Information Geometry of Wasserstein Divergence

Ryo Karakida[1(✉)] and Shun-ichi Amari[2]

[1] National Institute of Advanced Industrial Science and Technology, Koto-ku, Tokyo 135-0064, Japan
karakida.ryo@aist.go.jp
[2] RIKEN Brain Science Institute, Wako-shi, Saitama 351-0198, Japan
amari@brain.riken.jp

Abstract. There are two geometrical structures in a manifold of probability distributions. One is invariant, based on the Fisher information, and the other is based on the Wasserstein distance of optimal transportation. We propose a unified framework which connects the Wasserstein distance and the Kullback-Leibler (KL) divergence to give a new information-geometrical theory. We consider the discrete case consisting of n elements and study the geometry of the probability simplex S_{n-1}, the set of all probability distributions over n atoms. The Wasserstein distance is introduced in S_{n-1} by the optimal transportation of commodities from distribution $\boldsymbol{p} \in S_{n-1}$ to $\boldsymbol{q} \in S_{n-1}$. We relax the optimal transportation by using entropy, introduced by Cuturi (2013) and show that the entropy-relaxed transportation plan naturally defines the exponential family and the dually flat structure of information geometry. Although the optimal cost does not define a distance function, we introduce a novel divergence function in S_{n-1}, which connects the relaxed Wasserstein distance to the KL-divergence by one parameter.

1 Introduction

Information geometry studies invariant properties of a manifold of probability distributions, which are useful for various applications in statistics, machine learning, signal processing, optimization and others. Two geometrical structures have been introduced from two different backgrounds. One is constructed based on the invariance principle: The geometry is invariant under reversible transformations of random variables. We then have the Fisher information matrix as the unique invariant Riemannian metric (Rao 1945; Chentsov 1982; Amari 2016). Moreover, two dually coupled affine connections are given as invariant connections. These structures are useful for various applications. Another geometrical structure is introduced through the transportation problem. A distribution of commodities in a manifold is transported to another distribution. The transportation with the minimal cost defines a distance between the two distributions, called the Monge-Kantorovich-Wasserstein distance or earth-mover distance. This gives a tool to study the geometry of distributions taking the metric of the supporting manifold into account.

F. Nielsen and F. Barbaresco (Eds.): GSI 2017, LNCS 10589, pp. 119–126, 2017.
https://doi.org/10.1007/978-3-319-68445-1_14

Let $X = \{1, \cdots, n\}$ be the support of a probability measure $\boldsymbol{p}$. The invariant geometry gives a structure which is invariant under permutations of elements of X. It leads to an efficient estimator in statistical estimation. On the other hand, when we consider a picture over n^2 pixels $X = \{(ij); i, j = 1, \cdots, n\}$, neighboring pixels are close. A permutation of X destroys such a neighboring structure, so the invariance should not be required. The Wasserstein distance is responsible for such a structure. Therefore, it is useful for problems having neighboring structure in support X.

An interesting question arises how these two geometrical structures are related. They are useful structures in their own right, but it is intriguing to find a unified framework to include the two. For this purpose in mind, the present paper treats the discrete case over n elements, such that a probability distribution is given by a probability vector $\boldsymbol{p} = (p_1, \cdots, p_n)$ in the probability simplex S_{n-1}, letting a general case of continuous distributions over a manifold to be studied in future.

Cuturi (2013) modified the transportation problem such that the cost is minimized under the entropy constraint. This is called the entropy-relaxed optimal translation problem. In many applications, his group showed the quasi-distance defined by the entropy-constrained optimal solution gives superior properties to the information-geometric distance such as the KL divergence or the Hellinger distance. As an application, consider a set of normalized histograms over X. A clustering problem categorizes them in some classes such that a class consists of similar histograms. Since a histogram is regarded as an empirical probability distribution, the problem is formulated within the probability simplex S_{n-1} in the discrete case and the distances among supporting pixels play a fundamental role.

