Information Geometry for Regularized Optimal Transport and Barycenters of Patterns

Shun-ichi Amari
amari@brain.riken.jp
RIKEN, Wako-shi, Saitama 351-0198, Japan

Ryo Karakida
karakida.ryo@aist.go.jp
National Institute of Advanced Industrial Science and Technology,
Koto-ku, Tokyo 135-0064, Japan

Masafumi Oizumi
oizumi@araya.org
ARAYA, Minato-ku, Tokyo 105-0003, Japan

Marco Cuturi
cuturi@google.com
Google, 75009 Paris and CREST, ENSAE, 91120 Palaiseau, France

We propose a new divergence on the manifold of probability distributions, building on the entropic regularization of optimal transportation problems. As Cuturi (2013) showed, regularizing the optimal transport problem with an entropic term is known to bring several computational benefits. However, because of that regularization, the resulting approximation of the optimal transport cost does not define a proper distance or divergence between probability distributions. We recently tried to introduce a family of divergences connecting the Wasserstein distance and the Kullback-Leibler divergence from an information geometry point of view (see Amari, Karakida, & Oizumi, 2018). However, that proposal was not able to retain key intuitive aspects of the Wasserstein geometry, such as translation invariance, which plays a key role when used in the more general problem of computing optimal transport barycenters. The divergence we propose in this work is able to retain such properties and admits an intuitive interpretation.

1 Introduction

Two major geometrical structures have been introduced on the probability simplex, the manifold of discrete probability distributions. The first one is based on the principle of invariance, which requires that the geometry
between probability distributions must be invariant under invertible transformations of random variables. That viewpoint is the cornerstone of the theory of information geometry (Amari, 2016), which acts as a foundation for statistical inference. The second direction is grounded on the theory of optimal transport, which exploits prior geometric knowledge on the base pace in which random variables are valued (Villani, 2003). Computing optimal transport amounts to obtaining a coupling between these two random variables that is optimal in the sense that it has a minimal expected transportation cost between the first and second variables.

However, computing that solution can be challenging and is usually carried out by solving a linear program. Cuturi (2013) considered a relaxed formulation of optimal transport, in which the negative entropy of the coupling is used as a regularizer. We call that approximation of the original optimal transport cost the C function. Entropic regularization provides two major advantages: the resolution of regularized optimal transport problems, which relies on Sinkhorn’s algorithm (Sinkhorn, 1964), can be trivially parallelized and is usually faster by several orders of magnitude than the exact solution of the linear program. Unlike the original optimal transport geometry, regularized transport distances are differentiable functions of their inputs, even when the latter are discrete, a property that can be exploited in problems arising from pattern classification and clustering (Cuturi & Doucet, 2014; Cuturi & Peyré, 2016).

The Wasserstein distance between two distributions reflects the metric properties of the base space on which a pattern is defined. Conventional information-theoretic divergences such as the Kullback-Leibler (KL) divergence and the Hellinger divergence fail to capture the geometry of the base space. Therefore, the Wasserstein geometry is useful for applications where the geometry of the base space plays an important role, notably in computer vision, where an image pattern is defined on the two-dimensional base space $\mathbb{R}^2$. Cuturi and Doucet (2014) is a pioneering work, where the Wasserstein barycenter of various patterns can summarize the common shape of a database of exemplar patterns. Also, more advanced inference tasks use the C function as an output loss (Frogner, Zhang, Mobahi, Araya, & Poggio, 2015), a model fitting loss (Antoine, Marco, & Gabriel, 2016), or a way to learn mappings (Courty, Flamary, Tuia, & Rakotomamonjy, 2017). The Wasserstein geometry was also shown to be useful to estimate generative models, under the name of the Wasserstein GAN (Arjovsky, Chintala, & Bottou, 2017) and related contributions (Aude, Gabriel, & Marco, 2018; Salimans, Zhang, Radford, & Metaxas, 2018). It was also applied to economics and ecological systems (Muzellec, Nock, Patrini, & Nielsen, 2016). All in all, the field of “Wasserstein statistics” is becoming an interesting research topic for exploration.

The C function suffers, however, from a few issues. It is neither a distance nor a divergence, notably because comparing a probability measure with itself does not result in a null discrepancy, namely, if $p$ belongs to the
simplex, then \( C(p, p) \neq 0 \). More worrisome, the minimizer of \( C(p, q) \) with respect to \( q \) is not reached at \( q = p \). To solve these issues, we have proposed a first attempt at unifying the information and optimal transport geometrical structures in Amari, Karakida, and Oizumi (2018). However, the information-geometric divergence introduced in that previous work loses some of the beneficial properties inherent in the \( C \)-function. For example, the \( C \)-function can be used to extract a common shape as the barycenter of a number of patterns (Cuturi & Doucet, 2014), which our former proposal was not able to. Therefore, it is desirable to define a new divergence from \( C \), in the rigorous sense that it is minimized when comparing a measure with itself and, preferably, convex in both arguments, while still retaining the attractive properties of optimal transport.

We propose in this article such a new divergence between probability distributions \( p \) and \( q \) that is inspired by optimal transport while incorporating elements of information geometry. Its basic ingredient remains the entropic regularization of optimal transport. We show that the barycenters obtained with that new divergence are more sharply defined than those obtained with the original \( C \)-function, while still keeping the shape-location decomposition property. This article focuses on theoretical aspects, with only very preliminary numerical simulations. Detailed characteristics and simulations of the proposed divergence will be given in the future.

