

Convexity of Regularized Optimal Transport Dissimilarity Measures for Signed Signals

by

Christian Fowler

B.S., Mathematics, University of New Mexico, 2018

THESIS

Submitted in Partial Fulfillment of the Requirements for the Degree of

Master of Science

Mathematics

The University of New Mexico

Albuquerque, New Mexico

December 2022

For Cheryl

Acknowledgements

Thank you to Professor Mohammad Motamed and Professor Stephen Lau for your support and aid.

"If the calculus comes to vibrant life in celestial mechanics, as it surely does, then this is evidence that the stars in the sheltering sky have a secret mathematical identity, an aspect of themselves that like some tremulous night flower they reveal only when the mathematician whispers."

-David Berlinski

"A good writer possesses not only his own spirit but also the spirit of his friends."

-Friedrich Nietzsche

"Then, there is no madness, no raving lunacy, which such agitations do not bring forth. They fashion vain apparitions as in the dreams of sick men. When the soul is without a definite aim she gets lost; for, as they say, if you are everywhere you are nowhere"

-Michel De Montaigne

Convexity of Regularized Optimal Transport Dissimilarity Measures for Signed Signals

by

Christian Fowler

B.S., Mathematics, University of New Mexico, 2018

M.S., Mathematics, University of New Mexico, 2022

Abstract

Debiased Sinkhorn divergence (DS divergence) is a distance function of regularized optimal transport that measures the dissimilarity between two probability measures of optimal transport. This thesis analyzes the advantages of using DS divergence when compared to the more computationally expensive Wasserstein distance as well as the classical Euclidean norm. Specifically, theory and numerical experiments are used to show that Debiased Sinkhorn divergence has geometrically desirable properties such as maintained convexity after data normalization. Data normalization is often needed to calculate Sinkhorn divergence as well as Wasserstein distance, as these formulas only accept probability distributions as inputs and do not directly apply to signed data such as time signals and seismic waves; however, in doing so one may lose or distort information about the original signal. The investigations in this paper show that for high frequency signal inputs, Wasserstein distance may need a much more dramatic normalization compared to Debiased Sinkhorn in order to preserve convexity, leading to a loss of information about the original signal, the amplification of noise, and possibly machine overload, thus posing the desirability of the Debiased Sinkhorn divergence method.

Contents

Section 1

Introduction

Measuring the dissimilarity between two signals is an important and continually occurring problem in many fields of study [8, 1, 2, 3, 4, 5, 6, 7]. Optimal transport (OT) is an efficient way of handling this problem, as long as the inputted signals are probability distributions, i.e., their components sum to one and are all positive.
Debiased Sinkhorn divergence is a distance function that uses Sinkhorn divergence to measure the dissimilarity between two probability measures. Sinkhorn divergence is obtained by adding an entropic regularization term to the Kantorovich formulation of the optimal transport problem [1]. A main advantage of Sinkhorn divergence over Wasserstein distance, a commonly used metric, lies in its computability by an iterative algorithm known as Sinkhorn's matrix scaling algorithm, where each iteration involves two matrix-vector products. Sinkhorn is significantly less complex than Wasserstein distance yet still maintains the same desirable geometric properties [1, 2, 3, 8].

This paper aims to show convincing evidence that Debiased Sinkhorn divergence maintains convexity even when very high frequency signals are inputted and after data normalization is performed. Sinkhorn divergence and Wasserstein distance both require their inputs to be probability measures and to be entirely positive [1, 2, 3, 8], which not all signals are not guaranteed to be. To make them into acceptable inputs, data normalization is performed. Wasserstein distance is sensitive to this normalization, especially in the case of high frequency inputs, and often requires an extreme manipulation of the original signal to force convexity [1, 2, 3, 4]. Convexity is a very important property as with the absence of local extrema, there is no chance of an iterative method converging to the wrong solutions. Debiased Sinkhorn divergence does not suffer as much from this sensitivity to data normalization and maintains convexity without having drastically having to manipulate the original signal.

Section 2

Optimal Transport and Dissimilarity Measures

In this section we will discuss Kantorovich's optimal transport problem which will lead us to the formulation of the Wasserstein distance. Adding an entropic penalty term to the total transport cost allows us to approximate the solutions of the original OT problem, which leads to the formulation of Sinkhorn divergence and eventually Debiased Sinkhorn divergence.

As mentioned, Sinkhorn divergence and Wasserstein distance are types of dissimilarity measures of two probability distributions. Throughout this paper, we will be considering these probability distributions to be signals (one could think of them as seismic waves, for example).

2.1 Kantorovich's Optimal Transport Problem and Wasserstein Distance

Let $\mathcal{X}$ be a compact subset of Euclidean space $\mathbb{R}$. Also, let **f** and **g** be two n-dimensional probability vectors. This means they follow the structure $\Sigma_n \coloneqq \{\mathbf{f} \in \mathbb{R}_+^n : \mathbf{f}^{\mathrm{T}}\mathbf{1_n} = 1\}$, defined on $\mathcal{X}$. Here, $\mathbf{1_n}$ is the n-dimensional vector of ones. The two probability vectors **f** and **g** are assumed to be given as two sets of n discrete points $\{x_1, \dots, x_n\} \subset \mathcal{X}$ and $\{y_1, \dots, y_n\} \subset \mathcal{X}$ in $\mathbb{R}$, respectively.

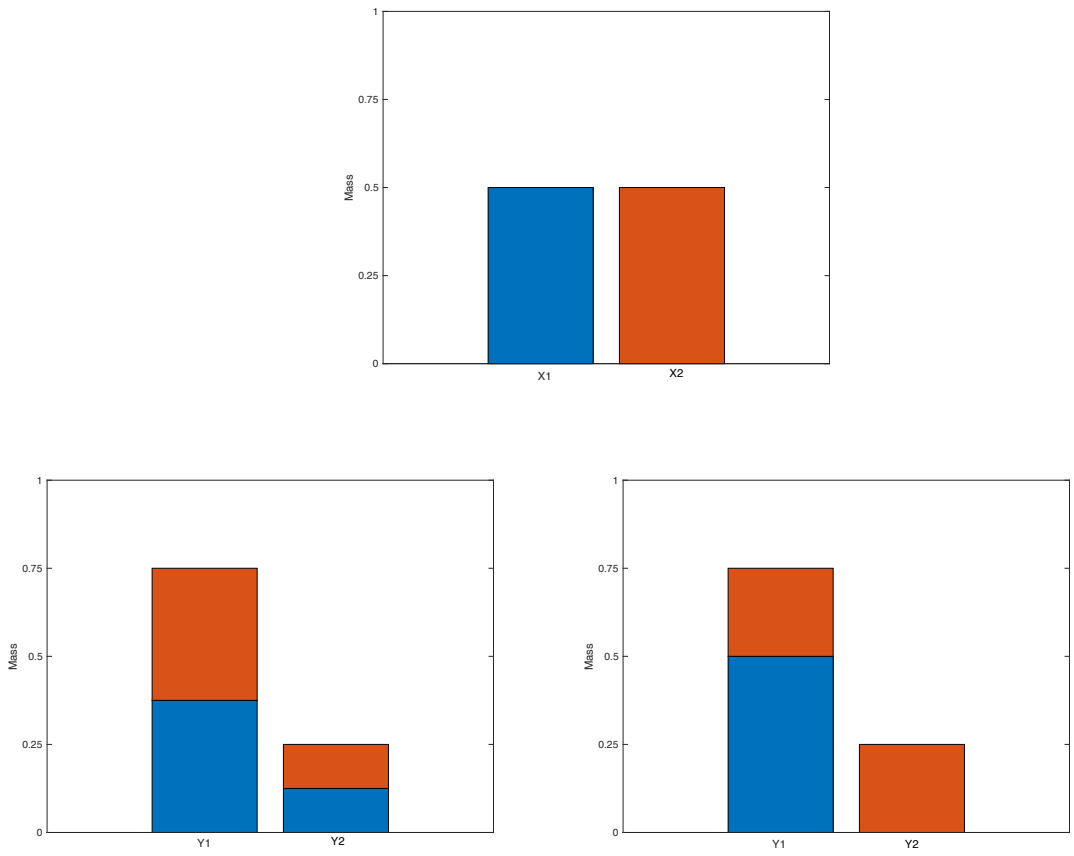


Figure 1: The top figure displays **f** which is consists of two components x_1 (blue) and x_2 (orange) each with a mass of $\frac{1}{2}$. The two color schemes in the bottom plots represent different ways to distribute the "mass" of x_1 and x_2 onto **g** which is comprised of $y_1 = \frac{3}{4}$ and $y_2 = \frac{1}{4}$.

Now, the goal is to figure out how much "work" it would take to map $\mathrm{x}_1, \ldots, \mathrm{x}_\mathrm{n}$ onto $\mathrm{y}_1, \ldots, \mathrm{y}_\mathrm{n}$ in an optimal fashion.

For this purpose, we will need to introduce a cost matrix defined by a distance function *d*.

$$C = [C_{ij}] \in \mathbb{R}_+^n \qquad C_{ij} = d(x_i, y_j)^p \quad i, j = 1, \ldots, n \;;\; p \in [1, \infty) \qquad (1)$$

Throughout this paper, we set $p = 2$ and we consider the Euclidean norm as the distance function

$$d(x_i, y_j)^2 \coloneqq (x_i - y_j)^2$$

We look for transport matrices $P \in \mathbb{R}_+^{nxn}$ where P_{ij} corresponds to the amount of "mass" we need to move from **f** at point x_i to **g** at point y_j. In order for this to be an admissible transport plan, the sum of rows of P_{ij} must be equal to **f** and the sum of columns must be equal to **g**. This is because all the mass taken from a point x_i must be equal to the mass at point x_i, and the mass taken to the target point y_j must be equal to the mass at the target point y_j. We call **f** and **g** the *marginals* of P [1]. With this we can now introduce the optimal transport problem, which is to find the optimal matrix P that transports **f** onto **g**. Let $P \in U(\mathbf{f}, \mathbf{g})$, where

$$U(\mathbf{f}, \mathbf{g}) \coloneqq \{P \in \mathbb{R}_+^{nxn}, P\mathbf{1_n} = \mathbf{f},\ P^T\ \mathbf{1_n} = \mathbf{g}\} \tag{2}$$

Here, $\mathbf{1_n}$ is the n-dimensional vector of ones. The "work" needed in using P as our transport plan is given by the Frobenius inner product $< P, C > = \sum_{i,j} P_{ij}\ C_{ij}$. We could conceptualize it as work being equivalent to mass multiplied by distance. Kantorovich's optimal transport problem aims to minimize this cost of transporting **f** onto **g**, which reads

$$T_c(\mathbf{f}, \mathbf{g}) \coloneqq \min_{P \in U(\mathbf{f}, \mathbf{g})} < P, C > \tag{3}$$

The Wasserstein distance of order p [1] is denoted

$$W_p(\mathbf{f}, \mathbf{g}) = \left(T_c(\mathbf{f}, \mathbf{g})\right)^{\frac{1}{p}} \tag{4}$$

2.2 Sinkhorn Divergence

One way to compute this solution is to regularize the problem and then try to approximate the solution of the regularized problem [1]. To perform this regularization, we use the idea from Cuturi [12] and add an entropic penalty term to the original problem and arrive at the following.

$$T_c^{\lambda}(\mathbf{f},\mathbf{g}) \coloneqq \min_{P\in U(\mathbf{f},\mathbf{g})} < P, C > -\frac{1}{\lambda}H(P) \tag{5}$$

Where $\lambda > 0$ is our regularization parameter, and $H(P)$ is the discrete entropy of the transport matrix.

$$H(P) \coloneqq -\sum_{i,j} P_{ij}\left(\log P_{ij} - 1\right) \tag{6}$$

Let P_λ be the optimal solution to the regularized problem. Then the Sinkhorn divergence of order p between $\mathbf{f}$ and $\mathbf{g}$ is given as [1],

$$S_{p,\lambda}(\mathbf{f},\mathbf{g}) \coloneqq < P_{\boldsymbol{\lambda}}, C >^{1/p} \tag{7}$$

Sinkhorn's Algorithm

Since we are interested in a constrained optimization problem, it is natural to use the Lagrangian. We first introduce two Lagrange multipliers $\hat{\mathbf{f}} \in \mathbb{R}^n$ and $\hat{\mathbf{g}} \in \mathbb{R}^n$ for the marginal constraints $P^T\mathbf{1_n} = \mathbf{f}$, $P\mathbf{1_n} = \mathbf{g}$.
The Lagrangian of (5) - (6) then reads

$$\boldsymbol{\mathcal{L}}\left(P,\hat{\mathbf{f}},\hat{\mathbf{g}}\right) = < P, C > -\frac{1}{\lambda}H(P) - \hat{\mathbf{f}}^{\mathrm{T}}(\,P\mathbf{1_n} - \mathbf{f}) - \hat{\mathbf{g}}(P^{\mathrm{T}}\mathbf{1_n} - \mathbf{g})$$

Setting $\partial_{P_{ij}}\boldsymbol{\mathcal{L}} = \mathrm{C}_{ij} + \frac{1}{\lambda}\log P_{ij} - \hat{\mathrm{f}}_i - \hat{\mathrm{g}}_j = 0$, we can solve for P_{ij}

$$\log P_{ij} = \lambda \hat{\mathrm{f}}_i + \lambda \hat{\mathrm{g}}_j - \lambda \mathrm{C}_{ij}$$

$$P_{ij} = e^{\lambda \hat{\mathbf{f}}_i + \lambda \hat{\mathbf{g}}_j - \lambda \mathrm{C}_{ij}}$$

Thus, one can write

$$P_{ij} = u_i \, Q_{ij} \, v_j \tag{8}$$

Where

$$u_i \coloneqq e^{\lambda \hat{\mathrm{f}}_i} \qquad Q_{ij} \coloneqq e^{-\lambda \mathrm{C}_{ij}} \qquad v_j \coloneqq e^{\lambda \hat{\mathrm{g}}_j} \tag{9}$$

Alternatively, one can write (8) - (9) in matrix factorization form [1], such that

$$P_\lambda = UQV \tag{10}$$

$$U = \mathrm{diag}(u_1, \dots, u_n) \qquad Q = [Q_{ij}] \qquad V = \mathrm{diag}(v_1, \dots, v_n) \tag{11}$$

Notice that due to the form of (8) and (9), and subsequently (10) and (11), that $P_\lambda \in \mathbb{R}_+^{nxn}$, i.e. P_λ is nonnegative. Also, the formulation of P_λ involves the multiplication of two nonnegative vectors called scaling vectors, which can be obtained using the marginal constraints,

$$P\mathbf{1_n} = \mathbf{f}, \qquad P^T\mathbf{1_n} = \mathbf{g}$$

$$UQV\mathbf{1_n} = \mathbf{f}, \qquad VQ^{\mathrm{T}}U\mathbf{1_n} = \mathbf{g} \tag{12}$$

And notice that

$$V\mathbf{1}_n = \boldsymbol{v}, \qquad U\mathbf{1}_n = \boldsymbol{u} \tag{13}$$

Therefore, one can write

$$\boldsymbol{u} \odot (Q\boldsymbol{v}) = \mathbf{f}, \qquad \boldsymbol{v} \odot (Q\boldsymbol{u}) = \mathbf{g} \tag{14}$$

Where $\odot$ denotes an entry-wise product, such as the operation " .* " in MATLAB.
Notice that (14) denotes two nonlinear equations involving nonnegative scalars and a nonnegative matrix. These equations can be solved using an iterative method known as Sinkhorn's algorithm (17), which begins with the following calculations [10, 11].

$$\boldsymbol{u}(i) = \mathbf{f} \oslash [Q\boldsymbol{v}(i-1)], \qquad \boldsymbol{v}(i) = \mathbf{g} \oslash [Q^{\mathrm{T}}\boldsymbol{u}(i)] \qquad\qquad i = 1, \dots, K \tag{15}$$

Where $\oslash$ denotes an entry-wise quotient, such as the operation " ./ " in MATLAB.
Now, when implementing this algorithm into a computer program, one will need to set a stopping criterion. A reasonable solution would be to measure the difference between our original inputs $\mathbf{f}$ and $\mathbf{g}$ and our most updated solutions $\boldsymbol{u} \odot (Q\boldsymbol{v})$ and $\boldsymbol{v} \odot (Q\boldsymbol{u})$ respectively, using the one-norm, and stop the algorithm once we have reached sufficient accuracy. More precisely, given a small tolerance $\epsilon_K > 0$,
we continue iterations until

$$\text{Max}\{\, ||\, \boldsymbol{u}(i) \odot [Q\boldsymbol{v}(i-1)] - \mathbf{f}||_1 \;,\; ||\boldsymbol{v}(i) \odot Q^{\mathrm{T}}\boldsymbol{u}(i) - \mathbf{g}\,||_1 \,\} \leq \epsilon_K \tag{16}$$

Once the scaling vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ have been computed to desired accuracy and putting the form of (8) - (11) into (5) - (6) and employing (14), we obtain the optimal transport cost in terms of two scaling vectors [1]. Once we have done this, we raise that cost to the power of $1/p$, and the p-Sinkhorn divergence is then

$$S_{p,\lambda} = \left[\frac{1}{\lambda}(\mathbf{f}^{\mathrm{T}} \log \boldsymbol{u} + \mathbf{g}^{\mathrm{T}} \log \boldsymbol{v} - 1)\right]^{1/p} \tag{17}$$

For the entirety of this paper, we will be considering the case where $p = 2$ in (4) and (17). One should note that the term "divergence" here is used in place of "distance" as a true distance function must satisfy the coincidence axiom and the triangle inequality.

