Noname manuscript No.
(will be inserted by the editor)

Information Geometry for Regularized Optimal Transport and Barycenters of Patterns

Shun-ichi Amari · Ryo Karakida · Masafumi Oizumi · Marco Cuturi

Abstract We propose a new divergence on the manifold of probability distributions, building upon the entropic regularization of optimal transportation problems. As shown in (Cuturi, 2013), regularizing the optimal transport problem with an entropic term is known to bring several computational benefits. However, because of that regularization, the resulting quantities do not define a proper distance or divergence between probability distributions. We have recently tried to introduce a family of divergences connecting the Wasserstein distance and the KL divergence from the information geometry point of view (see Amari et al. (2018)). However, that proposal was not able to retain key intuitive aspects of the Wasserstein geometry, such as translation invariances, which play a key role when used in the more general barycenter problem. The divergence we propose in this work is able to retain such properties and admits an intuitive interpretation.

Keywords Wasserstein distance · Kullback-Leibler divergence · Optimal transportation · Barycenter · Shape preservation

1 Introduction

Two major geometrical structures have been introduced on the probability simplex, the manifold of discrete probability distributions. The first one is

Shun-ichi Amari
2-1 Hirosawa, Wako-shi, Saitama, 351-0198, Japan
E-mail: amari@brain.riken.jp

Ryo Karakida
2-3-26 Aomi, Koto-ku, Tokyo, 135-0064, Japan

Masafumi Oizumi
2-8-10 Toranomon, Minato-ku, Tokyo, 105-0001, Japan

Marco Cuturi
CREST, ENSAE, Université Paris-Saclay

based on the principle of parameterization 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 space 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 metric 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 regularized optimal transport problem is usually easier and faster to compute than the solution of the linear program, and can be done using Sinkhorn's algorithm (1964); unlike the original optimal transport geometry, regularized transport distances are differentiable functions of their input, a property which can be exploited in problems arising from pattern classification and clustering (Cuturi and Avis, 2014; Cuturi and Peyré, 2016) as well as more advanced inference tasks that use the C function as an output loss (Frogner et al., 2015; Genevay et al., 2018), a model fitting loss (Rolet et al., 2016) or a way to learn mappings (Courty et al., 2017).

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 $\boldsymbol{p}$ belongs to the simplex, then $C(\boldsymbol{p}, \boldsymbol{p}) \neq 0$. More worryingly, the minimizer of $C(\boldsymbol{p}, \boldsymbol{q})$ with respect to $\boldsymbol{q}$ is not reached at $\boldsymbol{q} = \boldsymbol{p}$. To solve these issues, we have proposed a first attempt at unifying the information and optimal transport geometrical structures in (Amari et al., 2018). However, the information-geometric divergence introduced in that previous work loses some of the nice properties inherent to the C-function. For example, the C-function can be used to extract a common shape as the barycenter of several patterns (Cuturi and 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 paper such a new divergence between probability distributions $\boldsymbol{p}$ and $\boldsymbol{q}$ that is both 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, still keeping the shape-location decomposition property. We illustrate these new definitions with simple numerical illustrations.

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 $\boldsymbol{p}$, so that they end up being distributed as another distribution $\boldsymbol{q}$. In its full generality, such a transport can be carried out on a metric manifold, and both $\boldsymbol{p}$ and $\boldsymbol{q}$ be continuous measures on that manifold. We consider in this paper the discrete case, where that metric space is of finite size n, namely $X = \{1, 2, \ldots, n\}$. We normalize the total amount of commodities such that it sums to 1: $\boldsymbol{p}$ and $\boldsymbol{q}$ are therefore probability vectors of size n in the $n-1$ dimensional simplex,

$$S_{n-1} = \left\{ \boldsymbol{p} \in \mathbb{R}^n \mid \sum_i p_i = 1, \quad p_i \geq 0 \right\}. \tag{1}$$

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 only 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 $X \times X$ which describes at each entry P_{ij} the amount of commodities sent from bin i to bin j. Given a source distribution $\boldsymbol{p}$ and a target distribution $\boldsymbol{q}$, the set of transport plans allowing a transfer from $\boldsymbol{p}$ to $\boldsymbol{q}$ is defined as

$$U(\boldsymbol{p}, \boldsymbol{q}) = \left\{ \boldsymbol{P} \in \mathbb{R}_+^{n \times n} : \forall i \leq n, \sum_j P_{ij} = p_i, \quad \forall j \leq n, \sum_i P_{ij} = q_j \right\}. \tag{2}$$

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

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

$$\langle \boldsymbol{M}, \boldsymbol{P} \rangle = \sum_{ij} M_{ij} P_{ij}. \tag{3}$$

Its minimum among all feasible plans

$$W(\boldsymbol{p}, \boldsymbol{q}) = \min_{\boldsymbol{P} \in U(\boldsymbol{p}, \boldsymbol{q})} \langle \boldsymbol{M}, \boldsymbol{P} \rangle \tag{4}$$

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

$$H(\boldsymbol{P}) = -\sum_{ij} P_{ij} \log P_{ij}, \tag{5}$$

to consider the problem of minimizing

$$\mathcal{L}(\boldsymbol{P}) = \langle \boldsymbol{M}, \boldsymbol{P} \rangle - \lambda H(\boldsymbol{P}), \tag{6}$$

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

$$C_\lambda(\boldsymbol{p}, \boldsymbol{q}) = \min_{\boldsymbol{P} \in U(\boldsymbol{p}, \boldsymbol{q})} \mathcal{L}(\boldsymbol{P}). \tag{7}$$

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

The optimal transportation plan is given in the following Theorem (Cuturi and Peyré, 2016; Amari et al., 2018).

Theorem 1 The optimal transportation plan $\boldsymbol{P}^*_\lambda$ is given by

$$\boldsymbol{P}^*_\lambda = [c a_i b_j K_{ij}]_{ij}\,, \tag{8}$$

$$\boldsymbol{K} = \left[\exp\left(-\frac{M_{ij}}{\lambda}\right)\right]_{ij}, \tag{9}$$

where c, a normalization constant, and vectors $\boldsymbol{a}$, $\boldsymbol{b} \in S_{n-1}$ are determined from $\boldsymbol{p}$ and $\boldsymbol{q}$ 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 combine the KL divergence and the Wasserstein distance. However, these divergences are closer in spirit to the KL divergence, and lose therefore some crucial properties of the C-function. We define in this work a new family of divergences directly from the C-function.

