arXiv:1812.08113v1 [cs.LG] 19 Dec 2018

On The Chain Rule Optimal Transport Distance

Frank Nielsen[⋆] and Ke Sun[†]

[⋆]Sony Computer Science Laboratories, Inc.
[⋆]Tokyo, Japan
[⋆]Frank.Nielsen@acm.org
[†]Data61
[†]Australia
[†]sunk@ieee.org

Abstract

We define a novel class of distances between statistical multivariate distributions by solving an optimal transportation problem on their marginal densities with respect to a ground distance defined on their conditional densities. By using the chain rule factorization of probabilities, we show how to perform optimal transport on a ground space being an information-geometric manifold of conditional probabilities. We prove that this new distance is a metric whenever the chosen ground distance is a metric. Our distance generalizes both the Wasserstein distances between point sets and a recently introduced metric distance between statistical mixtures. As a first application of this Chain Rule Optimal Transport (CROT) distance, we show that the ground distance between statistical mixtures is upper bounded by this optimal transport distance, whenever the ground distance is joint convex. We report on our experiments which quantify the tightness of the CROT distance for the total variation distance and a square root generalization of the Jensen-Shannon divergence between mixtures.

Keywords: Optimal transport, Wasserstein distances, Information geometry, f-divergences, Total Variation, Jensen-Shannon divergence, Statistical mixtures, Joint convexity.

1 Introduction

Calculating (dis)similarities between statistical mixtures is a core primitive often met in statistics, machine learning, signal processing, and information fusion [3] among others. However, the usual information-theoretic Kullback-Leibler (KL) divergence (as known as relative entropy) or the f-divergences between statistical mixtures [28] do not admit closed-form formula, and is in practice approximated by costly Monte Carlo stochastic integration [28].

To tackle this computational tractability problem, two research directions have been considered in the literature: The first line of research consists in proposing some distances between mixtures that yield closed-form formula [23] (e.g., the Cauchy-Schwarz divergence or the Jensen quadratic Rényi divergence). The second line of research consists in lower and upper bounding the f-divergences between mixtures [28]. This is tricky when considering bounded divergences like the Total Variation (TV) distance or the Jensen-Shannon (JS) divergence that are upper bounded by 1 and $\log 2$, respectively.

When dealing with probability densities, two main classes of statistical distances have been widely studied in the literature:

1. The Information-Geometric (IG) invariant f-divergences [1] (characterized as the class of separable distances), and

2. The Wasserstein distances of Optimal Transport (OT) [37] which can be computationally accelerated using entropy regularization [4, 10] (Sinkhorn divergence).

In general, computing closed-form formula for the OT between parametric distributions is difficult. A closed-form formula is known for elliptical distributions [8] (that includes the multivariate Gaussian distributions), and the OT of multivariate continuous distributions can be calculated from the OT of their copulas [13].

The geometry induced by the distance is different in these two OT/IG cases. For example, consider location-scale families (or multivariate elliptical distributions):

1. For OT, the distance between any two members admit the *same* closed-form formula [8] (depending only on the mean and variance parameters, not on the type of location-scale family). The OT geometry of Gaussian distributions has positive curvature [41].

2. For any f-divergence, the information-geometric manifold has negative curvature [18] (hyperbolic geometry). It is known that for the Kullback-Leibler divergence, the manifold of mixtures with prescribed components is dually flat, and admits therefore an equivalent Bregman divergence [25].

In this paper, we build on the seminal work of Liu and Huang [21] that proposed a novel family of statistical distances for statistical mixtures by solving linear programs between [21] component weights of mixtures where the elementary distance between any two mixtures is prescribed. They proved that their distance between mixtures (that we term MCOT distance for Mixture Component Optimal Transport) is a metric whenever the elementary distance between mixture components is a metric. This framework also applies to semi-parametric mixtures obtained from Kernel Density Estimators [38] (KDEs).

We describe our main contributions as follows:

- We define the *Chain Rule Optimal Transport* (CROT) distance in Definition 1, and prove that it yields a metric whenever the distance between conditional distributions is a metric in §2.2 (Theorem 3). The CROT distance extends the Wasserstein/EMD distances and the MCOT distance between statistical mixtures. We further sketch show how to build recursively hierarchical families of CROT distances.

- We report a novel generic upper bound for statistical distances between mixtures [29] using CROT distances in §3 (Theorem 5) whenever the ground distance is joint convex.

- In §4, experiments highlight quantitatively the upper bound performance of the CROT distance for bounding the total variation distance and a generalization of the square root of the Jensen-Shannon distance.

2 The Chain Rule Optimal Transport (CROT) distance

2.1 Definition

We define a novel class of distances between statistical multivariate distributions. Recall the basic *chain rule* factorization of a joint probability distribution $p(x,y)$:

$$p(x,y) = p(y)p(x|y),$$

where probability $p(y)$ is called the *marginal probability*, and probability $p(x|y)$ is termed the *conditional probability*. Let $\mathcal{Y} = \{p(y)\}$ and $\mathcal{C} = \{p(x|y)\}$ denote the manifolds of marginal probability densities and conditional probability densities, respectively.

For example, for latent models like statistical mixtures or hidden Markov models [42, 39], x plays the role of the *observed variable* while y denotes the *hidden variable* [9] (unobserved so that inference has to tackle incomplete data, say, using the EM algorithm [6]).

