Approximate Bayesian Computation with the Sliced-Wasserstein Distance

Mathias Vigouroux*

April 5, 2024

Abstract

Approximate Bayesian Computation (ABC) is a technique for estimating posterior distributions in the presence of intractable likelihoods, leveraging synthetic data generation. Drawing on Optimal Transports, specifically the Wasserstein distance, as a means of comparing synthetic and real data. The advantages of different distances, especially for univariate data, is both analytically and numerically discussed. More computationally optimal than the Wasserstein distance, the Sliced Wasserstein, the study dive into the Sliced-Wasserstein ABC technique. A meticulous analysis of theoretical finding redemonstrates the convergence of SW-ABC posterior to the true posterior with decreasing acceptance thresholds. Empirical investigations allow to depict the robustness of the SW-ABC to difference form of covariance in the data, more precisely the decay of its eigenvalues. Empirical investigations, however, reveal no computational advantages of Sliced-Wasserstein distance over the regular Wasserstein distance, challenging theoretical expectations.

1 Introduction

Living and studying inside the École Normale Supérieure de Paris, just a few block from Place Monge, I owe to dedicate this introduction to the original problem that motivated optimal transport: the Monge Problem, which has been presented in class.

1.0.1 Historical Context: The Monge Problem

Originating from Gaspard Monge's work, the Monge problem addresses optimal mass transportation. Imagine piles of soil representing probability distributions. The goal is to minimize the work (distance) required to transform one landscape into another i.e. to displace sand from one place to the other. The aim is thus to find an optimal mapping (transport map) guiding each element in the source distribution to a specific location in the target distribution. This directly gives one important intuition about optimal transport: two grains of sand which have different initial locations can converge to the same location, however, one grain of sand cannot go to two different locations.

For a probability space Y with a distance function ρ, and two distributions μ and ν, the Monge problem seeks an optimal map $T : Y \to Y$ minimizing the cost:

$$\min_T \int_Y \rho(x, T(x))\, d\mu(x)$$

From the first glance it is possible to understand the importance of how this "work" or "distance" is calculated. Therefore, Optimal Transport is a mathematical field that really developped practical and theoretical tools to measure the distance between two distributions.

1.0.2 Generative models

Nowadays, the focus of scientist is seldomly to move pile of sands, yet, comparing densities remains of the utter importance.

*MVA, ENS-PLS: mathias.vigouroux@ens.psl.eu

One interesting application is to consider the problem of estimating the posterior distribution of some model parameters $\theta \in \mathbb{R}^{d_\theta}$ given n data points $y_{1:n} \in Y^n$. This distribution has a closed-form expression given by Bayes' theorem up to a multiplicative constant: $\pi(\theta|y_{1:n}) \propto \pi(y_{1:n}|\theta)\pi(\theta)$.

However, for many statistical models of interest, the likelihood $\pi(y_{1:n}|\theta)$ cannot be numerically evaluated in a reasonable amount of time, which prevents the application of classical likelihood-based approximate inference methods. Nevertheless, in various settings, even if the associated likelihood is numerically intractable, one can still generate synthetic data given any model parameter value. This generative setting gave rise to an alternative framework of likelihood-free inference techniques.

This is the realm of Approximate Bayesian Computation (ABC) methods. ABC methods approximate the posterior distribution by generating synthetic data from the model and retaining parameter values for which the synthetic data are close enough to the observed data.

The important question is then how to compare this generated data to the real data. The closeness is typically measured using a discrepancy measure between the two datasets, often reduced to summary statistics.However, these statistics are defined beforehand and might induce a loss of information, which has been shown to deteriorate the quality of the approximation.

Tools developed in Optimal Transports can thus be interesting to compute the distance between the synthetic data and the real data.

2 Link with the course and experimentation

2.1 Link with the course

At it has been shown previously, tools from Optimal Transports are necessary to compute the diferences between generated data and the data of interest. The first aim of this work will be to recall notations and notions from Optimal Transports that have been seen in class, the classic Wasserstein distance but also the push forward measure for instance, to introduce properly the technique used in the paper which is the Sliced-Wasserstein ABC technique. Therfore, compared to class the usage of optimal transport is not to transform one distribution into the other but rather to use it to define a threshold. This requires to use and understand the advantages and drawbacks of the Wasserstein distance versus the Sliced Wasserstein distance.

Then, this work will try to dissect meticulously one of the theoretical finding of the paper [1] which is the convergence of SW-ABC posterior to the true posterior when the threshold acceptance decrease. Even though this result was expected redoing the proof from scratch was a good way to scrap the literature and upskills with the theoretical tools of Optimal Transports that were presented in class.

2.2 Experimentation

After that, to gain a better intuition about the advantages of Wasserstein distance against others distances, a first part of the numerical part of this work tried to understand the differences between Wasserstein distances, the Kullback Leiberg divergence and the Maximum Mean Discrepancy. This first focus, will be done in low dimension, for $d = 1$ comparing two Gaussian distributions.

The paper of study [1], motivates the usage of a sliced Wasserstein distance for computational advantages compared to the regular Wasserstein distance. However, this hypotheses, given theoretically is never tested empirically, the core of my experimentation is therefore to try to assess this finding empirically. However, **I found no advantages, in computational times, to use the Sliced-Wasserstein distance compared to the regular Wasserstein Distance**. In particular, I tried to look at the impact of the covariance of higher dimensional ($d = 3$) Gaussian distributions. More precisely, the impact of the decay of the eigenvalues of the covariance matrix.

To finish with, I decided to give some direction that could link this work to interpretability of Neural Networks which is a field that I discovered during various class during the MVA and especially the Turing Seminar class.

3 Methodology

Let's now dive into the Optimal Transport framework to better grasp the tools.

3.1 General Notations

Consider a probability space $(\Omega, \mathcal{F}, P)$ with the associated expectation operator $\mathbb{E}$, on which all the random variables are defined. Let $(Y_k)_{k\in\mathbb{N}^*}$ be a sequence of independent and identically distributed (i.i.d.) random variables associated with some observations $(y_k)_{k\in\mathbb{N}^*}$ valued in $Y \subset \mathbb{R}^d$. Denote by μ^* the common distribution of $(Y_k)_{k\in\mathbb{N}^*}$, and by $\mathcal{P}(Y)$ the set of probability measures on Y. For any $n \in \mathbb{N}^*$, $\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^n \delta_{Y_i}$ denotes the empirical distribution corresponding to n observations.

Consider a statistical model $M_\Theta = \{\mu_\theta \in \mathcal{P}(Y) : \theta \in \Theta\}$ parameterized by $\Theta \subset \mathbb{R}^{d_\theta}$. We focus on parameter inference for purely generative models: for any $\theta \in \Theta$, we can draw i.i.d. samples $(Z_k)_{k\in\mathbb{N}^*} \in Y^{\mathbb{N}^*}$ from μ_θ, but the numerical evaluation of the likelihood is not possible or too expensive.

For any $m \in \mathbb{N}^*$, $\hat{\mu}_{\theta,m} = \frac{1}{m}\sum_{i=1}^m \delta_{Z_i}$ is the empirical distribution of m i.i.d. samples generated by μ_θ, $\theta \in \Theta$. We assume that:

- Y, endowed with the Euclidean distance ρ, is a Polish space,
- Θ, endowed with the distance ρ_Θ, is a Polish space,
- parameters are identifiable, i.e., $\mu_\theta = \mu_{\theta'}$ implies $\theta = \theta'$. $\mathcal{B}(Y)$ denotes the Borel σ-field of (Y, ρ).

