Info Geo (2018) 1:13–37
https://doi.org/10.1007/s41884-018-0002-8

RESEARCH PAPER

Information geometry connecting Wasserstein distance and Kullback–Leibler divergence via the entropy-relaxed transportation problem

Shun-ichi Amari[1,3] · Ryo Karakida[2] · Masafumi Oizumi[1,3]

Received: 5 October 2017 / Revised: 9 February 2018 / Published online: 23 March 2018

Abstract Two geometrical structures have been extensively studied for a manifold of probability distributions. One is based on the Fisher information metric, which is invariant under reversible transformations of random variables, while the other is based on the Wasserstein distance of optimal transportation, which reflects the structure of the distance between underlying random variables. Here, we propose a new information-geometrical theory that provides a unified framework connecting the Wasserstein distance and Kullback–Leibler (KL) divergence. We primarily considered a discrete case consisting of n elements and studied the geometry of the probability simplex S_{n-1}, which is the set of all probability distributions over n elements. The Wasserstein distance was introduced in S_{n-1} by the optimal transportation of commodities from distribution $\boldsymbol{p}$ to distribution $\boldsymbol{q}$, where $\boldsymbol{p}, \boldsymbol{q} \in S_{n-1}$. We relaxed the optimal transportation by using entropy, which was introduced by Cuturi. The optimal solution was called the entropy-relaxed stochastic transportation plan. The entropy-relaxed optimal cost $C(\boldsymbol{p}, \boldsymbol{q})$ was computationally much less demanding than the original Wasserstein distance but does not define a distance because it is not minimized at $\boldsymbol{p} = \boldsymbol{q}$. To define a proper divergence while retaining the computational advantage, we first introduced a divergence function in the manifold $S_{n-1} \times S_{n-1}$ composed of all optimal transportation plans. We fully explored the information geometry of the manifold of the optimal transportation plans and subsequently constructed a new one-parameter family of divergences in S_{n-1} that are related to both the Wasserstein distance and the KL-divergence.

✉ Shun-ichi Amari
amari@brain.riken.jp

[1] RIKEN Brain Science Institute, 2–1 Hirosawa, Wako-shi, Saitama 351-0198, Japan

[2] National Institute of Advanced Industrial Science and Technology (AIST), 2-3-26 Aomi, Koto-ku, Tokyo 135-0064, Japan

[3] Araya Inc, 2-8-10 Toranomon, Minato-ku, Tokyo 105-0001, Japan

Keywords Wasserstein distance · Kullback–Leibler divergence · Optimal transportation · Information geometry

1 Introduction

Information geometry [1] studies the properties of a manifold of probability distributions and is useful for various applications in statistics, machine learning, signal processing, and optimization. Two geometrical structures have been introduced from two distinct backgrounds. One is based on the invariance principle, where the geometry is invariant under reversible transformations of random variables. The Fisher information matrix, for example, is a unique invariant Riemannian metric from the invariance principle [1–3]. Moreover, two dually coupled affine connections are used as invariant connections [1,4], which are useful in various applications.

The other geometrical structure was introduced through the transportation problem, where one distribution of commodities is transported to another distribution. The minimum transportation cost defines a distance between the two distributions, which is called the Wasserstein, Kantorovich or earth-mover distance [5,6]. This structure provides a tool to study the geometry of distributions by taking the metric of the supporting manifold into account.

Let $\chi = \{1, \ldots, n\}$ be the support of a probability measure $\boldsymbol{p}$. The invariant geometry provides a structure that is invariant under permutations of elements of χ and results in an efficient estimator in statistical estimation. On the other hand, when we consider a picture over n^2 pixels $\chi = \{(ij); i, j = 1, \ldots, n\}$ and regard it as a distribution over χ, the pixels have a proper distance structure in χ. Spatially close pixels tend to take similar values. A permutation of χ destroys such a neighboring structure, suggesting that the invariance might not play a useful role. The Wasserstein distance takes such a structure into account and is therefore useful for problems with metric structure in support χ (see, e.g., [7–9]).

An interesting question is how these two geometrical structures are related. While both are important in their own respects, it would be intriguing to construct a unified framework that connects the two. With this purpose in mind, we examined the discrete case over n elements, such that a probability distribution is given by a probability vector $\boldsymbol{p} = (p, \ldots, p_n)$ in the probability simplex

$$S_{n-1} = \left\{ \boldsymbol{p} \;\middle|\; p_i > 0, \; \sum p_i = 1 \right\}. \tag{1}$$

It is easy to naively extend our theory to distributions over $\boldsymbol{R}^n$, ignoring mathematical difficulties of geometry of function spaces. See Ay et al. [4] for details. We consider Gaussian distributions over the one-dimensional real line X as an example of the continuous case.

Cuturi modified the transportation problem such that the cost is minimized under an entropy constraint [7]. This is called the entropy-relaxed optimal transportation problem and is computationally less demanding than the original transportation problem. In addition to the advantage in computational cost, Cuturi showed that the quasi-distance defined by the entropy-relaxed optimal solution yields superior results in many applications compared to the original Wasserstein distance and information-geometric divergences such as the KL divergence.

We followed the entropy-relaxed framework that Cuturi et al. proposed [7–9] and introduced a Lagrangian function, which is a linear combination of the transportation cost and entropy. Given a distribution $\boldsymbol{p}$ of commodity on the senders side and $\boldsymbol{q}$ on the receivers side, the constrained optimal transportation plan is the minimizer of the Lagrangian function. The minimum value $C(\boldsymbol{p}, \boldsymbol{q})$ is a function of $\boldsymbol{p}$ and $\boldsymbol{q}$, which we called the Cuturi function. However, this does not define the distance between $\boldsymbol{p}$ and $\boldsymbol{q}$ because it is non-zero at $\boldsymbol{p} = \boldsymbol{q}$ and is not minimized when $\boldsymbol{p} = \boldsymbol{q}$.

To define a proper distance-like function in S_{n-1}, we introduced a divergence between $\boldsymbol{p}$ and $\boldsymbol{q}$ derived from the optimal transportation plan. A divergence is a general metric concept that includes the square of a distance but is more flexible, allowing non-symmetricity between $\boldsymbol{p}$ and $\boldsymbol{q}$. A manifold equipped with a divergence yields a Riemannian metric with a pair of dual affine connections. Dually coupled geodesics are defined, which possess remarkable properties, generalizing the Riemannian geometry [1].

We studied the geometry of the entropy-relaxed optimal transportation plans within the framework of information geometry. They form an exponential family of probability distributions defined in the product manifold $S_{n-1} \times S_{n-1}$. Therefore, a dually flat structure was introduced. The m-flat coordinates are the expectation parameters $(\boldsymbol{p}, \boldsymbol{q})$ and their dual, e-flat coordinates (canonical parameters) are $(\boldsymbol{\alpha}, \boldsymbol{\beta})$, which are assigned from the minimax duality of nonlinear optimization problems. We can naturally defined a canonical divergence, that is the KL divergence $KL[(\boldsymbol{p}, \boldsymbol{q}) : (\boldsymbol{p}', \boldsymbol{q}')]$ between the two optimal transportation plans for $(\boldsymbol{p}, \boldsymbol{q})$ and $(\boldsymbol{p}', \boldsymbol{q}')$, sending $\boldsymbol{p}$ to $\boldsymbol{q}$ and $\boldsymbol{p}'$ to $\boldsymbol{q}'$, respectively.

To define a divergence from $\boldsymbol{p}$ to $\boldsymbol{q}$ in S_{n-1}, we used a reference distribution $\boldsymbol{r}$. Given $\boldsymbol{r}$, we defined a divergence between $\boldsymbol{p}$ and $\boldsymbol{q}$ by $KL[(\boldsymbol{r}, \boldsymbol{p}) : (\boldsymbol{r}, \boldsymbol{q})]$. There are a number of potential choices for $\boldsymbol{r}$: one is to use $\boldsymbol{r} = \boldsymbol{p}$ and another is to use the arithmetic or geometric mean of $\boldsymbol{p}$ and $\boldsymbol{q}$. These options yield one-parameter families of divergences connecting the Wasserstein distance and KL-divergence. Our work uncovers a novel direction for studying the geometry of a manifold of probability distributions by integrating the Wasserstein distance and KL divergence.

2 Entropy-constrained transportation problem

Let us consider n terminals $\chi = (X_1, \ldots, X_n)$, some of which, say $X_1, \ldots, X_s$, are sending terminals at which $p_1, \ldots, p_s$ $(p_i > 0)$ amounts of commodities are stocked. At the other terminals, $X_{s+1}, \ldots, X_n$, no commodities are stocked $(p_i = 0)$. These are transported within χ such that $q_1, \ldots, q_r$ amounts are newly stored at the receiving terminals $X_{j_1}, \ldots, X_{j_r}$. There may be overlap in the sending and receiving terminals, $\chi_S = \{X_1, \ldots, X_s\}$ and $\chi_R = \left\{X_{j_1}, \ldots, X_{j_r}\right\}$, including the case that $\chi_R = \chi_S = \chi$ (Fig. 1). We normalized the total amount of commodities to be equal to 1 so that $\boldsymbol{p} = (p_1, \ldots, p_s)$ and $\boldsymbol{q} = (q_1, \ldots, q_r)$ can be regarded as probability distributions in the probability simplex S_{s-1} and S_{r-1}, respectively,

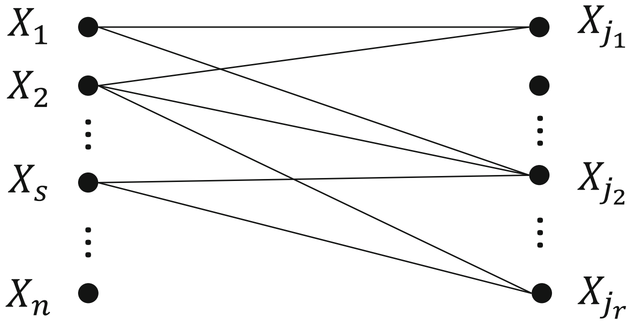

Fig. 1 Transportation from the sending terminals χ_S to the receiving terminals χ_R

$$\sum p_i = 1, \quad \sum q_i = 1, \quad p_i > 0, \quad q_i > 0. \tag{2}$$

Let S_{n-1} be the probability simplex over χ. Then $S_{s-1} \subset \bar{S}_{n-1}$, $S_{r-1} \subset \bar{S}_{n-1}$, where $\bar{S}_{n-1}$ is the closure of S_{n-1},

$$\bar{S}_{n-1} = \left\{ \boldsymbol{r} \;\middle|\; r_i \geq 0, \; \sum r_i = 1 \right\}. \tag{3}$$

It should be noted that if some components of $\boldsymbol{p}$ and $\boldsymbol{q}$ are allowed to be 0, we do not need to treat χ_S and χ_R separately, i.e., we can consider both χ_S and χ_R to be equal to χ. Under such a situation, we simply considered both $\boldsymbol{p}$ and $\boldsymbol{q}$ as elements of $\bar{S}_{n-1}$.