3 Divergence Derived from C-Function

The C-function $C_\lambda(\boldsymbol{p}, \boldsymbol{q})$ does not satisfy the requirements for a distance or divergence, which would be that for such a function Δ we have that for any $\boldsymbol{p}, \boldsymbol{q} \in S_{n-1}$

$$\Delta(\boldsymbol{p}, \boldsymbol{q}) \geq \Delta(\boldsymbol{p}, \boldsymbol{p}) = 0. \tag{10}$$

In order to find the minimizer $\boldsymbol{q}^*$ of $C_\lambda(\boldsymbol{p}, \boldsymbol{q})$ for a given $\boldsymbol{p}$, we use the exponentiated version $\boldsymbol{K}_\lambda$ of the cost $\boldsymbol{M}$ depending on λ given in (9).

We further define its conditional version,

$$\tilde{\boldsymbol{K}}_\lambda = \left[\frac{K_{ji,\lambda}}{K_{j\cdot}}\right]_{ij}, \tag{11}$$

$$K_{j\cdot} = \sum_i K_{ji}. \tag{12}$$

$\tilde{\boldsymbol{K}}_\lambda$ is a linear operator from S_{n-1} into S_{n-1}, and we will use for convenience the notation

$$\tilde{\boldsymbol{q}} = \tilde{\boldsymbol{K}}_\lambda \boldsymbol{q}. \tag{13}$$

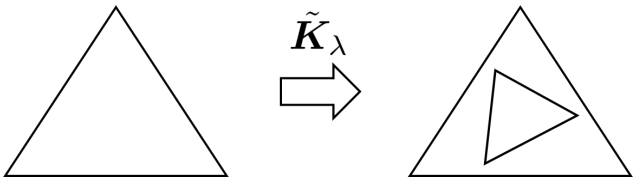

Fig. 1 Shrinkage operator $\tilde{K}_\lambda$.

$\tilde{\boldsymbol{K}}_\lambda$ is a monotonic shrinkage operator mapping S_{n-1} in its interior (see Fig. 1), and $\tilde{\boldsymbol{K}}_\lambda S_{n-1} \subset \tilde{\boldsymbol{K}}_{\lambda'} S_{n-1}$ for $\lambda > \lambda'$. When $\lambda = 0$, $\tilde{\boldsymbol{K}}_\lambda$ is the identity mapping

$$\tilde{\boldsymbol{K}}_0 \boldsymbol{q} = \boldsymbol{q}. \tag{14}$$

As λ tends to infinity, $\tilde{\boldsymbol{K}}_\lambda S_{n-1}$ converges to a single point, the center of S_{n-1}, $\mathbf{1}/n$, where $\mathbf{1} = (1, 1, \ldots, 1)^T$, and

$$\tilde{\boldsymbol{K}}_\infty = \frac{1}{n} [\mathbf{1} \cdots \mathbf{1}]. \tag{15}$$

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

Theorem 2 The minimizer of $C_\lambda(\boldsymbol{p}, \boldsymbol{q})$ with respect to $\boldsymbol{q}$, given $\boldsymbol{p}$, is

$$\boldsymbol{q}^* = \tilde{\boldsymbol{K}}_\lambda \boldsymbol{p}. \tag{16}$$

Proof By differentiation, we have the equation

$$\partial_{\boldsymbol{q}} C_\lambda(\boldsymbol{p}, \boldsymbol{q}^*) = 0 \tag{17}$$

to determine the minimizer $\boldsymbol{q}^*$. This gives the condition

$$b_i = 1, \tag{18}$$

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

$$P^*_{ij} = c a_i K_{ij}, \tag{19}$$

where suffix λ is omitted to alleviate notations. From

$$\sum_j c a_i K_{ij} = p_i, \tag{20}$$

we have

$$c a_i = \frac{p_i}{K_{i\cdot}}. \tag{21}$$

So

$$\boldsymbol{q}^* = \tilde{\boldsymbol{K}}_\lambda \boldsymbol{p}, \tag{22}$$

proving the theorem.

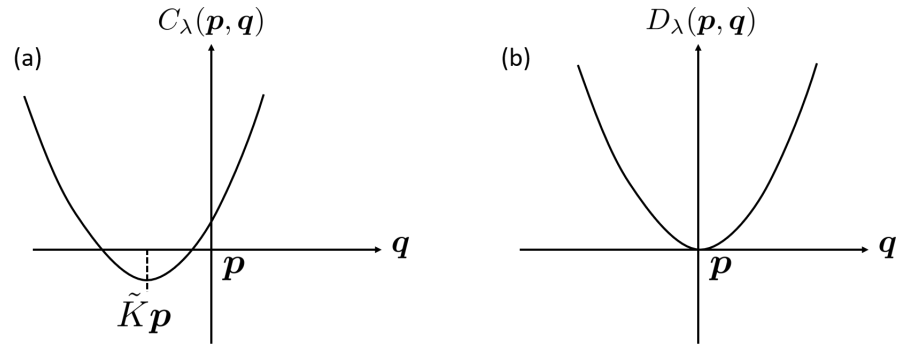

Fig. 2 Comparison of C_λ and D_λ. (a) C_λ as the function of $\boldsymbol{q}$. C_λ is minimized when $\boldsymbol{q} = \tilde{K}\boldsymbol{p}$. (c) D_λ as the function of $\boldsymbol{q}$. D_λ is minimized when $\boldsymbol{q} = \boldsymbol{p}$.

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

Definition 1

$$D_\lambda[\boldsymbol{p} : \boldsymbol{q}] = (1+\lambda)\left(C_\lambda(\boldsymbol{p}, \tilde{\boldsymbol{K}}_\lambda \boldsymbol{q}) - C_\lambda(\boldsymbol{p}, \tilde{\boldsymbol{K}}_\lambda \boldsymbol{p})\right). \tag{23}$$

Figure 2 compares C_λ and D_λ in S_{n-1}. The following theorem is obtained of which proof is given in Appendix I.

Theorem 3 $D_\lambda[\boldsymbol{p} : \boldsymbol{q}]$ is a convex function with respect to $\boldsymbol{p}$ and $\boldsymbol{q}$, satisfying the constraints (10). It converges to the Wasserstein distance as $\lambda \to 0$.