Remark. A Polish space is a separable completely metrizable topological space; that is, a space homeomorphic to a complete metric space that has a countable dense subset.

3.2 The Wasserstein distance

Given this notation let's now dive in the core Optimal Transport topic of use for ABC.

3.2.1 The Wasserstein Distance

The Wasserstein distance, denoted as $W_p(\mu, \nu)$, is defined for $p \geq 1$ and is applicable to probability measures μ, ν in the space $P_p(Y)$, where $P_p(Y) = \{\mu \in P(Y) : \int \|y - y_0\|^p d\mu(y) < +\infty\}$ for some $y_0 \in Y$.

The Wasserstein distance of order p between two probability measures $\mu, \nu \in P^p(Y)$ is defined as:

$$W_p(\mu, \nu) = \inf_{\gamma\in\Gamma(\mu,\nu)} \left(\int \|x - y\|^p d\gamma(x, y) \right)^{1/p},$$

where $\Gamma(\mu, \nu)$ is the set of probability measures γ on $Y \times Y$ satisfying:

$$\forall A \in B(Y), \quad \gamma(A \times Y) = \mu(A), \quad \gamma(Y \times A) = \nu(A).$$

Remark. Link to the Monge Problem : This question of Wasserstein distance can be linked to the historical problem of optimal transport. Indeed, Wasserstein distance extends the Monge problem to quantify dissimilarity between distributions. Thus, The Monge problem corresponds to Wasserstein distance when the order p is 1, since the Wasserstein distance measures the minimum work needed to transform one distribution into another. Formulated as a convex optimization problem, this distance captures both the minimum cost and the transport plan geometry.

3.2.2 Computational successes and challenges

However, there are remaining **computational challenges** for this distance. Indeed, evaluating the Wasserstein distance between multi-dimensional probability measures is generally numerically intractable. Thus, solving the optimization problem defined previously, involves computational costs in $O(n^3 \log(n))$ for empirical distributions over n samples.

Yet, for **one-dimensional measures** $\mu, \nu \in P_p(\mathbb{R})$ with $p \geq 1$, the Wasserstein distance $W_p(\mu, \nu)$ has a closed-form expression:

$$W_p(\mu, \nu) = \left(\int |F^{-1}(t) - G^{-1}(t)|^p \, dt \right)^{1/p},$$

where F^{-1} and G^{-1} are the quantile functions of μ and ν respectively.

This formula allows for the direct computation of the Wasserstein distance between one-dimensional probability measures, providing an analytical expression for the distance between the quantile functions of the distributions. Indeed, for empirical one-dimensional distributions, the closed-form expression can be efficiently approximated. To do so it is necessary to sort the n samples drawn from each distribution and computing the average cost between the sorted samples, resulting in $O(n \log(n))$ operations at worst.

This **computational efficiency for one dimensional measure** motivates the definition of the sliced Wasserstein distance in the next section.

3.3 The Sliced-Wasserstein distance

The Sliced-Wasserstein (SW) distance is an alternative optimal transport (OT) distance that reduces multi-dimensional distributions to one-dimensional representations through linear projections. It is defined using linear forms associated with unit vectors on the d-dimensional unit sphere. Here is a breakdown of the key concepts and the formulation of the Sliced-Wasserstein distance:

3.3.1 Definition

Denote by $S^{d-1} = \{u \in \mathbb{R}^d : \|u\|_2 = 1\}$ the d-dimensional unit sphere. For any $u \in S^{d-1}$, $u^\star$ is the linear form associated with u, defined as $u^\star(y) = \langle u, y \rangle$ for any $y \in Y$.

For $p \geq 1$, the Sliced-Wasserstein distance of order p between two probability measures $\mu, \nu \in P_p(Y)$ is defined as:

$$SW_p(\mu, \nu) = \int_{S^{d-1}} W_p^p(u^\star_\# \mu, u^\star_\# \nu) \, d\sigma(u),$$

where σ is the uniform distribution on S^{d-1}.

Here, $W_p(u^\star_\# \mu, u^\star_\# \nu)$ represents the Wasserstein distance of order p between the one-dimensional distributions obtained by projecting μ and ν onto the linear form associated with u. More precisely, $u^\star_\# \nu$ is often called the push-forward measure, of ν by $u^\star$.

3.3.2 Push - Forward Measure

For any measurable function $f : Y \to \mathbb{R}$ and $\zeta \in P(Y)$, $f^\sharp \zeta$ is the push-forward measure of ζ by f, defined as:

$$\forall A \in \mathcal{B}(\mathbb{R}), \quad (f^\sharp \zeta)(A) = \zeta(f^{-1}(A)),$$

where $f^{-1}(A) = \{y \in Y : f(y) \in A\}$.

Intuition : the core idea in push forward measure, or image measure is that we have a space of entrance, name Y here, on which we have a measure and we do not give explicitly a measure of the space of arrival of the function f. However, we do want to measure some ensemble in the space of arrival of the function. We will therefore push forward the measure in the space of arrival by looking at the pseudo inverse of an ensemble.

Example :In the case of the u*: lets give us a vector u in Y, now lets look at i the linear form associated with u. Let's focus to one value of this linear form, for instance we can look at the number 4 there might be a few $y \in Y$ such that $< u, y >= 4$. Let's suppose having a measure on Y denoted ζ. The goal is then to measure the size of this ensemble of $y \in Y$ that gave the number 4. To do so we will push forward the measure ζ with the function.

Remark. At first sight, the adjective "sliced" does not seem to properly describe what the distance represents. It might be more appropriate to talk about a "projected Wasserstein distance. However, in Fourier mode, it does result in a slicing, since $F(u^*_\# \mu)(s) = F\mu(su)$ since the argument in the exponential of the Fourier transform is $< u, s >$. [**?**]

3.4 Comparison between the Wasserstein distance and the Sliced-Wasserstein distance

SW_p is a distance on $P_p(Y)$ with lower computational requirements compared to the Wasserstein distance. As it has been before, the Sliced-Wasserstein distance provides a computationally more efficient alternative to the Wasserstein distance by reducing the dimensionality through linear projections and averaging the one-dimensional Wasserstein distances. It is applicable to probability measures in $P^p(Y)$, and the integration is approximated for practical implementation.

Moreover, in the paper [1], the authors illustrate the fact that Sliced-Wasserstein not only provides a computationally efficient alternative but also exhibits better statistical properties compared to the Wasserstein distance and its approximations. The use of Monte Carlo approximation in the Sliced-Wasserstein integral allows for practical implementation and comparison with other methods in estimating the scaling factor of the covariance matrix. In practice, the integration in the Sliced-Wasserstein distance is approximated using a finite-sample average, often employing a simple Monte Carlo (MC) scheme

To do so, the authors will use the fact that the Wasserstein distance between two Gaussian measures has an analytical formula, given by $W_2(\mu_{\sigma^\star}, \mu_\sigma) = d(\sigma^\star - \sigma)^2$, where $\mu_{\sigma^\star}$ represents the reference Gaussian measure. This analytical element will be used to approximate the Sliced-Wasserstein distance using a Monte Carlo (MC) approximation of:

$$SW_2(\mu_{\sigma^\star}, \mu_\sigma) = W_2(\mu_{\sigma^\star}, \mu_\sigma) \int_{S^{d-1}} (u^T u)\, d\sigma(u)$$

3.5 Approximate Bayesian Computation

After having define these optimal transport notions, the aim of the coming section is to provide more details about likelihood free inference, more precisely Approximate Bayesian Computation we have define the optimal transport notation lets go back to the likelihood free inference. As it as said in the introduction, ABC methods approximate the posterior distribution by generating synthetic data from the model and retaining parameter values for which the synthetic data are close enough to the observed data. The closeness is typically measured using a discrepancy measure between the two datasets, often reduced to summary statistics.