First, we state the generic definition of the *Chain Rule Optimal Transport* distance between joint distributions p and q (with $q(x,y) = q(y)q(x|y)$) as follows:

Definition 1 (CROT distance)**.** *Given two multivariate distributions p and q, we define the* Chain Rule Optimal Transport *(CROT) as follows:*

$$\begin{aligned} H_\delta(p,q) &:= \inf_r E_{r(y,z)}\left[\delta\Big(p(x|y), q(x|z)\Big)\right], && (1)\\ &= \inf_r \int r(y,z)\delta\Big(p(x|y), q(x|z)\Big)\,\mathrm{d}y\mathrm{d}z, && (2)\end{aligned}$$

where $\delta(\cdot,\cdot)$ is a ground distance defined on conditional density manifold $\mathcal{C} = \{p(x|y)\}$ (e.g., the Total Variation — TV), and $r \in \Gamma(p,q)$ (set of all probability measures on $\mathcal{Y}^2$ with marginals p and q) satisfying the following constraint:

$$\int r(y,z)\mathrm{d}z = p(y), \quad \int r(y,z)\mathrm{d}y = q(z), \tag{3}$$

When the ground distance δ is clear from the context, we write $H(p,q)$ for a shortcut of $H_\delta(p,q)$. Since $\int r(y,z)\mathrm{d}y\mathrm{d}z = 1$ and since $r(y,z) = p(y)q(z)$ is a feasible transport solution, we get the following upper bounds:

Property 2 (Upper bounds)**.** *The CROT is upper bounded by*

$$H_\delta(p,q) \le \int_y \int_z p(y)q(z)\delta\Big(p(x|y), q(x|z)\Big)\mathrm{d}y\mathrm{d}z \le \max_{y,z} \delta\Big(p(x|y), q(x|z)\Big).$$

Figure 1 illustrates the principle of the CROT distance. Another complementary motivation when dealing with statistical mixtures ispresented in §3

Let us notice that the CROT distance generalizes two distances met in the literature:

Remark 2.1 (CROT generalizes Wasserstein/EMD)**.** *In the case that $p(x|y) = \delta(y)$ (Dirac distributions), we recover the Wasserstein distance [41] between point sets (or Earth Mover Distance [36]), where $\delta(\cdot,\cdot)$ is the ground metric distance.*

Remark 2.2 (CROT generalizes MCOT)**.** *When both $p(y)$ and $q(z)$ are both (finite) categorical distributions, we recover the distance formerly defined in [21] that we term the MCOT distance in the remainder (for Mixture Component Optimal Transport).*

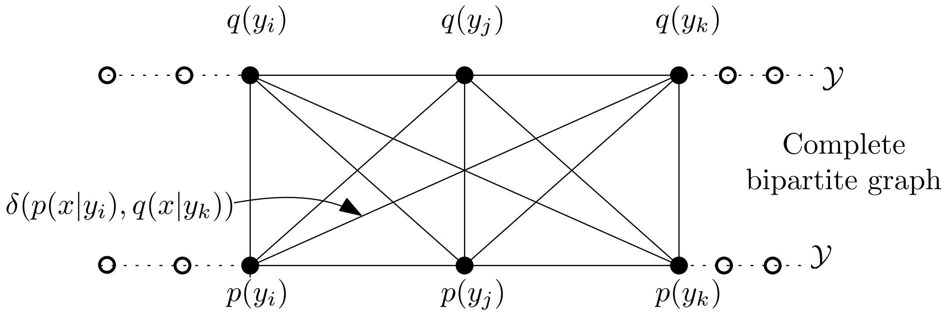


Figure 1: The CROT distance: Optimal matching of marginal densities wrt. a distance on conditional densities. We consider the complete bipartite graph with edges weighted by the distances δ between the corresponding conditional densities defined at edge vertices.

2.2 CROT is a metric when the ground distance is a metric

Theorem 3 (CROT metric). *$H_\delta(p, q)$ is a metric whenever $\delta(\cdot, \cdot)$ is a metric.*

Proof. We prove that $H(p, q)$ satisfies the following axioms of metric distances:

Non-negativity. As $\delta\left(p(x|y), q(x|y')\right) \geq 0$, we have $H_\delta(p, q) \geq 0$.

Law of indiscernibles. If $H_\delta(p, q) = 0$, as $\delta(\cdot, \cdot)$ is a metric, then the density $r(y, z)$ is concentrated on the region $p(x|y) = q(x|z)$ in $\mathcal{C}^2$. We therefore have

$$p(y)p(x|y) = \int r(y, z)\mathrm{d}z p(x|y) = \int r(y, z)p(x|y)\mathrm{d}z = \int r(y, z)q(x|z)\mathrm{d}y = q(z)q(x|z). \tag{4}$$

Symmetry.

$$H_\delta(p, q) = \inf_r \int r(y, z)\delta\left(p(x|y), q(x|z)\right) \mathrm{d}y\mathrm{d}z = \inf_r \int R(z, y)\delta\left(q(x|z), p(x|y)\right) \mathrm{d}y\mathrm{d}z \tag{5}$$
$$= H_\delta(q, p) \tag{6}$$

where $R(z, y) = r(y, z)$ s.t. $\int R(z, y)\mathrm{d}y = q(z)$ and $\int R(z, y)\mathrm{d}z = p(y)$.

Triangle inequality. The proof of the triangle inequality is not straightforward.

$$\begin{aligned}
H_\delta(p_1, p_2) + H_\delta(p_2, p_3) &= \inf_{r_1} E_{r_1(y,z)}\delta(p_1(x|y), p_2(x|z)) + \inf_{r_2} E_{r_2(y,z)}\delta(p_2(x|y), p_3(x|z)) \\
&= \inf_r E_{s(y_1,y_2,z)} \left(\delta(p_1(x|y_1), p_2(x|z)) + \delta(p_2(x|y_2), p_3(x|z))\right) \\
&\geq \inf_r E_{s(y_1,y_2,z)}\delta(p_1(x|y_1), p_3(x|z)) \\
&\geq \inf_r E_{r(y,z)}\delta(p_1(x|y), p_3(x|z)), \qquad (7)
\end{aligned}$$

where $s(p, q, r)$ denotes the set of all probability measures on $\mathcal{Y}^3$ with marginals p, q and r. $\square$