4 Behavior of $\tilde{\boldsymbol{K}}_\lambda$

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

We first consider the case for small λ to see how $\tilde{\boldsymbol{K}}_\lambda$ 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

$$\epsilon = \exp\left(-\frac{1}{\lambda}\right), \quad \lambda = -\frac{1}{\log \epsilon}, \tag{24}$$

we have

$$K_{ij} = \exp\left(-\frac{M_{ij}}{\lambda}\right) = \epsilon^{M_{ij}}, \tag{25}$$

$$\tilde{K}_{i|j} = \frac{\epsilon^{M_{ji}}}{\sum_i \epsilon^{M_{ji}}}. \tag{26}$$

Then, by neglecting higher-order terms of ϵ, we have

$$\tilde{K}_{i|j} = \begin{cases} 1 - |N(i)|\epsilon, & i = j, \\ \epsilon, & i \in N(j), \\ 0, & \text{otherwise}, \end{cases} \tag{27}$$

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, $|N(j)| = 4$, exept for boundary pixels. We see that $\tilde{K}_{i|j}$ is approximated by the discrete Laplacian operator Δ,

$$\tilde{\boldsymbol{K}} = (1 - \epsilon)\boldsymbol{I} + \epsilon\boldsymbol{\Delta}. \tag{28}$$

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

The inverse of $\tilde{\boldsymbol{K}}$ is

$$\tilde{\boldsymbol{K}}^{-1} = (1 + \epsilon)\boldsymbol{I} - \epsilon\boldsymbol{\Delta}. \tag{29}$$

This is the inverse diffusion operator, which sharpens $\boldsymbol{q}$ by emphasizing larger components.

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

$$M(\boldsymbol{\xi}, \boldsymbol{\xi}') = |\boldsymbol{\xi} - \boldsymbol{\xi}'|^2. \tag{30}$$

Then,

$$K_\lambda(\boldsymbol{\xi}, \boldsymbol{\xi}') = \exp\left(-\frac{|\boldsymbol{\xi} - \boldsymbol{\xi}'|^2}{\lambda}\right), \tag{31}$$

and we easily have

$$\tilde{K}_\lambda(\boldsymbol{\xi}|\boldsymbol{\xi}') = \frac{1}{(\sqrt{\pi\lambda})^n} \exp\left(-\frac{|\boldsymbol{\xi} - \boldsymbol{\xi}'|^2}{\lambda}\right). \tag{32}$$

This is a diffusion kernel. When $\boldsymbol{p}$ is a Gaussian distribution

$$p(\boldsymbol{\xi}) = \exp\left(-\frac{1}{2\sigma^2}|\boldsymbol{\xi}|^2\right), \tag{33}$$

we have

$$\tilde{K}_\lambda \boldsymbol{p} = \exp\left(-\frac{1}{2\tau^2}|\boldsymbol{\xi}|^2\right), \tag{34}$$

with

$$\tau^2 = \sigma^2 + \frac{\lambda}{2}. \tag{35}$$

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

We lastly consider the case when λ is large enough, studying the limiting behavior as $\lambda \to \infty$.

When λ is large, we expand $\boldsymbol{K}_\lambda$ as

$$K_{ij,\lambda} = \exp\left(-\frac{M_{ij}}{\lambda}\right) = 1 - \frac{M_{ij}}{\lambda}, \tag{36}$$

$$K_{j\cdot,\lambda} = n\left(1 - \frac{\bar{m}_{j\cdot}}{\lambda}\right), \qquad \bar{m}_{j\cdot} = \frac{1}{n}\sum_i M_{ji}, \tag{37}$$

obtaining

$$\tilde{K}_{i|j,\lambda} = \frac{K_{ji,\lambda}}{K_{j\cdot,\lambda}} = \frac{1}{n}\left(1 - \frac{\tilde{m}_{ij}}{\lambda}\right), \qquad \tilde{m}_{ij} = M_{ji} - \bar{m}_{j\cdot}. \tag{38}$$

Hence,

$$\tilde{\boldsymbol{K}}_\lambda(\boldsymbol{q}-\boldsymbol{p})_i = \frac{1}{n\lambda}\sum_j \tilde{m}_{ij}\left(q_j - p_j\right), \tag{39}$$

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

$$\tilde{M}_{jk} = \frac{1}{n}\sum_i \tilde{m}_{ij}\tilde{m}_{ik}. \tag{40}$$

Then, we have the following theorem, of which proof is given in Appendix II.

Theorem 4 When λ is large enough,

$$\lim_{\lambda\to\infty} D_\lambda[\boldsymbol{p}:\boldsymbol{q}] = \frac{1}{2}(\boldsymbol{q}-\boldsymbol{p})^T\tilde{\boldsymbol{M}}(\boldsymbol{q}-\boldsymbol{p}), \tag{41}$$

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

5 Right Barycenter of Patterns

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

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

$$F^r_D(S,\boldsymbol{q}) = \sum_i D_\lambda[\boldsymbol{p}_i:\boldsymbol{q}]. \tag{42}$$

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

$$F_C(S,\boldsymbol{q}) = \sum C_\lambda(\boldsymbol{p}_i,\boldsymbol{q}). \tag{43}$$

We call such a minimizer the C-barycenter $\boldsymbol{q}^*_C(S)$. Cuturi and Doucet showed that the C-barycenter can extract for some pattern families S a common shape. In particular, Cuturi and Doucet used a family S of deformed double rings at

various sizes and positions, whose barycenter $\boldsymbol{q}_C^*(S)$ turns out to be a standard double ring pattern. Other information-theoretic divergences such as the KL-divergence or Heillinger divergence are not able to recover such a common shape. We will show that the right D-barycenter $\boldsymbol{q}_D^*(S)$ exhibits the same property, but with a sharper solution.

The right D-barycenter minimizes

$$\frac{1}{1+\lambda}\sum D_\lambda(\boldsymbol{p}_i : \boldsymbol{q}) = \sum C_\lambda(\boldsymbol{p}_i, \tilde{\boldsymbol{K}}_\lambda \boldsymbol{q}) - \sum C_\lambda(\boldsymbol{p}_i, \tilde{\boldsymbol{K}}_\lambda \boldsymbol{p}_i). \tag{44}$$

The second term of the right-hand side of (44) does not depend on $\boldsymbol{q}$, so that it may be deleted for minimization. Let us put

$$\tilde{\boldsymbol{q}} = \tilde{\boldsymbol{K}}_\lambda \boldsymbol{q}. \tag{45}$$

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

$$\boldsymbol{q}_C^* = \tilde{\boldsymbol{K}}_\lambda \boldsymbol{q}_D^*, \tag{46}$$

provided (46) is solvable. In this case,

$$\boldsymbol{q}_D^* = \tilde{\boldsymbol{K}}_\lambda^{-1} \boldsymbol{q}_C^* \tag{47}$$

is a sharpened version of $\boldsymbol{q}_C^*$. However, (46) might not be always solvable.