2 C-Function: Entropy-Regularized Optimal Transportation Plan

The general transportation problem is concerned with the optimal transportation of commodities, initially distributed according to a distribution \( p \), so that they end up being distributed as another distribution \( q \). In its full generality, such a transport can be carried out on a metric manifold, and both \( p \) and \( q \) can be continuous measures on that manifold. We consider in this article the discrete case, where that metric space is of finite size \( n \), namely, \( X = \{1, 2, \ldots, n\} \). We normalize the total number of commodities such that it sums to 1; \( p \) and \( q \) are therefore probability vectors of size \( n \) in the \( n-1 \)-dimensional simplex,

\[
S_{n-1} = \left\{ p \in \mathbb{R}^n \mid \sum_i p_i = 1, \quad p_i \geq 0 \right\}.
\] (2.1)

We leave out extensions to the continuous case for future work. Let \( M_{ij} \) be the cost of transporting a unit of commodity from bin \( i \) to bin \( j \), usually defined as a distance (or a suitable power thereof) between points \( i \) and \( j \) in \( X \). In what follows, we assume that \( M_{ij} > 0 \) for \( i \neq j \) and \( M_{ii} = 0 \). A transportation plan \( P = (P_{ij}) \in \mathbb{R}_{+}^{n \times n} \) is a joint (probability) distribution over \( X \times X \) that describes at each entry \( P_{ij} \) the amount of commodities sent from bin \( i \) to bin \( j \). Given a source distribution \( p \) and a target distribution \( q \), the...
set of transport plans allowing a transfer from $p$ to $q$ is defined as

$$U(p, q) = \left\{ P \in \mathbb{R}^{n \times n}_+ : \forall i \leq n, \sum_{j=1}^{n} P_{ij} = p_i, \forall j \leq n, \sum_{i=1}^{n} P_{ij} = q_j \right\}.$$  \hfill (2.2)

Note that if a matrix is in $U(p, q)$, then its transpose is in $U(q, p)$.

The transportation cost of a plan $P$ is defined as the dot product of $P$ with the cost matrix $M = (M_{ij})$,

$$\langle M, P \rangle = \sum_{ij} M_{ij}P_{ij}. \hfill (2.3)$$

Its minimum among all feasible plans,

$$W(p, q) = \min_{P \in U(p, q)} \langle M, P \rangle, \hfill (2.4)$$

is the Wasserstein distance on $S_{n-1}$ parameterized by the ground metric (Cuturi, 2013) studied the regularization of that problem using entropy,

$$H(P) = - \sum_{ij} P_{ij} \log P_{ij}, \hfill (2.5)$$

to consider the problem of minimizing

$$\mathcal{L}(P) = \langle M, P \rangle - \lambda H(P), \hfill (2.6)$$

where $\lambda > 0$ is a regularization strength. We call its minimum the $C$-function:

$$C_{\lambda}(p, q) = \min_{P \in U(p, q)} \mathcal{L}(P). \hfill (2.7)$$

The $C_{\lambda}$ function is a useful proxy for the Wasserstein distance, with favorable computational properties, and has appeared in several applications as a useful alternative to information-geometric divergences such as the KL divergence.

The optimal transportation plan is given in the following theorem (Cuturi & Peyré, 2016; Amari et al., 2018).
Theorem 1. The optimal transportation plan \( P_\lambda^* \) is given by
\[
P_\lambda^* = \frac{ca_i b_j K_{ij}}{\lambda},
\]
(2.8)
\[
K = \exp\left(\frac{-M_{ij}}{\lambda}\right),
\]
(2.9)
where \( c, a, b \in S_{n-1} \) are determined from \( p \) and \( q \) such that the sender and receiver conditions (namely, marginal conditions) are satisfied.

In our previous paper (Amari et al., 2018), we studied the information geometry of the manifold of optimal transportation plans. We proposed a family of divergences that combines the KL divergence and the Wasserstein distance. However, these divergences are closer in spirit to the KL divergence and therefore lose some crucial properties of the \( C \)-function. In this work, we define a new family of divergences directly from the \( C \)-function.

3 Divergence Derived from \( C \)-Function

The \( C \)-function \( C_\lambda(p, q) \) does not satisfy the requirements for a distance or divergence. For such a function \( \Delta \), for any \( p, q \in S_{n-1} \),
\[
\Delta(p, q) \geq \Delta(p, p) = 0.
\]
(3.1)

In order to find the minimizer \( q^* \) of \( C_\lambda(p, q) \) for a given \( p \), we use the exponentiated version \( K_\lambda \) of the cost \( M \) depending on \( \lambda \) given in equation 2.9.

We further define its conditional version:
\[
\tilde{K}_\lambda = \begin{bmatrix} K_{ji, \lambda} \\ K_j \cdot \end{bmatrix},
\]
(3.2)
\[
K_j \cdot = \sum_i K_{ji}.
\]
(3.3)
\( \tilde{K}_\lambda \) is a linear operator from \( S_{n-1} \) into \( S_{n-1} \), and we will use for convenience the notation
\[
\tilde{q} = \tilde{K}_\lambda q.
\]
(3.4)
\( \tilde{K}_\lambda \) is a monotonic shrinkage operator mapping \( S_{n-1} \) in its interior (see Figure 1), and \( \tilde{K}_\lambda S_{n-1} \subset \tilde{K}_{\lambda'} S_{n-1} \) for \( \lambda > \lambda' \). When \( \lambda = 0 \), \( \tilde{K}_\lambda \) is the identity mapping
\[
\tilde{K}_0 q = q.
\]
(3.5)
As $\lambda$ tends to infinity, $\tilde{K}_\lambda S_{n-1}$ converges to a single point, the center of $S_{n-1}$, $1/n$, where $1 = (1, 1, \ldots, 1)^T$, and

$$\tilde{K}_\infty = \frac{1}{n} [1 \cdots 1].$$  \hfill (3.6)

Hence, for any $q$, $\tilde{K}_\infty q$ is the uniform distribution $1/n$.