2.3 Debiased Sinkhorn Divergence and its Properties

Notice that $S_\lambda(\mathbf{f}, \mathbf{f}) \neq 0$ due to the bias introduced by the entropic penalty term, thus not satisfying the coincidence axiom. To ensure that we are dealing with a true distance function, we need a way of debiasing the distance. Thus, throughout this paper we will be investigating the behavior and advantages of using what is known as *Debiased Sinkhorn divergence* (DS divergence), which for two probability vectors **f** and **g**, reads

$$DS(\mathbf{f}, \mathbf{g}) \coloneqq \sqrt{S_\lambda^2(\mathbf{f}, \mathbf{g}) - \frac{1}{2}\left(S_\lambda^2(\mathbf{f}, \mathbf{f}) + S_\lambda^2(\mathbf{g}, \mathbf{g})\right)} \qquad (18)$$

Note that we call this a "divergence" as opposed to a "distance" as one cannot analytically show that (18) satisfies the triangle inequality.

Theorem 1. Let $\mathcal{X}$ be a compact subset of Euclidean space $\mathbb{R}$ with a cost function $C(x, y)$ defined by (1) that induces, for $\lambda > 0$, a positive kernel defined by $k_\lambda(x, y) \coloneqq e^{-\lambda C(x,y)}$. Then, DS divergence defines a symmetric, positive definite, smooth distance function that is convex with respect to each of its input variables. It is a distance function in the sense that, for any probability measures f and g, $\mathrm{f} = \mathrm{g} \Leftrightarrow DS(\mathrm{f}, \mathrm{g}) = 0$. It is positive definite in the sense that $DS(\mathrm{f}, \mathrm{g}) \geq 0$, and symmetric in that $DS(\mathrm{f}, \mathrm{g}) = DS(\mathrm{g}, \mathrm{f})$.

Proof

An extensive proof of the above theorem is found in the proof of Theorem 1 in [9].
Note that DS divergence can accept continuous functions as inputs as well as vectors with a discrete index, which becomes noteworthy in the statement and proof of Theorem 3.

Section 3

Normalization and Convexity

3.1 Signed Signals

The fact that the inputs **f** and **g** are required to be positive probability distributions is the main problem of applying optimal transport to general signals. For example, oscillatory seismic waves where the signals dip into negative regions are not entirely positive nor are they normalized such that the components of the waves sum to 1. Thus, we must introduce a method of manipulating the data such that we attain acceptable probability distributions for vectors to be inputted into the Sinkhorn divergence algorithm [2].

3.2 Normalization

There are three popular methods of achieving signal positivity and forcing signals to be probability vectors [2, 3, 5]. We consider an exponential scaling (20), a linear scaling (21), and a newer type of exponential scaling that is now popular in practice [2] called softplus scaling (22). Given two signals **f** and **g** that are not probability distributions, we introduce the following normalizations:

$$\mathbf{f}_\delta = \frac{\sigma_\delta(\mathrm{f}(\mathbf{x}))}{\sum_i \sigma_\delta(\mathrm{f}(\mathrm{x}_i))} \tag{19}$$

Where

(i) σ_δ is one to one

(ii) σ_δ is a C^∞

For a hyperparameter $\delta > 0$, we introduce exponential scaling, linear scaling, and softplus scaling.

$$\sigma_{\delta,e}(\mathbf{f}) = \mathrm{e}^{\delta \mathbf{f}} \tag{20}$$

$$\sigma_{\delta,l}(\mathbf{f}) = \mathbf{f} + \delta \tag{21}$$

$$\sigma_{\delta,s}(\mathbf{f}) = \log\left(e^{\delta \mathbf{f}} + 1\right) \tag{22}$$

Note that when we write $e^{\delta \mathbf{f}}$, $\log\left(e^{\delta \mathbf{f}} + 1\right)$, $\mathbf{f} + \delta$, we are referring to component-wise exponentiation and component-wise addition, respectively.

One should be aware that when using (20) the normalization is very sensitive to the value of delta chosen. If a large delta is selected, the negative components of the signal will become suppressed, while the positive components become amplified which can lead to machine overflow. If delta is very small, (20) will shrink the data toward a value of one, and (22) will tend toward the value $\log(2)$. In the case of either a too small or too large choice delta, information about the original signal will be lost.

Since we are interested in analyzing the advantages of Debiased Sinkhorn divergence *after* its inputs have undergone normalization, we introduce the following. Let

$$DS_\delta(\mathbf{f}, \mathbf{g}) \coloneqq DS(\mathbf{f}_\delta, \mathbf{g}_\delta) \tag{23}$$

Where DS is given by (18) and $\mathbf{f}_\delta$ and $\mathbf{g}_\delta$ are given by (19) in Section 2.

Theorem 2: The normalized DS divergence (23) is a symmetric, positive definite, and smooth distance function with the normalization (19) under assumptions (i) and (ii).

Proof

Our goal is to show that $DS(\mathbf{f}_\delta, \mathbf{g}_\delta)$ being symmetric positive definite smooth distance function in terms of $\mathbf{f}_\delta$ and $\mathbf{g}_\delta$, as shown in Theorem 1, implies $DS_\delta(\mathbf{f}, \mathbf{g})$

is a symmetric positive definite smooth distance function in terms of $\mathbf{f}$ and $\mathbf{g}$.

Recall that if $\mathbf{f}$ and $\mathbf{g}$ are not probability vectors, then $DS(\mathbf{f}, \mathbf{g})$ is not well defined. Thus, we introduce $DS_\delta(\mathbf{f}, \mathbf{g}) \coloneqq DS(\mathbf{f}_\delta, \mathbf{g}_\delta)$ using any of the normalizations described by (20) through (22) on $\mathbf{f}$ and $\mathbf{g}$.

Part 1. We want to show that the normalized DS divergence is symmetrical, in effect, $DS_\delta(\mathbf{f}, \mathbf{g}) = DS_\delta(\mathbf{g}, \mathbf{f})$.

Well, $DS_\delta(\mathbf{f}, \mathbf{g}) = DS(\mathbf{f}_\delta, \mathbf{g}_\delta) = DS(\mathbf{g}_\delta, \mathbf{f}_\delta) = DS_\delta(\mathbf{g}, \mathbf{f})$

By symmetry of Sinkhorn divergence [1].

Part 2. We want to show that normalized DS divergence is smooth with respect to $\mathbf{f}$ and $\mathbf{g}$. Well, $DS_\delta(\mathbf{f}, \mathbf{g}) \coloneqq DS(\mathbf{f}_\delta, \mathbf{g}_\delta)$ is smooth with respect to $\mathbf{f}_\delta$ and $\mathbf{g}_\delta$ by Theorem 1. But $\mathbf{f}_\delta$ and $\mathbf{g}_\delta$ are smooth with respect to $\mathbf{f}$ and $\mathbf{g}$, respectively. Thus, $DS_\delta(\mathbf{f}, \mathbf{g})$ is smooth with respect to $\mathbf{f}$ and $\mathbf{g}$.

Part 3. We want to show that normalized DS divergence is positive definite, in effect that $DS_\delta(\mathbf{f}, \mathbf{g}) \geq 0$. Well, $DS_\delta(\mathbf{f}, \mathbf{g}) \coloneqq DS(\mathbf{f}_\delta, \mathbf{g}_\delta) \geq 0$ by Theorem 1.

Part 4. We want to show that normalized DS divergence satisfied the coincidence axiom, in effect,

$DS_\delta(\mathbf{f},\mathbf{g}) = 0 \Leftrightarrow \mathbf{f} = \mathbf{g}$.

Assume $DS_\delta(\mathbf{f},\mathbf{g}) = 0$. By definition, $DS_\delta(\mathbf{f},\mathbf{g}) = DS(\mathbf{f}_\delta,\mathbf{g}_\delta)$. Hence, $DS(\mathbf{f}_\delta,\mathbf{g}_\delta) = 0$. But we know that, by Theorem 1, $DS(\mathbf{f}_\delta,\mathbf{g}_\delta) = 0 \Leftrightarrow \mathbf{f}_\delta = \mathbf{g}_\delta$
Then because our chosen normalizations from (20) - (22) are smooth and one-to-one functions, we conclude that $\mathbf{f}_\delta = \mathbf{g}_\delta$ implies $\mathbf{f} = \mathbf{g}$.

Now assume $\mathbf{f} = \mathbf{g}$. Then, due to our smooth one-to-one normalizations from (20) – (22), $\mathbf{f}_\delta = \mathbf{g}_\delta$. We know from Theorem 1 that $\mathbf{f}_\delta = \mathbf{g}_\delta \Leftrightarrow DS(\mathbf{f}_\delta,\mathbf{g}_\delta) = 0$
And by definition, $DS(\mathbf{f}_\delta,\mathbf{g}_\delta) = DS_\delta(\mathbf{f},\mathbf{g})$. Thus, $DS_\delta(\mathbf{f},\mathbf{g}) = 0$

■

3.3 Convexity

In general normalization does not preserve convexity. We often need to use a parameter $\delta > 1$ in our normalization (18) through (20). The following theorem proves that Debiased Sinkhorn divergence indeed maintains convexity after softplus normalization (20) for δ large enough.

Theorem 3. Let f and g be two continuous and compactly supported functions on $\mathcal{X}$. Let f_δ, g_δ be the normalized functions of f and g based on softplus scaling (22). Then, $\exists\ \delta^* > 0$ such that $DS^2(\mathrm{f}_\delta,\mathrm{g}_\delta)$ is strictly convex with respect to f_δ and g_δ if $\delta > \delta^*$.

Proof

We know by Theorem 2 that $DS^2(\mathrm{f}_\delta, \mathrm{g}_\delta)$ is a smooth (and thus continuous) function with respect to δ. As $\delta \to \infty$,

$$\lim_{\delta\to\infty} \ \mathrm{f}_\delta$$

$$=$$

$$\lim_{\delta\to\infty} \ \frac{\log(\exp(\delta \mathrm{f}) + 1)}{< \log(\exp(\delta \mathrm{f}) + 1) >}$$

$$=$$

$$\lim_{\delta\to\infty} \ \frac{\log(\exp(\delta \mathrm{f}^+) + 1)}{< \log(\exp(\delta \mathrm{f}^+) + 1) >}$$

$$= \text{(L.H.)}$$

$$\lim_{\delta\to\infty} \ \frac{\frac{\mathrm{f}^+ \exp(\delta \mathrm{f}^+)}{\exp(\delta \mathrm{f}^+) + 1}}{< \frac{\mathrm{f}^+ \exp(\delta \mathrm{f}^+)}{\exp(\delta \mathrm{f}^+) + 1} >}$$

$$=$$

$$\frac{\mathrm{f}^+ \ \lim_{\delta\to\infty} \ \frac{\exp(\delta \mathrm{f}^+)}{\exp(\delta \mathrm{f}^+) + 1}}{\lim_{\delta\to\infty} \ < \mathrm{f}^+ \frac{\exp(\delta \mathrm{f}^+)}{\exp(\delta \mathrm{f}^+) + 1} >}$$

$$=$$

$$\frac{\mathrm{f}^+}{< \mathrm{f}^+ >}$$

$$:=$$

$$\tilde{\mathrm{f}}^+$$

And by the same procedure,

$$\lim_{\delta\to\infty} \ \mathrm{g}_\delta = \frac{\mathrm{g}^+}{< \mathrm{g}^+ >} := \tilde{\mathrm{g}}^+$$

Now, let $I(\mathrm{f}_\delta, \mathrm{g}_\delta) \coloneqq DS^2(\mathrm{f}_\delta, \mathrm{g}_\delta)$ and let $I_{\mathrm{g}_\delta \mathrm{g}_\delta}$ be the second partial derivative of I with respect to g_δ. Then since $DS^2\left(\tilde{\mathrm{f}}^+, \tilde{\mathrm{g}}^+\right)$ is strictly convex, $\lim_{\delta \to \infty} I_{\mathrm{g}_\delta \mathrm{g}_\delta} = I_{\tilde{\mathrm{g}}^+ \tilde{\mathrm{g}}^+}$ is positive definite. Since $I_{\mathrm{g}_\delta \mathrm{g}_\delta}$ is smooth and thus continuous with respect to δ, then $\exists\ \delta^* > 0$ such that that for $\delta > \delta^*$, $I_{\mathrm{g}_\delta \mathrm{g}_\delta}$ is positive definite, and $I(\mathrm{f}_\delta, \mathrm{g}_\delta)$ is convex with respect to g_δ. The same result holds with respect to f_δ.

■

Theorem 4. Let f and $\mathrm{g}(\theta)$, where $\theta \in \Theta$, be two continuous and compactly supported functions on $\mathcal{X}$ and let g be smooth with respect to θ. Let f_δ, $\mathrm{g}_\delta(\theta)$ be the normalized functions of f and $\mathrm{g}(\theta)$ based on softplus scaling (22). If $DS^2\left(\tilde{\mathrm{f}}^+, \tilde{\mathrm{g}}^+(\theta)\right)$ is convex with respect to θ, then $\exists\ \delta^* > 0$ such that $I(\mathrm{f}_\delta, \mathrm{g}_\delta) \coloneqq DS^2\left(\mathrm{f}_\delta, \mathrm{g}_\delta(\theta)\right)$ will remain strictly convex with respect to θ for $\delta > \delta^*$.

Proof

We know from Theorem 3 that $\lim_{\delta \to \infty} \mathrm{f}_\delta = \tilde{\mathrm{f}}^+$ and $\lim_{\delta \to \infty} \mathrm{g}_\delta(\theta) = \tilde{\mathrm{g}}^+(\theta)$.

Let the Hessian of I with respect to θ be denoted by $H(\theta, \delta)$.

$$H(\theta, \delta) = \begin{pmatrix} I_{\theta_1, \theta_1} & \cdots & I_{\theta_1, \theta_N} \\ \vdots & \ddots & \vdots \\ I_{\theta_N, \theta_1} & \cdots & I_{\theta_N, \theta_N} \end{pmatrix}$$

Notice that this a matrix valued continuous function in θ and δ, as $I(\mathrm{f}_\delta, \mathrm{g}_\delta)$ is smooth. Thus, $\lim_{\delta \to \infty} H(\theta, \delta) = H^+(\theta)$ where H^+ is the Hessian of $DS^2(\tilde{\mathrm{f}}^+, \tilde{\mathrm{g}}^+(\theta))$ and is hence symmetric positive definite with respect to θ as we assumed that $DS^2\left(\tilde{\mathrm{f}}^+, \tilde{\mathrm{g}}^+(\theta)\right)$ is convex. Then because of continuity of $H(\theta, \delta)$, $\exists\ \delta^* > 0$ such that for $\delta > \delta^*$, $H(\theta, \delta)$ is symmetric positive definite, and hence $DS^2\left(\mathrm{f}_\delta, \mathrm{g}_\delta(\theta)\right)$ is strictly convex with respect to θ.

■

Such results can be theoretically shown for softplus scaling (22). However, theory does not allow us to ensure that given $\delta > \delta^*$, DS divergence is convex in θ for our normalizations (20) and (21). The following section aims to convince the reader that normalized DS divergence indeed is convex in θ when using (20) and (21) through numerical representation.

Section 4

Numerical Experiments

4.1 A Numerical Example with Exponential Normalization

Let $\mathbf{f} = \mathrm{f}(\boldsymbol{x}) \in \mathbb{R}^n$ be a discrete three-pulse signal, where

$$\mathrm{f}(\boldsymbol{x}) = e^{-\left(\frac{x_i-0.4}{w}\right)} - e^{-\left(\frac{x_i-0.5}{w}\right)} + e^{-\left(\frac{x_i-0.6}{w}\right)}$$

$x_i = \frac{i-1}{n-1} \in [0,1], \quad i = 1, \dots, n$

Here, $w > 0$ is a positive constant that affects the frequency of the three pulses. A smaller w creates more high frequency pulses. Now consider **g**, which will simply be a shifted version of **f**, i.e.

$$\mathbf{g}(\theta) = [\mathrm{f}(\boldsymbol{x} - \theta)] \in \mathbb{R}^n$$

Where θ is our shift, and θ is subtracted from each value of x_i.

Our goal in using Sinkhorn divergence begins with altering this signal in a way so that all parts of the function are positive and its components add up to one. Thus, for this experiment we use exponential normalization (20) beginning with $\delta = 1$.

Figure 2 top shows the signals **f** and $\mathbf{g}(\theta = 0.3)$ for two different frequencies $w = 0.05$ (low frequency signals) and $w = 0.01$ (high frequency signals). In figure 2 bottom left and bottom right, allowing $-0.3 < \theta < 0.3$, we can compare three following dissimilarity measures.