We considered a transportation plan $\mathbf{P} = \left(P_{ij}\right)$ denoted by an $s \times r$ matrix, where $P_{ij} \geq 0$ is the amount of commodity transported from $X_i \in \chi_S$ to $X_j \in \chi_R$. The plan $\mathbf{P}$ was regarded as a (probability) distribution of commodities flowing from X_i to X_j, satisfying the sender and receivers conditions,

$$\sum_j P_{ij} = p_i, \quad \sum_i P_{ij} = q_j, \quad \sum_{ij} P_{ij} = 1. \tag{4}$$

We denoted the set of $\mathbf{P}$ satisfying Eq. (4) as $U(\boldsymbol{p}, \boldsymbol{q})$.

Let $\mathbf{M} = \left(m_{ij}\right)$ be the cost matrix, where $m_{ij} \geq 0$ denotes the cost of transporting one unit of commodity from X_i to X_j. We can interpret m_{ij} as a generalized distance between X_i and X_j. The transportation cost of plan $\mathbf{P}$ is

$$C(\mathbf{P}) = \langle \mathbf{M}, \mathbf{P} \rangle = \sum m_{ij} P_{ij}. \tag{5}$$

The Wasserstein distance between $\boldsymbol{p}$ and $\boldsymbol{q}$ is the minimum cost of transporting commodities distributed by $\boldsymbol{p}$ at the senders to $\boldsymbol{q}$ at the receivers side,

$$C_W(\boldsymbol{p}, \boldsymbol{q}) = \min_{\mathbf{P} \subset U(\boldsymbol{p}, \boldsymbol{q})} \langle \mathbf{M}, \mathbf{P} \rangle, \tag{6}$$

where min is taken over all $\mathbf{P}$ satisfying the constraints in Eq. (4) [5,6].

We considered the joint entropy of $\mathbf{P}$,

$$H(\mathbf{P}) = -\sum P_{ij} \log P_{ij}. \tag{7}$$

Given marginal distributions $\boldsymbol{p}$ and $\boldsymbol{q}$, the plan that maximizes the entropy is given by the direct product of $\boldsymbol{p}$ and $\boldsymbol{q}$,

$$\mathbf{P}_D = \boldsymbol{p} \otimes \boldsymbol{q} = \left(p_i q_j\right). \tag{8}$$

This is because the entropy of $\mathbf{P}_D$,

$$H\left(\mathbf{P}_D\right) = -\sum \mathrm{P}_{Dij} \log \mathrm{P}_{Dij} = H(\boldsymbol{p}) + H(\boldsymbol{q}), \tag{9}$$

is the maximum among all possible $\mathbf{P}$ belonging to $U(\boldsymbol{p}, \boldsymbol{q})$, i.e.,

$$H(\mathbf{P}) \leq H(\boldsymbol{p}) + H(\boldsymbol{q}) = H(\mathbf{P}_D), \tag{10}$$

where $H(\mathbf{P})$, $H(\boldsymbol{p})$ and $H(\boldsymbol{q})$ are the entropies of the respective distributions.

We consider the constrained problem of searching for $\mathbf{P}$ that minimizes $\langle \mathbf{M}, \mathbf{P} \rangle$ under the constraint $H(\mathbf{P}) \geq$ const. This is equivalent to imposing the condition that $\mathbf{P}$ lies within a KL-divergence ball centered at $\mathbf{P}_D$,

$$KL\left[\mathbf{P} : \mathbf{P}_D\right] \leq d \tag{11}$$

for constant d, because the KL-divergence from plan $\mathbf{P}$ to $\mathbf{P}_D$ is

$$KL\left[\mathbf{P} : \mathbf{P}_D\right] = \sum P_{ij} \log \frac{P_{ij}}{p_i q_j} = -H(\mathbf{P}) + H(\boldsymbol{p}) + H(\boldsymbol{q}). \tag{12}$$

The entropy of $\mathbf{P}$ increases within the ball as d increases. Therefore, this is equivalent to the entropy constrained problem that minimizes a linear combination of the transportation cost $\langle \mathbf{M}, \mathbf{P} \rangle$ and entropy $H(\mathbf{P})$,

$$F_\lambda(\mathbf{P}) = \langle \mathbf{M}, \mathbf{P} \rangle - \lambda H(\mathbf{P}) \tag{13}$$

for constant λ [7]. Here, λ is regarded as a Lagrangian multiplier for the entropy constraint and λ becomes smaller as d becomes larger.

3 Solution to the entropy-constrained problem: Cuturi function

Let us fix λ as a parameter controlling the magnitude of the entropy or the size of the KL-ball. When $\mathbf{P}$ satisfies the constraints in Eq. (4), minimization of Eq. (13) is formulated in the Lagrangian form by using Lagrangian multipliers α_i, β_j,

$$L_\lambda(\mathbf{P}) = \frac{1}{1+\lambda} \langle \mathbf{M}, \mathbf{P} \rangle - \frac{\lambda}{1+\lambda} H(\mathbf{P}) - \sum_{i,j} \left(\alpha_i + \beta_j\right) P_{ij}, \tag{14}$$

where λ in (13) is slightly modified. By differentiating Eq. (14) with respect to P_{ij}, we have

$$\frac{1+\lambda}{\lambda}\frac{\partial}{\partial P_{ij}}L_\lambda(\mathbf{P}) = \frac{1}{\lambda}m_{ij} + \log P_{ij} - \frac{1+\lambda}{\lambda}\left(\alpha_i + \beta_j\right) + 1. \tag{15}$$

By setting the above derivatives equal to 0, we have the following solution,

$$P_{ij} \propto \exp\left\{-\frac{m_{ij}}{\lambda} + \frac{1+\lambda}{\lambda}\left(\alpha_i + \beta_j\right)\right\}. \tag{16}$$

Let us put

$$K_{ij} = \exp\left\{-\frac{m_{ij}}{\lambda}\right\}, \tag{17}$$

$$a_i = \exp\left(\frac{1+\lambda}{\lambda}\alpha_i\right), \quad b_j = \exp\left(\frac{1+\lambda}{\lambda}\beta_j\right). \tag{18}$$

Then, the optimal solution is written as

$$P^*_{ij} = ca_i b_j K_{ij}, \tag{19}$$

where a_i and b_j are positive and correspond to the Lagrangian multipliers α_i and β_j to be determined from the constraints [Eq. (4)]. c is the normalization constant. Since $r + s$ constraints [Eq. (4)] are not independent because of the conditions that $\sum p_i = 1$ and $\sum q_j = 1$, we can use $b_r = 1$. Further, we noted that $\mu \boldsymbol{a}$ and $\boldsymbol{b}/\mu$ yield the same answer for any $\mu > 0$, where $\boldsymbol{a} = (a_i)$ and $\boldsymbol{b} = \left(b_j\right)$. Therefore, the degrees of freedom of $\boldsymbol{a}$ and $\boldsymbol{b}$ are $s - 1$ and $r - 1$, respectively. We can choose $\boldsymbol{a}$ and $\boldsymbol{b}$ such that they satisfy

$$\sum a_i = 1, \quad \sum b_j = 1. \tag{20}$$

Then, $\boldsymbol{a}$ and $\boldsymbol{b}$ are included in S_{s-1} and S_{r-1} respectively. We have the theorem below.

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

$$P^*_{\lambda ij} = ca_i b_j K_{ij}, \tag{21}$$

$$c = \frac{1}{\sum a_i b_j K_{ij}}, \tag{22}$$

where two vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ are determined from $\boldsymbol{p}$ and $\boldsymbol{q}$ using Eq. (4).

We have a generalized cost function of transporting $\boldsymbol{p}$ to $\boldsymbol{q}$ based on the entropy-constrained optimal plan $\mathbf{P}^*_\lambda(\boldsymbol{p}, \boldsymbol{q})$:

$$C_\lambda(\boldsymbol{p}, \boldsymbol{q}) = \frac{1}{1+\lambda}\left\langle \mathbf{M}, \mathbf{P}^*_\lambda \right\rangle - \frac{\lambda}{1+\lambda}H\left(\mathbf{P}^*_\lambda\right). \tag{23}$$

We called it the Cuturi function because extensive studies have been conducted by Cuturi and colleagues [7–9]. The function has been used in various applications as a measure of discrepancy between $\boldsymbol{p}$ and $\boldsymbol{q}$. The following theorem holds for the Cuturi function:

Theorem 2 *The Cuturi function $C_\lambda(\boldsymbol{p}, \boldsymbol{q})$ is a convex function of $(\boldsymbol{p}, \boldsymbol{q})$.*