3 CROT for statistical mixtures and Sinkhorn CROT

Consider two finite statistical mixtures $m_1(x) = \sum_{i=1}^{k_1} \alpha_i p_i(x)$ and $m_2(x) = \sum_{i=1}^{k_2} \beta_i q_i(x)$, not necessarily homogeneous nor of the same type. Let $[k]:=\{1,\ldots,k\}$. The *Mixture Component Optimal Transport* (MCOT) distance proposed in [21] amounts to solve a *Linear Program* (LP) with the following objective function to minimize:

$$H_\delta(p,q) = \sum_{i=1}^{k_1}\sum_{j=1}^{k_2} w_{ij}\delta(p_i,q_j), \tag{8}$$

satisfying the following constraints:

$$\begin{aligned} w_{ij} &\geq 0, \quad \forall i \in [k_1], j \in [k_2] &(9)\\ \sum_{l=1}^{k_2} w_{il} &= \alpha_i, \quad \forall i \in [k_1] &(10)\\ \sum_{l=1}^{k_1} w_{lj} &= \beta_j, \quad \forall j \in [k_2]. &(11)\end{aligned}$$

By defining $U(\alpha,\beta)$ to be set of non-negative matrices $W=[w_{ij}]$ with $\sum_{l=1}^{k_2} w_{il} = \alpha_i$ and $\sum_{l=1}^{k_1} w_{lj} = \beta_j$ (transport polytope [5]), we get the equivalent compact definition of MCOT/CROT:

$$H_\delta(m_1 : m_2) = \min_{W\in U(\alpha,\beta)} \sum_{i=1}^{k_1}\sum_{j=1}^{k_2} w_{ij}\delta(p_i,q_j). \tag{12}$$

When the ground distance δ is asymmetric, we shall use the ':' notation instead of the ',' notation for separating arguments.

In general, the LP problem (with $k_1 \times k_2$ variables and inequalities, k_1+k_2 equalities whom k_1+k_2-1 are independent) delivers an optimal soft assignment of mixture components with exactly k_1+k_2-1 nonzero coefficients[1] in matrix $W=[w_{ij}]$. The complexity of linear programming in n variables with b bits using Karmarkar's interior point methods is polynomial, in $O(n^{\frac{7}{2}}b^2)$ [19].

Observe that we necessarily have:

$$\max_{j\in[k_2]} w_{ij} \geq \frac{\alpha_i}{k_2},$$

and similarly that:

$$\max_{i\in[k_1]} w_{ij} \geq \frac{\beta_j}{k_1}.$$

Note that $H(m,m)=0$ since $w_{ij}=\delta_{ij}$ where δ_{ij} denotes the Krönecker symbol: $\delta_{ij}=1$ iff $i=j$, and 0 otherwise.

We can interpret MCOT as a discrete optimal transport between (non-embedded) histograms. When $k_1=k_2=d$, the transport polytope is the polyhedral set of non-negative $d\times$ matrices:

[1]A LP in d-dimensions has its solution located at a vertex of a polytope, described by the intersection of $d+1$ hyperplanes (linear constraints).

$$U(\alpha,\beta) = \{P \in \mathbb{R}_+^{d\times d} : P1_d = \alpha, P^\top 1_d = \beta\},$$

and

$$H_\delta(m_1 : m_2) = \min_{P\in U(\alpha,\beta)} \langle P, W\rangle,$$

where $\langle A, B\rangle = \mathrm{tr}(A^\top B)$ is the Fröbenius inner product of matrices, and $\mathrm{tr}(A)$ the matrix trace. This OT can be calculated using the network simplex in $O(d^3 \log d)$ time.

Cuturi [5] showed how to relax the objective function in order to get fast calculation using the Sinkhorn divergence:

$$S_\delta(m_1 : m_2) = \min_{P\in U_\lambda(\alpha,\beta)} \langle P, W\rangle, \tag{13}$$

where $U_\lambda(\alpha,\beta){:=}\{P \in U(\alpha,\beta) : \mathrm{KL}(P : \alpha\beta^\top) \leq \lambda\}$. The KL divergence between two $k\times k$ matrices $M = [m_{i,j}]$ and $M' = [m'_{i,j}]$ is defined by

$$\mathrm{KL}(M : M'){:=}\sum_{i,j} m_{i,j} \log \frac{m_{i,j}}{m'_{i,j}},$$

with the convention that $0 \log \frac{0}{0} = 0$. The Sinkhorn divergence is calculated using the equivalent dual Sinkhorn divergence by using matrix scaling algorithms (e.g., the Sinkhorn-Knopp algorithm).

Because the minimization is performed on $U_\lambda(\alpha,\beta) \subset U(\alpha,\beta)$, we have

$$H_\delta(m_1 : m_2) \leq S_\delta(m_1 : m_2). \tag{14}$$

3.1 Upper bounding statistical distances between mixtures with CROT

First, let us report the basic upper bounds for MCOT mentioned earlier in Property 2. The objective function is upper bounded by:

$$H(m_1, m_2) \leq \sum_{i=1}^{k_1}\sum_{j=1}^{k_2} \alpha_i\beta_j\delta(p_i, q_j) \leq \max_{i\in[k_1], j\in[k_2]} \delta(p_i, q_j). \tag{15}$$

Now, when the conditional density distance δ is *separate convex* (i.e., meaning convex in both arguments), we get the following *Separate Convexity Upper Bound* (SCUB):

$$(SCUB): \quad \delta(m_1 : m_2) \leq \sum_{i=1}^{k_1}\sum_{j=1}^{k_2} \alpha_i\beta_j\delta(p_i : q_j). \tag{16}$$

For example, norm-induced distances or f-divergences [26] are separate convex distances.