We follow the entropy-relaxed framework of Cuturi (2013), Cuturi and Avis (2014), Cuturi and Peyré (2016), etc. and introduce a Lagrangian function which is a linear combination of the transportation cost and the entropy. Given distribution $\boldsymbol{p}$ of commodities at the sender and $\boldsymbol{q}$ at the receiver, the optimal transportation plan is the minimizer of the Lagrangian function. We reveal that it is a convex function of $\boldsymbol{p}$ and $\boldsymbol{q}$ so it defines a dually flat geometric structure in $S_{n-1} \times S_{n-1}$. The m-flat coordinates are $(\boldsymbol{p}, \boldsymbol{q})$ and their dual, e-flat coordinates $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ are given from the Lagrangian duality of nonlinear optimization problems. The set of the optimal transportation plans is an exponential family with the canonical parameters $(\boldsymbol{\alpha}, \boldsymbol{\beta})$, where the expectation parameters are $(\boldsymbol{p}, \boldsymbol{q})$. Furthermore, we introduce a novel divergence between $\boldsymbol{p}$ and $\boldsymbol{q}$ in S_{n-1}. It connects the relaxed Wasserstein distance to the KL-divergence by a one parameter family. Our divergence will be expected to be useful for practical applications, because a divergence is a general concept including the square of a distance and more flexible admitting non-symmetricity between $\boldsymbol{p}$ and $\boldsymbol{q}$.

2 Entropy-Constrained Transportation Problem

Let us consider n terminals $X = (X_1, \cdots, X_n)$ at which amounts $p_1, \cdots, p_n$ of commodities are stocked. We transport them within X such that amounts

$q_1, \cdots, q_n$ are newly stored at $X_1, \cdots, X_n$. We normalize the total amount to be equal to 1, so $\boldsymbol{p} = (p_1, \cdots, p_n)$ and $\boldsymbol{q} = (q_1, \cdots, q_n)$ are regarded as probability distributions in the probability simplex S_{n-1},

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

We consider a transportation plan $\mathbf{P} = (P_{ij})$, where P_{ij} is the amount of commodity transported from X_i to X_j. A plan $\mathbf{P}$ is regarded as a joint probability distribution of commodities flowing from X_i to X_j, satisfying the sender and receiver constraints,

$$\sum_j P_{ij} = p_i, \quad \sum_i P_{ij} = q_j. \tag{2}$$

The set of $\mathbf{P}$'s satisfying (2) is denoted by $U(\boldsymbol{p}, \boldsymbol{q})$.

Let $\mathbf{c} = (c_{ij})$ be the cost matrix, where c_{ij} denotes the cost of transporting one unit of commodities from X_i to X_j.

The transportation cost is defined by

$$c(\mathbf{P}) = \langle \mathbf{c}, \mathbf{P} \rangle = \sum c_{ij} P_{ij}. \tag{3}$$

The Wasserstein distance is defined by the minimal cost of transporting distribution $\boldsymbol{p}$ at the senders to $\boldsymbol{q}$ at the receivers,

$$c(\boldsymbol{p}, \boldsymbol{q}) = \min_{\mathbf{P} \subset U(\boldsymbol{p}, \boldsymbol{q})} \langle \mathbf{c}, \mathbf{P} \rangle, \tag{4}$$

where min is taken over all $\mathbf{P}$ satisfying constraints (2). See e.g., Villani (2013).

Given $\boldsymbol{p}$ and $\boldsymbol{q}$, let us consider a special transportation plan $\mathbf{P}_D$ defined by the direct product of $\boldsymbol{p}$ and $\boldsymbol{q}$,

$$\mathbf{P}_D = \boldsymbol{p} \otimes \boldsymbol{q} = (p_i q_j). \tag{5}$$

This plan transports commodities from each sender to the receivers according to the receiver distribution $\boldsymbol{q}$, irrespective of $\mathbf{c}$. The entropy of $\mathbf{P}_D$,

$$H(\mathbf{P}_D) = -\sum \mathrm{P}_{Dij} \log \mathrm{P}_{Dij} = H(\boldsymbol{p}) + H(\boldsymbol{q}), \tag{6}$$

is the minimum among all $\mathbf{P}$'s belonging to $U(\boldsymbol{p}, \boldsymbol{q})$, because of $H(\mathbf{P}) \leq H(\boldsymbol{p}) + H(\boldsymbol{q})$, where $H(\mathbf{P})$, $H(\boldsymbol{p})$ and $H(\boldsymbol{q})$ are the respective entropies and the equality holds for $\mathbf{P} = \mathbf{P}_D$.