Proof Let $\mathbf{P}_1^*$ and $\mathbf{P}_2^*$ be the optimal solutions of transportation problems $(\boldsymbol{p}_1, \boldsymbol{q}_1)$ and $(\boldsymbol{p}_2, \boldsymbol{q}_2)$, respectively. For scalar $0 \leq \nu \leq 1$, we use

$$\bar{\mathbf{P}} = \nu \mathbf{P}_1^* + (1-\nu)\mathbf{P}_2^*. \tag{24}$$

We have

$$\begin{aligned}
&\nu C_\lambda\left(\boldsymbol{p}_1 : \boldsymbol{q}_1\right) + (1-\nu) C_\lambda\left(\boldsymbol{p}_2 : \boldsymbol{q}_2\right)\\
&= \frac{1}{(1+\lambda)}\left\{\nu\langle \mathbf{M}, \mathbf{P}_1^*\rangle + (1-\nu)\langle \mathbf{M}, \mathbf{P}_2^*\rangle\right\} - \frac{\lambda}{1+\lambda}\left\{\nu H\left(\mathbf{P}_1^*\right) + (1-\nu) H\left(\mathbf{P}_2^*\right)\right\}\\
&\geq \frac{1}{(1+\lambda)}\langle \mathbf{M}, \bar{\mathbf{P}}\rangle - \frac{\lambda}{1+\lambda} H\left(\bar{\mathbf{P}}\right),
\end{aligned} \tag{25}$$

because $H(\mathbf{P})$ is a concave function of $\mathbf{P}$. We further have

$$\begin{aligned}
&\frac{1}{(1+\lambda)}\langle \mathbf{M}, \bar{\mathbf{P}}\rangle - \frac{\lambda}{1+\lambda} H\left(\bar{\mathbf{P}}\right) \geq \min_{\mathbf{P}}\left\{\frac{1}{(1+\lambda)}\langle \mathbf{M}, \mathbf{P}\rangle - \frac{\lambda}{1+\lambda} H(\mathbf{P})\right\}\\
&\quad = C_\lambda\left\{\nu \mathbf{p}_1 + (1-\nu)\mathbf{p}_2, \nu \boldsymbol{q}_1 + (1-\nu)\boldsymbol{q}_2\right\},
\end{aligned} \tag{26}$$

since the minimum is taken for $\mathbf{P}$ transporting commodities from $\nu \boldsymbol{p}_1 + (1-\nu)\boldsymbol{p}_2$ to $\nu \boldsymbol{q}_1 + (1-\nu)\boldsymbol{q}_2$. Hence, the convexity of C_λ is proven. □

When $\lambda \to 0$, it converges to the original Wasserstein distance $C_W(\boldsymbol{p}, \boldsymbol{q})$. However, it does not satisfy important requirements for “distance”. When $\boldsymbol{p} = \boldsymbol{q}$, C_λ is not equal to 0 and does not take the minimum value, i.e., there are some $\boldsymbol{q}$ $(\neq \boldsymbol{p})$ that yield smaller C_λ than $\boldsymbol{q} = \boldsymbol{p}$:

$$C_\lambda(\boldsymbol{p}, \boldsymbol{p}) > C_\lambda(\boldsymbol{p}, \boldsymbol{q}). \tag{27}$$

4 Geometry of optimal transportation plans

We first showed that a set of optimal transportation plans forms an exponential family embedded within the manifold of all transportation plans. Then, we studied the invariant geometry induced within these plans. A transportation plan $\mathbf{P}$ is a probability distribution over branches (i, j) connecting terminals of $\chi_i \in \chi_S$ and $\chi_j \in \chi_R$. Let x denote branches (i, j). We used the delta function $\delta_{ij}(x)$, which is 1 when x is (i, j) and 0 otherwise. Then, $\mathbf{P}$ is written as a probability distribution of the random variable x,

$$P(x) = \sum_{i,j} P_{ij}\delta_{ij}(x). \tag{28}$$

By introducing new parameters

$$\theta^{ij} = \log \frac{P_{ij}}{P_{sr}}, \quad \boldsymbol{\theta} = \left(\theta^{ij}\right), \tag{29}$$

it is rewritten in parameterized form as

$$P(x, \boldsymbol{\theta}) = \exp\left\{\sum_{i,j} \theta^{ij}\delta_{ij}(x) + \log P_{sr}\right\}. \tag{30}$$

This shows that the set of transportation plans is an exponential family, where θ^{ij} are the canonical parameters and $\eta_{ij} = P_{ij}$ are the expectation parameters. They form an $(sr - 1)$-dimensional manifold denoted by S_{TP}, because $\theta^{sr} = 0$.

The transportation problem is related to various problems in information theory such as the rate-distortion theory. We provide detailed studies on the transportation plans in the information-geometric framework in Sect. 7, but here we introduce the manifold of the optimal transportation plans, which are determined by the senders and receivers probability distributions $\boldsymbol{p}$ and $\boldsymbol{q}$.

The optimal transportation plan specified by $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ in Eq. (16) is written as

$$P_\lambda(x, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \exp\left[\sum_{i,j}\left\{\frac{1+\lambda}{\lambda}\left(\alpha_i + \beta_j\right) - \frac{m_{ij}}{\lambda}\right\}\delta_{ij}(x) - \frac{1+\lambda}{\lambda}\psi_\lambda\right]. \tag{31}$$

The notation ψ is a normalization factor called the potential function which is defined by

$$\psi_\lambda(\boldsymbol{\alpha}, \boldsymbol{\beta}) = -\frac{\lambda}{1+\lambda}\log c, \tag{32}$$

where c is calculated by taking the summation over all of x,

$$c = \sum_{x \in (\chi_S, \chi_R)} \exp\left[\sum_{i,j}\left\{\frac{1+\lambda}{\lambda}\left(\alpha_i + \beta_j\right) - \frac{m_{ij}}{\lambda}\right\}\delta_{ij}(x)\right]. \tag{33}$$

This corresponds to the free energy in physics.

Using

$$\theta^{ij} = \frac{1+\lambda}{\lambda}\left(\alpha_i + \beta_j\right) - \frac{m_{ij}}{\lambda}, \tag{34}$$

we see that the set $S_{OTP,\lambda}$ of the optimal transformation plans is a submanifold of S_{TP}. Because Eq. (34) is linear in $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, $S_{OTP,\lambda}$ itself is an exponential family, where the canonical parameters are $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ and the expectation parameters are $(\boldsymbol{p}, \boldsymbol{q}) \in S_{s-1} \times S_{r-1}$. This is confirmed by

$$\mathrm{E}\left[\sum_j \delta_{ij}(x)\right] = p_i, \tag{35}$$

$$\mathrm{E}\left[\sum_i \delta_{ij}(x)\right] = q_j, \tag{36}$$

where E denotes the expectation. Because of $\boldsymbol{p} \in S_{s-1}$ and $\boldsymbol{q} \in S_{r-1}$, $S_{OPT,\lambda}$ is a $(r+s-2)$-dimensional dually flat manifold, We can use $\alpha_s = \beta_r = 0$ without loss of generality, which corresponds to using $a_s = b_r = 1$ instead of the normalization $\sum a_i = \sum b_j = 1$ of $\boldsymbol{a}$ and $\boldsymbol{b}$.

In a dually flat manifold, the dual potential function φ_λ is given from the potential function ψ_λ as its Legendre dual, which is given by

$$\varphi_\lambda(\boldsymbol{p}, \boldsymbol{q}) = \boldsymbol{p} \cdot \boldsymbol{\alpha} + \boldsymbol{q} \cdot \boldsymbol{\beta} - \psi_\lambda(\boldsymbol{\alpha}, \boldsymbol{\beta}). \tag{37}$$

When we use new notations $\boldsymbol{\eta} = (\boldsymbol{p}, \boldsymbol{q})^T$, $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\beta})^T$, we have

$$\psi_\lambda(\boldsymbol{\theta}) + \varphi_\lambda(\boldsymbol{\eta}) = \boldsymbol{\theta} \cdot \boldsymbol{\eta}, \tag{38}$$

which is the Legendre relationship between $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$, and we have the following theorem:

Theorem 3 *The dual potential φ_λ is equal to the Cuturi function C_λ.*

Proof Direct calculation of Eq. (37) gives

$$\begin{aligned}\varphi_\lambda(\boldsymbol{p}, \boldsymbol{q}) &= \boldsymbol{p} \cdot \boldsymbol{\alpha} + \boldsymbol{q} \cdot \boldsymbol{\beta} - \psi_\lambda(\boldsymbol{\alpha}, \boldsymbol{\beta}) \\ &= \frac{1}{1+\lambda} \langle \mathbf{M}, \mathbf{P} \rangle + \sum_{i,j} P_{ij} \left\{ \left(\alpha_i + \beta_j\right) - \frac{1}{1+\lambda} m_{ij} - \psi_\lambda \right\} \\ &= \frac{1}{1+\lambda} \langle \mathbf{M}, \mathbf{P} \rangle + \frac{\lambda}{1+\lambda} \sum_{i,j} P_{ij} \left(\log a_i + \log b_j - \frac{m_{ij}}{\lambda} + \log c \right) \\ &= C_\lambda(\boldsymbol{p}, \boldsymbol{q}). \end{aligned} \tag{39}$$

□

We summarize the Legendre relationship below.

Theorem 4 *The dual potential function φ_λ (Cuturi function) and potential function (free energy, cumulant generating function) ψ_λ of the exponential family $S_{OPT,\lambda}$ are both convex, connected by the Legendre transformation,*

$$\boldsymbol{\theta} = \nabla_{\boldsymbol{\eta}} \varphi_\lambda(\boldsymbol{\eta}), \quad \boldsymbol{\eta} = \nabla_{\boldsymbol{\theta}} \psi_\lambda(\boldsymbol{\theta}), \tag{40}$$

or

$$\boldsymbol{\alpha} = \nabla_{\boldsymbol{p}} \varphi_\lambda(\boldsymbol{p}, \boldsymbol{q}), \quad \boldsymbol{\beta} = \nabla_{\boldsymbol{q}} \varphi_\lambda(\boldsymbol{p}, \boldsymbol{q}), \tag{41}$$

$$\boldsymbol{p} = \nabla_{\boldsymbol{\alpha}} \psi_\lambda(\boldsymbol{\alpha}, \boldsymbol{\beta}), \quad \boldsymbol{q} = \nabla_{\boldsymbol{\beta}} \psi_\lambda(\boldsymbol{\alpha}, \boldsymbol{\beta}). \tag{42}$$

Since $S_{OPT,\lambda}$ is dually flat, we can introduce a Riemannian metric and cubic tensor. The Riemannian metric $\mathbf{G}_\lambda$ is given to $S_{s-1} \times S_{r-1}$ by

$$\mathbf{G}_\lambda = \nabla_{\boldsymbol{\eta}} \nabla_{\boldsymbol{\eta}} \varphi_\lambda(\boldsymbol{\eta}) \tag{43}$$

in the $\boldsymbol{\eta}$-coordinate system $(\boldsymbol{p}, \boldsymbol{q})$. Its inverse is

$$\mathbf{G}_\lambda^{-1} = \nabla_{\boldsymbol{\theta}} \nabla_{\boldsymbol{\theta}} \psi_\lambda(\boldsymbol{\theta}). \tag{44}$$

Calculating Eq. (44) carefully, we have the following theorem:

Theorem 5 *The Fisher information matrix $\mathbf{G}_\lambda^{-1}$ in the $\boldsymbol{\theta}$-coordinate system is given by*

$$\mathbf{G}_\lambda^{-1} = \left[\begin{array}{c|c} p_i \delta_{ij} - p_i p_j & P_{ij} - p_i q_j \\ \hline P_{ij} - p_i q_j & q_i \delta_{ij} - q_i q_j \end{array} \right]. \tag{45}$$

Remark 1 The $\boldsymbol{p}$-part and $\boldsymbol{q}$-part of $\mathbf{G}_\lambda^{-1}$ are equal to the corresponding Fisher information in S_{s-1} and S_{r-1} in the e-coordinate systems.

