COMMUN. MATH. SCI.
Vol. 6, No. 4, pp. 915–926

METRICS DEFINED BY BREGMAN DIVERGENCES[*]

PENGWEN CHEN[†], YUNMEI CHEN[‡], AND MURALI RAO[§]

Abstract. Bregman divergences are generalizations of the well known Kullback-Leibler divergence. They are based on convex functions and have recently received great attention. We present a class of "squared root metrics" based on Bregman divergences. They can be regarded as natural generalization of Euclidean distance. We provide necessary and sufficient conditions for a convex function so that the square root of its associated average Bregman divergence is a metric.

Key words. Metrics, Bregman divergence, convexity

AMS subject classifications. 26D10, 94A15

Analysis of noise-corrupted data is difficult without interpreting the data to have been randomly drawn from some unknown distribution with unknown parameters. The most common assumption on noise is Gaussian distribution. However, this assumption may be inappropriate if data is binary-valued, integer-valued or nonnegative. A Gaussian distribution is a member of the exponential family of distributions. Other members of this family are, for example the Poisson and the Bernoulli disctributions, which are better suited for integer and binary data. Exponential families and Bregman divergences (Definition 1.1) have a very intimate relationship. There exists a unique Bregman divergence corresponding to every regular exponential family [13, 3]. More precisely, the log-likelihood of an exponential family distribution can be represented by a sum of a Bregman divergence and a function that does not depend on the distribution parameter. Hence, a Bregman divergence provides a likelihood distance for an exponential family in some sense. This property has been used in generalizing principal component analysis to the exponential family [7].

The Bregman divergence however is not a metric, because it is not symmetric, and does not satisfy the triangle inequality.

Consider the case of Kullback-Leibler divergence (defined in Definition 1.4) [8]. It is not a metric. However, as proved in [11], the square root of the Jensen-Shannon divergence $\frac{1}{2}\big(KL\big(f,\frac{1}{2}(f+g)\big)+KL\big(g,\frac{1}{2}(f+g)\big)\big)$ is a metric. Moreover, it is always finite for any two densities. In fact, the Jensen-Shannon divergence is nothing but an averaged Bregman divergence associated with the convex function $x\log x$. It is very natural to ask whether square roots of other averaged Bregman divergences also are metric, which is the main motivation of this work. We will provide a sufficient and necessary condition on the associated convex function, such that the square root of the corresponding averaged Bregman divergence is a metric. Clearly the justification of the triangle inequality is the only nontrivial part.

One of the most critical properties of a metric is the triangle inequality, which ensures that if both a, b and b, c are "close", so are a, c. This property has many applications. For instance, an important task in pattern recognition is finding the nearest neighbor in a multidimensional vector space. One efficient method of finding

[*] Received: January 22, 2008; accepted: September 15, 2008. Communicated by Shi Jin.

[†] Department of Mathematics, University of Connecticut, Storrs, CT 06269, USA (pengwen@math.uconn.edu).

[‡] Department of Mathematics, University of Florida, Gainesville, FL 32611, USA (yun@math.ufl.edu). Supported by NIH R01 NS052831-01 A1.

[§] Department of Mathematics, University of Florida, Gainesville, FL 32611, USA (rao@math.ufl.edu).

nearest neighbors is through the construction of a so-called metric tree. Given a metric space with N objects we can arrange them into a metric tree with height $\approx \log_2 N$. The triangle inequality then saves a lot of effort in finding the nearest neighbor. The total number of distance computations is reduced from N to merely $\log_2 N$. In this work we provide a large class of metrics for construction of metric trees.

In this paper, our main contributions are summarized briefly as follows: divergence to the power α can be a metric, only if α is at most a half. In some sense, the power $\frac{1}{2}$ is critical. We provide a necessary and sufficient condition characterizing Bregman divergences which, when averaged and taken to the power one half, are metrics.

1. Preliminaries and notations

In this paper, we adopt the following notation:

$$\Omega : \text{the interior domain of the strictly convex function } F, \text{ i.e. } \{x : |F(x)| < \infty\}. \quad (1.1)$$

Usually, Ω is the whole real line or positive half line.

Definition 1.1 (Bregman divergence). *Bregman divergence is defined as $B_F(x,y) := F(x) - F(y) - (x-y)F'(y)$, for any strictly convex function F. For the sake of simplicity, we now assume that all the convex functions considered are smooth, i.e., in C^∞. We will discuss this restriction in Sec. 2.5. For some properties of Bregman divergences, we refer interested readers to [1, 5]. $B_F(x,y)$ is in general is not symmetric. It can be symmetrized in many ways. However the following procedure will be found to be highly rewarding.*

Given x,y define

$$m_F(x,y) := min_z \frac{1}{2}(B_F(x,z) + B_F(y,z)). \quad (1.2)$$

The Lemma below is known. We provide the proof for convenience.

Lemma 1.2.

$$B_F(x,z) \geq 0, \textit{and equality holds if and only if } x = z. \quad (1.3)$$

For all z, $0 < p < 1$ and $q = 1-p$ we have

$$pB_F(x,z) + qB_F(y,z) \geq pB_F(x, px+qy) + qB_F(y, px+qy). \quad (1.4)$$

In particular
$m_F(x,y) = \frac{1}{2}(F(x)+F(y)) - F(\frac{1}{2}(x+y))$*, and* $m_F(x,y) \geq 0$ *with equality iff* $x = y$.

