International Journal of Pattern Recognition and Artificial Intelligence
World Scientific Publishing Company

Optimization of the cost function in the Monge-Kantorovich Problem (MKP) under the Monge condition

Laurent Oudre
Institut Télécom, TELECOM ParisTech, CNRS/LTCI
37-39 rue Dareau - 75014 Paris, France
laurent.oudre@gmail.com

This article presents a method for adapting the cost function in the Monge-Kantorovich Problem (MKP) to a classification task. More specifically, we introduce a criterion that allows to learn a cost function which tends to produce large *distance* values for elements belonging to different classes and small *distance* values for elements belonging to the same class. Under some additional constraints (one of them being the well-known Monge condition), we show that the optimization of this criterion writes as a linear programming problem. Experimental results on synthetic data show that the output optimal cost function provides good retrieval performances in presence of two types of perturbations commonly found in histograms. When compared to a set of various commonly used cost functions, our optimal cost function performs as good as the best cost function of the set, which shows that it can adapt well to the task. Promising results are also obtained on real data for two-class image retrieval based on grayscale intensity histograms.

Keywords: Classification; Wasserstein distance; pattern recognition; machine learning; instrument classification

1. Introduction

The Monge-Kantorovich mass transportation problem (MKP) has been given much interest for the last past years because of the large number of mathematical implications raised by this problem but also because of the wide scope of possible applications (statistics, image processing, economics, fluid mechanics, dynamical systems...) [1,2].

Given two probability spaces (X, μ) and (Y, ν) and a non-negative measurable cost function $c : X \times Y \to \mathbb{R}^+$, the Monge-Kantorovich mass transportation problem consists in finding the optimal transportation cost:

$$MK_c(\mu, \nu) = \inf_{\pi \in \Pi(\mu,\nu)} \int_{X \times Y} c(x, y)\, d\pi(x, y) \tag{1}$$

where $\Pi(\mu, \nu)$ is the set of all probability measures on $X \times Y$ with marginals μ on X and ν on Y. The elements of $\Pi(\mu, \nu)$ are called *transportation plans*: if $\pi^* \in \Pi(\mu, \nu)$ attains the above minimum, we say that π^* determines an *optimal transportation plan* with respect to c for μ and ν. The mathematical properties

and interpretations of these optimal transportation plans and costs (according to the chosen cost function c or to the structure of measure spaces X and Y) are beyond the scope of this article and are described in [1,2]. Intuitively, the quantity $MK_c(\mu, \nu)$ can be interpreted as a *distance* between μ and ν (the use of the word *distance* is improper as this quantity does not *in general* satisfy all the axioms of a distance). Yet, if $X = Y$ and if the cost function can be written as $c(x, y) = d(x, y)^p$ with d a distance on X and $p \in [1, \infty[$ the quantity:

$$W_p(\mu, \nu) = \left(\inf_{\pi \in \Pi(\mu,\nu)} \int_{X \times X} d(x, y)^p \, d\pi(x, y) \right)^{\frac{1}{p}} \tag{2}$$

satisfies the axioms of a distance and is called the *Wasserstein distance* of order p.

In particular, the Wasserstein distance W_1 defined with the L1-norm $d(x, y) = |x - y|$ (which is commonly simply called *Wasserstein distance*) has been used for classification, clustering, computer vision, segmentation and pattern recognition [3,4,5] of different types of data such as images [6,7,8] or sounds [9,10] because of its nice interpretation in terms of mass transportation. In particular, the Wasserstein distance has proven its utility when dealing with histograms (which can be seen as discrete probability distributions when normalized). Yet, most of the time the use of the L1-norm as a cost function only constitutes a default choice which has empirically proved to provide good performances on a large range of problems. Nevertheless, given a classification or retrieval task, this choice in cost function might not be optimal as the L1-norm is non-informative and inherently assumes that all points of X have the same role.

This article investigates the possibility of adapting the choice of the cost function c to a considered two-class classification problem. More specifically, our aim is to define a criterion depending on some MK_c quantities calculated on training data and to optimize it so as to compute an optimal cost function c for this classification problem. For sake of simplicity and clarity, the space X we consider in this article is discrete and composed of N-bins histograms ($N \in \mathbb{N}^*$):

$$X = \left\{ f \in \mathbb{R}^N ; \forall i \in \ 1, \ldots, N \quad f_i \geq 0 \text{ and } \sum_{i=1}^{N} f_i = 1 \right\} \tag{3}$$

In this discrete framework, and under some assumptions on the cost function (among which the Monge condition), it will be shown that this complex and unsolvable criterion optimization problem becomes a linear programing problem. Through this procedure, the output optimal Monge cost is therefore learned so as to adapt to a given task.

This article is organized as follows. Section 2 introduces the criterion used for the evaluation of the optimal cost function and the associated constraints that enable to optimize it. Section 3 describes our experimental protocol and the data used for testing. Sections 4 & 5 present and discuss the results we obtain with our optimal

Monge costs on respectively synthetic and real data and compare their performances to those obtained with other well-known and commonly used distances and cost functions.

2. Method

2.1. *Discrete framework*

Consider two real-valued, positive and normalized histograms f and g of size N. These histograms can be seen as discrete probability distributions, and thus the discrete MKP (1) can be written:

$$MK_c(f,g) = \min_{\substack{\pi_{i,j}\geq 0 \\ \sum_j \pi_{i,j}=f_i \\ \sum_i \pi_{i,j}=g_j}} \sum_{i,j} c(i,j)\ \pi_{i,j}. \tag{4}$$

Given an optimal transportation plan $\pi^*_{i,j}$ with respect to c for f and g this quantity writes as a simple linear sum:

$$MK_c(f,g) = \sum_{i,j} c(i,j)\ \pi^*_{i,j}. \tag{5}$$

Our aim is to use this quantity –that shall be called *distance* in the following despite the fact that it does not *in general* constitute a proper distance– in order to perform a two-class histogram classification.