Remark 2 The $\boldsymbol{p}$-part and the $\boldsymbol{q}$-part of $\mathbf{G}_\lambda$ are not equal to the corresponding Fisher information in the m-coordinate system. This is because $(\boldsymbol{p}, \boldsymbol{q})$-part of $\mathbf{G}$ is not 0.

We can similarly calculate the cubic tensor,

$$\mathbf{T} = \nabla\nabla\nabla\psi_\lambda \tag{46}$$

but we have not shown the results here.

From the Legendre pair of convex functions φ_λ and ψ_λ, we can also introduce the canonical divergence between two transportation problems $(\boldsymbol{p}, \boldsymbol{q})$ and $(\boldsymbol{p}', \boldsymbol{q}')$,

$$D_\lambda\left[(\boldsymbol{p}, \boldsymbol{q}) : \left(\boldsymbol{p}', \boldsymbol{q}'\right)\right] = \psi_\lambda(\boldsymbol{\alpha}, \boldsymbol{\beta}) + \varphi_\lambda(\boldsymbol{p}', \boldsymbol{q}') - \boldsymbol{\alpha} \cdot \boldsymbol{p}' - \boldsymbol{\beta} \cdot \boldsymbol{q}' \tag{47}$$

where $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ corresponds to $(\boldsymbol{p}, \boldsymbol{q})$. This is the KL-divergence between the two optimal transportation plans,

$$D_\lambda\left[(\boldsymbol{p}, \boldsymbol{q}) : \left(\boldsymbol{p}', \boldsymbol{q}'\right)\right] = KL[P_\lambda(\boldsymbol{p}, \boldsymbol{q}) : P_\lambda(\boldsymbol{p}', \boldsymbol{q}')]. \tag{48}$$

5 λ-divergences in S_{n-1}

5.1 Derivation of λ-divergences

We define a divergence between $\boldsymbol{p} \in S_{n-1}$ and $\boldsymbol{q} \in S_{n-1}$ using the canonical divergence in the set $S_{OTP,\lambda}$ of the optimal transportation plans [Eq. (48)]. For the sake of simplicity, we hereafter only study the case $\chi_S = \chi_R = \chi$. We introduce a reference distribution $\boldsymbol{r} \in S_{n-1}$ and define the $\boldsymbol{r}$-referenced divergence between $\boldsymbol{p}$ and $\boldsymbol{q}$ by

$$D_{\boldsymbol{r},\lambda}[\boldsymbol{p} : \boldsymbol{q}] = \gamma_\lambda KL\left[\mathbf{P}^*_\lambda(\boldsymbol{r}, \boldsymbol{p}) : \mathbf{P}^*_\lambda(\boldsymbol{r}, \boldsymbol{q})\right], \tag{49}$$

where γ_λ is a scaling factor, which we discuss later, and $\mathbf{P}^*_\lambda(\boldsymbol{r}, \boldsymbol{p})$ is the optimal transportation plan from $\boldsymbol{r}$ to $\boldsymbol{p}$. Note that its dual

$$\tilde{D}_{\boldsymbol{r},\lambda}[\boldsymbol{p}, \boldsymbol{q}] = \gamma_\lambda KL\left[\mathbf{P}^*_\lambda(\boldsymbol{r}, \boldsymbol{q}) : \mathbf{P}^*_\lambda(\boldsymbol{r}, \boldsymbol{p})\right] \tag{50}$$

is another candidate. There are other combinations but we study only Eq. (49) as the first step.

There are various ways of choosing a reference distribution $\boldsymbol{r}$. We first considered the simple choice of $\boldsymbol{r} = \boldsymbol{p}$, yielding the following λ-divergence:

$$D_\lambda[\boldsymbol{p} : \boldsymbol{q}] = \gamma_\lambda KL\left[\mathbf{P}^*_\lambda(\boldsymbol{p}, \boldsymbol{p}) : \mathbf{P}^*_\lambda(\boldsymbol{p}, \boldsymbol{q})\right]. \tag{51}$$

Theorem 6 *$D_\lambda[\boldsymbol{p} : \boldsymbol{q}]$ with the scaling factor $\gamma_\lambda = \frac{\lambda}{1+\lambda}$ is given by*

$$D_\lambda[\boldsymbol{p} : \boldsymbol{q}] = C_\lambda(\boldsymbol{p}, \boldsymbol{p}) - C_\lambda(\boldsymbol{p}, \boldsymbol{q}) - \nabla_{\boldsymbol{q}} C_\lambda(\boldsymbol{p}, \boldsymbol{q}) \cdot (\boldsymbol{p} - \boldsymbol{q}), \tag{52}$$

which is constructed from the Cuturi function.

Proof The optimal transportation plans are rewritten by the $\boldsymbol{\theta}$ coordinates in the form

$$\frac{\lambda}{1+\lambda} \log \mathbf{P}^*_\lambda(\boldsymbol{p}, \boldsymbol{p})_{ij} = \alpha'_i + \beta'_j - \frac{m_{ij}}{\lambda} - \psi'_\lambda, \tag{53}$$

$$\frac{\lambda}{1+\lambda} \log \mathbf{P}^*_\lambda(\boldsymbol{p}, \boldsymbol{q})_{ij} = \alpha_i + \beta_j - \frac{m_{ij}}{\lambda} - \psi_\lambda. \tag{54}$$

Then, we have

$$\begin{aligned} D_\lambda[\boldsymbol{p} : \boldsymbol{q}] &= \boldsymbol{p} \cdot \boldsymbol{\alpha}' + \boldsymbol{p} \cdot \boldsymbol{\beta}' - \psi'_\lambda - \boldsymbol{p} \cdot \boldsymbol{\alpha} - \boldsymbol{q} \cdot \boldsymbol{\beta} - \psi_\lambda - (\boldsymbol{p} - \boldsymbol{q}) \cdot \boldsymbol{\beta} \\ &= \varphi_\lambda(\boldsymbol{p}, \boldsymbol{p}) - \varphi_\lambda(\boldsymbol{p}, \boldsymbol{q}) - \nabla_{\boldsymbol{q}} \varphi_\lambda(\boldsymbol{p}, \boldsymbol{q}) \cdot (\boldsymbol{p} - \boldsymbol{q}). \end{aligned} \tag{55}$$

Since we showed that $\varphi_\lambda = C_\lambda$ in Theorem 3, we obtain Eq. (52). □

This is a divergence function satisfying $D_\lambda[\boldsymbol{p} : \boldsymbol{q}] \geq 0$, with equality when and only when $\boldsymbol{p} = \boldsymbol{q}$. However, it is not a canonical divergence of a dually flat manifold. The Bregman divergence derived from a convex function $\tilde{\varphi}(\boldsymbol{p})$ is given by

$$\tilde{D}_\lambda[\boldsymbol{p} : \boldsymbol{q}] = \tilde{\varphi}(\boldsymbol{p}) - \tilde{\varphi}(\boldsymbol{q}) - \nabla_{\boldsymbol{p}}\tilde{\varphi}(\boldsymbol{q}) \cdot (\boldsymbol{p} - \boldsymbol{q}). \tag{56}$$

This is different from Eq. (52), which is derived from $\varphi_\lambda(\boldsymbol{p}, \boldsymbol{q})$. Thus, we call $D_\lambda[\boldsymbol{p} : \boldsymbol{q}]$ Bregman-like divergence.

In the extremes of λ, the proposed divergence $D_\lambda[\boldsymbol{p} : \boldsymbol{q}]$ is related to the KL-divergence and Wasserstein distance in the following sense:

1. When $\lambda \to \infty$, D_λ converges to $KL[\boldsymbol{p} : \boldsymbol{q}]$. This is because $\mathbf{P}^*$ converges to $\boldsymbol{p} \otimes \boldsymbol{q}$ in the limit and we easily have

$$KL[\boldsymbol{p} \otimes \boldsymbol{p} : \boldsymbol{p} \otimes \boldsymbol{q}] = KL[\boldsymbol{p} : \boldsymbol{q}]. \tag{57}$$

2. When $\lambda \to 0$, D_λ with $\gamma_\lambda = \lambda/(1+\lambda)$ converges to 0, because $KL\left[\mathbf{P}_0^*(\boldsymbol{p}, \boldsymbol{p}) : \mathbf{P}_0^*(\boldsymbol{p}, \boldsymbol{q})\right]$ takes a finite value (see Example 1 in the next section). This suggests that it is preferable to use a scaling factor other than $\gamma_\lambda = \lambda/(1+\lambda)$ when λ is small. When $\lambda = 0$, $C_\lambda = \varphi_\lambda$ is not differentiable. Hence, we cannot construct the Bregman-like divergence from C_0 [Eq. (52)] in a simple example given in Sect. 5.3.