The basic ABC algorithm involves iteratively drawing candidate parameters from a prior distribution, generating synthetic data from the model with these parameters, and accepting the parameters if the data discrepancy measure is below a specified tolerance threshold. The algorithm returns samples from the approximate posterior distribution.

More precisely, the algorithm does the following :

ABC Algorithm:

- The algorithm draws a candidate parameter θ' from a prior distribution π.
- Synthetic data $z_{1:m} = (z_i)_{i=1}^m$ is generated from the model with parameters θ'.
- The candidate parameter θ' is accepted if the data discrepancy measure $D(s(y_{1:n}), s(z_{1:m}))$ is below a tolerance threshold ϵ.
- The process is repeated to obtain samples from the approximate posterior distribution.

The acceptance rule in the ABC algorithm is based on a data discrepancy measure D between the observed data $y_{1:n}$ and the synthetic data $z_{1:m}$ and s is a summary statistics function that reduces the high-dimensional datasets to a lower-dimensional space ($\mathbb{R}^{d_s}$).

3.6 Wasserstein - ABC

Selecting appropriate summary statistics is there fore of the utter importance since its choice influences the quality of the approximate posterior distribution.

Therefore, Wasserstein-ABC is proposed as a method to mitigate the loss of information associated with inadequate summary statistics. It utilizes Wasserstein distance, defined in the previous section to quantify the closeness between the observed and synthetic data.

To handle datasets of varying sizes efficiently, two approximations can be introduced [**?**][9]. The goal is to improve the efficiency and accuracy of ABC methods, especially when dealing with high-dimensional datasets.

- the **Hilbert Distance** is an approximation of the Wasserstein distance, that extends the computation of W_p in 1D to higher dimensions. It involves sorting samples based on their projection obtained through the Hilbert space-filling curve. This alternative can be computed in O(n log(n)), making it scalable to larger datasets. However, it yields accurate approximations primarily for low dimensions.
- the **Swapping Distance** is an another approximation introduced, which relies on an iterative greedy swapping algorithm. However, each iteration of the algorithm requires n^2 operations, making it computationally expensive. Moreover, there is no guarantee of convergence to W_p with the swapping distance.

3.7 Sliced Wasserstein - ABC

Then the authors of [1] propose SW-ABC (Sliced-Wasserstein ABC) which is their main contribution. It is proposed as a variant of Approximate Bayesian Computation (ABC) inspired by the success both the usage of Wasserstein distance for ABC algorithm and the advantages of Sliced-Wasserstein (SW) compared to Wassersten distance itself. Thus, compared to Wasserstein - ABC, In SW-ABC, empirical distributions are compared using the Sliced-Wasserstein distance, instead of W_p, SW-ABC employs SW_p for comparing empirical distributions within the ABC framework. The resulting posterior distribution, the SW-ABC posterior, is formulated similarly to standard ABC, with SW_p replacing the traditional discrepancy measure. SW-ABC offers improved scalability to both data size and dimension, leveraging the efficiency of SW.

4 Theoretical Guarantees

The very first guarantee that one would want to get using ABC methods would be that at least, when the threshold of acceptace of the distance between the generated data and the original data tends to zero, the posterior found from the ABC computation do tends to the true posterior.

This theoretical section will therefore dissect meticulously one of the theoretical finding of the paper [1] which is the convergence of SW-ABC posterior to the true posterior when the threshold acceptance decrease. Even though this result was expected redoing the proof from scratch was a good way to scrap the literature and gain more ease with the theoretical tools of Optimal Transports.

4.1 Convergence in the case of the Wasserstein distance

To begin with [**?**] demonstrate the convergence of the ABC posterior in the case of using the Wasserstein distance. Since the Sliced-Wasserstein distance is derived from the Wasserstein distance, starting with guarantees on the Wasserstein itself is the natural thing to have some solid background.

Lemma 4.1 (Convergence Conditions for ABC Posterior under Fixed Observations fro the Wasserstein distance)**.** *$\mathcal{H}_1$: Suppose $\mu_\theta^{(n)}$ has a continuous density $f_\theta^{(n)}$, with*

$$\sup_{\theta \in H \setminus N_H} f_\theta^{(n)}(y_{1:n}) < \infty,$$

where $N_H \subset H$ such that $\pi(N_H) = 0$.
$\mathcal{H}_2$: Assume, the existence of $\epsilon > 0$ such that

$$\sup_{\theta \in H \setminus N_H} \sup_{z_{1:n} \in A^\epsilon} f_\theta^{(n)}(z_{1:n}) < \infty,$$

where $A^\epsilon = \{z_{1:n} : D(y_{1:n}, z_{1:n}) \leq \epsilon\}$.

$\mathcal{H}_3$: Also, suppose D is continuous in the sense that $D(y_{1:n}, z_{1:n}) \to D(y_{1:n}, x_{1:n})$ whenever $z_{1:n} \to x_{1:n}$ componentwise for a metric. If either

- $\mathcal{L}_1$: *$f_\theta^{(n)}$ is n exchangeable, i.e., $f_\theta^{(n)}(y_{1:n}) = f_\theta^{(n)}(y_{\sigma(1:n)})$ for any $\sigma \in S_n$, and $D(y_{1:n}, z_{1:n}) \neq 0$ if and only if $z_{1:n} = y_{\sigma(1:n)}$ for some $\sigma \in S_n$*
- $\mathcal{L}_2$: *$D(y_{1:n}, z_{1:n}) \neq 0$ if and only if $z_{1:n} = y_{1:n}$,*

then, with fixed $y_{1:n}$, the ABC posterior converges strongly to the true posterior as $\epsilon \to 0$

Proof. The proof can be found in [2]. □

4.2 The case of the Sliced-Wasserstein distance

It is know possible to show the main theoretical finding of the paper [1] which the convergence of the ABC posterior using the Sliced-Wasserstein distance .

Proposition 4.2 (Asymptotic Consistency of SW-ABC Posterior)**.** *Suppose the $\mathcal{H}_1$ and $\mathcal{H}_2$ of the previous lemma 4.1 verified, with D being SW_p and $y_{1:n}$ fixed.*

*Then, the **SW-ABC posterior** $\pi^\epsilon_{y1:n}(B)$ converges to the true posterior $\pi(B|y_{1:n})$ as ϵ approaches zero, for any measurable $B \subset \Theta$.*

Proof. **Proof Overview:** The main goal is to use the previously defined lemma.
One hypotheses is missing which is $\mathcal{H}_3$, and then it is necessary to either check $\mathcal{L}_1$ or $\mathcal{L}_2$.
In particular, the aim is to verify twos conditions from the lemma:
- $\mathcal{H}_3$ Show continuity of SW_p through the convergence of empirical measures under W_p metric.
- $\mathcal{L}_2$ $SW_p(\mu_{\text{empirical}}, \mu_{\theta,\text{empirical}}) = 0$ if and only if $\mu_{\text{empirical}} = \mu_{\theta,\text{empirical}}$.

