Review. A Wasserstein-type distance in the space of Gaussian Mixture Models

Dimitri Meunier
Ensae & ENS Paris Saclay
dimitri.meunier@ensae.fr

Abstract

In this report, we review the article [5]. Optimal Transport distances are not always easy to compute. In the raw form of Optimal Transport, one has to solve a Linear Program whose cost can quickly become prohibitive whenever the size of the support of these measures or the histograms dimension exceeds a few hundred. Motivated by a need for approximation and simplification the authors introduce an approximation of the Wasserstein distance on the set of Gaussian Mixture Models. By restricting the set of possible coupling measures to the set of Gaussian Mixture distributions they make the new distance fast compute. It boils down to solve a low dimensional linear program between the probability vectors of the GMM. To illustrate that their method scale to transport between large cloud points, they apply it to the problem of color transfer with high resolution images. They also show connections with the Wasserstein distance and derive a theoretical bound on the approximation error.

1 Introduction

Optimal transport is a set of mathematical tools to solve the problem of transporting masses from a source distribution to a target distribution in a mass preserving manner with minimum cost. Initially created to solve the problem of transporting amount of dirt, it as been extended to the transport of any distribution onto another. OT delivers tools to compare distributions and the resulting distances share good geometrical properties.

Despite good theoretical properties [17], computing optimal transport distances remained intractable when the size of the cloud points we aim to transport exceeded a certain limit. The main distance from Optimal Transport is called the Wasserstein distance and to compute it one has to solve a linear program. Even if fast methods exist for this task, it can remain slow. They have been successes to tackle this problem introducing entropic regularisation for fast computation with the Sinkhorn algorithm [4].

[5] builds on the work of [3] and offers a different perspective for approximating Optimal Transport. Instead of adding a penalty to the objective function, they constrain the solution to have a specific shape: a Gaussian Mixture Distribution. Mixture Models are convex combinations of distributions from the same parametric family. They are powerful and flexible models that can represent the presence of subpopulations within an overall population, without requiring that an observed data set should identify the sub-population to which an individual observation belongs. They can therefore be used to unsupervised clustering and for density approximation. Indeed, any distribution can be approximated arbitrarily well by a finite mixture of normal densities [11]. Gaussian Mixture Models are a specific example of mixture models where each component is Gaussian. Manipulating GMM is practical because most computation involving them are easy and they can model well multidimensional continuous distribution. Optimal transport computation between GMM is not explicit, but it is between Gaussian Variables. We will see that it is of essential importance to build computationally efficient OT-like distances.

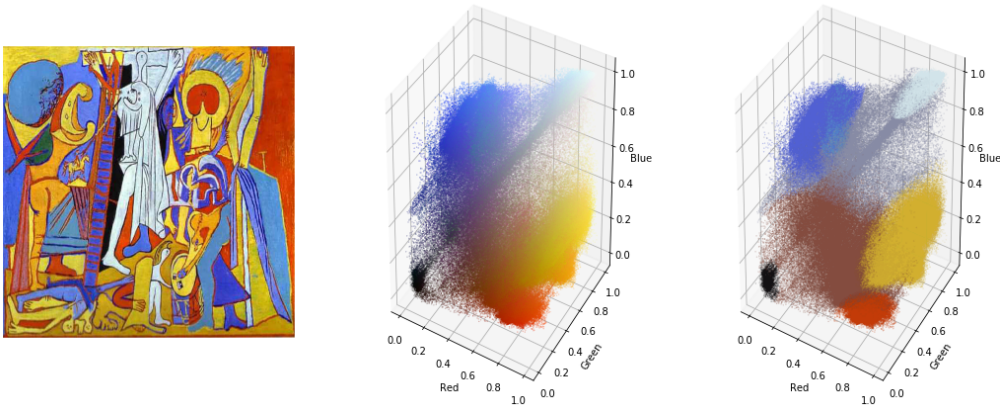


Figure 1: Illustration of a fitted GMM. (*left*) Crucifixion by Pablo Picasso (1930), size=600*600 pixels, source: WikiArt. (*Middle*) Colour histogram of pixels in the RGB space, 360,000 points. (*Right*) Colour histogram approximated by a GMM with 8 classes, it took less than 5.5 seconds to fit the 600*600 points using a Scikit-Learn implementation [14] on CPU.

As pointed out in [3], in data science, meaningful data always have some structure. In many cases, the data are sparsely distributed among subgroups. The difference between data within a subgroup is way less significant than that between subgroups. For such data set, a mixture model can be well fitted. The most well known algorithm to fit A GMM to data is the Expectation-Maximisation algorithm [1]. It gives a powerful and fast way of fitted a GMM to a wide range of datasets. GMM combined with the EM algorithm can perform unsupervised learning and clustering that perform well on high dimensional tasks such as imaging (see figure 1).

Except in special cases, the optimal coupling from the Wasserstein distance between two GMM is not a GMM itself. From this observation, the authors suggest a constrained version of the Wasserstein distance, denoted the Mixture Wasserstein distance (MW_2), that forces the resulting optimal coupling to stay in the space of GMM. The main result of the article is the proof that the MW_2 distance is equivalent to the discrete formulation found by [3]. It boils down to solve a low dimensional linear program that connects the probability vectors of the two GMM and where the cost matrix is the pairwise Wasserstein distances between the Gaussian components. Using the continuous formulation of the distance, they show tight connections to the Wasserstein distance that were not mentioned in the literature formerly. They also introduce a generalisation to the multi marginal and the barycenter settings and a bound on the approximation error of MW_2. They demonstrate the usefulness of this approach in a computationally intensive setting with a practical example on color transfer. Finally, they explain how the results could be extended to other Mixture Models.

In the next section, we will first review the main components of Optimal Transport. The key element is the Wasserstein distance that we need to adapt for the case of the non compact space $\mathcal{X} = \mathbb{R}^d$. Secondly, we will recall and show the main results of Optimal Transport with Gaussian distributions. Next, we will explain the proof of two important properties of GMM: a density property and an identifiability property. The later is essential to build the Mixture Wasserstein distance. We will also explain how the EM algorithm can be use to fit a GMM to a dataset as it will be useful for our numerical experiments. Then, we will detail the key result of the paper: the derivation of the Mixture Wasserstein distance and how it is linked to the distance derived in [3]. Finally, we will introduce two numerical contributions. It is mentioned in the paper that other Mixtures than the GMM are conceivable. Among the family of elliptic distributions the Mixture of T-distributions is the best applicant, we will give numerical evidence of this fact. We will also replicate the *color transfer* experiment and make a comparison with the entropic approach that works by applying gradient descent on the Sinkhorn divergence loss. The comparison is based on computational speed and quality of the results.

Notations

- $||.||$, Euclidian norm in $\mathbb{R}^d$
- $\mathcal{M}(\mathcal{X})$ set of arbitrary measures on $\mathcal{X}$, $\mathcal{M}_+(\mathcal{X})$ set of positive measures on $\mathcal{X}$, $\mathcal{M}_+^1(\mathcal{X})$ set of probability measures on $\mathcal{X}$
- $\mathcal{C}(\mathcal{X})$ set of continuous function on $\mathcal{X}$
- Push-forward operator $T_\#$. For $T\colon \mathcal{X} \mapsto \mathcal{Y}$, the push forward measure $\beta = T_\#\alpha \in \mathcal{M}(\mathcal{Y})$ of some $\alpha \in \mathcal{M}(\mathcal{X})$ satisfies $\beta(B) = \alpha(T^{-1}(B)) \quad \forall B \subset \mathcal{Y}$
- $\mathbb{1}_p$ denotes a column vector of ones of length p
- $\forall x, y \in \mathbb{R}^d \quad P_t(x, y) = (1-t)x + ty$
- $g_{m,\Sigma}$ density of the multivariate Gaussian density with mean $\mu \in \mathbb{R}^d$ and covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$ (the dimension is implicite)

2 Main review

2.1 Definitions and review of the OT notations

In this part we review some notations and key concepts of Optimal Transport that will be useful in later sections. Since we will be manipulating Gaussian Mixtures Models, we will always consider $\mathcal{X} = \mathbb{R}^d$ which is not compact. We thus have to take extra care and we will refer to the definitions of [17].