For the particular KL divergence

$$\mathrm{KL}(p : q){:=}\int p(x) \log \frac{p(x)}{q(x)} \mathrm{d}x,$$

and when $k_1 = k_2$, we get the following upper bound using the log-sum inequality [7, 27]:

$$\mathrm{KL}(m_1 : m_2) \leq \mathrm{KL}(\alpha : \beta) + \sum_{i=1}^{k} \alpha_i \mathrm{KL}(p_i : q_i), \tag{17}$$

Since this holds for any permutation of σ of mixture components, we can tight this upper bound by minimizing over all permutations:

$$\mathrm{KL}(m_1 : m_2) \leq \min_\sigma \mathrm{KL}(\alpha : \sigma(\beta)) + \sum_{i=1}^k \alpha_i \mathrm{KL}(p_i : \sigma(q_i)). \tag{18}$$

The best permutation σ can be computed using the Hungarian cubic time algorithm [40, 35, 15, 14] (with cost matrix $C = [c_{ij}]$, and $c_{ij} = \mathrm{kl}(\alpha_i : \beta_j) + \alpha_i \mathrm{KL}(p_i : q_j)$ with $\mathrm{kl}(a : b) = a \log \frac{a}{b}$).

Now, let us further rewrite $m_1(x) = \sum_{i=1}^{k_1} \sum_{j=1}^{k_2} w_{i,j} p_i(x)$ with $\sum_{j=1}^{k_2} w_{i,j} = \alpha_i$, and $m_2(x) = \sum_{i=1}^{k_1} \sum_{j=1}^{k_2} w'_{i,j} q_j(x)$ with $\sum_{i=1}^{k_1} w'_{i,j} = \beta_j$. That is, we can interpret $m_1(x) = \sum_{i=1}^{k_1} \sum_{j=1}^{k_2} w_{i,j} p_{i,j}(x)$ and $m_2(x) = \sum_{i=1}^{k_1} \sum_{j=1}^{k_2} w'_{i,j} q_{i,j}(x)$ as mixtures of $k = k_1 \times k_2$ (redundant) components $\{p_{i,j}(x) = p_i(x)\}$ and $\{q_{i,j}(x) = q_j(x)\}$, and apply the upper bound of Eq. 17 for the "best split" of matching mixture components $\sum_{j=1}^{k_2} w_{i,j} p_i(x) \leftrightarrow \sum_{j=1}^{k_1} w'_{j,i} q_i(x)$:

$$\mathrm{KL}(m_1 : m_2) \leq \min_{w \in U(\alpha,\beta)} \sum_{i=1}^{k_1} \sum_{j=1}^{k_2} w_{i,j} \log \frac{w_{i,j}}{w'_{j,i}} + \sum_{i=1}^{k_1} \sum_{j=1}^{k_2} w_{ij} \mathrm{KL}(p_i : q_j), \tag{19}$$

Let

$$O(m_1 : m_2) = \min_{w \in U(\alpha,\beta)} \sum_{i=1}^{k_1} \sum_{j=1}^{k_2} w_{i,j} \log \frac{w_{i,j}}{w'_{j,i}} + \sum_{i=1}^{k_1} \sum_{j=1}^{k_2} w_{ij} \mathrm{KL}(p_i : q_j). \tag{20}$$

Then it follows that

$$\mathrm{KL}(m_1 : m_2) \leq O(m_1 : m_2) \leq \sum_{i=1}^{k_1} \sum_{j=1}^{k_2} w_{i,j} \log \frac{w_{i,j}}{w'_{j,i}} + H_{\mathrm{KL}}(m_1, m_2). \tag{21}$$

Thus CROT allows to upper bound the KL divergence between mixtures. The technique of rewriting mixtures as mixtures of $k = k_1 \times k_2$ redundant components bears some resemblance with the variational upper bound on the KL between mixtures proposed in [16] that requires to iterate until convergence an update of the variational upper bound.

In fact, the CROT distance provides a good upper bound on the distance between mixtures provided the base distance δ is joint convex [2, 33].

Definition 4 (Joint convex distance)**.** *A distance $D(\cdot : \cdot)$ is joint convex if and only if*

$$D((p_1 p_2)_\alpha : (q_1 q_2)_\alpha) \leq (D(p_1 : p_2) D(p_2 : q_2))_\alpha, \quad \forall \alpha \in [0, 1],$$

where $(ab)_\alpha := (1-\alpha)a + \alpha b$.

The f-divergences $I_f(p : q) = \int p(x) f(q(x)/p(x)) \mathrm{d}x$ (for a convex generator $f(u)$ satisfying $f(1) = 0$) are joint convex distances [31]. For mixtures with same weights but different component basis and a joint convex distance D (e.g., KL), we get $D(\sum_{i=1}^k w_i p_i : \sum_{i=1}^k w_i q_i) \leq \sum_{i=1}^k \alpha_i D(p_i : q_i)$.