The image $\tilde{\boldsymbol{K}}_\lambda S_{n-1}$ is a simplex sitting inside S_{n-1}. Since the C-barycenter $\boldsymbol{q}_C^*$ is not necessarily inside $\tilde{\boldsymbol{K}}_\lambda S_{n-1}$, we need to solve the D-barycenter problem (44) under the constraint that $\boldsymbol{q}$ is constrained inside $\tilde{\boldsymbol{K}}_\lambda S_{n-1}$. When the C-barycenter $\boldsymbol{q}_C^*$ is inside $\tilde{\boldsymbol{K}}_\lambda S_{n-1}$, the D-barycenter $\boldsymbol{q}_D^*$ is simply given by (46), which is more localized or sharper than $\boldsymbol{q}_C^*$. When $\boldsymbol{q}_C^*$ is not inside $\tilde{\boldsymbol{K}}_\lambda S_{n-1}$, the solution of (44) is on the boundary of the simplex $\tilde{\boldsymbol{K}}_\lambda S_{n-1}$, which implies that some components of $\boldsymbol{q}_D^*$ are forced to be equal to 0.

Theorem 5 The right D-barycenter $\boldsymbol{q}_D^*(S)$ 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 $\boldsymbol{p}_i$ in S, but does not depend on their location, namely that 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 $p(\boldsymbol{\xi}) = p(x, y)$ on the (x, y)-plane, where $\boldsymbol{\xi} = (x, y)$. The center $\boldsymbol{\xi}_p$ of $p(x, y)$ is defined by

$$\boldsymbol{\xi}_p = \int \boldsymbol{\xi} p(\boldsymbol{\xi}) d\boldsymbol{\xi}. \tag{48}$$

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

$$T_{\bar{\boldsymbol{\xi}}} p(\boldsymbol{\xi}) = p(\boldsymbol{\xi} - \bar{\boldsymbol{\xi}}). \tag{49}$$

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

$$T_{\bar{\boldsymbol{\xi}}} P(\boldsymbol{\xi}, \boldsymbol{\xi}') = \bar{P}(\boldsymbol{\xi}, \boldsymbol{\xi}') = P\left(\boldsymbol{\xi}, \boldsymbol{\xi}' - \bar{\boldsymbol{\xi}}\right), \tag{50}$$

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

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

$$m(\boldsymbol{\xi}, \boldsymbol{\xi}') = \|\boldsymbol{\xi} - \boldsymbol{\xi}'\|^2. \tag{51}$$

Then, the direct cost of transportation is

$$\langle M, P \rangle = \int m(\boldsymbol{\xi}, \boldsymbol{\xi}') P(\boldsymbol{\xi}, \boldsymbol{\xi}') \, d\boldsymbol{\xi} d\boldsymbol{\xi}'. \tag{52}$$

The cost of the shifted plan is

$$\begin{aligned}
\langle M, T_{\bar{\boldsymbol{\xi}}} P \rangle &= \int m(\boldsymbol{\xi}, \boldsymbol{\xi}') P\left(\boldsymbol{\xi}, \boldsymbol{\xi}' - \bar{\boldsymbol{\xi}}\right) d\boldsymbol{\xi} d\boldsymbol{\xi}' && (53)\\
&= \int m\left(\boldsymbol{\xi}, \boldsymbol{\xi}'' + \bar{\boldsymbol{\xi}}\right) P(\boldsymbol{\xi}, \boldsymbol{\xi}'') \, d\boldsymbol{\xi} d\boldsymbol{\xi}'' && (54)\\
&= \int \left\{ \|\boldsymbol{\xi} - \boldsymbol{\xi}''\|^2 + \left\|\bar{\boldsymbol{\xi}}\right\|^2 - 2\bar{\boldsymbol{\xi}} \cdot (\boldsymbol{\xi} - \boldsymbol{\xi}'') \right\} P(\boldsymbol{\xi}, \boldsymbol{\xi}'') \, d\boldsymbol{\xi} d\boldsymbol{\xi}'' && (55)\\
&= \langle M, P \rangle + \|\bar{\boldsymbol{\xi}}\|^2 - 2\bar{\boldsymbol{\xi}} \cdot (\boldsymbol{\xi}_p - \boldsymbol{\xi}_q). && (56)
\end{aligned}$$

Note that a shift does not change the entropy

$$H\{P(\boldsymbol{\xi}, \boldsymbol{\xi}')\} = H\left\{T_{\bar{\boldsymbol{\xi}}} P(\boldsymbol{\xi}, \boldsymbol{\xi}')\right\}. \tag{57}$$

When $\boldsymbol{\xi}_p = \boldsymbol{\xi}_q$, two patterns $\boldsymbol{p}$ and $\boldsymbol{q}$ are said to be co-centered. For two co-centered $\boldsymbol{p}, \boldsymbol{q}$, we have

$$\mathcal{L}(P_{\boldsymbol{p},\boldsymbol{q}}) \geq \mathcal{L}\left(P_{\boldsymbol{p},\tilde{K}_\lambda \boldsymbol{p}}\right), \tag{58}$$

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

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

$$\bar{\boldsymbol{\xi}} = \boldsymbol{\xi}_q - \boldsymbol{\xi}_p, \tag{59}$$

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

$$\begin{aligned}
P_{\boldsymbol{p},\boldsymbol{q}} &= T_{-\bar{\boldsymbol{\xi}}} P_{\boldsymbol{p},\bar{\boldsymbol{q}}}, && (60)\\
\boldsymbol{q} &= T_{-\bar{\boldsymbol{\xi}}} \bar{\boldsymbol{q}}. && (61)
\end{aligned}$$

We have

$$C_\lambda\left(T_{\boldsymbol{\xi}}P\right) = C_\lambda(P) + \left\|\bar{\boldsymbol{\xi}}\right\|^2 - 2\bar{\boldsymbol{\xi}}\cdot\left(\boldsymbol{\xi}_p - \boldsymbol{\xi}_q\right). \tag{62}$$

