Thesis Outline

Working Title: Sinkhorn Divergence Driven Bayesian Inversion

**Abstract:** This paper will outline a Sinkhorn Divergence driven Bayesian inversion framework. Conventionally a Gaussian Driven Bayesian framework is used when preforming Bayesian inversion. A major issue with this Gaussian framework is that the Gaussian likelihood, driven by the L2 norm, is not affected by phase shift in a given signal. This is a very important issue that is not addressed using the Gaussian Likelihood but has been addressed using a Wasserstein driven Bayesian framework. However, the Wasserstein Framework still has an issue because it assumes statistical independence when multidimensional signals are analyzed. This assumption of statistical independence cannot always be made when analyzing signals where multiple detectors are recording one event, say from seismic event. The Wasserstein metric can be generalized to multidimensional signals, getting rid of the assumption of statistical independence of multiple signals. However, implementation of this multidimensional Wasserstein metric is very computationally expensive, making it unreasonable for Bayesian inversion. Sinkhorn Divergence offers an approximation to the multidimensional Wasserstein metric while remaining relatively cheap computationally. This allows for the creation of a Sinkhorn Divergence driven Bayesian framework that will be formulated and analyzed in this paper.

**Introduction:** In most applications of Bayesian inversion, a Gaussian likelihood function is used to formulate a Bayesian inversion framework. This is often because the Gaussian, or sometimes called Normal, likelihood is the most common likelihood that a Bayesian framework is built on. The Gaussian likelihood implements the L2 norm. This norm is a standard choice in certain specific applications but fails to account for phase differences in a given signal. Because of this, the Gaussian likelihood can produce a false minimum that a Bayesian Inversion Algorithm can become trapped in, producing an incorrect posterior distribution. Through the work in [1] we can see that a way to avoid this issue all together is to use a different likelihood, namely a Quadratic Wasserstein likelihood. This likelihood has one major advantage over the Gaussian likelihood. This advantage is that the Wasserstein metric not only measures the difference in amplitude of two signals, but also the distance in phase. This means that depending on the application, the Wasserstein Likelihood can produce a better posterior compared to the Gaussian Likelihood. This idea is explored in detail in [1]. The Wasserstein metric is used to create the Wasserstein Likelihood and in [1], only the one-dimensional Wasserstein metric is used. This is because implementation of a multidimensional Wasserstein metric would be very computationally expensive. So, instead of implementing the multidimensional Wasserstein metric, we typically use the one-dimensional Wasserstein metric and make some assumptions on the signal being analyzed. The main assumption is that the signals being analyzed are statistically independent. This assumption is made so that when a multidimensional signal is analyzed, we can assume that the likelihood of the signal is a product of the likelihoods of the one-dimensional signals. This assumption allows the one-dimensional Wasserstein Likelihood to be applied to multidimensional signals. However, this assumption of statistical independence cannot always be made for a given set of signals. This is where the Sinkhorn Divergence has a major advantage. Sinkhorn Divergence can be viewed as an approximation to the multidimensional Wasserstein metric [2], allowing for the creation of a Sinkhorn Divergence driven Bayesian framework. This Sinkhorn Framework has the advantage that it does not require the assumption of statistical independence to work. This is because Sinkhorn Divergence acts as an approximation to the multidimensional Wasserstein metric, eliminating the need for statistical independence assumptions. Sinkhorn has an advantage over the multidimensional Wasserstein metric in that it is far less computationally costly. Because the Sinkhorn approximation costs less than the multidimensional Wasserstein metric, it is a viable option in a Bayesian inversion framework. And, because the Sinkhorn Divergence does not need the statistical independence assumption, it can be a better choice when analyzing a multidimensional signal compared to the Wasserstein metric. We create a Sinkhorn Divergence Bayesian framework by modifying the Wasserstein framework from [1], changing it so that no assumptions are made about statistical independence.

The rest of the paper will explore this new Sinkhorn Divergence Bayesian framework and compare it to both the Gaussian and Wasserstein frameworks. Section 2 will outline the general Bayesian Inversion Problem. Section 3 will show the new likelihood structure of the Sinkhorn Divergence likelihood. Section 4 will show the formulation of a Markov Chain Monte Carlo algorithm that implements the new Sinkhorn Divergence likelihood. And section 5 will show numerical examples that implement this new Sinkhorn Divergence driven Bayesian inversion.