2.2. *Computation of the optimal cost*

Consider a finite training database composed of N-bins histograms. Each of these elements belongs either to class $\mathcal{A}$ or class $\mathcal{B}$. Our aim is to find a non-negative measurable cost function c and an associated *distance* MK_c which insure that:

- Two elements belonging to different classes are *far away* according to MK_c
- Two elements belonging to the same class are *close* according to MK_c

In order to perform this task, we propose to maximize the following criterion with respect to c:

$$\lambda_c = \alpha\lambda_c^{inter} - (1-\alpha)\lambda_c^{intra} \tag{6}$$

where $\alpha \in [0,1]$ and the terms λ_c^{inter} and λ_c^{intra} respectively write:

$$\lambda_c^{inter} = \frac{1}{|\mathcal{A}||\mathcal{B}|} \sum_{a\in\mathcal{A}} \sum_{b\in\mathcal{B}} MK_c(a,b) \tag{7}$$

and

$$\lambda_c^{intra} = \frac{1}{2|\mathcal{A}|^2} \sum_{a\in\mathcal{A}} \sum_{a'\in\mathcal{A}} MK_c(a,a') + \frac{1}{2|\mathcal{B}|^2} \sum_{b\in\mathcal{B}} \sum_{b'\in\mathcal{B}} MK_c(b,b') \tag{8}$$

and stand for the average interclass and intraclass *distance*.

The optimal cost c^* is chosen such that:

$$c^* = \underset{c}{\operatorname{argmax}}\ \lambda_c \tag{9}$$

In (6), α is a weight parameter which controls the relative influence of the interclass and intraclass *distances*. When $\alpha = 0$ we favour the coherence of the classes while when $\alpha = 1$ we seek for the separability of the classes. Note that our criterion is based on the same principle as many statistical based discrimination criteria such as the Fisher criterion or the maximum margin criterion, i.e. maximizing interclass and minimizing intra-class. Yet, there are two major differences between our criterion and those used in the state-of-the-art:

- Our criterion relies only the quantities MK_c and not on data-related parameters (such as means, variances or barycenters). Indeed, classical criteria are based on well-defined distances, which is unfortunately not our case. In particular, according to the cost function c, the calculation of the means, variances or data-related parameters in the sense of the so-defined *distance* MK_c is *a priori* not tractable. For example, the definition of the barycenter, even when the cost function is trivial (e.g. $c(i,j) = |i-j|$) is still a tricky problem [6].
- Our criterion writes as a linear sum of the quantities MK_c. This point is crucial for the tractability and the solvability of our complex optimization problem. In particular, due to the linear optimization performed within our method (see Section 2.3), the use of margins would have made the problem unsolvable.

2.3. *Monge condition*

Unfortunately, the optimization of criterion (9) is not feasible in the general case, as the optimal transportation plan $\pi^*_{i,j}$ used in the computation of (5) depends on the two histograms f and g but also on the used cost function c. This is a non-convex problem, that can be addressed for instance using alternating optimization w.r.t π and c, but in this case, there is no guaranty of convergence to a fixed point. In order to numerically optimize expression (9), we need to add some constraints on the cost function c.

The cost function $c : \mathbb{N}^* \times \mathbb{N}^* \to \mathbb{R}^+$ satisfies the *Monge condition* if:

$$\forall (i,j) \in \mathbb{N}^* \times \mathbb{N}^* \quad c(i,j) + c(i+1,j+1) \leq c(i+1,j) + c(i,j+1). \tag{10}$$

Note that this condition is somehow interpretable as a *convexity* property. More generally, it can be proved that every function $c(i,j) = \Phi(|f(i) - f(j)|)$ with Φ being a convex and increasing function on $\mathbb{R}^+$ and f being a monotonous function on $\mathbb{R}^+$ satisfies this condition.

Interestingly, if the cost function c satisfies the Monge condition then according to [1,11]:

- There exists a simple procedure (called the *Northwest Corner Rule*) to find an optimal solution to the transportation problem described in (4) (This procedure actually boils down to histogram specification and LookUp Table techniques with linear time complexity).
- The optimal transportation plan $\pi_{i,j}^{f\to g}$ output by this procedure does not depend on the used cost function c.

Note that these results are particularly interesting for our problem since, as will be seen, the independence of $\pi_{i,j}^{f\to g}$ with respect to c transforms our complex and intractable optimization problem into a simple and solvable linear programming problem.

Given a cost function c satisfying the Monge condition, the expression of the criterion (6) now rewrites:

$$\lambda_c = \sum_{i,j} c(i,j)\left[\alpha\pi_{i,j}^{inter} - (1-\alpha)\pi_{i,j}^{intra}\right] \tag{11}$$

where $\pi_{i,j}^{inter}$ and $\pi_{i,j}^{intra}$ are defined as:

$$\pi_{i,j}^{inter} = \frac{1}{|\mathcal{A}||\mathcal{B}|}\sum_{a\in\mathcal{A}}\sum_{b\in\mathcal{B}}\pi_{i,j}^{a\to b} \tag{12}$$

and

$$\pi_{i,j}^{intra} = \frac{1}{2|\mathcal{A}|^2}\sum_{a\in\mathcal{A}}\sum_{a'\in\mathcal{A}}\pi_{i,j}^{a\to a'} + \frac{1}{2|\mathcal{B}|^2}\sum_{b\in\mathcal{B}}\sum_{b'\in\mathcal{B}}\pi_{i,j}^{b\to b'} \tag{13}$$

and can respectively be interpreted as the global transportation from $\mathcal{A}$ on $\mathcal{B}$ and as the sum of the global transportations from $\mathcal{A}$ to $\mathcal{A}$ and from $\mathcal{B}$ to $\mathcal{B}$.

The simplified version of criterion λ_c in (11) now linearly depends on c, and thus shall be maximized through classical linear programming methods under the Monge and the non-negativity constraints.

