SIMILARITIES BETWEEN INFORMATION GEOMETRY AND EUCLIDEAN GEOMETRY

TOMOHIRO NISHIYAMA

Abstract. Dually flat spaces play a key role in the differential geometrical approach in statistics (information geometry) and many divergences have been studied as an amount which measures the discrepancy between two probability distributions. In a dually flat space, a canonical divergence is naturally introduced and it satisfies a relational expression called triangular relation. This can be regarded as a generalization of the law of cosines in Euclidean space. We introduce a new divergence corresponding to the squared distance in a dually flat space. For this divergence, we show that more relational equations and theorems similar to Euclidean space hold and study the relationship with other divergences.

Keywords: information geometry, parallelogram law, polarization identity, Jeffreys, Jensen-Shannon, divergence.

1. Introduction

Given a probability distribution $p_\theta(x)$, we consider a manifold M with a parameter vector $\boldsymbol{\theta} = (\theta^1, \theta^2, \cdots, \theta^n) \in \mathbb{R}^n$ as a coordinate system. Manifolds with probability distributions as elements are called statistical manifolds, and dually flat space is an important concept of it [1]. In a dually flat space, it is possible to introduce a dual coordinate system $\boldsymbol{\eta} = (\eta_1, \eta_2, \cdots, \eta_n)$ and the convex functions called potential $\psi(\theta)$ and $\phi(\eta)$.

For example, the manifold of the exponential family $p_\theta(x) = \exp(C(x) + \sum_{i=1}^n \theta^i F_i(x) - \psi(\theta))$ is a dually flat space for $\boldsymbol{\theta}$ and $\boldsymbol{\eta} = E[\mathbf{F}(x)]$, where x is a random variable, $F_i(x), C(x)$ are known functions and $E[\cdot]$ indicates expected value. Similarly, the manifold of the mixture family $p_\eta(x) = p_0(x) + \sum_i \eta_i(p_i(x) - p_0(x))$ is a dually flat space, where $p_i(x) (i = 0, 1, 2, \cdots, n)$ are probability distributions.

Given two probability distributions, the Kullback-Leibler divergence(KL-divergence) has long been known as an amount which measures the discrepancy. In a dually flat space, we may define an amount called a canonical divergence, and in exponential families the canonical divergence is consistent with the KL-divergence [1–4]. The canonical divergence satisfies relational expressions called triangular relations for three points P, Q and $R \in M$.This can be regarded as a generalization of the law of cosines in Euclidean space. When a curve PQ and a curve QR are "orthogonal" , the triangular relation becomes the same expression as the Pythagorean theorem [1]. This generalized Pythagorean theorem is an important role in the projection theorem.

The present paper aims at studying the properties similar to Euclidean space in dually flat spaces. We introduce a new divergence called "affine divergence " which can be expressed as an inner product of θ-coordinate and its dual η-coordinate. The affine divergence satisfies triangular relational expression as well as the canonical divergence.

We study the behavior of the affine divergence on θ and η-geodesics and show that the affine divergence satisfies the same summation formula as the squared

Euclidean distance. We further show that the generalized parallelogram law and the generalized polarization identity hold like Euclidean space for the sum of vectors in θ and η-coordinate systems.

Using the sum of vectors of the affine coordinates, we introduce new dual divergences called ψ and ϕ-divergences.These divergences can be represented only by potentials. The ϕ-divergence is consistent with Jensen-Shannon divergence for mixture families. We derive inequalities that holds between the affine divergence and ψ, ϕ-divergence. Table1. shows the summary of the properties of divergences.

Table 1. The summary of the properties of divergences

In this table, the bold fonts indicate the divergences which can be represented by the affine coordinates or potentials. With the canonical divergence as example, the table indicates the canonical divergence is consistent with the KL-divergence for the exponential families and satisfy generalized the law of cosines as the property similar to Euclidean geometry.The blank cells indicate unknown.

	exponential families	mixture families	properties similar to Euclidean geometry
canonical	Kullback-Leibler		law of cosines
affine	Jeffreys		law of cosines summation formula on geodesics parallelogram law polarization identity
phi		skew Jensen-Shannon	

2. Canonical divergence

Let M be a dually flat space, we may define dual affine connections ∇, ∇^* and a Riemannian metric g_{ij}. There exist dual affine coordinate systems θ, η and dual convex functions $\psi(\theta), \phi(\eta)$ on M. An affine coordinate system θ corresponds to the connection ∇, and η corresponds to the connection ∇^*. $\psi(\theta)$ and $\phi(\eta)$ are in a relationship of Legendre transformation with each other.

$$\phi(\eta) = \theta^i \eta_i - \psi(\theta), \tag{1}$$

where we use Einstein notation for $i = (1, 2, \cdots, n)$. We denote $\frac{\partial}{\partial \theta^i}$ as ∂_i and $\frac{\partial}{\partial \eta_i}$ as ∂^i. Then, equations

$$\theta^i = \partial^i \phi \tag{2}$$

$$\eta_i = \partial_i \psi \tag{3}$$

hold. A relationships between the Riemannian metric and affine coordinates are

$$g_{ij} = \partial_i \eta_j \tag{4}$$

$$g^{ij} = \partial^i \theta^j. \tag{5}$$

For two points $P, Q \in M$, we may define a divergence $D(P\|Q)$ called the canonical divergence as follows [1, 2].

$$D(P\|Q) = \psi(\theta(P)) + \phi(\eta(Q)) - \theta^i(P)\eta_i(Q), \tag{6}$$

The canonical divergence is not symmetric with respect to P and Q.

