COMMUN. MATH. SCI.
Vol. 6, No. 4, pp. 927–948

METRICS DEFINED BY BREGMAN DIVERGENCES: PART 2[*]

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

Abstract. Bregman divergences have played an important role in many research areas. Divergence is a measure of dissimilarity and by itself is not a metric. If a function of the divergence is a metric, then it becomes much more powerful. In Part 1 we have given necessary and sufficient conditions on the convex function in order that the square root of the averaged associated divergence is a metric. In this paper we provide a min-max approach to getting a metric from Bregman divergence. We show that the "capacity" to the power $1/e$ is a metric.

Key words. Metrics, Bregman divergence, triangle inequality, Kullback-Leibler divergence, Shannon entropy, capacity

AMS subject classifications. 26D10, 94A15

1. Introduction

Kullback-Liebler divergence was introduced to overcome some difficulties associated with the Shannon entropy. Bregman divergences are generalizations of this notion and are associated with strictly convex functions. In the last decade, Bregman divergences [3] have become an important tool in many research areas. For instance, several specific Bregman divergences, such as Itakura-Saito distance [4, 13], Kullback-Leibler divergence [8, 9], and Mahalanobis distance [16] have been used in machine learning as the distortion functions (or loss functions) for clustering tasks. These divergences have also been used in generalizations of principal component analysis to data with distributions belonging to the exponential family [7]. However, they are not metrics, because they are neither symmetric nor satisfy the triangle inequality.

The Jensen-Shannon divergence [15] defined by

$$\frac{1}{2}\left(KL\left(f,\frac{1}{2}(f+g)\right)+KL\left(g,\frac{1}{2}(f+g)\right)\right),$$

where KL is Kullback-Leibler divergence, is nonnegative, symmetric, bounded and vanishes only if $f=g$. Thus it has two of the three properties of a metric.

In [11], it was proved that the square root of Jensen-Shannon divergence is a metric. Jensen-Shannon divergence is an "averaged Bregman divergence" associated with the convex function $x\log x$. It is natural to ask whether the square root of other averaged Bregman divergences also satisfy the triangle inequality. In [19], we gave a sufficient and necessary condition on the associated convex function, in order that the square root of the averaged Bregman divergence is a metric. Clearly the justification of the triangle inequality is the only nontrivial part.

Triangle inequalities provide valuable information in pattern recognition research area. For instance, a popular method is to start with a similarity criterion given through a user-defined distance function and search for the nearest neighbor. Most often this happens in a multidimensional vector space. In the last decade, many

[*]Received: April 2, 2008; accepted: September 15, 2008. Comminicated by Shi Jin.

[†]Department of Mathematics, University of Connecticut, Storrs, CT, 06268, 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).

efficient algorithms have been proposed to find a nearest neighbor in a variety of metric spaces [17, 18]. One economical method of finding the nearest neighbor is through constructing a so-called metric tree, of say N objects, with height $\approx \log_2 N$. Then use of triangle inequality saves a lot of effort in finding the nearest neighbor. Indeed the total distance computations are reduced from N to $\log_2 N$. Our current work provides a large class of metrics for the choice of metric trees.

In this paper we provide a min-max approach to obtaining a metric from Bregman divergence. This method is sometimes used to measure the discrimination of probability distributions symmetrically. In information theory this leads to the notion of capacity, which has played an important role [8]. Capacity may be thought of as a minimal enclosing Kullback-Leibler divergence sphere. However, it is also known that capacity itself is not a metric [10]. In this paper, we show that the capacity to the power $1/e$ is a metric.

This paper is organized as follows. In section 2, we start with the definition of min-max procedure, and illustrate its relationship to the typical min-max problem in convex programming. In sectio 3, we derive necessary and sufficient conditions for the min-max procedure to lead to a metric in case of scalars. In section 4, we generalize this result to the case of vector spaces. Our conjecture is that the condition found also works for the vector space case. However, the computations become very complex. Therefore we will concentrate on one dimension, i.e., real numbers.

2. Preliminaries and problem formulation

In this paper, we adopt the following notations:

$\mathbf{R}$: real numbers,

RHS: the right-hand side of the equation,

LHS: the left-hand side of the equation,

Ω: the interior domain of the associated convex function$f$,i.e., $\{x : |f'(x)| < \infty\}$.

DEFINITION 2.1 (Bregman divergence). *The 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 assume all the convex functions mentioned in this paper are smooth, i.e., C^∞.

In general, $B_f(x,y)$ is not symmetric. We consider two kinds of averaging procedures to achieve symmetry averaging. The first procedure is

$$m_f(x,y) = \frac{1}{2}(B_f(x,z) + B_f(y,z)),\ z = \frac{1}{2}(x+y). \tag{2.1}$$

We can reformulate this as $m_f(x,y) = \frac{1}{2}(f(x)+f(y)) - f(\frac{1}{2}(x+y))$; This has been discussed in Part 1 [19].

The second procedure is min-maximizing. The inspiration for this is the following fact:

Given any N points $x_i, i = 1,2,...,n$ in R^N, the center and radius of the smallest sphere containing these points are simultaneously determined by $\min_{z \in R^N} \max_i |z - x_i|$.

For now, we consider only two elements of R^1 and define

$$m_f(x,y) = \min_z \max_{0 \le p \le 1} (pB_f(x,z) + (1-p)B_f(y,z)). \tag{2.2}$$

We will show that

$$m_f(x,y)=\max_{0\le p\le 1}(pB_f(x,z)+(1-p)B_f(y,z)),\text{ with } z=px+(1-p)y. \tag{2.3}$$

We can rewrite the above as

$$m_f(x,y)=\max_{0\le p\le 1, z=px+(1-p)y}(pf(x)+(1-p)f(y)-f(z)).$$

The auxiliary variable z is called the **center** of $m_f(x,y)$. For more properties of Bregman divergences, we refer interested readers to see [1, 5].

In this paper we show that $\sqrt{m_f}$ is a metric if and only if $3(\log f'')''\ge((\log f'')')^2$.

To this end, we need to verify that given three real numbers x,y,z

- $\sqrt{m_f(x,y)}\ge 0$, '=' holds only when $x=y$.
- $\sqrt{m_f(x,y)}=\sqrt{m_f(y,x)}$.
- $\sqrt{m_f(x,y)}+\sqrt{m_f(y,z)}\ge\sqrt{m_f(x,z)}$.

Clearly, we have the second property. From the forthcoming lemma, the first property follows.

Lemma 2.2. *The following are well known, important properties for Bregman divergences.*

$$B_f(x,z)\ge 0,''='' \text{ holds if and only if } x=z, \tag{2.4}$$

and

$$pB_f(x,z)+(1-p)B_f(y,z)\ge pB_f(x,px+(1-p)y)+(1-p)B_f(y,px+(1-p)y). \tag{2.5}$$

Therefore we have $m_f(x,y)\ge 0$, and equality holds only when $x=y$.
The proof can be found in [19].

Thus we only need to examine the third property: the triangle inequality. The following lemma establishes that if a power of $m_f(x,y)$ is a metric, then the power cannot exceed 1/2. Also, it will be clear from the lemma, that we only need to verify the case

$$\sqrt{m_f(x,z)}+\sqrt{m_f(z,y)}\ge\sqrt{m_f(x,y)},\text{ if } x<z<y. \tag{2.6}$$

Lemma 2.3. *Suppose $f(x)$ is a strictly convex, smooth (at least three times differentiable) function on an open set Ω. Denote*

$$m_f(p;x,y):=pf(x)+(1-p)f(y)-f(px+(1-p)y),$$

where $0<p<1$, and $x,y\in\Omega$. Then we have the following facts (the proof can be found in [19]).

(1). $m_f(p;x,y)\ge 0$, and equality holds if and only if $x=y$.

(2). Monotonicity: If $x<y<z, x,y,z\in\Omega$, then $m_f(p;x,y)<m_f(p;x,z), m_f(p;y,z)<m_f(p;x,z)$, which gives $\sqrt{m_f(p;x,y)}<\sqrt{m_f(p;x,z)}+\sqrt{m_f(p;z,y)}$, $\sqrt{m_f(p;y,z)}<\sqrt{m_f(p;y,x)}+\sqrt{m_f(p;x,z)}$.

(3). Suppose we know the triangle inequality holds for some positive r_0,

$$m_f(p;x,y)^{r_0} \le m_f(p;x,z)^{r_0} + m_f(p;z,x)^{r_0}. \tag{2.7}$$

Then the triangle inequality still holds for all $0 \le r < r_0$

$$m_f(p;x,y)^{r} \le m_f(p;x,z)^{r} + m_f(p;z,x)^{r}. \tag{2.8}$$

(4). Maximal possible exponent: if there exists a small neighborhood $(0,\epsilon)$ *such that*

$$m_f(p;x-a,x+a)^r \le m_f(p;x-a,x)^r + m_f(p;x,x+a)^r \text{ holds for all } a \in (0,\epsilon), \tag{2.9}$$

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

Thus, maximizing on $0<p<1$, then property (2) tells us that if $x<y<z, x,y,z \in \Omega$, then $\sqrt{m_f(x,y)} < \sqrt{m_f(x,z)} + \sqrt{m_f(z,y)}$, $\sqrt{m_f(y,z)} < \sqrt{m_f(y,x)} + \sqrt{m_f(x,z)}$.