2.4. *Additional constraints*

We have already discussed the issues raised by the use of MK_c as a distance, but we can force this quantity to behave *almost* like a distance by adding some constraints on c within the linear programming problem. Namely, we can make MK_c satisfy two out of the three axioms of a distance by applying the following constraints:

- *identity of indiscernibles:* $\forall i \in \mathbb{N}^* \;\; c(i,i) = 0$
- *symmetry:* $\forall (i,j) \in \mathbb{N}^* \times \mathbb{N}^* \;\; c(i,j) = c(j,i)$

Also, without loss of generality and for sake of reproducibility, we can constrain that $\sum_{i,j} c(i,j) = 1$, which allows to balance the expression (5) (as we know that by definition $\sum_{i,j}\pi^*(i,j) = 1$) and insures that $0 \leq MK_c(f,g) \leq 1$.

2.5. *Summary of the algorithm*

Consider that we have a training database composed of N_{train} histograms for each class $\mathcal{A}$ and $\mathcal{B}$, the computation of the optimal cost c^* is performed as follows:

(1) Calculate all the $3 \times N_{train}^2$ optimal transport plans $\pi_{i,j}^{f \to g}$ between all pairs of elements in class $\mathcal{A}$ and $\mathcal{B}$ thanks to the Northwest Corner Rule [1] ($\mathcal{A} \to \mathcal{A}$, $\mathcal{A} \to \mathcal{B}$, $\mathcal{B} \to \mathcal{B}$).
(2) Calculate $\pi_{i,j}^{inter}$ and $\pi_{i,j}^{intra}$ as defined in (12) and (13).
(3) Given a value for α, optimize the linear criterion λ_c in (11) w.r.t c thanks to a large-scale optimization method[a]:

$$c^* = \underset{\substack{c(i,j)+c(i+1,j+1) \leq c(i+1,j)+c(i,j+1) \\ c(i,i)=0 \\ c(i,j)=c(j,i) \\ \sum_{i,j} c(i,j)=1}}{\operatorname{argmax}} \sum_{i,j} c(i,j) \left[\alpha \pi_{i,j}^{inter} - (1-\alpha)\pi_{i,j}^{intra}\right] \quad (14)$$

Due to the symmetry and the fact that the diagonal is equal to zeros, we only have $\frac{N(N-1)}{2}$ coefficients to compute. The normalization constraint writes as a $1 \times \frac{N(N-1)}{2}$ equality constraint in the matrix form and the non-negativity and the Monge constraints write as a $\left(\frac{N(N-1)}{2} - 1\right)^2 \times \frac{N(N-1)}{2}$ inequality constraint in the matrix form.

3. Evaluation: two-class histogram retrieval

Let us test the performances of optimal Monge cost functions on a basic two-class retrieval task. Consider a database Ω, composed of two classes $\mathcal{A}$ and $\mathcal{B}$, each of them containing respectively $|\mathcal{A}|$ and $|\mathcal{B}|$ histograms ($|\Omega| = |\mathcal{A}| + |\mathcal{B}|$). Each histogram $h^{(n)} \in \Omega$ has a binary label $l(h^{(n)})$. By convention, we state that $l(h) = 0$ if $h \in \mathcal{A}$ and $l(h) = 1$ if $h \in \mathcal{B}$. The retrieval task is performed as follows:

- We randomly select N_{train} histograms in each class $|\mathcal{A}|$ and $|\mathcal{B}|$, that will constitute our training database Ω^{train} ($|\Omega^{train}| = 2N_{train}$).
- We evaluate the optimal Monge cost function c^* on Ω^{train} through the process described in Section 2.5.
- We randomly select N_{test} histograms in each class $|\mathcal{A}|$ and $|\mathcal{B}|$, that will constitute our test database Ω^{test} ($|\Omega^{test}| = 2N_{test}$), making sure that these histograms do not belong to Ω^{train}.
- We calculate the *distances* MK_{c^*} between the histograms of Ω^{test} and all the histograms of Ω thanks to expression (5) and to the cost function c^* (note that this computation is tractable and low time-consuming since c^* satisfies the Monge condition and that an optimal transportation plan can easily be computed by the Northwest Corner algorithm).

[a]We used the Matlab function `linprog` within the Optimization Toolbox for our tests.

- Then, for histogram $h^{(n)} \in \Omega^{test}$ and for $m \in [\![1, |\Omega|]\!]$, we define $\omega^{(n)}(m)$ as the set containing the m histograms closest to histogram $h^{(n)}$ according to *distance* MK_{c^*}. We define the precision and the recall respectively as:

$$p_m^{(n)} = \frac{\#\left\{h \in \omega^{(n)}(m); l(h) = l(h^{(n)})\right\}}{m} \tag{15}$$

and

$$r_m^{(n)} = \frac{\#\left\{h \in \omega^{(n)}(m); l(h) = l(h^{(n)})\right\}}{\#\left\{h \in \Omega; l(h) = l(h^{(n)})\right\}} \tag{16}$$

- Finally, for $m \in [\![1, |\Omega|]\!]$, we define the average precision, recall and F-measure as:

$$p_m^{av} = \frac{1}{2N_{test}} \sum_{n=1}^{2N_{test}} p_m^{(n)} \qquad r_m^{av} = \frac{1}{2N_{test}} \sum_{n=1}^{2N_{test}} r_m^{(n)} \tag{17}$$

$$f_m^{av} = \frac{1}{2N_{test}} \sum_{n=1}^{2N_{test}} \frac{2\, p_m^{(n)} r_m^{(n)}}{p_m^{(n)} + r_m^{(n)}} \tag{18}$$

Note that this retrieval task is not actually directly linked to the criterion used in Section 2. The hypothesis tested with this experiment is that the optimal cost c^* tends to give large *distance* values for elements belonging to different classes and small *distance* values for elements belonging to the same class. Threfore, if the optimal cost c^* satisfies this hypothesis, for $h^{(n)} \in \mathcal{A}$ the set $\omega^{(n)}(|\mathcal{A}|)$ should contain all the histograms belonging to class $\mathcal{A}$, thus giving large precision, recall and F-measure.