**Theorem 2.** The minimizer of $C_\lambda(p, q)$ with respect to $q$, given $p$, is

$$q^* = \tilde{K}_\lambda p.$$  \hfill (3.7)

**Proof.** By differentiation, we have the equation

$$\partial_q C_\lambda(p, q^*) = 0$$  \hfill (3.8)

to determine the minimizer $q^*$. This gives the condition

$$b_i = 1,$$  \hfill (3.9)

for $q^*$ (see the duality theorem in Amari et al., 2018). Hence, the optimal transportation plan from $p$ to $q^*$ is given by

$$P^*_{ij} = ca_i K_{ij},$$  \hfill (3.10)

where suffix $\lambda$ is omitted to alleviate notations. From

$$\sum_j ca_i K_{ij} = p_i,$$  \hfill (3.11)

we have

$$ca_i = \frac{p_i}{K_i}.$$  \hfill (3.12)
Figure 2: Comparison of $C_\lambda$ and $D_\lambda$. (a) $C_\lambda$ as the function of $q$. $C_\lambda$ is minimized when $q = \bar{K}p$. (b) $D_\lambda$ as the function of $q$. $D_\lambda$ is minimized when $q = p$.

So

$$q^* = \bar{K}_\lambda p,$$  

(3.13)

proving the theorem.

We define a new family of divergences $D_\lambda(p, q)$ that also depend on $\lambda$.

**Definition 1.**

$$D_\lambda[p : q] = (1 + \lambda) \left( C_\lambda(p, \bar{K}_\lambda q) - C_\lambda(p, \bar{K}_\lambda p) \right).$$  

(3.14)

Figure 2 compares $C_\lambda$ and $D_\lambda$ in $S_{n-1}$. The following theorem is obtained (the proof is given in appendix A).

**Theorem 3.** $D_\lambda[p : q]$ is a convex function with respect to $p$ and $q$, satisfying the constraints of equation 3.1. It converges to the Wasserstein distance as $\lambda \to 0$.

**4 Behavior of $\bar{K}_\lambda$**

The divergence $D_\lambda$ is defined through $\bar{K}_\lambda$. We study properties of $\bar{K}_\lambda$, including two limiting cases of $\lambda \to 0, \infty$.

We first consider the case for small $\lambda$ to see how $\bar{K}_\lambda$ behaves, assuming that $X$ has a graphical structure. We assume that $M_{ii} = 0, M_{ij} = 1$ when $i$ is a nearest neighbor of $j$, and $M_{ij} > 1$ otherwise. By putting

$$\epsilon = \exp \left( -\frac{1}{\lambda} \right), \quad \lambda = -\frac{1}{\log \epsilon},$$  

(4.1)

we have

$$K_{ij} = \exp \left( -\frac{M_{ij}}{\lambda} \right) = \epsilon^{M_{ij}},$$  

(4.2)
\[ \tilde{K}_{ij} = \frac{\epsilon^{M_{ji}}}{\sum_i \epsilon^{M_{ji}}} . \quad (4.3) \]

Then, by neglecting higher-order terms of \( \epsilon \), we have

\[
\tilde{K}_{ij} = \begin{cases} 
1 - |N(i)|\epsilon, & i = j, \\
\epsilon, & i \in N(j), \\
0, & \text{otherwise,}
\end{cases} \quad (4.4)
\]

where \( N(j) \) is the set of nearest neighbors of \( j \). When \( X \) is a plane consisting of \( n \times n \) pixels, a neighbor consists of four pixels, \( |N(j)| = 4 \), except for boundary pixels. We see that \( \tilde{K}_{ij} \) is approximated by the discrete Laplacian operator \( \Delta \),

\[
\tilde{K} = (1 - \epsilon)I + \epsilon \Delta . \quad (4.5)
\]

This shows that \( \tilde{K} \) is a diffusion operator, flattening pattern \( q \), that is, shifting \( q \) toward the uniform distribution \( 1/n \).

The inverse of \( \tilde{K} \) is

\[
\tilde{K}^{-1} = (1 + \epsilon)I - \epsilon \Delta . \quad (4.6)
\]

This is the inverse diffusion operator, which sharpens \( q \) by emphasizing larger components.

In order to make clear the character of diffusion without assuming \( \lambda \) is small, we consider a continuous pattern \( p = p(\xi), \xi \in \mathbb{R}^n \) and that metric

\[
M(\xi, \xi') = |\xi - \xi'|^2 . \quad (4.7)
\]

Then,

\[
K_{ij}(\xi, \xi') = \exp \left( -\frac{|\xi - \xi'|^2}{\lambda} \right) , \quad (4.8)
\]

and we easily have

\[
\tilde{K}_{ij}(\xi|\xi') = \frac{1}{(\sqrt{\pi} \lambda)^n} \exp \left( -\frac{|\xi - \xi'|^2}{\lambda} \right) . \quad (4.9)
\]
This is a diffusion kernel. When $p$ is a gaussian distribution,

$$
p(\xi) = \exp \left( -\frac{1}{2\sigma^2} |\xi|^2 \right), \quad (4.10)
$$

we have

$$
\tilde{K}_\lambda p = \exp \left( -\frac{1}{2\tau^2} |\xi|^2 \right), \quad (4.11)
$$

with

$$
\tau^2 = \sigma^2 + \frac{\lambda}{2}. \quad (4.12)
$$

Hence, $\tilde{K}_\lambda p$ is gaussian, where the variance $\tau^2$ is increased by $\lambda/2$, blurring the original $p$.