2.1. Examples of Bregman divergences.

The above definition of Bregman divergence is defined between scalars, but it can be easily generalized to vectors. Let x,y be vectors in $\mathbf{R}^n$ with components $\{x_1,\dots,x_n\}$, $\{y_1,\dots,y_n\}$. The vector version Bregman divergence can be defined as

$$B_f(x,y) := \sum_{i=1}^n f(x_i) - f(y_i) - (x_i - y_i) f'(y_i).$$

In the literature, three Bregman divergences have played a more important role than the rest: Euclidean distance, Kullback-Leibler divergence (I-divergence), and Itakura-Saito distance. Their associated strictly convex functions are $f(x)=x^2$, $x\log x$, and $-\log x$, respectively.

Definition 2.4. *We are given two non-negative vectors* $x := (x_1,...,x_n), y := (y_1,...,y_n) \in \mathbf{R}^n$. *The Kullback-Leibler divergence is defined as*

$$KL(x,y) := \sum_{i=1}^n x_i \log \frac{x_i}{y_i}, \textit{when } \sum_{i=1}^n x_i = \sum_{i=1}^n y_i = 1. \tag{2.10}$$

The I-divergence is defined as

$$CKL(x,y) := \sum_{i=1}^n x_i \log \frac{x_i}{y_i} - \sum_{i=1}^n x_i + \sum_{i=1}^n y_i. \tag{2.11}$$

The Itakura-Saito distance is defined as

$$IS(x,y) := \sum_{i=1}^n -\log \frac{x_i}{y_i} + \sum_{i=1}^n \frac{x_i}{y_i} - 1. \tag{2.12}$$

2.2. The min-max procedure for m_f.

In this section, we describe the min-max procedure for obtaining a metric. This metric acting on any two points (scalars or vectors) can be regarded as the "radius" of the smallest sphere enclosing these two points.

[**Numbers:**] Let us consider the min-max procedure between two numbers first.

Lemma 2.5. *For $0 \le p \le 1$, and $q = 1-p$, denote $m_f(p;x,y) := pB_f(x,px+qy) + qB_f(y,px+qy) = pf(x)+qf(y)-f(px+qy)$, and*

$$m_f(x,y) := \max_{0\le p\le 1, p+q=1} m_f(p;x,y). \tag{2.13}$$

Suppose $x \ne y$. Then the maximizer $(p,q) = (p^*,q^*)$ of $m(p;x,y)$ can be solved uniquely by solving the equation:

$$f(x)-f(y)-(x-y)f'(p^*x+q^*y) = 0, \tag{2.14}$$

and as $(p,q) = (p^*,q^*)$, we have $m_f(p^*;x,y) = B_f(x,p^*x+q^*y) = B_f(y,p^*x+q^*y)$, i.e., equal divergences.

Proof. $p=0$ or $q=0$ implies $m_f(p;x,y)=0$, so we may assume $p\ne 0, p \ne 1$. Thus, we have $\frac{\partial m_f(p;x,y)}{\partial p} = 0$ for some $p = p^*, q = q^*$, which implies equation (2.14). By the Mean Value Theorem and the strict convexity of f, we know that p^*, q^* exist and are unique. Also, we can rewrite this equation as

$$f(x)-f(p^*x+q^*y)-(x-(p^*x+q^*y))f'(p^*x+q^*y) \tag{2.15}$$

$$= f(y)-f(p^*x+q^*y)-(y-(p^*x+q^*y))f'(p^*x+q^*y), \tag{2.16}$$

i.e., $B_f(x,p^*x+q^*y) = B_f(y,p^*x+q^*y)$, equal divergences. Since $m_f(x,y) = \max_{0\le p\le 1,p+q=1}(pB_f(x,px+qy)+qB_f(y,px+qy))$, we have $m_f(x,y) = B_f(x,p^*x+q^*y) = B_f(y,p^*x+q^*y)$. □

The following remark provides an alternate viewpoint.

Remark 2.6. Consider the problem

$$\min_z R, \text{subject to } B_f(x,z) \le R, B_f(y,z) \le R. \tag{2.17}$$

By standard convex programming arguments [2], this can be reformulated as

$$\max_{0\le p\le 1, p+q=1} m_f(p;x,y). \tag{2.18}$$

Thus $m_f(x,y) = B_f(x,p^*x+q^*y) = B_f(y,p^*x+q^*y)$ may be thought of as the radius of the smallest "sphere" enclosing x,y in the sense of Bregman divergence.

[**Vectors:**] We outline a generalization of this procedure to vectors to obtain a new type of metric in higher dimensional spaces, which we call the min-max metric.

Consider a strictly convex function F with its effective domain $\Omega := \{x \in \mathbf{R} : F(x) < \infty, F'(x) < \infty\}$. We are given vectors $f,g,l,a \in \mathbf{R}^n$, with their components $f_{(j)}, g_{(j)}, l_{(j)} \in \Omega$, and $a_{(j)} > 0$ for $j = 1,\ldots,n$. $M_F(f,g;a)$ is defined as the smallest R, such that

$$\min_{l\in\Omega^n} R, \text{subject to } B_F(f,l;a) \le R, B_F(g,l;a) \le R, \tag{2.19}$$

where the vector version of the Bregman divergence is defined as

$$B_F(f,l;a) := \sum_{j=1}^n (F(f_{(j)}) - F(l_{(j)}) - (f_{(j)} - l_{(j)})F'(l_{(j)}))a_{(j)}. \tag{2.20}$$

We can also consider the case of multiple vectors. Define

$$M_F(f_1,...,f_m;a) := \max_{p\in\Delta} \sum_{i=1}^{m} p_i B_F(f_i,<pf>;a),$$

where the convex set Δ is defined by

$$\Delta := \{(p_1,\dots,p_m) : \sum_{i=1}^{m} p_i = 1, \text{ and each } p_i \text{ is nonnegative number}\}.$$

Let $<pf> := \sum_{i=1}^{m} p_i f_i \in \Omega^n$. The following holds.

Lemma 2.7. *Consider vectors $f_1,...,f_m,h \in \Omega^n$ with an integer $m \geq 2$, and vectors $a \in \mathbf{R}^n$ with components $a_j > 0$, and $p \in \Delta$. Then we have*

$$\sum_{i=1}^{m} p_i B_F(f_i,h;a) \geq \sum_{i=1}^{m} p_i B_F(f_i,<pf>;a), \tag{2.21}$$

and equality holds only when $h = <pf>$.

Proof.

$$LHS - RHS = B_F(<pf>,h;a) \geq 0. \tag{2.22}$$

□

Theorem 2.8 (Properties of $M_F(f_1,\dots,f_m;a)$). *Given m vectors $f_1,\dots,f_m$ that are not all identical, let $p^* = (p_1^*,...,p_m^*)$ be the maximizer of $M_F(f_1,\dots,f_m;a)$. Denote $M(p) = \sum_{i=1}^{m} p_i B_F(f_i,<pf>;a)$. Then we have*

1.$M_F(f_1,\dots,f_m;a) > 0$.

2.*For all i with $0 < p_i^* < 1$, we have the same constant $\lambda := B_F(f_i,<p^*f>;a)$.*

3.$M(p^*) = B_F(f_i,<p^*f>;a)$, *for all i with $p_i^* > 0$.*

Proof. Let $p = (1,0,...,0)$, then $M(p) = B_F(f_1,f_1;a) = 0$. Since $M(p)$ is strictly concave, then $M(p^*) > M(p) = 0$, which means that p^* is not one of the vertices of the domain of p, and $M_F(f_1,\dots,f_m;a) > 0$.
Let

$$L(p,\lambda) = \sum_{i=1}^{m} p_i B_F(f_i,<pf>;a) - \lambda\left(\sum_{i=1}^{m} p_i - 1\right).$$

For those i, with $0 < p_i^* < 1$, we have $\frac{\partial \mathrm{L}}{\partial p_i}(p^*) = 0$, which implies that

$$\sum_{j=1}^{n} (F(f_i) - f_i F'(<p^*f>))_{(j)} a_{(j)} - \lambda = 0, i.e., \tag{2.23}$$

$$B_F(f_i,<p^*f>;a) = \lambda - \sum_{j=1}^{n} (F(<p^*f>) - <p^*f> F'(<p^*f>))_{(j)} a_{(j)}. \tag{2.24}$$

Thus, $B_F(f_i,<p^*f>;a)$ is a constant, independent of i.

Note that since we are maximizing a concave functional L, we always have $\frac{\partial L}{\partial p_i}(p^*)\leq 0$. Thus, the statement (2) is the optimal condition for p^*.

Now, multiplying the statement (2) by p_i^*, and summing over all i, we have $\lambda = M(p^*)$, and

$$M(p^*)=B_F(f_i,<p^*f>;a),\ \text{for all } i \text{ with } p_i^*>0. \tag{2.25}$$

□

In other words, $B_F(f_i,<p^*f>;a)$ is constant on the set of i such that $p_i^*>0$. Let R denote this number. Then for each j with $B_F(f_j,p^*f;a)<R$, we must have $p_j^*=0$.