We consider a constrained problem of searching for $\mathbf{P}$ that minimizes $\langle \mathbf{c}, \mathbf{P} \rangle$ within a KL-divergence ball centered at $\mathbf{P}_D$,

$$KL[\mathbf{P} : \mathbf{P}_D] \leq d \tag{7}$$

for constant d. As d increases, the entropy of $\mathbf{P}$ increases within the ball. This is equivalent to the entropy constrained problem that minimizes a linear combination of the transportation cost $\langle \mathbf{c}, \mathbf{P} \rangle$ and entropy $H(\mathbf{P})$,

$$F_\lambda(\mathbf{P}) = \frac{1}{1+\lambda} \langle \mathbf{c}, \mathbf{P} \rangle - \frac{\lambda}{1+\lambda} H(\mathbf{P}) \tag{8}$$

for constant λ (Cuturi 2013). Here, λ is a Lagrangian multiplier and λ becomes smaller as d becomes larger.

3 Solution of Entropy-Constrained Problem

Since $\mathbf{P}$ satisfies constraints (2), by using Lagrange multipliers α_i, β_j, minimization of (8) is formulated in the Lagrangian form,

$$L_\lambda(\mathbf{P}) = \frac{1}{1+\lambda}\langle \mathbf{c}, \mathbf{P}\rangle - \frac{\lambda}{1+\lambda}H(\mathbf{P}) - \sum_{i,j}(\alpha_i + \beta_j)\, P_{ij}. \tag{9}$$

Let us fix λ, considering it as a parameter controlling the magnitude of the entropy or the size of the KL-ball. By differentiating (9) with respect to P_{ij}, we have the following solution,

$$P_{ij} = \exp\left\{-\frac{c_{ij}}{\lambda} + \frac{1+\lambda}{\lambda}(\alpha_i + \beta_j) - 1\right\}. \tag{10}$$

Let us put

$$a_i = \exp\left(\frac{1+\lambda}{\lambda}\alpha_i\right) \quad b_j = \exp\left(\frac{1+\lambda}{\lambda}\beta_j\right), \quad K_{ij} = \exp\left\{-\frac{c_{ij}}{\lambda}\right\}, \tag{11}$$

and the optimal solution is written as

$$P^*_{\lambda ij} \propto a_i b_j K_{ij}, \tag{12}$$

where a_i and b_j correspond to the Lagrange multipliers α_i and β_j to be determined from the constraints (2). Note that $2n$ constraints (2) are not independent. Because of $\sum p_i = 1$, we can obtain a_n by $a_n = 1 - \sum_{i\neq n} a_i$. Further, we note that $\mu\boldsymbol{a}$ and $\boldsymbol{b}/\mu$ give the same answer for any $\mu > 0$, where $\boldsymbol{a} = (a_i)$ and $\boldsymbol{b} = (b_j)$. Therefore, the degrees of freedom of $\boldsymbol{a}$ and $\boldsymbol{b}$ are $2(n-1)$, which are to be determined from $\boldsymbol{p}$ and $\boldsymbol{q}$ of which degrees of freedom are also $2(n-1)$. Therefore, we may choose $\boldsymbol{a}$ and $\boldsymbol{b}$ such that they satisfy

$$\sum a_i = 1, \quad \sum b_j = 1. \tag{13}$$

Then, $\boldsymbol{a}, \boldsymbol{b} \in S_{n-1}$ and we have the following theorem.

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

$$P^*_{\lambda ij} = c a_i b_j K_{ij}, \tag{14}$$

$$c = \frac{1}{\sum a_i b_j K_{ij}}, \tag{15}$$

where two vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ are determined from $\boldsymbol{p}$ and $\boldsymbol{q}$.

Cuturi (2013) obtained the above $P^*_{\lambda ij}$ and applied the Sinkhorn-Knopp algorithm to iteratively compute $\boldsymbol{a}$ and $\boldsymbol{b}$.

The following lemma is useful for later calculations.

Lemma 1. The optimal value

$$\varphi_\lambda(\boldsymbol{p}, \boldsymbol{q}) = \min F_\lambda(\mathbf{P}) \tag{16}$$

is given by

$$\varphi_\lambda(\boldsymbol{p}, \boldsymbol{q}) = \frac{\lambda}{1+\lambda}\left(\sum p_i \log a_i + \sum q_j \log b_j + \log c\right). \tag{17}$$

Proof. We first calculate $H\left(\mathbf{P}^*_\lambda\right)$. Substituting (15) in $H\left(\mathbf{P}^*_\lambda\right)$, we have

$$H\left(\mathbf{P}^*_\lambda\right) = -\sum_{ij} \mathrm{P}^*_{\lambda ij}\left(-\frac{c_{ij}}{\lambda} + \log c a_i b_j\right) \tag{18}$$

$$= \frac{1}{\lambda}\langle \mathbf{c}, \mathbf{P}^*_\lambda \rangle - \sum p_i \log a_i - \sum q_j \log b_j - \log c. \tag{19}$$

Hence, (17) follows.