We denote $\psi(\theta(P))$ as $\psi(P)$ and $\phi(\eta(P))$ as $\phi(P)$. For the exponential family $p_\theta(x) = \exp(C(x) + \theta^i F_i(x) - \psi(\theta))$, The Riemannian metric can be expressed as

$$g_{ij} = E[\partial_i l_\theta \partial_j l_\theta] \tag{7}$$

$$l_\theta = \log p_\theta(x). \tag{8}$$

The right hand side of (7) is Fisher information matrix. The canonical divergence is consistent with the KL-divergence. The KL-divergence is

$$D_{KL}(P\|Q) = \sum_i p_{\theta(P)}(x_i) \log \frac{p_{\theta(P)}(x_i)}{p_{\theta(Q)}(x_i)} \tag{9}$$

for discrete distribution, and

$$D_{KL}(P\|Q) = \int p_{\theta(P)}(x) \log \frac{p_{\theta(P)}(x)}{p_{\theta(Q)}(x)} dx \tag{10}$$

for continuous distribution. The canonical divergence have distance-like properties(Property 1. and 2.), and the properties similar to Euclidean space(Property 3. and 4.).

Property 1. $D(P\|Q) \geq 0$

Property 2. $D(P\|Q) = 0 \iff P = Q$

Property 3. Triangular relation

$$D(P\|Q) + D(Q\|R) - D(P\|R) = (\eta_i(R) - \eta_i(Q))(\theta^i(P) - \theta^i(Q)) \tag{11}$$

This formula plays important roles in this paper. When the dual geodesic connecting P and Q is orthogonal at Q to the dual geodesic connecting Q and R, the generalized Pythagorean Theorem holds.

$$D(P\|Q) + D(Q\|R) = D(P\|R) \tag{12}$$

Because coordinates $\boldsymbol{\theta} \equiv \{\theta^i\}$ and $\boldsymbol{\eta} \equiv \{\eta_i\}$ are affine, a curve represented in the form

$$\theta^i(Q(t)) = a^i t + \theta^i(P) \tag{13}$$

is a geodesic for ∇-connection and

$$\eta_i(Q(t)) = a_i t + \eta_i(P) \tag{14}$$

is a geodesic for ∇^*-connection, where $\{a^i\}$ and $\{a_i\}$ are constant vectors in $\mathbb{R}^n$, $t \in \mathbb{R}$ is a parameter along the geodesic.

Propertiy 4. For the point $Q(t)$ on θ-geodesic, $\theta^i(Q(t)) = a^i t + \theta^i(P)$, the following equation holds.

$$D(P||Q(T)) = a^i a^j \int_0^T t g_{ij}(\theta(Q(t)))dt \tag{15}$$

3. Affine divergence

3.1. **Definition of the affine divergence.** In this section we introduce new divergence called "affine divergence" as an inner product of dual affine coordinates.

First, we show the affine divergence satisfies three distance axioms except for the triangle inequality, and the affine divergence satisfies the properties similar to Euclidean space on geodesics.

Second, we prove the generalized parallelogram law and the generalized polarization identity.

Definition 1. A **semimetric** on M is a function $d : X \times X \to \mathbb{R}$ that satisfies

$$d(P, Q) \geq 0 \tag{16}$$

$$d(P, Q) = 0 \iff P = Q \tag{17}$$

$$d(P, Q) = d(Q, P). \tag{18}$$

Definition 2. We define the **affine divergence** $D_A : M \times M \to \mathbb{R}$ as follows.

$$D_A(P, Q) = (\eta_i(Q) - \eta_i(P))(\theta^i(Q) - \theta^i(P)) \tag{19}$$

The affine divergence is an inner product of dual affine coordinates. In a self-dual space($\theta^i = \eta_i$ for all i), the affine divergence is consistent with the squared Euclidean distance $\sum_i (\theta^i(Q) - \theta^i(P))^2$.

Proposition 1. The affine divergence can be expressed as the sum of the canonical divergence.

$$D_A(P, Q) = D(P\|Q) + D(Q\|P) \tag{20}$$

Proof. The result follows by substituting $P = R$ in the triangular relation(11) and using (19).

For the exponential family $p_\theta(x) = \exp(C(x) + \theta^i F_i(x) - \psi(\theta))$, the affine divergence can be expressed as follows.

$$D_A(P, Q) = D_{KL}(P\|Q) + D_{KL}(Q\|P) \tag{21}$$

The right hand side of this equation is the Jeffreys divergence [5, 7].

Proposition 2. The affine divergence satisfies semimetric axioms.
Proof. By using Property1., Property2. of the canonical divergence and the definition of the affine divergence(20), equation (16) and (17) follow. Equation (18) is trivial by (20).

3.2. **Properties along geodesics.** We show that the canonical divergence and the affine divergence monotonically increase along θ and η-geodesics. For three points $P, Q, R \in M$ on the same geodesic, we show that the same relational expression as Euclidean space holds.