5.2 Other choices of reference distribution $\boldsymbol{r}$

We can consider other choices of the reference distribution $\boldsymbol{r}$. One option is choosing $\boldsymbol{r}$, which minimizes the KL-divergence.

$$\tilde{D}_\lambda[\boldsymbol{p} : \boldsymbol{q}] = \gamma_\lambda \min_{\boldsymbol{r}} KL\left[\mathbf{P}_\lambda^*(\boldsymbol{r}, \boldsymbol{p}) : \mathbf{P}_\lambda^*(\boldsymbol{r}, \boldsymbol{q})\right]. \tag{58}$$

However, obtaining the minimizer $\boldsymbol{r}$ is not computationally easy. Thus, we can simply replace the optimal $\boldsymbol{r}$ with the arithmetic mean or geometric mean of $\boldsymbol{p}$ and $\boldsymbol{q}$. The arithmetic mean is given by the m-mixture midpoint of $\boldsymbol{p}$ and $\boldsymbol{q}$,

$$\boldsymbol{r} = \frac{1}{2}(\boldsymbol{p} + \boldsymbol{q}). \tag{59}$$

The geometric mean is given by the e-midpoint of $\boldsymbol{p}$ and $\boldsymbol{q}$,

$$\boldsymbol{r} = c\left(\sqrt{p_i q_i}\right), \tag{60}$$

where c is the normalization constant.

5.3 Examples of λ-divergence

Below, we consider the case where $\boldsymbol{r} = \boldsymbol{p}$. We show two simple examples, where $D_\lambda(\boldsymbol{p}, \boldsymbol{q})$ can be analytically computed.

Example 1 Let $n = 2$ and

$$m_{ii} = 0, \quad m_{ij} = 1 \quad (i \neq j). \tag{61}$$

We use $a_2 = b_2 = 1$ for normalization,

$$P_{ij} = ca_i b_j K_{ij}, \tag{62}$$

$$K_{ij} = \exp\left\{-\frac{m_{ij}}{\lambda}\right\} = \begin{bmatrix} 1 & \varepsilon \\ \varepsilon & 1 \end{bmatrix}, \tag{63}$$

$$\varepsilon = \exp\left\{-\frac{1}{\lambda}\right\}. \tag{64}$$

Note that $\varepsilon \to 0$ as $\lambda \to 0$.

When $\lambda > 0$, the receiver conditions require

$$cab + ca\varepsilon = p, \tag{65}$$

$$cab + cb\varepsilon = q, \tag{66}$$

where we use $a = a_1$, $b = b_1$ and

$$c = \frac{1}{ab + \varepsilon(a+b) + 1}. \tag{67}$$

Solving the above equations, we have

$$a = \frac{z - (q-p)/\varepsilon}{2(1-p)}, \tag{68}$$

$$b = \frac{z + (q-p)/\varepsilon}{2(1-q)}, \tag{69}$$

where

$$z = -\varepsilon(1-p-q) + \sqrt{(q-p)^2/\varepsilon^2 + \varepsilon^2(1-p-q)^2 + 2p(1-p) + 2q(1-q)}.$$

We can show $D_\lambda[\boldsymbol{p} : \boldsymbol{q}]$ explicitly by using the solution, although it is complicated. When $\lambda = 0$, we easily have

$$C_0(p, q) = |p - q|, \tag{70}$$

where $\boldsymbol{p} = (p, 1-p)$ and $\boldsymbol{q} = (q, 1-q)$. $C_0(p, q)$ is piecewise linear, and cannot be used to construct a Bregman-like divergence. However, we can calculate the limiting case of $\lambda \to 0$ because the optimal transportation plans $\mathbf{P}^*$ where λ is small are directly calculated by minimizing $C_\lambda(\boldsymbol{p}, \boldsymbol{q})$ as

$$\mathbf{P}_{\lambda}^{*}(\boldsymbol{p}, \boldsymbol{p}) = \begin{bmatrix} p & 0 \\ 0 & 1-p \end{bmatrix} + \begin{bmatrix} -\varepsilon & \varepsilon \\ \varepsilon & -\varepsilon \end{bmatrix}, \tag{71}$$

$$\mathbf{P}_{\lambda}^{*}(\boldsymbol{p}, \boldsymbol{q}) = \begin{bmatrix} p & 0 \\ q-p & 1-q \end{bmatrix} + \begin{bmatrix} -\varepsilon^2 & \varepsilon^2 \\ \varepsilon^2 & -\varepsilon^2 \end{bmatrix}. \tag{72}$$

where we set $q > p$. The limit of KL divergence is given by

$$\lim_{\lambda \to 0} KL[\mathbf{P}_{\lambda}^{*}(\boldsymbol{p}, \boldsymbol{p}) : \mathbf{P}_{\lambda}^{*}(\boldsymbol{p}, \boldsymbol{q})] = \begin{cases} p \log \frac{p}{q} & (p \geq q), \\ (1-p) \log \frac{1-p}{1-q} & (p < q). \end{cases} \tag{73}$$

In the general case of $n \geq 2$, the optimal transportation plan is $\mathbf{P}_0^{*}(\boldsymbol{p}, \boldsymbol{p}) = (p_i \delta_{ij})$. The diagonal parts of the optimal $\mathbf{P}_0^{*}(\boldsymbol{p}, \boldsymbol{q})$ are $\min\{p_i, q_i\}$ when $m_{ii} = 0, \ \ m_{ij} > 0 \ \ (i \neq j)$. Thus, the KL divergence is given by

$$KL[\mathbf{P}_0^{*}(\boldsymbol{p}, \boldsymbol{p}) : \mathbf{P}_0^{*}(\boldsymbol{p}, \boldsymbol{q})] = \sum_{i; p_i > q_i} p_i \log \frac{p_i}{q_i}. \tag{74}$$

Remark that when $\lambda \to \infty$,

$$\lim_{\lambda \to \infty} KL[\mathbf{P}_{\lambda}^{*}(\boldsymbol{p}, \boldsymbol{p}) : \mathbf{P}_{\lambda}^{*}(\boldsymbol{p}, \boldsymbol{q})] = \sum_{i} p_i \log \frac{p_i}{q_i}. \tag{75}$$

Example 2 We take a family of Gaussian distributions $N\left(\mu, \sigma^2\right)$,

$$p\left(x \ ; \ \mu, \sigma^2\right) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \tag{76}$$

on the real line $\chi = \{x\}$, extending the discrete case to the continuous case. We transport $p\left(x \ ; \ \mu_p, \sigma_p^2\right)$ to $q\left(x \ ; \ \mu_q, \sigma_q^2\right)$, where the transportation cost is

$$m(x, y) = |x - y|^2. \tag{77}$$

Then, we have

$$K(x, y) = \exp\left\{-\frac{(x-y)^2}{2\lambda^2}\right\}, \tag{78}$$

where we use $2\lambda^2$ instead of previous λ for the sake of convenience.

The optimal transportation plan is written as

$$P^{*}(x, y) = ca(x)b(y)K(x, y), \tag{79}$$

where a and b are determined from

$$\int ca(x)b(y)K(x,y)dy = p(x), \tag{80}$$

$$\int ca(x)b(y)K(x,y)dx = q(x). \tag{81}$$

The solutions are given in the Gaussian framework, $x \sim N\left(\tilde{\mu}, \tilde{\sigma}^2\right)$, $y \sim N\left(\tilde{\mu}', \tilde{\sigma}'^2\right)$. As derived in Appendix A, the optimal cost and divergence are as follows:

$$\begin{aligned} C_\lambda(p,q) = \frac{1}{1+\lambda}\Bigg[&\left(\mu_p - \mu_q\right)^2 + \sigma_p^2 + \sigma_q^2 + \frac{\lambda}{2}\left(1 - \sqrt{1+X}\right) \\ &-\lambda\left\{\log\sigma_p\sigma_q + \frac{1}{2}\log 8\pi^2 e^2 - \frac{1}{2}\log\left(1+\sqrt{1+X}\right)\right\}\Bigg], \end{aligned} \tag{82}$$

$$\begin{aligned} D_\lambda[p:q] = \gamma_\lambda\Bigg[&\frac{1}{2}\left(\sqrt{1+X} - \sqrt{1+X_p}\right) + \log\frac{\sigma_q}{\sigma_p} + \frac{1}{2}\log\frac{1+\sqrt{1+X_p}}{1+\sqrt{1+X}} \\ &+\frac{1+\sqrt{1+X}}{4}\left\{\frac{\left(\mu_p-\mu_q\right)^2}{\sigma_q^2} + \frac{\sigma_p^2}{\sigma_q^2} - 1\right\}\Bigg], \end{aligned}$$

$$\text{where} \qquad X = \frac{16\sigma_p^2\sigma_q^2}{\lambda^2} \quad X_p = \frac{16\sigma_p^4}{\lambda^2}. \tag{83}$$

Note that $D_\lambda = KL\left[\mathbf{P}_\lambda^*(\boldsymbol{p},\boldsymbol{p}) : \mathbf{P}_\lambda^*(\boldsymbol{p},\boldsymbol{q})\right]$ diverges to infinity in the limit of $\lambda \to 0$ because the support of the optimal transport $\mathbf{P}_\lambda^*(\boldsymbol{p},\boldsymbol{q})$ reduces to a 1-dimensional subspace. To prevent D_λ from diverging and to make it finite, we set the scaling factor as $\gamma_\lambda = \frac{\lambda}{1+\lambda}$. In this case, D_λ is equivalent to the Bregman-like divergence of the Cuturi function as shown in Theorem 6. With this scaling factor γ_λ, D_λ in the limits of $\lambda \to \infty$ and $\lambda \to 0$ is given by

$$\lim_{\lambda\to\infty} D_\lambda = \frac{1}{2}\left\{\frac{(\mu_p-\mu_q)^2}{\sigma_q^2} + \frac{\sigma_p^2}{\sigma_q^2} - 1\right\} + \log\frac{\sigma_q}{\sigma_p} = KL[p:q], \tag{84}$$

$$\lim_{\lambda\to 0} D_\lambda = \frac{\sigma_p}{\sigma_q}(\mu_p-\mu_q)^2 + \frac{\sigma_p}{\sigma_q}(\sigma_p-\sigma_q)^2. \tag{85}$$

6 Applications of λ-divergence

6.1 Cluster center (barycenter)

Let $\boldsymbol{q}_1, \ldots, \boldsymbol{q}_k$ be k distributions in S_{n-1}. Its λ-center is defined by $\boldsymbol{p}^*$, which minimizes the average of λ-divergences from $\boldsymbol{q}_i$ to $\boldsymbol{p} \in S_{n-1}$,

$$\boldsymbol{p}^* = \arg\min_{\boldsymbol{p}} \sum D_\lambda[\boldsymbol{q}_i : \boldsymbol{p}]. \tag{86}$$

The center is obtained from

$$\partial_{\boldsymbol{p}} \sum_i D_\lambda \left[\boldsymbol{q}_i : \boldsymbol{p}\right] = 0, \tag{87}$$

which yields the equation to give $\boldsymbol{p}^*$,

$$\sum \mathbf{G}\left(\boldsymbol{q}_i, \boldsymbol{p}^*\right)\left(\boldsymbol{q}_i - \boldsymbol{p}^*\right) = 0, \tag{88}$$

where

$$\mathbf{G}(\boldsymbol{q}, \boldsymbol{p}) = \nabla_{\boldsymbol{p}} \nabla_{\boldsymbol{p}} \varphi_\lambda(\boldsymbol{q}, \boldsymbol{p}). \tag{89}$$

It is known [5] that the mean (center) of two Gaussian distributions $N\left(\mu_1, \sigma_1^2\right)$ and $N\left(\mu_2, \sigma_2^2\right)$ over the real line $\chi = \mathbf{R}$ is Gaussian $N\left(\frac{\mu_1+\mu_2}{2}, \frac{(\sigma_1+\sigma_2)^2}{4}\right)$, when we use the square of the Wasserstein distance W_2^2 with the cost function $|x_1 - x_2|^2$. It would be interesting to see how the center changes depending on λ based on $D_\lambda[\boldsymbol{p} : \boldsymbol{q}]$.