In Euclidean space, by convex programming [2] the radius R of the sphere containing m vectors $v_1,...,v_m$ in $\mathbf{R}^n$ may be characterized by

$$R=\min_{v\in\mathbf{R}^N}\max_i\|v_i-v\|. \tag{2.26}$$

In the case of Bregman divergences, we have

Theorem 2.9.

$$M_F(f_1,...,f_m;a)=\min_{h\in\Omega^n}\max_i B_F(f_i,h;a). \tag{2.27}$$

Proof. By the previous lemma, since $\sum_{i=1}^m p_i=1$

$$\sum_{i=1}^m p_iB_F(f_i,<pf>;a)\leq\sum_{i=1}^m p_iB_F(f_i,h;a)\leq\max_i B_F(f_i,h;a) \tag{2.28}$$

Now, taking the maximum in p on both sides and minimizing among all possible $h\in\Omega^n$, we have

$$M_F(f_1,...,f_n;a)\leq\min_h\max_i B_F(f_i,h;a), \tag{2.29}$$

and by the previous theorem, we know that when $h=<pf>$, equality holds. □

3. Min-max metric in case of numbers

Based on Lemma 2.3, our task is to find the condition on f which ensures that for any $a<b<c$, we have

$$\sqrt{m_f(a,b)}+\sqrt{m_f(b,c)}-\sqrt{m_f(a,c)}\geq 0. \tag{3.1}$$

In the next lemma, we consider three numbers that are close to each other. Then we use Taylor's expansion to get the leading term's coefficient in the LHS of equation (3.1). This coefficient needs to be nonnegative, giving us a necessary condition. Since the computation is quite lengthy, we do it in the appendix.

Lemma 3.1 (Necessary condition). *If $\sqrt{m_f}$ is a metric, then*

$$-4\left(\frac{f'''}{f''}\right)^2+3\frac{f''''}{f''}\geq 0,\ \ \textit{or}\ 3(\log f'')''-((\log f'')')^2\geq 0. \tag{3.2}$$

Let us examine the case when equality holds. Denote $z:=(\log f'')'$. Then we have $3z'-z^2\geq 0$. Clearly equality holds when either $z(x)=-\frac{3}{x-c_1}$, with $x\geq c_1$, or $z(x)=0$ for all x. In the first case, we have $f''(x)=c_2(x-c_1)^{-3}$ with $c_2>0$, i.e., $f(x)=\frac{c_2}{x-c_1}+c_3x+c_4$. In the second case, we have $f(x)=c_2x^2+c_3x+c_4$. In both cases, $c_1,c_2>0,c_3,c_4$ are constants of integration.

Next, we can solve for the maximizer p in $m_f(p;a,b)$ by equation (2.14). In the first case, let $f(x):=\frac{1}{x}$, then $f'(x)=-\frac{1}{x^2}$. For any two points $(a,\frac{1}{a}),(b,\frac{1}{b})$, equation (2.14) yields that $f'(pa+(1-p)b)=(f(a)-f(b))/(a-b)=-\frac{1}{ab}$. Thus, $p=\sqrt{b}/(\sqrt{a}+\sqrt{b})$, and

$$\begin{aligned}&m_f(a,b)=pf(a)+(1-p)f(b)-f(pa+(1-p)b)\\&=\frac{1}{a}\left(1-\frac{\sqrt{a}}{\sqrt{a}+\sqrt{b}}\right)+\frac{1}{b}\left(1-\frac{\sqrt{b}}{\sqrt{a}+\sqrt{b}}\right)-\frac{1}{\sqrt{ab}}=\left(\frac{1}{\sqrt{a}}-\frac{1}{\sqrt{b}}\right)^2.\end{aligned}$$

Hence, its square root is a metric, and for any given numbers $a<b<c$ we have $\sqrt{m_f(a,b)}+\sqrt{m_f(b,c)}=\sqrt{m_f(a,c)}$.

In the second case, we get the Euclidean distance, and we have $\sqrt{m_f(a,b)}+\sqrt{m_f(b,c)}=\sqrt{m_f(a,c)}$ and $p=1/2$ in both cases.

Definition 3.2. *Define a class of functions* $\mathbb{F}'$ *as* $\{f:3(\log f'')''\geq((\log f'')')^2$, *with* $f''>0\}$. *Define the sub-class of functions* $\mathbb{G}$ *by* $\{f:f''>0,\ 3(\log f'')''=((\log f'')')^2\}$. *This class* $\mathbb{G}$ *is in fact the same as the class* $\{g:g(x)=\frac{c_2}{x-c_1}+c_3x+c_4,x>c_1\}$ $\cup\{g:g(x)=c_2x^2+c_3x+c_4\}\cup\{g:g(x)=\frac{c_2}{x-c_1}+c_3x+c_4,x<c_1\}$, *here* $c_2>0,c_1,c_3,c_4$ *are scalars.*

Remark 3.3. Examples of f with $(\log f'')''\geq((\log f'')')^2/3$:

$f(x)$	$(\log f'')$	$(\log f'')'$	$(\log f'')''$
$x^\alpha/(\alpha(\alpha-1)),\alpha=2,\alpha\geq -1$	$(\alpha-2)\log x$	$(\alpha-2)/x$	$-(\alpha-2)/x^2$
$x\log x-x$	$-\log x$	$-1/x$	$1/x^2$
$-\log x$	$-2\log x$	$-2/x$	$2/x^2$
$1/(2x)$	$-3\log x$	$-3/x$	$3/x^2$

Definition 3.4. *We say that* f *is in the class* $\mathbb{F}$ *if* $\sqrt{m_f}$ *is a metric,* f *is smooth (at least four times differentiable), and* $f''>0$.

Now, we will point out two important relations between $\mathbb{F}'$ and $\mathbb{G}$.

Lemma 3.5. *Let four points* $x_1<x_2<x_3<x_4$ *and* $f\in\mathbb{F}'$ *be given. Then there are scalars* $c_1,c_2\neq 0,c_3,c_4$ *such that one of the functions* $\frac{c_2}{x-c_1}+c_3x+c_4$ *(the RECIPROCAL CASE) or* $c_2x^2+c_3x+c_4$ *(the QUADRATIC CASE) agrees with* f *at exactly these* 4 *points.*

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

$$y_{k+1}-y_k=f(x_{k+1})-f(x_k)=\frac{c_2}{x_{k+1}-c_1}-\frac{c_2}{x_k-c_1}+c_3(x_{k+1}-x_k),k=1,2,3. \tag{3.3}$$

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

$$\frac{y_{k+3}-y_{k+2}}{x_{k+3}-x_{k+2}}-\frac{y_{k+2}-y_{k+1}}{x_{k+2}-x_{k+1}}=-\frac{c_2}{(x_{k+3}-c_1)(x_{k+2}-c_1)}+\frac{c_2}{(x_{k+2}-c_1)(x_{k+1}-c_1)}. \tag{3.4}$$

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

$$\frac{r_{3,4}-r_{2,3}}{x_4-x_2}\left(\frac{r_{2,3}-r_{1,2}}{x_3-x_1}\right)^{-1} = \frac{x_1-c_1}{x_4-c_1}. \tag{3.5}$$

The left hand side (LHS) of equation (3.5) is positive because the function f is strictly convex. Therefore either $c_1 < x_1 < x_4$ or $c_1 > x_4 > x_1$.

Note that as $c_1 \to \infty$, $LHS \to 1^+$, and as $c_1 \to x_4^+$, $LHS \to \infty$. Similarly, as $c_1 \to -\infty$, $LHS \to 1^-$, and as $c_1 \to x_1^-$, $LHS \to 0^+$. Thus if $LHS \neq 1$, the existence of c_1 is guaranteed by the Intermediate Value theorem, and c_1 is unique because the right hand side is an increasing function of c_1. Once we have c_1, we can trace the steps backwards to solve for c_2, c_3, c_4 (using equations (3.4), (3.3)) and get the RECIPROCAL CASE.

If $LHS = 1$, then c_1 cannot be determined. In this case, follow the same steps as above replacing the reciprocal by the quadratic to see that we obtain the QUADRATIC CASE. □

LEMMA 3.6. *We are given any $f \in \mathbb{F}'$, and $g \in \mathbb{G}$. Then $H = f - g$ vanishes at most at four points or it vanishes identically on a segment and is positive outside this segment. If H vanishes at four adjacent points $\{x_1 < \ldots < x_4\}$ and nowhere else then H is positive and convex outside $[x_1, x_4]$.*

If H vanishes at the four points $x_1 < x_2 < x_3 < x_4$, then $f - g$ takes values with signs $+,-,+,-,+$ on $(-\infty, x_1)$, (x_1, x_2), ... (x_4, ∞).

Proof. If $H = 0$ at 5 points then H' vanishes at least at 4 points, and $H'' = f'' - g''$ vanishes at least at 3 points, say $\eta_1 < \eta_2 < \eta_3$. This implies that $\log(f''/g'')$ vanishes at 3 points, and $(\log(f''/g''))'$ vanishes at least at 2 points, say $\xi_1 < \xi_2$ with $\eta_1 \leq \xi_1 \leq \eta_2 \leq x_2 \leq \eta_3$. Denote $y(x) := (\log f(x)'')'$, $z(x) := (\log g(x)'')'$, then $y' \geq y^2/3 \geq 0$, (y increases monotonically), $z' = z^2/3$, and we know that $y - z$ vanishes at ξ_1, ξ_2.

There are three possibilities for z, depending on the sign of the initial condition $z(x_0)$: either $z(x_0) < 0$, $z(x_0) = 0$, or $z(x_0) > 0$. Their solutions are $z = -3/(x-c)$ with $x > c$, $z = 0$, or $z = -3/(x-c)$ with $x < c$, respectively (here c is chosen to satisfy the initial condition). The discussion in the first case can be applied to the third case if we replace $f(x), g(x)$ with $f(-x), g(-x)$ to get $z(-x_0) < 0$. Hence, we only need to discuss the cases $z(x_0) > 0$ and $z(x_0) = 0$.