Proposition 3. Let $\{a^i\}$ and $\{a_i\}$ be constant vectors in $\mathbb{R}^n$ and t be $t \in \mathbb{R}$. When points $P, Q \in M$ are on θ-geodesic

$$\theta^i(Q(t)) = a^i t + \theta^i(P), \tag{22}$$

the affine divergence can be expressed as

$$D_A(P, Q(T)) = T a^i a^j \int_0^T g_{ij}(\theta(t)) dt. \tag{23}$$

When $P, Q \in M$ are on η-geodesic

$$\eta_i(Q(t)) = a_i t + \eta_i(P), \tag{24}$$

the affine divergence can be expressed as

$$D_A(P, Q(T)) = T a_i a_j \int_0^T g^{ij}(\eta(t)) dt. \tag{25}$$

Because g_{ij} and g^{ij} are positive definite, the affine divergence monotonically increases with T.

Proof. We prove for θ.The same is true of η.
By combining

$$\frac{d\eta_i(\theta(t))}{dt} = a^j \partial_j \eta_i(\theta(t)) \tag{26}$$

and (4) yields

$$\eta_i(Q(T)) - \eta_i(P) = a^j \int_0^T g_{ij}(\theta(t)) dt. \tag{27}$$

By substituting (27) to (19), we have the result. Equation (25) corresponds to Property 4. of the canonical divergence.

Corollary 1. Let $\{a_i\}$ be a constant vector in $\mathbb{R}^n$ and t be $t \in \mathbb{R}$. When points $P, Q \in M$ are on η-geodesic

$$\eta_i(Q(t)) = a_i t + \eta_i(P), \tag{28}$$

the canonical divergence can be expressed

$$D(P\|Q(T)) = a_i a_j \int_0^T dt \int_0^t dt' g^{ij}(\eta(t')). \tag{29}$$

Because g_{ij} is positive definite, the canonical divergence $D(P\|Q(T))$ monotonically increases with T.

Proof. Denoting ∂_Q^i as $\frac{\partial}{\partial \eta_i(Q)}$ and combining the canonical divergence definition(6) and (2) yields

$$\partial_Q^i D(P\|Q(t)) = \theta(Q(t))^i - \theta(P)^i. \tag{30}$$

Hence, we have

$$\frac{dD(P\|Q(t))}{dt} = \frac{1}{t}(\eta(Q(t))_i - \eta(P)_i)(\theta(Q(t))^i - \theta(P)^i) = \frac{1}{t} D_A(P, Q(t)). \tag{31}$$

By substituting (25) to this equation and integrating with respect to t, we have the result.

Lemma 1. Let t be $t \in \mathbb{R}$ and P, R be points on a dually flat space M. For a point $\boldsymbol{\theta}(Q) = (1-t)\boldsymbol{\theta}(P) + t\boldsymbol{\theta}(R)$ on θ-geodesic,

$$D(P\|R) = D(P\|Q) + D(Q\|R) + \frac{t}{1-t} D_A(R, Q) \tag{32}$$

holds. For a point $\boldsymbol{\eta}(Q) = (1-t)\boldsymbol{\eta}(P) + t\boldsymbol{\eta}(R)$ on η-geodesic,

$$D(P\|R) = D(P\|Q) + D(Q\|R) + \frac{1-t}{t} D_A(P, Q) \tag{33}$$

holds.

Proof. We prove for θ-geodesic.The same is true of η-geodesic.
By triangular relation (11), we have

$$D(P\|R) = D(P\|Q) + D(Q\|R) + (\eta_i(R) - \eta_i(Q))(\theta^i(Q) - \theta^i(P)). \tag{34}$$

By assumption,

$$\theta^i(Q) - \theta^i(P) = \frac{t}{1-t}(\theta^i(R) - \theta^i(Q)) \tag{35}$$

holds. By substituting (35) to (34), the result follows.
When $t \in (0, 1)$, taking into account $D_A(R, Q) \geq 0$, we have

$$D(P\|R) \geq D(P\|Q) + D(Q\|R). \tag{36}$$

Corollary 2. Let t be $t \in \mathbb{R}$ and P, R be points on a dually flat space M. For a point $\boldsymbol{\theta}(Q) = (1-t)\boldsymbol{\theta}(P) + t\boldsymbol{\theta}(R)$ on θ-geodesic,

$$D(R\|P) = D(Q\|P) + D(R\|Q) + \frac{1-t}{t} D_A(Q, P) \tag{37}$$

holds.
For a point $\boldsymbol{\eta}(Q) = (1-t)\boldsymbol{\eta}(P) + t\boldsymbol{\eta}(R)$ on η-geodesic,

$$D(R\|P) = D(Q\|P) + D(R\|Q) + \frac{t}{1-t} D_A(R, Q) \tag{38}$$

holds.

Proof. We exchange P and R in (34), we can prove corollary2. in the same way as Lemma1.

Theorem 1. Let t be $t \in \mathbb{R}$ and P, R be points on a dually flat space M. For a point $\boldsymbol{\theta}(Q) = (1-t)\boldsymbol{\theta}(P) + t\boldsymbol{\theta}(R)$ on θ-geodesic or $\boldsymbol{\eta}(Q) = (1-t)\boldsymbol{\eta}(P) + t\boldsymbol{\eta}(R)$ on η-geodesic,

$$D_A(P,R) = \frac{1}{t} D_A(P,Q) + \frac{1}{1-t} D_A(Q,R) \tag{39}$$

holds.