4. Results on synthetic data

4.1. *Synthetic data*

The synthetic data used for our tests is largely inspired by Rabin et al[12]. Each class corresponds to a set of realizations of a mixture of two Gaussians whose parameters depend on the class. For each class, a set of 5 parameters are defined: w is the relative weight of the two mixture components and (μ_1, σ_1) and (μ_2, σ_2) are respectively the parameters of component 1 & 2 (see Figure 1). For each class, the histograms are created by generating $N_{sample} = 1000$ samples of the Gaussian mixture in $[0, 1]$ (according to the parameters of the class) and computing the associated histogram quantized on $N_{bin} = 100$ bins. This histogram is then normalized so as it sums to 1.

In our simulations, the database Ω is composed of $|\Omega|$=600 histograms (300 for each class), and we have $N_{train} = 50$ and $N_{test} = 20$ (see Section 3 for the definition of these terms).

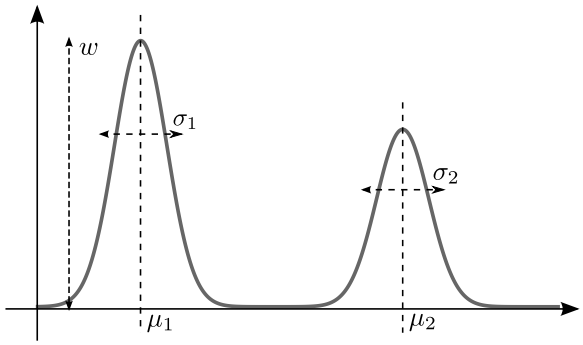

Fig. 1. The Gaussian mixture depends on 5 parameters: means μ_1 and μ_2, standard deviations σ_1 and σ_2, and a weight parameter w.

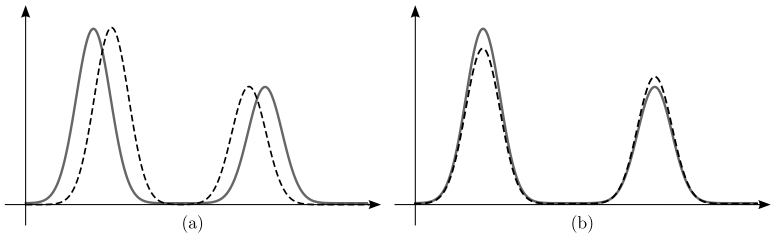

Fig. 2. Two types of perturbations: (a) Histogram shift (b) Histogram weight variability

4.2. *Tests*

For our simulations, we investigate two types of perturbations which are frequently observed in histograms (see Figure 2): the intraclass shift variability (slight changes in mean within the class) and the intraclass weight variability (slight changes in weight within the class). Our aim is twofold: to study the behavior and the robustness of our optimal costs for different choices of parameter α in (11) and to compare them to some usual cost functions.

4.2.1. *Task 1: Histogram shift*

In this context, the Gaussian mixtures of class $\mathcal{A}$ and $\mathcal{B}$ have the same means but different weights. For each histogram, a small perturbation is applied on the means so as to introduce intraclass shift variability. Namely, we set $w^A = 0.6$, $w^B = 0.8$, $\sigma_1^A = \sigma_2^A = \sigma_1^B = \sigma_2^B = 0.05$, $\mu_1^A = \mu_1^B = 0.2 + \epsilon$ and $\mu_2^A = \mu_2^B = 0.7 + \epsilon$ when ϵ is a random variable uniformly drawn in $[-0.1, 0.1]$.

Let us first investigate on the influence of parameter α in (11), i.e. how the taking into account of the interclass and/or intraclass contributions can affect the shape of optimal cost c^* and the retrieval scores. Transportation plans $\pi_{i,j}^{inter}$ and $\pi_{i,j}^{intra}$ in (11) (which can schematically be interpreted as the average interclass and

intraclass transportations plans) are displayed on the top of Figure 3. The intraclass transportation plan $\pi_{i,j}^{intra}$ (Figure 3(a)) clearly shows the variations introduced in the means: the two spots centered at bins 20 and 70 (which corresponds to values 0.2 and 0.7) are spread. Intuitively, the anti-diagonal width (top-right towards bottom-left) of these spots corresponds to the introduced variation in mean (around 20 bins i.e 0.2) and the diagonal width (top-left towards bottom-right) corresponds to the width of the support of the Gaussian. Also, one spot is darker (and thus has larger values): it is due to the larger weights given to the first mixture component (0.6 and 0.8). The interclass transportation plan $\pi_{i,j}^{inter}$ (Figure 3(b)) shows four spots which are likely to represent the four possible transports between the two mixtures components of the two classes.

The optimal costs obtained for $\alpha = 0, 0.5$ and 1 are presented on the top of Figure 4: interestingly, the shape of the obtained cost functions seems to strongly depends on parameter α. For $\alpha = 0$ the cost has small values around the diagonal so as not to take into account the intraclass variability (which according to Figure 3(a) is spread around the diagonal). For $\alpha = 0.5$ and $\alpha = 1$, the costs have a binary structure. Properly speaking, for $\alpha = 0.5$, the cost is only non-null if a bin in $1, 36$ is moved to $57, 100$ (and reciprocally), that is to say if a bin in the first component of the mixture is moved to the second one. Considering two histograms belonging to the same class, in theory no transports are needed since they have the same means and the same mixture weights (both mixture component will not move): the transportation cost between them is therefore approximately null. On the contrary, considering two histograms belonging to different classes (and thus having the same means but different mixture weights), a transport between the first mixture components and the second mixture components will occur (intuitively, the extra or missing 0.2 weight of the first component will move to the other component) and thus the transportation cost will be non-null. In consequence, this binary rule enables to detect the changes in mixture weights. For $\alpha = 1$, this binary rule is changed: it now considers segments $1, 19$ and $20, 100$. The limit between the two segments lies around 0.2 i.e. on the first component of the mixture. This is probably due to the high values associated to the corresponding spot in Figure 3(b) which mislead the optimization of our cost by assuming that all relevant information is located on the first mixture component.

By calculating the maximum value of the averaged F-measure defined in (18) for different values of α, we observe the same phenomenon: when α is too large, the optimal cost function only considers the first mixture component and thus misses some relevant information. The value $\alpha = 0.5$ gives the best results for this task (maximum value of the F-measure equal to 99.94%), which tends to show that both terms of criterion (11) are useful.