Theorem 5 (Upper Bound on Joint Convex Mixture Distance (UBJCMD))**.** *Let $m_1(x) = \sum_{i=1}^{k_1} \alpha_i p_i(x)$ and $m_2(x) = \sum_{i=1}^{k_2} \beta_i q_i(x)$ be two finite mixtures, and $\delta(\cdot, \cdot)$ any joint convex statistical base distance. Then CROT $H_\delta(m_1 : m_2)$ upper bounds the distance $\delta(m_1, m_2)$ between mixtures:*

$$(JCUB): \quad \delta(m_1 : m_2) \leq H_\delta(m_1 : m_2). \tag{22}$$

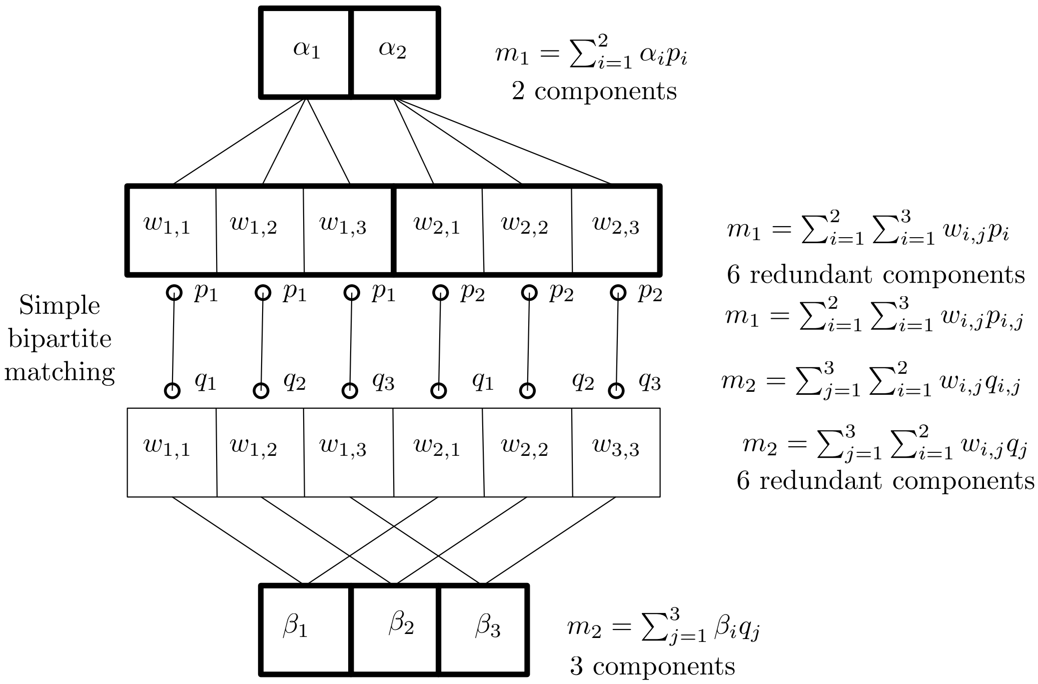


Figure 2: An interpretation of CROT by rewriting the mixtures $m_1 = \sum_{i=1}^{k_1} \sum_{j=1}^{k_2} w_{i,j} p_{i,j}$ and $m_2 = \sum_{i=1}^{k_1} \sum_{j=1}^{k_2} w_{i,j} q_{i,j}$ with $p_{i,j} = p_i$ and $q_{i,j} = q_j$ and using the joint convexity of the base distance δ.

Proof.

$$
\begin{aligned}
\delta(m_1 : m_2) &= \delta\left(\sum_{i=1}^{k_1} \alpha_i p_i, \sum_{j=1}^{k_2} \beta_j q_j\right) \\
&= \delta\left(\sum_{i=1}^{k_1} \sum_{j=1}^{k_2} w_{i,j} p_{i,j} : \sum_{i=1}^{k_1} \sum_{j=1}^{k_2} w_{i,j} q_{i,j}\right) \\
&\leq \sum_{i=1}^{k_1} \sum_{j=1}^{k_2} w_{i,j} \delta(p_{i,j} : q_{i,j}), \\
&\leq \sum_{i=1}^{k_1} \sum_{j=1}^{k_2} w_{i,j} \delta(p_i : q_j) =: H_\delta(m_1, m_2).
\end{aligned}
$$

□

Notice that $H_\delta(m_1, m_2) \neq H_\delta(m_2, m_1)$ for asymmetric base distance δ.

Conversely, CROT yields a lower bound for joint concave distances (e.g., fidelity in quantum computing [30]).

Figure 2 illustrates the CROT distance between statistical mixtures (not having the same number of components).

4 Experiments

4.1 Total Variation distance

Since TV is a metric f-divergence [17] bounded in $[0,1]$, so is MCOT. The closed-form formula for the total variation between univariate Gaussian distributions is reported in [24] (using the erf function), and the other formula for the total variation between Rayleigh distributions and Gamma distributions is given in [29].

Figure 3a illustrates the performances of the various lower/upper bounds on the total variation between mixtures of Gaussian, Gamma, and Rayleigh distributions with respect to the true value which is estimated using Monte Carlo samplings.

The acronyms of the various bounds are as follows:

- CELB: Combinatorial Envelope Lower Bound [28] (applies only for 1D mixtures)
- CEUB: Combinatorial Envelope Upper Bound [28] (applies only for 1D mixtures)
- CGQLB: Coarse-Grained Quantization Lower Bound [28] for 1000 bins (applies only for f-divergences that satisfy the information monotonicity property)
- CROT: Chain Rule Optimal Transport H_δ (this paper)
- Sinkhorn CROT: Entropy-regularized CROT [5] $S_\delta \leq H_\delta$, with $\lambda = 1$ and $\epsilon = 10^{-8}$ (for convergence of the Skinhorn-Knopp iterative matrix scaling algorithm).

Next, we consider the renown MNIST handwritten digit database [20]: A dataset of 70000 handwritten digit 28×28 grey images.[2] We learn GMMs composed of multivariate Gaussian distributions with a diagonal covariance matrix from this MNIST database using PCA (dimension reduction from original dimension $d = 28 \times 28 = 784$ to reduced dimension D) as explained in the caption of Table 1. We used the Expectation-Maximization (EM) algorithm implementation of scikit-learn [32].