Proof. Now

$$B_F(x,z) = \frac{1}{2}(x-z)^2 F''(\xi), \text{ for some } \xi \in [x,z]. \quad (1.5)$$

The function F is strictly convex, $F''(\xi) > 0$, thus we have $B_F(x,z) \geq 0$, and equality holds only when $z = x$.

For the second statement, since F is convex, and $pB_F(x,px+qy)+qB_F(y,px+qy)=pF(x)+qF(y)-F(px+qy)$, we have

$$\begin{aligned}&pB_F(x,z)+qB_F(y,z)-(pB_F(x,px+qy)+qB_F(y,px+qy))\\ &=F(px+qy)-F(z)-(px+qy-z)F'(z)=B_F(px+qy,z)\geq 0.\end{aligned}$$

Thus, $pB_F(x,z)+qB_F(y,z)\geq pB_F(x,px+qy)+qB_F(y,px+qy)$. □

If $F(x)=x^2$, then $B_F(x,y)=(x-y)^2$, and $\sqrt{m_F(x,y)}$, is a metric. But for an arbitrary convex function, this square root function, $\sqrt{m_F(x,y)}$, is not necessarily a metric (see Remark 2.1). Our goal is to discuss the conditions on the convex function such that $\sqrt{m_F(x,y)}$ is a metric.

1.1. Three important Bregman divergences. Bregman divergences corresponding to the three strictly convex functions x^2, $x\log x$, and $-\log x$ satisfy the homogeneity condition:

$$B_F(kx,ky)=k^\alpha B_F(x,y) \tag{1.6}$$

for all $x,y,k>0$ with α equal to 2,1 and 0 respectively.

In fact, they are the only ones modulo affine additions with this property among all Bregman divergences.

It is easily seen that the Bregman divergence associated with a convex function is not affected by the addition of an affine function to that convex function.

Lemma 1.3. *If the Bregman divergence $B_F(x,y)$ satisfies the homogeneity condition*

$$\textit{with } \alpha=\begin{cases}2,\\1,\\0,\end{cases} \textit{then } F=\begin{cases}x^2,\\x\log x,\\-\log x.\end{cases}$$

The statements hold modulo affine functions.

Proof. Let $B_F(x,y)$ be of order α. Then

$$k^\alpha(F(x)-F(y)-(x-y)F'(y))=F(kx)-F(ky)-k(x-y)F'(ky). \tag{1.7}$$

Differentiating with respect to x twice, we have

$$k^\alpha F''(x)=k^2F''(kx). \tag{1.8}$$

Let $x=1$, then

$$k^{\alpha-2}F''(1)=F''(k). \tag{1.9}$$

Now, if $\alpha=1$, by integrating twice we have

$$F(k)=c_1k\log k+c_2k+c_3,\ B_F(x,y)=c_1(x\log\frac{x}{y}-(x-y)); \tag{1.10}$$

if $\alpha=0$, then

$$F(k)=-c_1\log k+c_2k+c_3,\ B_F(x,y)=c_1\left(-\log\frac{x}{y}-1+\frac{x}{y}\right); \tag{1.11}$$

if $\alpha \neq 0,1$, then

$$F(k) = \frac{c_1}{\alpha(\alpha-1)} k^\alpha + c_2 k + c_3,\ B_F(x,y) = \frac{c_1}{\alpha(\alpha-1)}(x^\alpha - \alpha x y^{\alpha-1} + (\alpha-1) y^\alpha). \quad (1.12)$$

(Here c_1, c_2, c_3 are some constants). □

These divergences can be generalized from the scalar case to the vector case as follows. In the vector case, the Bregman divergence with $F = x\log x$ is called I-divergence, and this Bregman divergence with $F = -\log x$ is called IS divergence.

Definition 1.4. *Given two non-negative vectors $x := (x_1, ..., x_n), y := (y_1, ..., y_n) \in \mathbf{R}^n$, the I-divergence is defined as*

$$CKL(x,y) := \sum_{i=1}^n x_i \log \frac{x_i}{y_i} - \sum_{i=1}^n x_i + \sum_{i=1}^n y_i. \quad (1.13)$$

The Itakura-Saito divergence [4, 14] is defined as

$$IS(x,y) := \sum_{i=1}^n -\log \frac{x_i}{y_i} + \sum_{i=1}^n \frac{x_i}{y_i} - 1. \quad (1.14)$$

The I-divergence reduces to the well known Kullback-Leibler divergence:

$$KL(x,y) := \sum_{i=1}^n x_i \log \frac{x_i}{y_i},$$

if $\sum_{i=1}^n x_i = \sum_{i=1}^n y_i = 1$.

1.2. Bregman divergences v.s. exponential families. Here, we indicate the implications of Bregman divergences in the theory of exponential families.

Definition 1.5 (Exponential family). *([6]) Let ν be a $\sigma-$finite measure on the Borel subsets of $\mathbf{R}^n$, and be absolutely continuous relative to Lebesgue measure (dx). Define the set of natural parameters by $\mathbb{N} = \{\theta \in \mathbf{R}^n : \int \exp(\theta \cdot x)\nu(dx) < \infty\}$. For a natural parameter θ let $\lambda(\theta) = \int \exp(\theta \cdot x)\nu(dx)$, and let $\psi(\theta) = \log \lambda(\theta)$ be the cumulant generating function.*

Define $p_\theta(x) = \exp(\theta \cdot x - \psi(\theta)), \theta \in \mathbb{N}$. This family of probability measures is called an exponential family. Finally let $p_0(x) := \frac{d\nu}{dx}$.