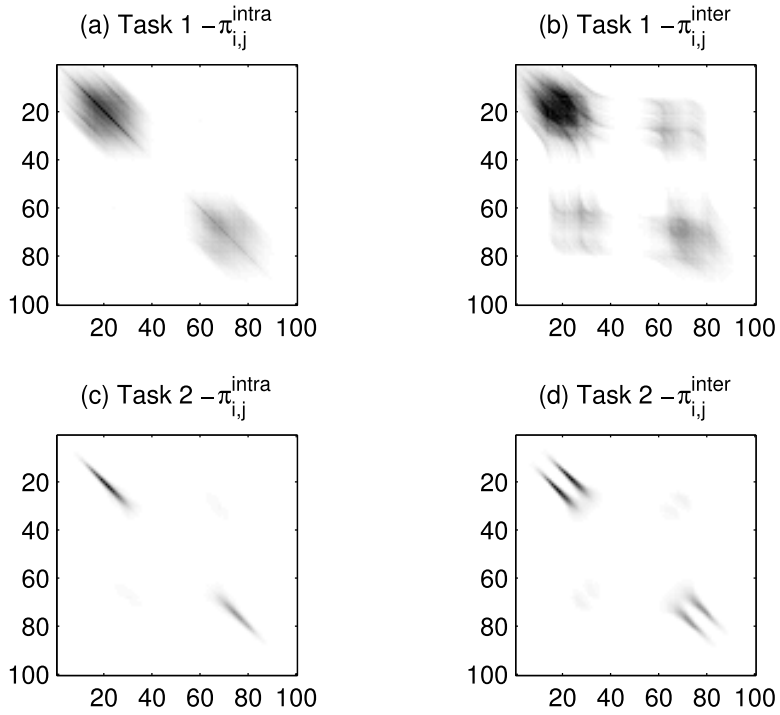

Fig. 3. Optimal Monge transportations plans $\pi_{i,j}^{inter}$ and $\pi_{i,j}^{intra}$ (see (11)) obtained for Tasks 1 & 2.

4.2.2. *Task 2: Histogram weight variability*

In this context, the two Gaussian mixtures have the same weights but different means. For each histogram, a small perturbation is applied on the mixture weight so as to introduce intraclass weight variability. Namely, we set $\sigma_1^A = \sigma_2^A = \sigma_1^B = \sigma_2^B = 0.05$, $\mu_1^A = 0.25$, $\mu_1^B = 0.2$, $\mu_2^A = 0.75$, $\mu_2^B = 0.7$ and $w^A = w^B = 0.6 + \epsilon$, when ϵ is a random variable uniformly drawn in $[-0.1, 0.1]$.

On the bottom of Figure 3 are displayed the interclass and intraclass transportation plans. While the intraclass transportation plan of Task 2 (Figure 3(c)) has the same structure than the one of Task 1 (Figure 3(a)) (except for the variations in mean which no longer occur), the interclass ones (Figure 3(b) & 3(d)) strongly differ: we can clearly see the two different sets of means (around bins 20/25 and 70/75). Note that only the anti-diagonal width of the spots has changed, as the standard deviations and thus the width of the support are still the same.

The optimal costs can be seen on the bottom of Figure 4: for $\alpha = 0$ we observe the same phenomenon than on Task 1 (low costs on the diagonal). For $\alpha = 0.5$ and $\alpha = 1$, we recognize the same binary structure than on Task 1 except than the transports with non-null costs are now those from $1, 21$ to $22, 100$. The interpretation is still the same: schematically, these costs only consider the first

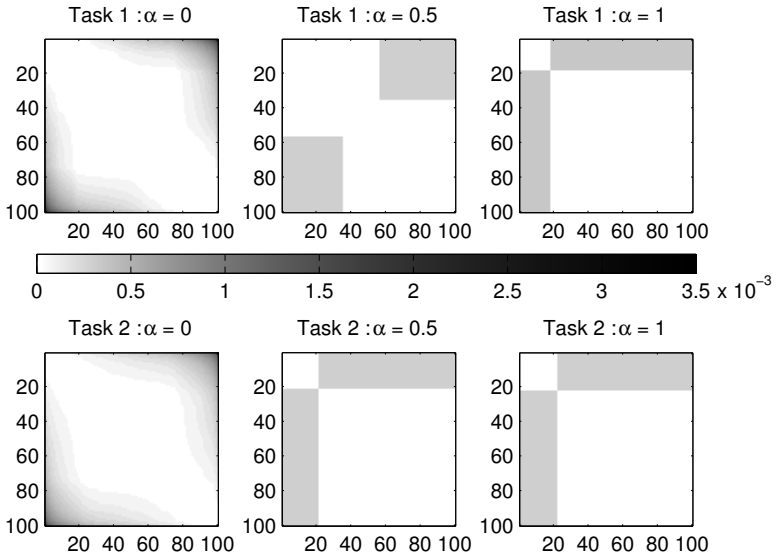

Fig. 4. Optimal costs c^* obtained for Tasks 1 & 2 and different values of α

component, assuming that if a transport is needed between the area associated to class $\mathcal{A}$ (mean ≥ 0.22) and the one of class $\mathcal{B}$ (mean ≤ 0.21), then the histograms belong to different classes. Note that this binary rule adapts well to our task as it is possible to separate the histograms by only comparing the mean of their first component.

Numerical results confirm the intuition than only considering the first component is sufficient for this retrieval task. Note that in this case, α has only a small influence on the scores (the maximum value of the F-measure is approximately equal to 99.90% for all α).

4.3. *Influence of N_{train} and α*

In the results previously displayed, the size of the training database for each class was $N_{train} = 50$. In order to investigate how crucial this parameter is, we propose to launch the exact same experiment (same database, same test data) but using less histograms for training. Figure 5 shows the maximum value of the averaged F-measure for each task, and different values of N_{train} and α.