We have an important relation that $C_\lambda\left(P_{\boldsymbol{p},\boldsymbol{q}}\right)$ is decomposed into a sum of the shape deformation cost among co-centered patterns and the positional transportation cost, as

$$C_\lambda\left(P_{\boldsymbol{p},\boldsymbol{q}}\right) = C_\lambda\left(P_{\boldsymbol{p},\bar{\boldsymbol{q}}}\right) + \left\|\boldsymbol{\xi}_p - \boldsymbol{\xi}_q\right\|^2 \tag{63}$$

because of

$$\boldsymbol{\xi}_p = \boldsymbol{\xi}_{\bar{q}}. \tag{64}$$

Lemma 1 Given $\boldsymbol{p}$, the C_λ-optimal pattern $q(\boldsymbol{\xi})$ transporting $\boldsymbol{p}$ to $\boldsymbol{q}$ is a shifted and blurred version of $p(\boldsymbol{\xi})$,

$$q^*(\boldsymbol{\xi}) = T_{\boldsymbol{\xi}_p - \boldsymbol{\xi}_q}\tilde{K}_\lambda p(\boldsymbol{\xi}), \tag{65}$$

not depending on the locations $\boldsymbol{\xi}_p$ and $\boldsymbol{\xi}_q$.

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

$$F_C(\boldsymbol{q}) = \sum_{i=1}^{n} C_\lambda\left(P_{\boldsymbol{p}_i,\boldsymbol{q}}\right), \quad F_D(\boldsymbol{q}) = \sum_{i=1}^{n} D_\lambda\left(P_{\boldsymbol{p}_i,\boldsymbol{q}}\right), \tag{66}$$

The center of $p_i(\boldsymbol{\xi})$ is denoted by $\boldsymbol{\xi}_i = \boldsymbol{\xi}_{p_i}$. Before solving the problem, let us respectively shift $p_i(\boldsymbol{\xi})$ from $\boldsymbol{\xi}_i$ to $\boldsymbol{\xi}_0$, a fixed common location, such that all shifted $\bar{p}_i(\boldsymbol{\xi})$'s are co-centered. Let $q_C^*(\boldsymbol{\xi})$ and $q_D^*(\boldsymbol{\xi})$ be the barycenters of all co-centered $\bar{p}_i(\boldsymbol{\xi})$, which do not depend on the locations of $p_i(\boldsymbol{\xi})$ but their shapes.

Theorem 6 (Shape-Location Separation Theorem) The barycenters $q_C^*(\boldsymbol{\xi})$ and $q_D^*(\boldsymbol{\xi})$ of $p_1(\boldsymbol{\xi}), \cdots, p_n(\boldsymbol{\xi})$ are located at the barycenter of $\boldsymbol{\xi}_1, \cdots, \boldsymbol{\xi}_n$ and their shapes are given by the respective barycenters of $\bar{p}_1(\boldsymbol{\xi}), \cdots, \bar{p}_n(\boldsymbol{\xi})$.

Proof From (63), we have

$$F_C(\boldsymbol{q}) = \sum C_\lambda\left(P_{\boldsymbol{p}_i,\bar{\boldsymbol{q}}_i}\right) + \sum \left\|\boldsymbol{\xi}_{p_i} - \boldsymbol{\xi}_q\right\|^2, \tag{67}$$

$$F_D(\boldsymbol{q}) = \sum C_\lambda\left(P_{\boldsymbol{p}_i,\tilde{K}_\lambda\bar{\boldsymbol{q}}}\right) + \sum \left\|\boldsymbol{\xi}_{p_i} - \boldsymbol{\xi}_q\right\|^2, \tag{68}$$

where $\bar{\boldsymbol{q}}_i$ is the shifted version of $\boldsymbol{q}$ to the center of $\boldsymbol{p}_i$. Here, the shape and location of $\boldsymbol{q}$ is separated. Minimizing the first term, we have $\boldsymbol{q}^*$ which is the respective barycenters of the shapes of co-centered $\boldsymbol{p}_1, \cdots, \boldsymbol{p}_n$. The second term gives the barycenter of locations $\boldsymbol{\xi}_1, \cdots, \boldsymbol{\xi}_n$.

Corollary 1 When $\boldsymbol{p}_i$ are shifts of an identical $\boldsymbol{p}$, their right D-barycenter has the same shape as the original $\boldsymbol{p}$, whereas the C-barycenter is a blurred version of $\boldsymbol{p}$,

$$\boldsymbol{q}_C^* = \boldsymbol{K}_\lambda\boldsymbol{p}. \tag{69}$$

We show a simple example where $\boldsymbol{p}_i$ are shifted $\boldsymbol{p}$, a cat shape (Fig. 3a). Its C-barycenter has a blurred shape $\boldsymbol{K}_\lambda\boldsymbol{p}$ (Fig. 3b) tending to the uniform distribution as $\lambda \to \infty$. However, the shape of the right D-barycenter is exactly the same as $\boldsymbol{p}$ (Fig. 3c).

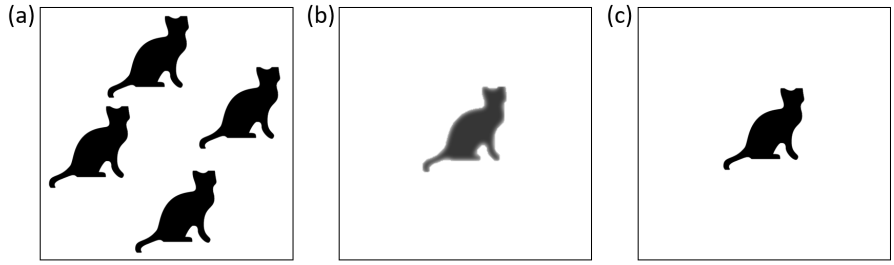

Fig. 3 (a) Cat images. (b) The C-barycenter of panel (a). The C-barycenter has a blurred shape $\boldsymbol{K}_\lambda \boldsymbol{p}$ tending to the uniform distribution as $\lambda \to \infty$. (c) The D-barycenter of panel (a). The shape of the right D-barycenter is exactly the same as the original shape $\boldsymbol{p}$.