Proof. Taking the sum of (32) and (37), the result follows.
Theorem1. holds for the points P, Q, R on the same straight line and the squared Euclidean distance.

3.3. **Generalized expansion formula for the sum of vectors, parallelogram law, polarization identity.** For the affine divergence, we show that the generalized law of cosines holds as well as the canonical divergence. We consider about the sum of vectors $\boldsymbol{\theta}$ or $\boldsymbol{\eta}$, we show that the generalized expansion formula for the sum of vectors, the parallelogram law and the polarization identity hold for the affine divergence.

Definition 3. We define $\langle , \rangle : M \times M \to \mathbb{R}$ as

$$\langle Q, R \rangle_P \equiv \frac{1}{2}(\theta^i(Q) - \theta^i(P))(\eta_i(R) - \eta_i(P)) + \{Q \leftrightarrow R\}. \tag{40}$$

Symbol $\{Q \leftrightarrow R\}$ means replacement of R and Q.

When $Q = R$,

$$\langle Q, Q \rangle_P = D_A(P,Q) \tag{41}$$

holds.For a self-dual space($\theta^i = \eta_i$ for all i), this is consistent with a dot product.

Corollary 3. (Generalized law of cosines)
For points $P, Q, R \in M$,

$$D_A(P,Q) + D_A(Q,R) - 2\langle P, R \rangle_Q = D_A(P,R) \tag{42}$$

holds.

Proof. By exchanging P and R in triangular relation(11) and taking the sum with original triangular relation (11), we show this corollary. Equation (42) is the generalized law of cosines.

Theorem 2. (Generalized expansion formula for the sum of vectors)
When points $P, Q, R, S \in M$ satisfy $\boldsymbol{\theta}(P) + \boldsymbol{\theta}(R) = \boldsymbol{\theta}(Q) + \boldsymbol{\theta}(S)$ or $\boldsymbol{\eta}(P) + \boldsymbol{\eta}(R) = \boldsymbol{\eta}(Q) + \boldsymbol{\eta}(S)$, the following equations hold.

$$D_A(P,R) = D_A(P,Q) + D_A(P,S) + 2\langle Q, S \rangle_R \tag{43}$$

$$D_A(P,R) = D_A(R,Q) + D_A(R,S) + 2\langle Q, S \rangle_P \tag{44}$$

Proof. We prove for $\boldsymbol{\theta}(P) + \boldsymbol{\theta}(R) = \boldsymbol{\theta}(Q) + \boldsymbol{\theta}(S)$. The same is true of $\boldsymbol{\eta}(P) + \boldsymbol{\eta}(R) = \boldsymbol{\eta}(Q) + \boldsymbol{\eta}(S)$. By the definition of the affine divergence

$$D_A(P,R) = (\eta_i(R) - \eta_i(P))(\theta^i(R) - \theta^i(P)) \tag{45}$$

and the assumption $\boldsymbol{\theta}(P) + \boldsymbol{\theta}(R) = \boldsymbol{\theta}(Q) + \boldsymbol{\theta}(S)$, $D_A(P,R)$ can be expressed as

$$D_A(P,R) = (\eta_i(R) - \eta_i(P))(\theta^i(Q) - \theta^i(P) + \theta^i(S) - \theta^i(P)). \tag{46}$$

Furthermore, the following relational equation holds.

$$\begin{aligned}(\eta_i(R) - \eta_i(P))(\theta^i(Q) - \theta^i(P)) &= (\eta_i(R) - \eta_i(Q))(\theta^i(Q) - \theta^i(P)) + D_A(P,Q) \\ &= (\eta_i(R) - \eta_i(Q))(\theta^i(R) - \theta^i(S)) + D_A(P,Q)\end{aligned} \tag{47}$$

We now use the assumption $\boldsymbol{\theta}(P)+\boldsymbol{\theta}(R)=\boldsymbol{\theta}(Q)+\boldsymbol{\theta}(S)$ again. In the same way, we obtain the following relational equation.

$$\begin{aligned}(\eta_i(R)-\eta_i(P))(\theta^i(S)-\theta^i(P)) &= (\eta_i(R)-\eta_i(S))(\theta^i(S)-\theta^i(P))+D_A(P,S)\\ &= (\eta_i(R)-\eta_i(Q))(\theta^i(R)-\theta^i(Q))+D_A(P,S) \quad (48)\end{aligned}$$

Substituting (47) and (48) to (46), and using (40), we prove that (43) holds. Because the assumption is symmetric with respect to P and R, we exchange P and R in (43) and we prove that (44) holds.
Theorem 2. is a generalization of the expansion formula for the squared norm of the sum of vectors.

$$\|(y-x)+(z-x)\|^2=\|y-x\|^2+\|z-x\|^2+2(y-x,z-x), \quad (49)$$

where $(,)$ is a dot product and $\|\cdot\|$ is a Euclidean norm.