In the first case, we have $z = -3/(x-c)$ with some constant c to be determined, and (c, ∞) is the domain of g. Then z is negative in this domain. Hence, the zeroes of $y(x) - z(x)$ lie in the domain $D := \{x > c : y(x) < 0\}$. Note D is a connected set because $3y' \geq y^2$ implies that y increases monotonically. On D, we have

$$\left(-\frac{1}{y}\right)' \geq \frac{1}{3}, \left(-\frac{1}{z}\right)' = \frac{1}{3}, \text{ and so } \left(-\frac{1}{y} + \frac{1}{z}\right)' \geq 0. \tag{3.6}$$

Recall that $y - z = 0$ at ξ_1, ξ_2, i.e., $1/y - 1/z$ vanishes at ξ_1, ξ_2. Combining this with equation (3.6), we conclude that $-1/y + 1/z = 0$, i.e., $y = z$ on $[\xi_1, \xi_2]$. We conclude that $f'' = c_0 g''$ with some constant of integration c_0. Moreover, $f''(\eta_2) = g''(\eta_2)$ and $\xi_1 \leq \eta_2 \leq \xi_2$, implying that $c_0 = 1$. Since $f' - g'$ has zeros in $[\xi_1, \xi_2]$, we must have $f = g$ in $[\xi_1, \xi_2]$.

Now, we consider the second case where there are exactly 4 zeros, say x_1, x_2, x_3, x_4. Since $(-1/y + 1/z)' \geq 0$, $y - z$ has signs $-,+$, in $(-\infty, b)$, (b, ∞), where b is the zero of $y - z = 0$. Therefore $\log f'' - \log g''$ takes values with signs $+,-,+$. Finally, $f - g$

takes values with signs $+,-,+,-,+$ on $(-\infty,x_1)$, (x_1,x_2), ... (x_4,∞).

If we solve the equation $z'=z^2/3$ with $z(x_0)=0$, we have $z(x)=0$ for all $x\in\mathbf{R}$. Since y increases monotonically, the equality $y=z$ at ξ_1,ξ_2 implies $y=z=0$ on the whole interval $[\xi_1,\xi_2]$. Using the same arguments as in the case $z(x_0)<0$ above, we find $f=g$ on the whole interval $[\xi_1,\xi_2]$. Now suppose there are only 4 zeros. Since $z(x)=0$ for all x, $y-z$ monotonically increases. Using similar arguments as above, we can show that $f-g$ takes values with signs $+,-,+,-,+$ on $(-\infty,x_1)$, (x_1,x_2), ... (x_4,∞). □

Note that this lemma also implies that given any $f\in\mathbb{F}'$, and three distinct numbers a,b,c, we can construct a unique $g\in\mathbb{G}$ such that $f(a)=g(a)$, $f'(a)=g'(a)$, $f(b)=g(b)$ and $f(c)=g(c)$.

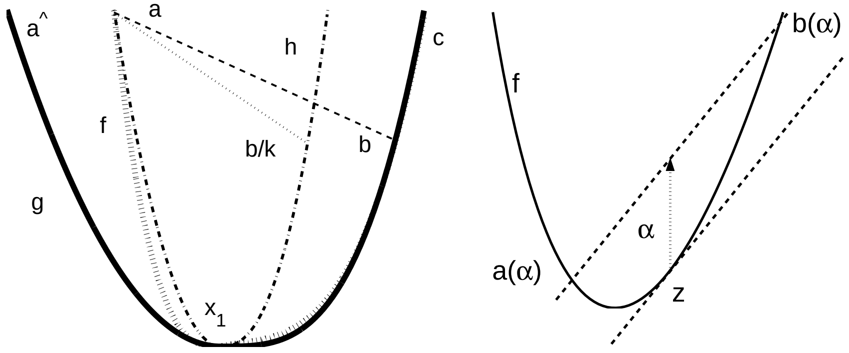

Fig. 3.1. *Figure for Thm. 3.8 (Left) and figure for Setting 4.1 (Right)*

Here are several properties of $m_f(\cdot,\cdot)$.

Lemma 3.7. *We are given numbers a,b, and $k\neq 0$, and a strictly convex function h.*

- *Let $g(x)=h(x-c_0)+c_1x+c_2$, where c_0,c_1,c_2 are scalars. Then $m_g(a+c_0,b+c_0)=m_h(a,b)$.*
- *Let $g(x)=h(kx)$, then $m_g(a,b)=m_h(a/k,b/k)$, and if $h\in\mathbb{G}$, then $g\in\mathbb{G}$.*
- *We are given three numbers $a<b\leq b_1$ and a strictly convex function f. Assume $f(a)=h(a)\geq f(b_1)=h(b)$, and $f\leq h$ on $[a,b]$. Then $m_f(a,b_1)\geq m_h(a,b)$.*

Proof. The first two statements are trivial. Hence, we only prove the third statement. Let the line connecting two points $(a,h(a))$ and $(b,h(b))$ be $y=s_1x+t_1$, and the the line connecting the two points $(a,f(a))$ and $(b_1,f(b_1))$ be $y=s_2x+t_2$. By assumption, both slopes s_1,s_2 satisfy the relation $s_1\leq s_2\leq 0$, and so $s_1x+t_1\leq s_2x+t_2$ for any $x\in[a,b]$.

Consider the two regions $D_1:=\{(x,y):h(x)\leq y\leq s_1x+t_1\}$ and $D_2:=\{(x,y):f(x)\leq y\leq s_2x+t_2\}$. The validity of $s_1x+t_1\leq s_2x+t_2$ and $h(x)\geq f(x)$ for any $x\in[a,b]$ implies that $D_1\subset D_2$. Therefore the largest vertical line segment in D_1 cannot be longer than the largest vertical line segment in D_2. That is, $m_h(a,b)\leq m_g(a,b_1)$. □

Now, we are ready to show the sufficient condition, i.e., $\mathbb{F}'\subset\mathbb{F}$.

THEOREM 3.8. *If f is four times differentiable, $f''>0$ and $3(\log f'')''\geq((\log f'')')^2$, then $\sqrt{m_f(a,b)}+\sqrt{m_f(b,c)}\geq\sqrt{m_f(a,c)}$ for any $a<b<c\in\Omega$.*

Proof. Let x_1 be the center of $m_f(a,c)$. Without loss of generality, assume that $a<x_1\leq b<c$, otherwise replace $f(x)$ with $f(-x)$. Also assume $f(a)=f(c)$, otherwise replace $f(x)$ with $f(x)-(x-a)(f(c)-f(a))/(c-a)$. Then $f'(x_1)=0$.

By Lemma 3.6, there exists a $g\in\mathbb{G}$ agreeing with f at x_1,b,c, and tangent at x_1 (i.e., $g'(x_1)=0$). Since g is convex and $g'(x_1)=0$, we can find a point $\widehat{a}$, such that $g(\widehat{a})=f(c)=g(c)$. Since f,g agree at x_1,b,c, and are tangential at x_1, by Lemma 3.6 $f-g$ takes values with signs $+,+,-,+$ on $(-\infty,x_1),(x_1,b),(b,c),(c,\infty)$. Thus $\widehat{a}\leq a<x_1$. Since $g(\widehat{a})=f(a)=f(c)=g(c)$, $g(x_1)=f(x_1)$ and $g'(x_1)=f'(x_1)$, we have

$$m_f(a,c)=m_g(\widehat{a},c). \tag{3.7}$$

Further, since $f\leq g$ on $[b,c]$,

$$m_f(b,c)\geq m_g(b,c). \tag{3.8}$$

Next, we claim $m_g(\widehat{a},b)\leq m_f(a,b)$. To see this, set $h(x):=g(kx)$, where $k:=\widehat{a}/a\geq 1$. Then by Lemma 3.7 we have $h\in\mathbb{G}$, and

$$m_h(a,b/k)=m_h(\widehat{a}/k,b/k)=m_g(\widehat{a},b). \tag{3.9}$$

Since $f\in\mathbb{F},h\in\mathbb{G}$, there exists a $\widehat{c}$ with $f(\widehat{c})=h(\widehat{c})$ such that $f-h>0$ on $(\widehat{c},\infty)$. Moreover, $h(c)=g(kc)\geq g(c)=f(c)$, so that $\widehat{c}>c$. Since f,h agree at $a,\widehat{c}$ and are tangential at x_1, then by Lemma 3.6 $f-h$ takes values with signs $+,-,-,+$ on $(-\infty,a)$, (a,x_1), $(x_1,\widehat{c})$, $(\widehat{c},\infty)$. Therefore we have $f\leq h$ in the interval $[a,b]$. Furthermore, because $f(a)=h(a)\geq f(b)=h(b/k)$, the line segment connecting $(a,f(a)),(b,f(b))$ lies above the segment connecting $(a,h(a)),(b/k,h(b/k))$. Hence by Lemma 3.7, we have $m_f(a,b)\geq m_h(a,b/k)$. Using equation (3.9), we have

$$m_f(a,b)\geq m_h(a,b/k)=m_g(\widehat{a},b). \tag{3.10}$$

Finally, from (3.7), (3.8), (3.10), and $\sqrt{m_g(\widehat{a},c)}=\sqrt{m_g(\widehat{a},b)}+\sqrt{m_g(b,c)}$ for any $g\in\mathbb{G}$, we have

$$\sqrt{m_f(a,c)}=\sqrt{m_g(\widehat{a},c)}=\sqrt{m_g(\widehat{a},b)}+\sqrt{m_g(b,c)}\leq\sqrt{m_f(a,b)}+\sqrt{m_f(b,c)}. \tag{3.11}$$

□

REMARK 3.9. 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 $\alpha\in[0,1]$ (the proof can be found in [19]).