4 Exponential Family of Optimal Transportation Plans

A transportation plan $\mathbf{P}$ is a probability distribution over branches (i, j) connecting terminals i and j. Let x denote branches and $\delta_{ij}(x) = 1$ when x is (i, j) and 0 otherwise. Then $\mathbf{P}$ is a probability distribution of random variable x,

$$P(x) = \sum_{i,j=1}^{n} P_{ij}\delta_{ij}(x). \tag{20}$$

By introducing new parameters

$$\theta^{ij} = \log \frac{P_{ij}}{P_{nn}}, \quad \boldsymbol{\theta} = \left(\theta^{ij}\right), \tag{21}$$

it is rewritten in a parameterized form as

$$P(x, \boldsymbol{\theta}) = \exp\left\{\sum_{i,j} \theta^{ij}\delta_{ij}(x) + \log P_{nn}\right\}. \tag{22}$$

This shows that the set of transportation plans is an exponential family, where θ^{ij} are the canonical parameters and $\eta_{ij} = P_{ij}$ the expectation parameters. They form an $\left(n^2 - 1\right)$-dimensional manifold denoted by S_{TP}, because $\theta^{nn} = 0$.

An optimal transportation plan is specified by $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ in (10), $\boldsymbol{\alpha} = (\alpha_i)$, $\boldsymbol{\beta} = (\beta_j)$ which are determined from $(\boldsymbol{p}, \boldsymbol{q})$. It is written as

$$P(x, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \exp\left[\sum_{i,j}\left\{\frac{\lambda+1}{\lambda}\left(\alpha_i + \beta_j\right) - \frac{c_{ij}}{\lambda}\right\}\delta_{ij}(x) - \frac{\lambda+1}{\lambda}\psi\right], \tag{23}$$

where

$$\psi(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{\lambda}{1+\lambda} \log \sum_{i,j} \exp\left\{ \frac{\lambda+1}{\lambda} (\alpha_i + \beta_j) - \frac{c_{ij}}{\lambda} \right\} \tag{24}$$

is the normalization factor. By putting

$$\theta^{ij} = \frac{1+\lambda}{\lambda} (\alpha_i + \beta_j) - \frac{c_{ij}}{\lambda}, \tag{25}$$

we see that the set S_{OTP} of optimal transformation plans is a submanifold of S_{TP}. Because (25) is linear in $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, S_{OTP} itself is an exponential family, where the canonical parameters are $(1+\lambda)/\lambda$ times $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ and the expectation parameters are $(\boldsymbol{p}, \boldsymbol{q}) \in S_{n-1} \times S_{n-1}$, since

$$E\left[\sum_j \delta_{ij}(x) \right] = p_i, \tag{26}$$

$$E\left[\sum_i \delta_{ij}(x) \right] = q_j. \tag{27}$$

Since each of $\boldsymbol{p}$, $\boldsymbol{q} \in S_{n-1}$ has $n-1$ degrees of freedom, S_{OPT} is a $2(n-1)$-dimensional dually flat manifold. We may put $\alpha_n = \beta_n = 0$ without loss of generality, which correspond to putting $a_n = b_n = 1$ instead of $\sum a_i = \sum b_j = 1$.

We calculate the relaxed cost function $\varphi_\lambda(\boldsymbol{p}, \boldsymbol{q})$ corresponding to $\mathbf{P}(\boldsymbol{\alpha}, \boldsymbol{\beta})$. We then have

$$\varphi_\lambda(\boldsymbol{p}, \boldsymbol{q}) = \frac{1}{1+\lambda} \langle \boldsymbol{c}, \mathbf{P} \rangle + \frac{\lambda}{1+\lambda} \sum_{i,j} P_{ij} \left\{ \frac{1+\lambda}{\lambda} (\alpha_i + \beta_j) - \frac{c_{ij}}{\lambda} - \frac{1+\lambda}{\lambda} \psi_\lambda \right\} \tag{28}$$

$$= \boldsymbol{p} \cdot \boldsymbol{\alpha} + \boldsymbol{q} \cdot \boldsymbol{\beta} - \psi_\lambda(\boldsymbol{\alpha}, \boldsymbol{\beta}). \tag{29}$$

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

$$\psi_\lambda(\boldsymbol{\theta}) + \varphi_\lambda(\boldsymbol{\eta}) = \boldsymbol{\theta} \cdot \boldsymbol{\eta}, \tag{30}$$

which is the Legendre relation between $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$. Thus, we have the following theorem.