Theorem 3. (Generalized parallelogram law)
When points $P,Q,R,S\in M$ satisfy $\boldsymbol{\theta}(P)+\boldsymbol{\theta}(R)=\boldsymbol{\theta}(Q)+\boldsymbol{\theta}(S)$ or $\boldsymbol{\eta}(P)+\boldsymbol{\eta}(R)=\boldsymbol{\eta}(Q)+\boldsymbol{\eta}(S)$, the following equations holds.

$$D_A(P,Q)+D_A(Q,R)+D_A(R,S)+D_A(S,Q)=D_A(P,R)+D_A(Q,S) \quad (50)$$

The left hand side is the sum of four sides of rectangle $PQRS$ and the right hand side is the sum of diagonal lines of rectangle $PQRS$.
Proof. Applying Corollary3. for points Q,R and S, we have

$$2\langle Q,S\rangle_R=D_A(R,Q)+D_A(R,S)-D_A(Q,S). \quad (51)$$

Substituting (51) to (43), we have the result.
Theorem 3. is a generalization of parallelogram law,

$$2(\|y-x\|^2+\|z-x\|^2)=\|(y-x)+(z-x)\|^2+\|(z-x)-(y-x)\|^2. \quad (52)$$

Corollary 4. (Generalized polarization identity)
When points $P,Q,R,S\in M$ satisfy $\boldsymbol{\theta}(P)+\boldsymbol{\theta}(R)=\boldsymbol{\theta}(Q)+\boldsymbol{\theta}(S)$ or $\boldsymbol{\eta}(P)+\boldsymbol{\eta}(R)=\boldsymbol{\eta}(Q)+\boldsymbol{\eta}(S)$, the following equations holds.

$$2(\langle Q,S\rangle_P+\langle Q,S\rangle_R)=D_A(P,R)-D_A(Q,S). \quad (53)$$

Proof.
By using (44)and (51), the result follows.
Corollary 4 is a generalization of polarization identity,

$$4(y-x,z-x)=\|(y-x)+(z-x)\|^2-\|(z-x)-(y-x)\|^2. \quad (54)$$

Corollary 5. When points $P,Q,R,S\in M$ satisfy $\boldsymbol{\theta}(P)+\boldsymbol{\theta}(R)=\boldsymbol{\theta}(Q)+\boldsymbol{\theta}(S)$ or $\boldsymbol{\eta}(P)+\boldsymbol{\eta}(R)=\boldsymbol{\eta}(Q)+\boldsymbol{\eta}(S)$, the following equations holds.

$$\langle Q,S\rangle_P+\langle Q,S\rangle_R+\langle P,R\rangle_Q+\langle P,R\rangle_S=0 \quad (55)$$

Proof. Exchanging $(P,R)\leftrightarrow(Q,S)$ in (53) and taking the sum with (53), we have the result.
Corollary 5. is a generalization that the sum of adjacent interior angles of a parallelogram in Euclidean space is π.

4. ψ AND ϕ-DIVERGENCES

In this section, we newly introduce dual divergences by using the sum of vectors of affine coordinates. We call these divergence "ψ-divergence" and "ϕ-divergence". The ψ and ϕ-divergences can be expressed by potential functions. We show the ψ and ϕ-divergences are a kind of the skew Jensen divergence [7] and show that the ϕ-divergence is consistent with the (skew) Jensen-Shannon divergence(JS-divergence) for mixture families. Next, we derive inequalities that holds between ψ or ϕ-divergence and the affine divergence, and show that the inequality is consistent with the generalized Lin's inequality for probability distributions of discrete random variable X over $X = \{0, 1, 2 \cdots n\}$. For propbability distributions p and q, Lin's inequality [6] is

$$D_{JS}(p, q) \leq \frac{1}{4} D_J(p, q), \tag{56}$$

where D_{JS} is the JS-divergence

$$D_{JS}(p, q) \equiv \frac{1}{2} D_{KL}(p \| \frac{p+q}{2}) + \frac{1}{2} D_{KL}(q \| \frac{p+q}{2}) \tag{57}$$

and $D_J(p, q) \equiv D_{KL}(p\|q) + D_{KL}(q\|p)$ is the Jeffreys divergence mentioned in section 2.

Proposition 4. Let manifold M be a dual flat space and a, b be $a, b \in \mathbb{R}$. When points $P, Q, R \in M$ satisfy $\boldsymbol{\theta}(R) = a\boldsymbol{\theta}(P) + b\boldsymbol{\theta}(Q)$, the following equations holds.

$$aD(P\|R) + bD(Q\|R) = (a + b - 1)\phi(R) + a\psi(P) + b\psi(Q) - \psi(R) \tag{58}$$

When points $P, Q, R \in M$ satisfy $\boldsymbol{\eta}(R) = a\boldsymbol{\eta}(P) + b\boldsymbol{\eta}(Q)$, the following equations holds.

$$aD(R\|P) + bD(R\|Q) = (a + b - 1)\psi(R) + a\phi(P) + b\phi(Q) - \phi(R) \tag{59}$$

Proof.
We prove for $\boldsymbol{\theta}(R) = a\boldsymbol{\theta}(P) + b\boldsymbol{\theta}(Q)$.The same is true of $\boldsymbol{\theta}(R) = a\boldsymbol{\theta}(P) + b\boldsymbol{\theta}(Q)$. By (6), we have

$$D(P\|R) = \psi(P) + \phi(R) - \theta^i(P)\eta_i(R) \tag{60}$$

$$D(Q\|R) = \psi(Q) + \phi(R) - \theta^i(Q)\eta_i(R). \tag{61}$$

By combining these equations and assumption, we have

$$aD(P\|R) + bD(Q\|R) = a\psi(P) + b\psi(Q) + (a + b)\phi(R) - \theta^i(R)\eta_i(R). \tag{62}$$

By using (1), the result follows. If $a \geq 0$ and $b \geq 0$, (58) and (59) are nonnegative.