Finally, we consider the case when $\lambda$ is large enough, studying the limiting behavior as $\lambda \to \infty$. When $\lambda$ is large, we expand $K_{ij}$ as

$$
K_{ij,\lambda} = \exp \left( -\frac{M_{ij}}{\lambda} \right) = 1 - \frac{M_{ij}}{\lambda}, \quad (4.13)
$$

$$
K_{j,\lambda} = n \left( 1 - \bar{m}_j \right), \quad \bar{m}_j = \frac{1}{n} \sum_i M_{ji}, \quad (4.14)
$$

obtaining

$$
\tilde{K}_{ij,\lambda} = \frac{K_{j,\lambda}}{K_{i,\lambda}} = \frac{1}{n} \left( 1 - \bar{m}_{ij} \right), \quad \bar{m}_{ij} = M_{ji} - \bar{m}_j. \quad (4.15)
$$

Hence,

$$
\tilde{K}_\lambda (q - p)_i = \frac{1}{n\lambda} \sum_j \bar{m}_{ij} (q_j - p_j), \quad (4.16)
$$

showing that this is of order $1/\lambda$. Let $\tilde{M}$ be the moment matrix defined by

$$
\tilde{M}_{jk} = \frac{1}{n} \sum_i \bar{m}_{ij} \bar{m}_{jk}. \quad (4.17)
$$

Then, we have the following theorem (the proof is given in appendix B).
**Theorem 4.** When $\lambda$ is large enough,

$$
\lim_{\lambda \to \infty} D_\lambda[p : q] = \frac{1}{2}(q - p)^T \tilde{M}(q - p),
$$

which is a squared energy distance defined by the moment matrix $\tilde{M}$.

## 5 Right Barycenter of Patterns

We consider the barycenter of image patterns represented as probability measures $p = p(\xi)$ on the plane $\xi = (x, y) \in \mathbb{R}^2$ using divergence $D_\lambda$. The plane is discretized into a grid of $n \times m$ pixels, and therefore $p$ is a probability vector of size $nm$.

Let us consider a family $S$ of patterns, $S = (p_i)_{i=1}^N$. A right $D$-barycenter $q^*_D(S)$ of these patterns is the minimizer of

$$
F_D(S, q) = \sum_i D_\lambda[p_i : q].
$$

Cuturi and Doucet (2014) used $C_\lambda(p, q)$ to define the following barycenter, as a minimizer of

$$
F_C(S, q) = \sum_i C_\lambda(p_i, q).
$$

We call such a minimizer the $C$-barycenter $q^*_C(S)$. Cuturi and Doucet (2014) showed that the $C$-barycenter can extract a common shape for some pattern families $S$. In particular, they used a family $S$ of deformed double rings of various sizes and positions, whose barycenter $q^*_C(S)$ turns out to be a standard double-ring pattern. Other information-theoretic divergences such as the KL divergence or Hellinger divergence are not able to recover such a common shape. We will show that the right $D$-barycenter $q^*_D(S)$ exhibits the same property, but with a sharper solution.

The right $D$-barycenter minimizes

$$
\frac{1}{1 + \lambda} \sum_i D_\lambda(p_i : q) = \sum_i C_\lambda(p_i, \tilde{K}_\lambda q) - \sum_i C_\lambda(p_i, \tilde{K}_\lambda p_i).
$$

The second term of the right-hand side of equation 5.3 does not depend on $q$, so it may be deleted for minimization. Let us put

$$
\tilde{q} = \tilde{K}_\lambda q.
$$

Then the $D$-barycenter is derived from the $C$-barycenter by

$$
q^*_C = \tilde{K}_\lambda q^*_D.
$$
provided equation 5.5 is solvable. In this case,

$$q_D^* = K_x^{-1} q_C^*$$  \hfill (5.6)

is a sharpened version of $q_C^*$. However, equation 5.5 might not always be solvable.

The image $K_x S_{n-1}$ is a simplex sitting inside $S_{n-1}$. Since the $C$-barycenter $q_C^*$ is not necessarily inside $K_x S_{n-1}$, we need to solve the $D$-barycenter problem, equation 5.3, under the constraint that $q$ is constrained inside $K_x S_{n-1}$. When the $C$-barycenter $q_C^*$ is inside $K_x S_{n-1}$, the $D$-barycenter $q_D^*$ is simply given by equation 5.5, which is more localized or sharper than $q_C^*$. When $q_C^*$ is not inside $K_x S_{n-1}$, the solution of equation 5.3 is close to the boundary of the simplex $K_x S_{n-1}$, giving a sharper version.

**Theorem 5.** The right $D$-barycenter $q_D^*(S)$ of $S$ is a sharper (more localized) version of the $C$-barycenter.

Agueh and Carlier (2011) showed that when the ground metric is the quadratic Euclidean distance, the $W$-barycenter has the property that its shape is determined from the shapes of each element $p_i$ in $S$, but does not depend on their location; namely, it is translation invariant. The $C$-barycenter also inherits this property, and we show that so does the right $D$-barycenter.