- ($\mathcal{H}_3$):

  Show the continuity of SW_p through the convergence of empirical measures under W_p metric. In other words, if $(z^k_{1:m})_{k\in\mathbb{N}}$ converges to $z_{1:m}$ in the metric ρ, then, for any empirical distribution $\hat{\mu}_n$,

  $$\lim_{k\to\infty} SW_p(\hat{\mu}_n, \hat{\mu}^k_{\theta,m}) = SW_p(\hat{\mu}_n, \hat{\mu}_{\theta,m}),$$

  where $\hat{\mu}^k_{\theta,m}$ is the empirical measure of $z^k_{1:m}$.

  To do so we need to use tow intermediary steps $(\mathcal{H}^1_3)$ and $(\mathcal{H}^2_3)$.

  Let $y' \in Y$ and $\psi : Y \to \mathbb{R}$ be a continuous function such that $|\psi(y)| \leq K(1 + \rho(y', y)^p)$ with $K \in \mathbb{R}$.

  ($\mathcal{H}^1_3$): Convergence of Empirical Measures:

  As $(z^k_{1:m})$ converges to $z1 : m$ in ρ, we have $\lim_{k\to\infty} \int \psi \, d\mu_{\theta_k,m} = \int \psi \, d\mu_{\theta,m}$ due to the continuity of ψ .

  This implies that $\mu_{\theta_k,m}$ weakly converges to $\mu_{\theta,m}$ in $P^p(Y)$ [[3], Definition 6.8], equivalent to $\lim_{k\to\infty} W_p(\mu_{\theta_k,m}, \mu$ 0 [[3], Theorem 6.9].

  ($\mathcal{H}^2_3$) Application of Triangle Inequality:

  Applying the triangle inequality and [[4], Proposition 5.1.3], we find $C \geq 0$ such that for any empirical measure $\mu_{\text{empirical}}$:

  $$|SW_p(\mu_{\text{empirical}}, \mu_{\theta_k,m}) - SW_p(\mu_{\text{empirical}}, \mu_{\theta,m})| \leq SW_p(\mu_{\theta_k,m}, \mu_{\theta,m}) \leq C_p(\mu_{\theta_k,m}, \mu_{\theta,m})$$

  We conclude that $\lim_{k\to\infty} SW_p(\mu_{\text{empirical}}, \mu_{\theta_k,m}) = SW_p(\mu_{\text{empirical}}, \mu_{\theta,m})$, making condition (ii) applicable.

- ($\mathcal{L}_2$):

  This condition directly follows from the fact that SW_p is a distance measure [[4], Proposition 5.1.2]. Thus it is defined. Therefore, the distance SW_p between two empirical measures is zero if and only if the measures are identical. More precisely lets show $SW_p(\mu, \nu) = 0$ implies $\mu = \nu$. But if $SW_p(\mu, \nu) = 0$, then $u^*_\# \mu = u^*_\# \nu$ for almost every $u \in S^{d-1}$, and this, in turn, yields:

$$
\begin{aligned}
F_\mu(su) &:= \int_{\mathbb{R}^d} e^{-2i\pi s\langle u|x\rangle}\, d\mu(x) \\
&= F(u^*_\# \mu)(s) \\
&= F(u^*_\# \nu)(s) \\
&= F_\nu(su).
\end{aligned}
$$

Since the Fourier transform is injective, we conclude that $\mu = \nu$.

- **Conclusion:**

  Therefore, the two missing conditions from 4.1 are verified.

□

5 Numerical Results

Two experiments were conducted to complete the ones provided in the initial paper of [1].

The first one, is to identify in 1D the advantages of the Wasserstein distance compared to other distances such as the Kuback-Leiberg divernce or the Maximum Mean Discrepancy. The second one is to look at the computational efficiency of W-ABC, SW-ABC compared each other in quite "low" but higher dimension then 1D.

5.1 Experimentations in 1D: comparison to $\mathcal{N}(0, 1)$

The paper motivated SW by the fact that W_p has an easy to compute and efficient formulation in 1D.

Moreover, for uni-dimensional distribution the Sliced Wasserstein distance are equal to each others.

The aim was to put forward the fact that the Wasserstein distance can be easy calculated and depicts a "smooth" distance (in the sense that it doe not exploses compared to other distance such as the D_{KL}) or that it can be analytically computed (compared to the MMD) which enable better interpretations and guaranties of the proprieties of the distance.

Indeed, in the paper of [1] they only demonstrate and do not try to experiment these properties of the Wasserstein distance.

5.1.1 The Wasserstein distance

Since we are in the simplest case, comparing two univariate Gaussian it is possible to compute analytically the Wasserstein distance between the two.

Indeed, for two univariate Gaussian distributions with parameters (μ_1, σ_1) and (μ_2, σ_2), the cumulative distribution functions (CDFs) are given by:

$$
F_1(x) = \frac{1}{2}\left(1 + \operatorname{erf}\left(\frac{x - \mu_1}{\sqrt{2}\sigma_1}\right)\right)
$$

$$
F_2(x) = \frac{1}{2}\left(1 + \operatorname{erf}\left(\frac{x - \mu_2}{\sqrt{2}\sigma_2}\right)\right)
$$

where erf is the error function.

As it was said previously, for univariate distributions, the Wasserstein distance between these distributions is given by the integral of the absolute difference between their inverse CDFs:

$$
W_1(N(\mu_1, \sigma_1), N(\mu_2, \sigma_2)) = \int |F_1^{-1}(p) - F_2^{-1}(p)|\, dp
$$

Now, for Gaussian distributions, the inverse CDF (quantile function) can be expressed in terms of the mean (μ) and standard deviation (σ):

$$F^{-1}(p) = \mu + \sigma \cdot \text{erfinv}(2p - 1)$$

Since `numpy` already implement the quantile function, it is possible to directly integrate this formula.

However, an even more detailed formula can be derived. Indeed, it is sufficient to substitute this expression into the Wasserstein distance formula and perform the integration, and you'll arrive at the formula I provided earlier:

$$W_1(N(\mu_1, \sigma_1), N(\mu_2, \sigma_2)) = \frac{|\mu_1 - \mu_2|}{\sqrt{2(\sigma_1^2 + \sigma_2^2)}}$$

This formula essentially represents the normalized absolute difference between the means, taking into account the variability (standard deviation) of the distributions. It provides a measure of how much mass needs to be transported from one location to another to transform one distribution into the other.

5.1.2 The Kullback-Leiber divergence

On of the most wide spread measure of the distance between two distribution is the Kullback-Leiber divergence.