Denote the expectation parameter by $\bar{x}(\theta) := \int x p_\theta(x) dx$. It can be shown that $\mathbb{N}$ is a convex set and $\psi(\theta)$ is a convex function.

Now let F be the conjugate function of ψ, i.e. $F(\bar{x}) = \sup_{\theta \in \mathbb{N}} \{\bar{x} \cdot \theta - \psi(\theta)\}$. It can be shown that $\bar{x}(\theta) = \nabla \psi(\theta)$, $\theta(\bar{x}) = \nabla F(\bar{x})$, and $F(\bar{x}) = \bar{x} \cdot \theta - \psi(\theta)$. Based on these relations, we have

$$x \cdot \theta - \psi(\theta) = F(\bar{x}) + (x - \bar{x}) \cdot F'(\bar{x}) = -B_F(x, \bar{x}) + F(x). \quad (1.15)$$

Hence,

$$-\log p_\theta(x) = -\log p_0(x) + \theta \cdot x - \psi(\theta) = -\log p_0(x) + B_F(x, \bar{x}) - F(x). \quad (1.16)$$

Thus the negative log-likelihood can always be written as a Bregman divergence plus a term that is constant with respect to θ and which therefore can be ignored [7, 5, 1].

The following consideration is relevant and of interest: Consider two observed events $x_1,x_2\in dom(F)$. Considering each $x_i, i=1,2$ we maximize the estimators θ_i separately. In dual form, we have that $\bar{x}=x_i$ minimizes $-\log p_\theta(x_i)$. Now suppose these two events happen independently. Then, finding a single estimator $\theta=\theta_0$ to maximize the likelihood $p_\theta(x_1)p_\theta(x_2)$ is equivalent to finding a $\bar{x}$ to minimize $B_F(x_1,\bar{x})+B_F(x_2,\bar{x})$. Clearly, $\bar{x}=\frac{1}{2}(x_1+x_2)$ is the minimizer, and the likelihood ratio is given by $\frac{p_{\theta_1}(x_1)p_{\theta_2}(x_2)}{p_{\theta_0}(x_1)p_{\theta_0}(x_2)}$. This ratio is always greater than or equal to 1. In fact our function $m_F(x_1,x_2)$ is a half of the log-likelihood-ratio,

$$\frac{1}{2}\left(\log\frac{p_{\theta_1}(x_1)p_{\theta_2}(x_2)}{p_{\theta_0}(x_1)p_{\theta_0}(x_2)}\right)=\frac{1}{2}(B_F(x_1,\frac{1}{2}(x_1+x_2))+B_F(x_2,\frac{1}{2}(x_1+x_2)))=m_F(x_1,x_2). \tag{1.17}$$

As shown in paper [5], the associated function F for Gaussian distributions, Poison distributions, Bernoulli distributions, exponential distributions, and multinomial distributions can be determined to be among Euclidean distance, Kullback-Leibler divergence, and Itakura-Saito distance. In this paper, we will show that those $\sqrt{m_F}$ are metrics. Thus, these log-likelihood-ratios are squared metrics.

2. Square root metrics

2.1. The critical power $1/2$. In the case of Euclidean distance, $F(x)=x^2$, $m_F(x,y)$ itself is not a metric, but its square root is a metric. Therefore, we will examine the necessary condition that $(m_F(x,y))^r$ is a metric in the next lemma.

Lemma 2.1. *Suppose $F(x)$ is a strictly convex, smooth (at least four times differentiable) function on an open set Ω which will either be the entire line or a half line.*

Denote $m(p;x,y):=pF(x)+qF(y)-F(px+qy), p+q=1, 0<p<1$, $x,y\in\Omega$. Then we have the following facts.

1. *$m(p;x,y)\geq 0$, equality holds if and only if $x=y$.*
2. *Monotonicity: If $x<y<z$, $x,y,z\in\Omega$, then $m(p;x,y)<m(p;x,z), m(p;y,z)<m(p;x,z)$. In particular, $\sqrt{m(p;x,y)}<\sqrt{m(p;x,z)}+\sqrt{m(p;z,y)}$, $\sqrt{m(p;y,z)}<\sqrt{m(p;y,x)}+\sqrt{m(p;x,z)}$.*
3. *Suppose we know the triangle inequality holds for some positive r_0:*

$$m(p;x,y)^{r_0}\leq m(p;x,z)^{r_0}+m(p;z,x)^{r_0}, \tag{2.1}$$

then the triangle inequality still holds for any r with $0\leq r<r_0$:

$$m(p;x,y)^r\leq m(p;x,z)^r+m(p;z,x)^r. \tag{2.2}$$

4. *$1/2$ is the maximum possible value of r: if there exists a small neighborhood $(0,\epsilon)$, such that*

$$m(p;x-a,x+a)^r\leq m(p;x-a,x)^r+m(p;x,x+a)^r \text{ holds for all } a\in(0,\epsilon), \tag{2.3}$$

then we must have $\frac{1}{2}\geq r\geq 0$.