Let us consider a pattern $p(\xi) = p(x, y)$ on the $(x, y)$-plane, where $\xi = (x, y)$. The center $\xi_p$ of $p(x, y)$ is defined by

$$\xi_p = \int p(\xi) d\xi.$$  \hfill (5.7)

Given a $\tilde{\xi}$, we define a shift operator $T_{\tilde{\xi}}$, which shifts $p(x, y)$ by $\tilde{\xi} = (\tilde{x}, \tilde{y})$:

$$T_{\tilde{\xi}} p(\xi) = p(\xi - \tilde{\xi}).$$  \hfill (5.8)

Let $P(\xi, \xi')$ be a transportation plan from $p(\xi)$ to $q(\xi)$. When $q(\xi)$ is shifted as $T_{\tilde{\xi}} q(\xi)$, we naturally define the transportation plan $T_{\tilde{\xi}} P(\xi, \xi')$:

$$T_{\tilde{\xi}} P(\xi, \xi') = P(\xi, \xi' - \tilde{\xi}).$$  \hfill (5.9)

which transports $p(\xi)$ to $T_{\tilde{\xi}} q(\xi)$.

We study how the transportation cost changes by a shift. As recalled from above, we use the squared Euclidean distance as the ground cost:

$$m(\xi, \xi') = \|\xi - \xi'\|^2.$$  \hfill (5.10)
Then the direct cost of transportation is

$$\langle M, P \rangle = \int m(\xi, \xi') P(\xi, \xi') d\xi d\xi'. \quad (5.11)$$

The cost of the shifted plan is

$$\langle M, T_\xi P \rangle = \int m(\xi, \xi') P(\xi, \xi' - \xi) d\xi d\xi' \quad (5.12)$$

$$= \int m(\xi, \xi'' + \xi) P(\xi, \xi'') d\xi d\xi'' \quad (5.13)$$

$$= \int (\|\xi - \xi''\|^2 + \|\xi\|^2 - 2\xi \cdot (\xi - \xi'')) P(\xi, \xi'') d\xi d\xi'' \quad (5.14)$$

$$= \langle M, P \rangle + \|\xi\|^2 - 2\xi \cdot (\xi_p - \xi_q). \quad (5.15)$$

Note that a shift does not change the entropy,

$$H\{P(\xi, \xi')\} = H\{T_\xi P(\xi, \xi')\}. \quad (5.16)$$

When $\xi_p = \xi_q$, two patterns $p$ and $q$ are said to be co-centered. For two co-centered $p, q$, we have

$$\mathcal{L}(P_{p,q}) \geq \mathcal{L}(P_{p,\xi_p p}), \quad (5.17)$$

where $P_{p,q}$ is a transportation plan sending $p$ to $q$. Hence, $q^* = K_\lambda p$ minimizes the transportation cost among all co-centered $q$.

We fix $\xi_q$ and search for the optimal shape $q^*(\xi)$ located at $\xi_q$ that minimizes the transportation cost $C_\lambda(P_{p,q})$ from $p$ to $q$. In order to derive the optimal shape, we shift $q$ by $T_{\tilde{\xi}}$,

$$\tilde{\xi} = \xi_q - \xi_p, \quad (5.18)$$

such that $p$ and $\tilde{q} = T_{\tilde{\xi}} q$ are co-centered. Then, for a plan $P_{p,q}$, we have $P_{p,q}$, which is the shifted plan of $P_{p,q}$ by $-\tilde{\xi}$:

$$P_{p,q} = T_{-\tilde{\xi}} P_{p,q}, \quad (5.19)$$

$$q = T_{-\tilde{\xi}} \tilde{q}. \quad (5.20)$$

We have

$$C_\lambda(T_{\tilde{\xi}} P) = C_\lambda(P) + \|\tilde{\xi}\|^2 - 2\tilde{\xi} \cdot (\xi_p - \xi_q). \quad (5.21)$$
We have an important relation that $C_\lambda(P_{p,q})$ is decomposed into a sum of the shape deformation cost among co-centered patterns and the positional transportation cost as

$$C_\lambda(P_{p,q}) = C_\lambda(P_{p,q}) + \|\xi_p - \xi_q\|^2$$  (5.22)

because of

$$\xi_p = \xi_q.$$  (5.23)

**Lemma 1.** Given $p$, the $C_\lambda$-optimal pattern $q(\xi)$ transporting $p$ to $q$ is a shifted and blurred version of $p(\xi)$,

$$q^*(\xi) = T_{\xi_p - \xi_q} \bar{K}_p p(\xi),$$  (5.24)

not depending on the locations $\xi_p$ and $\xi_q$.

Let $p_1(\xi), \ldots, p_n(\xi)$ be $n$ local patterns. We search for their $C$- and right $D$-barycenters $q(\xi)$ that minimize

$$F_C(q) = \sum_{i=1}^n C_i(P_{p_i,q}), \quad F_D(q) = \sum_{i=1}^n D_i(P_{p_i,q}).$$  (5.25)

The center of $p_i(\xi)$ is denoted by $\xi_i = \xi_{p_i}$. Before solving the problem, let us respectively shift $p_i(\xi)$ from $\xi_i$ to $\xi_0$, a fixed common location, such that all shifted $\bar{p}_i(\xi)$’s are co-centered. Let $q^*_C(\xi)$ and $q^*_D(\xi)$ be the barycenters of all co-centered $\bar{p}_i(\xi)$, which do not depend on the locations of $p_i(\xi)$ but their shapes.