- L_2 norm: $L_2 \coloneqq \|\mathbf{f} - \mathbf{g}(\theta)\|_2^2$
- Normalized quadratic Wasserstein distance: $W^2\big(\mathbf{f}_\delta, \mathbf{g}_\delta(\theta)\big)$
- Normalized quadratic Debiased Sinkhorn divergence: $DS^2\big(\mathbf{f}_\delta, \mathbf{g}_\delta(\theta)\big)$

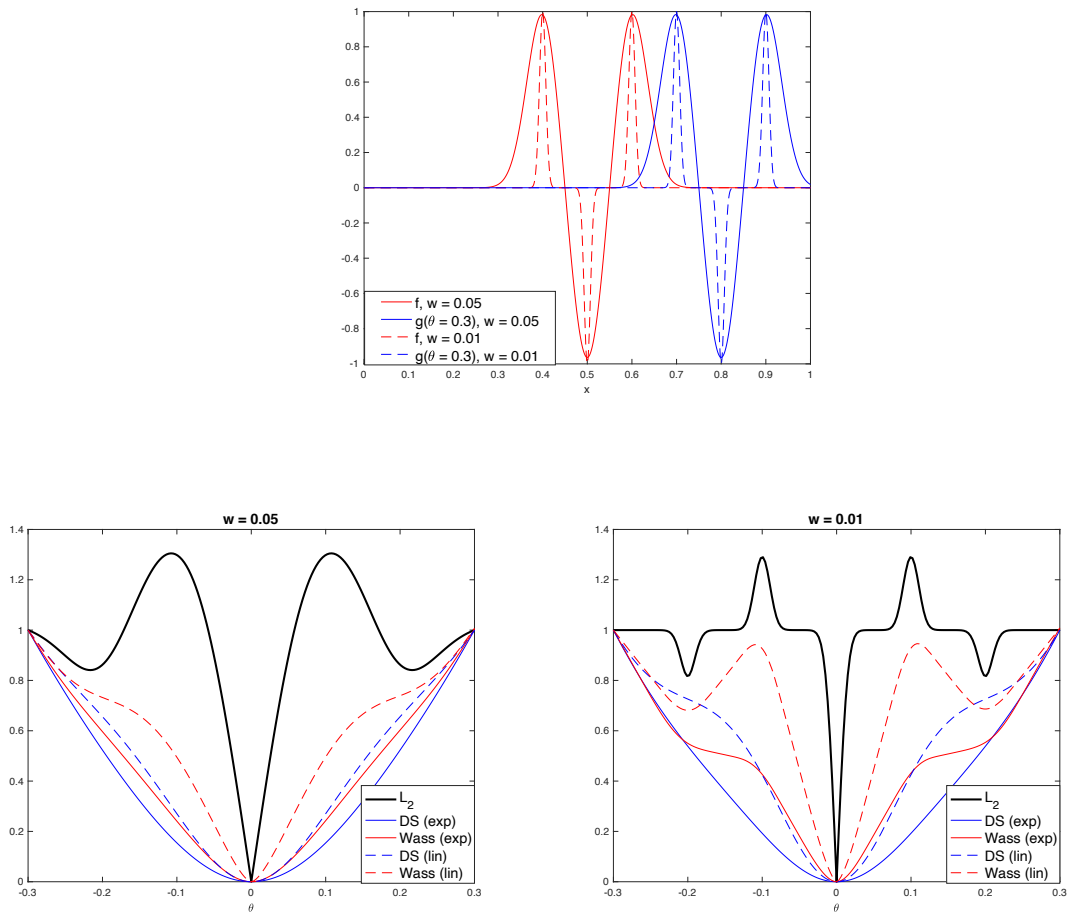


Figure 2: The top figure shows two types of low frequency (solid curves) and high frequency (dashed curves) signals. Normalized DS divergence, normalized Wasserstein distance, and the L_2 norm are shown in the lower left for low frequency signals and in the lower right for high frequency signals.

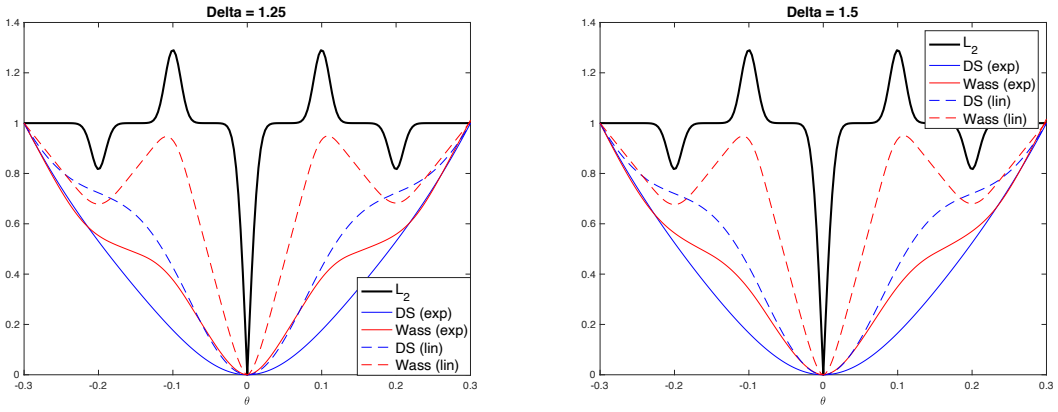
As we can see in Figure 2, using the L_2 norm (while commonly used and inexpensive to compute) produces local extrema, thus losing convexity. This is an issue because when attempting to find an absolute minimum using an iterative method, such local extrema may lead to an algorithm converging to the wrong solution. We see that in the high frequency and low frequency case, the L_2 norm fails to maintain convexity for even a small range of θ, though it is

slightly less extreme in the low frequency case. The normalized Wasserstein distance remains quasi-convex after data normalization in the low frequency case only. As the frequency of the input signals increases, the normalized Wasserstein distance begins producing these undesirable local extrema. In contrast, the normalized Debiased Sinkhorn method proves to maintain convexity (or quasi-convexity) even in the case when the inputs are high frequency signals.

In this paper, we are interested in examining how adjusting the hyperparameter δ may improve the convexity of normalized Wasserstein distance; however, this poses two problems. The first is that a value of $\delta = 1$ may be sufficient for Wasserstein to maintain convexity if the signals are of low enough frequency – however, it is unclear how much larger delta must become in order to force convexity for higher frequency inputs. One might posit that we simply always set δ to be a large value, forcing convexity for even high frequency inputs, which leads us to the second issue. Notice that when using exponential scaling, a large value of delta will drastically shrink the negative portions of the original signal, and the positive parts will become rapidly become amplified. We examine in the following delta study whether we can force convexity for normalized Wasserstein distance without setting δ too large.

Delta Study:

We are interested in studying how increasing values of δ may change this convexity issue. Below are the results.



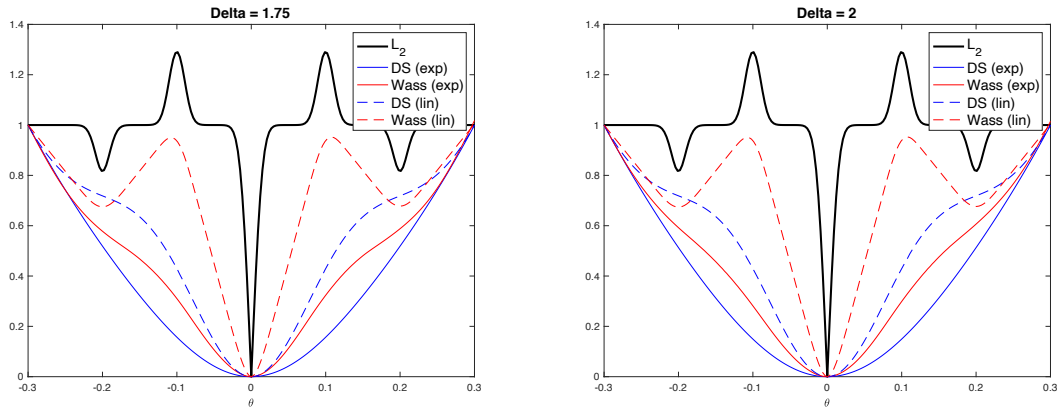


Figure 3: The above figures compare the normalized Debiased Sinkhorn distance (blue) to the normalized Wasserstein distance (red) and the classical Euclidean norm (black) for increasing values of delta. We use two methods of normalization – exponential normalization (solid curves) and linear normalization (dashed curves).

From the results above, we have failed to introduce a value for δ that forces convexity when using linear scaling on the Wasserstein distance that is not much larger than 1. In the case of exponential scaling, we were able to obtain obvious quasi-convexity with $\delta = 1.5$ for both normalized Wasserstein distance and DS divergence. In all cases, normalized Debiased Sinkhorn divergence outperformed the other methods, regardless of the choice of scaling.

The example above alone should not be sufficient in convincing the reader of the advantage of using Debiased Sinkhorn divergence as an alternative to Wasserstein distance. Throughout this paper, we will introduce other high frequency signals and compare the performance of normalized Wasserstein distance and Debiased Sinkhorn divergence for various values of δ. We will also analyze what happens when we consider θ as a frequency parameter, amplitude parameter, dilation parameter, as well as a phase shift.

4.2 Ricker Wavelets

To investigate this phenomenon of convexity further, we aim to analyze other simulated waves that we might encounter. We begin with a simple ricker wavelet, which is shown below.

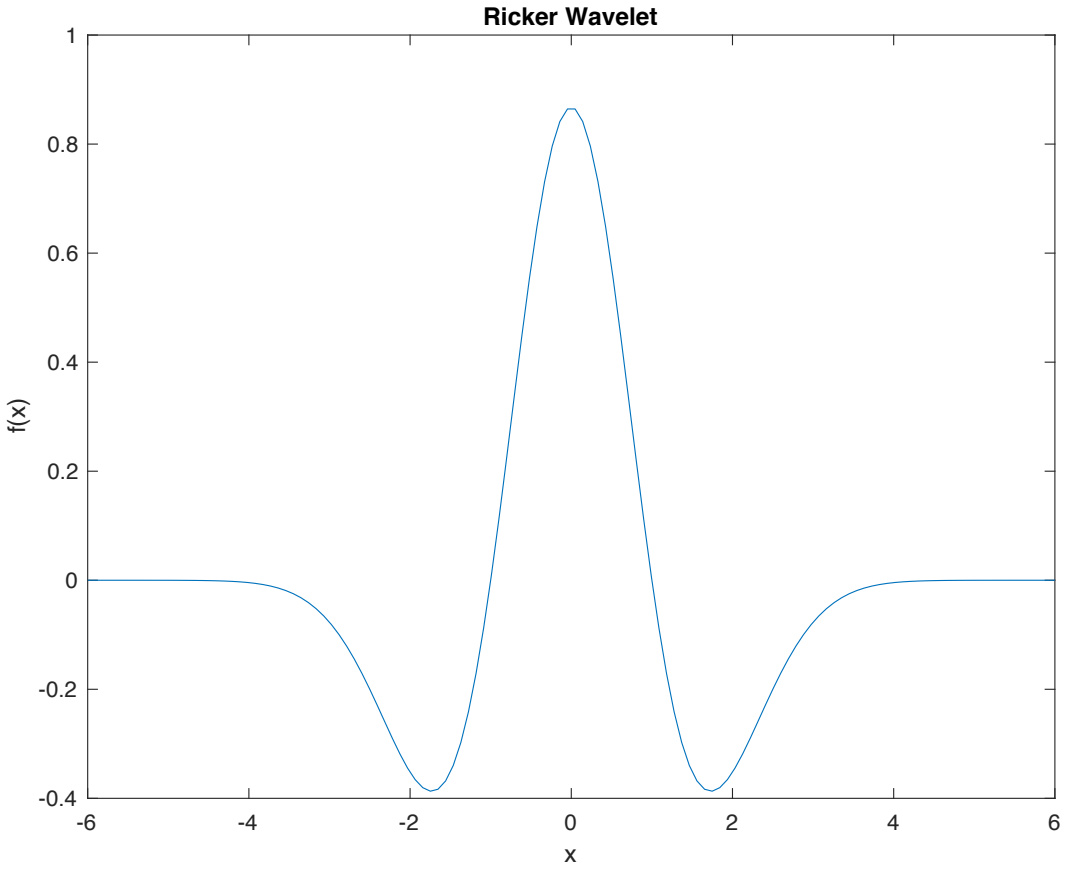


Figure 4: The standard Ricker Wavelet is shown for $x \in [-6, 6]$

A good way to simulate different kinds of waves is to use linear combinations of Ricker wavelets with m terms. This can be constructed using the following form.

$$\mathrm{f}(\boldsymbol{x}; a, w) \coloneqq \sum_{k=1}^{m} \tilde{\mathrm{f}}_k(\boldsymbol{x}; a, w)$$

Where

$$\tilde{\mathrm{f}}_k(\boldsymbol{x}; a, w) \coloneqq \frac{2\xi_k a}{\sqrt{3\alpha_k w}\pi^{\frac{1}{4}}}\left(1-\left(\frac{x_i - s_k}{\alpha_k w}\right)^2\right)e^{-\frac{x_i^2}{2(\alpha_k w)^2}}$$

$$i = 1, \dots n$$

Throughout this paper, we will have $m = 5$, with ξ_k, α_k, and s_k fixed such that,

$$[\xi_1, \xi_2, \xi_3, \xi_4, \xi_5] = [1, 1, 1, 0.5, 0.3]$$

$$[\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5] = [4, 3, 2, 1, 0.5]$$

$$[s_1, s_2, s_3, s_4, s_5] = [-1, 3, 1, 5, -3]$$

Note that a is an amplitude parameter, and w is a frequency parameter. Smaller w will result in higher frequency waves while larger w will result in lower frequency waves.

We need to generate **g**, a wave to compare with **f**. Furthermore,

In this paper, we will consider several different types of manipulations using a parameter, θ. We will analyze phase shifts, amplitude manipulation, frequency manipulation, and dilation. Thus, we will have,

$$\mathbf{g}_s(\theta) \coloneqq \mathrm{f}(\boldsymbol{x} - \theta; a, w) \tag{25}$$

$$\mathbf{g}_\mathrm{a}(\theta) \coloneqq \mathrm{f}(\boldsymbol{x};\ \theta a, w) \tag{26}$$

$$\mathbf{g}_w(\theta) \coloneqq \mathrm{f}(\boldsymbol{x}, a, \theta w) \tag{27}$$

$$\mathbf{g}_\mathrm{d}(\theta) \coloneqq \mathrm{f}(\theta \boldsymbol{x}, a, w) \tag{28}$$

Where the normalization for **f** and **g**, in all of these cases, follow (19). Note that in the case of (25), we are referring to a component-wise subtraction of θ.

4.3 Exponential Normalization on Ricker Wavelets

Example 1 (Phase Shift for High, Low, and Mixed Frequency Waves):

In this example, we will study how the normalized Debiased Sinkhorn divergence, the normalized Wasserstein distance, and the classical L_2 norm behave when given inputs of $\mathbf{f}_\delta$ and $\mathbf{g}_\delta$ that represent very high frequency, somewhat mixed frequency, as well as very low frequency signals. In order to capture and represent all the pulses of the signals (particularly in the case of the high frequency wave) we need **x** to be large, so we now set **x** to be a vector ranging from -1 to 1 with $n = 2^{10}$ evenly spaced entries.

As mentioned, we will consider three different waves, and we will study them separately.

For our high frequency wave, let

$$w = 0.1$$

For our mixed frequency wave, let

$$w = 0.5$$

For our low frequency wave, let

$$w = 1$$

After creating our linear combination, we receive the following figures displaying **f** and the scaled version $\mathbf{f}_\delta$ using (20) for $\delta = 1$.

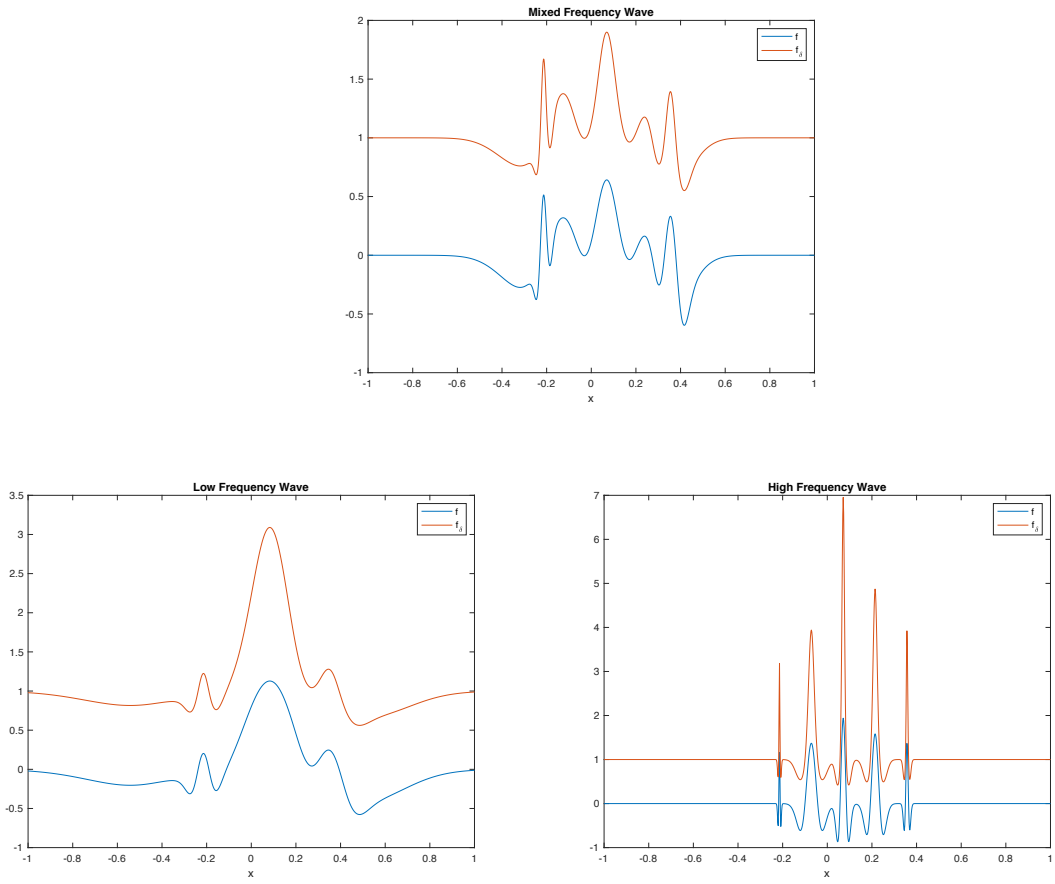


Figure 5: The figures display **f** (blue) compared with the exponentially normalized version $\mathbf{f}_\delta$ (orange) for mixed frequency (top) low frequency (bottom left) and high frequency (bottom right) waves for $\delta = 1$ in (20).