Proof. (1): It follows by the convexity of the function F.

(2): Without loss of generality, assume $z > y > x$. Then

$$m(p;x,z) - m(p;x,y) = \int_y^z \frac{\partial m(p;x,\xi)}{\partial \xi} d\xi = \int_y^z qF'(\xi) - qF'(px+q\xi) d\xi > 0. \tag{2.4}$$

Here $qF'(\xi) - qF'(px+q\xi) = pq(\xi - x)F''(\zeta) > 0$, for some $\zeta \in [px+q\xi, \xi]$. The other inequality follows in a similar fashion.

(3) : Based on the result in (2), the only nontrivial case is $x < z < y$. By assumption $0 \leq r < r_0$, and $m(p;x,z)^{r_0} + m(p;z,y)^{r_0} \geq m(p;x,y)^{r_0}$. Now we use the simple fact: if a, b, c are positive and satisfy $a+b > c$ then for each $0 < r < 1$, $a^r + b^r > c^r$.

(4): Suppose that the triangle inequality holds for the points $x-a, x, x+a$ for all positive numbers a in some neiborhood of zero. Using Taylor's expansion about x, we estimate:

$$\begin{aligned} m(p;x-a,x+a) &= 2pqF''(x)a^2 + O(a^3) \\ m(p;x-a,x) &= \frac{pq}{2}F''(x)a^2 + O(a^3) \\ m(p;x,x+a) &= \frac{pq}{2}F''(x)a^2 + O(a^3), \end{aligned}$$

Using this and the triangle inequality:

$$m(p;x-a,x)^r + m(p;x,x+a)^r - m(p;x-a,x+a)^r \tag{2.5}$$

$$= (pqF''(x)a^2)^r (2(2)^{-r} - 2^r) + O(a^{2r+1}) \geq 0. \tag{2.6}$$

The validity of this for all small ϵ leads to $2 \geq 2^{2r}$. □

Our goal is to show that $\sqrt{m_F(x,y)}$ is a metric. Recall $m_F(x,y) := \frac{1}{2}B_F(x, \frac{1}{2}(x+y)) + \frac{1}{2}B_F(y, \frac{1}{2}(x+y))$. This is not true for arbitrary strictly convex functions F as will be seen later. Nonnegativity and symmetry properties are clear. According to the previous lemma, no exponent larger than 1/2 enables the triangle inequality. In the following, we will show that the necessary and sufficient condition that $\sqrt{m_F(x,y)}$ is a metric is $(\log F'')'' \geq 0$.

As noted before the only case of interest is $a < b < c$. So our proof will focus on this case.

2.2. Necessary condition($\mathbb{F} \subset \mathbb{F}'$).

Recall $B_F(x,y) := F(x) - F(y) - (x-y)F'(y)$, and $m_F(x,y) := \frac{1}{2}(B_F(x, \frac{1}{2}(x+y)) + B_F(y, \frac{1}{2}(x+y))) = \frac{1}{2}(F(x)+F(y)) - F(\frac{x+y}{2})$.

Definition 2.2. *We say that F is in a class $\mathbb{F}$ if $\sqrt{m_F}$ is a metric.*

The set $\mathbb{F}$ is non-empty since it contains the function x^2. In this section we show that this set in fact contains the set $\{F : (\log F'')'' \geq 0\}$.

Here are several properties of $\mathbb{F}$.

Lemma 2.3.

1. *$B_{c_1F(x)+c_2x+c_3}(\cdot,\cdot) = c_1 B_F(\cdot,\cdot)$. So also $m_{c_1F(x)+c_2x+c_3}(\cdot,\cdot) = m_F(\cdot,\cdot)$.*
2. *$F(x) \in \mathbb{F}$ if and only if $c_1F(x) + c_2x + c_3 \in \mathbb{F}$.*
3. *$e^x \in \mathbb{F}$ and also $x^2 \in \mathbb{F}$.*

$F(x)$	$(\log F'')$	$(\log F'')''$	Ω
e^x	x	0	$\mathbf{R}$
$x^\alpha/(\alpha(\alpha-1)), 2\geq\alpha>1,$	$(\alpha-2)\log x$	$-(\alpha-2)/x^2$	$\mathbf{R}$
$x\log x - x$	$-\log x$	$1/x^2$	$(0,\infty)$
$-\log x$	$-2\log x$	$2/x^2$	$(0,\infty)$
$1/(2x)$	$-3\log x$	$3/x^2$	$(0,\infty)$

TABLE 2.1. *several examples of F with* $(\log F'')'' \geq 0$

Proof. The first two statements are trivial. If $F(x) := \exp(x)$, $m_F(x,y) = \frac{1}{2}(e^x + e^y) - e^{\frac{x+y}{2}} = \frac{1}{2}(e^{\frac{x}{2}} - e^{\frac{y}{2}})^2$, so that its square root is a metric.

When $F(x) := x^2$, $B_F(x,y) = (x-y)^2$, then $m_F(x,y) = (\frac{x-y}{2})^2$. So x^2, e^x both are in $\mathbb{F}$. □

The next theorem is the necessary condition. We do the proof in the appendix due to its lengthy algebraic manipulations.

THEOREM 2.4 (Necessary condition for $F \in \mathbb{F}$). *$F''F'''' \geq (F''')^2$, i.e., $(\log F'')'' \geq 0$ is a necessary condition that $\sqrt{m_F}$ is a metric.*

Note that the following statement is not always true: for any strictly convex function F, we can find an $r > 0$, such that m_F^r is a metric. We provide a simple example here.