We consider the center of two Gaussian distributions $\boldsymbol{q}_1$ and $\boldsymbol{q}_2$, defined by

$$\boldsymbol{\eta}_p = \arg\min_{\boldsymbol{p}} \sum D_\lambda \left[\boldsymbol{p} : \boldsymbol{q}_i\right]. \tag{90}$$

When $\lambda \to 0$ and $\lambda \to \infty$, we have

$$\lambda \to \infty : \sigma_p^2 = \frac{2\sigma_{q_1}^2 \sigma_{q_2}^2}{\sigma_{q_1}^2 + \sigma_{q_2}^2}, \quad \mu_p = \frac{\sigma_{q_2}^2 \mu_{q_1} + \sigma_{q_1}^2 \mu_{q_2}}{\sigma_{q_1}^2 + \sigma_{q_2}^2}, \tag{91}$$

$$\lambda \to 0 : \sigma_p = \frac{2\sigma_{q_1}\sigma_{q_2}}{\sigma_{q_1} + \sigma_{q_2}}, \quad \mu_p = \frac{\sigma_{q_2}\mu_{q_1} + \sigma_{q_1}\mu_{q_2}}{\sigma_{q_1} + \sigma_{q_2}}. \tag{92}$$

However, if we use C_λ instead of D_λ the centers are

$$\lambda \to \infty : \sigma_p = \lambda, \tag{93}$$

$$\lambda \to 0 : \sigma_p = \frac{\sigma_{q_1} + \sigma_{q_2}}{2}, \tag{94}$$

which are not reasonable for large λ.

6.2 Statistical estimation

Let us consider a statistical model M,

$$M = \{p(\boldsymbol{x}, \boldsymbol{\xi})\} \tag{95}$$

parameterized by $\boldsymbol{\xi}$. An interesting example is the set of distributions over $\chi = (0, 1)^n$, where $\boldsymbol{x}$ is a vector random variable defined on the n-cube χ.

The Boltzmann machine M is its submodel, consisting of probability distributions which do not include higher-order interaction terms of random variables x_i,

$$p(\boldsymbol{x}, \boldsymbol{\xi}) = \exp\left\{\sum b_i x_i + \sum_{i<j} w_{ij} x_i x_j - \psi\right\}, \tag{96}$$

where $\boldsymbol{\xi} = \left(b_i, w_{ij}\right)$. The transportation cost is

$$\mathbf{m}(\boldsymbol{x}, \boldsymbol{y}) = \sum_i |x_i - y_i|, \tag{97}$$

which is the Hamming distance [10].

Let $\hat{\boldsymbol{q}} = \hat{\boldsymbol{q}}(\boldsymbol{x})$ be an empirical distribution. Then, D_λ-estimator $\boldsymbol{p}^* = \boldsymbol{p}^*(\boldsymbol{x}, \boldsymbol{\xi}^*) \in M$ is defined by

$$\boldsymbol{p}\left(\boldsymbol{x}, \boldsymbol{\xi}^*\right) = \arg\min_{\boldsymbol{\xi}} D_\lambda\left[\hat{\boldsymbol{q}} : p(\boldsymbol{x}, \boldsymbol{\xi})\right]. \tag{98}$$

Differentiating D_λ with respect to $\boldsymbol{\xi}$, we obtain the following theorem:

Theorem 7 *The λ-estimator $\boldsymbol{\xi}^*$ satisfies*

$$\mathbf{G}\left(\hat{\boldsymbol{q}}, \boldsymbol{p}\right)\left(\boldsymbol{p} - \hat{\boldsymbol{q}}\right)\frac{\partial p(\boldsymbol{x}, \boldsymbol{\xi}^*)}{\partial \boldsymbol{\xi}} = 0. \tag{99}$$

6.3 Pattern classifier

Let $\boldsymbol{p}_1$ and $\boldsymbol{p}_2$ be two prototype patterns of categories C_1 and C_2. A separating hyper-submanifold of the two categories is defined by the set of $\boldsymbol{q}$ that satisfy

$$D_\lambda\left[\boldsymbol{p}_1 : \boldsymbol{q}\right] = D_\lambda\left[\boldsymbol{p}_2 : \boldsymbol{q}\right] \tag{100}$$

or

$$D_\lambda\left[\boldsymbol{q} : \boldsymbol{p}_1\right] = D_\lambda\left[\boldsymbol{q} : \boldsymbol{p}_2\right]. \tag{101}$$

It would be interesting to study the geometrical properties of the λ-separating hyper-submanifold (Fig. 2).

7 Information geometry of transportation plans

We provide a general framework of the transportation plans from the viewpoint of information geometry. The manifold of all transportation plans is a probability simplex $M = S_{n^2-1}$ consisting of all the joint probability distributions $\mathbf{P}$ over $\chi \times \chi$. It is dually flat, where m-coordinates are $\eta_{ij} = P_{ij}$, from which P_{nn} is determined.

$$\sum P_{ij} = 1. \tag{102}$$

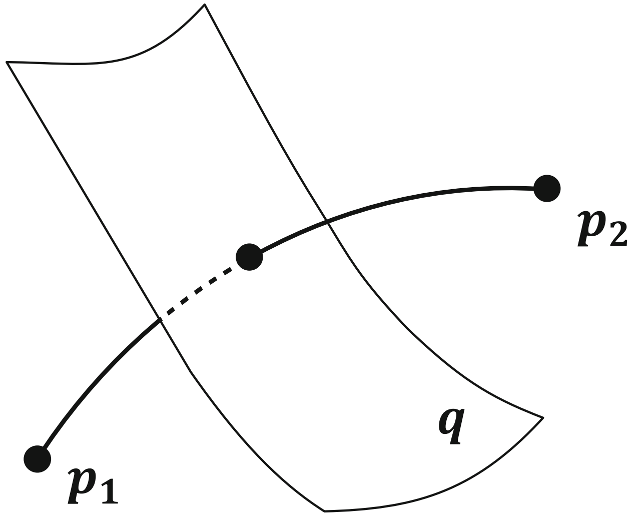

Fig. 2 λ-separating hyperplane

The corresponding e-coordinates are log P_{ij} divided by P_{nn} as

$$\theta^{ij} = \log \frac{P_{ij}}{P_{nn}}. \tag{103}$$

We considered three problems in $M = S_{n^2-1}$, when the cost matrix $\mathbf{M} = \left(m_{ij}\right)$ is given.

7.1 Free problem

Minimize the entropy-relaxed transportation cost $\varphi_\lambda(\mathbf{P})$ without any constraints on $\mathbf{P}$. The solution is

$$\mathbf{P}^*_{\text{free}} = \exp\left(-\frac{m_{ij}}{\lambda} - \frac{1+\lambda}{\lambda}\psi\right) = c\mathbf{K}, \tag{104}$$

where c is a normalization constant. This clarifies the meaning of the matrix $\mathbf{K}$ [Eq. (17)], i.e., $\mathbf{K}$ is the optimal transportation plan for the free problem.

7.2 Rate-distortion problem

We considered a communication channel in which $\boldsymbol{p}$ is a probability distribution on the senders terminals. The channel is noisy and P_{ij}/p_i is the probability that x_j is received when x_i is sent. The costs m_{ij} are regarded as the distortion of x_i changing to x_j. The rate distortion-problem in information theory searches for $\mathbf{P}$, which minimizes the mutual information of the sender and receiver under the constraint of distortion $\langle \mathbf{M}, \mathbf{P} \rangle$. The problem is formulated by maximizing $\varphi_\lambda(\mathbf{P})$ under the senders constraint $\boldsymbol{p}$, where $\boldsymbol{q}$ is free (R. Belavkin, personal communication; see also [16]).

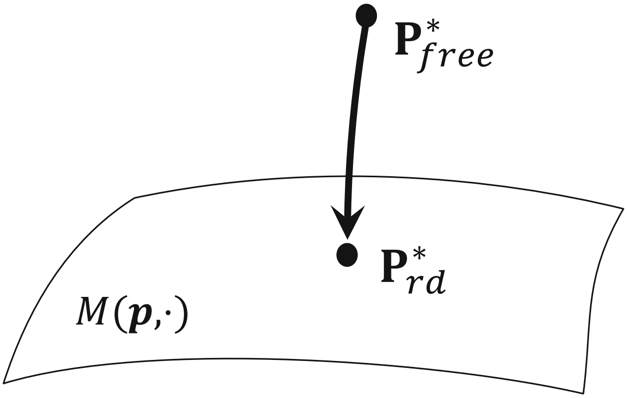

Fig. 3 *e*-projection in the rate-distortion problem

The optimal solution is given by

$$\mathbf{P}^*_{rd} = \left(ca_i K_{ij}\right), \tag{105}$$

since $\boldsymbol{q}$ is free and $\boldsymbol{\beta} = 0$ or $b_j = 1$. a_i are determined from $\boldsymbol{p}$ such that the senders condition

$$c \sum_j a_i K_{ij} = p_i \tag{106}$$

is satisfied. Therefore, the dual parameters a_i are given explicitly as

$$ca_i = \frac{p_i}{\sum_j K_{ij}}. \tag{107}$$

Let $M(\boldsymbol{p}, \cdot)$ be the set of plans that satisfy the senders condition

$$\sum_j P_{ij} = p_i. \tag{108}$$

Then, we will see that $\mathbf{P}^*_{rd}$ is the *e*-projection of $\mathbf{P}^*_{\text{free}}$ to $M(\boldsymbol{p}, \cdot)$. The *e*-projection is explicitly given by Eq. (107) (Fig. 3).