We now generalize to vectors.

THEOREM 3.10. *Let $F\in\mathbb{F}$ and $f,g,l,a\in\mathbf{R}^n$, with their $i-$th components denoted by f_i,g_i,l_i,a_i respectively. Denote*

$$M_F(f,g;a):=\sum_{i=1}^{n}m_F(f_i,g_i)a_i,\ \textit{then}\ \sqrt{M_F(f,g;a)}\leq\sqrt{M_F(f,l;a)}+\sqrt{M_F(l,g;a)}. \tag{3.12}$$

Proof.

$$\sqrt{M_F(f,g_i)} = \sqrt{\sum_{i=1}^{n} m_F(f_i,g_i)a_i} = \sqrt{\sum_{i=1}^{n}\left(\sqrt{m_F(f_i,g_i)}\right)^2 a_i},$$

$$\leq \sqrt{\sum_{i=1}^{n}\left(\sqrt{m_F(f_i,l_i)}+\sqrt{m_F(l_i,g_i)}\right)^2 a_i}, \textit{ by the assumption } F\in\mathbb{F},$$

$$\leq \sqrt{\sum_{i=1}^{n}\left(\sqrt{m_F(f_i,l_i)}\right)^2 a_i} + \sqrt{\sum_{i=1}^{n}\left(\sqrt{m_F(l_i,g_i)}\right)^2 a_i}, \textit{ by Minkowski inequality,}$$

$$= \sqrt{\sum_{i=1}^{n} m_F(f_i,l_i)a_i} + \sqrt{\sum_{i=1}^{n} m_F(l_i,g_i)a_i} = \sqrt{M_F(f,l)} + \sqrt{M_F(l,g)}.$$

□

4. Generalization to vector spaces: capacity, min-max IS distance

In this section, we describe a family of metrics based on min-max Bregman divergences between vectors or functions. Our conjecture is that the condition found in the previous section should work in case of vectors and functions as well. Due to the complexity of this problem, we generalize this result only to two separate important cases. One is "capacity" associated with $f = x\log x$ and the other is min-max IS distance associated with $f = -\log x$. First, we need to analyze the range of the maximizer p^*.

4.1. Range of the maximizer p^*.

The main purpose of this subsection is to provide a bound on the actual range of p^*. This result will be used to prove that the capacity to the power $1/e$ is a metric.

Notation 4.1. Let the domain of a strictly convex function f be (γ_1,γ_2). Fix a point $z\in(\gamma_1,\gamma_2)$, and $s := f'(z)$.

Consider the family of parallel lines $\{y = sx+\alpha$, with $\alpha\in(\alpha_0,\alpha_1)$, $\alpha_0 := f(z) - sz, \alpha_1 := \min_{i=1,2}(f(\gamma_i) - s\gamma_i)\}$.

For each $\alpha\in(\alpha_0,\alpha_1)$, any line $y = sx+\alpha$ intersects $y = f(x)$ twice, say at $a(\alpha)$ and $b(\alpha)$, with $a(\alpha)\leq b(\alpha)$. Then z is the center of $m_f(a(\alpha),b(\alpha))$ for all $\alpha\in(\alpha_0,\alpha_1)$. We have $f(a(\alpha)) - sa(\alpha) = \alpha$, $f(b(\alpha)) - sb(\alpha) = \alpha$. Differentiating with respect to α, we have

$$\frac{da}{d\alpha} = \frac{1}{f'(a)-s}, \frac{db}{d\alpha} = \frac{1}{f'(b)-s}. \tag{4.1}$$

The maximizer $p(\alpha) := \arg\max_p m(p;a(\alpha),b(\alpha))$. Solving $z = p(\alpha)a(\alpha) + (1 - p(\alpha))b(\alpha)$, we find $p(\alpha) = (b(\alpha)-z)/(b(\alpha)-a(\alpha))$.

Lemma 4.2. *Suppose $f'''\leq 0$, and $f''>0$. If $a<b$ and $p^* := \arg\max_p m(p;a,b)$, then $p^* > \frac{1}{2}$. Using 4.1, $z = p(\alpha)a(\alpha)+(1-p(\alpha))b(\alpha)$ is fixed for all α, and $p'(\alpha)\geq 0$. Therefore, if $\gamma_1 > -\infty$, we have*

$$\sup_x p^*(x) = \sup_x \arg\max m(p;\gamma_1,x), \ \inf_x p^*(x) = 1 - \sup_x p^*(x). \tag{4.2}$$

Proof. Note that the maximizer $p = p^*$ satisfies

$$\frac{f(b)-f(a)}{b-a} - f'(p^*a+(1-p^*)b) = 0. \tag{4.3}$$

By Taylor's expansion, as $|a-b|$ gets close to 0 we have

$$\frac{1}{2}(b-a)f''(a) - ((1-p^*)(b-a)f''(a)) + O(|a-b|^2) = 0, \tag{4.4}$$

which implies $p^* \to \frac{1}{2}$, as $a \to b$.

For the second statement, we proceed as follows. In the setting of 4.1, consider the intersections $(a(\alpha), b(\alpha))$ with $b(\alpha) \geq a(\alpha)$, $a(\alpha_0) = b(\alpha_0) = z$. Since by assumption $f''' \leq 0$, f' is concave.

$$f'(x) - f'(z) \geq \frac{f'(b)-f'(z)}{b-z}(x-z), \text{ for } x \in (z,b), \tag{4.5}$$

which implies

$$B_f(b,z) = f(b) - f(z) - (b-z)f'(z) = \int_z^b (f'(x)-f'(z))dx \tag{4.6}$$

$$\geq \frac{f'(b)-f'(z)}{b-z} \int_z^b (x-z)dx = \frac{1}{2}(f'(b)-f'(z))(b-z). \tag{4.7}$$

Similarly,

$$B_f(a,z) = \int_a^z (f'(z)-f'(x))dx \leq \frac{1}{2}(f'(z)-f'(a))(z-a). \tag{4.8}$$

Since $B_f(b,z) = B_f(a,z)$, using the above two equations we have

$$(f'(b)-f'(z))(b-z) - (f'(z)-f'(a))(z-a) \leq 0. \tag{4.9}$$

On the other hand, since $p(\alpha)$ satisfies $z = p(\alpha)a(\alpha) + (1-p(\alpha))b(\alpha)$, differentiating with respect to α we have

$$0 = p'a + pa' - p'b + (1-p)b' = p'(a-b) + \frac{1}{b-a}\left(\frac{b-z}{f'(a)-f'(z)} + \frac{z-a}{f'(b)-f'(z)}\right). \tag{4.10}$$

We have

$$p' = \frac{1}{(b-a)^2} \cdot \frac{(b-z)(f'(b)-f'(z)) - (a-z)(f'(a)-f'(z))}{(f'(b)-f'(z))(f'(a)-f'(z))}. \tag{4.11}$$

Since $f'' > 0$,

$$(f'(b)-f'(z))(f'(a)-f'(z)) = \frac{(f'(b)-f'(z))(f'(a)-f'(z))}{(b-z)(a-z)}(b-z)(a-z) < 0.$$

Finally, using equation (4.9), equation (4.11) implies $p'(\alpha) > 0$.

Given $a < b$, assume $p^* = \arg\max_p m(p; a, b)$, and let $z(\alpha_0) = p^*a + (1-p^*)b$. Then because $p'(\alpha) > 0$, we have $p^* > p(\alpha_0) = \frac{1}{2}$, so $p^* > q^*$. □

Remark 4.3. By the above result, if $f(x):=x\log x$, then $\gamma_1=0$. p^* is bounded by the sup and the inf of $\{\arg\max_p m_p(0,a), a>0\}$. Since

$$m(p;0,a)=p\cdot 0+qa\log a-qa\log qa=-aq\log q, \tag{4.12}$$

we have

$$\arg\max_p m(p;0,a)=1-\frac{1}{e}, \text{ thus } \frac{1}{e}\leq p^*\leq 1-\frac{1}{e}. \tag{4.13}$$

Of course, this result does not depend on the base of the logarithm.

Remark 4.4. Let us consider a subset of Bregman divergences, $f(x):=\frac{1}{\alpha(\alpha-1)}x^\alpha$, $\alpha>1$, and restrict its domain to $(0,\infty)$. Since the maximizer (p,q) of $m(p;0,a)$ satisfies $a^\alpha-0^\alpha=(a-0)\alpha(p\cdot 0+qa)^{\alpha-1}$, we obtain $q=\alpha^{-\frac{1}{\alpha-1}}$. Note that this result is independent of a. Interestingly, as $\alpha\to 1$ we have $q=\lim_{\alpha\to 1}(1+(\alpha-1))^{\frac{-1}{\alpha-1}}=1/e$, which is exactly the conclusion in the previous remark.

We extend the above to functions on a measure space.