**Theorem 6** (shape-location separation theorem). The barycenters $q^*_C(\xi)$ and $q^*_D(\xi)$ of $p_1(\xi), \ldots, p_n(\xi)$ are located at the barycenter of $\xi_1, \ldots, \xi_n$, and their shapes are given by the respective barycenters of $\bar{p}_1(\xi), \ldots, \bar{p}_n(\xi)$.

**Proof.** From equation 5.22, we have

$$F_C(q) = \sum_{i=1}^n C_i(P_{p_i,q}) + \sum \|\xi_{p_i} - \xi_q\|^2,$$  (5.26)

$$F_D(q) = \sum_{i=1}^n C_i(P_{p_i,\bar{K}_p q}) + \sum \|\xi_{p_i} - \xi_q\|^2,$$  (5.27)

where $\bar{q}_i$ is the shifted version of $q$ to the center of $p_i$. Here, the shape and location of $q$ are separated. Minimizing the first term, we have $q^*$, which is the respective barycenters of the shapes of co-centered $p_1, \ldots, p_n$. The second term gives the barycenter of locations $\xi_1, \ldots, \xi_n$. □
Corollary 1. When $p_i$ are shifts of an identical $p$, their right $D$-barycenter has the same shape as the original $p$, whereas the $C$-barycenter is a blurred version of $p$:

$$q^*_C = K_\lambda p. \quad (5.28)$$

We show a simple example where $p_i$ are shifted $p$, a cat shape (see Figure 3a). Its $C$-barycenter has a blurred shape $K_\lambda p$ (see Figure 3b) tending to the uniform distribution as $\lambda \to \infty$. However, the shape of the right $D$-barycenter is exactly the same as $p$ (see Figure 3c).

6 Left Barycenter of Patterns

Since we use the assymetric divergence $D_\lambda$ to define a barycenter, we may consider another barycenter by putting the unknown barycenter in the left argument of $D_\lambda$.

We consider again a family $S$ of patterns, $S = (q_i)_{i=1,...,N}$. A barycenter $p$ of these patterns based on divergence $D_\lambda$ is defined by the minimizer of

$$F^l_D(S, p) = \sum_i D_\lambda[p : q_i]$$

$$= \sum_i [C_\lambda(p, \tilde{K}_\lambda q_i) - C_\lambda(p, \tilde{K}_\lambda p)] \quad (6.2)$$

and is called the left $D$-barycenter. We propose to solve that problem using the accelerated gradient descent approach outlined in Cuturi and Doucet (2014), with two differences: all examples $q_i$ must be smoothed beforehand following an application of $\tilde{K}_\lambda$, and the gradient now incorporates not only terms resulting from $C_\lambda(p, \tilde{K}_\lambda q_i)$ but also from $-C_\lambda(p, \tilde{K}_\lambda p)$, which, as explained in equation A.3, is simply minus the KL divergence between $p$ and the vector $\tilde{K}_\lambda 1$, the entropy of $p$ plus the dot product between $p$ and $\log(\tilde{K}_\lambda 1)$. As a result the gradient of $-C_\lambda(p, \tilde{K}_\lambda p)$ is equal to, up to a

---

Figure 3: (a) Cat images. (b) The $C$-barycenter of panel a. The $C$-barycenter has a blurred shape $K_\lambda p$, more blurred as $\lambda \to \infty$. (c) The $D$-barycenter of panel a. The shape of the right $D$-barycenter is exactly the same as the original shape $p$. 

---

$840$ S.-i. Amari, R. Karakida, M. Oizumi, and M. Cuturi
constant term, $-\log(p) + \log(K_\lambda 1)$, which tends to sharpen further the $C_\lambda$ barycenter. Note that this approach, namely, adding the entropy to the regularized Wasserstein barycenter problem, was used in a heuristic way by Solomon et al. (2015), who called it *entropic sharpening*, without proving that the resulting problem was convex. Our work shows that up to a given strength, the entropic sharpening of regularized Wasserstein barycenters remains a convex problem.

The shape-location separation theorem, theorem 6, also holds for the left barycenter. We note that $\tilde{q}_i$ and $K_\lambda q_i$ have the same center. So we shift all $\tilde{q}_i = K_\lambda q_i$ to a common location $\xi$. Since the additional term $C_\lambda(p, K_\lambda q)$ is shift invariant, the terms concerning the shapes and locations are separated, as shown in equation 6.1. An example of the left $D$-barycenter of four patterns is shown in Figures 4 and 5.

7 Preliminary Simulations

It is important to study properties of the two $D$-barycenters, theoretically as well as by simulations. However, we need enough computational facilities for large-scale simulations. Here, we use the simple exponentiated gradient descent algorithm to obtain the $D$-barycenters as a preliminary study. We will explore more detailed study in forthcoming papers.