Definition 4. Let manifold M be a dual flat space. For points $P, Q, R \in M$ which satisfy $\boldsymbol{\theta}(R) = (1 - \alpha)\boldsymbol{\theta}(P) + \alpha\boldsymbol{\theta}(Q)$ and parameter $\alpha \in (0, 1)$, let us define the α-skew ψ-divergence as

$$D_\psi^{(\alpha)}(P\|Q) \equiv (1 - \alpha)D(P\|R) + \alpha D(Q\|R). \tag{63}$$

For points $P, Q, R \in M$ which satisfy $\boldsymbol{\eta}(R) = (1 - \alpha)\boldsymbol{\eta}(P) + \alpha\boldsymbol{\eta}(Q)$ and parameter $\alpha \in (0, 1)$, let us define the α-skew ϕ-divergence as

$$D_\phi^{(\alpha)}(P\|Q) \equiv (1 - \alpha)D(R\|P) + \alpha D(R\|Q). \tag{64}$$

From the definition and Property 1. and 2. of the canonical divergence, we can easily confirm the α-skew ψ and ϕ-divergence satisfy

$$D_\theta^{(\alpha)}(P\|Q) \geq 0 \tag{65}$$

$$D_\theta^{(\alpha)}(P\|Q) = 0 \iff P = Q \tag{66}$$

$$D_\eta^{(\alpha)}(P\|Q) \geq 0 \tag{67}$$

$$D_\eta^{(\alpha)}(P\|Q) = 0 \iff P = Q. \tag{68}$$

$$\tag{69}$$

Corollary 6. The α-skew ψ and ϕ-divergence can be expressed as follows.

$$D_\psi^{(\alpha)}(P\|Q) = (1 - \alpha)\psi(P) + \alpha\psi(Q) - \psi(R) \tag{70}$$

$$D_\phi^{(\alpha)}(P\|Q) = (1 - \alpha)\phi(P) + \alpha\phi(Q) - \phi(R) \tag{71}$$

Proof.
In Proposition 4., substituting $a = 1 - \alpha$ and $b = \alpha$ and using the definition of the α-skew ψ and ϕ-divergences, the result follows.

Because $\psi(\theta)$ and $\phi(\eta)$ are convex functions, these divergences are a kind of the Jensen divergences [7]. Potentials ψ and ϕ are equal to $\frac{1}{2}\sum_i(\theta^i)^2$ in a self-dual flat space, ψ and ϕ-divergences are equal to $\frac{\alpha(1-\alpha)}{2}\sum_i(\theta^i(Q) - \theta^i(P))^2$ in a self-dual flat space.

Proposition 5. Let $p_i(x)(i = 1, 2, \cdots, n)$ be probability distributions, and let probability distribution $p_\eta(x)$ be the mixture family as follows.

$$p_\eta(x) = p_0(x) + \sum_i \eta_i(p_i(x) - p_0(x)) \tag{72}$$

For the dually flat space M with the above affine coordinate $\boldsymbol{\eta}$, the α-skew ϕ-divergence is consistent with the α-skew JS-divergence as follows.

$$D^{(\alpha)}_\phi(P\|Q) = D^{(\alpha)}_{JS}(p_{\eta(P)}\|p_{\eta(Q)}) \tag{73}$$

$$D^{(\alpha)}_{JS}(p_{\eta(P)}\|p_{\eta(Q)}) = (1-\alpha)D_{KL}(p_{\eta(P)}\|(1-\alpha)p_{\eta(P)} + \alpha p_{\eta(Q)}) + \alpha D_{KL}(p_{\eta(Q)}\|(1-\alpha)p_{\eta(P)} + \alpha p_{\eta(Q)})$$

Proof.
For mixture families, the potential $\phi(\eta)$ equals to $-H(p_\eta)$, where $H(p)$ is entropy of probability distribution p. Hence, the α-skew ϕ-divergence can be expressed as

$$D^{(\alpha)}_\phi(P\|Q) = H(p_{\eta(R)}) - (1-\alpha)H(p_{\eta(P)}) - \alpha H(p_{\eta(Q)}). \tag{74}$$

Furthermore, for the point $R \in M$ which satisfies $\boldsymbol{\eta}(R) = (1-\alpha)\boldsymbol{\eta}(P) + \alpha\boldsymbol{\eta}(Q)$, the following equation holds.

$$\begin{aligned} p_{\eta(R)} &= p_0(x) + \sum_i \eta_i(R)(p_i(x) - p_0(x)) \qquad (75) \\ &= (1-\alpha)(p_0(x) + \sum_i \eta_i(P)(p_i(x) - p_0(x))) + \alpha(p_0(x) + \sum_i \eta_i(Q)(p_i(x) - p_0(x))) \\ &= (1-\alpha)p_{\eta(P)} + \alpha p_{\eta(Q)} \end{aligned}$$