Theorem 4.5. *Given functions $f(x),g(x)$ and a convex function F, let*

$$p^*(x):=\arg\max_p(pB_F(f(x),pf(x)+qg(x))+qB_F(g(x),pf(x)+qg(x))), \tag{4.14}$$

with $p+q=1$, and let

$$p^*:=arg\max_p\int pB_F(f(x),pf(x)+qg(x))+qB_F(g(x),pf(x)+qg(x))d\mu, \tag{4.15}$$

with $p+q=1$. We assume that this integral well-defined (for instance, under the condition $\mu(\Omega)<\infty$).
The above lemma gives a bound for p^*:

$$\sup_{x\in\Omega}p^*(x)\geq p^*\geq\inf_{x\in\Omega}p^*(x). \tag{4.16}$$

Proof. For the sake of simplicity, let f,g denote $f(x)$ and $g(x)$.
By equation (2.14), $p^*(x)$ and p^* satisfy the equations

$$F(f)-F(g)=(f-g)F'(p^*(x)f+q^*(x)g) \tag{4.17}$$

and

$$\int F(f)-F(g)d\mu=\int(f-g)F'(p^*f+q^*g)d\mu. \tag{4.18}$$

Integrating the first equation, and subtracting the second one, we have

$$\int(f-g)\left(F'(p^*(x)f+q^*(x)g)-F'(p^*f+q^*g)\right)d\mu=0. \tag{4.19}$$

By the Mean Value Theorem, there exists an $\eta(x)$ lying between $p^*(x),p^*$ such that

$$F'(p^*(x)f+q^*(x)g)-F'(p^*f+q^*g)=(p^*(x)-p^*)(f-g)F''(\eta(x)f+(1-\eta(x))g). \tag{4.20}$$

Hence, we have

$$\int(f-g)^2F''(\eta(x)f+(1-\eta(x))g)(p^*(x)-p^*)d\mu=0. \tag{4.21}$$

But since $F''>0$ and $f-g$ does not vanish everywhere, we conclude that p^* must lie between $\sup_{x\in\Omega}p^*(x)$, and $\inf_{x\in\Omega}p^*(x)$. ∎

4.2. Min-max Kullback-Leibler divergence: capacity, $F(x)=x\log x$. We start with the following setting.

Let $L^1_+(\Omega):=\{f\in L^1(\Omega),f\geq 0\}$ $f_1,\dots,f_n\in L^1_+(\Omega)$ with $n\geq 2$, and $p\in\Delta$, where $\Delta:=\{(p_1,p_2,\dots,p_n)\in\mathbf{R}^n,p_i\geq 0,\sum_{i=1}^n p_i=1\}$. Also let $<p,f>:=\sum_{i=1}^n p_i f_i,$. We define a functional

$$M_p(f_1,...,f_n):=\int_\Omega\sum_{i=1}^n p_i f_i\log\frac{f_i}{<p,f>}d\mu, \tag{4.22}$$

and the "capacity" functional

$$M(f_1,...,f_n):=\max_{p\in\Delta}M_p(f_1,...,f_n). \tag{4.23}$$

Finally let p^* be the maximizer of $M_p(f)$.

The reason why we call $M_p(f)$ capacity function is that this quantity is indeed the capacity of a discrete memoryless channel, whose stochastic matrix is given by f_i (See [8]).

Our goal in this subsection is to show that $M(f,g)^{\frac{1}{e}}$ is a metric.(It is known that capacity itself is not a metric.) To this end, we will first prove the following statement. Given any $a,b,c\in R^+$ with $a<c<b$, and writing $m(p;a,b):=pa\log a+qb\log b-(pa+qb)\log(pa+qb)$, we have

$$(m(p;a,b))^r\leq(m(p;a,c))^r+(m(p;c,b))^r,\ \text{ with } r:=\min(p,q),q:=1-p. \tag{4.24}$$

Now we examine the function $g_r:\mathbf{R}^+\setminus\{1\}$. This is related to the derivative of $m(p;a,c)^r$, that is, for $0<r\leq\frac{1}{2}$,

$$g_r(p;x):=\left(\frac{\partial m(p;a,c)^r}{\partial c}\right)_{a=x,c=1}=rq\log\frac{1}{q+px}\left(q\log\frac{1}{q+px}+px\log\frac{x}{q+px}\right)^{r-1}. \tag{4.25}$$

The proofs of these two lemmas are quite lengthy, so they are given in the appendix.

Lemma 4.6 (Behavior of $g_r(x)$).

1. *$g_r(p;x)$ has only one discontinuity at $x=1$.*
$$lim_{x\to 1^\mp}g_r(p;x)=\begin{cases}\pm\frac{\sqrt{pq}}{\sqrt{2}}, & if\ r=\frac{1}{2}.\\ \pm\infty, & if\ r\leq\frac{1}{2}.\end{cases}$$
2. *For $r=\min(p,q)$, the derivative $\frac{d}{dx}g_r(p;x)$ is positive for $x\in\mathbf{R}^+\setminus\{1\}$, thus $g_r(p;x)$ increases monotonically.*

Lemma 4.7. *For positive numbers a,b,c, and $0<p<1$, $p+q=1$, let $r(p)=\min(p,q)$. Then we have*

$$m(p;a,b)^r\leq m(p;a,c)^r+m(p;c,b)^r. \tag{4.26}$$

Since the maximizer p^* always lies between $\frac{1}{e}$ and $1-\frac{1}{e}$, we have

$$m(p;a,b)^{\frac{1}{e}}\leq m(p;a,c)^{\frac{1}{e}}+m(p;c,b)^{\frac{1}{e}}. \tag{4.27}$$

Theorem 4.8. $M(g,k)^{\frac{1}{e}}\leq M(g,h)^{\frac{1}{e}}+M(h,k)^{\frac{1}{e}}$.

Proof. By Remark 4.3, we have $\frac{1}{e}\leq p\leq 1-\frac{1}{e}$, and using Lemma 4.7 and Minkowski inequality, we obtain this result. □

4.3. Min-max is distance. The previous section, in particular implies that the square root of mini-max IS distance of positive numbers is a metric, since $-\log x \in \mathbb{F}$. Here, we will show that it is also a metric in the vector case. Note that the IS divergence containing a vector with any zero component is infinity, and then obviously the triangle inequality holds. Let $f,g,l,a \in \mathbf{R^n}$ have positive components. Subscript $(\cdot)_i$ will denote their $i-$th component. Denote $M_F(f,g) := \sum_{i=1}^n m_F(f_i,g_i)a_i$, $F(x) = -\log x$. In order to verify the triangle inequality for functions,

$$\sqrt{M_F(f,g)} \leq \sqrt{M_F(f,l)} + \sqrt{M_F(g,l)}, \tag{4.28}$$

our strategy will be to show that it holds even in the worst case, namely, the case where the LHS is maximized with RHS fixed. Note that it is easy to see that $m_F(x,y) = m_F(x/y,1)$ for two numbers $x>0, y>0$, so without loss of generality, we can assume $l_i := 1$ for all i. We will show that in the worst case f,g are both parallel to $E := (1,1,...,1)$ so that verifying the triangle inequality is equivalent to proving the triangle inequality for scalars, which was done in the previous section.

Given two fixed positive numbers L_f, L_g, we like to maximize $M_F(f,g)$ among the set $D := \{(f/l, g/l) : M_F(f,l) \leq L_f, M_F(g,l) \leq L_g\}$, here $(f/l)_i := f_i/l_i, (g/l)_i := g_i/l_i$. We form the Lagrangian:

$$M_F(f,g) - \mu_1 M_F(f,l) - \mu_2 M_F(g,l), \tag{4.29}$$

with two Lagrange multipliers $\mu_1 \geq 0, \mu_2 \geq 0$, and

$$M_F(f,g) = \max_{p_1+q_1=1} \sum_{i=1}^n (p_1 F(f_i) + q_1 F(g_i) - F(p_1 f_i + q_1 g_i))a_i, \tag{4.30}$$

$$M_F(f,l) = \max_{p_2+q_2=1} \sum_{i=1}^n (p_2 F(f_i) + q_2 F(l_i) - F(p_2 f_i + q_2 l_i))a_i, \tag{4.31}$$

$$M_F(g,l) = \max_{p_3+q_3=1} \sum_{i=1}^n (p_3 F(g_i) + q_3 F(l_i) - F(p_3 g_i + q_3 l_i))a_i. \tag{4.32}$$

Since D is a compact set, there exists a maximizer. Now taking variation with respect to each component of f,g,l, at the maximizer $(f,g,l=E)$, for each i one has

$$p_1(F'(f_i) - F'(p_1 f_i + q_1 g_i)) - \mu_2 q_3 (F'(f_i) - F'(p_3 l_i + q_3 f_i)) = 0, \tag{4.33}$$

$$q_1(F'(g_i) - F'(p_1 f_i + q_1 g_i)) - \mu_1 p_2 (F'(g_i) - F'(p_2 g_i + q_2 l_i)) = 0, \tag{4.34}$$

$$\mu_1 q_2 (F'(l_i) - F'(p_2 g_i + q_2 l_i)) + \mu_2 p_3 (F'(l_i) - F'(p_3 l_i + q_3 f_i)) = 0. \tag{4.35}$$

Denoting $\lambda_1 := \mu_1 q_2/(\mu_2 p_3)$, $\lambda_2 := \mu_1 p_2/q_1$, and $\lambda_3 := \mu_2 q_3/p_1$, we have $\lambda_1 \geq 0, \lambda_2 \geq 0, \lambda_3 \geq 0$, and

$$(F'(f_i) - F'(p_1 f_i + q_1 g_i)) - \lambda_3 (F'(f_i) - F'(p_3 l_i + q_3 f_i)) = 0, \tag{4.36}$$
$$(F'(g_i) - F'(p_1 f_i + q_1 g_i)) - \lambda_2 (F'(g_i) - F'(p_2 g_i + q_2 l_i)) = 0, \tag{4.37}$$
$$\lambda_1 (F'(l_i) - F'(p_2 g_i + q_2 l_i)) + (F'(l_i) - F'(p_3 l_i + q_3 f_i)) = 0. \tag{4.38}$$