Theorem 2. The relaxed cost function φ_λ and the free energy (cumulant generating function) ψ_λ of the exponential family are both convex, connected by the Legendre transformation,

$$\boldsymbol{\theta} = \nabla_{\boldsymbol{\eta}} \varphi_\lambda(\boldsymbol{\eta}), \quad \boldsymbol{\eta} = \nabla_{\boldsymbol{\theta}} \psi_\lambda(\boldsymbol{\theta}), \tag{31}$$

$$\boldsymbol{\alpha} = \nabla_{\boldsymbol{p}} \varphi_\lambda(\boldsymbol{p}, \boldsymbol{q}), \ \boldsymbol{\beta} = \nabla_{\boldsymbol{q}} \varphi_\lambda(\boldsymbol{p}, \boldsymbol{q}), \tag{32}$$

$$\boldsymbol{p} = \nabla_{\boldsymbol{\alpha}} \psi_\lambda(\boldsymbol{\alpha}, \boldsymbol{\beta}), \ \boldsymbol{q} = \nabla_{\boldsymbol{p}} \psi_\lambda(\boldsymbol{\alpha}, \boldsymbol{\beta}). \tag{33}$$

The Riemannian metric $\mathbf{G}_\lambda$ is given to $S_{n-1} \times S_{n-1}$ by

$$\mathbf{G}_\lambda = \nabla_{\boldsymbol{\eta}} \nabla_{\boldsymbol{\eta}} \varphi_\lambda(\boldsymbol{\eta}) \tag{34}$$

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

$$\mathbf{G}_\lambda^{-1} = \nabla_{\boldsymbol{\theta}} \nabla_{\boldsymbol{\theta}} \psi_\lambda(\boldsymbol{\theta}). \tag{35}$$

In addition, we can calculate $\mathbf{G}_\lambda^{-1}$ explicitly from (24).

Theorem 3. The Fisher information matrix $\mathbf{G}_\lambda^{-1}$ in the $\boldsymbol{\theta}$-coordinate system is given by

$$\mathbf{G}_\lambda^{-1} = \frac{1+\lambda}{\lambda} \left\{ \begin{bmatrix} \mathrm{diag}(\boldsymbol{p}) & \mathbf{P} \\ \mathbf{P}^{\mathrm{T}} & \mathrm{diag}(\boldsymbol{q}) \end{bmatrix} - \begin{bmatrix} \boldsymbol{p} \\ \boldsymbol{q} \end{bmatrix} [\boldsymbol{p}^T \ \boldsymbol{q}^T] \right\}, \tag{36}$$

or in the component form as

$$\mathbf{G}_\lambda^{-1} = \frac{1+\lambda}{\lambda} \begin{bmatrix} p_i \delta_{ij} - p_i p_j & P_{ij} - p_i q_j \\ P_{ij} - p_i q_j & q_i \delta_{ij} - q_i q_j \end{bmatrix}. \tag{37}$$

Remark 1. The $\boldsymbol{p}$-part of $\mathbf{G}_\lambda^{-1}$ is a scalar multiple of the Fisher information of $\boldsymbol{p}$ in S_{n-1} in the e-coordinate system. So is the $\boldsymbol{q}$-part. They are independent of the cost matrix c_{ij}, but the off-diagonal blocks of $\mathbf{G}_\lambda^{-1}$ depend on it.

Remark 2. The $\boldsymbol{p}$-part of $\mathbf{G}_\lambda$ is not equal to the Fisher information of $\boldsymbol{p}$ in the m-coordinate system. It is the $\boldsymbol{p}$-part of the inverse of $\mathbf{G}_\lambda^{-1}$, depending on $\boldsymbol{q}$, too.