The exponentiated gradient updates the current $q(p)$ to the next $q$ as follows:

Update: $q'_i \leftarrow q_i \exp(-\eta \nabla D)$, \hspace{1cm} (7.1)

Normalize: $q^*_i = \frac{q'_i}{\sum q'_j}$, \hspace{1cm} (7.2)

giving the next candidate $q^*$ in the case of the right $D$-barycenter, where $\nabla$ is the gradient operator $\partial/\partial q_i$. Here, $\eta$ is a learning constant. (The same holds
Figure 5: We consider in this example the squared-Euclidean distance on grids as the ground metric cost. We study the iso-barycenter of the four shapes presented in Figure 4 using two different discrepancy functions, $C_\lambda$ and $D_\lambda$. (Left) Barycenter obtained using algorithm 2 of Solomon et al. (2015), itself a specialized version of Benamou, Carlier, Cuturi, Nenna, and Peyré’s (2014) algorithm. We used a $10^{-10}$ tolerance for the $l_1$ norm between two successive iterates as a stopping criterion. (Right) $D_\lambda$ left-barycenter obtained with our accelerated gradient approach. We use a jet color map to highlight differences in the support of the barycenters. As expected, the entropy of the $C_\lambda$ barycenter is higher than that of the $D_\lambda$ left barycenter, since the latter optimization incorporates a penalization term, $-C_\lambda(p, \tilde{K}_\lambda p)$, which is equivalent to penalizing the entropy of the solution $p$. This difference in entropy results in subtle yet visible differences between the two solutions, with sharper edges for the $D_\lambda$ left barycenter.

by using $p$ instead of $q$ for the left $D$-barycenter.) As initial $q(p)$, we used the $C$-barycenter, which is easy to compute using Benamou et al. (2014).

We study a toy model searching for the barycenters of the two patterns: a large square and a small square, shown in Figure 6a. We compare the $C$-, right $D$-, and left $D$-barycenters shown in Figure 6b. Both $D$-barycenters give a sharper result compared to the $C$-barycenter for various $\lambda$. However, the right and left $D$-barycenters are very similar. Further studies are necessary to compare the two.

8 Conclusion

We defined new divergences between two probability distributions based on the $C$-function, which is the entropy-regularized cost function (Cuturi, 2013). Although it is useful in many applications, it does not satisfy the criterion of a distance or divergence. We defined a new divergence function $D_\lambda$ derived from $C_\lambda$, which works better than the original $C_\lambda$ for some problems—in particular, the barycenter problem. We proved that the minimizer of $C_\lambda(p, q)$ is given by $\tilde{K}_\lambda p$, where $\tilde{K}_\lambda$ is a diffusion operator.
depending on the base metric $M$. We studied properties of $\tilde{K}_\lambda$ showing how it changes as $\lambda$ increases, elucidating properties of $D_\lambda$.

We applied $D_\lambda$ to obtain the barycenter of a cluster of image patterns. It is proved that the right and left $D$-barycenters keep a good property that the shape and locations of patterns are separated, a merit of the $C$-function based barycenter. Moreover, the $D$-barycenters give even a sharper shape than the $C$-barycenter.

Computing $D$-barycenters provides a few novel numerical challenges. The algorithm we have proposed, essentially a simple multiplicative gradient descent, is not as efficient as that proposed by Benamou et al. (2014) for the $C$-barycenter. We believe there is room to improve our approach and leave this for future work.

Appendix A: Proof of Convexity of $D_\lambda$

Let us put the Hessian of the cost function as follows:

$$G_\lambda = \frac{\partial^2 C_\lambda(p, \tilde{K}_\lambda q)}{\partial \eta \partial \eta^T} = \begin{bmatrix} X & Y \\ Y^T & Z \end{bmatrix}, \quad (A.1)$$
expressed by block matrices $X$, $Y$, $Z$ where $\eta = (p, q)^T$. Because $C_\lambda$ is strictly convex (Amari et al., 2018), $G_\lambda$ is positive definite and the block component $Z$ is also regular and positive definite.

By using

$$p_{ij}^* = p_i \frac{K_{ij}}{K_i}, \quad (A.2)$$

we get

$$C_\lambda(p, \tilde{K}_\lambda p) = \lambda \sum_{i=1}^{n} p_i \ln \frac{p_i}{K_i}. \quad (A.3)$$

Therefore, its Hessian becomes

$$G'_\lambda = \frac{\partial^2 C_\lambda(p, \tilde{K}_\lambda p)}{\partial \eta \partial \eta^T} = \lambda \begin{pmatrix} \text{diag} \left( \frac{1}{p_i} \right) + \frac{1}{p_n} 11^T & O \\ O & O \end{pmatrix}, \quad (A.4)$$

where $O$ is the zero matrix and $\text{diag}(p_i)$ represents a diagonal matrix whose $ij$ component is given by $p_i \delta_{ij}$. Let us put $R = G_\lambda - G'_\lambda$. The determinant of $R$ is given by

$$\det(R) = \det(Z) \det(R'), \quad (A.5)$$

where we put

$$R' = X - YZ^{-1}Y^T - \lambda \left( \text{diag} \left( \frac{1}{p_i} \right) + \frac{1}{p_n} 11^T \right). \quad (A.6)$$

Because $Z$ is positive definite, the positive definiteness of $R$ is equivalent to that of $R'$. As derived in Amari et al. (2018),

$$G^{-1}_\lambda = \frac{1}{1 + \lambda} \begin{pmatrix} p_i \delta_{ij} - p_i p_j & P_{ij} - p_i q_i \\ P_{ji} - q_i p_j & q_i \delta_{ij} - q_i q_j \end{pmatrix}, \quad (A.7)$$

and we can represent the $(p, p)$-block part of $G^{-1}_\lambda$ by using the block components of $G_\lambda$ as follows:

$$(X - YZ^{-1}Y^T)^{-1} = \frac{1}{1 + \lambda} \left( \text{diag}(p_i) - pp^T \right). \quad (A.8)$$
By using the Sherman-Morrison formula, we get

$$X - YZ^{-1}Y^T = (1 + \lambda) \left( \text{diag} \left( \frac{1}{p_i} \right) + \frac{1}{p_n} \mathbf{1} \mathbf{1}^T \right).$$  \hfill (A.9)

Finally, we obtain

$$R' = \text{diag} \left( \frac{1}{p_i} \right) + \frac{1}{p_n} \mathbf{1} \mathbf{1}^T.$$  \hfill (A.10)

Because this $R'$ is positive definite, $R$ is also positive definite and $D\lambda$ is strictly convex.