Note that for our high frequency wave, even to the human eye the dissimilarity measure looks to be very difficult to compute when compared to the low or even mixed frequency simulations. Solely based off these images, a hypothesis could be formulated that very high frequency signals could pose difficulties when conducting a dissimilarity measure. Thus, one could assume that suboptimal metrics and distance functions may lose convexity and therefore become undesirable after normalization. Our goal is to explore whether the Debiased Sinkhorn divergence may be the most optimal measure.

For the construction of the other signal **g**, we will use (25) where θ is our shift.
The following graph shows the two signals in the same figure with $\theta = 4$.

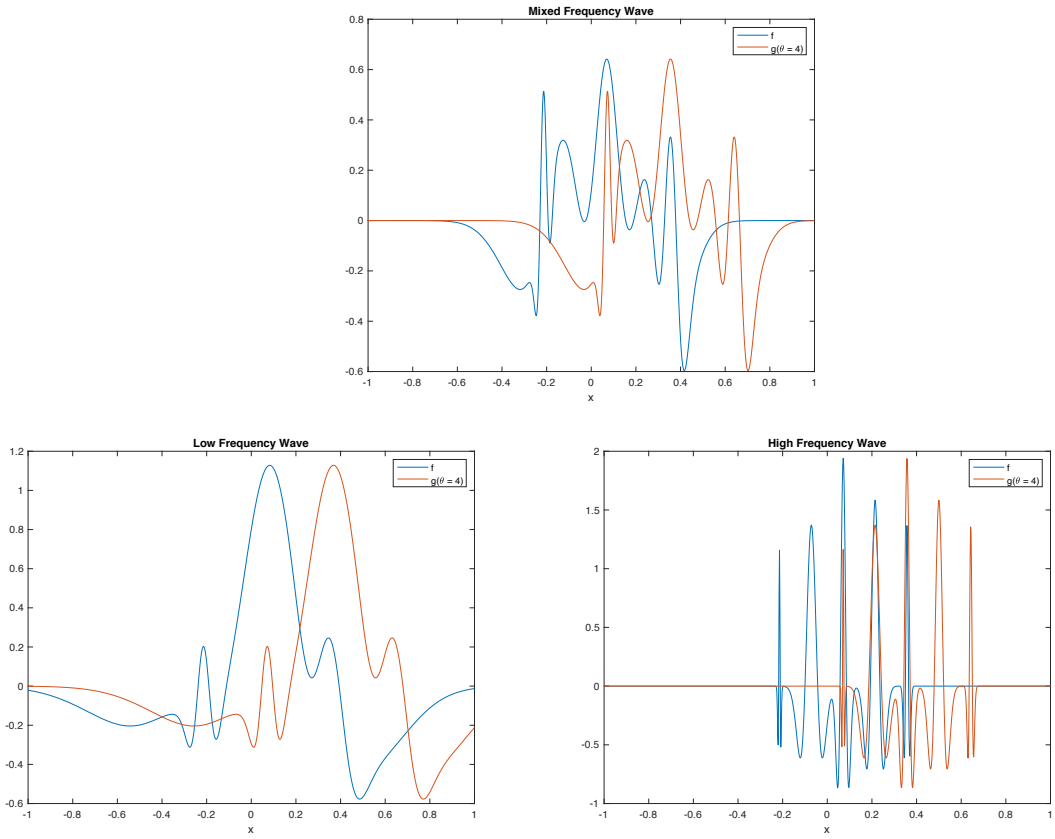


Figure 6: The above figures show **f** (blue) compared to the shifted signal **g** (orange) with a shift of $\theta = 4$ for our high mixed frequency wave (top) our low frequency (bottom left) wave and our high frequency wave (bottom right.)

We want to study the dissimilarity between **f** and **g** with a broader range of shifts, θ, compared to that in Section 4.1. We now allow 151 choices of θ ranging from -4 to 4. Now that we have more complicated signals, we proceed to create the following figures that compare the normalized Debiased Sinkhorn divergence with the classical L_2 norm as well as the normalized Wasserstein distance when using $\delta = 1$ in (20).

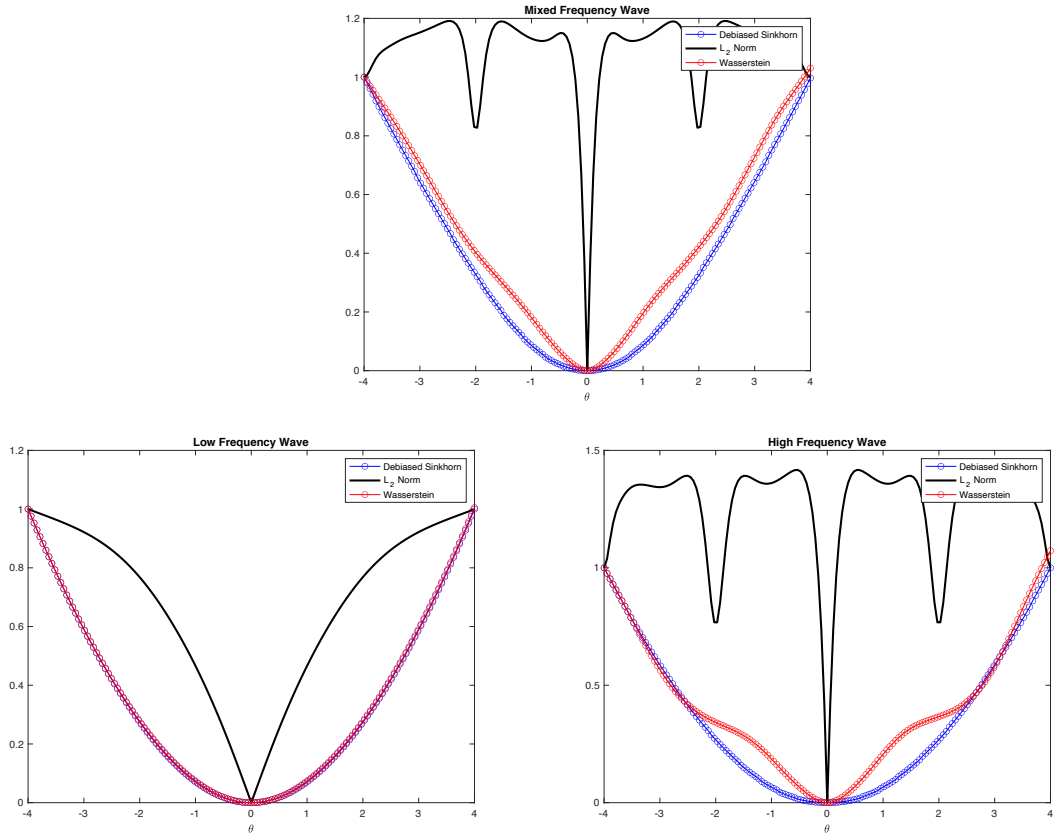


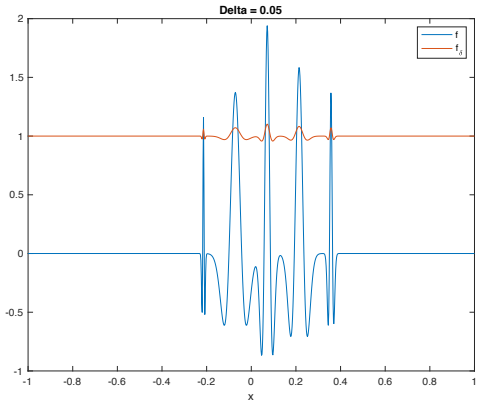
Figure 7: The above figures compare the convexity of the normalized Debiased Sinkhorn divergence (blue) against the classical L_2 norm (black) and the normalized Wasserstein distance (red) for our high mixed frequency wave (top) our low frequency (bottom left) wave and our high frequency wave (bottom right) when setting $\delta = 1$ in (20).

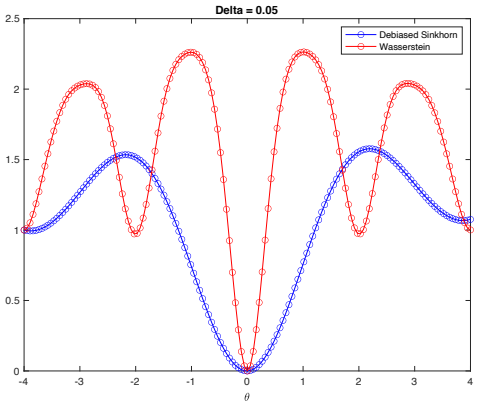
Upon inspection we can clearly see the benefits of using the normalized Debiased Sinkhorn divergence as opposed to the classic L_2 norm, as the L_2 norm quickly loses convexity as we deviate from $\theta = 0$, creating local extrema. This effect is more profound with higher frequency waves. A much closer contender is found in the normalized Wasserstein distance, shown as the red line in the above figure. Wasserstein is still quasi-convex in this example and does not pose any problems.

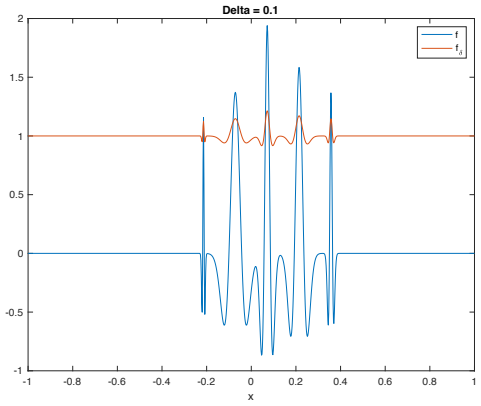
Notice that the above graphs used the exponential scaling (20) with $\delta = 1$. In the next example, we investigate how this is affected when we choose smaller and larger values of δ for our high frequency wave, as such waves seem to be slightly more of a challenge for our dissimilarity measures.

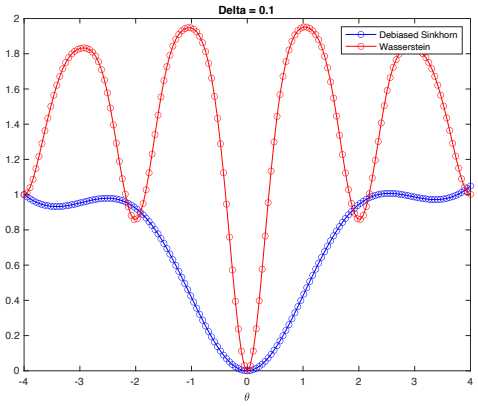
Example 2 (δ Study for Phase Shifts in High Frequency Waves)

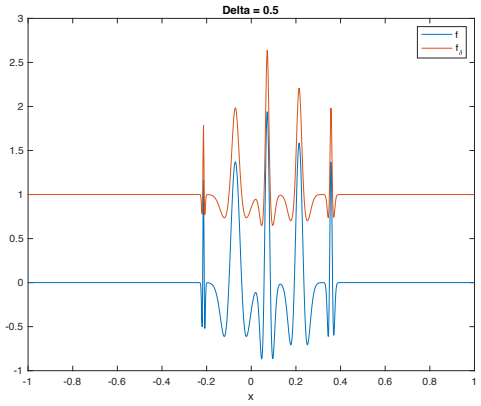
In the following example we will investigate how different δ values impact the convexity of the normalized Wasserstein distance as opposed to the normalized Debiased Sinkhorn divergence for the high frequency waves shown in Figure 6.

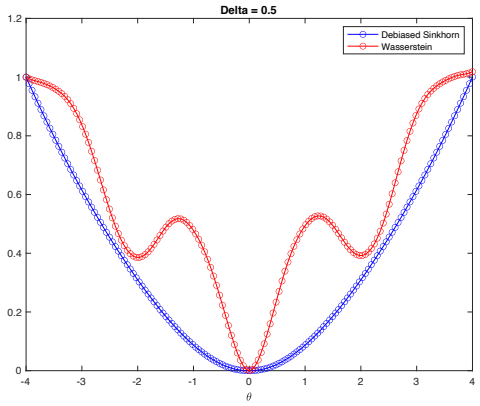


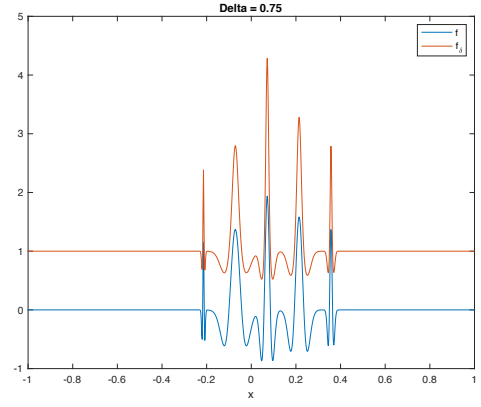


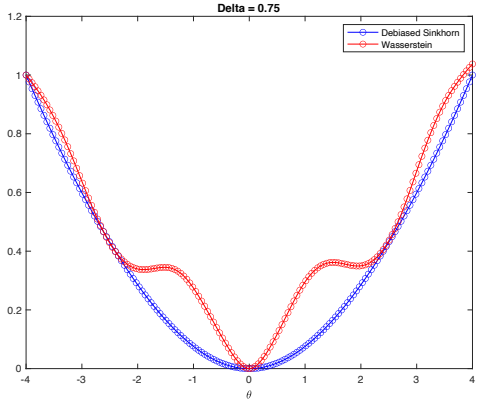


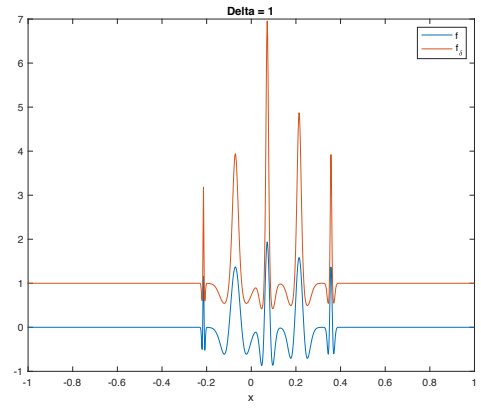


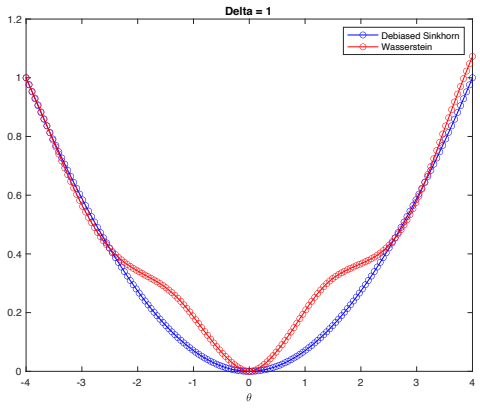

Delta = 0.5
f
f_δ
x


Delta = 0.5
Debiased Sinkhorn
Wasserstein
θ


Delta = 0.75
f
f_δ
x


Delta = 0.75
Debiased Sinkhorn
Wasserstein
θ


Delta = 1
f
f_δ
x


Delta = 1
Debiased Sinkhorn
Wasserstein
θ

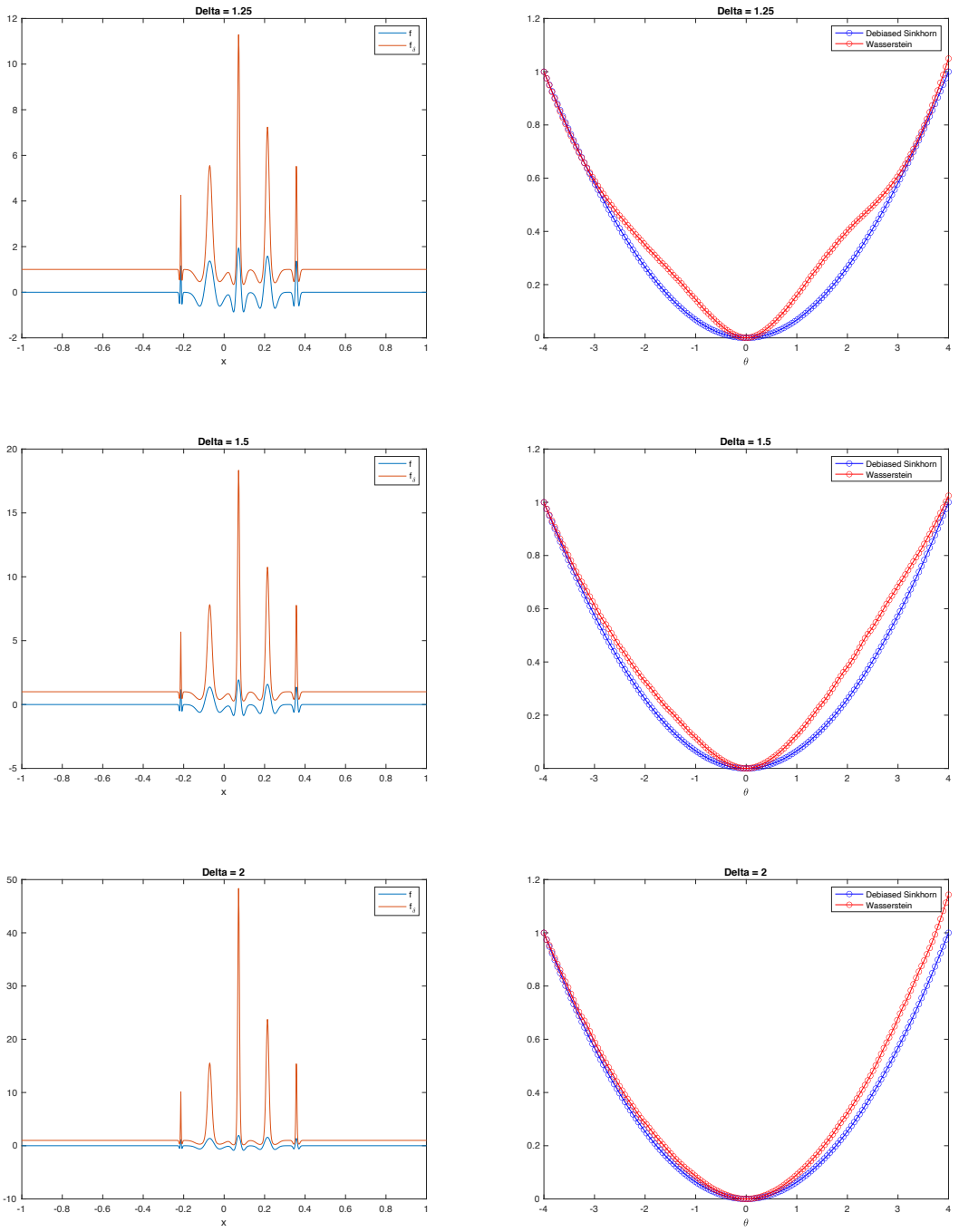


Figure 8: The figures on the right compare the normalized Debiased Sinkhorn divergence (blue) to the normalized Wasserstein distance (red) for our high frequency waves for increasing values of δ when considering θ as a phase shift (25). The figures on the left display the effect that increasing values of delta have on our exponential scaling $\mathbf{f}_\delta$ (orange) in comparison to the original signal $\mathbf{f}$ (blue.)