Intuition: It is the difference of the entropy $(-log(p_x))$of two distributions, weighted by the probabilities of the first distribution to take into account that some event might not occur

More formally, for two probability measures P and Q that posses densities p and q with regard to the Lebesgues measure, the Kullback-Leiber divergence is defined as :

$$D_{\mathrm{KL}}(p \,||\, q) = \int p(x) \log\left(\frac{p(x)}{q(x)}\right) dx$$

Remark. This tools mainly based on entropic considerations can also be interpretated in term of Optimal Transport, more specifically in term of regularized optimal transport. Indeed, regularized optimal transport can be seen as a projection according to the Kullback-Leiber divergence in the sens that given a convex set $\mathcal{C} \subset \mathbb{R}^N$, the projection according to the Kullback-Leiber divergence is defined as

$$Proj_{\mathcal{C}}^{KL}(\xi) = arg\min_{\pi \in \mathcal{C}} KL(\pi|\xi).$$

In the special case of two Gaussians, the objective is therefore to derive $D_{\mathrm{KL}}(N(\mu_1, \sigma_1^2) \,||\, N(\mu_2, \sigma_2^2))$.

To do so, lets substitute the density formula of two Gaussians.

$$p(x) = \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left(-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right)$$
$$q(x) = \frac{1}{\sqrt{2\pi\sigma_2^2}} \exp\left(-\frac{(x-\mu_2)^2}{2\sigma_2^2}\right)$$

Taking the logarithms yields,

$$\log\left(\frac{p(x)}{q(x)}\right) = -\frac{1}{2}\log\left(\frac{\sigma_2^2}{\sigma_1^2}\right) - \frac{(x-\mu_1)^2}{2\sigma_1^2} + \frac{(x-\mu_2)^2}{2\sigma_2^2}$$

Integrating the previous formula yields,

$$D_{\mathrm{KL}}(N(\mu_1, \sigma_1^2) \,||\, N(\mu_2, \sigma_2^2)) = -\frac{1}{2}\log\left(\frac{\sigma_2^2}{\sigma_1^2}\right) + \frac{1}{2}\frac{\sigma_2^2}{\sigma_1^2} + \frac{1}{2}\left(\frac{\mu_1 - \mu_2}{\sigma_1^2}\right)^2 - \frac{1}{2}$$

By simplifying the above formulation, one can show that :

$$D_{\mathrm{KL}}(N(\mu_1, \sigma_1^2) \,||\, N(\mu_2, \sigma_2^2)) = \frac{1}{2}\left(\frac{\sigma_2^2}{\sigma_1^2} + \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2} - 1 + \log\left(\frac{\sigma_1^2}{\sigma_2^2}\right)\right)$$

This completes the derivation of the KL divergence between two univariate Gaussian distributions. The formula obtained quantifies the information gain or loss when approximating one Gaussian distribution by another, considering both mean and variance differences.

5.1.3 Maximum Mean Discrepancy (MMD)

The third type of distance that will be used is the Maximum Mean Discrepancy (MMD).

Some notions needs to be recalles. First, an Hilbert space which is a complete inner product space, which means it has a notion of distance and angle. In the context of MMD, H is a Hilbert space of functions defined on a set X.

Then, a Reproducing Kernel (k). It is a positive definite kernel function $k : X \times X \to \mathbb{R}$ associated with H. This kernel captures the similarity between elements in X.

This yields the definition of the MMD. Given two probability distributions P and Q over the set X, the MMD between P and Q with respect to the kernel k is defined as:

$$\text{MMD}^2(P, Q) = \mathbb{E}_{x,x' \sim P}[k(x, x')] + \mathbb{E}_{y,y' \sim Q}[k(y, y')] - 2\mathbb{E}_{x \sim P, y \sim Q}[k(x, y)]$$

Where: - $\mathbb{E}$ denotes the expectation. - x, x' are samples from P. - y, y' are samples from Q.

Moreover, the MMD can be expressed in terms of the feature embeddings of distributions in the Reproducing Kernel Hilbert Space (RKHS).

If $\phi : X \to H$ is the feature map associated with the kernel k, then the MMD can be expressed as:

$$\text{MMD}^2(P, Q) = \|\mathbb{E}_{x \sim P}[\phi(x)] - \mathbb{E}_{y \sim Q}[\phi(y)]\|^2$$

Remark. The choice of the kernel function k is crucial and depends on the characteristics of the data. Common choices include the Gaussian kernel, polynomial kernels, etc.

The previoulsy given definition allows to define an estimation method for the MMD, the empirical MMD.

Given samples $X = \{x_1, x_2, \ldots, x_m\}$ from P and $Y = \{y_1, y_2, \ldots, y_n\}$ from Q, the empirical MMD is computed as:

$$\text{MMD}^2(X, Y) = \frac{1}{m(m-1)} \sum_{i \neq j} k(x_i, x_j) + \frac{1}{n(n-1)} \sum_{i \neq j} k(y_i, y_j) - \frac{2}{mn} \sum_{i,j} k(x_i, y_j)$$

MMD measures the dissimilarity between two distributions by comparing their feature embeddings in a Hilbert space. Even though, it is a powerful tool for non-parametric statistical testing, generative model evaluation, and domain adaptation, no analytical formula can be derived from it.

5.1.4 Illustration of the differences

Implementation: For this first experimentation, the distribution of reference will be the Normal distribution of mean zero and standard deviation of one ($\mathcal{N}(0, 1)$). The second one will still be a Gaussian but with a mean $\mu_2 \in \{-3, 3\}$ and a standard deviation in $\mu_2 \in \{-0.5, 3\}$.

These univariate experiment can be found in my first notebook. The aim of this first notebook was to try to redo everything from scratch, not relying on precomputed libraries for optimal transport such as `Py OT`.

Thanks to the fact that the distribution are Gaussian and univariate it is possible to vary both the mean and the standard devaition and see how the variance of the parameters of the two distribution lead to variations for the different distances.

Intepretation:

Compared to the MMD, both D_{KL} and W_1 do have a well defined analytical formula which enables to better undertsand the properties of these two distance and to be more computationally efficient. Graphically on a heatmap where the distance is plotted against the mean and the variance 1, this can be seen by the monotonic decrease of these two first distances. On the other hand, the necessary statistical estimations of the MMD yields a more granular and dotted aspected, revealing the little statistical incertitude on the estimaton of this distance.

However, the D_{KL} exploses quite quick when the distributions are of low variance, even though means are not really far. These extremes values in the edges completely crush the finer differences in the distances around the distributions. One could say that the D_{KL} is a quite rough measure of distance since it specifically focuses on the edges (in terms of how far the distributions are). On the other hand, both MMD and W_1

allow to keep some importance of little pertubertion around the distrbituion of interest, giving therefore a more suitable tools for a precise estimation.

This is in accordance with the finding of [1] especially in the first figure where they show that compared to the D_{KL} which exploses around the true parameter, the Wasserstein and Sliced-Wasserstein are much flatten.