Definition 2.1 *Wasserstein space.* $\mathcal{P}_p(\mathbb{R}^d) = \left\{\mu |\ \mu \textit{ probability measure, } \int_{\mathbb{R}^d} ||x||^p d_\mu(x) < +\infty\right\}$

As explain in [17], working in the Wasserstein space ensures that the Wasserstein distance (see below) is finite. GMM are in $\mathcal{P}_p(\mathbb{R}^d)$ for all p because they are finite combinations of Gaussian distributions that admit exponential moments.

If μ_0 and μ_1 are two probability measures in $\mathcal{P}_p(\mathbb{R}^d)$, we define $\Pi(\mu_0, \mu_1) \in \mathcal{P}_p(\mathbb{R}^d \times \mathbb{R}^d)$ the subset of probability distributions γ on $\mathbb{R}^d \times \mathbb{R}^d$ such that $P_0\#\gamma = \mu_0$ and $P_1\#\gamma = \mu_1$. This is equivalent to the following statement: if X, Y are two random variables such as $(X, Y) \sim \gamma \in \mathcal{P}_p(\mathbb{R}^d \times \mathbb{R}^d)$ then $X \sim \mu_0$ and $Y \sim \mu_1$.

Definition 2.2 *p-Wasserstein distance.* $\forall(\mu_0, \mu_1) \in \mathcal{P}_p(\mathbb{R}^d)^2$

$$\begin{aligned} W_p^p(\mu_0, \mu_1) &= \inf_{\gamma \in \Pi(\mu_0,\mu_1)} \int_{\mathbb{R}^d \times \mathbb{R}^d} ||y_0 - y_1||^p d\gamma(y_0, y_1) \\ &= \inf_{X\sim\mu_0, Y\sim\mu_1} \mathbb{E}[||X - Y||^p] \end{aligned}$$

It is called the p-Wasserstein distance because it provides a distance on $\mathcal{P}_p(\mathbb{R}^d)$. The symmetry and the separativity are trivial and the proof that the p-Wasserstein distance satisfies the triangular inequality can be found in [17] (proof 6.1). They use a technical tool called the *gluing lemma*. To be properly called a distance, we also have to show that it is always finite. First, let's not that the constraint set is never empty since the independent coupling $\mu_0 \otimes \mu_1$ is in $\Pi(\mu_0, \mu_1)$. Indeed, $\forall A, B \in \mathcal{B}(\mathbb{R}^d)$,

$$P_0\#(\mu_0 \otimes \mu_1)(A) = (\mu_0 \otimes \mu_1)(P_0^{-1}(A)) = (\mu_0 \otimes \mu_1)(A \times \mathbb{R}^d) = \mu_0(A) \times \mu_1(\mathbb{R}^d) = \mu_0(A)$$

and similarly $P_1\#(\mu_0 \otimes \mu_1)(B) = \mu_1(B)$. Secondly, if $(\mu_0, \mu_1) \in \mathcal{P}_p(\mathbb{R}^d)^2$ and $X \sim \mu_0, Y \sim \mu_1$ then, using Jensen's inequality with the convex function $x \mapsto x^p$ and the triangular inequality,

$$W_p^p(\mu_0, \mu_1) \leq \mathbb{E}[||X - Y||^p] \leq 2^p \mathbb{E}[\big(\frac{||X|| + ||Y||}{2}\big)^p] \leq 2^{p-1}\big(\mathbb{E}[||X||^p] + \mathbb{E}[||Y||^p]\big) < +\infty$$

We thus conclude that $(\mathcal{P}_p(\mathbb{R}^d), W_p)$ is a metric space. We will show in later section that it is also separable.

Definition 2.2 is called the Kantorovich formulation of Optimal Transport. When the cost is $c(x,y) = ||x-y||^p$ and p=2, Brenier's theorem ([15] Theorem 1 from [2]) tells us that the Monge formulation (remark 2.7 [15]) and the Kantorovich formulation are equal and in particular that the optimal coupling γ^* is **unique** and linked to a **Monge map** – a map T such that $T_{\#}\mu_0 = \mu_1$ – by the formula $\gamma^* = (Id, T)_{\#}\mu_0$ i.e.

$$\int_{\mathbb{R}^d \times \mathbb{R}^d} h(x,y) d\pi(x,y) = \int_{\mathbb{R}^d} h(x, T(x)) d\mu_0$$

Thus, γ^* is a deterministic coupling ([17] definition 1.2), and has two useful interpretations: if $(X,Y) \sim \gamma^*$ then $Y = T(X)$ and the law γ^* is supported on the graph of T. We also have that T is the unique Monge map that is a gradient of a convex function.

Using this results, the rest of the review will focus on the manipulation of the 2-Wasserstein distance: the space is $\mathcal{X} = \mathbb{R}^d$ and the cost function is the square Euclidian distance: $c(x,y) = ||x-y||^2$.

We end this section with a property that is useful for calculus with Gaussian distributions (this is a consequence of [16] remark 1.9) that say that to compute W_2 it is enough to compute the distance with centered distributions.

Proposition 1 *Translations. Let T_t be the translation operator $T_t \colon x \mapsto x - t$, $m_\mu = \int_{\mathcal{X}} x d\mu(x)$ and $\tilde{\mu} = T_{m_\mu} \# \mu$, then,*

$$W_2^2(\mu_0, \mu_1) = W_2^2(\tilde{\mu}_0, \tilde{\mu}_1) + ||m_{\mu_0} - m_{\mu_1}||^2$$

We deliberately omit the multi-marginal and the barycenters formulation of Optimal Transport as we will not use it in the numerical experiments.

2.2 Optimal Transport between Gaussian distributions

The main advantage of using Gaussian distributions is that most of the operations related to Optimal Transport are explicit. We will see that this ease of use will help to get simplified methods in the case of GMM. In this section, we recall the main Optimal Transport properties for Gaussian distributions with proofs.

We recall the probabilistic point of view of push forwards as it allows to reduce the size of the proofs: if $X \sim \alpha$ and $Y \sim \beta$ then $\beta = T_{\#}\alpha$ iff $Y = T(X)$.

Proposition 2 *Optimal Transport for Gaussians. If $\alpha \sim \mathcal{N}(m_\alpha, \Sigma_\alpha)$ and $\beta \sim \mathcal{N}(m_\beta, \Sigma_\beta)$ are two Gaussians in $\mathbb{R}^d$, then the optimal Monge map between them is,*

$$T \colon x \mapsto m_\beta + A(x - m_\alpha)$$
$$A = \Sigma_\alpha^{-\frac{1}{2}} \left(\Sigma_\alpha^{\frac{1}{2}} \Sigma_\beta \Sigma_\alpha^{\frac{1}{2}}\right)^{\frac{1}{2}} \Sigma_\alpha^{-\frac{1}{2}}$$

Proof 1 *T is the gradient of the quadratic function $\psi \colon x \mapsto m_\beta^T x + \frac{1}{2}(x - m_\alpha)^T A (x - m_\alpha)$ and since A is a positive symmetric matrix, ψ is convex. Thus, by Brenier's Theorem, it is sufficient to show that $T_{\#}\alpha = \beta$ to obtain that T is the unique Monge Map.*
Using the probabilistic point of view of push-forwards, we have to show that if $X \sim \alpha$, then $Y = T(X) \sim \beta$. Since X is a Gaussian distribution and T is linear, Y is also a Gaussian distribution and we have,

$$\mathbb{E}[Y] = m_\beta + A(m_\alpha - m_\alpha) = m_\beta$$
$$\mathbb{V}[Y] = A\mathbb{V}[X]A^T = \Sigma_\alpha^{-\frac{1}{2}} \left(\Sigma_\alpha^{\frac{1}{2}} \Sigma_\beta \Sigma_\alpha^{\frac{1}{2}}\right)^{\frac{1}{2}} \Sigma_\alpha^{-\frac{1}{2}} \Sigma_\alpha \Sigma_\alpha^{-\frac{1}{2}} \left(\Sigma_\alpha^{\frac{1}{2}} \Sigma_\beta \Sigma_\alpha^{\frac{1}{2}}\right)^{\frac{1}{2}} \Sigma_\alpha^{-\frac{1}{2}} = \Sigma_\beta$$

Thus $Y \sim \beta$ and T is the unique Monge Map.