REMARK 2.1 (A counterexample). Let $F(x) := \sqrt{1+x^2}$, $\Omega := \mathbf{R}$ then there exists no positive exponent such that m_F^r is a metric. The reasoning is as follows.

Consider three numbers $-a, 0, a$, with $a > 0$. We have

$$m_F(-a,a) = \sqrt{1+a^2} - 1, m_F(0,a) = m_F(0,-a) = \frac{1}{2}(\sqrt{1+a^2}+1) - \sqrt{1+\frac{a^2}{4}}. \quad (2.7)$$

To ensure that the triangle inequality $m_F(a,0)^r + m_F(-a,0) \geq m_F(a,-a)^r$ holds, we need $2(m_F(a,0)/m_F(a,-a))^r \leq 1$. But no $r > 0$ can satisfy this because

$$\frac{m_F(0,a)}{m_F(-a,a)} = \frac{\frac{1}{2} + \frac{1}{4a} - \frac{1}{a} + O(\frac{1}{a^2})}{a + \frac{1}{2a} - 1 + O(\frac{1}{a^2})} \approx \frac{1}{2a} \to 0, a \to \infty. \quad (2.8)$$

Hence as $a \to \infty$, the set of possibilities for r approaches the sole number 0.

Note that this function F does not satisfy the condition $(\log F'')'' \geq 0$.

REMARK 2.2. In Table 2.1, we list several examples of F with $(\log F'')'' \geq 0$. Note that if a strictly convex function F satisfies $(\log F'')'' = 0$, then we have $\log F''(x) = c_1 x + c_2$. Therefore either $F(x) = e^{c_1 x + c_2}/c_1^2 + c_3 x + c_4$, with $c_1 \neq 0$, or $F(x) = c_2 x^2 + c_3 x + c_4$, with $c_i, i = 1, ..., 4$ some scalars. As shown in Lemma 2.3, these functions belong to $\mathbb{F}$, and for any numbers $a < b < c$, we have $\sqrt{m_F(a,b)} + \sqrt{m_F(b,c)} = \sqrt{m_F(a,c)}$.

DEFINITION 2.5. *Denote the class of functions $\{F : F'' > 0, (\log F'')'' \geq 0\}$ by $\mathbb{F}'$. We will also need the class $\mathbb{G}$ of functions $\{G : G'' > 0, (\log G'')'' = 0\}$. We have shown above that*

$$\mathbb{G} = \{G : G(x) = e^{c_1 x + c_2} + c_3 x + c_4, c_1 \neq 0\} \cup \{G : G(x) = c_2 x^2 + c_3 x + c_4, c_2 > 0\}, \quad (2.9)$$

where c_1, c_2, c_3, c_4 are some scalars.

Note that for any function $g \in \mathbb{G}$, and $a \le b \le c$, we have

$$\sqrt{m_G(a,b)} + \sqrt{m_G(b,c)} = \sqrt{m_G(a,c)}. \tag{2.10}$$

In the appendix, we will show that the set $\mathbb{G}$ is the same as the set $\{G : \sqrt{m_G(a,b)} + \sqrt{m_G(b,c)} = \sqrt{m_G(a,c)}$, for all numbers $a < b < c$ in $\Omega\}$ Intuitively, the set $\mathbb{G}$ is part of the 'boundary' of the set $\mathbb{F}'$, and we have 'triangle equality' on the set $\mathbb{G}$.

We will point out one important relation between the set $\mathbb{F}'$ and the set $\mathbb{G}$ in the next two lemmas.

Lemma 2.6. *Consider any $G \in \mathbb{G}$, and any $F \in \mathbb{F}'$. Then $H = F - G$ vanishes at most at 4 points or it vanishes identically on a segment and is positive outside this segment. If H vanishes at the 4 adjacent points $\{a_1 < \dots < a_4\}$ and nowhere else then H is positive and convex outside $[a_1, a_4]$.*

Proof. If $H = 0$ at five points then H' vanishes at least at 4 points, and $H'' = F'' - G''$ vanishes at least at 3 points. This implies that $\log(F''/G'')$, vanishes at 3 points. Since $\log(F''/G'')$ is convex, the vanishing of this function at more than two points implies that it is zero in an interval and positive outside. Then $H'' = F'' - G''$ also vanishes in an interval and is positive ouside implying in particular that H is also convex. Since it vanishes at 5 points it must vanish in an interval and be positive outside.

The proof is similar for the next statement. Indeed, if H vanishes at the 4 adjacent points $\{a_1 < \dots < a_4\}$ then H'' vanishes at two points in the interval $[a_1, a_4]$. Then also $\log(F''/G'')$ vanishes at two points strictly inside the interval $[a_1, a_4]$. Again, because $\log(F''/G'')$ is convex, there are points x, y such that $a_1 < x < y < a_4$ and $\log(F''/G'')$ is strictly positive outside the interval $[x, y]$. As before we conclude $H'' > 0$ and hence H is convex outside the interval $[x, y]$. Since H vanishes at a_1 and a_4 it must be strictly positive outside the interval $[a_1, a_4]$. This completes the proof. □

The above lemma is optimal in a sense. In fact, given any function $F \in \mathbb{F}'$, and any 4 points, $a_k, k = 1, \dots, 4$, there exists a function $G \in \mathbb{G}$ which agrees with F exactly at these 4 points.