6 Left Barycenter of Patterns

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

We consider again a family S of patterns, $S = (\boldsymbol{q}_i)_{i=1,\dots,N}$. The barycenter $\boldsymbol{p}$ of these patterns based on divergence D_λ is defined by the minimizer of

$$F_D^l(S, \boldsymbol{p}) = \sum_i D_\lambda[\boldsymbol{p} : \boldsymbol{q}_i], \tag{70}$$

and is called the left D-varycenter. We propose to solve that problem using the accelerated gradient descent approach outlined in (Cuturi and Doucet, 2014), with two differences: all examples $\boldsymbol{q}_i$ must be smoothed beforehand following an application of $\tilde{\boldsymbol{K}}_\lambda$, and the gradient incorporates now not only terms resulting from $C_\lambda(\boldsymbol{p}, \tilde{\boldsymbol{K}}_\lambda \boldsymbol{q}_i)$, but also from $-C_\lambda(\boldsymbol{p}, \tilde{\boldsymbol{K}}_\lambda \boldsymbol{p})$ which, as explained in Equation (73), is simply minus the Kullback-Leibler divergence between $\boldsymbol{p}$ and the vector $\tilde{\boldsymbol{K}}_\lambda \mathbf{1}$, that is the entropy of $\boldsymbol{p}$ plus the dot product between $\boldsymbol{p}$ and $\log(\tilde{\boldsymbol{K}}_\lambda \mathbf{1})$. As a result the gradient of $-C_\lambda(\boldsymbol{p}, \tilde{\boldsymbol{K}}_\lambda \boldsymbol{p})$ is equal to, up to a constant term, $-\log(\boldsymbol{p}) + \log(\tilde{\boldsymbol{K}}_\lambda \mathbf{1})$, which tends to sharpen further any iterate compared to the simple minimization of the C_λ 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.

It might be easier to calculate numerically, but it does not have the shape-location separation property. An example of the left D-barycenter of four patterns is shown in Fig. 4.

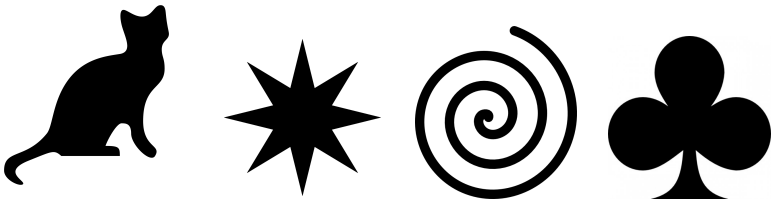

Fig. 4 Four shapes considered in our experiments to compute barycenters according to C_λ or D_λ

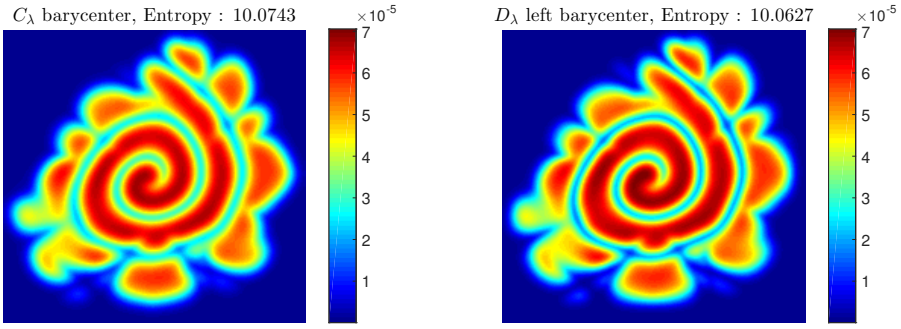

Fig. 5 We consider in this example the squared-Euclidean distance on grids as the ground metric cost. We study the iso-barycenter of the four shapes presented in Figure 4 using two different discrepancy functions, C_λ and D_λ. *(left)* barycenter obtained using Algorithm 2 of Solomon et al. (2015), itself a specialized version of Benamou et al.'s algorithm (2014). We used a 10^-10 tolerance for the l_1 norm between two successive iterates as a stopping criterion). *(right)* D_λ left-barycenter obtained with our accelerated gradient approach. We use a `jet` colormap to highlight differences in the support of the barycenters. As expected, the entropy of the C_λ barycenter is higher than that of the D_λ left barycenter, since the latter optimization incorporates a penalization term, $-C_\lambda(\boldsymbol{p}, \tilde{\boldsymbol{K}}_\lambda \boldsymbol{p})$, which is equivalent to penalizing the entropy of the solution $\boldsymbol{p}$. This difference in entropy results in subtle yet visible differences between the two solutions, with sharper edges for the D_λ left barycenter.

Conclusions

We defined a new divergence 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 criteria of a distance or divergence. We defined a new divergence function D_λ derived from C_λ, which works better than the original C_λ for some problems, in particular, the barycenter problem. We proved that the minimizer of $C_\lambda(\boldsymbol{p}, \boldsymbol{q})$ is given by $\tilde{\boldsymbol{K}}_\lambda \boldsymbol{p}$, where $\tilde{\boldsymbol{K}}_\lambda$ is a diffusion operator depending on the base metric $\boldsymbol{M}$. We studied properties of $\tilde{\boldsymbol{K}}_\lambda$ showing how it changes as λ increases, elucidating properties of D_λ.

We applied D_λ to obtain the barycenter of a cluster of image patterns. It is proved that the right D-barycenter keeps a good property that the shape and locations of patterns are separated, which is a merit of the C-function based

barycenter. Moreover, the D-barycenter gives even a sharper shape than the C-barycenter.

We cannot touch upon computational aspects of the D-barycenter, because this is a theoretical paper proposing a new divergence and its properties. We also defined the left D-barycenter, because D_λ is an asymmetric divergence. This is computationally easy to calculate. However, it remains as our future study to explore its properties.

Appendix I: Proof of convexity of D_λ

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

$$\boldsymbol{G}_\lambda = \frac{\partial^2 C_\lambda(\boldsymbol{p}, \tilde{\boldsymbol{K}}_\lambda \boldsymbol{q})}{\partial \boldsymbol{\eta} \partial \boldsymbol{\eta}^T} = \begin{bmatrix} \boldsymbol{X} & \boldsymbol{Y} \\ \boldsymbol{Y}^T & \boldsymbol{Z} \end{bmatrix}, \tag{71}$$

expressed by block matrices $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}$ where $\boldsymbol{\eta} = (\boldsymbol{p}, \boldsymbol{q})^T$. Because C_λ is strictly convex (Amari et al., 2018), $\boldsymbol{G}_\lambda$ is positive definite and the block component $\boldsymbol{Z}$ is also regular and positive definite.