The optimal coupling can be retrieve with $\gamma = (Id, T)_{\#}\alpha$ wich is equivalent to $(X,Y) = (X, T(X))$. Since T is linear, (X,Y) is thus a degenerate Gaussian distribution.

Proposition 3 *2-Wasserstein distance. If $\alpha \sim \mathcal{N}(m_\alpha, \Sigma_\alpha)$ and $\beta \sim \mathcal{N}(m_\beta, \Sigma_\beta)$ are two Gaussians in $\mathbb{R}^d$, then*

$$W_2^2(\alpha, \beta) = ||m_\alpha - m_\beta||^2 + \textit{Tr}\left(\Sigma_\alpha + \Sigma_\beta - 2\left(\Sigma_\alpha^{\frac{1}{2}} \Sigma_\beta \Sigma_\alpha^{\frac{1}{2}}\right)^{\frac{1}{2}}\right)$$

$\mathcal{B}(\Sigma_\alpha, \Sigma_\beta) := Tr\left(\Sigma_\alpha + \Sigma_\beta - 2\big(\Sigma_\alpha^{\frac{1}{2}}\Sigma_\beta\Sigma_\alpha^{\frac{1}{2}}\big)^{\frac{1}{2}}\right)$ *is the Bures metric.*

Proof 2 *First, from the invariance property of the 2-Wasserstein distance we have that*

$$W_2^2(\alpha, \beta) = ||m_\alpha - m_\beta||^2 + W_2^2(\hat{\alpha}, \hat{\beta})$$

where $\hat{\alpha}$ and $\hat{\beta}$ are the centered version of α and β. Let's show that $W_2^2(\hat{\alpha}, \hat{\beta}) = \mathcal{B}(\Sigma_\alpha, \Sigma_\beta)$. From the last proposition, the optimal map between $\hat{\alpha}$ and $\hat{\beta}$ is $T(x) = Ax$, $A = \Sigma_\alpha^{-\frac{1}{2}}\big(\Sigma_\alpha^{\frac{1}{2}}\Sigma_\beta\Sigma_\alpha^{\frac{1}{2}}\big)^{\frac{1}{2}}\Sigma_\alpha^{-\frac{1}{2}}$. Thus, using repeatedly the cyclic invariance property of the trace,

$$\begin{aligned}
W_2^2(\hat{\alpha}, \hat{\beta}) &= \mathbb{E}_{X\sim\hat{\alpha}}[||X - T(X)||^2] = \mathbb{E}[||(I_d - A)X||^2] = \mathbb{E}[X^T(I_d - A)^T(I_d - A)X] \\
&= \mathbb{E}[Tr((XX^T)(I_d - A)^T(I_d - A))] = Tr\big(\mathbb{E}[XX^T](I_d - A)^2\big) = Tr\big(\Sigma_\alpha(I_d - A)^2\big) \\
&= Tr\big(\Sigma_\alpha(I_d - 2A + A^2)\big) = Tr\big(\Sigma_\alpha\big) + Tr\big(A\Sigma_\alpha A\big) - 2Tr\big(\Sigma_\alpha\Sigma_\alpha^{-\frac{1}{2}}\big(\Sigma_\alpha^{\frac{1}{2}}\Sigma_\beta\Sigma_\alpha^{\frac{1}{2}}\big)^{\frac{1}{2}}\Sigma_\alpha^{-\frac{1}{2}}\big) \\
&= Tr\big(\Sigma_\alpha\big) + Tr\big(\Sigma_\beta\big) - 2Tr\big(\big(\Sigma_\alpha^{\frac{1}{2}}\Sigma_\beta\Sigma_\alpha^{\frac{1}{2}}\big)^{\frac{1}{2}}\big) = \mathcal{B}(\Sigma_\alpha, \Sigma_\beta)
\end{aligned}$$

In the case where the covariance matrices are singular, one can use the Moore-Penrose pseudo inverse to compute the inverse of the square roots. If both distributions are Dirac masses — seen as a degenerate Gaussian distribution where $\Sigma = 0$ — the distance is simply the Euclidian distance between the means.

We end this section with the notion of **displacement interpolation** that will be useful for the numerical experiments.

Definition 2.3 *Displacement Interpolation. If γ is an optimal transport plan for W_2 between two distributions μ_0 and μ_1, the path $(\mu_t)_{t\in[0,1]}$, $\mu_t = P_t\#\gamma$ is the displacement interpolation between μ_0 and μ_1 and is the solution of*

$$\arg\min_\rho (1-t)W_2^2(\mu_0, \rho) + tW_2^2(\mu_1, \rho)$$

If an optimal Monge map T exists between μ_0 and μ_1, then, $\mu_t = ((1-t)Id + tT)\#\mu_0$.

We saw previously that if $X \sim \alpha$, $Y \sim \beta$ are two Gaussian random variables then the optimal coupling is $(X, T(X)) \sim \gamma$. The random variable $P_t(X, T(X)) = (1-t)X + tT(X) \sim P_t\#\gamma$ and since P_t is linear $P_t(X, T(X))$ is also a Gaussian variable.

$$\begin{aligned}
\mu_t &= \mathbb{E}[P_t(X, T(X))] = (1-t)\mathbb{E}[X] + t\mathbb{E}[T(X)] = (1-t)\mu_\alpha + t\mu_\beta \\
\Sigma_t &= \mathbb{V}[P_t(X, T(X)] = (1-t)^2\Sigma_\alpha + t^2\mathbb{V}[T(X)] + t(1-t)[Cov(X, T(X)) + Cov(T(X), X)] \\
&= ((1-t)Id + tC)\Sigma_\alpha((1-t)Id + tC)
\end{aligned}$$

$C = \Sigma_\alpha^{\frac{1}{2}}\big(\Sigma_\alpha^{\frac{1}{2}}\Sigma_\beta\Sigma_\alpha^{\frac{1}{2}}\big)^{-\frac{1}{2}}\Sigma_\alpha^{\frac{1}{2}}$

It shows that the displacement interpolation between Gaussian distributions is also extremely simple to compute as we only have to compute μ_t and Σ_t.

2.3 Gaussian Mixture Models and the EM algorithm

In this section, we focus on Gaussian Mixture Models (GMM) and two properties that motivate their use. We also recall what is the Expectation-Maximisation algorithm.