The next lemma gives the proof.

Lemma 2.7. *Let 4 points $x_1 < x_2 < x_3 < x_4$ and $F \in \mathbb{F}'$ be given. Then there are scalars $c_1 \neq 0, c_2, c_3, c_4$ such that one of the functions $e^{c_1 x + c_2} + c_3 x + c_4$ (the exponential case) or $c_2 x^2 + c_3 x + c_4$ (the quadratic case) agrees with F at exactly these 4 points.*

Proof. Let $y_k = F(x_k), k = 1, \dots, 4$. Assuming that the assertion holds, we must have

$$y_{k+1} - y_k = F(x_{k+1}) - F(x_k) = e^{c_1 x_{k+1} + c_2} - e^{c_1 x_k + c_2} + c_3(x_{k+1} - x_k), k = 1, 2, 3. \tag{2.11}$$

So, for $k = 1, 2$ we also have

$$\frac{y_{k+2} - y_{k+1}}{x_{k+2} - x_{k+1}} - \frac{y_{k+1} - y_k}{x_{k+1} - x_k} = e^{c_2} \left(\frac{e^{c_1 x_{k+2}} - e^{c_1 x_{k+1}}}{x_{k+2} - x_{k+1}} - \frac{e^{c_1 x_{k+1}} - e^{c_1 x_k}}{x_{k+1} - x_k} \right). \tag{2.12}$$

For the sake of notational simplicity, put $r_{k,k+1} := \frac{y_{k+1} - y_k}{x_{k+1} - x_k}$. Then we have

$$\frac{r_{3,4} - r_{2,3}}{r_{2,3} - r_{1,2}} = \frac{(e^{c_1 x_4} - e^{c_1 x_3})/(x_4 - x_3) - (e^{c_1 x_3} - e^{c_1 x_2})/(x_3 - x_2)}{(e^{c_1 x_3} - e^{c_1 x_2})/(x_3 - x_2) - (e^{c_1 x_2} - e^{c_1 x_1})/(x_2 - x_1)}. \tag{2.13}$$

The numerators and denominators on both sides of the above equation are positive because the functions involved are strictly convex. The right side approaches ∞ and 0 as c_1 tends to ∞ and $-\infty$ respectively. The existence of c_1 is guaranteed by the Intermediate Value Theorem. Only if c_1 is not equal to zero we can trace the steps backwards to solve for c_2, c_3, c_4 and get the exponential case. If $c_1 = 0$, however c_2 cannot be determined and the exponential function is actually linear. Using L'Hôpital's rule on equation (2.13) (twice with $c_1 \to 0$), we can see that c_1 cannot be zero unless $(r_{3,4} - r_{2,3})/(x_4 - x_2) = (r_{2,3} - r_{1,2})/(x_3 - x_1)$. If the above equation holds, follow the same steps as above replacing the exponential by the quadratic to see that the quadratic case holds. □

2.3. Sufficient condition($\mathbb{F}' \subset \mathbb{F}$). The triangle inequality, for $\sqrt{m_F}$ involves the function values of F at 6 points including 4 "interior points" $((a+b)/2)$, $((b+c)/2)$, $((a+c)/2)$, b, and two end points a, c. According to Lemma 2.7, we know that there is a function $G \in \mathbb{G}$ whose function values at these 4 interior points agree with those of F and at the two end points are smaller than those of F.

Now, we are ready to prove the triangle inequality.

LEMMA 2.8. *Let $a < b < c$, and consider two convex functions F, G, satisfying $F(x) = G(x)$ at $x = b, (a+b)/2, (b+c)/2, (a+c)/2$, and $F(x) \geq G(x)$ at $x = a, c$. If $\sqrt{m_G(a,b)} + \sqrt{m_G(b,c)} \geq \sqrt{m_G(a,c)}$, then $\sqrt{m_F(a,b)} + \sqrt{m_F(b,c)} \geq \sqrt{m_F(a,c)}$.*

Proof. Since $F = G$ at b and $(a+b)/2$ we have $m_F(a,b) = (F(a) - G(a))/2 + m_G(a,b) := x + X$ with $x \geq 0$. Similarly $m_F(b,c) = (F(c) - G(c))/2 + m_G(b,c) := y + Y$ and $m_F(a,c) = (F(a) - G(a))/2 + (F(c) - G(c))/2 + m_G(a,c) := x + y + Z$. By assumption $\sqrt{X} + \sqrt{Y} \geq \sqrt{Z}$. This easily implies $\sqrt{x+X} + \sqrt{y+Y} \geq \sqrt{x+y+Z}$. This concludes the proof. □

We have thus proved

THEOREM 2.9 (Sufficient condition for $F \in \mathbb{F}$). *The condition $(logF''(x))'' \geq 0$ is sufficient condition for the square root $\sqrt{m_F}$ to be a metric. In other words, the classes $\mathbb{F}, \mathbb{F}'$ coincide.*

2.4. Extention to functions. Now, we extend this result to functions. Suppose we are given functions $g, h, k \in L^1$ in some measure space S with a measure μ. Suppose $\int_S m_F(g(x), k(x)) d\mu$, $\int_S m_F(g(x), h(x)) d\mu$, $\int_S m_F(h(x), k(x)) d\mu$ are all well-defined and finite. Denote $M_F(k,g) := \int_S m_F(k(x), g(x)) d\mu$, $M_F(g,h) := \int_S m_F(g(x), h(x)) d\mu$, $M_F(k,h) := \int_S m_F(k(x), h(x)) d\mu$.