We approximate the TV between D-dimensional GMMs using Monte Carlo by performing stochastic integration of the following integrals: Let $r(x) = \frac{p(x)+q(x)}{2}$,

$$\begin{aligned}\mathrm{TV}(p,q) :=& \frac{1}{2}\int |p(x)-q(x)|\mathrm{d}x = \frac{1}{2}\int r(x)\left|\frac{p(x)}{r(x)} - \frac{q(x)}{r(x)}\right|\mathrm{d}x \\ &= \frac{1}{4}\int p(x)\left|\frac{p(x)-q(x)}{r(x)}\right|\mathrm{d}x + \frac{1}{4}\int q(x)\left|\frac{p(x)-q(x)}{r(x)}\right|\mathrm{d}x.\end{aligned}$$

Furthermore, we have:

$$\frac{p(x)-q(x)}{r(x)} = 2\frac{p(x)-q(x)}{p(x)+q(x)} = 2\frac{\frac{p(x)}{q(x)}-1}{\frac{p(x)}{q(x)}+1}.$$

The results are obtained using POT [11] (Python Optimal Transport).

Our experiments yield the following observations: As the sample size τ decreases, the TV distances between GMMs turn larger because the GMMs are pulled towards the two different empirical distributions. As the dimension D increases, TV increases because in a high dimensional space the GMM components are less likely to overlap. We check that CROT-TV is an upper bound of TV. We verify that Sinkhorn divergences are upper bounds of CROT.

[2] `http://yann.lecun.com/exdb/mnist/`

Table 1: Distances between two GMMs with 10 components each estimated on PCA-processed MNIST dataset. D is the dimensionality of the PCA. Parameter $0 < \tau \leq 1$ is the relative sample size used to estimated the GMMs. The two GMMs are estimated based on non-overlapping samples. For each configuration, the GMMs are repeatedly estimated based on 10 different random split of the MNIST dataset. The mean $\pm$ std is based on 50 independent runs.

	TV	CROT-TV	Sinkhorn ($\lambda = 5$)	Sinkhorn ($\lambda = 1$)
$D = 50$, $\tau = 1$	0.363 ± 0.0921	0.445 ± 0.121	0.446 ± 0.121	0.806 ± 0.0351
$D = 50$, $\tau = 0.1$	0.549 ± 0.064	0.646 ± 0.0846	0.648 ± 0.0853	0.862 ± 0.027
$D = 10$, $\tau = 1$	0.199 ± 0.0877	0.325 ± 0.149	0.325 ± 0.149	0.807 ± 0.0393

4.2 Square root of the symmetric α-Jensen-Shannon divergence

TV is bounded in $[0, 1]$ which makes it difficult to appreciate the quality of the CROT upper bounds in general. We shall consider a different parametric distance D_α that is upper bounded by an arbitrary bound: $D_\alpha(p, q) \leq C_\alpha$.

It is well known that the square root of the Jensen-Shannon divergence is a metric [12] (satisfying the triangle inequality). In [22], a generalization of the Jensen-Shannon divergence was proposed, given by

$$\mathrm{JS}_\alpha(p : q) := \mathrm{KL}(p : (pq)_\alpha) + \mathrm{KL}(q : (pq)_\alpha), \tag{23}$$

where $(pq)_\alpha := (1 - \alpha)p + \alpha q$. JS_α unifies (twice) the Jensen-Shannon divergence (obtained when $\alpha = \frac{1}{2}$) with the Jeffreys divergence ($\alpha = 1$) [22]. A nice property is that the skew K-divergence is upper bounded as follows:

$$\mathrm{KL}(p : (pq)_\alpha) \leq \int p \log \frac{p}{(1-\alpha)p} \leq -\log(1 - \alpha)$$

for $\alpha \in (0, 1)$, so that $\mathrm{JS}_\alpha[p : q] \leq -2\log(1 - \alpha)$ for $\alpha \in (0, 1)$.

Thus, we have the square root of the symmetrized α-divergence that is upper bounded by

$$\sqrt{\mathrm{JS}_\alpha(p : q)} \leq C_\alpha = \sqrt{-2\log(1 - \alpha)}.$$

However, $\sqrt{\mathrm{JS}_\alpha[p : q]}$ is not a metric in general [31]. Indeed, in the extreme case of $\alpha = 1$, it is known that any positive power of the Jeffreys divergence does not yield a metric.

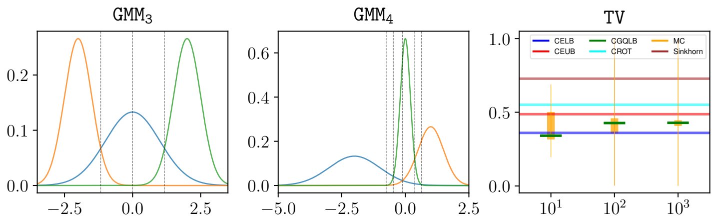
Observe that JS_α is a f-divergence since $K_\alpha(p : q) := \mathrm{KL}(p : (pq)_\alpha)$ is a f-divergence for the generator $f(u) = -\log((1 - \alpha) + \alpha u)$, and we have $\mathrm{KL}(q : (pq)_\alpha) = K_{1-\alpha}(q : p)$. Since $I_f(q : p) = I_{f^\diamond}(p : q)$ for $g(u) = uf(1/u)$, it follows that the f-generator f_{JS_α} for the JS_α divergence is:

$$f_{\mathrm{JS}_\alpha}(u) = -\log\left((1 - \alpha) + \alpha u\right) - \log\left(\alpha + \frac{1-\alpha}{u}\right). \tag{24}$$

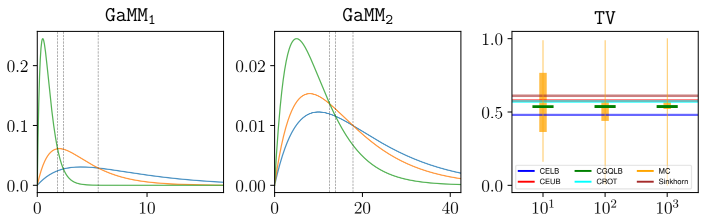
Figure 3b displays the experimental results obtained for the α-JS divergences.