**Bayesian Inversion:**

In many problems in stochastic processes, we are tasked with finding the conditional probability of the model parameter given an observed quantity. This is done through Bayes’ Theorem which allows us to calculate the conditional probability of an event occurring. In Bayesian inversion, we are using Bayes’ Theorem to calculate the conditional probability of the model parameter vector $\_theta\_ Contained in\_THETA\_Subset\_R^m$. Let $g=(g1,…,gn)\_contained in\_R^n$ be vector of $n$ observed quantities and let $f=(f1,…,fn)\_contained in\_R^n$ be a vector of $n$ predicted quantities created by a forward predictive model depending on the parameter vector $theta$. That is to say that:

$f=f(theta): THETA->R^n

Now, applying Bayes’ Theorem to solve for the conditional probability of $theta$ given $g$, written as $pi(theta|g)$ we have:

$BAYES THEROEM$

Where $pi(theta|g)$ is the posterior, $pi(g|theta)$ is the likelihood, and $pi(theta)$ is the prior distribution of $theta$. $The denominator of BAYES THEROEM$ is sometimes written as $pi(g)$ which through some calculation can be shown to be the same as $The denominator of BAYES THEROEM$. Since $pi(g)$ is independent of $theta$, it can be viewed as a scaling constant to ensure that the posterior obtained from Bayes’ Theorem is consistent with the definition of a probability density function (i.e. integrates to one and is positive for $\_any\_theta \_contained in\_ THETA$). This means that $BAYES THEROM \_is proportional to\_ pi(g|theta)pi(theta)$.

This now leaves us with the task of finding distributions for the prior and likelihood. The prior, pi(theta), is a distribution that we obtain from prior knowledge about the model parameter $theta$. As an example, if we know that $theta \_is contained in\_ (1,3)$ then one choice for a prior could be $theta~uniform(1,3)$.

*Likelihood structure:*

The choice of likelihood is a fundamental step in Bayesian Inversion and is one of the main aspects of Bayesian Inversion that can be changed. Often, the choice of a likelihood is based on the noise structure that the analyzed signal is expected to have. Under the assumption that we have a simple additive noise structure, the Gaussian likelihood is the most common choice. Assuming that we have measurement noise ${eps\_1,…,eps\_n}$ that appears in the measured quantities ${g1,…,gn}$ the additive noise is assumed to be normally distributed with mean zero and standard deviation $sigma$. That is to say:

$g\_i=f\_i(theta)+eps\_i$, $eps\_i~normal(0, sigma)$, $i=1,…,n$

Where n is the number of measured quantities. This noise structure can be handled well by the Gaussian likelihood:

$GAUSIAN LIKLIHOOD$

Such a likelihood structure also requires an assumption of statistical independence for the signals being analyzed. This is seen in the fact that the overall likelihood, $pi(g|theta)$ is equal to the product of the individual likelihood of $g\_i$:

$L(theta):=pi(g|theta)=\_Pi\_pi(gi|theta)$ (IN PAPER REFRENCE 1)

This illuminates two major issues with the Gaussian Likelihood. One issue is that the noise structure may not be realistic because of its simplicity. For example, if we have a set of two dimensional signals $g(x1,x2)$ with an additive Gaussian noise structure

$g(x1\_i,x2\_j)=f(x1\_i,x2\_j;theta)+eps\_ij$

We can easily show that the Gaussian Likelihood fails to predict the correct values of theta for specific problems that contain this noise structure (see section on numerical experiments for further illustration).

Another issue with the Gaussian likelihood is that the assumption of statistical independence may not be mathematically consistent, again depending on the application. We have Statistical independence when the probability of two events occurring is equal to the product of the individual probabilities of each event occurring independently:

$P(A\_and\_B)=P(A)P(B)$

Another way to say this is that statistical independence exists when the occurrence of one event does not affect the probability of the other event occurring. Many applications in stochastic processes do not have statistical independence, leading to a desire to create a Bayesian framework that does not assume statistical independence and that can handle more complicated noise structures.

**Concepts from Optimal Transport:**

In this section, two metrics from optimal transport will be explored. These metrics are the key to creating a new Bayesian Inversion framework that satisfies the need to have a framework that does not assume statistical independence and handles complicated noise structure well. First, the quadratic Wasserstein metric will be explored. Then, the Debiased Sinkhorn Divergence (DSD) will be explored and will be shown to approximate the Wasserstein metric. These metrics will then be used in the next section to create likelihood functions that can be used in Bayesian inversion.

One of the main desirable traits for a likelihood function used in Bayesian Inversion is convexity for the type of problems the framework is applied to. This is because a convex function has a more well-defined minimum value compared to a non-convex function, allowing the Bayesian inversion algorithm to converge to the correct minimum and not fall into a false minimum. The likelihood functions in this paper that employ the Wasserstein metric and the DSD have this convexity property with respect to the phase shift, phase dilation, and amplitude change in the simulated and measured signals [1]. This is the motivation behind using the Wasserstein metric and DSD in a likelihood function.