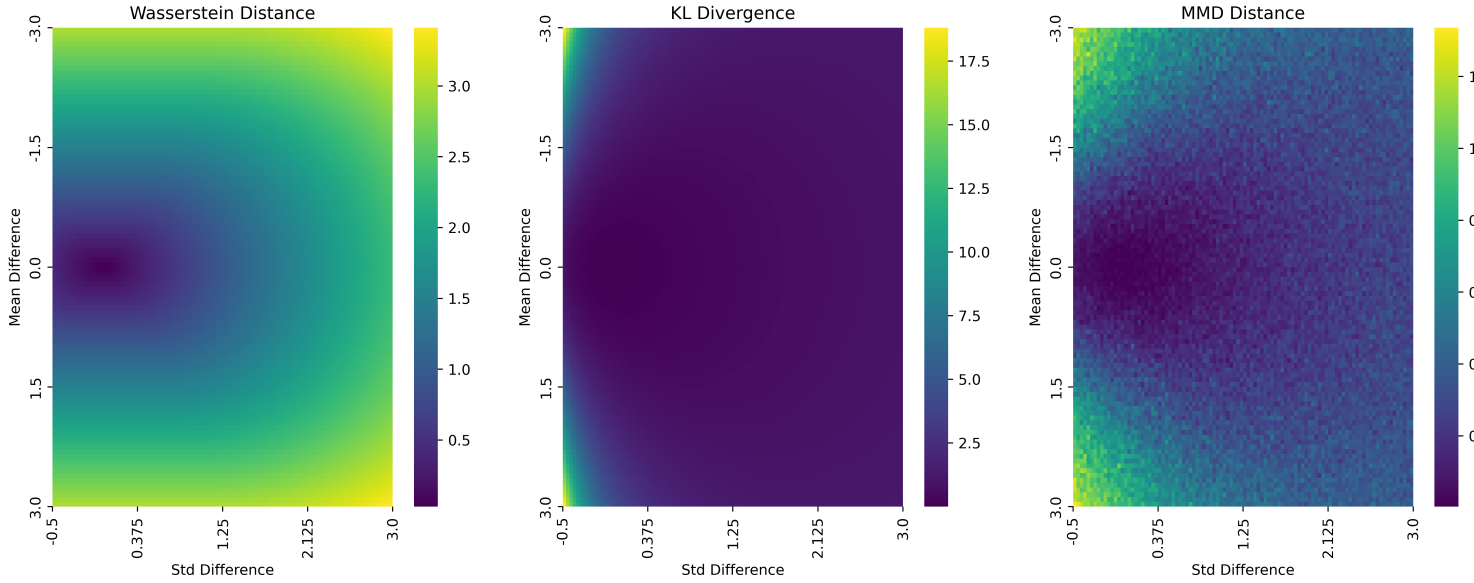

Figure 1: Comparing the Wasserstein distance, the Kullback-Leiber divergence, and the MMD distance.

5.2 Higher dimension

5.3 Challenges in the paper

In higher dimension, the paper [1] explain the statistically advantages of using the sliced Wasserstein distance and Wasserstein distance and how the distance convergences through time for the ABC algorithm.

However, **they do not adress the main motivation of their paper which was to say that the Sliced-Wasserstein distance was computationally more efficient than the Wasserstein distance itself...**

The aim of this section was therefore to test this hypothesis.

5.4 Implementation

The code can be found in this second notebook

5.5 For one given pair of Gaussian

Lets start with two bivariate Gaussian.

P_{true} is drawn from a gausian with $\mu^* = [-0.7, 0.1]$ and a covariance matrix given by

$$\sigma^* = \begin{bmatrix} 1. & 0.5 \\ 0.5 & 1. \end{bmatrix}$$

Different metrics can be used for the ABC algorithm, the Sliced Wasserstein distance and the Wasserstein distance, but also the Euclidian Norm or just the mean between the two distrution (called the Sufficient Summary, even though it might be biased).

Since in the first notebook I recomputed from scratch all of the Optimal Transport from `numpy` for this second experiment I decided to directly use the implementation made by the paper [1] which can be know found in the `PyABC` library, for Approximate Bayesian Computation, combined with the `Py OT` library for optimal transport.

With a Gaussian prior having the same mean and variance of 0.25 (independance), it is possible to see that each of the ABC are able to capture the ground truth 2. This was done for 100 iterations.

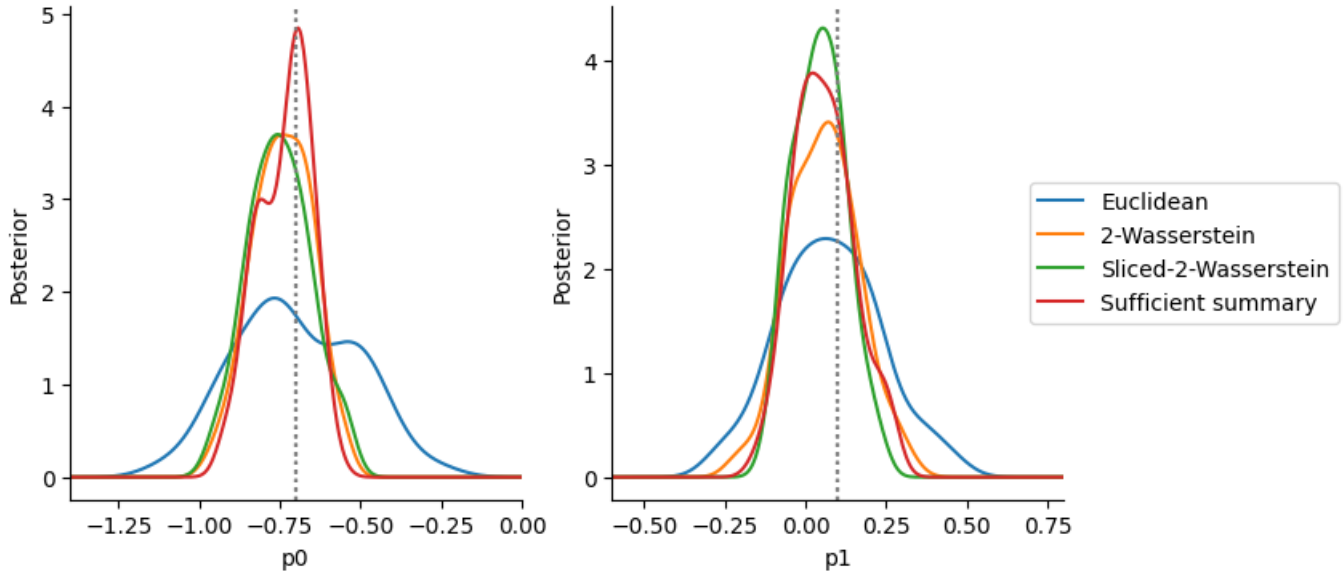

Figure 2: Convergence of different ABC methods

This is in adequation with the finding of the paper [1].

However, if one look at the computation time that it cost to run the sliced Wasserstein distance compared to Wasserstein, it is possible to see that it takes a longer time to execute the SW-ABC than the noraml W-ABC.3 Naturally the execution of these two methods is higher than the ABC with euclidian or sumary statistics since these later two do not need any optimization to be calculated.