7.3 Transportation problem

A transportation plan satisfies the senders and receivers conditions. Let $M(\cdot, \boldsymbol{q})$ be the set of plans that satisfies the receivers conditions

$$\sum_i P_{ij} = q_j. \tag{109}$$

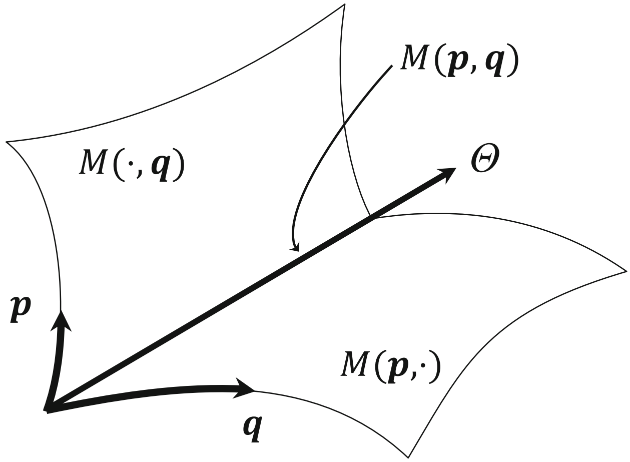
$M(\boldsymbol{p},\boldsymbol{q})$
$M(\cdot,\boldsymbol{q})$
Θ
$\boldsymbol{p}$
$M(\boldsymbol{p},\cdot)$
$\boldsymbol{q}$

Fig. 4 m-flat submanifolds in the transportation problem

Then, the transportation problem searches for the plan that minimizes the entropy-relaxed cost in the subset

$$M(\boldsymbol{p},\boldsymbol{q}) = M(\boldsymbol{p},\cdot) \cap M(\cdot,\boldsymbol{q}). \tag{110}$$

Since the constraints Eqs. (108) and (109) are linear in the m-coordinates $\mathbf{P}$, $M(\boldsymbol{p},\cdot)$, $M(\cdot,\boldsymbol{q})$ and $M(\boldsymbol{p},\boldsymbol{q})$ are m-flat submanifolds (Fig. 4).

Since $\boldsymbol{p}$ and $\boldsymbol{q}$ are fixed, $M(\boldsymbol{p},\boldsymbol{q})$ is of dimensions $(n-1)^2$, in which all the degrees of freedom represent mutual interactions between the sender and receiver. We define them by

$$\Theta_{ij} = \log \frac{P_{ij}P_{nn}}{P_{in}P_{nj}}, \quad i,j = 1,\ldots,n-1. \tag{111}$$

They vanish for $\mathbf{P}_D = \boldsymbol{p} \otimes \boldsymbol{q}$, as is easily seen Eq. (111). Since Θ_{ij} are linear in $\log P_{ij}$, the submanifold $E\left(\Theta_{ij}\right)$, in which Θ_{ij}'s take fixed values but $\boldsymbol{p}$ and $\boldsymbol{q}$ are free, is an $2(n-1)$-dimensional e-flat submanifold.

We introduce mixed coordinates

$$\Xi = \left(\boldsymbol{p},\boldsymbol{q},\Theta_{ij}\right) \tag{112}$$

such that the first $2(n-1)$ coordinates $(\boldsymbol{p},\boldsymbol{q})$ are the marginal distributions in the m-coordinates and the last $(n-1)^2$ coordinates Θ are interactions in the e-coordinates given in Eq. (111). Since the two complementary coordinates are orthogonal, we have orthogonal foliations of S_{n^2-1} [1] (Fig. 5).

Given two vectors $\boldsymbol{a} = (a_i)$ and $\boldsymbol{b} = \left(b_j\right)$, we considered the following transformation of $\mathbf{P}$,

$$T_{\boldsymbol{ab}}\mathbf{P} = \left(ca_ib_j\mathbf{P}_{ij}\right), \tag{113}$$

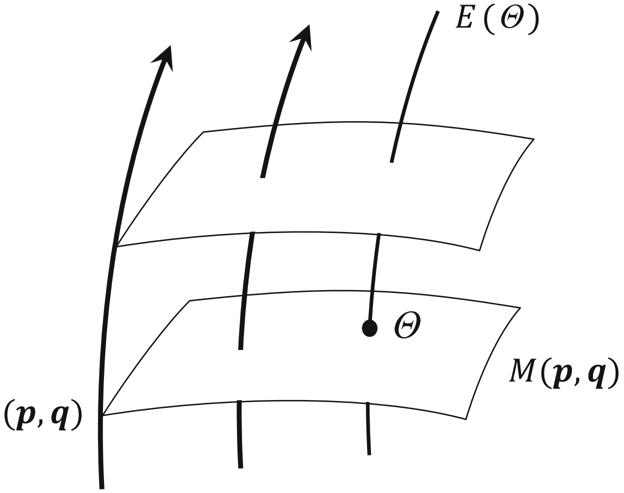
$E(\Theta)$
Θ
$M(\boldsymbol{p},\boldsymbol{q})$
$(\boldsymbol{p},\boldsymbol{q})$

Fig. 5 Orthogonal foliations of S_{n^2-1} with the mixed coordinates

where c is a constant determined from the normalization condition,

$$c\sum_{i,j} a_i b_j P_{ij} = 1. \tag{114}$$

Ξ is the mixed coordinates of $\mathbf{P}$ and m-flat submanifold $M(\boldsymbol{p},\boldsymbol{q})$, defined by fixing the first $2(n-1)$ coordinates, is orthogonal to e-flat submanifold $E(\Theta)$, defined by making the last $(n-1)^2$ coordinates equal to Θ_{ij}. This is called the RAS transformation in the input-output analysis of economics.

Lemma *For any* $\boldsymbol{a}$, $\boldsymbol{b}$, *transformation* $T_{\boldsymbol{ab}}$ *does not change the interaction terms* Θ_{ij}. *Moreover, the e-geodesic connecting* $\mathbf{P}$ *and* $T_{\boldsymbol{ab}}\mathbf{P}$ *is orthogonal to* $M(\boldsymbol{p},\boldsymbol{q})$.

Proof By calculating the mixed coordinates of $T_{\boldsymbol{ab}}\mathbf{P}$, we easily see that the Θ-part does not change. Hence, the e-geodesic connecting $\mathbf{P}$ and $T_{\boldsymbol{ab}}\mathbf{P}$ is given, in terms of the mixed coordinates, by keeping the last part fixed while changing the first part. This is included in $E(\Theta)$. Therefore, the geodesic is orthogonal to $M(\boldsymbol{p},\boldsymbol{q})$. □

Since the optimal solution is given by applying $T_{\boldsymbol{ab}}$ to $\mathbf{K}$, even when $\mathbf{K}$ is not normalized, such that the terminal conditions [Eq. (4)] are satisfied, we have the following theorem:

Theorem 8 *The optimal solution* $\mathbf{P}^*$ *is given by e-projecting* $\mathbf{K}$ *to* $M(\boldsymbol{p},\boldsymbol{q})$.

7.4 Iterative algorithm (Sinkhorn algorithm) for obtaining *a* and *b*

We need to calculate $\boldsymbol{a}$ and $\boldsymbol{b}$ when $\boldsymbol{p}$ and $\boldsymbol{q}$ are given for obtaining the optimal transportation plan. The Sinkhorn algorithm is well known for this purpose [5]. It is an iterative algorithm for obtaining the e-projection of $\mathbf{K}$ to $M(\boldsymbol{p},\boldsymbol{q})$.

Let $T_{A\cdot}$ be the e-projection of $\mathbf{P}$ to $M(\boldsymbol{p}, \cdot)$ and let $T_{\cdot B}$ be the e-projection to $M(\cdot, \boldsymbol{q})$. From the Pythagorean theorem, we have

$$KL\left[T_{A\cdot}\mathbf{P} : \mathbf{P}\right] + KL\left[\mathbf{P}^* : T_{A\cdot}\mathbf{P}\right] = KL\left[\mathbf{P}^* : \mathbf{P}\right], \tag{115}$$

where $\mathbf{P}^* = T_{\boldsymbol{ab}}\mathbf{P}$ is the optimal solution; that is, the e-projection of $\mathbf{K}$ to $M(\boldsymbol{p}, \boldsymbol{q})$. Hence, we have

$$KL\left[\mathbf{P}^* : T_{A\cdot}\mathbf{P}\right] \leq KL\left[\mathbf{P}^* : \mathbf{P}\right] \tag{116}$$

and the equality holds when and only when $\mathbf{P} \in M(\boldsymbol{p}, \cdot)$. The e-projection of $\mathbf{P}$ decreases the dual KL-divergence to $\mathbf{P}^*$. The same property holds for the e-projection to $M(\cdot, \boldsymbol{q})$. The iterative e-projections of $\mathbf{K}$ to $M(\boldsymbol{p}, \cdot)$ and $M(\cdot, \boldsymbol{q})$ converges to the optimal solution $\mathbf{P}^*$.