By using

$$P^*_{ij} = p_i \frac{K_{ij}}{K_{i\cdot}}, \tag{72}$$

we get

$$C_\lambda(\boldsymbol{p}, \tilde{\boldsymbol{K}}_\lambda \boldsymbol{p}) = \lambda \sum_{i=1}^{n} p_i \ln \frac{p_i}{K_{i\cdot}}. \tag{73}$$

Therefore, its Hessian becomes

$$\boldsymbol{G}'_\lambda = \frac{\partial^2 C_\lambda(\boldsymbol{p}, \tilde{\boldsymbol{K}}_\lambda \boldsymbol{p})}{\partial \boldsymbol{\eta} \partial \boldsymbol{\eta}^T} = \lambda \begin{bmatrix} \mathrm{diag}\left(\frac{1}{p_i}\right) + \frac{1}{p_n} \mathbf{1}\mathbf{1}^T & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{O} \end{bmatrix}, \tag{74}$$

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

$$\det(\boldsymbol{R}) = \det(\boldsymbol{Z}) \det(\boldsymbol{R}'), \tag{75}$$

where we put

$$\boldsymbol{R}' = \boldsymbol{X} - \boldsymbol{Y}\boldsymbol{Z}^{-1}\boldsymbol{Y}^T - \lambda \left(\mathrm{diag}\left(\frac{1}{p_i}\right) + \frac{1}{p_n} \mathbf{1}\mathbf{1}^T \right). \tag{76}$$

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

$$\mathbf{G}_\lambda^{-1} = \frac{1}{1+\lambda} \left[\begin{array}{c|c} p_i \delta_{ij} - p_i p_j & P_{ij} - p_i q_j \\ \hline P_{ji} - q_i p_j & q_i \delta_{ij} - q_i q_j \end{array} \right], \tag{77}$$

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

$$(\boldsymbol{X}-\boldsymbol{Y}\boldsymbol{Z}^{-1}\boldsymbol{Y}^T)^{-1}=\frac{1}{1+\lambda}\left(\mathrm{diag}(p_i)-\boldsymbol{p}\boldsymbol{p}^T\right). \tag{78}$$

By using the Sherman-Morrison formula, we get

$$\boldsymbol{X}-\boldsymbol{Y}\boldsymbol{Z}^{-1}\boldsymbol{Y}^T=(1+\lambda)\left(\mathrm{diag}\left(\frac{1}{p_i}\right)+\frac{1}{p_n}\mathbf{1}\mathbf{1}^T\right). \tag{79}$$

Finally, we obtain

$$\boldsymbol{R}'=\mathrm{diag}\left(\frac{1}{p_i}\right)+\frac{1}{p_n}\mathbf{1}\mathbf{1}^T. \tag{80}$$

Because this $\boldsymbol{R}'$ is positive definite, $\boldsymbol{R}$ is also positive definite and D_λ is strictly convex.

By using Eq.(14) and $C_0(\boldsymbol{p},\boldsymbol{p})=0$, we can confirm that D_λ converges to the Wasserstein distance as $\lambda\to 0$.

Appendix II: Proof of D_λ approaching an energy function for large λ

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

$$\begin{aligned}\frac{1}{(1+\lambda)}D_\lambda[\boldsymbol{p},\boldsymbol{q}] &= C_\lambda\left(\boldsymbol{p},\tilde{\boldsymbol{K}}_\lambda\boldsymbol{p}+\tilde{\boldsymbol{K}}(\boldsymbol{q}-\boldsymbol{p})\right)-C_\lambda\left(\boldsymbol{p},\tilde{\boldsymbol{K}}_\lambda\boldsymbol{p}\right) &(81)\\ &= \partial_{\boldsymbol{q}}C_\lambda\left(\boldsymbol{p},\tilde{\boldsymbol{K}}_\lambda\boldsymbol{p}\right)\cdot\tilde{\boldsymbol{K}}_\lambda(\boldsymbol{q}-\boldsymbol{p})\\ &\quad+\frac{1}{2}\partial_{\boldsymbol{q}}\partial_{\boldsymbol{q}}C\left(\boldsymbol{p},\tilde{\boldsymbol{K}}_\lambda\boldsymbol{p}\right)\cdot\tilde{\boldsymbol{K}}_\lambda(\boldsymbol{q}-\boldsymbol{p})\otimes\tilde{\boldsymbol{K}}_\lambda(\boldsymbol{q}-\boldsymbol{p}) &(82)\\ &= \frac{1}{2}\partial_{\boldsymbol{q}}\partial_{\boldsymbol{q}}C\left(\boldsymbol{p},\tilde{\boldsymbol{K}}_\lambda\boldsymbol{p}\right)\cdot\tilde{\boldsymbol{K}}_\lambda(\boldsymbol{q}-\boldsymbol{p})\otimes\tilde{\boldsymbol{K}}_\lambda(\boldsymbol{q}-\boldsymbol{p}), &(83)\end{aligned}$$

because of

$$\partial_{\boldsymbol{q}}C_\lambda\left(\boldsymbol{p},\tilde{\boldsymbol{K}}_\lambda\boldsymbol{p}\right)=0, \tag{84}$$

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

When λ is large, we expand $\boldsymbol{K}_\lambda$ as

$$\begin{aligned}K_{ij,\lambda} &= \exp\left\{-\frac{M_{ij}}{\lambda}\right\}=1-\frac{M_{ij}}{\lambda}, &(85)\\ K_{j\cdot,\lambda} &= n\left(1-\frac{\bar{m}_{\cdot j}}{\lambda}\right), \qquad \bar{m}_{j\cdot}=\frac{1}{n}\sum_i M_{ji}, &(86)\end{aligned}$$

obtaining

$$\tilde{K}_{i|j,\lambda}=\frac{K_{ji}}{K_{j\cdot}}=\frac{1}{n}\left(1-\frac{\tilde{m}_{ij}}{\lambda}\right), \qquad \tilde{m}_{ij}=M_{ji}-\bar{m}_{j\cdot}. \tag{87}$$

Hence

$$\tilde{\boldsymbol{K}}_\lambda(\boldsymbol{q}-\boldsymbol{p})_i = \frac{1}{n\lambda}\sum_j \tilde{m}_{ij}\left(q_j - p_j\right). \tag{88}$$