Definition 2.4 *Gaussian Mixture Models. Let $K \geq 1$ be the number of components in the mixture. A GMM of size K on $\mathbb{R}^d$ is a probability distribution μ on $\mathbb{R}^d$ that can be written as*

$$\mu = \sum_{k=1}^{K} \pi_k \mu_k \quad s.t. \quad \mu_k = \mathcal{N}(m_k, \Sigma_k) \quad \pi \succeq 0 \quad \pi^T \mathbb{1}_K = 1$$

Let's remark that it is fairly easy to simulate observations from a Gaussian Mixture Model. One can simulate according to a discrete distribution on $\{1, \cdots, K\}$ – with probability π – and according to the output k simulate from $\mathcal{N}(m_k, \Sigma_k)$ – with the Box-Muller Transform for example. It is however not straight forward to estimate the parameters of the GMM. The direct maximisation of the likelihood for inference does not lead to a closed form solution. The trick is to introduce latent variables that attribute to each observation its component. The introduction of the latent variables allows to cast the maximisation of the likelihood as a problem that is solvable by the EM algorithm. EM is efficient for maximization problem with latent – unobserved – variables where if the latent variables were observed the likelihood would be easy to solve.

Expectation-Maximisation Algorithm The EM algorithm is implemented in all good Machine Learning frameworks such as Scikit-Learn [14] and is really fast (see figure 1 for an example). The EM method is not just useful for Gaussian Mixtures Models; it's a class of algorithms or a meta-algorithm that deals with missing/hidden information via a specific type of marginalization. The lecture notes [12] gives an excellent introduction to general EM and its derivation.

If the likelihood of the model can be written in the form

$$L(X_{1:n}|\theta) = \int_{\mathcal{Z}} L(X_{1:n}, z|\theta) dz$$

where $\mathcal{Z}$ is the space of latent variables and where $L(X_{1:n}, z|\theta)$ is easy to maximize in θ. Then the following algorithm allows to *climb* the likelihood and make inferences on the parameters. Indeed, the (E) and (M) steps can be seen as coordinate ascent on the space of parameters and on the space of distributions for the latent variables of a specific objective function (see [12]).

Algorithm 1: Generic EM Algorithm

Result: $\theta^{(T)}$ argmax of the log likelihood (potentially just a local argmax)
Initialization: ϵ, $\theta^{(0)}$;
while $l(\theta^{(t+1)}) - l(\theta^{(t)}) > \epsilon$ **do**
  (E-Step);
  $\forall i \in [\![1, n]\!]$

$$Q_i(.) = f(.|X_i, \theta^{(t)})$$

  (M-Step);

$$\theta^{(t+1)} = \arg\max_{\theta} \sum_{i=1}^{n} \mathbb{E}_{Z \sim Q_i} \left[\log f(X_i, Z|\theta) \right]$$

end

This algorithm works because one can show that for each step of the EM algorithm $l(\theta^{(t+1)}) \geq l(\theta^{(t)})$. However, this does not guarantee that EM leads to the global maximum of the likelihood and extra care has to be taken on the initialization of the parameters to keep good convergence properties [11].

EM for Gaussians EM for Gaussians is omnipresent in the literature because the (E) and (M) step becomes explicit. It can be shown using simple derivations with Langrangian multipliers.

Algorithm 2: EM for GMMs

Result: Parameters of the GMM
initialization: ϵ, $p_k^{(0)}$, $\mu_k^{(0)}$, $\Sigma_k^{(0)}$;
while $l(\theta^{(t+1)}) - l(\theta^{(t)}) > \epsilon$ **do**
(E-Step);
$\forall i \in [\![1, n]\!], \forall k \in [K]$

$$Q_i(k) = \frac{g_{\mu_k^{(t)}, \Sigma_k^{(t)}}(X_i) p_k^{(t)}}{\sum_{l=1}^{K} g_{\mu_l^{(t)}, \Sigma_l^{(t)}}(X_i) p_l^{(t)} p_l^{(t)}}$$

(M-Step);
$\forall k \in [K]$

$$p_k^{(t+1)} = \frac{1}{n} \sum_{i=1}^{n} Q_i(k)$$

$$\mu_k^{(t+1)} = \frac{\sum_{i=1}^{n} Q_i(k) X_i}{\sum_{i=1}^{n} Q_i(k)}$$

$$\Sigma_k^{(t+1)} = \frac{\sum_{i=1}^{n} Q_i(k)(X_i - \mu_k^{(t+1)})(X_i - \mu_k^{(t+1)})^T}{\sum_{i=1}^{n} Q_i(k)}$$

end

On this form, there exists far parallelization schemes that allow the EM to be extremely fast. However, the quality of the optimum reached as well as the initialisation procedure is a discussed topic. It's worth noting that Optimal Transport can help to supply a better approach. [10] uses the sliced Wasserstein distance to perform inference on GMM.

We will now introduce two properties of GMMs. The later is essential for the derivation of the Mixture Wasserstein distance and the first one motivates the use of GMM in the Wasserstein space. We only sketch the proofs as [5] gave complete proofs that do not necessitate any extension.

Density The following proposition shows that mixture of Dirac are good approximator of densities in the Wasserstein space and therefore GMM share the same property.

Proposition 4 *$GMM_d(\infty) = \bigcup_{K \geq 0} GMM_d(K)$ is dense in $\mathcal{P}_p(\mathbb{R}^d)$ for the metric W_p.*

The proof uses the fact that any convex combination of Dirac masses is a special case of degenerate GMM where every covariance matrix is zero. They show that the set of convex combinations of Dirac is dense in $\mathcal{P}_p(\mathbb{R}^d)$ for the metric W_p. This part is an adaptation of the Theorem 6.16 from [17] where they show that the Wasserstein space is separable when equipped with the Wasserstein distance. To perform that, we start by noting that if a random variable $X \sim \mu$ is in $\mathcal{P}_p(\mathbb{R}^d)$ then using the dominated convergence theorem, for any $\epsilon > 0$, we can control the value of its moment integral outside a compact set K.

$$\int_{\mathbb{R}^d \setminus K} ||x||^p d_\mu(x) \leq \epsilon$$

Then, using a covering of K with disjoint balls, they explicitly construct a convex combination of Dirac masses as a transformation of X: $\phi(X)$ such that they can control the Wasserstein distance $W_p(\mu, \phi_{\#}\mu) \leq f(\epsilon) \quad \lim_{\epsilon \to 0^+} f(\epsilon) = 0$. This shows that μ can be approximated, with arbitrarily precision, by a finite combination of Dirac masses. Furthermore, it is also possible to show that the space is complete [17].

Identifiability Property The following result is fundamental to derive the discrete formulation of the Mixture Wasserstein distance.

Proposition 5 *The set of finite Gaussian Mixtures is identifiable, in the sense that two mixtures $\mu = \sum_{k=1}^{K_0} \pi_0^k \mu_0^k$, $\mu = \sum_{k=1}^{K_1} \pi_1^k \mu_1^k$, written such that all components are pairwise distinct , are*

equal if and only if $K_0 = K_1$ and we can reorder the indexes such that for all k, $\pi_0^k = \pi_1^k$, $\mu_0^k = \mu_1^k$, $\Sigma_0^k = \Sigma_1^k$.

In dimension 1, we can recover the identifiability by looking at the density of the mixtures at infinity. In $\mathbb{R}^d$, $d > 1$, it is sufficient to look at the asymptotic behaviour in all directions of $\mathbb{R}^d$ and using the fact that we have identifiability in $\mathbb{R}$.

2.4 Mixture Wasserstein Distance

We saw that the optimal coupling for the Wasserstein distance between two Gaussian distributions is also a (degenerate) Gaussian distribution. This is not the case anymore if we step up to GMM. In fact, [5] shows that except in the case where the target GMM is an affine transformation of the source GMM, or they are both mixtures of Dirac masses, then the optimal transport plan is not a GMM itself. It makes the Wasserstein distance between GMM hard to compute when using large histograms. It is the main motivation of the authors for the next definition where they modify the Wasserstein distance by forcing the couplings to stay in the GMM family of distributions.