For Task 1, we see that for $\alpha > 0.5$, performances are poor no matter the size of the training database (for large values of α the optimal cost function only considers the first mixture component, which is not enough to accurately perform this task). For $\alpha = 0$, the performances fluctuate with the size of the training database until $N_{train} = 20$, which is probably due to the fact that since only the intra-class information is available, the training process is influenced by the perturbations and even a few number of pathological cases can affect the training. For $\alpha = 0.25$ and

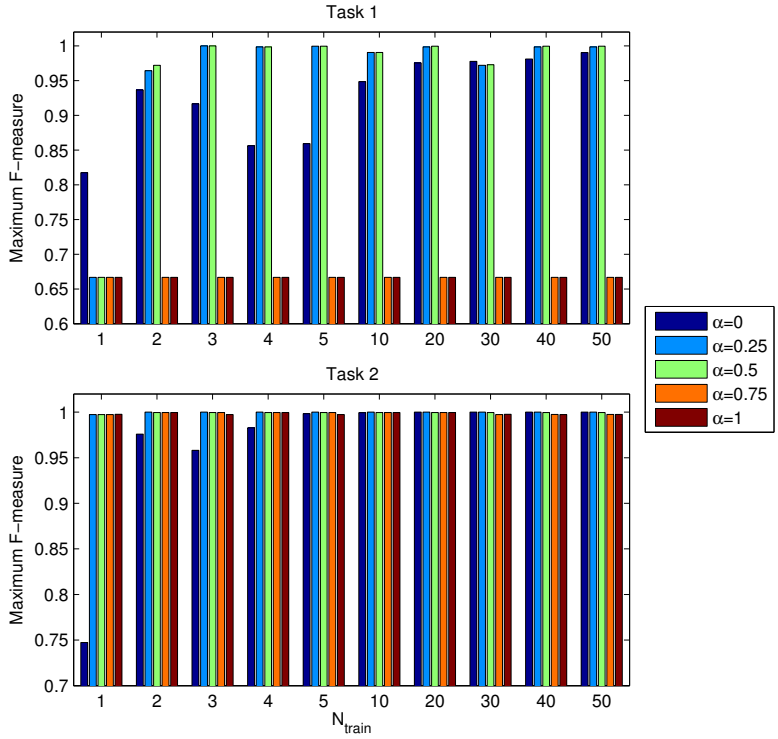

Fig. 5. Influence of the size of the training database N_{train} and of parameter α on the retrieval performances (maximum value of the averaged F-measure) for Task 1 & 2.

$\alpha = 0.5$, we see that after $N_{train} = 3$ the performances remain almost identical, which shows that a very few number of training histograms is sufficient to perform this task.

For Task 2 and $\alpha \neq 0$, only $N_{train} = 1$ histogram is sufficient for the algorithm to learn how to perform the task. For $\alpha = 0$, the performances fluctuate until N_{train}=5 (the fact that this value is greater than the one for $\alpha \neq 0$ is probably due to the same reasons already explained for Task 1).

These results are interesting, since they tend to show that for both these tasks, the algorithm is able to perform very well with only a very small number of training samples (which could also allow to save some computational time).

4.4. *Comparison with commonly used distances and cost functions*

The optimal Monge cost c^* (with $N_{train} = 50$ and $\alpha = 0.5$) is to be compared to different cost functions or distances commonly used in the image retrieval community:

- L1-distance, Euclidean distance, Cosine distance, Chi-2 distance

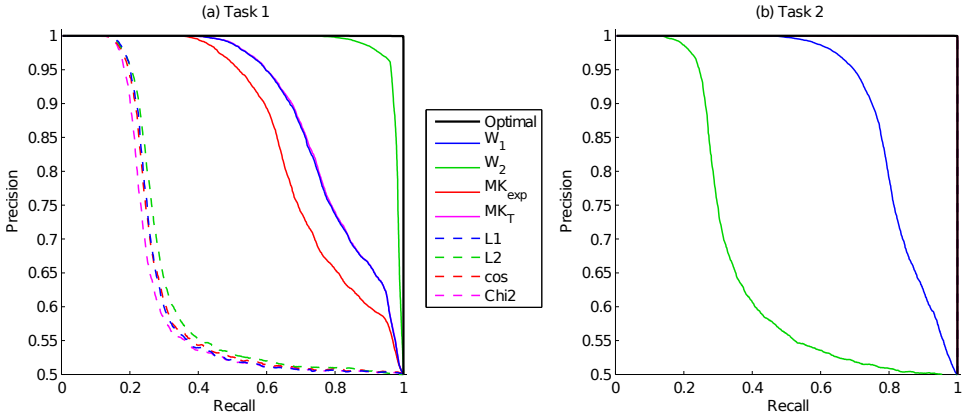

Fig. 6. Precision-recall curves for Task 1 & 2 and different distances and cost functions. Note that on (a), the plots of W_1 and MK_T almost overlap and that on (b), plots of *Optimal*, MK_{exp}, MK_T, $L1$, $L2$, *cos* and $Chi2$ are the same.

- Wasserstein distances W_p (see (2)):

$$c(i,j) = |i-j|^p \text{ with } p = 1 \text{ or } 2$$

- Exponential cost functions $MK_{exp\beta}$ [13]:

$$c(i,j) = 1 - e^{-\frac{|i-j|}{\beta}}$$

- Thresholded cost functions $MK_{T\tau}$ [14,15]:

$$c(i,j) = \min(|i-j|, \tau)$$

Note that distance $MK_{T\tau}$ is proportional to the $L1$ norm for $\tau = 1$ and that only the Wasserstein distances are based on a cost which satisfies the Monge condition. All optimal transportation costs are computed by using the FastEMD[b] package for Matlab [14,15]. For the $MK_{exp\beta}$ and $MK_{T\tau}$ distances, the optimal values of parameters β and τ are learnt so as to maximize the maximum value of the averaged F-measure on the training database Ω^{train}. The precision-recall curves for Task 1 & 2 are plotted respectively on Figure 6(a) & 6(b).