Substituting this equation to (74), we have

$$D^{(\alpha)}_\phi(P\|Q) = H((1-\alpha)p_{\eta(P)} + \alpha p_{\eta(Q)})) - (1-\alpha)H(p_{\eta(P)}) - \alpha H(p_{\eta(Q)}). \tag{76}$$

On the other hand, the α-skew JS-divergence for continuous distribution is

$$\begin{aligned} &(1-\alpha)\int dx p_{\eta(P)} \ln(\frac{p_{\eta(P)}}{(1-\alpha)p_{\eta(P)} + \alpha p_{\eta(Q)}}) + \alpha\int dx p_{\eta(Q)} \ln(\frac{p_{\eta(Q)}}{(1-\alpha)p_{\eta(P)} + \alpha p_{\eta(Q)}}) \qquad (77) \\ &= H((1-\alpha)p_{\eta(P)} + \alpha p_{\eta(Q)})) - (1-\alpha)H(p_{\eta(P)}) - \alpha H(p_{\eta(Q)}). \end{aligned}$$

This equation holds for discrete distribution, too. Comparing this equation with (76), we have the result.

Theorem 4. Let P, Q be the points in a dually flat space M. For the affine divergence and the α-skew ψ or ϕ-divergence, the following inequality holds.

$$\alpha(1-\alpha)D_A(P\|Q) \geq D^{(\alpha)}_\psi(P\|Q) \tag{78}$$

$$\alpha(1-\alpha)D_A(P\|Q) \geq D^{(\alpha)}_\phi(P\|Q) \tag{79}$$

Proof.
Let R be a point which satisfies $\boldsymbol{\theta}(R) = (1-\alpha)\boldsymbol{\theta}(P) + \alpha\boldsymbol{\theta}(Q)$. From (63) and using $D(P\|Q) \leq D_A(P, Q)$, we have

$$D^{(\alpha)}_\psi(P\|Q) = (1-\alpha)D(P\|R) + \alpha D(Q\|R) \leq (1-\alpha)D_A(P, R) + \alpha D_A(R, Q). \tag{80}$$

Using Theorem 1., we have

$$(1-\alpha)D_A(P,R)+\alpha D_A(R,Q)=\alpha(1-\alpha)(\frac{1}{\alpha}D_A(P,R)+\frac{1}{1-\alpha}D_A(R,Q)) \quad (81)$$
$$\leq \alpha(1-\alpha)D_A(P,Q)$$

From this equation, the result follows. The same is true of ϕ-divergence.

Proposition 6. (generalized Lin's inequality)For probability distributions of discrete random variable X over $X=\{0,1,2\cdots n\}$, (79) is consistent with generalized Lin's inequality.

$$\alpha(1-\alpha)D_J(p_{\eta(P)},p_{\eta(Q)})\geq D_{JS}^{(\alpha)}(p_{\eta(P)}\|p_{\eta(Q)}) \quad (82)$$

Proof.
Probability mass function $p(x)$ can be expressed as

$$p(x)=\exp(\sum_{i=1}^{n}\theta^i\delta(x-i)+\ln(p_0)) \quad (83)$$
$$\delta(x)=0(x\neq 0)$$
$$\delta(x)=1(x=0),$$

where $\theta^i=\ln\frac{p_i}{p_0}$ and $p_0=1-\sum_{i=1}^{n}p_i$. Hence, $p(x)$ is an exponential family. Because η_i satisfy $\eta_i=E[\delta(x-x_i)]=p_i$, $p(x)$ can be represented as

$$p(x)=(1-\sum_{i=1}^{n}\eta_i)\delta(x)+\sum_{i=1}^{n}\eta_i\delta(x-i). \quad (84)$$

Hence, $p(x)$ is a mixture family, too. For the mixture family, by Proposition 5., the α-skew ϕ-divergence is consistent with the α-skew JS-divergence. Furthermore, For the exponential family, the affine divergence is consistent with the Jeffreys divergence. By combining these relations with Theorem 4., the result follows. If $\alpha=\frac{1}{2}$, (82) is consistent with Lin's inequality.

5. Examples

In this section, we show concrete examples of the affine divergence.

5.1. **Normal distribution.**

As representative example of continuous distribution, we think 1-dimentional normal distribution.

$$p_\theta(x)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp(-\frac{(x-\mu)^2}{2\sigma^2}) \quad (85)$$

The relations between θ, η and σ,μ are

$$\theta^1=-\frac{1}{2\sigma^2} \quad (86)$$
$$\theta^2=\frac{\mu}{\sigma^2}$$

$$\eta_1=\sigma^2+\mu^2 \quad (87)$$
$$\eta_2=\mu.$$

We calculate the affine divergence of normal distribution as follows.

$$D_A(P,Q)=(\eta_i(Q)-\eta_i(P))(\theta^i(Q)-\theta^i(P))=\frac{({\sigma_Q}^2-{\sigma_P}^2)^2}{2{\sigma_P}^2{\sigma_Q}^2}+(\mu_Q-\mu_P)^2(\frac{1}{2{\sigma_P}^2}+\frac{1}{2{\sigma_Q}^2}) \quad (88)$$