Notice immediately that for δ small, the amplitude of $\mathbf{f}_\delta$ is now smaller than that of $\mathbf{f}$ and all of the data shrinks toward the value 1. For for δ large, $\mathbf{f}_\delta$ is greatly amplified and quickly dwarfs $\mathbf{f}$. Recall that when delta was equal to 1 the amplitudes were unchanged. This is due to the nature of our exponential scaling. Having a large δ is therefore problematic because upon computing $\mathbf{f}_\delta$ we lose information about the large negative and positive sections of $\mathbf{f}$, as the negative parts shrink to zero and the positive parts are greatly amplified. However, a significantly large delta is required for the normalized Wasserstein distance to achieve strict convexity. Therefore, we see the advantage of using normalized DS divergence as an alternative, which achieved convexity for a smaller value of δ.

Example 3 (δ Study for Amplitude Changes in High Frequency Waves)

In this section we investigate how the convexity of the normalized Debiased Sinkhorn divergence, the normalized Wasserstein distance, and the L_2 norm is affected when considering changes in amplitude. In this example we now let $\mathbf{g}(\theta) \coloneqq \mathbf{g}_a(x;\theta)$ from (26). Because we do not want to introduce negative amplitudes (which is not realistic) we shrink the range of θ, now letting it consist of 151 choices ranging from 0.5 to 1.5.

A graph of $\mathbf{f}$ and $\mathbf{g}$ is presented below, with $\theta = 1.5$.

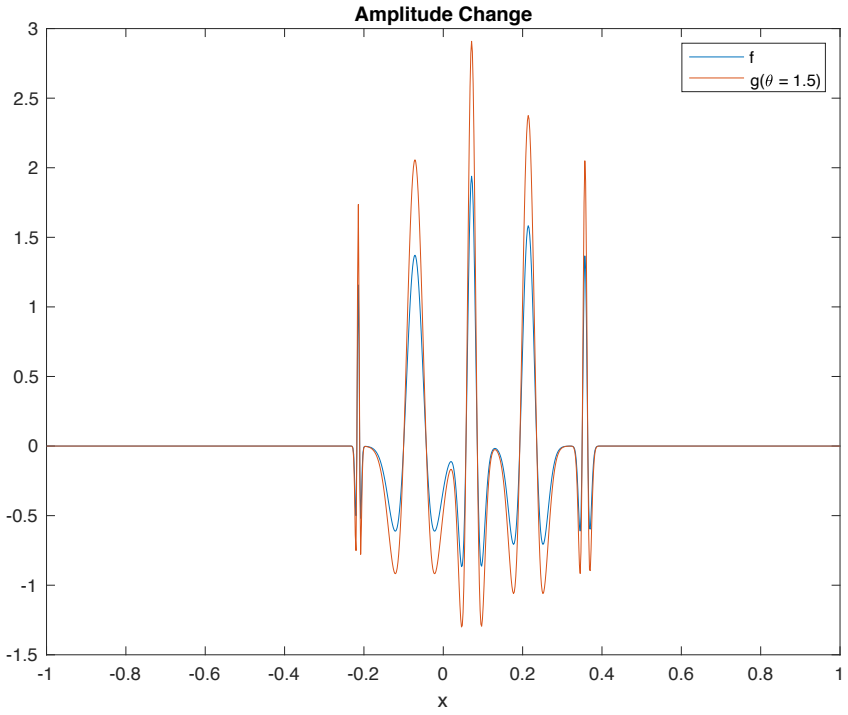
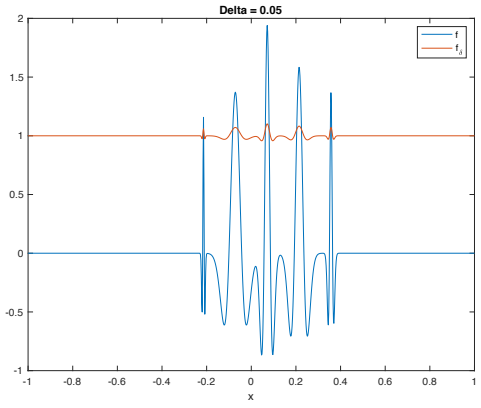
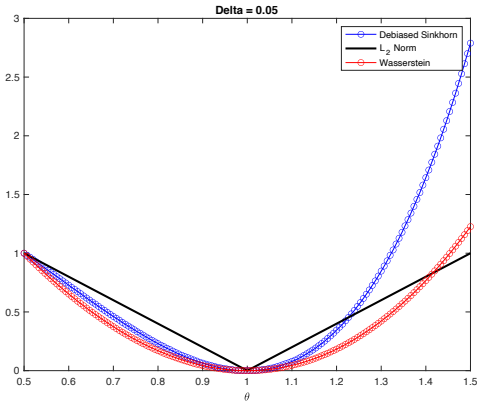
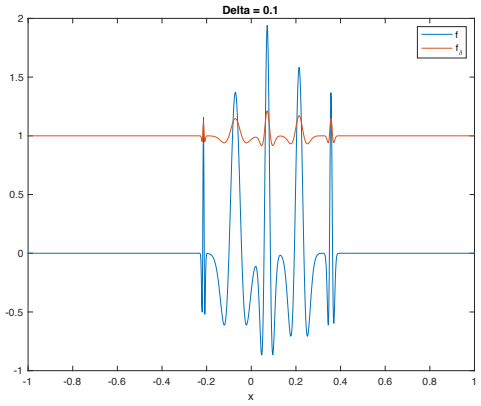


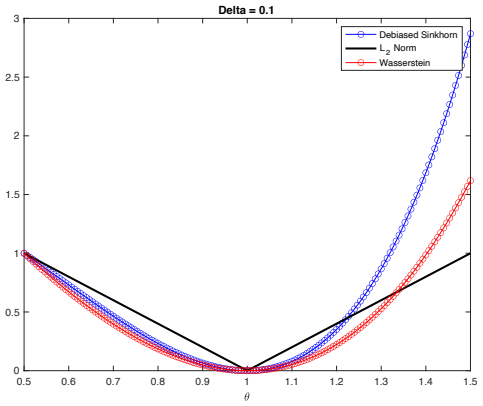
Figure 9: The above figure compares **f** and $\mathbf{g}(\theta)$ for $\theta = 1.5$. Here, θ effects the amplitude of the signal (26).

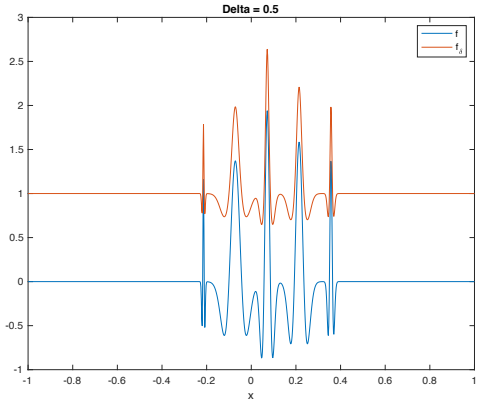
We now investigate how different δ values affect how our dissimilarity measures behave when considering these changes in amplitude.

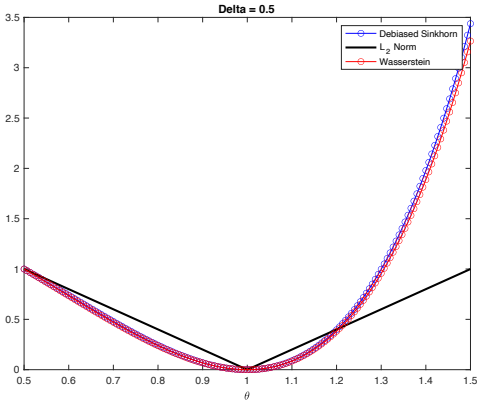


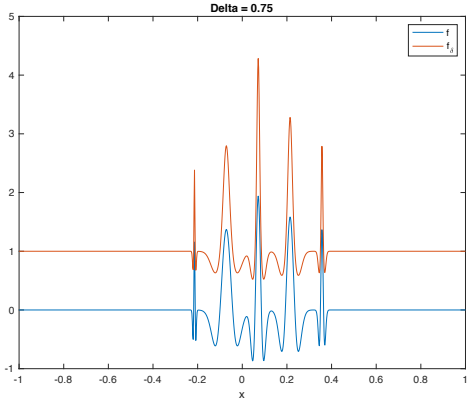


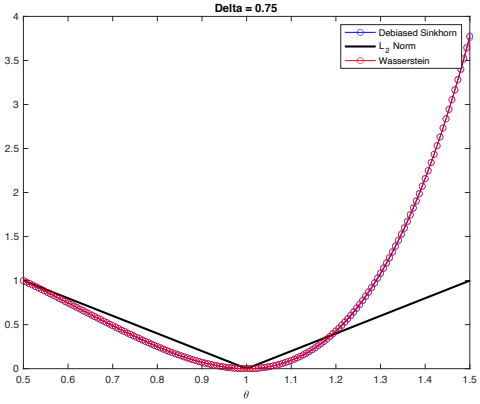

Delta = 0.1
f
f_δ
x

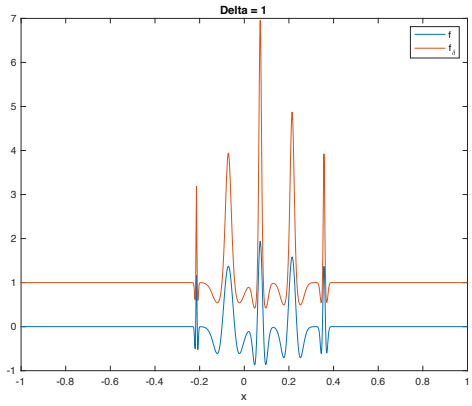

Delta = 0.1
Debiased Sinkhorn
L_2 Norm
Wasserstein
θ

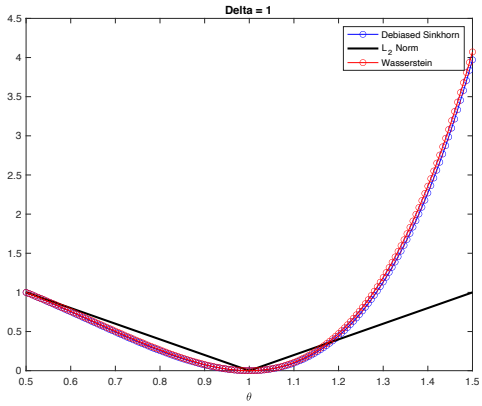

Delta = 0.5
f
f_δ
x

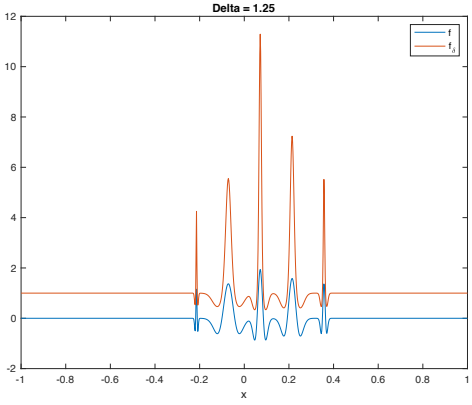

Delta = 0.5
Debiased Sinkhorn
L_2 Norm
Wasserstein
θ

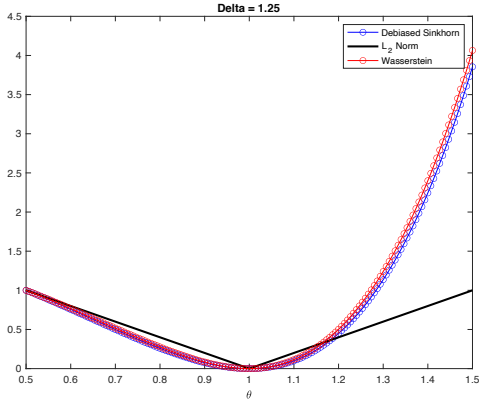

Delta = 0.75
f
f_δ
x

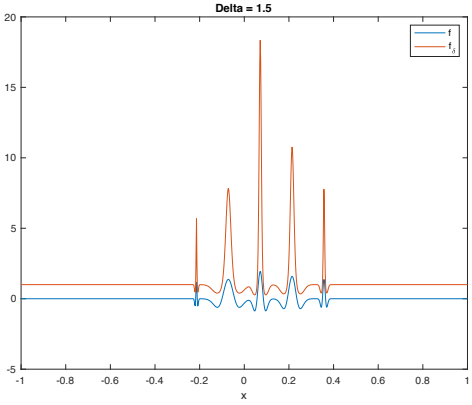

Delta = 0.75
Debiased Sinkhorn
L_2 Norm
Wasserstein
θ

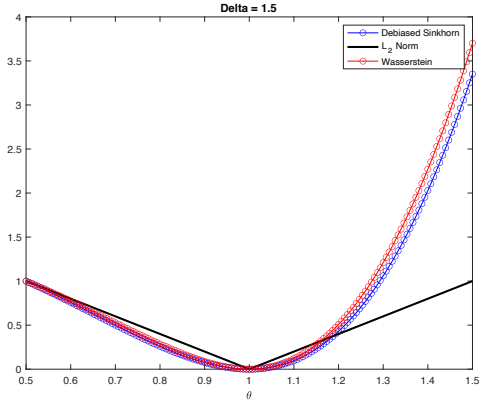

Delta = 1
f
f_δ
x


Delta = 1
Debiased Sinkhorn
L_2 Norm
Wasserstein
θ


Delta = 1.25
f
f_δ
x


Delta = 1.25
Debiased Sinkhorn
L_2 Norm
Wasserstein
θ


Delta = 1.5
f
f_δ
x


Delta = 1.5
Debiased Sinkhorn
L_2 Norm
Wasserstein
θ

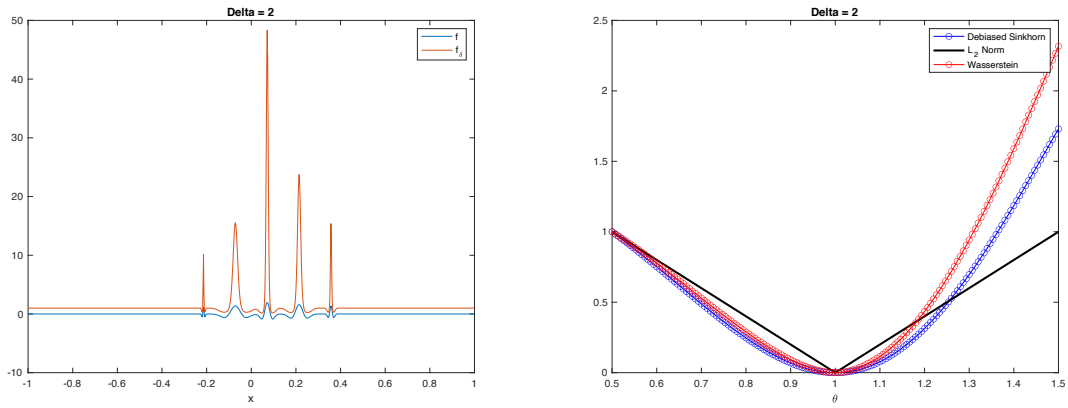


Figure 10: The figures on the right compare the normalized Debiased Sinkhorn divergence (blue) to normalized Wasserstein distance (red) and the L_2 norm (black) for our high frequency wave for increasing values of delta when considering θ as an amplitude parameter (26). The figures on the left display the effect that increasing values of delta have on our exponential scaling $\mathbf{f}_\delta$ (orange) in comparison to the original signal $\mathbf{f}$ (blue.)