Definition 2.5 *Mixture Wasserstein distance. If μ_0 and μ_1 are two GMM distributions*

$$MW_2^2(\mu_0, \mu_1) = \inf_{\gamma \in \Pi(\mu_0, \mu_1) \cap GMM_{2d}(\infty)} \int_{\mathbb{R}^d \times \mathbb{R}^d} ||y_0 - y_1||^2 d\gamma(y_0, y_1)$$

Let's remark first that the constraint space is not empty since if $\mu_0 = \sum_{k=1}^{K_0} \pi_0^k \mu_0^k$, $\mu_1 = \sum_{k=1}^{K_1} \pi_1^k \mu_1^k$, then the independent coupling $\mu_0 \otimes \mu_1$ is in $\text{GMM}_{2d}(\infty)$. Indeed,

$$\mu_0 \otimes \mu_1(A \times B) = \mu_0(A)\mu_1(B) = \sum_{k,l=1}^{K_0,K_1} \pi_0^k \pi_1^l \mu_0^k(A)\mu_1^l(B)$$

$\sum_{k,l=1}^{K_0,K_1} \pi_0^k \pi_1^l = 1$ and since $\mu_0^k \otimes \mu_1^l$ is the independent coupling of two Gaussian distributions it is also a Gaussian distribution. It proves that $\mu_0 \otimes \mu_1$ is also a GMM (with $K_0 \times K_1$ components). Secondly, the constraint set being smaller we have $W_2(\mu_0, \mu_1) \leq MW_2(\mu_0, \mu_1)$.

The next proposition gives a direct proof that MW_2 is a distance. The problem of finding MW_2 can be casted as a Kantorovich problem between discrete histograms. If μ_0 and μ_1 are two Gaussian Mixture distributions, then finding MW_2 is equivalent to solve a Kantorovich problem between their probability vectors with the cost matrix being the pairwise 2-Wasserstein distances between the Gaussian distributions of the mixtures.

Proposition 6 *Let $\mu_0 = \sum_{k=1}^{K_0} \pi_0^k \mu_0^k$, $\mu_1 = \sum_{k=1}^{K_1} \pi_1^k \mu_1^k$ be two Gaussian Mixtures, then,*

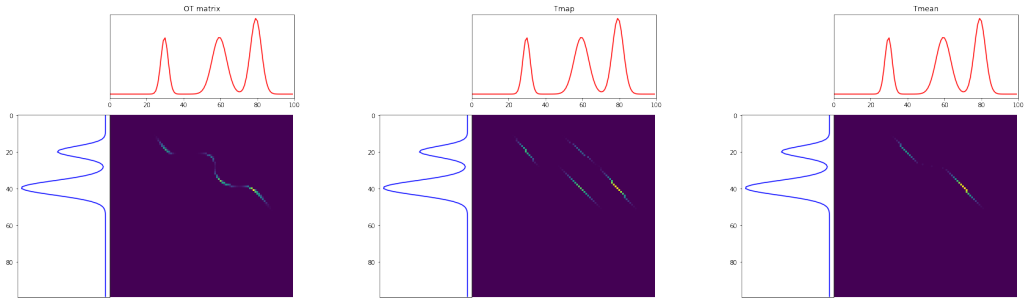
$$MW_2^2(\mu_0, \mu_1) = \min_{P \in U(\pi_0, \pi_1)} \sum_{k,l} P_{k,l} W_2^2(\mu_0^k, \mu_1^k)$$

where $U(\pi_0, \pi_1) = \left\{P \in \mathbb{R}^{K_0 \times K_1} \quad P_{i,j} \geq 0 \quad P\mathbb{1}_{K_1} = \pi_0 \quad P^T\mathbb{1}_{K_0} = \pi_1\right\}$. Moreover, if P^* is a minimizer and $T_{k,l}$ is the W_2-Monge map between μ_0^k and μ_1^l, the optimal coupling

$$\gamma^*(x, y) = \sum_{k,l} P_{k,l}^* g_{m_0^k, \Sigma_0^k}(x) \delta_{y = T_{k,l}(x)}$$

is a solution of the first formulation.

Casting the problem to this form as numerous advantages. First, it is faster to compute than using the 2-Wasserstein distance. The computation of the pairwise cost matrix could be easily parallized and then it remains only to solve a low dimensional linear program. Since it is casted as a Kantorovich problem, one can apply any developed trick from the literature to speed up computing. One idea that as not been mentioned in the paper would be to apply the Sinkhorn algorithm on the remaining linear program. Secondly, for a discrete Kantorovich problem with n source points and m target points it exists a solution that is a sparse matrix with less that $n + m - 1$ non zero elements. Thus, the optimal coupling is a Gaussian mixture with less than $K_0 + K_1 - 1$ components. Finally, it gives a direct proof that MW_2 is a distance.



(a) Monge Map (b) Mixture Coupling (c) Tmean

Figure 2: Different transport between two GMM. μ_0 as two components and μ_1 three. (*left*) Optimal transport between the two GMMs. We see that it is supported on a line which is the Monge map. (*center*) Optimal Mixture transport. It is not supported on the graph of a function. It is composed of a mixture of 4 degenerated Gaussian distributions. (*right*) Tmean transportation map.

Their results is connected to the result of [3] that have previously found an identical discrete distance. However, their continuous formulation allows to make a precise comparison to the 2-Wasserstein distance, the authors shows that W_2 and MW_2 differ by a factor that depends only on the trace of the covariance matrices of the mixtures.

$$M_2(\mu_0,\mu_1) \leq MW_2(\mu_0,\mu_1) \leq W_2(\mu_0,\mu_1) + \sum_{i=0,1} \left(2\sum_{k=1}^{K_i} \pi_i^k Tr(\Sigma_i^k)\right)^{\frac{1}{2}}$$

If we consider mixtures of Dirac masses, W_2 and MW_2 are thus the same.

Another advantage of the discrete formulation is that it generalises to the displacement interpolation. If w^* is an optimal solution of the discrete formulation then the displacement interpolation between the mixtures is

$$\mu_t = P_t \# \gamma^* = \sum_{k,l} w^*_{k,l} \mu_t^{k,l}$$

Where $\mu_t^{k,l}$ is the displacement interpolation between μ_0^k and μ_1^l as described in the last section. Since w^* has less than $K_0 + K_1 - 1$ components, μ_t is also a GMM with less than $K_0 + K_1 - 1$ components.

2.5 Transport map from MW_2

One drawback of using an approximation of the Wasserstein distance is that the obtained coupling is not of the form $(Id, T)\#\mu_0$. We thus have to define a transport that is not necessarily a Monge map. Similarly to using a Barycentric projection in an entropic regularization of OT, it is possible to define a proxy from the MW_2 distance.

$$T_{mean}(x) = \mathbb{E}_\gamma[Y|X=x]$$

Since $\gamma^*(x,y) = \sum_{k,l} P^*_{k,l} g_{m_0^k,\Sigma_0^k}(x)\delta_{y=T_{k,l}(x)}$ is the distribution of (X,Y) we see that $Y|X=x$ is a discrete variable that takes value in the set $\{T_{k,l}(x) \quad k \in [K_0] \quad l \in [K_1]\}$. We thus have,

$$T_{mean}(x) = \sum_{k,l} T_{k,l}(x) \frac{\gamma^*(x,Y)}{g_0(x)}$$

where g_0 is the density of the GMM μ_0. This is a proxy for a Monge map and it is not necessary the gradient of a convex function nor nor we have $T_{mean}\#\mu_0 = \mu_1$. This is not the only possibility but that the one we retained for the numerical experiments.