For points $P, Q, R \in M$ which satisfies $\boldsymbol{\theta}(R) = a\boldsymbol{\theta}(P) + b\boldsymbol{\theta}(Q)$, an expected value and a variance are

$$\frac{1}{{\sigma_R}^2} = \frac{a}{{\sigma_P}^2} + \frac{b}{{\sigma_Q}^2}. \tag{89}$$

$$\mu_R = \frac{a{\sigma_Q}^2\mu_P + b{\sigma_P}^2\mu_Q}{a{\sigma_Q}^2 + b{\sigma_P}^2}. \tag{90}$$

Equation(89) means weighted harmonic mean and μ_R is a point to internally divide the straight line $\overline{\mu_P\mu_Q}$ into $b{\sigma_P}^2 : a{\sigma_Q}^2$.
For the point $P, Q, R \in M$ which satisfies $\boldsymbol{\eta}(R) = a\boldsymbol{\eta}(P) + b\boldsymbol{\eta}(Q)$, an expected value and a variance are

$${\sigma_R}^2 = a{\sigma_P}^2 + b{\sigma_Q}^2 + a(1-a)\mu_P^2 + b(1-b)\mu_Q^2 - 2ab\mu_P\mu_Q \tag{91}$$

$$\mu_R = a\mu_P + b\mu_Q. \tag{92}$$

For these quantities, if $a = t, b = 1-t, t \in \mathbb{R}$, Theorem 1. hods. In the case of $a = b = 1$, Theorem 2., the generalized parallelogram law(Theorem 3.) and the generalized polarization identity(Corollary 4.) hold.

5.2. **Binomal distribution.** As representative example of discrete distribution, we think 1-dimentional binomal distribution.

$$p_\theta(k) = \binom{n}{k} p^k (1-p)^{(n-k)}, \tag{93}$$

where n is constant and $p \in [0,1]$. The relations between θ, η and p are

$$\theta = \ln\frac{p}{1-p} \qquad \eta = np. \tag{94}$$

We calculate the affine divergence of binomal distribution as follows.

$$D_A(P,Q) = (\eta(Q) - \eta(P))(\theta(Q) - \theta(P)) = n(p_Q - p_P)\ln\frac{p_Q(1-p_P)}{p_P(1-p_Q)}. \tag{95}$$

For the point $P, Q, R \in M$ which satisfies $\boldsymbol{\theta}(R) = a\boldsymbol{\theta}(P) + b\boldsymbol{\theta}(Q)$, the parameter p is

$$p_R = \frac{\exp(\lambda)}{1+\exp(\lambda)} \qquad \lambda = \frac{p_P^a p_Q^b}{(1-p_P)^a(1-p_Q)^b}. \tag{96}$$

For the point $P, Q, R \in M$ which satisfies $\boldsymbol{\eta}(R) = a\boldsymbol{\eta}(P) + b\boldsymbol{\eta}(Q)$, the parameter p is

$$p_R = ap_P + bp_Q. \tag{97}$$

The above operation is defined only for a, b which satisfies $p_R \in [0,1]$. For these quantities, if $a = t, b = 1-t, t \in \mathbb{R}$, Theorem 1. hods. In the case of $a = b = 1$, Theorem 2., the generalized parallelogram law(Theorem 3.) and the generalized polarization identity(Corollary4.) hold.

6. Conclusion

We have introduced the affine divergence and studied the properties similar to Euclidean space in a dually flat space. We have shown the affine divergence satisfies semimetric axioms and triangular relations as well as the canonical divergence. We also have shown that both the canonical divergence and the affine divergence increase monotonically along geodesics of affine coordinates. As a property peculiar to the affine divergence on geodesics, we have shown that the same summation formula as the squared Euclidean distance hold for three points on the same geodesic.

In addition, for the sum of vectors in affine coordinate systems, we have shown that the generalized expansion formula, the generalized parallelogram law and the generalized polarization identity hold.

The affine divergence can be expressed by the dual affine coordinates. By this analogy, we have introduced ψ and ϕ-divergence which can be expressed by the potentials. We have derived inequalities between the affine and ψ or ϕ-divergence and have shown the relations between the affine, ϕ-divergence and the KL-divergence, the Jeffreys divergence, the Jensen-Shannon divergence.

It is expected that dually flat spaces further have structures similar to Euclidean space and there are more relations between divergences.

References

[1] Shun-ichi Amari. *Information geometry and its applications.* Springer, 2016.
[2] Shun-ichi Amari and Andrzej Cichocki. Information geometry of divergence functions. *Bulletin of the Polish Academy of Sciences: Technical Sciences*, 58(1):183–195, 2010.
[3] Shun-ichi Amari and Hiroshi Nagaoka. *Methods of information geometry*, volume 191. American Mathematical Soc., 2007.
[4] Lev M Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR computational mathematics and mathematical physics*, 7(3):200–217, 1967.
[5] Harold Jeffreys. An invariant form for the prior probability in estimation problems. *Proc. R. Soc. Lond. A*, 186(1007):453–461, 1946.
[6] Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151, 1991.
[7] Frank Nielsen. A family of statistical symmetric divergences based on jensen's inequality. *arXiv preprint arXiv:1009.4004*, 2010.