THEOREM 2.10. *Assume $M_F(k,g), M_F(g,h), M_F(k,h)$ are well-defined and finite, and $F \in \mathbb{F}$, i.e., $(\log F'')'' \geq 0$. Then we have $\sqrt{M_F(g,k)} \leq \sqrt{M_F(g,h)} + \sqrt{M_F(h,k)}$.*

Proof.

$$\sqrt{M_F(g,k)} = \sqrt{\int m_F(g,k) d\mu} = \sqrt{\int \left(\sqrt{m_F(g,k)}\right)^2 d\mu},$$

$$\leq \sqrt{\int \left(\sqrt{m_F(g,h)} + \sqrt{m_F(h,k)}\right)^2 d\mu}, \text{ by the assumption } F \in \mathbb{F},$$

$$\leq \sqrt{\int \left(\sqrt{m_F(g,h)}\right)^2 d\mu} + \sqrt{\int \left(\sqrt{m_F(h,k)}\right)^2 d\mu}, \text{ by Minkowski inequality,}$$

$$= \sqrt{\int m_F(g,h) d\mu} + \sqrt{\int m_F(h,k) d\mu} = \sqrt{M_F(g,h)} + \sqrt{M_F(h,k)}.$$

□

Remark 2.3. *The set $\mathbb{F}$ is convex, i.e., if F_1,F_2 both belong to $\mathbb{F}$, then $\alpha F_1+(1-\alpha)F_2\in\mathbb{F}$ for $\alpha\in[0,1]$.*

Proof. We prove a more general result. If $F_i, i=1,2$ satisfy $F_i''F_i''''-k(F_i''')^2\geq 0$ with a number $k\in\mathbf{R}$, then $F:=\alpha F_1+(1-\alpha)F_2$ also satisfies this inequality. The computation follows.

$$\begin{aligned}
&F''F''''-k(F''')^2\\
&=\alpha^2(F_1''F_1''''-k(F_1''')^2)+(1-\alpha)^2(F_2''F_2''''-k(F_2''')^2)\\
&\quad+\alpha(1-\alpha)(F_1''F_2''''+F_1''''F_2''-2kF_1'''F_2''')\\
&\geq 0+0+\alpha(1-\alpha)(2\sqrt{F_1''F_1''''F_2''F_2''''}-2kF_1'''F_2''')\\
&\geq 2\alpha(1-\alpha)(\sqrt{k^2(F_1'''F_2''')^2}-kF_1'''F_2''')\geq 0.
\end{aligned}$$

□

Also, if $F\in\mathbb{F}$ and $\alpha>0$, then $\alpha F\in\mathbb{F}$. Thus, the set $\mathbb{F}$ is a convex cone.

2.5. Do we really need $F\in C^4$? All the arguments we made are based on F being smooth: at least four times differentiable. However, it is not necessary. But we do need $F\in C^2$. A simple example is $F(x)=|x|$: $\sqrt{m_F(0,1)}+\sqrt{m_F(-1,0)}=0<\sqrt{m_F(-1,1)}$, the triangle inequality does not hold.

In fact, to ensure $F\in\mathbb{F}$ we simply need that F is twice differentiable and $\log F''$ is convex. The proof is given below.

Theorem 2.11. *If F is twice differentiable and $\log F''$ is convex, then $F\in\mathbb{F}$.*

Proof. Let $\bar{F}:=\log F''$. Then clearly $\bar{F}$ is continuous. Let $F^\epsilon:=\eta_\epsilon\star\bar{F}$, where $\eta_\epsilon(x):=\eta(x/\epsilon)/\epsilon$, and $\eta(x)$ is the standard mollifier, defined as follows:

$$\eta(x):=C\exp(1/(|x|^2-1)) \text{ if } -1<x<1, \text{ and } \eta(x):=0 \text{ otherwise.} \tag{2.14}$$

Then we have $F^\epsilon\in C^\infty$, and $F^\epsilon\to F$ uniformly on compact sets (see [12, p. 630]). Integrating $\exp(F^\epsilon)$ twice, we get a function f^ϵ such that $\log(f^\epsilon)''=F^\epsilon$. Hence, for any numbers a,b, we have $m_{f^\epsilon}(a,b)\to m_F(a,b)$, as $\epsilon\to 0$. (Note that although given a fixed F^ϵ, f^ϵ is determined modulo a linear function. Thus all f^ϵ lead to identical m_{f^ϵ}.)

It is easy to see that F^ϵ is convex. Thus $f^\epsilon\in\mathbb{F}$. Therefore, given any $a<b<c$ in Ω, we have the triangle inequality

$$\sqrt{m_{f^\epsilon}(a,b)}+\sqrt{m_{f^\epsilon}(b,c)}\geq\sqrt{m_{f^\epsilon}(a,c)}. \tag{2.15}$$

Let $\epsilon\to 0$, to get the desired triangle inequality: $\sqrt{m_F(a,b)}+\sqrt{m_F(b,c)}\geq\sqrt{m_F(a,c)}$. This proves $F\in\mathbb{F}$. □

Appendix A. Proof of the necessary condition.