We have

$$\partial_{\boldsymbol{q}}\partial_{\boldsymbol{q}} C\left(\boldsymbol{p}, \tilde{\boldsymbol{K}}_\lambda \boldsymbol{p}\right) = \partial_{\boldsymbol{q}}\partial_{\boldsymbol{q}} C_\lambda\left(\boldsymbol{p}, \boldsymbol{1}\right) + O\left(\frac{1}{\lambda}\right). \tag{89}$$

So we calculate $\partial_{\boldsymbol{q}}\partial_{\boldsymbol{q}} C_\lambda(\boldsymbol{p}, \tilde{\boldsymbol{q}})$ when

$$\tilde{q}_i = \left(\tilde{\boldsymbol{K}}_\lambda \boldsymbol{p}\right)_i = \frac{1}{n} + \varepsilon_i \tag{90}$$

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

The optimal transportation plan for $\boldsymbol{p} \to \boldsymbol{q}$ is

$$p^*_{ij} = c a_i b_j \tag{91}$$

when $\lambda \to \infty$, because $K_{ij,\lambda} = 1$. Hence,

$$P^*_{ij} = p_i q_j. \tag{92}$$

We have already obtained the inverse of $\partial_{\boldsymbol{\eta}}\partial_{\boldsymbol{\eta}} C_\lambda(\boldsymbol{p}, \boldsymbol{q})$ in (45) of the previous paper

$$\boldsymbol{G}^{-1}_\lambda = \frac{1}{1+\lambda}\left[\begin{array}{c|c} p_i\delta_{ij} - p_i p_j & P^*_{ij} - p_i q_j \\ \hline P^*_{ji} - q_i p_j & q_i\delta_{ij} - q_i q_j \end{array}\right]. \tag{93}$$

Hence, it is block-diagonal $\left(P^*_{ij} - p_i q_j = 0\right)$, and the $(\boldsymbol{q}, \boldsymbol{q})$-part of $\boldsymbol{G}_\lambda$ $(\lambda \to \infty)$ is

$$\boldsymbol{G}_{\lambda,\boldsymbol{qq}} = (1+\lambda)\left[q_i\delta_{ij} - q_i q_j\right]^{-1} \tag{94}$$

$$= (1+\lambda)\left(\mathrm{diag}\left(\frac{1}{q_i}\right) + \frac{1}{q_n}\boldsymbol{1}\boldsymbol{1}^T\right). \tag{95}$$

In our case of $\boldsymbol{q} = \boldsymbol{1}/n$,

$$\boldsymbol{G} = \partial_{\boldsymbol{q}}\partial_{\boldsymbol{q}} C = (1+\lambda)(n\delta_{ij} + n). \tag{96}$$

We finally calculate

$$\begin{aligned}
&\boldsymbol{G}\cdot\tilde{\boldsymbol{K}}_\lambda(\boldsymbol{q}-\boldsymbol{p})\otimes\tilde{\boldsymbol{K}}_\lambda(\boldsymbol{q}-\boldsymbol{p}) \\
&= \frac{(1+\lambda)n}{n^2\lambda^2}\sum \tilde{m}_{ij}\tilde{m}_{ik}\left(q_j - p_j\right)\left(q_k - p_k\right) + \frac{(1+\lambda)n}{n^2\lambda^2}\{\sum_{ij}\tilde{m}_{ij}\left(q_j - p_j\right)\}^2 \\
&= \frac{1+\lambda}{\lambda^2}(\boldsymbol{q}-\boldsymbol{p})^T\tilde{\boldsymbol{M}}(\boldsymbol{q}-\boldsymbol{p}),
\end{aligned}$$

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

Bibliography

M Agueh and G Carlier. Barycenters in the wasserstein space. SIAM J. on Mathematical Analysis, 43:904–924, 2011.

Shun-ichi Amari. Information Geometry and Its pplications. Springer, 2016.

Shun-ichi Amari, Ryo Karakida, and Masafumi Oizumi. Information geometry connecting wasserstein distance and kullback–leibler divergence via the entropy-relaxed transportation problem. Information Geometry, 2018. doi: 10.1007/s41884-018-0002-8.

J.-D. Benamou, G. Carlier, M. Cuturi, L. Nenna, and G. Peyré. Iterative Bregman projections for regularized transportation problems. arXiv preprint arXiv:1412.5154, 2014.

N. Courty, R. Flamary, D. Tuia, and A. Rakotomamonjy. Optimal transport for domain adaptation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(9):1853–1865, Sept 2017. ISSN 0162-8828. doi: 10.1109/TPAMI.2016.2615921.

M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In Advances in Neural Information Processing Systems, pages 2292–2300, 2013.

M. Cuturi and D. Avis. Ground metric learning. The Journal of Machine Learning Research, 2014.

M. Cuturi and A. Doucet. Fast computation of Wasserstein barycenters. In Proceedings of the 31st International Conference on Machine Learning (ICML-14), pages 685–693, 2014.

Marco Cuturi and Gabriel Peyré. A smoothed dual approach for variational wasserstein problems. SIAM Journal on Imaging Sciences, 9(1):320–343, 2016.

Charlie Frogner, Chiyuan Zhang, Hossein Mobahi, Mauricio Araya, and Tomaso A Poggio. Learning with a wasserstein loss. In Advances in Neural Information Processing Systems, pages 2053–2061, 2015.

Aude Genevay, Gabriel Peyre, and Marco Cuturi. Learning generative models with sinkhorn divergences. In Amos Storkey and Fernando Perez-Cruz, editors, Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics, volume 84 of Proceedings of Machine Learning Research, pages 1608–1617, Playa Blanca, Lanzarote, Canary Islands, 09–11 Apr 2018. PMLR.

Antoine Rolet, Marco Cuturi, and Gabriel Peyr. Fast dictionary learning with a smoothed wasserstein loss. In Arthur Gretton and Christian C. Robert, editors, Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, volume 51 of Proceedings of Machine Learning Research, pages 630–638, Cadiz, Spain, 09–11 May 2016. PMLR.

Richard Sinkhorn. A relationship between arbitrary positive matrices and doubly stochastic matrices. The annals of mathematical statistics, 35(2): 876–879, 1964.

J. Solomon, F. de Goes, G. Peyré, M. Cuturi, A. Butscher, A. Nguyen, T. Du, and L. Guibas. Convolutional Wasserstein distances: Efficient optimal trans-

portation on geometric domains. ACM Transactions on Graphics (Proc. SIGGRAPH 2015), 2015.

Cédric Villani. Topics in Optimal Transportation, volume 58. AMS Graduate Studies in Mathematics, 2003.