Figure 2 illustrates the difference between Tmean, the Mixture Wasserstein coupling γ^* and the Monge map T^* from the Wasserstein distance. We will use T_{mean} in the color transfer experiment since we need a way to move each pixel to a single position. The choice of the transport can lead to different behaviour for image tasks thus it is preferable to test every possibility.

2.6 Elliptical Mixtures

An interesting point that is raised in the article is that the results could be extended to other mixtures of distributions. The first thing we want is that the Wasserstein distance between the components of the mixture is easy to compute. Fortunately, [9] shows that the formula from proposition 3 is also valid if we take elements from the same elliptical distributions. Elliptical distributions are of the form,

$$\forall x \in \mathbb{R}^d \quad f_{m,\Sigma}(x) = C_{h,d,\Sigma} h((x-m)^T \Sigma^{-1}(x-m))$$

where $m \in \mathbb{R}^d$, Σ is a positive definite symmetric matrix and h is a given function from $[0, +\infty)$ to $[0, +\infty)$.

For example, taking $h(x) = \exp(-t/2)$ gives Gaussian distributions, taking $h(x) = \mathbb{1}\{x \leq 1\}$ gives uniform distributions on ellipsoids and taking $h(x) = (1 + t/v)^{-(v+d)/2}$ gives the family of t-distributions with v degrees of freedom.

Secondly, to obtain the discrete formulation of proposition 5 we need to generalize the proof of the proposition 6. For that, we need that the marginal of the components of each distribution is from the same family. For example, the marginals of a Gaussian variables are all Gaussian variables. We have shown that the Mixture Wasserstein distance was well-defined because $\Pi(\mu_0, \mu_1) \cap \text{GMM}_{2d}(\infty)$ contains $\mu_0 \otimes \mu_1$. This might not stay true if instead of $\text{GMM}_{2d}(\infty)$ we consider another family that does not share the marginal consistency property. The marginal constituency is satisfied by the t-distributions. Indeed if Z is a multivariate student with scaling parameter Σ, location parameter m and v degrees of freedom, it exists independent variables (X, Y), $X \sim \mathcal{N}_d(0, \Sigma)$ and $Y \sim \chi^2_v$ such that

$$Z = \frac{X}{\sqrt{Y/v}} + m$$

Thus if $a \in \mathbb{R}^d$,

$$\hat{Z} = a^T Z = \frac{a^T X}{\sqrt{Y/v}} + a^T m$$

and since $a^T X$ is a univariate Gaussian variable, $\hat{Z}$ is a univariate t-distribution.
We also need the identifiability property and the authors claim that using a similar proof than for Gaussian distributions, one can also prove it for Student variables. We numerically illustrate in the next section that Student Mixtures can be used easily without changing the formula of the Mixture Wasserstein distance.

2.7 Approximation

The main interest of GMM is that they can approximate any (smooth) distribution if given enough components. Thus even if working with data that are not generated by GMM, one can approximate them by a GMM and then compute the Mixture Wasserstein distance. In some applications we could need to retain a large number of components when approximating our data with GMM. It could therefore stay slow to solve the resulting linear program. One could therefore apply an entropic regularisation with the Sinkhorn algorithm to accelerate the process. Since we did not need to use large GMM in our numerical experiments we did not use this extension, but we think this should be try, besides, it is not mentioned in the article.

3 Numerical Experiments

We present two numerical experiments using the MW_2 distance. First we illustrate that we can interpolate between T-Mixtures with exactly the same method as for GMM. Secondly, we discuss the color transfer experiment and compare the Mixture Wasserstein results with the Wassertein Flow for matching that apply a gradient descent on the Sinkhorn divergence. For all experiments we adapted the code of [6] which is the Github repository for the experiments of [5]. We used the POT library that gives helper function for Optimal Transport related tasks [8] for the first experiment and the GeomLoss library for the second experiment [7]. We also refer to [3] for more experiments with MW_2.

3.1 MW_2 for T-Mixtures

As mentioned in the last section it is possible to apply the MW_2 distance between mixture of T-distributions. T-distributions are related to Gaussian distributions as a T-distribution with infinite degrees of freedom is a Gaussian distribution. The lower the degree the heavier their tails, the extreme case being a degree of freedom of 1 which gives a Cauchy distribution.

Figure 3 presents the interpolation between two GMMs (first row) and two T-Mixtures (second row). For the GMM and the T-Mixtures we chose the same mean and covariance for a comparison purpose. Each T-distribution has 3 degrees of freedom. Since we put the same moments. they share the same pairwise Wasserstein distances between their components. Thus, to compute the interpolation between the T-Mixtures, no modification of the code is needed, except for the theoretical densities.

We see that the interpolation is almost the same between the mixtures. When we compare the two center frames, it seems that the interpolation between the T-mixtures is less smooth. This is probably due to their heavier tails. In some applications if the dataset is closer to a T-Mixture, one should try to use a T-Mixture approximation instead of GMM and solve the same discrete linear program.

In dimension 2, there is almost no visual differences between the contour plot of the GMM and the T-Mixtures. We refer to the 2D experiment of the paper [5] where they interpolate between two GMM in dimension 2.

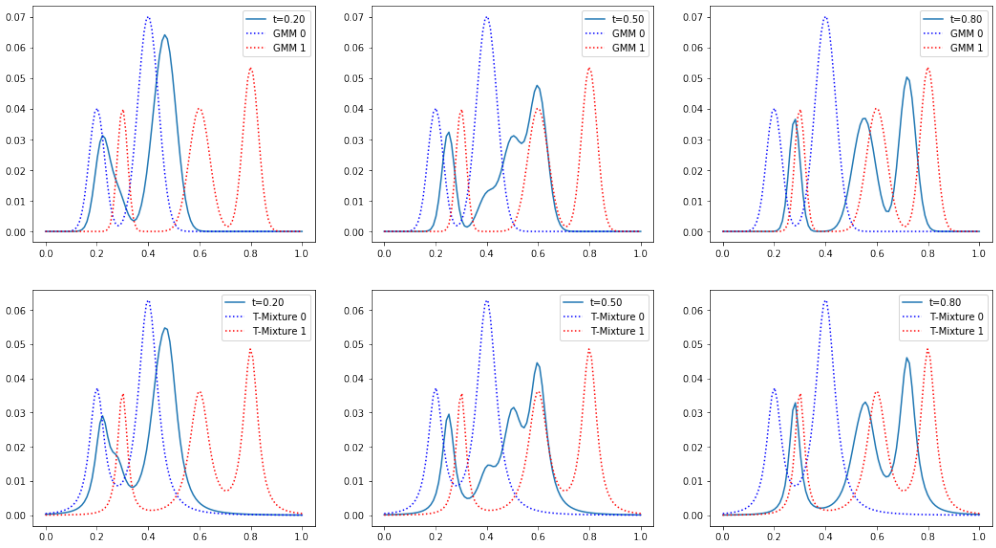


Figure 3: Interpolation between two GMM (first row) and two T-mixtures (second row). From left to right it is the interpolation for t=0.2, t=0.5, t=0.8. Gaussian distributions and T-distributions are both in the elliptical family, thus, the derivation is exactly the same.