The first observation is that in both cases the optimal cost performs better or as well as the best of the others tested distances and cost functions. One can also remark that no commonly used cost function is able to perform efficiently on both tasks (even for $MK_{exp\beta}$ and $MK_{T\tau}$ where the parameters are learnt on the training database): Wasserstein distances suit well to Task 1 but not to Task 2 while other cost functions obtain good performances on Task 2 but not on Task 1. This demonstrates both the utility of adapting the cost function to the task and

[b]http://www.cs.huji.ac.il/∼ofirpele/FastEMD/

the fact that the optimization process proposed in this article is able to compute efficient cost functions.

While the data used for testing is rather basic, it is also interesting to notice that as far as Task 1 is concerned, all the tested and commonly used cost function fail at obtaining *perfect* performances (90-degree angle), while our optimal cost does. Task 2 seems easier to perform, since only Wasserstein distances fail to give optimal results. These results are promising, as in most classification or retrieval tasks, one does not *a priori* know which kind of perturbation can be found within the elements of the same class or which characteristics allow to separate elements of different classes. This uncertainty can make the choice of the cost function tricky and results most of the times in choosing a default cost function (such as the L1-norm) which offers a good compromise for a large range of problems. But as seen on Figure 6, the W_1 distance (commonly called Wasserstein distance) only offers acceptable results while the adaptation of the cost function allows to give optimal results.

5. Results on real data

We propose to test our method for image retrieval using grayscale intensity histograms. The protocol used in this experiment is identical to the one described in Section 3 (two-class retrieval task).

5.1. *Real data*

We use here the same data already used by Pele & Werman[15] when testing their EMD implementations, composed of 773 images from the COREL database[16c]. Details on the database are provided on Table 1. Each image is first converted into grayscale, and its intensity histogram is calculated on N_b=256 bins, then normalized.

5.2. *Results and comparison with commonly used distances and cost functions*

Since our method only deals with two-class tasks, the 10 classes of the database will be processed two by two, which defines 45 two-class retrieval tasks. The sizes of the training and test databases are respectively N_{train}=30 and N_{test}=5. For each of these 45 retrieval tasks, the parameters β and τ of $MK_{exp\beta}$ and $MK_{T\tau}$, as well as the optimal cost c^* are learnt on the training database. We use the basic configuration ($\alpha = 0.5$) for this comparison. For each task, we compare the best F-measure obtained by all others distances or cost functions (MK_{exp}, MK_T, $L1$,

[c]Data can be found on *http://www.seas.upenn.edu/ ofirpele/FastEMD/ and http://wang.ist.psu.edu/ jwang/test1.tar*

Label	Description	Number of images
A	People in Africa	50
B	Beaches	79
C	Outdoor Buildings	62
D	Buses	98
E	Dinosaurs	100
F	Elephants	85
G	Flowers	92
H	Horses	80
I	Mountains	63
J	Food	64

Table 1. Description of the image database.

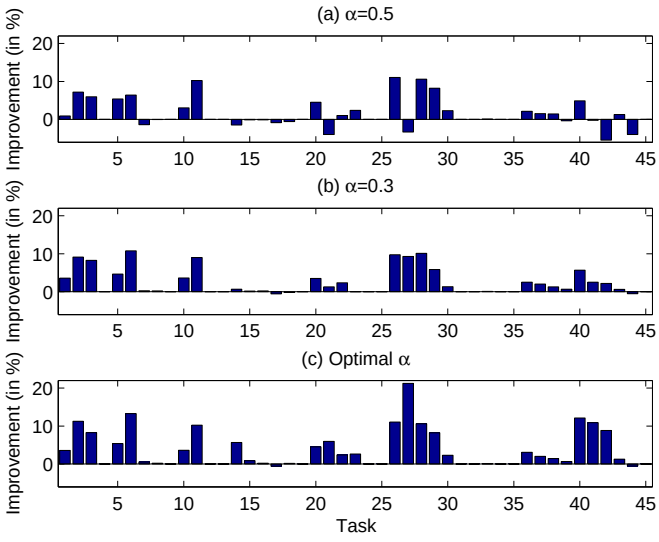

Fig. 7. Comparison between the best F-measure obtained by others distances or cost functions (MK_{exp}, MK_T, $L1$, $L2$, cos, $Chi2$) to the one obtained by the optimal cost on the image retrieval tasks with different values of α and $N_{train} = 30$

$L2$, cos, $Chi2$) to the one obtained by our optimal cost. Results are displayed on Figure 7(a).

We see on this figure that on 32 out of the 45 tasks, the optimal cost gives best or equal performances compared to all other tested distance and cost functions. The improvement is greater that 5% on 8 tasks and greater than 10% on 3 tasks. The average F-measure for the 45 tasks is 75.85% for our optimal cost and $\alpha = 0.5$. If we

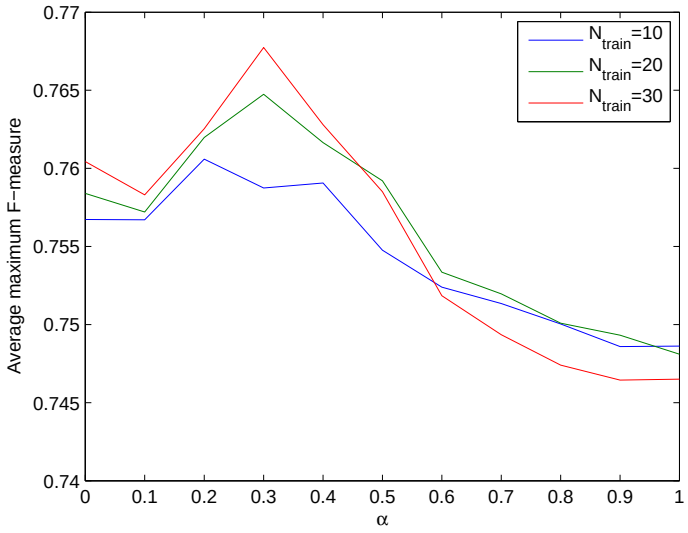

Fig. 8. Influence of the size of the training database N_{train} and of parameter α on the retrieval performances (average on all 45 tasks of the maximum values of averaged F-measure) for the image retrieval task.

compare this score to the one obtained by the cost function giving the second best overall results (which is MK_{exp} with an average F-measure of 73.89%), we see that using the optimal cost function improves the overall performances by approximately 2%.