We notice that when considering an amplitude change (26), all three distance functions maintain convexity for each of the various values of δ shown in Figure 10. Therefore we may conclude that when considering amplitude change, there is no specific advantage to using normalized DS divergence over the other measures when it comes to maintaining or achieving convexity.

Example 4 (δ Study for Frequency Changes in High Frequency Waves)

We now consider what happens when we manipulate the frequency parameter w of $\mathbf{f}$, so we consider $\mathbf{g}(\theta) \coloneqq \mathbf{g}_w(x;\theta)$ from (27). Note that changes in w will affect amplitude as well as frequency of $\mathbf{f}$. The parameter w also appears under a radical, so we need to avoid introducing a negative θ. For this purpose, we will conduct our delta study using 151 choices of θ ranging from 0.5 to 1.5.

A graph of $\mathbf{f}$ and $\mathbf{g}$ is presented below, with $\theta = 1.5$.

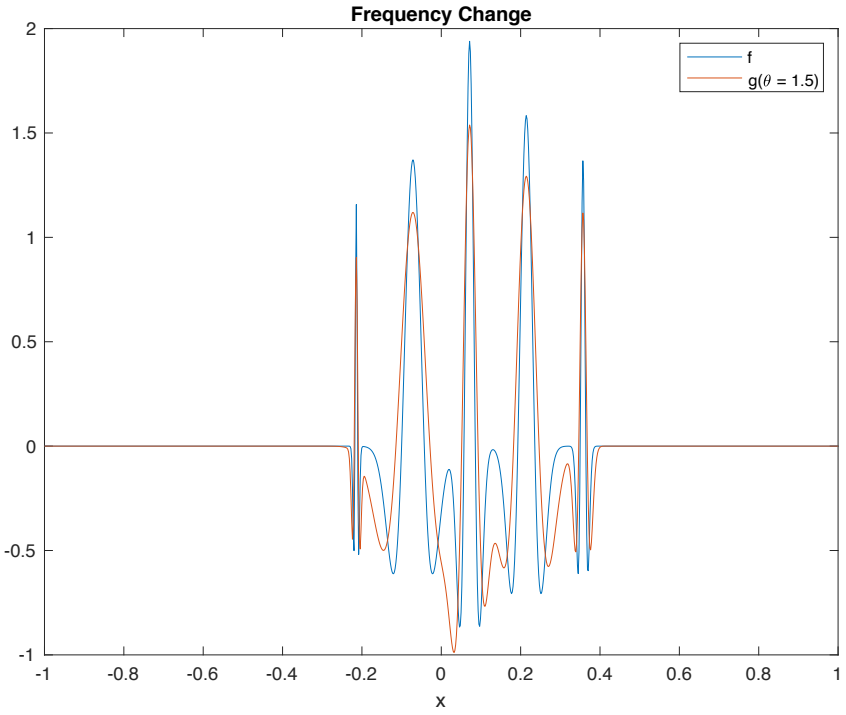
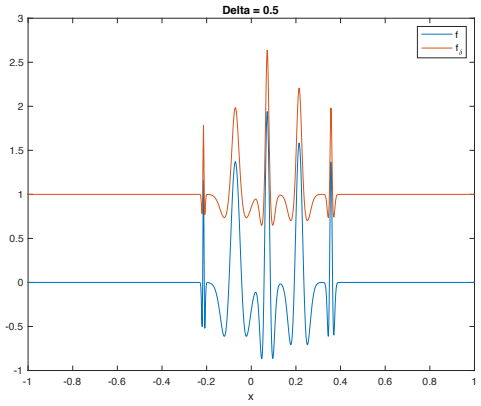
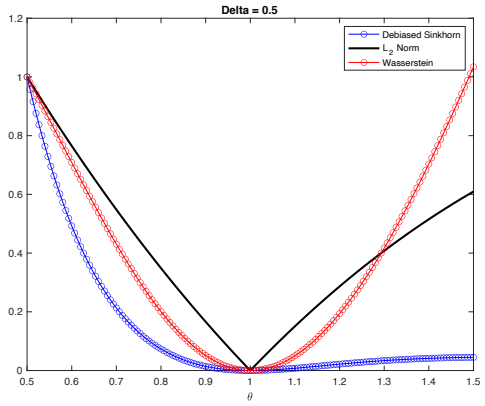
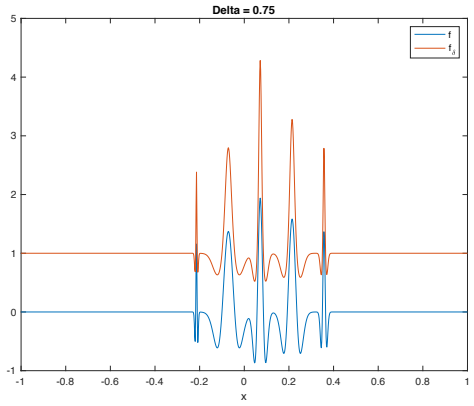


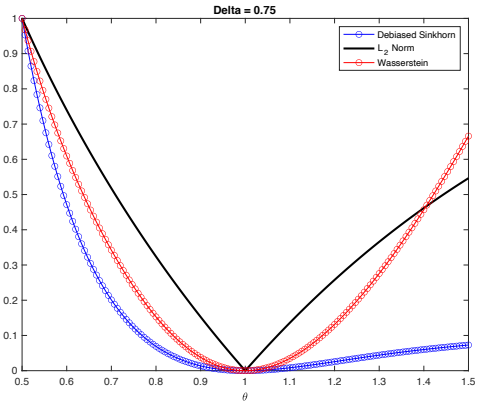
Figure 11: The above figure compares $\mathbf{f}$ and $\mathbf{g}(\theta)$ for $\theta = 1.5$. Here, θ effects the frequency of the signal (27).

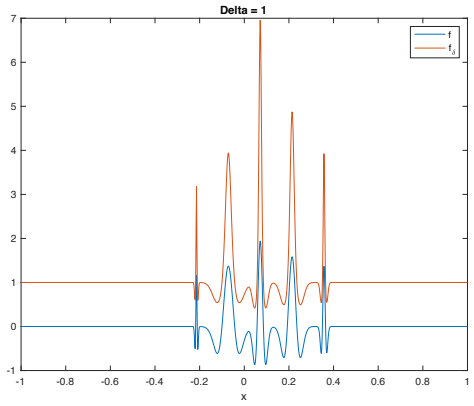
Our delta study is shown below.

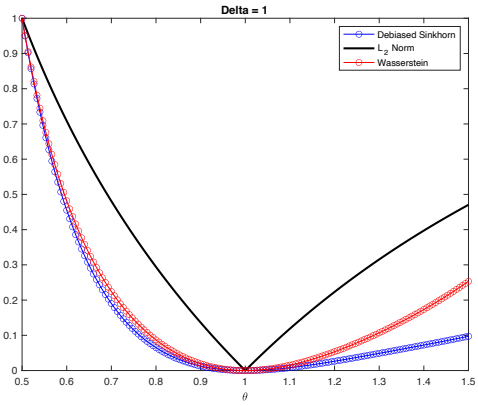


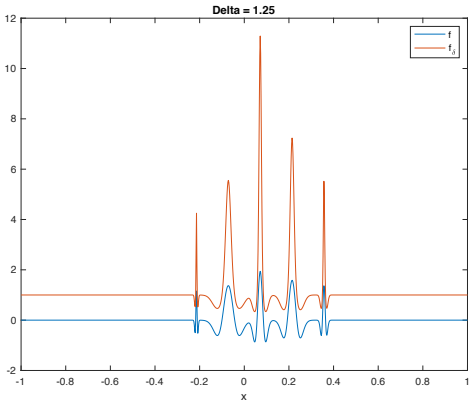


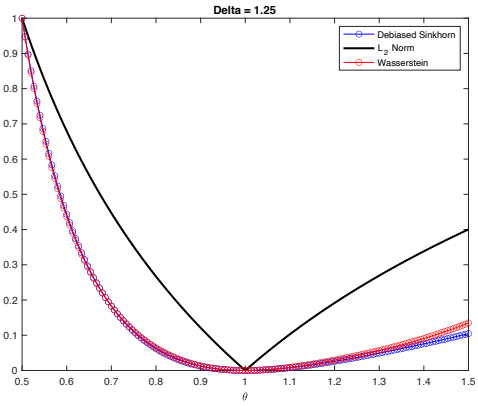

Delta = 0.75
f
f_δ
x


Delta = 0.75
Debiased Sinkhorn
L_2 Norm
Wasserstein
θ


Delta = 1
f
f_δ
x


Delta = 1
Debiased Sinkhorn
L_2 Norm
Wasserstein
θ


Delta = 1.25
f
f_δ
x


Delta = 1.25
Debiased Sinkhorn
L_2 Norm
Wasserstein
θ

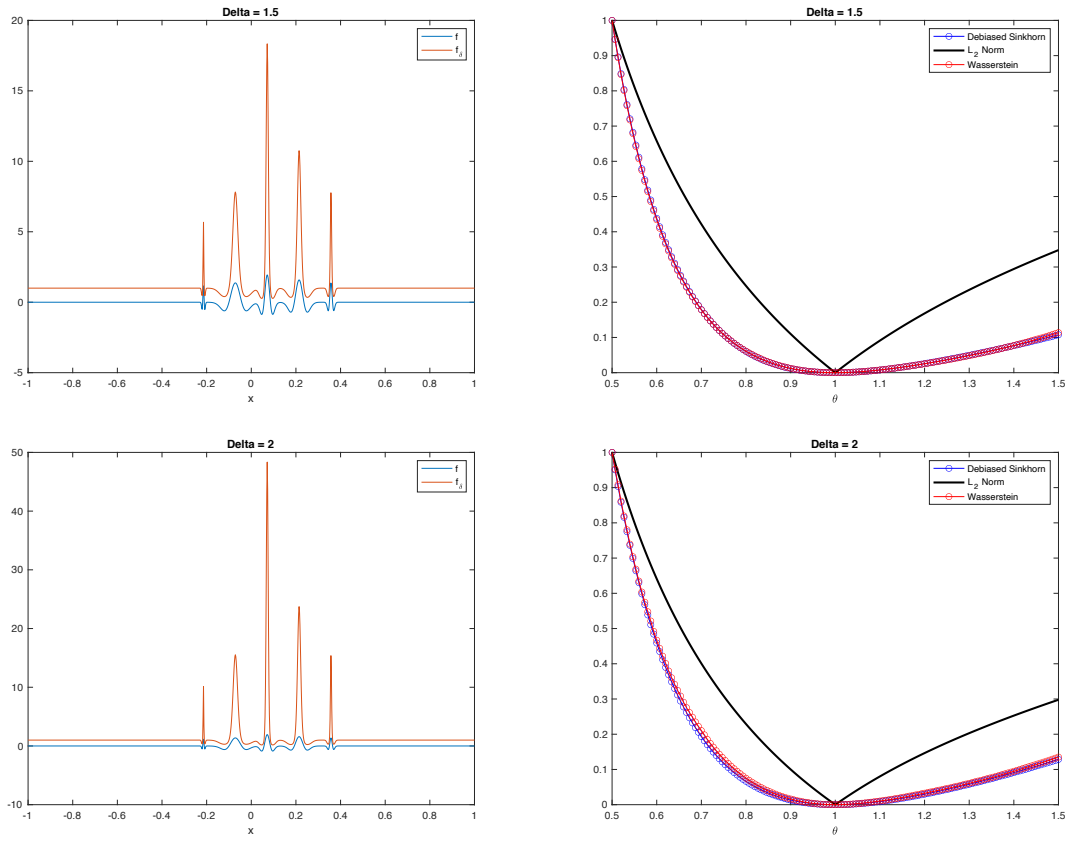
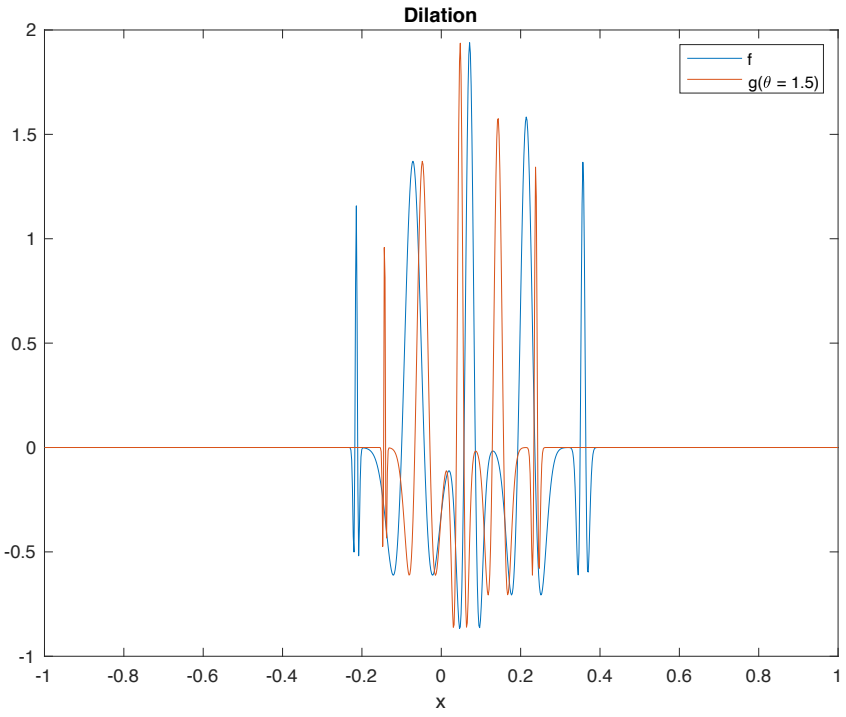


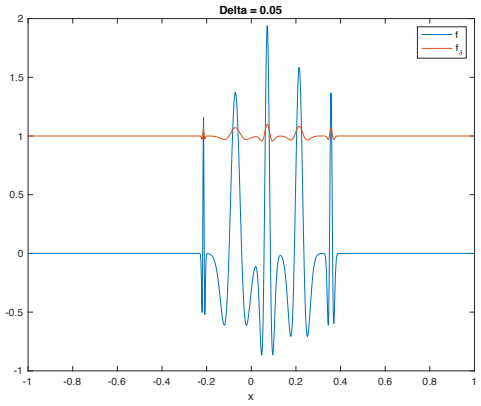
Figure 12: The figures on the right compare the normalized Debiased Sinkhorn divergence (blue) to the normalized Wasserstein distance (red) and the L_2 norm (black) for our high frequency wave for increasing values of delta when considering θ as a frequency parameter (27). The figures on the left display the effect that increasing values of delta have on our exponential scaling $\mathbf{f}_\delta$ (orange) in comparison to the original signal $\mathbf{f}$ (blue.)

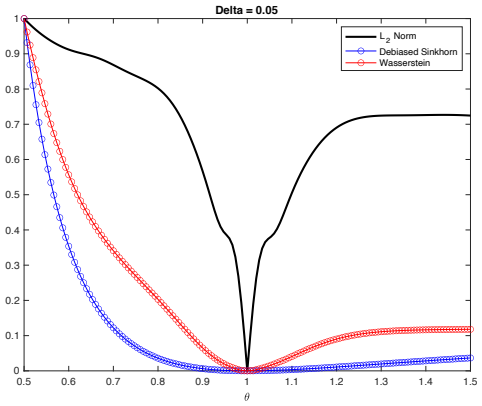
Again, we see that in the case of changing the frequency of $\mathbf{f}$, all three distance functions perform similarly for various values of δ. There is no major advantage to using normalized DS divergence over the other measures in this case with regards to convexity.