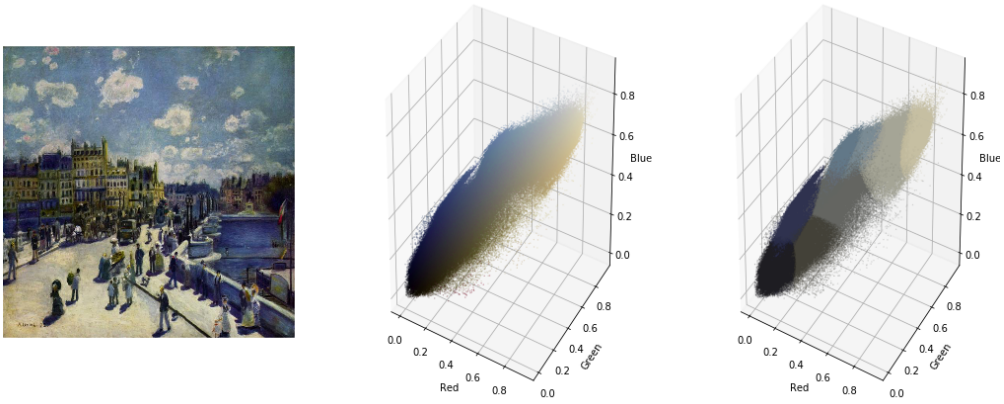


Figure 4: (*left*) Le Pont-Neuf by Auguste Renoir (1872), size=600*600 pixels, source: WikiArt. (*Middle*) Colour histogram of pixels in the RGB space. (*Right*) Colour histogram approximated by a GMM with 8 classes.

3.2 Color Transfer

We will apply the color transfer problem to demonstrate that the MW_2 distance can lead to extremely fast approximations of Optimal Transport with large cloud points.

The task is to transfer the colors of one source image to a target image while preserving structure. For this, we consider the color space of each image. Our source and target samples are clouds of 3D points, each of whom encodes the RGB color of a pixel in a standard test image. We can then define a pair of discrete probability measures on our color space $[0, 1]^3$.

$$\alpha = \frac{1}{N}\sum_{i=1}^{N} \delta_{x_i} \qquad \beta = \frac{1}{M}\sum_{i=1}^{M} \delta_{y_i}$$

As a source image, we consider Picasso's painting 1, and as a target image Renoir's painting 4. Figure 1 and 4 illustrate their distribution α (for Picasso) and β (for Renoir). Both images are of sized (600*600), thus α and β are discrete distributions of 360,000 Dirac masses. Due to the size of the cloud points it is inconceivable to solve directly the Kantorocich problem in it's simpliest form. We will consider two approaches, first we will apply the methodology of [5] with MW_2: in a first step we approximate α and β by GMM using the Expectation-Maximization algorithm and then we fit the MW_2 distance and make the transfer. Secondly, we will compare the results with the one using entropic regularization, we will apply Wasserstein Flow for matching with the gradient of the Sinkhorn Divergence.

To perform Wasserstein Flow for matching we have to minimize the following energy function:

$$\mathcal{E}(z) := S_\epsilon\left(\frac{1}{N}\sum_i \delta_{z_i}, \frac{1}{M}\sum_i \delta_{y_i}\right)$$

where z is initialized to be α and S_ϵ is the Sinkhorn divergence:

$$S_\epsilon(\alpha, \beta) = W_\epsilon(\alpha, \beta) - W_\epsilon(\alpha, \alpha)/2 - W_\epsilon(\beta, \beta)/2 \qquad W_\epsilon(\alpha, \beta) := \langle P, C\rangle - \epsilon \mathrm{KL}(P|\alpha\beta^\top)$$

C is the matrix of pairwise distance between points in $\mathbb{R}^3$, P is the coupling matrix that solves the entropic optimal transport and KL is the Kullback Leibler divergence. The major advantage of using entropic regularization is that contrary to W_2 or MW_2, W_ϵ is differentiable. Thus, we can perform gradient descent to minimize $\mathcal{E}$. Secondly, we introduced the Sinkhorn divergence because contrary to W_ϵ, S_ϵ defines a positive definite approximation of Optimal Transport Wasserstein cost.

Color Transfer with MW_2 The only parameters that we have to tweak is K_0 and K_1 the number of components we choose to approximate α and β with GMM. We always set $K = K_0 = K_1$ and

Figure 5: Source images after transport for values of K in $[3, 8, 15, 100]$. Computation time for the EM are respectively 1.72, 9.22, 22.7 and 163.8 seconds, for the computation of the linear program it is 0.099, 0.65, 2.53 and 146.47 seconds.

present the result for different values of K. All computations are done on CPU with the package POT [8] to solve the linear programm and Scikit-Learn [14] for the GMM approximations.

Figure 5 shows the resulting images after transport for different values of K. We see that $K = 8$ is already enough for a visually convincing result, there is almost no visual difference between the color transport for $K = 8, 15, 100$. On figure 6, we show the target histogram and the histogram of the source after transport for $K = 3, 8, 100$. This is more helpful to evaluate the quality as we see how close is $T_{mean}\#(\alpha)$ to β. For $K = 3$, we see a common artefact that was not mention in the paper: if α, β takes support in a compact set (here $[0, 1]^3$) there is no guarantee that $T_{mean}\#(\alpha)$ lies in the same compact set because there is no guarantee that $T_{mean}\#(\alpha) = \beta$. Thus, we had to clip the values between $[0, 1]^3$ and that is why we see an accumulation of points on the border of the histogram for $K = 3$.

One last remark is that when the number of classes we choose grows, both the time for the EM and to solve the linear program increase. For $K = 8$ the total time for the method is less than 10 seconds but for $K = 100$ it takes around 5 minutes. For the EM part, there not much do do except trying a better implementation or another approximation method such as [10]. For the linear programming part, we could use an entropic regularization and Sinkhorn algorithm to solve the linear program faster. We did not do it since $K = 100$ is already more than enough for the task and the 100*100 linear program is solved under three minutes.

Color Transfer with S_ϵ Due to the size of the cloud points it is not possible to solve the Wasserstein Flow Matching on CPU. However, GeomLoss [7] supply helper functions that allow to run the gradient descent of the Sinkhorn divergence on GPU with fast and memory efficient algorithms.

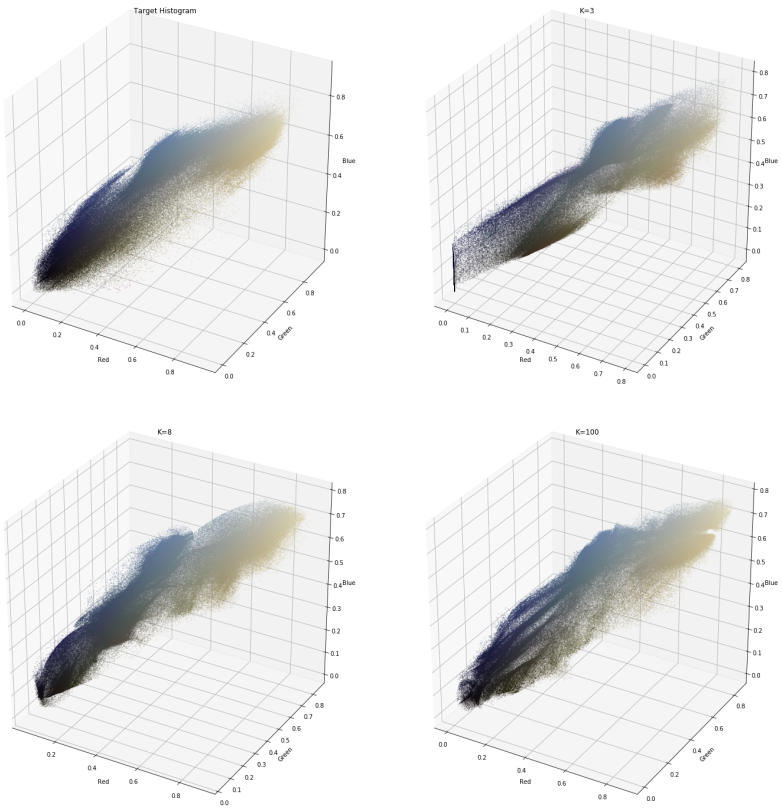


Figure 6: (*top left*) Target Histogram, (*top right*), (*bottom left*), (*bottom right*) respectively the source histograms after transport with the MW_2 distance for $K = 3, 8, 100$.