5.3. *Influence of N_{train} and α*

We now investigate how parameters N_{train} and α can influence the performances of the optimal cost for the image retrieval task. Figure 8 displays the maximum F-measure averaged on all 45 tasks obtained with N_{train}=10, 20 & 30 and various values of α. We see that for this experiment, the value $\alpha = 0.5$ (used for the comparison with other distances and cost functions) is not optimal and that better performances can be obtained with $\alpha = 0.3$. The size of the training database has a limited influence on the performances since using N_{train}=10 only decreases the performances to less than 1%. This tend to show that our method only needs a few training histograms to perform accurately. Interestingly, note that the poorest score (obtained for $N_{train} = 30$ and $\alpha = 1$) is still better than the best of the other tested distances and cost functions (74.64% vs. 73.89% for MK_{exp}).

We propose to see how α can influence the comparison with common distances and cost functions, and we display on Figure 7(b) the improvements obtained on the 45 tasks with the best configuration ($\alpha = 0.3$). These results are especially promising, since the improvement is greater or equal to 0 in 42 out of 45 tasks

(32 for $\alpha = 0.5$). Also, negative improvements are smaller and less frequent with $\alpha = 0.3$. This is interesting since it proves the interest of our optimization process: even if we select for each task the best possible common distance and cost function, our optimal cost still performs better or only slightly worse in all cases.

For all previous experiments, we let the value of α identical for each task. But in practice, it is possible to learn the best value for α on each task by using the value maximizing the averaged F-measure on the training database with $N_{train} = 30$ (this is actually how we learnt parameters β and τ for $MK_{exp\beta}$ and $MK_{T\tau}$). The improvements obtained when using for each task the optimal α are presented on Figure 7(c). We see that the improvement is now greater or equal to 0 in 43 out of 45 tasks, greater than 5% in 14 tasks, greater than 10% in 8 tasks and greater than 20% in 1 task. Note that in this case, the average F-measure for our optimal cost function is 78.13%, while the average F-measure obtained by selecting the best common distance or cost function for each task is 74.32%, which proves that the good performances of our method are not only due to the training effect, but also to the optimization process.

6. Conclusion

In this article, we have presented an investigation on the optimization of the cost function in the Monge-Kantorovich Problem (MKP). The framework we introduce for this optimization (maximization criterion and Monge condition) enables to simplify an intractable and unsolvable problem into a simple linear programming problem. This process is applied to a basic two-class histogram retrieval with synthetic and real data. Results show that not only the adaptation of the cost function is useful in order to increase the robustness regarding intraclass perturbations, but also that our method enables to produce an efficient adapted cost function. These preliminary results are promising, as in many cases the choice of the cost function is a tricky process and that commonly used cost functions do not necessary guarantee good performances. The ideas introduced in this article shall be developed by considering other tasks and types of data and by adapting the method to multi-class problems.

Acknowledgments

This work was funded by ANR program TECSAN-SVELTE. The author wants to thank Jérémie Jakubowicz, Pascal Bianchi and Thomas Courtat for some helpful discussions and comments.

References

1. S. Rachev, L. Rüschendorf, Mass transportation problems, Vol. 1, Springer Verlag, 1998.
2. C. Villani, Optimal transport: old and new, Springer Verlag, 2008.

3. T. Chan, S. Esedoglu, K. Ni, Histogram based segmentation using wasserstein distances, Lecture Notes in Computer Science (2010) 4485 (2007) 1–12.
4. A. Irpino, R. Verde, Dynamic clustering of interval data using a wasserstein-based distance, Pattern Recognition Letters 29 (11) (2008) 1648–1658.
5. A. Irpino, R. Verde, F. De Carvalho, Dynamic clustering of histogram data based on adaptive squared wasserstein distances, to appear in Pattern Recognition.
6. J. Rabin, G. Peyré, J. Delon, M. Bernot, Wasserstein barycenter and its application to texture mixing, in: Proceedings of the Third International Conference on Scale Space and Variational Methods in Computer Vision (SSMV 11), Vol. 8, 2011.
7. A. Zamolotskikh, P. Cunningham, An assessment of alternative strategies for constructing EMD-based kernel functions for use in an SVM for image classification, in: Proceedings of the International Workshop on Content-Based Multimedia Indexing (CBMI), Bordeaux, France, 2007, pp. 11–17.
8. H. Ling, K. Okada, An efficient earth mover's distance algorithm for robust histogram comparison, IEEE Transaction on Pattern Analysis and Machine Intelligence 29 (5) (2007) 840–853.
9. H. Sakoe, S. Chiba, Dynamic programming algorithm optimization for spoken word recognition, IEEE Transactions on Acoustics, Speech and Signal Processing 26 (1) (1978) 43–49.
10. R. Typke, P. Giannopoulos, R. Veltkamp, F. Wiering, R. Van Oostrum, Using transportation distances for measuring melodic similarity, in: Proceedings of the International Conference on Music Information Retrieval (ISMIR), Baltimore, USA, 2003, pp. 107–114.
11. A. Hoffman, On simple linear programming problems. convexity., in: Proceedings of Symposia in Pure Mathematics, Vol. 7, Providence, R.I., 1961, pp. 317–327.
12. J. Rabin, J. Delon, Y. Gousseau, Transportation distances on the circle, to appear in Journal of Mathematical Imaging and Vision.
13. Y. Rubner, C. Tomasi, L. Guibas, The earth mover's distance as a metric for image retrieval, International Journal of Computer Vision 40 (2) (2000) 99–121.
14. O. Pele, M. Werman, A linear time histogram metric for improved sift matching, in: Proceedings of the European Conference on Computer Vision (ECCV), 2008, pp. 495–508.
15. O. Pele, M. Werman, Fast and robust earth mover's distances, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2009, pp. 460–467.
16. J. Wang, J. Li, G. Wiederhold, SIMPLIcity: Semantics-sensitive integrated matching for picture libraries, IEEE Transactions on Pattern Analysis and Machine Intelligence 23 (9) (2001) 947–963.