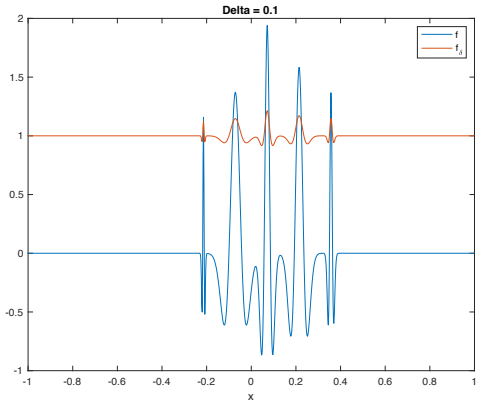
Example 5 (δ Study for Dilation in High Frequency Waves)

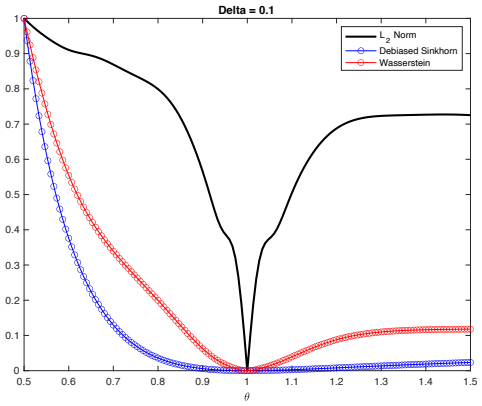
In this example we want to explore how our different dissimilarity measures are affected when we dilate our signal $\mathbf{f}$. Thus, let $\mathbf{g}(\theta) \coloneqq \mathbf{g}_d(x;\theta)$ from (28). A graph of $\mathbf{f}$ and $\mathbf{g}$ is shown below, with $\theta = 1.5$

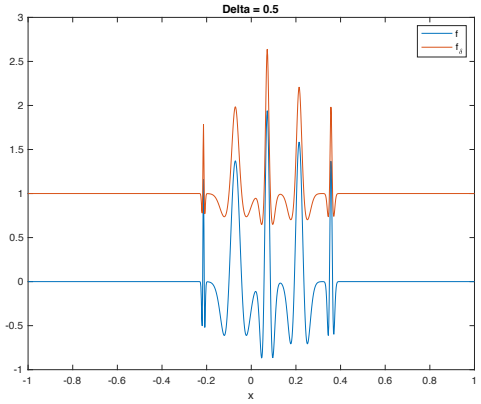

Dilation
f
g(θ = 1.5)
x

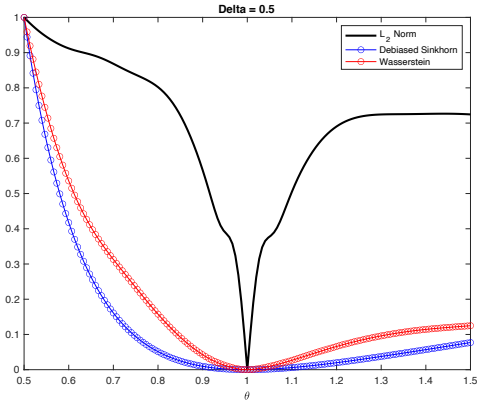

Delta = 0.05
f
f_1
x

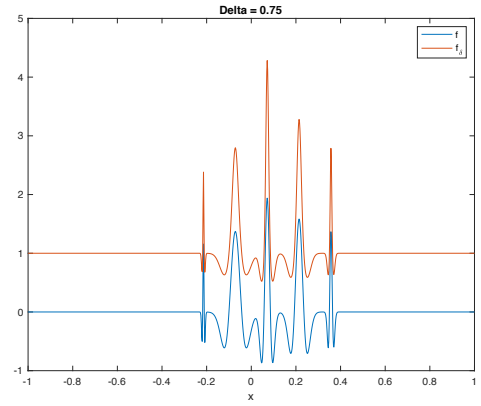

Delta = 0.05
L_2 Norm
Debiased Sinkhorn
Wasserstein
θ

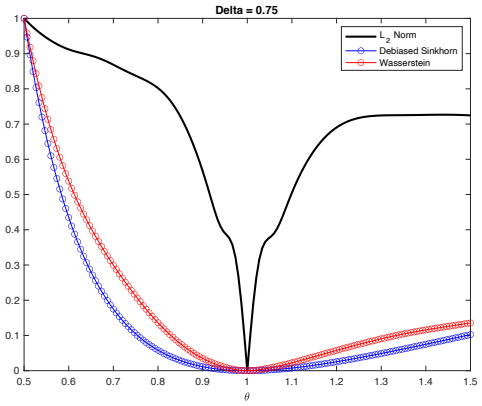

Delta = 0.1
f
f_δ
x


Delta = 0.1
L_2 Norm
Debiased Sinkhorn
Wasserstein
θ


Delta = 0.5
f
f_δ
x


Delta = 0.5
L_2 Norm
Debiased Sinkhorn
Wasserstein
θ


Delta = 0.75
f
f_δ
x


Delta = 0.75
L_2 Norm
Debiased Sinkhorn
Wasserstein
θ

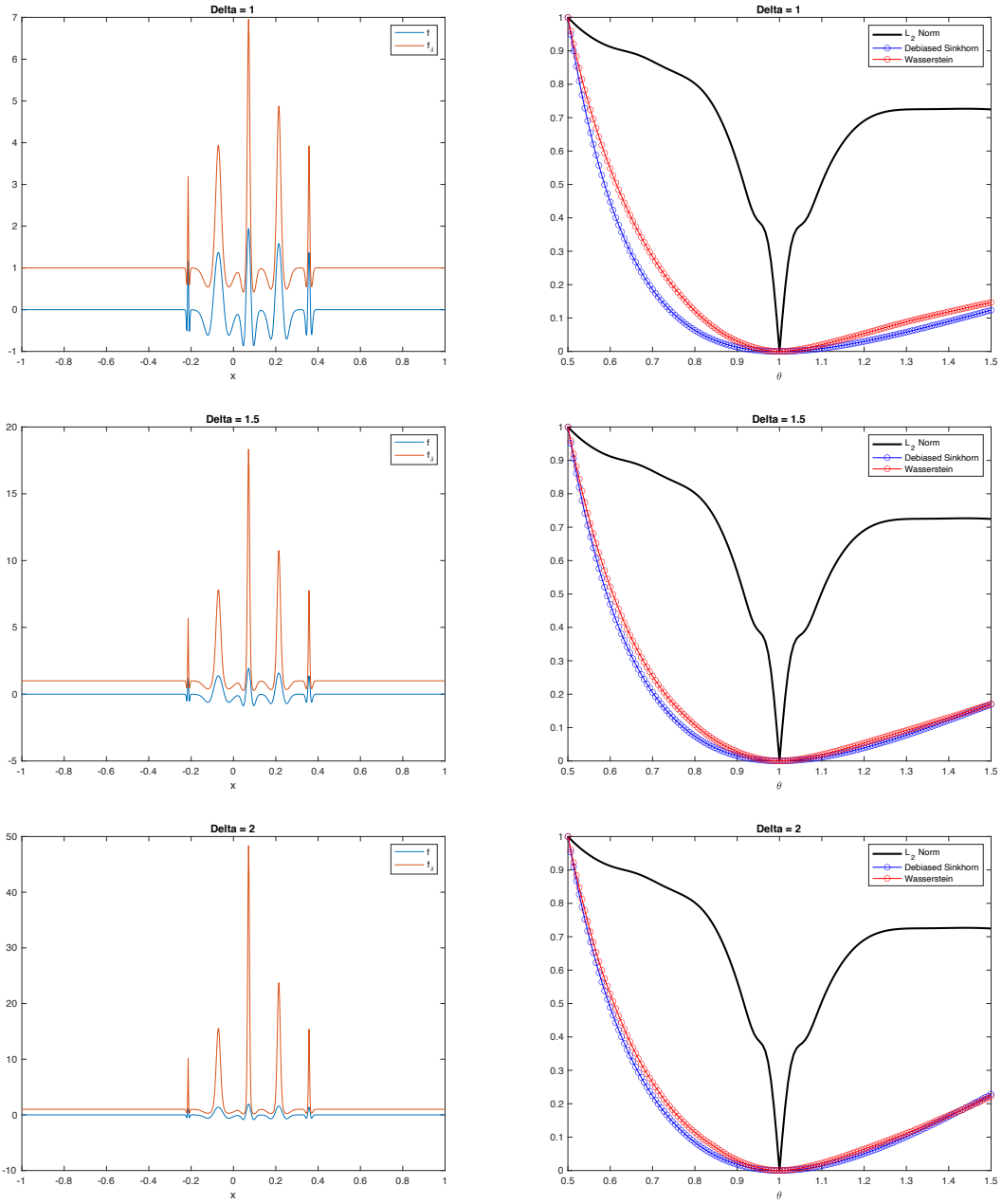


Figure 13: The figures on the right compare the normalized Debiased Sinkhorn divergence (blue) to the normalized Wasserstein distance (red) and the L_2 norm (black) for our high frequency wave for increasing values of δ when considering θ as a dilation parameter (28). The figures on the left display the effect that increasing values of δ have on our exponential scaling $\mathbf{f}_\delta$ (orange) in comparison to the original signal $\mathbf{f}$ (blue.)

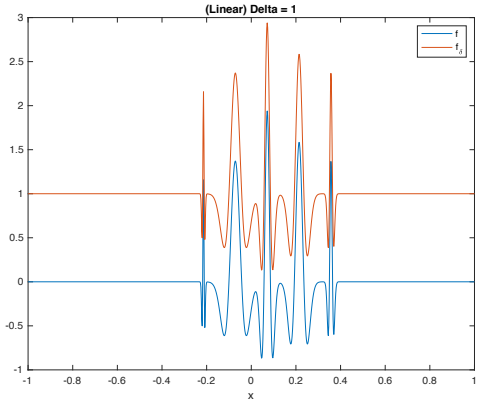
In this case we do see a small advantage in using normalized DS divergence as compared to the normalized Wasserstein distance, and there is certainly an advantage in using the previous measures over the classical L_2 norm, which fails to maintain convexity as θ approaches 0.

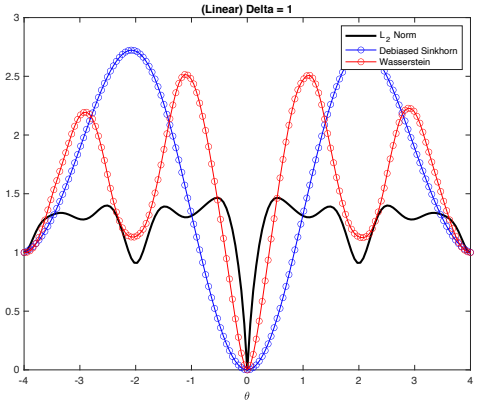
4.4 Linear and Softplus Normalization on Ricker Wavelets

Recall that in (20) we used exponentiation in order to ensure that our signals $\mathbf{f}$ and $\mathbf{g}$ were entirely positive, i.e. $\mathbf{f}, \mathbf{g} \in \mathbb{R}_+^n$ (which is needed to perform Debiased Sinkhorn divergence as well as Wasserstein distance). In this example, we continue to use our high frequency waves Figure 6 and the θ parameters presented in Example 1, i.e. θ is a phase shift (25), and we explore two other ways of ensuring that $\mathbf{f}$ and $\mathbf{g}$ are shifted in the positive direction. Specifically, we are interested in exploring the methods of using linear scaling (21) and softplus scaling (22) using the normalization described in (19).

Linear Scaling:

Once again, we compare the performances of normalized Debiased Sinkhorn divergence, the L_2 norm, and the normalized Wasserstein distance for different values of δ. Note that for this linear scaling, δ needs to be sufficiently large in order to ensure that $\mathbf{f}$ and $\mathbf{g}$ are entirely positive signals. Also note that in the linear case, the difference between $\mathbf{f}$ and $\mathbf{f}_\delta$ (and by the same token $\mathbf{g}$ and $\mathbf{g}_\delta$) is less sensitive to large delta, whereas in the exponential case even $\delta = 2$ created an extreme loss of information about the original signal.





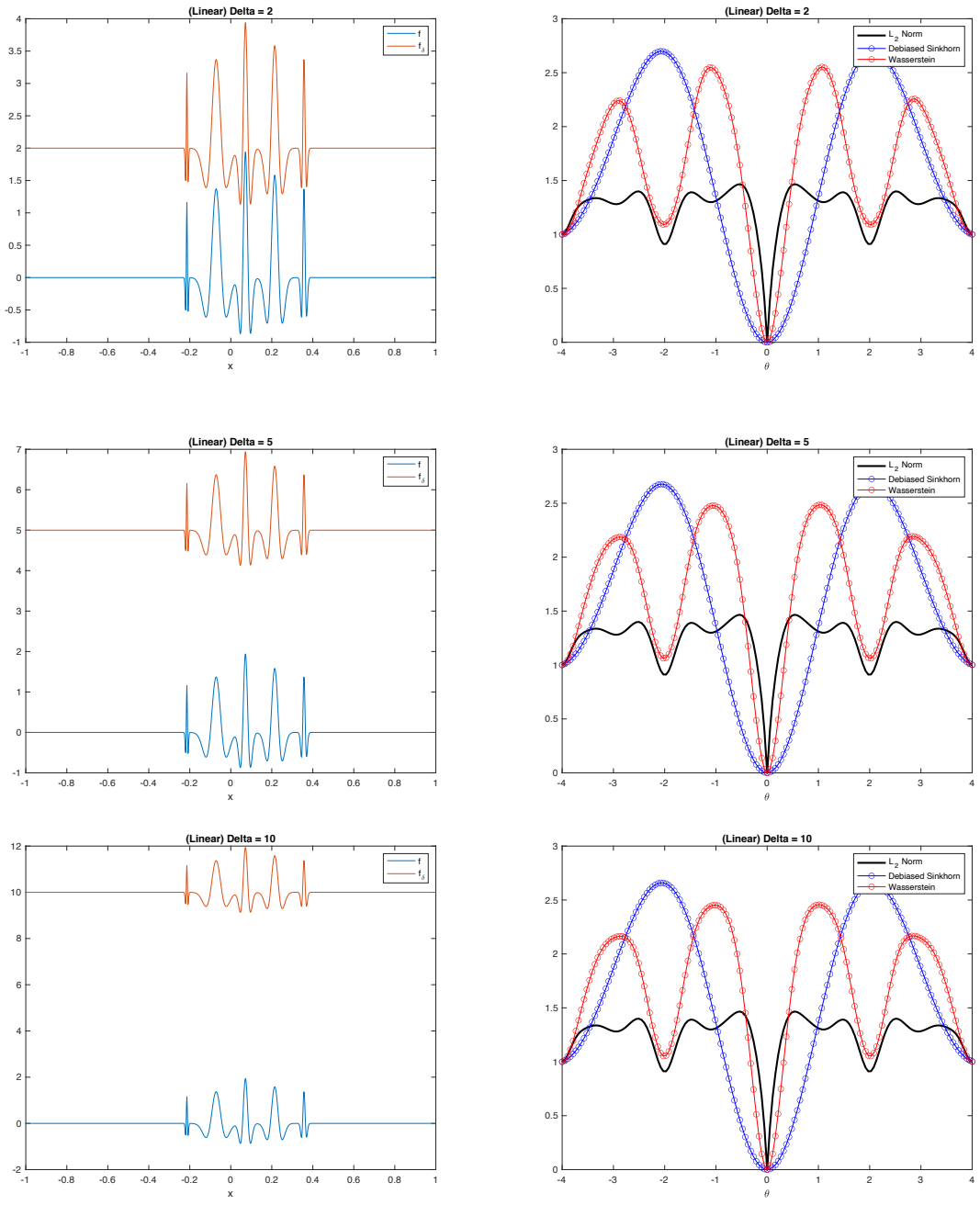


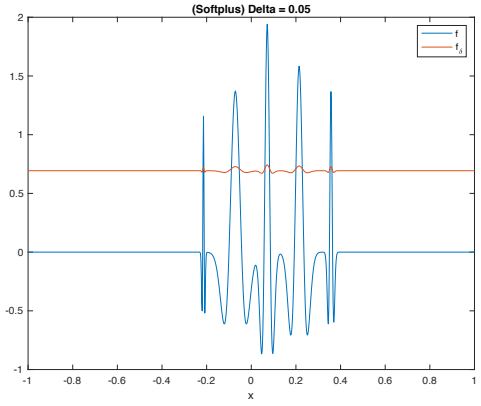
Figure 14: The figures on the right compare the normalized Debiased Sinkhorn divergence (blue) to the normalized Wasserstein distance (red) and the L_2 norm (black) for our high frequency wave for increasing values of δ when considering θ as a phase shift (25). The figures on the left display the effect that increasing values of δ have on our linear scaling $\mathbf{f}_\delta$ (orange) in comparison to the original signal $\mathbf{f}$ (blue.)

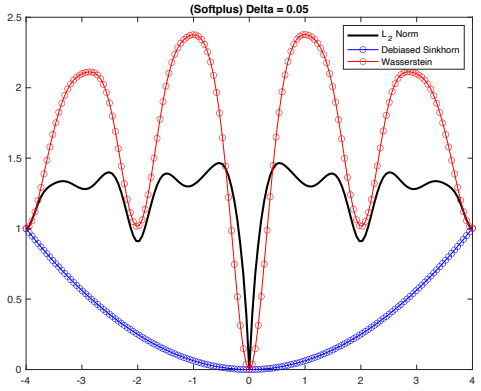
We observe that, while linear normalization is less disruptive to the original signal, for high frequency waves all three of our distance measures fail to achieve convexity regardless of the

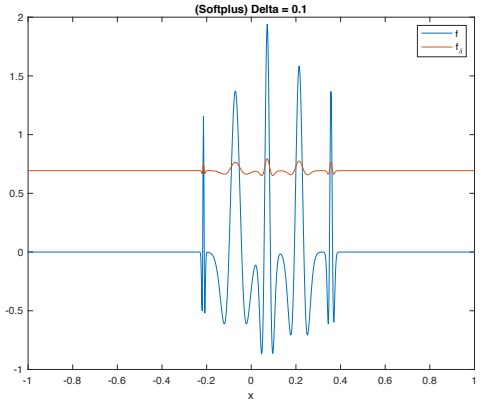
value of δ. In the case of having high frequency signals as inputs, using linear normalization appears to be an undesirable method.