Figure 7 presents the results for different values of ϵ. ϵ controls how well we approximate the Optimal Transport Wasserstein problem. The closer ϵ is to 0 the more precise the approximation is but the slower runs the algorithm. For each ϵ we run 10 iterations of gradient descent. For $\epsilon = 1$, the approximation is too brutal and the color are still too close to the original image. For $\epsilon = 0.05$ we get a really accurate approximation. The computation time for $\epsilon = 1, 0.1, 0.01$ are respectively 0.21, 0.22 and 6.22 seconds per iteration which give a total time of 2.1, 2.2 and 62.2 seconds.

Figure 8 presents the resulting images after running gradient descent with $\epsilon = 0.05$. We run 10 iterations of the gradient descent but wee see that the results are already convincing after one iteration.

Conclusion To conclude this experiment, both approach work well with large cloud points in reasonable time. We observe that the approximation with MW_2 is a bit less precise but run well on CPU. For a more precise approximation if one has access to a GPU, it might be preferable to use a gradient descent on the Sinkhorn divergence. Nonetheless, both approaches give the same visually convincing results.

We choose to work on painting because there is less challenge than with realistic images. For accurate colour transfer on high resolution images we need to add other enhancements. [13] devoted a thesis on the problematic of Optimal Transport for image processing and give example of tricks that work well on colour transfer.

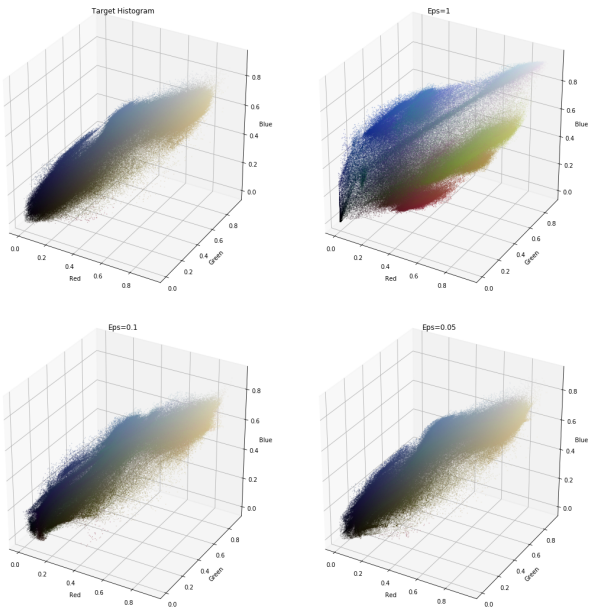


Figure 7: (*top left*) Target Histogram, (*top right*), (*bottom left*), (*bottom right*) respectively the source histograms after transport with Wasserstein Flow Matching for $\epsilon = 1, 0.1, 0.01$ and 10 iterations of the gradient descent.

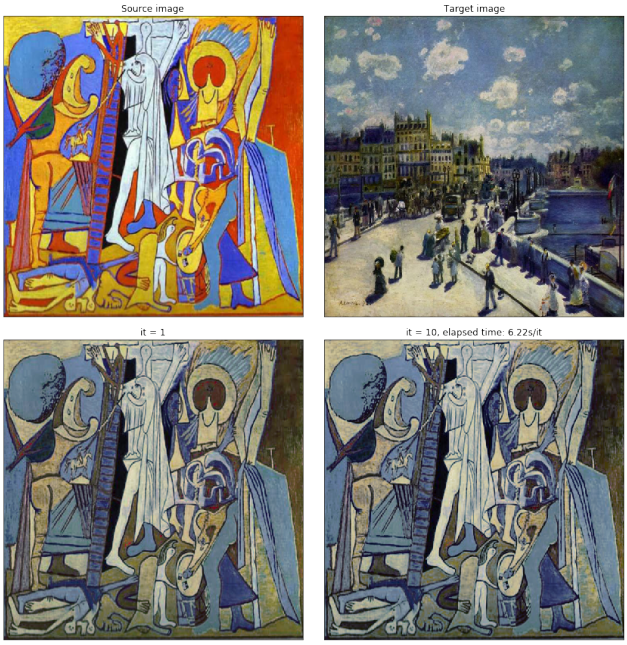


Figure 8: (*top left*) Source image, (*top right*) Target image, (*bottom left*) transported source image after one iteration of the gradient descent, (*bottom right*) transported source image after ten iterations of the gradient descent.

4 Conclusion and perspectives

We reviewed the main results of [5] that gives a rigorous continuous formulation of the discrete distance of [3]. We have shown by taking advantage of the explicit OT formulation between Gaussian distributions that restricting the Wasserstein distance on the set of Gaussian Mixtures allows to derive a discrete approximation of the Wasserstein distance that is fast to compute. We also demonstrate with numerical results that the Mixture Wasserstein distance can be extended to more general families such as the T-Mixtures. Secondly, we compared the results of the method with the entropic regularization approach on the computationally demanding problem of color transfer. We have shown that it lead to similar results in quality, however, the later is more demanding as it necessitates to use a GPU. We see this as the main advantage of MW_2, its computation is so fast that it only requires a CPU.

References

[1] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin, Heidelberg: Springer-Verlag, 2006. ISBN: 0387310738.

[2] Yann Brenier. "Polar factorization and monotone rearrangement of vector-valued functions". In: *Communications on pure and applied mathematics* 44.4 (1991), pp. 375–417.

[3] Yongxin Chen, Tryphon Georgiou, and Allen Tannenbaum. "Optimal transport for Gaussian mixture models". In: *IEEE Access* PP (Oct. 2017). DOI: 10.1109/ACCESS.2018.2889838.

[4] Marco Cuturi. "Sinkhorn Distances: Lightspeed Computation of Optimal Transportation Distances". In: *Advances in Neural Information Processing Systems* 26 (June 2013).

[5] Julie Delon and Agnes Desolneux. "A Wasserstein-type distance in the space of Gaussian Mixture Models". In: (July 2019).

[6] Julie Delon and Agnes Desolneux. *gmmot*. https://github.com/judelo/gmmot. 2019.

[7] Jean Feydy et al. "Interpolating between Optimal Transport and MMD using Sinkhorn Divergences". In: *The 22nd International Conference on Artificial Intelligence and Statistics*. 2019, pp. 2681–2690.

[8] R'emi Flamary and Nicolas Courty. *POT Python Optimal Transport library*. 2017. URL: https://github.com/rflamary/POT.

[9] Matthias Gelbrich. "On a formula for the L2 Wasserstein metric between measures on Euclidean and Hilbert spaces". In: *Mathematische Nachrichten* 147.1 (1990), pp. 185–203.

[10] Soheil Kolouri, Gustavo K Rohde, and Heiko Hoffmann. "Sliced wasserstein distance for learning gaussian mixture models". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 3427–3436.

[11] G. J. McLachlan and D. Peel. *Finite mixture models*. New York: Wiley Series in Probability and Statistics, 2000.

[12] Andrew Ng. *Lecture notes in Machine Learning CS229*. 2019.

[13] Nicolas Papadakis. "Transport Optimal pour le Traitement d'Images". Habilitation à diriger des recherches. Université de Bordeaux, Dec. 2015. URL: https://hal.archives-ouvertes.fr/tel-01246096.

[14] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[15] Gabriel Peyré. *Course Notes on Computational Optimal Transport*. 2019.

[16] Gabriel Peyré and Marco Cuturi. "Computational Optimal Transport". In: *Foundations and Trends in Machine Learning* 11 (2018), pp. 355–607.

[17] C Villani. "Optimal transport – Old and new". In: vol. 338. Jan. 2008, pp. xxii+973. DOI: 10.1007/978-3-540-71050-9.