By using equation 3.5 and $C_0(p, p) = 0$, we can confirm that $D\lambda$ converges to the Wasserstein distance as $\lambda \to 0$.

Appendix B: Proof of $D\lambda$ Approaching an Energy Function for Large $\lambda$

We expand $D\lambda[p, q]$ in terms of $\tilde{K}_\lambda(q - p)$ as

$$\frac{1}{(1 + \lambda)} D\lambda[p, q] = C_\lambda(p, \tilde{K}_\lambda p + \tilde{K}(q - p)) - C_\lambda(p, \tilde{K}_\lambda p)$$

$$= \partial_q C_\lambda(p, \tilde{K}_\lambda p) \cdot \tilde{K}_\lambda(q - p)$$

$$+ \frac{1}{2} \partial_q \partial_q C(p, \tilde{K}_\lambda p) \cdot \tilde{K}_\lambda(q - p) \otimes \tilde{K}_\lambda(q - p)$$

$$= \frac{1}{2} \partial_q \partial_q C(p, \tilde{K}_\lambda p) \cdot \tilde{K}_\lambda(q - p) \otimes \tilde{K}_\lambda(q - p),$$  \hfill (B.1)

because of

$$\partial_q C_\lambda(p, \tilde{K}_\lambda p) = 0,$$  \hfill (B.2)

where $\otimes$ is the tensor product. Higher-order terms are neglected.

When $\lambda$ is large, we expand $K_\lambda$ as

$$K_{ij,\lambda} = \exp \left\{ - \frac{M_{ij}}{\lambda} \right\} = 1 - \frac{M_{ij}}{\lambda},$$  \hfill (B.5)

$$K_{j,\lambda} = n \left( 1 - \frac{\bar{m}_j}{\lambda} \right), \quad \bar{m}_j = \frac{1}{n} \sum_i M_{ji},$$  \hfill (B.6)
obtaining
\[ \tilde{K}_{ij,\lambda} = \frac{K_{ij}}{K_j} = \frac{1}{n} \left( 1 - \frac{\bar{m}_{ij}}{\lambda} \right), \quad \bar{m}_{ij} = M_{ji} - \bar{m}_j. \]  
(B.7)

Hence,
\[ \tilde{K}_\lambda(q-p)_i = \frac{1}{n\lambda} \sum_j \bar{m}_{ij} (q_j - p_j). \]  
(B.8)

We have
\[ \partial_q \partial_q C(p, \tilde{K}_\lambda p) = \partial_q \partial_q C_\lambda(p, 1) + O\left(\frac{1}{\lambda}\right). \]  
(B.9)

So we calculate \( \partial_q \partial_q C_\lambda(p, \tilde{q}) \) when
\[ \tilde{q}_i = (\tilde{K}_\lambda p)_i = \frac{1}{n} + \varepsilon_i \]  
(B.10)

and expand it up to \( O(\varepsilon^2) \).

The optimal transportation plan for \( p \rightarrow q \) is
\[ p_{ij}^* = ca_ib_j \]  
(B.11)

when \( \lambda \rightarrow \infty \), because \( K_{ij,\lambda} = 1 \). Hence,
\[ p_{ij}^* = p_iq_j. \]  
(B.12)

We have already obtained the inverse of \( \partial_q \partial_q C_\lambda(p, q) \) in equation 45 of the previous paper:
\[ G^{-1}_\lambda = \frac{1}{1 + \lambda} \left[ \frac{p_i\delta_{ij} - p_i p_j}{P_{ij}^* - q_i p_j} \frac{P_{ij}^* - p_i q_j}{q_i \delta_{ij} - q_i q_j} \right]. \]  
(B.13)

Hence, it is block-diagonal \( (P_{ij}^* - p_iq_j = 0) \), and the \( (q, q) \)-part of \( G_\lambda (\lambda \rightarrow \infty) \) is
\[ G_{\lambda,qq} = (1 + \lambda) \left[ q_i \delta_{ij} - q_i q_j \right]^{-1} \]  
(B.14)
\[ = (1 + \lambda) \left( \text{diag} \left( \frac{1}{q_i} \right) + \frac{1}{q_n} \mathbf{1}\mathbf{1}^T \right). \]  
(B.15)
In our case of \( q = 1/n \),
\[
G = \partial_q \partial_q C = (1 + \lambda)(n\delta_{ij} + n). \tag{B.16}
\]

We finally calculate
\[
G \cdot \tilde{K}_\lambda(q - p) \otimes \tilde{K}_\lambda(q - p)
= \left(1 + \frac{\lambda}{n^2} \right)^2 \sum \tilde{m}_{ij} \tilde{m}_{ik} (q_j - p_j) (q_k - p_k)
+ \left(1 + \frac{\lambda}{n^2} \right)^2 \sum \tilde{m}_{ij} (q_j - p_j)
= \frac{1 + \lambda}{\lambda^2} (q - p)^T \tilde{M}(q - p).
\]

Note that \( \sum_i \tilde{m}_{ij} = 0 \).

References


Received April 30, 2018; accepted December 8, 2018.