5 λ-Divergence in S_{n-1}

The relaxed Wasserstein distance $\varphi_\lambda(\boldsymbol{p} : \boldsymbol{q})$ does not satisfy a criterion of divergence, i.e. $\varphi_\lambda(\boldsymbol{p} : \boldsymbol{p}) \neq 0$, because $\varphi_\lambda(\boldsymbol{p} : \boldsymbol{q})$ is minimalized at $\boldsymbol{q} \neq \boldsymbol{p}$ in general. In contrast, the original Wasserstein distance literally satisfies the criteria of distance and those of divergence. To recover the property of divergence in the relaxed form, we introduce a canonical divergence between two transportation plans $(\boldsymbol{p}, \boldsymbol{p})$ and $(\boldsymbol{p}, \boldsymbol{q})$, which is composed of the Legendre pair of the convex functions φ_λ and ψ_λ (Amari 2016):

$$D_\lambda[\boldsymbol{p} : \boldsymbol{q}] = \psi_\lambda(\boldsymbol{\alpha}, \boldsymbol{\beta}) + \varphi_\lambda(\boldsymbol{p}, \boldsymbol{p}) - \boldsymbol{\alpha} \cdot \boldsymbol{p} - \boldsymbol{\beta} \cdot \boldsymbol{p}, \tag{38}$$

where $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ corresponds to $(\boldsymbol{p}, \boldsymbol{q})$. We call this a λ-divergence $D_\lambda[\boldsymbol{p} : \boldsymbol{q}]$ in S_{n-1} from $\boldsymbol{p}$ to $\boldsymbol{q}$. It connects the Wasserstein distance and the KL-divergence in the following way.

This λ-divergence can be transformed into a Bregman-like divergence with the relaxed cost function φ_λ (not a Bregman divergence constructed from a convex function of a single variable $\boldsymbol{q}$):

$$D_\lambda[\boldsymbol{p} : \boldsymbol{q}] = \varphi_\lambda(\boldsymbol{p}, \boldsymbol{p}) - \varphi_\lambda(\boldsymbol{p}, \boldsymbol{q}) - \langle \nabla_{\boldsymbol{q}} \varphi_\lambda(\boldsymbol{p}, \boldsymbol{q}), \boldsymbol{p} - \boldsymbol{q} \rangle. \tag{39}$$

As easily confirmed by substituting (14) to (38), the λ-divergence is equivalent to the KL-divergence between the two transportation plans, up to a constant factor:

$$D_\lambda[\boldsymbol{p} : \boldsymbol{q}] = \frac{\lambda}{1+\lambda} KL\left[\mathbf{P}' : \mathbf{P}\right], \tag{40}$$

where $\mathbf{P}'$ and $\mathbf{P}$ are the optimal plans from $\boldsymbol{p}$ to $\boldsymbol{p}$ and $\boldsymbol{p}$ to $\boldsymbol{q}$, respectively. It is easy to see that $D_\lambda[\boldsymbol{p} : \boldsymbol{q}]$ satisfies the criteria of divergence. However, it is not dually flat in general.

Let us consider the case of $\lambda \to \infty$. Then,

$$\mathbf{P}' = (p_i p_j), \quad \mathbf{P} = (p_i q_j), \tag{41}$$

and hence

$$D_\lambda[\boldsymbol{p} : \boldsymbol{q}] = KL[\boldsymbol{p} : \boldsymbol{q}], \tag{42}$$

converging to the KL-divergence of S_{n-1}.

6 Conclusions

We have opened a new way of studying the geometry of probability distributions. We showed that the entropy-relaxed transportation plan in a probability simplex naturally defines the exponential family and the dually flat structure of information geometry. We also introduced a one-parameter family which connects the relaxed Wasserstein distance to the KL-divergence.

It remains as future problems to extend the information geometry of the relaxed Wasserstain distance into a general case of continuous distributions on a metric manifold. Another direction of research is to study the geometrical properties of the manifold through the new family of λ-divergence and to apply it to various practical applications, where some modifications of D_λ might be useful.

References

Amari, S.: Information Geometry and Its Applications. Springer, Tokyo (2016)
Chentsov, N.N.: Statistical Decision Rules and Optimal Inference. Nauka (1972). Translated in English. AMS (1982)
Cuturi, M.: Sinkhorn distances: light speed computation of optimal transport. In: Advances in Neural Information Processing Systems, pp. 2292–2300 (2013)
Cuturi, M., Avis, D.: Ground metric learning. J. Mach. Learn. Res. **15**, 533–564 (2014)
Cuturi, M., Peyré, G.: A smoothed dual formulation for variational Wasserstein problems. SIAM J. Imaging Sci. **9**, 320–343 (2016)
Rao, C.R.: Information and accuracy attainable in the estimation of statistical parameters. Bull. Calcutta Math. Soc. **37**, 81–91 (1945)
Villani, C.: Topics in Optimal Transportation. Graduate Studies in Math. AMS (2013)