5 Conclusion and perspectives

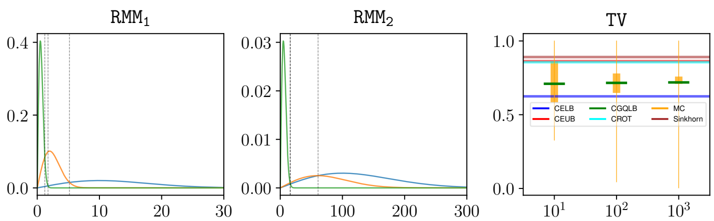
In this work, we defined the generic *Chain Rule Optimal Transport* (CROT) distance (Definition 1) H_δ for a ground distance δ that encompasses the Wasserstein distance between point sets (Earth



(1)

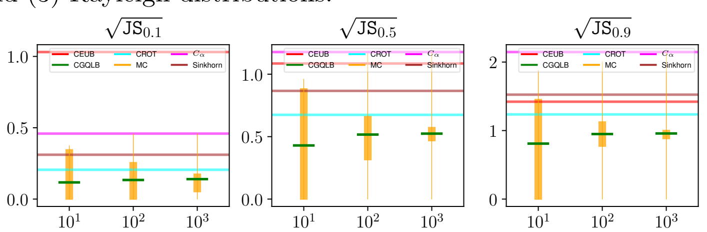


(2)

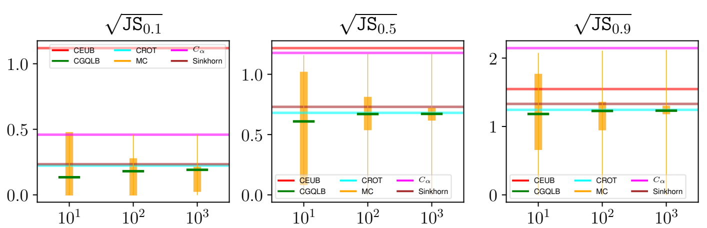


(3)

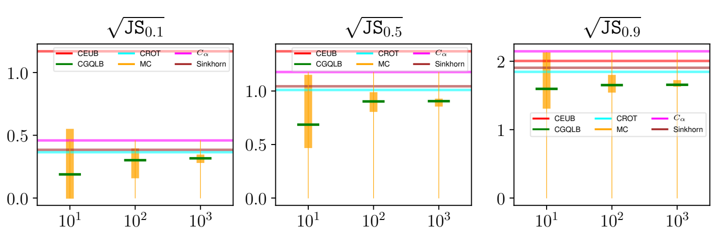
(a) Performance of the CROT distance and the Sinkhorn CROT distance for upper bounding the total variation distance between mixtures of (1) Gaussian, (2) Gamma, and (3) Rayleigh distributions.



(1)



(2)



(3)

(b) Performance of the CROT distance and the Sinkhorn CROT distance for upper bounding the square root of the α-Jensen-Shannon distance between mixtures of (1) Gaussian, (2) Gamma, and (3) Rayleigh distributions.

Figure 3: Experimental results.

Mover Distance [36]) and the *Mixture Component Optimal Transport* (MCOT) distance [21], and proved that H_δ is a metric whenever δ is a metric (Theorem 3). We then dealt with statistical mixtures, and showed that $H_\delta(m_1 : m_2) \geq \delta(m_1 : m_2)$ (Theorem 5) whenever δ is joint convex. This holds in particular for statistical f-divergences $I_f(p : q) = \int p(x) f(q(x)/p(x)) \mathrm{d}x$:

$$H_{I_f}(m_1 : m_2) \geq I_f(m_1 : m_2).$$

We also considered the smoothened Sinkhorn CROT distance $S_\delta(m_1 : m_2)$ for fast calculations of $H_\delta(m_1 : m_2)$ via matrix scaling algorithms (Sinkhorn-Knopp algorithm), with $H_\delta(m_1 : m_2) \leq S_\delta(m_1 : m_2)$.

There are many venues to explore for further research. For example, we may consider infinite Gaussian mixtures [34], the chain rule factorization for d-variate densities: This gives rise to a hierarchy of CROT distances. Another direction is to explore the use of the CROT distance in deep learning.

The smooth (dual) Sinkhorn divergence has also been shown experimentally (MNIST classification) to improve over the EMD in applications [5]. It would be also interesting to consider the Sinkhorn CROT vs CROT in applications [21] that deal with mixtures of features.

Acknowledgments

Frank Nielsen thanks Steve Huntsman for pointing out reference [21] to his attention.

References

[1] Shun-ichi Amari. *Information Geometry and Its Applications.* Applied Mathematical Sciences. Springer Japan, 2016.

[2] Heinz H Bauschke and Jonathan M Borwein. Joint and separate convexity of the Bregman distance. In *Studies in Computational Mathematics*, volume 8, pages 23–36. Elsevier, 2001.

[3] Kuo-Chu Chang and Wei Sun. Scalable fusion with mixture distributions in sensor networks. In *11th International Conference on Control Automation Robotics & Vision (ICARCV)*, pages 1251–1256, 2010.

[4] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in neural information processing systems*, pages 2292–2300, 2013.

[5] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in neural information processing systems*, pages 2292–2300, 2013.

[6] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977.

[7] Minh N Do. Fast approximation of Kullback-Leibler distance for dependence trees and hidden Markov models. *IEEE signal processing letters*, 10(4):115–118, 2003.

[8] D. C. Dowson and B.V. Landau. The fréchet distance between multivariate normal distributions. *Journal of multivariate analysis*, 12(3):450–455, 1982.