Proof. The necessary condition in fact comes from the leading coefficient of a Taylor expansion.

First we consider the special case of three numbers $x-a,x,x+a$, with a positive and close to 0. By a Taylor expansion,

$$F(x+a)=F(x)+F'(x)a+\frac{F''(x)}{2}a^2+\frac{F'''(x)}{3!}a^3+\frac{F''''(x)}{4!}a^4+O(a^5), \tag{A.1}$$

Therefore

$$\frac{1}{2}(F(x+a)+F(x-a))-F(x)=\frac{F''(x)}{2}a^2+\frac{F''''(x)}{4!}a^4+O(a^5). \tag{A.2}$$

We also have

$$\begin{aligned}\frac{1}{2}(F(x+a)+F(x))-F(x+a/2)&=\frac{F''(x+\frac{a}{2})}{2}(\frac{a}{2})^2+\frac{F''''(x+\frac{a}{2})}{4!}(\frac{a}{2})^4+O(a^5)\\ &=\frac{F''(x)}{8}a^2+\frac{F'''(x)}{16}a^3+\frac{F''''(x)}{64}a^4+\frac{F''''(x)}{16\cdot 4!}a^4+O(a^5).\end{aligned}$$

Thus, if the triangle inequality holds, $0\leq\sqrt{m_F(x+a,a)}+\sqrt{m_F(x,x-a)}-\sqrt{m_F(x+a,x-a)}$, and we have

$$\begin{aligned}0&\leq\sqrt{(\frac{1}{2}(F(x+a)+F(x))-F(x+a/2))}+\sqrt{(\frac{1}{2}(F(x-a)+F(x))-F(x-a/2))}\\ &\quad-\sqrt{(\frac{1}{2}(F(x+a)+F(x-a))-F(x))}\\ &=\sqrt{\frac{F''(x)}{2}}a^2\frac{1}{16}\left(-\frac{1}{2}(\frac{F'''(x)}{F''(x)})^2+\frac{F''''}{2F''}\right)+O(a^3).\end{aligned}$$

Letting a tend to 0 leads to $F''F''''\geq(F''')^2$. This can be rewritten as $(F'''/F'')'\geq 0$, i.e. $(\log F'')''\geq 0$. □

Based on this proof, $(\log F'')''=0$ is a necessary condition for the equation

$$\sqrt{m_F(x+a,a)}+\sqrt{m_F(x,x-a)}=\sqrt{m_F(x+a,x-a)}. \tag{A.3}$$

This also implies that the set $\mathbb{G}$ coincides with the set $\{G:\sqrt{m_G(a,b)}+\sqrt{m_G(b,c)}=\sqrt{m_G(a,c)}$, for all numbers $a<b<c\}$.

REFERENCES

[1] K.S. Azoury and M.K. Warmuth, *Relative loss bounds for on-line density estimation with the exponential family of distributions*, Machine Learning, 43, 211-246, 2001.

[2] D.P. Bertsekas, *Nonlinear Programming*, Athena Scientific, 2003.

[3] L.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 Comput. Math. Phys., 7, 200-217, 1967.

[4] A. Buzo, A.H. Gray, R.M. Gray and J.D. Markel, *Speech coding based upon vector quantization*, IEEE Trans. on Acoustics, Speech and Signal Processing, 28(5), 562-574, 1980.

[5] A. Banerjee, S. Merugu, I. S. Dhillon and J. Ghosh, *Clustering with Bregman Divergences*, Machine Learning Research, 6, 1705-1749, October 2005.

[6] L.D. Brown, *Fundamentals of Statistical Exponential Families*, Institute of Math. Statistics, 1986.

[7] M. Collins, S. Dasgupta and R. Schapire, *A generalization of principal component analysis to the exponential family*, In Proc. of the Annual Conf. on NIPS, 2001.

[8] T.M. Cover and J.A. Thomas: *Elements of Information Theory*, New York, Wiley & Sons, 1991.

[9] I. Dhillon, S. Mallela and R. Kumar, *A divisive information–theoretic feature clustering algorithm for text classification*, Machine Learning Research, 3(4), 1265-1287, 2003.

[10] I. Csiszar and J. Korner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, Academic Press, 1981.

[11] D.M. Endres and J.E. Schindelin, *A new metric for probability distributions*, IEEE Trans. Inform. Theory, 49(7), 1858-1860, 2003.

[12] L.C. Evans, *Partial Differential Equations*, American Mathemtical Society, 1998.

[13] J. Forster and M.K. Warmuth, *Relative expected instantaneous loss bounds*, Proc. of the 13th Annual Conf. on Computational Learning Theory, 90-99, 2000.

[14] F. Itakura and S. Saito, *Analysis synthesis telephony based upon maximum likelihood method*, Repts. of the 6th Internat'l. Cong. Acoust., Y. Kohasi ed., Tokyo, C-5-5, C17-20, 1968.

[15] J. Uhlmann, *Satisfying general proximity/similarity queries with metric trees*, Information Processing Letters, 175-179, 1991.

[16] P.N. Yianilos *Data structures and algorithms for nearest neighbor search in general metric spaces*, ACM-SIAM Symp., Discrete Algorithms, 311-321, 1993.