LEMMA 4.9. *Now, in this case $F(x)=-\log x$, and there are at most two solutions to the equations (4.36)-(4.38), one of which is $(f_i,g_i)=(l_i,l_i)$. Dropping those identical components, we can assume that the maximizer is $f=\alpha_f E, g=\alpha_g E, l=E$, with two scalars $\alpha_f\neq 1, \alpha_g\neq 1$.*

Proof. Obviously $f_i=g_i=l_i$ is a trivial solution. Suppose $M_F(f,l)\neq 0$, $M_F(g,l)\neq 0$, then there exist other solutions satisfying the system: for each i,

$$\frac{f_i-g_i}{p_1f_i+q_1g_i}-\lambda_3\left(\frac{f_i-l_i}{p_3l_i+q_3f_i}\right)=0, \tag{4.39}$$

$$\frac{g_i-f_i}{p_1f_i+q_1g_i}-\lambda_2\left(\frac{g_i-l_i}{p_2g_i+q_2l_i}\right)=0, \tag{4.40}$$

$$\frac{l_i-f_i}{p_3l_i+q_3f_i}+\lambda_1\left(\frac{l_i-g_i}{p_2g_i+q_2l_i}\right)=0. \tag{4.41}$$

Note that these three equations must be dependent with $\lambda_2/\lambda_1=\lambda_3$, otherwise this system has only the trivial solution. On the other hand, since (p_1,q_1) is the maximizer of $M_F(f,g)$, by equation (2.14), we have

$$\sum_{i=1}^n(\log f_i-\log g_i)a_i=\sum_{i=1}^n\left(\frac{f_i-g_i}{p_1f_i+q_1g_i}\right)a_i. \tag{4.42}$$

Likewise we have similar equations corresponding to $M_F(f,l), M_F(g,l)$. Substituting these in system (4.39)-(4.41), we have

$$\sum_{i=1}^n(\log l_i-\log f_i)a_i+\lambda_1\sum_{i=1}^n(\log l_i-\log g_i)a_i=0, \tag{4.43}$$

and

$$\sum_{i=1}^n(\log g_i-\log f_i)a_i+\lambda_2(\sum_{i=1}^n(\log l_i-\log g_i)a_i=0. \tag{4.44}$$

Using these two equations, and rewriting $\log g_i-\log f_i=(\log l_i-\log f_i)-(\log l_i-\log g_i)$, we can conclude that $\lambda_2=\lambda_1+1$.
Denote $x:=g_i/l_i-1$, $y=f_i/l_i-1$. Then the system (4.39)-(4.41) becomes

$$\begin{cases}\frac{y}{1+q_3y}+\lambda_1\frac{x}{1+p_2x}=0,\\ \frac{y-x}{1+x+p_1(y-x)}+\lambda_2\frac{x}{1+p_2x}=0.\end{cases}$$

Or

$$\begin{cases}y+(p_2+q_3\lambda_1)xy+\lambda_1x=0,\\ y+x(\lambda_2-1)+x^2(\lambda_2-p_2-p_1\lambda_2)+xy(p_1\lambda_2+p_2)=0.\end{cases}$$

Subtracting the first equation from the second, we have

$$\begin{cases}y+(p_2+q_3\lambda_1)xy+\lambda_1x=0,\\ x=0,\text{ or } y(p_1\lambda_2-q_3\lambda_1)+x(q_1\lambda_2-p_2)+(\lambda_2-\lambda_1-1)=0.\end{cases}$$

In the case where $x=0$, we have $(x,y)=(0,0)$, and $f_i=g_i=l_i$, i.e., the trivial solution.

In the second case, using the result $\lambda_2=\lambda_1+1$, we see that this describes a line through the origin. Note that the first equation represents one branch of hyperbolic curve passing through the origin, and a straight line can intersect this curve at most twice. Clearly one of the intersections is (0,0).

Therefore, only two solutions exist for the system (4.36)–(4.38), and one of them is $(f_i,g_i,l_i)=(1,1,1)$. □

THEOREM 4.10. *Given positive vectors $f,g,l,a\in R^n$, denote*

$$IS(f,g):=\max_{p+q=1}\sum_{i=1}^{n}(-p\log(f_i)-q\log(g_i)+\log(pf_i+qg_i))a_i. \tag{4.45}$$

Based on the previous lemma, we have the triangle inequality: $\sqrt{IS(f,g)}\le\sqrt{IS(f,l)}+\sqrt{IS(l,g)}$.

Appendix A. Proof of necessary condition. This necessary condition in fact comes from the leading coefficient of Taylor's expansion.

Consider the special case of three numbers $z-a,z,z+b$, with $a>0,b>0$ and close to zero, where a,b satisfy $y:=B_f(z-a,z)=B_f(z+b,z)=m_f(z-a,z+b)$. Without loss of generality, we may assume $f(z)=0,f(z-a)=f(z+b)=y\ge0$, and $f'(z)=0$. Otherwise, replace $f(x)$ by $f(x)-f(z)-(x-z)f'(z)$. Here we will derive a relation between a and b, of the type $a=k_1b+k_2b^2+O(b^3)$ for some constants k_1,k_2 to be determined. Since $f(z-a)=f(z+b)=y$, taking Taylor's expansion around z, we have

$$\frac{f''}{2}a^2-\frac{f'''}{6}a^3+O(a^4)=\frac{f''}{2}b^2+\frac{f'''}{6}b^3+O(b^4)=y. \tag{A.1}$$

(Hereafter f'',f''',f'''' refer to $f''(z),f'''(z),f''''(z)$.) By comparing the coefficients of b^2,b^3 in equation (A.1), we have $k_1=1$ and $k_2=\frac{f'''}{3f''}$, thus we have $a=b+\frac{f'''}{3f''}b^2+O(b^3)$.

Assume the triangle inequality holds for $z-a,z,z+b$:

$$\sqrt{m_f(z,z+b)}+\sqrt{m_f(z-a,z)}\ge\sqrt{m_f(z-a,z+b)}. \tag{A.2}$$

Now let $z+c,z+d$ be the centers of $m_f(z,z-a)$ and $m_f(z,z+b)$. Then we have

$$-af'(z-c)=f(z-a)-f(z)=f(z-a)=y,\, bf'(z+d)=f(z+b)=y,$$

and the triangle inequality (A.2) can be rewritten as

$$\sqrt{\frac{c}{a}f(z-a)-f(z-c)}+\sqrt{\frac{d}{b}f(z+b)-f(z+d)}\ge\sqrt{y}. \tag{A.3}$$

i.e.,

$$\sqrt{\frac{c}{a}-\frac{f(z-c)}{f(z-a)}}+\sqrt{\frac{d}{b}-\frac{f(z+d)}{f(z+b)}}-1\ge0. \tag{A.4}$$

Next, we write each term of equation (A.4) in terms of the variable b. Consider the Taylor's expansion of $f'(z+d)=\frac{f(z+b)}{b}$ around z. We can express d in terms of b:

$$d=\frac{1}{2}b+\frac{f'''}{24f''}b^2+\frac{1}{48}\left(\frac{f''''}{f''}-\left(\frac{f'''}{f''}\right)^2\right)b^3+O(b^4). \tag{A.5}$$

Moreover,

$$\frac{f(z+d)}{f(z+b)} = \frac{d^2 f'' + d^3 f'''/3 + d^4 f''''/12}{b^2 f'' + b^3 f'''/3 + b^4 f''''/12} + O(\text{ higher order terms})$$
$$= \frac{1}{4} + \frac{b^2}{48}\left(\frac{1}{12}\left(\frac{f'''}{f''}\right)^2 + \left(\frac{f''''}{f''} - \left(\frac{f'''}{f''}\right)^2\right) + \left(\frac{1}{2}\left(\frac{f'''}{f''}\right)^2 - \frac{3}{4}\frac{f''''}{f''}\right) + \right) + O(b^3).$$

Hence,

$$\frac{d}{b} - \frac{f(z+d)}{f(z+b)} = \frac{1}{4} + \frac{f'''}{f''}\left(\frac{b}{24}\right) - \frac{b^2}{48}\left(\frac{1}{12}\left(\frac{f'''}{f''}\right)^2 + \left(\frac{1}{2}\left(\frac{f'''}{f''}\right)^2 - \frac{3}{4}\frac{f''''}{f''}\right)\right) + O(b^3),$$

and

$$\sqrt{\frac{d}{b} - \frac{f(z+d)}{f(z+b)}} = \frac{1}{2} + \frac{f'''}{f''}\left(\frac{b}{24}\right) - \frac{b^2}{48}\left(\frac{2}{3}\left(\frac{f'''}{f''}\right)^2 - \left(\frac{3}{4}\frac{f''''}{f''}\right)\right) + O(b^3).$$

Similarly, we have

$$\sqrt{\frac{c}{a} - \frac{f(z-c)}{f(z-a)}} = \frac{1}{2} - \frac{a}{24}\frac{f'''}{f''} - \frac{a^2}{48}\left(\frac{2}{3}\left(\frac{f'''}{f''}\right)^2 - \left(\frac{3}{4}\frac{f''''}{f''}\right)\right) + O(a^3)$$
$$= \frac{1}{2} - \frac{b}{24}\frac{f'''}{f''} - \frac{b^2}{48}\left(\frac{2}{3}\left(\frac{f'''}{f''}\right)^2 - \left(\frac{3}{4}\frac{f''''}{f''}\right)\right) - \frac{b^2}{72}\left(\frac{f'''}{f''}\right)^2 + O(b^3). \tag{A.6}$$