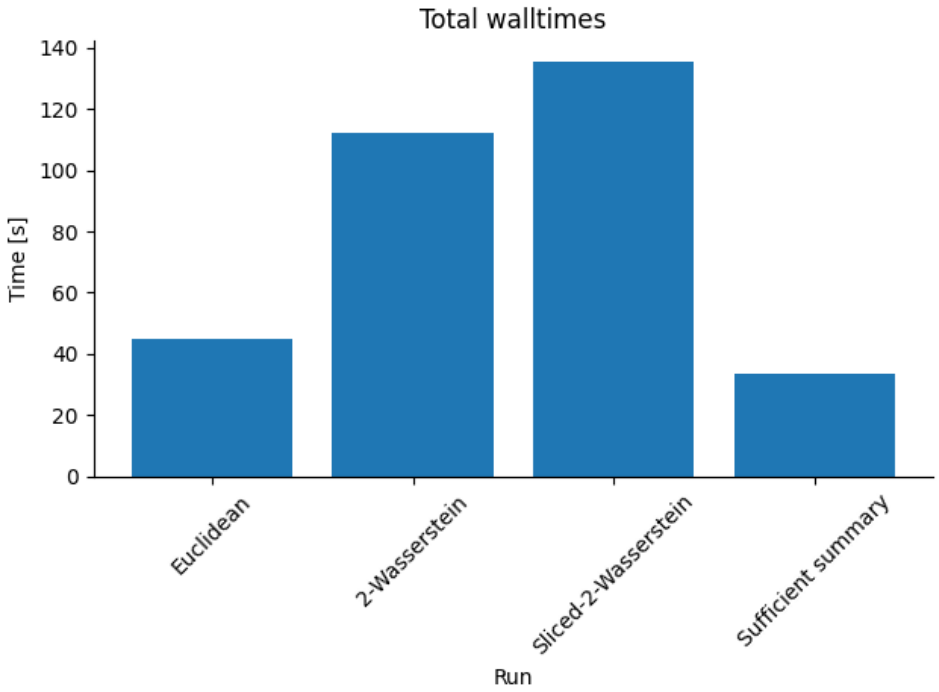

Figure 3: Difference in time execution of ABC, for a bivariate Gaussian

The initial motivation of the paper to introduced the Sliced-Wasserstein distance is therefore not verified... and was not clearly tested in the paper which, in my opinion is quite shady...

5.6 Assessing the importance of the Covariance matrix

One natural hypothesis about the negative result would be that the datastructure was somehow "bad" for the the Sliced Wasserstein distance. Following the recommendations in the guideline of my mini-project I decided to try to study the importance of the decay of the eigen values of the covariance matrix to study these different types of ABC. In particular, while the convergence of the different ABC methods are qualitatively checked, the aim was to try to understand if there might conditions where the it is possible to see a faster Sliced - Wasserstein distance compared to the standard Wasserstein distance

5.6.1 The covariance matrices

To ensure that the eigenvalues of the covariance matrix of the dataset has different decay of eigen values the following was implemented.

The Gaussian where in $\mathbb{R}^3$, with 0 mean for all coordinates. The eigenvalues where chosen to exponentially decay with a decay factor, drawn from `decayArray = np.linspace(0.001, 3, 6)`. 6 decay values wre tested because it made already the code run for 1 hour on Google Collaboratory. The aim is not do heavy computation. The bounds of decay where inspected visually in order to have "flat, linearly decreasing, and strongly decreasing" eigenvalues 4.

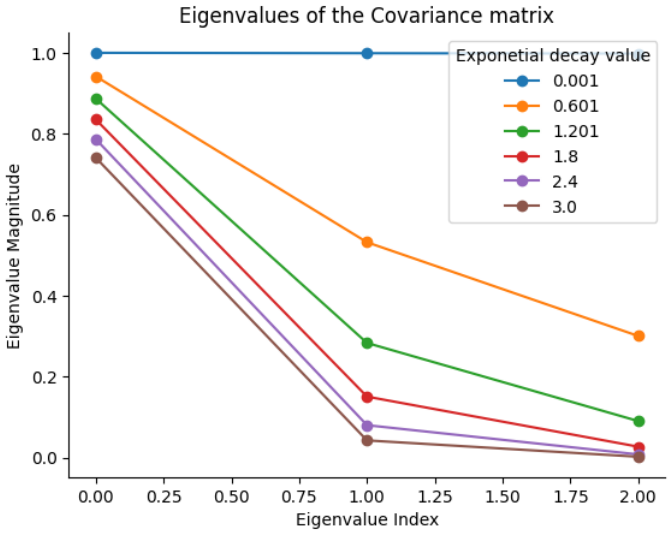

Figure 4: Decay of the eigenvalues of the covariance matrix

Remark. The range of the eigenvalues could be selected by selecting the range of values on which an exponential decay was applyied. The range was selected in order not to have any null eigen values (which would mean a totally predictive distribution along one axis).

In order not to have just independent Gaussians (due to a the diagonal covariance matrix), some noise was added. To do so the covariance matrix was multplied by both side by a random but orthonormal matrix to ensure that the covariance matrices are semi-definite positive. This random orthonormal matrix was created by using the QR decomposition of a random matrix and taking the first element, Q, which is orthonormal.

In the end, the covariance matrices Cov, for a given exponential decay, τ were defined by :

$$Cov(\tau) = Qdiag((\lambda_i)_{i \in I_\tau})Q^T$$

where I_τ is the set of diagonal matrix decaying exponentially.

5.7 Trying to asses the impact of the covariance matrix

To asses the impact of the covariance matrix and the decay of its eigenvalues, the idea was just to loop the previously notebook for different values of covariance matrix. For the computation time, clear pattern could

emerge **??**. The notation and the pictures are therefore the same as before, but looped, to better read the label one can thus rely on the previous figures.

This once again goes against the main motivation of the paper [1], which was once again not tested in the paper.

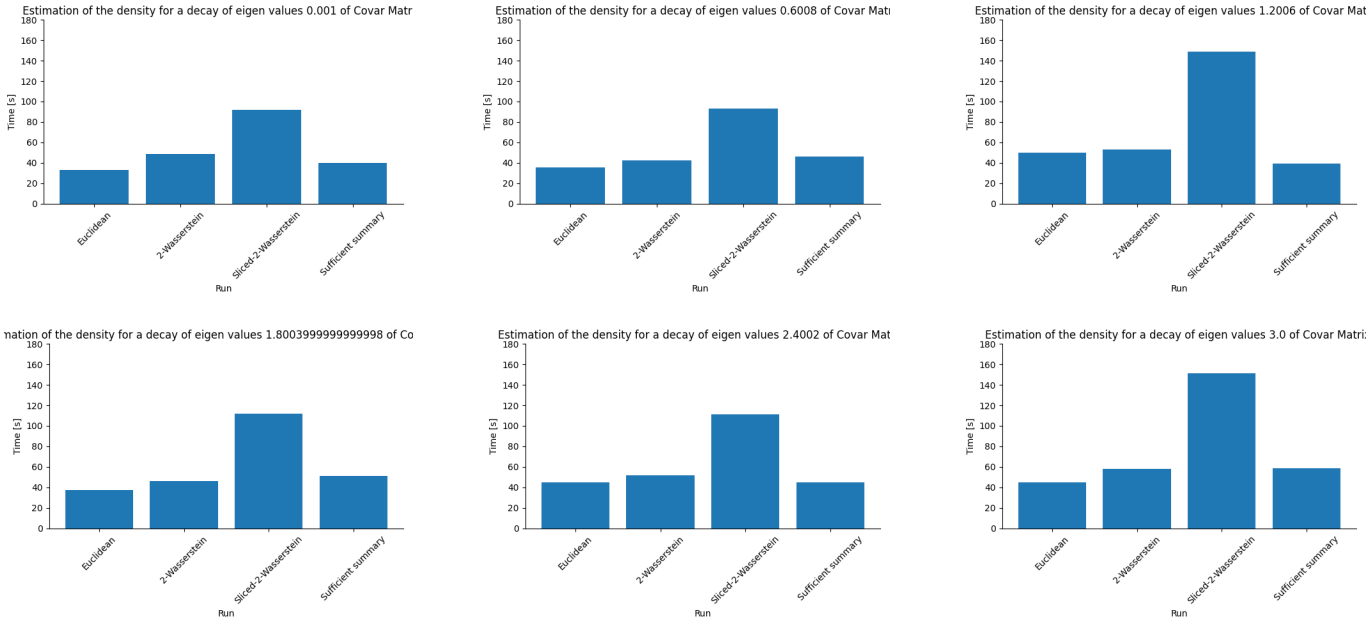

Figure 5: Eigen Value decay does not seems to impact the computation time

Qualitatively, it is also possible to inspect how the estimation of the distribution along the different axis is impacted by different decay of the eigenvalues **??**.