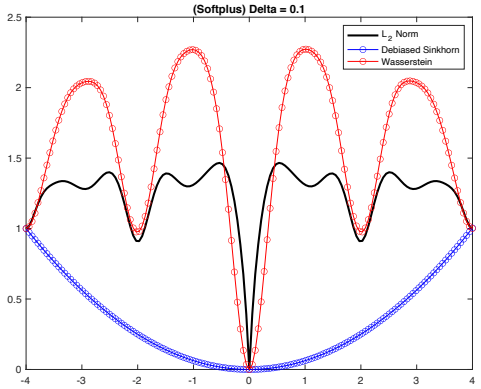
Softplus Scaling:

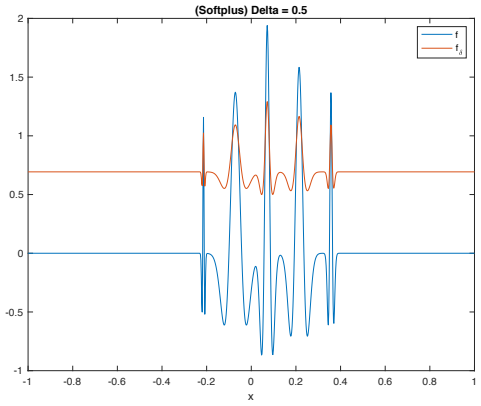
Softplus scaling (22) is a newer method of forcing our signals to become entirely positive. Due to its form, we do not have issues with setting δ to be a small value. Thus, we begin with $\delta = 0.05$ as we did with our investigations with exponential normalization and study how the convexity of normalized DS divergence, normalized Wasserstein distance, and the L_2 norm are affected by increasing values of δ. Note that we again face an issue with setting δ too small or too large, as information about the original signal becomes corrupted as it did in the case of exponential normalization (20); however, in the case of δ large, the disruption is not as robust when using softplus scaling. Below are the results.

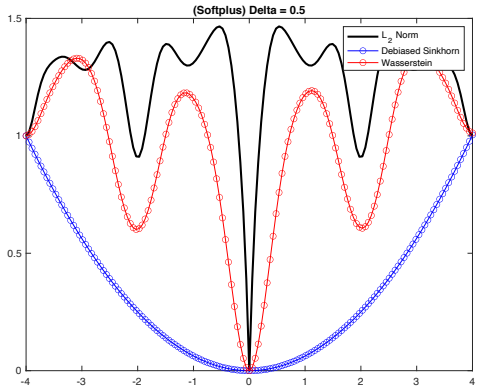


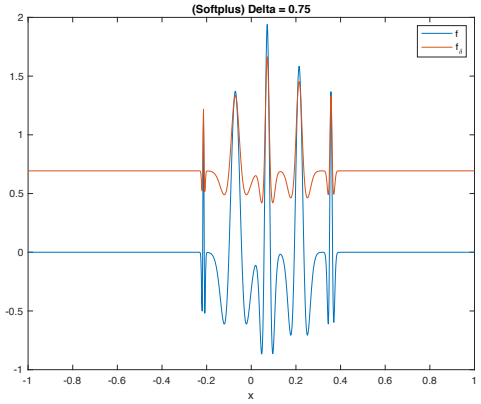


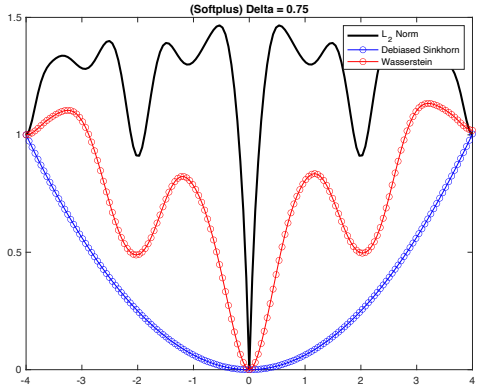


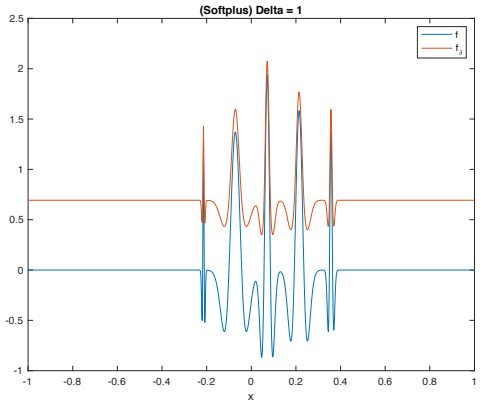


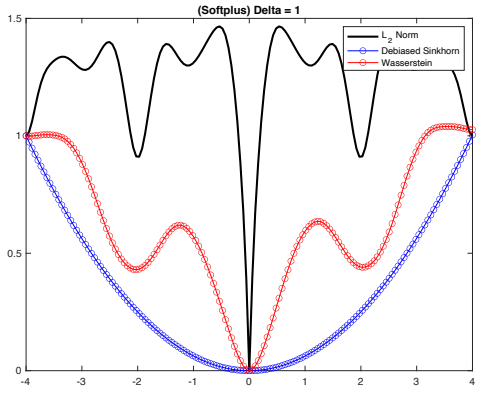

(Softplus) Delta = 0.5
f
f_δ
x

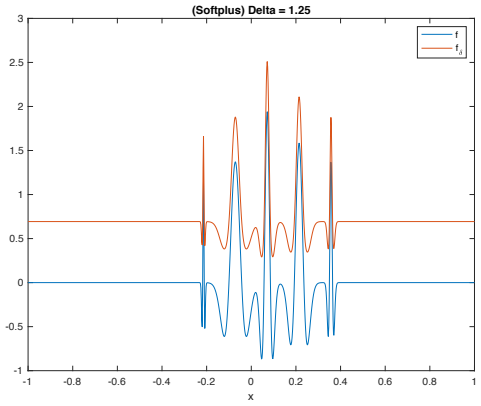

(Softplus) Delta = 0.5
L_2 Norm
Debiased Sinkhorn
Wasserstein

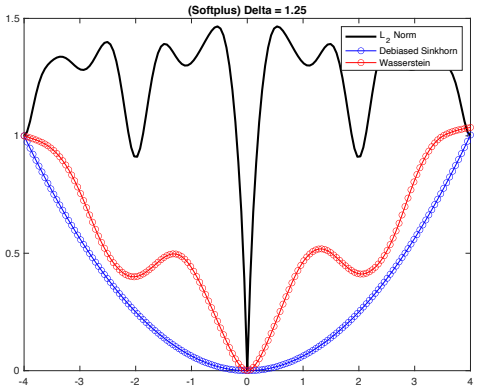

(Softplus) Delta = 0.75
f
f_δ
x

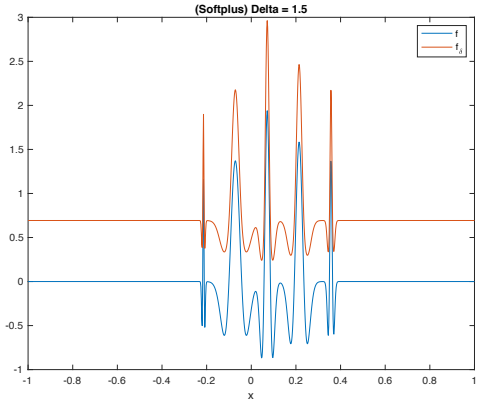

(Softplus) Delta = 0.75
L_2 Norm
Debiased Sinkhorn
Wasserstein

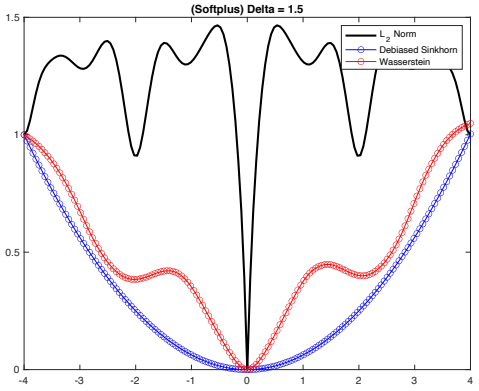

(Softplus) Delta = 1
f
f_δ
x

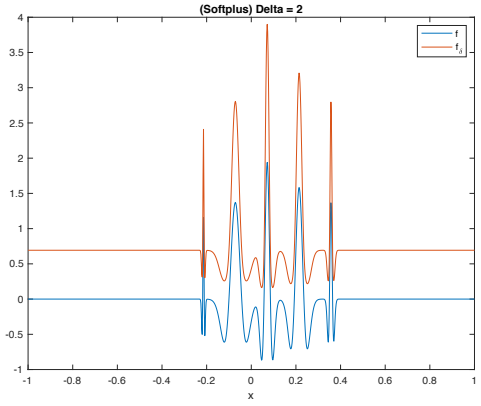

(Softplus) Delta = 1
L_2 Norm
Debiased Sinkhorn
Wasserstein

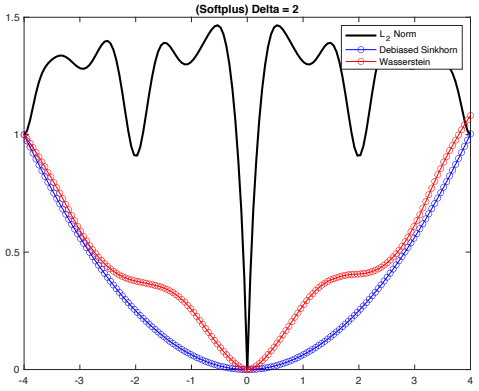

(Softplus) Delta = 1.25
f
f_Δ
x


(Softplus) Delta = 1.25
L_2 Norm
Debiased Sinkhorn
Wasserstein


(Softplus) Delta = 1.5
f
f_Δ
x


(Softplus) Delta = 1.5
L_2 Norm
Debiased Sinkhorn
Wasserstein


(Softplus) Delta = 2
f
f_Δ
x


(Softplus) Delta = 2
L_2 Norm
Debiased Sinkhorn
Wasserstein

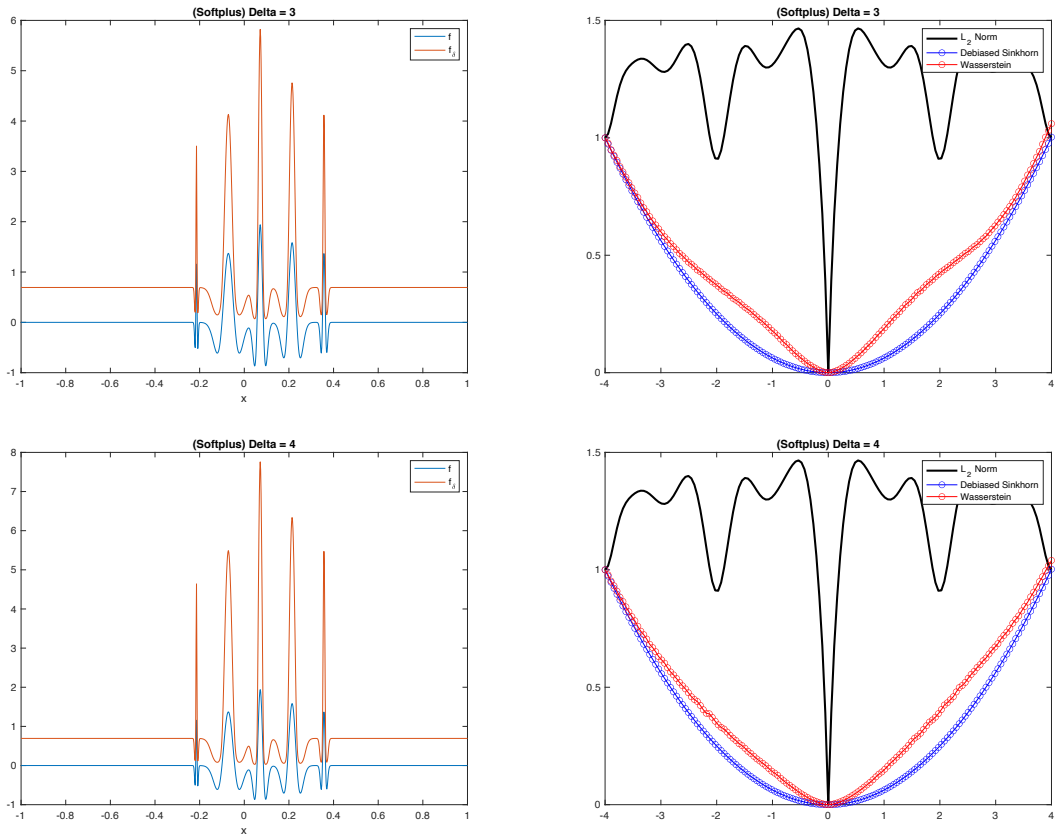


Figure 15: The figures on the right compare the normalized Debiased Sinkhorn divergence (blue) to the normalized Wasserstein distance (red) and the L_2 norm (black) for our high frequency wave for increasing values of δ when considering θ as a phase shift (25). The figures on the left display the effect that increasing values of δ have on our softplus scaling $\mathbf{f}_\delta$ (orange) in comparison to the original signal $\mathbf{f}$ (blue.)

The results above robustly show the advantages of using normalized DS divergence as a measure over the normalized Wasserstein distance and the L_2 norm when using softplus scaling as a means of normalization. Regardless of the value of δ, normalized DS divergence maintained convexity. In contrast, the normalized Wasserstein distance demanded a δ value significantly larger than 1 to achieve strict convexity. However, in the case of softplus scaling, having a large δ does not disrupt the original signal as much as in the case of exponential scaling.

Section 5

Conclusion

The normalized Debiased Sinkhorn divergence offers many advantages over the commonly used normalized Wasserstein distance and the classical Euclidean norm. In the case of very high frequency signal inputs, normalized DS divergence requires far less extreme data normalization to achieve convexity. In Section 3, where we considered high frequency waves as inputs and used exponential scaling as a means of data normalization, we observed some instances where normalized DS divergence achieved convexity for significantly less extreme values of δ than was needed for the normalized Wasserstein distance, which as mentioned in the introduction leads to loss of information, noise amplification, and can lead to machine overflow. In the case where the two high frequency signals differed by a phase shift, this advantage was abundantly clear. In the cases where the difference between the two signals was a matter of amplitude, frequency, or dilation the advantage of using normalized Debiased Sinkhorn divergence was not as clear in terms of convexity. However, normalized DS divergence is still a less computationally costly method compared to the normalized Wasserstein distance and maintained convexity for the same values of δ. In Section 4 we considered using linear and softplus scaling as a means of data normalization as opposed to exponential scaling. We continued to analyze two signals that differed by a phase shift, as it produced the most robust results in Section 3. When using linear scaling, we observed that, while linear scaling poses less issues when it comes to corrupting our original data, we were unable to achieve convexity with any of our measures regardless of the value of δ chosen. In the case of softplus scaling, normalized DS divergence proved to be desirable over the normalized Wasserstein distance. In this case a large value of δ was needed for normalized Wasserstein to achieve convexity; however, softplus scaling is not as sensitive to large δ values as exponential scaling. Normalized DS divergence maintained convexity regardless of the value of δ when using softplus scaling. Thus, when considering high frequency signals as inputs, using softplus scaling to normalize the data and using Debiased Sinkhorn as a measure proves to be the most optimal choice when considering these kinds of dissimilarity problems.

Section 6

Future Work

In Section 4, where we analyzed the effects of using linear and softplus scaling as a means for data normalization, we only considered a phase shift as the difference between the two signals because that seemed to produce the most compelling evidence in Section 3. Further research could be done in how differences in amplitude, frequency, and dilation affect normalized DS divergence and the normalized Wasserstein distance when considering linear and softplus scaling.

Finally, the inputs given to the normalized Debiased Sinkhorn divergence and Wasserstein distance throughout this paper were all one-dimensional probability vectors. Further research could be done comparing these two methods when given two dimensional inputs (such as 2D images) or even higher dimensional inputs (e.g., in the case of FMRI images).

References

[1] M. Motamed. A Hierarchically Low-Rank Optimal Transport Dissimilarity Measure for Structured Data. *Bit Numer Math*, https://doi.org/10.1007/s10543-022-00937-9, 2022

[2] Engquist and Yang. Optimal Transport Based Seismic Inversion: Beyond Cycle Skipping. https://arxiv.org/pdf/2002.00031.pdf, 2021.

[3] Yang. Analysis and Application of Optimal Transport For Challenging Seismic Inverse Problems. https://arxiv.org/pdf/1902.01226.pdf, 2019

[4] B. Engquist and Y. Yang. Seismic Imaging and Optimal Transport. *Communications in Information and Systems*. https://arxiv.org/pdf/1808.04801.pdf, 2018

[5] Engquist and Yang. Seismic Inversion and the Data Normalization for Optimal Transport. https://arxiv.org/pdf/1810.08686.pdf, 2018

[6] Engquist and Froese. Application of the wasserstein metric to seismic signals. *Communications in Mathematical Sciences*, 12(5):979–988, 2014.

[7] Y. Yang, B. Engquist. Analysis of optimal transport and related misfit functions in full-waveform inversion. *GEOPHYSICS* 83: A7-A12, 2018.

[8] H. Janati, M. Cuturi, A. Gramfort. Debiased Sinkhorn Baycenters. *Proceedings of the 37th International Conference on Machine Learning*, PMLR 119:4692-4701, 2020.

[9] J. Feydy, T. Séjourné, F. Vialard, S. Amari, A. Trouvé, G. Peyré. Interpolating between Optimal Transport and MMD using Sinkhorn Divergences. https://arxiv.org/pdf/1810.08278.pdf, 2018

[10] R. Sinkhorn. A relationship between arbitrary positive matrices and doubly stochastic matrices. Annals of Mathematical Statistics, 35:876–879, 1964.

[11] G. Peyré and M. Cuturi. Computational optimal transport. Foundations and Trends in Machine Learning, 11:355–607, 2019.

[12] Cuturi, M.: Sinkhorn distances: lightspeed computation of optimal transport. Adv. Neural Inf. Process. Syst. 26, 2292–2300 (2013)