[9] B. Everett. *An introduction to latent variable models*. Springer Science & Business Media, 2013.

[10] Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-Ichi Amari, Alain Trouvé, and Gabriel Peyré. Interpolating between optimal transport and MMD using Sinkhorn divergences. *arXiv preprint arXiv:1810.08278*, 2018.

[11] R'emi Flamary and Nicolas Courty. POT python optimal transport library, 2017.

[12] Bent Fuglede and Flemming Topsoe. Jensen-Shannon divergence and Hilbert space embedding. In *International Symposium on Information Theor (ISIT 2004)*, page 31. IEEE, 2004.

[13] N. Ghaffari and S. Walker. On Multivariate Optimal Transportation. *ArXiv e-prints*, January 2018.

[14] Jacob Goldberger and Hagai Aronowitz. A distance measure between GMMs based on the unscented transform and its application to speaker recognition. In *INTERSPEECH European Conference on Speech Communication and Technology,*, pages 1985–1988, 2005.

[15] Jacob Goldberger, Shiri Gordon, and Hayit Greenspan. An efficient image similarity measure based on approximations of KL-divergence between two Gaussian mixtures. In *IEEE International Conference on Computer Vision (ICCV)*, page 487. IEEE, 2003.

[16] John R. Hershey and Peder A. Olsen. Approximating the Kullback-Leibler divergence between Gaussian mixture models. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 4, pages IV–317. IEEE, 2007.

[17] Mohammadali Khosravifard, Dariush Fooladivanda, and T Aaron Gulliver. Confliction of the convexity and metric properties in f-divergences. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, 90(9):1848–1853, 2007.

[18] Fumiyasu Komaki. Bayesian prediction based on a class of shrinkage priors for location-scale models. *Annals of the Institute of Statistical Mathematics*, 59(1):135–146, 2007.

[19] Bernhard Korte and Jens Vygen. Linear programming algorithms. In *Combinatorial Optimization*, pages 75–102. Springer, 2018.

[20] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[21] Zhu Liu and Qian Huang. A new distance measure for probability distribution function of mixture type. In *International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 616–619. IEEE, 2000.

[22] Frank Nielsen. A family of statistical symmetric divergences based on Jensen's inequality. *arXiv preprint arXiv:1009.4004*, 2010.

[23] Frank Nielsen. Closed-form information-theoretic divergences for statistical mixtures. In *Pattern Recognition (ICPR), 2012 21st International Conference on*, pages 1723–1726. IEEE, 2012.

[24] Frank Nielsen. Generalized Bhattacharyya and Chernoff upper bounds on bayes error using quasi-arithmetic means. *Pattern Recognition Letters*, 42:25–34, 2014.

[25] Frank Nielsen and Gaëtan Hadjeres. Monte Carlo information geometry: The dually flat case. *arXiv preprint arXiv:1803.07225*, 2018.

[26] Frank Nielsen and Richard Nock. On the chi square and higher-order chi distances for approximating f-divergences. *IEEE Signal Processing Letters*, 21(1):10–13, 2014.

[27] Frank Nielsen and Richard Nock. On $w-$mixtures: Finite convex combinations of prescribed component distributions. *CoRR*, abs/1708.00568, 2017.

[28] Frank Nielsen and Ke Sun. Guaranteed bounds on information-theoretic measures of univariate mixtures using piecewise log-sum-exp inequalities. *Entropy*, 18(12):442, 2016.

[29] Frank Nielsen and Ke Sun. Guaranteed deterministic bounds on the total variation distance between univariate mixtures. In *IEEE Machine Learning in Signal Processing (MLSP)*, pages 1–6, 2018.

[30] Michael A Nielsen and Isaac Chuang. Quantum computation and quantum information, 2002.

[31] Ferdinand Österreicher and Igor Vajda. A new class of metric divergences on probability spaces and its applicability in statistics. *Annals of the Institute of Statistical Mathematics*, 55(3):639–653, 2003.

[32] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in Python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.

[33] József Pitrik and Dániel Virosztek. On the joint convexity of the Bregman divergence of matrices. *Letters in Mathematical Physics*, 105(5):675–692, 2015.

[34] Carl Edward Rasmussen. The infinite Gaussian mixture model. In *Advances in neural information processing systems*, pages 554–560, 2000.

[35] Douglas A Reynolds, Thomas F Quatieri, and Robert B Dunn. Speaker verification using adapted Gaussian mixture models. *Digital signal processing*, 10(1-3):19–41, 2000.

[36] Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. The earth mover's distance as a metric for image retrieval. *International journal of computer vision*, 40(2):99–121, 2000.

[37] Filippo Santambrogio. Optimal transport for applied mathematicians. *Birkäuser, NY*, pages 99–102, 2015.

[38] Olivier Schwander and Frank Nielsen. Learning mixtures by simplifying kernel density estimators. In *Matrix Information Geometry*, pages 403–426. Springer, 2013.

[39] Jorge Silva and Shrikanth Narayanan. Upper bound Kullback-Leibler divergence for hidden Markov models with application as discrimination measure for speech recognition. In *IEEE International Symposium on Information Theory (ISIT)*, pages 2299–2303. IEEE, 2006.

[40] Yoram Singer and Manfred K Warmuth. Batch and on-line parameter estimation of Gaussian mixtures based on the joint entropy. In *Advances in Neural Information Processing Systems*, pages 578–584, 1999.

[41] Asuka Takatsu. Wasserstein geometry of Gaussian measures. *Osaka Journal of Mathematics*, 48(4):1005–1026, 2011.

[42] Li Xie, Valery A. Ugrinovskii, and Ian R. Petersen. Probabilistic distances between finite-state finite-alphabet hidden Markov models. *IEEE transactions on automatic control*, 50(4):505–511, 2005.