What is possible to see is that the Sliced-Wasserstein distance remain more consistent why the increase of the decay of the eigenvalues of the covariance matrix. This illustrates that the Sliced- Wasserstein distance have good statistical properties.

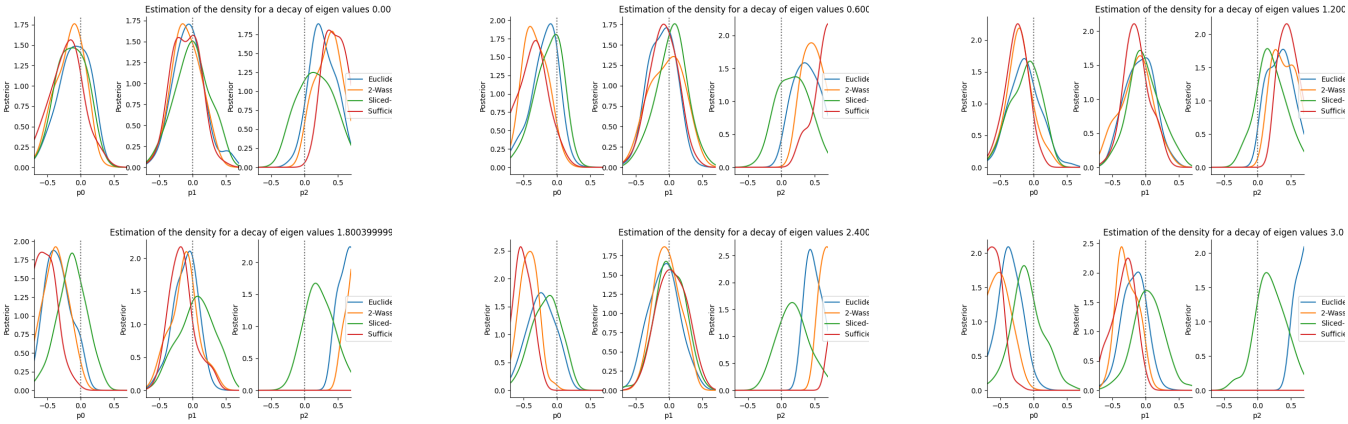

Figure 6: The Sliced Wasserstein distance stays consistent compared to the others

6 Conclusion and Perspective

6.1 Conclusion

In conclusion, the paper explores the application of Approximate Bayesian Computation (ABC) methods to estimate the posterior distribution of model parameters in cases where the likelihood function is computationally intractable. The use of synthetic data generation within this framework, while addressing the challenges of intractable likelihoods, raises the critical question of how to effectively compare generated data

to real data. Traditional summary statistics may induce information loss, and the paper suggests that tools from Optimal Transports, specifically the Wasserstein distance, offer a promising alternative for measuring the distance between synthetic and real data.The contributions of this work involve a thorough examination of Optimal Transports concepts, focusing on the classic Wasserstein distance and introducing the Sliced-Wasserstein ABC technique. Theoretical analysis delves into the convergence of the SW-ABC posterior to the true posterior as the threshold acceptance decreases. Additionally, a comparative exploration of Wasserstein distance, Kullback Leibler divergence, and Maximum Mean Discrepancy in low dimensions helps build intuition about their advantages and differences. The numerical part of the work further investigates the paper's suggestion of using Sliced-Wasserstein distance for computational advantages. Surprisingly, the empirical findings indicate no computational advantages of Sliced-Wasserstein distance over the regular Wasserstein distance, even when considering the impact of covariance in higher dimensions.

6.2 Perspective : Usage of ABC computation for interpretability of LLMs

Lastly, I would like to suggests potential directions linking these findings to the interpretability of Neural Networks, expanding the scope of the study to broader applications beyond statistical inference. One possibility would be to use Approximate Bayesian Computation (ABC) to estimate parameters within black-box (or grey box) models for improved interpretability. The final aim would not be to try to find a parameter in the dataset but rather to find the parameters that the neural network had learned.

For instance, lets take the case of Large Language Models (LLMs). One could select a topic of interest and give some text to our language model. Let's denote this dataset $\mathcal{D}$. Then, this would yields a certain distribution of data points in the embedding space of the LLM. Then, one could use Approximate Bayesian Computation with the generated data being text generated by an other LLM. The discrepancy measure being computed in the embedding space of the first LLM in order to gain a mapping to what really the first LLM is considering similar to the dataset $\mathcal{D}$. Then the role of the first LLM and the second LLM are reversed and this would be a way to measure how the representation of two LLMs could be relatively different. For instance, if the dataset is only of the letter A, the estimated dataset for one LLM could be $ABCDEF$, showing that this LLM has learned knowledge related to grammar, and maybe for the other it would yields $ACGTCG$ saying that the second LLM has more knowledge in biology.

More formally :

Let's pretend to have a set of textual data T. The aim would be to try to understand what the model M_1 compared to M_2 and if they share the same representation.

try to understand if what the model M_2 is saying to M_1 is what M_2 thinks. Otherwise, it could "lie". Deception in LLM is a very important topic in interpretability [5]

First, we would use M_2, to generate text data T_{M_2}. Then we could go to embedding space of M_1, calculate $D_1(T, T_{M_2})$. From that create an ensemble$S_{M_2} := \{T_{M_2} : D_1(T, T_{M_2}) \leq \epsilon\}$ Then, in the embedding of M_2, calculate $D_2(T, S_{M_2})$.

References

[1] Kimia Nadjahi, Valentin De Bortoli, Alain Durmus, Roland Badeau, Umut Şimşekli, *Approximate Bayesian Computation with the Sliced-Wasserstein Distance*. ArxiV, 2019.

[2] E. Bernton, P. E. Jacob, M. Gerber, and C. P. Robert, *Approximate bayesian computation with the wasserstein distance*. Series B (Statistical Methodology), vol. 81, no. 2, pp. 235–269, 2019.

[3] C. Villani, *Optimal Transport: Old and New,e*. Grundlehren der mathematischen Wissenschaften. Springer, 2009 edition, Sept. 2008.

[4] N. Bonnotte, *Unidimensional and Evolution Methods for Optimal Transportation*. PhD thesis, 2013

[5] C. Burns, H. Ye, D. Klein, J. Steinhardt *Discovering Latent Knowledge in Language Models Without Supervision*. ArxiV 2022.