*Wasserstein Metric:*

The Wasserstein metric is a distance function defined by the minimization of the cost of turning one probability distribution into the other [1]. Suppose we have two discrete time signals $f,g\_contained in\_R^N$ with discrete time steps {t1,…,tN}. The two signals $f and g$ need some preliminary altering before they can be implemented into the Wasserstein metric. Since the Wasserstein metric is a measure of the distance between two probability distributions, we need to alter the signals since they are likely not probability distributions. We need to ensure that the signals are always non-negative and that the $sum(f\_i)=1$ and $sum(g\_i)=1$ to remain consistent with the definition of a probability mass function (PMF). This can be done several different ways, but the one that will be used in this paper is to shift the signals by some constant, and then normalize.

First, choose a constant value $c$ such that $c>min(g,f)$. This ensures that $f\_i+c>0$ and $g\_i+c>0$ for all $i=1,…,N$. Next, normalize the two signals which creates two new signals that are now in the form of a probability distribution:

$f\_bar=(f\_i+c)/sum(f\_i+c)$ and $g\_bar=(g\_i+c)/sum(g\_i+c)$

Now we have two PMFs which can be used to create two discrete cumulative density functions (CDFs).

$F\_i=sum(f\_bar\_k)$, $G\_i=sum(f\_bar\_k)$, $i=1,..,N$

where $f\_bar\_k$ and $g\_bar\_k$ are the $kth$ component of their respective functions. This allows us to define the discrete quadratic Wasserstein metric between two signals, $f and g$:

$Wasserstein Metric$

Where $T=G^-1 \_of\_ F$ is the optimal map from $f\_bar$ to $g\_bar$. Note that this is a formulation for the single dimensional Wasserstein metric. Applications for of this single dimensional metric in multidimensional problems will be addressed in the next section.

***%Here needs to be info on Sinkhorn divergence and DSD. Also, the relationship %between DSD and the Wasserstein metric.***

**Optimal Transport Based Bayesian Inversion:**

We can now create an exponential likelihood function based on the DSD. We will first look at the likelihood function for the one-dimensional Wasserstein metric, and derive a new likelihood function for the DSD based on the Wasserstein likelihood function. The Wasserstein likelihood is as follows:

$L\_wass(theta)=pi(g|theta)=s^N\*exp(-s\*dw(f(theta),g))$

where s is a hyperparameter that will be found through the Markov Chain Monte Carlo (MCMC) sampling algorithm discussed later in this paper. Note that the one-dimensional Wasserstein likelihood can be applied to multidimensional problems if statistical independence is assumed. This is the reason that we have the term $s^N$ in the Wasserstein likelihood, since this is considering the product of other one-dimensional exponential likelihood. As discussed in the previous section, the DSD approximates the multidimensional Wasserstein metric, meaning that we do not need the product of multiple likelihoods in order to analyze multidimensional problems. This ultimately simplifies the exponential likelihood, thus creating an exponential likelihood with the DSD:

$L\_DSD=pi(g|theta)=s\*exp(-s\*DSD(f(theta),g))$

where s is again a hyperparameter that will be found through the MCMC sampling algorithm discussed later in this paper. Note here that $s$ is not raised to the power of $N$ since we are no longer looking at the product of multiple likelihoods. See equation (IN PAPER REFRENCE 1) for details.

*Convexity of DSD Likelihood:*

As it has been stated earlier in this paper, one important feature of the DSD likelihood is the convexity with respect to phase shift, phase dilation, and amplitude change. This convexity will be tested by applying the DSD likelihood to a test problem. A comparison will be made between the DSD likelihood and the Wasserstein and L2 likelihoods.

Suppose that the original signal $f$ is:

$f(t)=…$

And a shifted version of the signal (representative of noise perhaps) g is:

$g(t)=…$

Where s is the factor in which the signal is shifted.

[figure1]

Note that there are many ways to normalize the signals for the DSD likelihood and the Wasserstein likelihood. The option that has been used in this example is linear scaling. That is:

$f^hat=f+c/<f+c>$

$g^hat=g+c/<g+c>$

Where $c$ is some constant chosen to ensure that both $f+c>0$ and $g+c>0$. The signals are then normalized to ensure that they can now be viewed as probability distributions.

As seen in [figure1], the Wasserstein likelihood and DSD likelihood both show convexity for this specific example. The L2 likelihood produced many minima which, in a Bayesian Inversion application, could produce an incorrect posterior. This example shows the advantage of using the DSD likelihood over the L2 likelihood for inversion problems that involve phase shift.

**Numerical Algorithm:**

In the section the Metropolis-Hastings-within-Gibbs (MH within G) sampling algorithm along with the DSD likelihood to create a numerical algorithm for the DSD Bayesian framework. MH within G is a Markov chain Monte Carlo method that updates probabilities based on a selection process (see Algorithm 1). Looking at the posterior in the Bayesian Inversion, note that:

$pi(theta|g) \_proportional to\_ pi(g|theta)pi(theta)$

This means that for a Bayesian algorithm we need a likelihood and prior. This is where we decide to use the DSD likelihood and employ a know algorithm like MH within G.

*Gibbs Sampler:*