Now, equation (A.4) becomes

$$-\frac{b^2}{24}\left(\frac{2}{3}\left(\frac{f'''}{f''}\right)^2 - \left(\frac{3}{4}\frac{f''''}{f''}\right)\right) - \left(\frac{f'''}{f''}\right)^2 \frac{b^2}{24\cdot 3}$$
$$= \frac{b^2}{24}\left(-\left(\frac{f'''}{f''}\right)^2 + \left(\frac{3}{4}\frac{f''''}{f''}\right)\right) \geq 0, \text{ as } b \to 0.$$

Thus we arrive at a necessary condition (A.4):

$$-4\left(\frac{f'''}{f''}\right)^2 + 3\frac{f''''}{f''} \geq 0, i.e., \quad 3(\log f'')'' - ((\log f'')')^2 \geq 0. \tag{A.7}$$

Proof of Lemma 4.6. The first statement can be verified by Taylor's Expansion. Since

$$\begin{cases} q\log\frac{1}{px+q} & = -pq(x-1) + \frac{p^2q}{2}(x-1)^2 + \dots \\ px\log\frac{x}{q+px} & = pq(x-1) + \frac{pq^2}{2}(x-1)^2 + \dots, \end{cases}$$

then

$$\lim_{x\to 1} g_r(p;x) = \lim_{x\to 1}\frac{-pqr(x-1)+\dots}{(\frac{pq}{2}(x-1)^2)^{1-r}} = \begin{cases} -\sqrt{\frac{pq}{2}}sgn(x-1), & \text{if } r = \frac{1}{2} \\ -sgn(x-1)\infty, & \text{if } r < \frac{1}{2}. \end{cases}$$

The second statement: differentiating, we have $\frac{dg_r}{dx} = pqrm(p;x,1)^{-2+r}f(x)$. Hence, $\frac{d}{dx}g_r > 0$ if and only if $f > 0$, here

$$f(x) := (-1+r)\log\frac{1}{q+px}\log\frac{x}{q+px} - \frac{q}{q+px}\log\frac{1}{q+px} - \frac{px}{q+px}\log\frac{x}{q+px}. \tag{A.8}$$

Now we observe that the positivity of $f(x)$ ensures the monotonic increasing behavior of $g_r(p;x)$, which then guarantees the triangle inequality.
Differentiating again,

$$f'(x) = \frac{q(q(-1+r)+prx)\log\frac{1}{q+px} - px(qr+p(-1+r)x)\log\frac{x}{q+px}}{x(q+px)^2}. \tag{A.9}$$

First, since $f(1)=0, f'(1)=0$, it suffices to show that if $x>1$, then $f'(x)>0$, and if $x<1$, then $f'(x)<0$.
There are two cases for r. Suppose $p<q$, then $r=p$, and

$$f'(x) = \frac{q((-q^2+p^2x)\log\frac{1}{q+px} - p^2x(1-x)\log\frac{x}{q+px})}{x(q+px)^2}. \tag{A.10}$$

Now to simplify the numerator of $f'(x)$, let $s=\frac{x}{q+px}, t=\frac{1}{q+px}$. Then $ps+qt=1, x=\frac{s}{t}$, and

$$\begin{aligned}
&\frac{(-q^2+p^2x)\log\frac{1}{q+px} - p^2x(1-x)\log\frac{x}{q+px}}{q+px}\\
&= t\left(-q^2+p^2\frac{s}{t}\right)\log t - p^2\left(1-\frac{s}{t}\right)s\log s\\
&= \left(\frac{1}{t}-1\right)(ps\log s+qt\log t)+(p-q)\log t\\
&= \begin{cases}\geq 0, \text{ if } x>1, \text{ i.e., } t<1<s.\\ \leq 0, \text{ if } x<1, \text{ i.e., } t>1>s.\end{cases}
\end{aligned}$$

Suppose $p>q$, then $r=q$, and

$$f'(x) = \frac{q(-pq+pqx)\log\frac{1}{q+px} - px(q^2-p^2x)\log\frac{x}{q+px}}{x(q+px)^2}. \tag{A.11}$$

Using the same substitution, we have

$$\begin{aligned}
&-\frac{q^2(1-x)\log\frac{1}{q+px} + (q^2-p^2x)x\log\frac{x}{q+px}}{q+px}\\
&= -q^2\left(1-\frac{s}{t}\right)t\log t - \left(q^2-\frac{p^2s}{t}\right)s\log s\\
&= -q\left(1-\frac{s}{t}\right)(ps\log s+qt\log t)+\frac{(p-q)}{t}s\log s\\
&= \begin{cases}\geq 0, \text{ if } x>1, \text{ i.e., } t<1<s.\\ \leq 0, \text{ if } x<1, \text{ i.e., } t>1>s.\end{cases}
\end{aligned}$$

In the above discussion, we have shown that $f'(x)>0$ if $x>1$, and $f'(x)<0$ if $x<1$. Hence, from $f(1)=0, f'(1)=0$, we conclude that $f(x)\geq 0$, and $g(x)$ increases monotonically except for the jump at $x=1$.

Proof of Lemma 4.7. Without loss of generality, assume $a<b$. Since the triangle inequality holds in the cases where $a<b<c$ and $c<a<b$, we only need to show

$$m(p;a,b)^r \leq m(p;a,c)^r + m(p;c,b)^r, \text{where } a<c<b. \tag{A.12}$$

To this end, we claim the right-hand side is concave in c. First, differentiate the right-hand side (denoted by RHS) with respect to c. We claim that its derivative decreases as c increases.
Indeed,

$$\frac{\partial}{\partial c}RHS = \frac{rq\log\frac{c}{ap+cq}}{m(p;a,c)^{1-r}} + \frac{rp\log\frac{c}{cp+bq}}{m(p;c,b)^{1-r}} = \frac{rq\log\frac{1}{\frac{ap}{c}+q}}{(cm(p;\frac{a}{c},1))^{1-r}} + \frac{rp\log\frac{1}{p+\frac{bq}{c}}}{(cm(q;\frac{b}{c},1))^{1-r}}, \quad \text{(A.13)}$$

i.e.,

$$c^{1-r}\frac{\partial RHS}{\partial c} = \frac{rq\log\frac{1}{\frac{ap}{c}+q}}{(m(p;\frac{a}{c},1))^{1-r}} + \frac{rp\log\frac{1}{\frac{bq}{c}+p}}{(m(q;\frac{b}{c},1))^{1-r}}. \quad \text{(A.14)}$$

Using the definition of g in the previous lemma,

$$g_r(p;x) := rq\log\frac{1}{q+px}\left(q\log\frac{1}{q+px} + px\log\frac{x}{q+px}\right)^{r-1}, \quad \text{(A.15)}$$

and denoting $x := a/c, x<1, \beta := \frac{b}{a} > 1$, (then $\beta\cdot x = b/c$, $x\in(\frac{1}{\beta},1), \beta x\in(1,\beta)$), then we have that

$$c^{1-r}\frac{\partial RHS}{\partial c} = g_r(p;x) + g_r(q;\beta x). \quad \text{(A.16)}$$

Since $g_r(p;x)$ is nonzero with a jump at $x=1$, and its derivative is nonnegative. If $r<\frac{1}{2}$, then

$$\begin{cases} RHS\to+\infty & \text{as } x\to 1^-, \\ RHS\to-\infty & \text{as } x\to \frac{1}{\beta}^+, \end{cases}$$

and if $r=\frac{1}{2}$, then

$$\begin{cases} RHS\to+\sqrt{\frac{pq}{2}} & \text{as } x\to 1^-, \\ RHS\to-\sqrt{\frac{pq}{2}} & \text{as } x\to \frac{1}{\beta}^+. \end{cases}$$

By Lemma 4.6 we know that $g_r(p;x)+g_r(q;\beta x)$ is monotonically increasing as x increases from $\frac{1}{\beta}$ to 1, thus it has at most one sign change between $x=\frac{1}{\beta}$ and $x=1$.

Note as c decreases, x increases. When the sign change happens, the derivative $c^{1-r}\frac{\partial RHS}{\partial c}$ goes from negative to positive as x goes from $\frac{1}{\beta}$ to 1, (i.e., c decreases from b to a). Also $\frac{\partial RHS}{\partial c}<0$ at $c=b$, and $\frac{\partial RHS}{\partial c}>0$ at $c=a$, which implies that there is one and only one sign change in $c\in[a,b]$, which is a maximizer, not a minimizer. Thus, RHS has two local minima at $c=a$ and $c=b$.

Finally, when $c=a$, or $c=b$, then $m(p;a,c)^r+m(p;c,b)^r$ becomes $m(p;a,b)^r$. So we have established the inequality.

By Lemma 2.3, clearly we have the second statement.

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

[13] 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.

[14] Y. Linde, A. Buzo and R.M. Gray, *An algorithm for vector quantizer design*, IEEE Trans. on Comm., 28(1), 84-95, 1980.

[15] J. Lin, *Divergence measures based on Shannon entropy*, IEEE Trans. on Information Theory, 37(1), 145-151, 1991.

[16] P.C. Mahalanobis, *On the generalized distance in statistics*, Proceedings of National Institute of Science of India, 12, 49-55, 1936.

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

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

[19] P. Chen, Y. Chen and M. Rao, *Metrics defined by Bregman divergences*, Comm. Math. Sci., 6, 915-926, 2008.