It is difficult to have an explicit expression of the e-projection of $\mathbf{P}$ to $M(\boldsymbol{p}, \boldsymbol{q})$, but those of e-projections to $M(\boldsymbol{p}, \cdot)$ and $M(\cdot, \boldsymbol{q})$ are easily obtained. The e-projection of $\mathbf{P}$ to $M(\boldsymbol{p}, \cdot)$ is given by

$$T_{A\cdot}\mathbf{P} = \left(a_i P_{ij}\right), \tag{117}$$

where $\boldsymbol{a}$ is given explicitly by

$$a_i = \frac{p_i}{\sum_j P_{ij}}. \tag{118}$$

Similarly, the e-projection to $M(\cdot, \boldsymbol{q})$ is given by

$$T_{\cdot B}\mathbf{P} = \left(b_j P_{ij}\right), \tag{119}$$

with

$$b_j = \frac{q_j}{\sum_i P_{ij}}. \tag{120}$$

Therefore, the iterative algorithm, which is known as the Sinkhorn Algorithm [7,12] of e-projection from $\mathbf{K}$ is formulated as follows:

Iterative e-projection algorithm

1. Begin with $\mathbf{P}_0 = \mathbf{K}$.
2. For $t = 0, 1, 2, \ldots$, e-project P_{2t} to $M(\boldsymbol{p}, \cdot)$ to obtain

$$\mathbf{P}_{2t+1} = T_{A\cdot}\mathbf{P}_{2t}. \tag{121}$$

3. To obtain $\mathbf{P}_{2t+2}$, e-project $\mathbf{P}_{2t+1}$ to $M(\cdot, \boldsymbol{q})$,

$$\mathbf{P}_{2t+2} = T_{\cdot B}\mathbf{P}_{2t+1}. \tag{122}$$

4. Repeat until convergence.

Figure 6 schematically illustrates the iterative e-projection algorithm for finding the optimal solution $\mathbf{P}^*$.

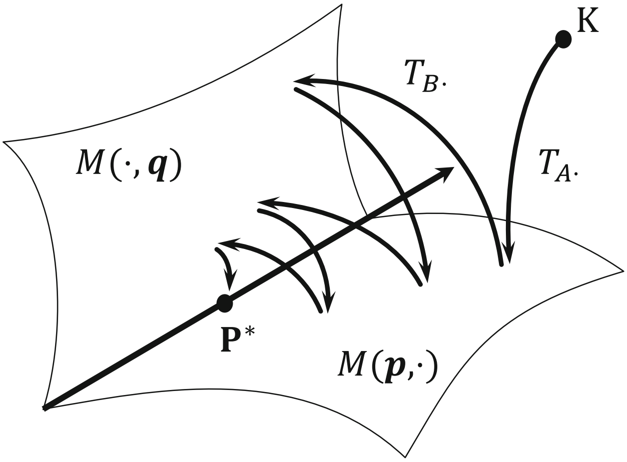

Fig. 6 Sinkhorn algorithm as iterative *e*-projections

8 Conclusions and additional remarks

We elucidated the geometry of optimal transportation plans and introduced a one-parameter family of divergences in the probability simplex which connects the Wasserstein distance and KL-divergence. A one-parameter family of Riemannian metrics and dually coupled affine connections were introduced in S_{n-1}, although they are not dually flat in general. We uncovered a new way of studying the geometry of probability distributions. Future studies should examine the properties of the λ-divergence and apply these to various problems. We touch upon some related problems below.

1. *Uniqueness of the optimal plan*
The original Wasserstein distance is obtained by solving a linear programming problem. Hence, the solution is not unique in some cases and is not necessarily a continuous function of **M**. However, the entropy-constrained solution is unique and continuous with respect to **M** [7]. While $\varphi_\lambda(\boldsymbol{p}, \boldsymbol{q})$ converges to $\varphi_0(\boldsymbol{p}, \boldsymbol{q})$ as $\lambda \to 0$, $\varphi_0(\boldsymbol{p}, \boldsymbol{q})$ is not necessarily differentiable with respect to $\boldsymbol{p}$ and $\boldsymbol{q}$.

2. *Integrated information theory of consciousness*
Given a joint probability distribution **P**, the amount of integrated information is measured by the amount of interactions of information among different terminals. We used a disconnected model in which no information is transferred through branches connecting different terminals. The geometric measure of integrated information theory is given by the KL-divergence from **P** to the submanifold of disconnected models [13,14]. However, the Wasserstein divergence can be considered as such a measure when the cost of transferring information through different terminals depends on the physical positions of the terminals [15]. We can use the entropy-constrained divergence D_λ to define the amount of information integration.

3. *f-divergence*
We used the KL-divergence in a dually flat manifold for defining D_λ. It is pos-

sible to use any other divergences, for example, the f-divergence instead of KL-divergence. We would obtain similar results.

4. *q-entropy*
Muzellec et al. used the α-entropy (Tsallis q-entropy) instead of the Shannon entropy for regularization [16]. This yields the q-entropy-relaxed framework.

5. *Comparison of C_λ and D_λ*
Although D_λ satisfies the criterion of a divergence, it might differ considerably from the original C_λ. In particular, when $C_\lambda(\boldsymbol{p}, \boldsymbol{q})$ includes a piecewise linear term such as $\sum d_i |p_i - q_i|$ for constant d_i, D_λ defined in Eq. (52) eliminates this term. When this term is important, we can use $\{C_\lambda(\boldsymbol{p}, \boldsymbol{q})\}^2$ instead of $C_\lambda(\boldsymbol{p}, \boldsymbol{q})$ for defining a new divergence D_λ in Eq. (52). In our accompanying paper [17], we define a new type of divergence that retains the properties of C_λ and is closer to C_λ.

Appendix: The proof of Example 2

Let us assume that functions $a(x)$ and $b(y)$ are constrained into Gaussian distributions: $a(x) = N\left(\tilde{\mu}, \tilde{\sigma}^2\right)$, $b(y) = N\left(\tilde{\mu}', \tilde{\sigma}'^2\right)$. This means that the optimal plan $P^*(x, y)$ is also given by a Gaussian distribution $N(\boldsymbol{\mu}, \Sigma)$. The marginal distributions p and q require the mean value of the optimal plan to become

$$\boldsymbol{\mu} = [\mu_p \;\; \mu_q]^T. \tag{A.1}$$

It is also necessary for the diagonal part of the covariance matrix to become

$$\Sigma_{11} = \sigma_p^2, \tag{A.2}$$

$$\Sigma_{22} = \sigma_q^2. \tag{A.3}$$

Because the entropy-relaxed optimal transport is given by Eq. (79), Σ is composed of $\tilde{\sigma}^2$ and $\tilde{\sigma}'^2$ as follows:

$$\Sigma_{11} = \frac{\tilde{\sigma}^2\left(2\tilde{\sigma}'^2 + \lambda\right)}{2\left(\tilde{\sigma}^2 + \tilde{\sigma}'^2\right) + \lambda}, \tag{A.4}$$

$$\Sigma_{22} = \frac{\tilde{\sigma}'^2\left(2\tilde{\sigma}^2 + \lambda\right)}{2\left(\tilde{\sigma}^2 + \tilde{\sigma}'^2\right) + \lambda}. \tag{A.5}$$

Solving Eqs. (A.4, A.5) under the conditions given in Eqs. (A.2, A.3), we have

$$\tilde{\sigma}^2 = \left\{ \frac{1}{2\sigma_p^2}\left(1+\sqrt{1+X}\right) - \frac{2}{\lambda} \right\}^{-1}, \tag{A.6}$$

$$\tilde{\sigma}'^2 = \left\{ \frac{1}{2\sigma_q^2}\left(1+\sqrt{1+X}\right) - \frac{2}{\lambda} \right\}^{-1}, \tag{A.7}$$

$$\text{where} \qquad X = \frac{16\sigma_p^2\sigma_q^2}{\lambda^2}. \tag{A.8}$$

Substituting the mean [Eq. (A.1)] and variances [Eqs. (A.6, A.7)] into the definition of the cost [Eq. (23)], after straightforward calculations, we get Eq. (82). In general, the η coordinates of the Gaussian distribution q are given by $(\eta_1, \eta_2) = (\mu_q, \mu_q^2 + \sigma_q^2)$. After differentiating $C_\lambda(p, q)$ with the η coordinates and substituting them into Eq. (52), we get Eq. (83).

References

1. Amari, S.: Information Geometry and Its Applications. Springer, Japan (2016)
2. Chentsov, N.N.: Statistical Decision Rules and Optimal Inference. Nauka (1972) **(translated in English, AMS (1982))**
3. Rao, C.R.: Information and accuracy attainable in the estimation of statistical parameters. Bull. Calcutta Math. Soc. **37**, 81–91 (1945)
4. Ay, N., Jost, J., Le, H.V., Schwachhöfer, L.: Information Geometry. Springer, Cham (2017)
5. Santambrogio, F.: Optimal Transport for Applied Mathematicians. Birkhauser, Basel (2015)
6. Villani, C.: Topics in Optimal Transportation. Graduate Studies in Math. AMS, Providence (2013)
7. Cuturi, M.: Sinkhorn distances: light speed computation of optimal transport. In: Advances in Neural Information Processing Systems, pp. 2292–2300 (2013)
8. Cuturi, M., Avis, D.: Ground metric learning. J. Mach. Learn. Res. **15**, 533–564 (2014)
9. Cuturi, M., Peyré, G.: A smoothed dual formulation for variational Wasserstein problems. SIAM J. Imaging Sci. **9**, 320–343 (2016)
10. Montavon, G., Muller K., Cuturi, M.: Wasserstein training for Boltzmann machines (2015). arXiv:1507.01972v1
11. Belavkin, R.V.: Optimal measures and Markov transition kernels. J. Glob. Optim. **55**, 387–416 (2013)
12. Sinkhorn, R.: A relationship between arbitrary positive matrices and doubly stochastic matrices. Ann. Math. Stat. **35**, 876–879 (1964)
13. Oizumi, M., Tsuchiya, N., Amari, S.: Unified framework for information integration based on information geometry. Proc. Natl. Acad. Sci. **113**, 14817–14822 (2016)
14. Amari, S., Tsuchiya, N., Oizumi, M.: Geometry of information integration (2017). arXiv:1709.02050
15. Oizumi, M., Albantakis, L., Tononi, G.: From the phenomenology to the mechanisms of consciousness: integrated information theory 3.0. PLoS Comput. Biol. **10**, e1003588 (2014)
16. Muzellec, B., Nock, R., Patrini, G., Nielsen, F.: Tsallis regularized optimal transport and ecological inference (2016). arXiv:1609.04495v1
17. Amari, S, Karakida, R. Oizumi, M., Cuturi, M.: New divergence derived from Cuturi function **(in preparation)**