On conformal divergences and their population minimizers

Richard Nock, Frank Nielsen, Shun-ichi Amari

Abstract

Total Bregman divergences are recent regularizations of Bregman divergences that are invariant to rotations of the natural space, a very desirable property in many signal processing applications. Clustering with total Bregman divergences often display superior results compared to ordinary Bregman divergences (Liu, 2011), and raises, for a wide applicability in clustering, the question of the complete characterization of the left and right population minimizers — total Bregman divergences are indeed not symmetric.

In this paper, we focus on the left and right population minimizers for conformal divergences, a super family of total Bregman divergences with interesting geometric features. Our results are obtained in Amari’s recently coined $(u, v)$-geometric structure, a generalization of dually flat affine connections. We characterize, analytically and geometrically, the population minimizers. We also show that conformal divergences (resp. total Bregman divergences) are essentially exhaustive for their left (resp. right) population minimizers. We provide new results and extend previous results on the robustness to outliers of the left and right population minimizers, and discuss further the $(u, v)$ geometric structure relation in the context of clustering. Additional results are also presented.

Index Terms

Bregman divergences, total Bregman divergences, $(u, v)$-geometric structure.

I. INTRODUCTION

Bregman divergences are probably the most widely used distortion measure in clustering algorithms. Popular algorithms like $k$-means and Expectation Maximization can be formalized as computing quantities that are minimizers of the average (sample-wise) divergence to a sample [1], [2], [3]; we call these quantities population minimizers. Left and right population minimizers are respectively the sample $f$-mean and the sample average (Bregman divergences are in general not symmetric). It has been shown that modulo technical assumptions, Bregman divergences are exhaustive for their right population minimizer: any divergence whose population minimizer is the sample average is a Bregman divergence [4], [5]. This shows that the scope of $k$-means and EM is wide and encompasses all domains whose “natural” distortion measures rely on Bregman divergences, such as signal processing, Euclidean geometry, information theory, statistics, etc. .

R. Nock is with CEREMIA-Université Antilles-Guyane, France. E-mail:rnock@martinique.univ-ag.fr.
F. Nielsen is with Sony Computer Science Laboratories, Inc, Tokyo, Japan. E-mail:nielsen@csl.sony.co.jp.
S.-i. Amari is with the Riken Brain Science Institute, Japan. E-mail:amari@brain.riken.jp.
There has been a recent burst of interest in new classes of divergences, built from Bregman divergences, known as total Bregman divergences [6], [7], [8], [9], [10], [11], [12]. These divergences are invariant to particular transformations of the natural space. Experimentally speaking, clustering with their left population minimizers yields significantly improved results compared to (ordinary) Bregman divergences [6]. These results exploit the fact that the left population minimizers of total Bregman divergences are weighted generalized $f$-means [6]. In the general context of clustering, and also for reasons related to statistics and maximum likelihood estimation [13], it is important to characterize further the population minimizers of total Bregman divergences: important questions include the characterization of their right population minimizers and the exhaustivity of these divergences for their population minimizers.

In this paper, we address these questions in a setting which generalizes in two ways total Bregman divergences. First, we consider a superclass of total Bregman divergences that we define as conformal divergences, which is a strict generalization of total Bregman divergences and ordinary Bregman divergences. Second, we consider a coordinate system which is not the usual dually flat affine coordinate system of (total, ordinary) Bregman divergences, but a recent generalization in information geometry due to Amari, known as the $(u,v)$-geometric structure [14], [15], in which two coordinate mappings $u$ and $v$ define the gradient of the generator of the Bregman divergence.

In this generalized setting, our main contribution includes:

- the characterization of the right population minimizers for total Bregman divergences;
- the characterization of the right population minimizer for an interesting $L_p$ generalization of total Bregman divergences;
- a proof that conformal divergences are exhaustive for their left population minimizers;
- a proof that total Bregman divergences are exhaustive for their right population minimizers;
- the robustness analysis of the left and right population minimizers for conformal divergences, which generalizes results known for total Bregman divergences [6].

Our contribution also includes results pertinent for clustering, such as (i) a proof that the $(u,v)$-geometric structure sometimes describe an equivalence relation which might be useful in the context of clustering; (ii) a proof that the square loss in $v$-coordinates is the only 1D symmetric conformal divergence in the $(u,v)$ geometric structure; (iii) a discussion on population minimizers for a further extension involving the recently coined scaled Bregman divergences (that generalize Csiszár’s $f$-divergences) [13].

The paper is structured as follows. The following Section gives definitions. Section III is devoted to left population minimizers of conformal divergences in the $(u,v)$-geometric structure. Section IV studies right population minimizers for conformal divergences, and Section V extends the results to the $(u,v)$ geometric structure. Section VI studies the robustness of the population minimizers and Section VII discusses our results. A last section concludes.

In order not to laden the paper’s body, some proofs have been postponed to an appendix in Section X.
II. Definitions

Throughout this paper, bold faces denote column vectors, such as $\mathbf{0}$ for the null vector, while capitals, like $J$ or $H$ (respectively Jacobian and Hessian) denote matrices. Coordinates are noted in exponent, such as $x^1, x^2, \ldots, x^d$ for vector $x \in \mathbb{R}^d$, where $d \geq 1$.

A (right-sided) conformal divergence, $D_{\varphi, g}$, is parameterized by two real-valued functions $\varphi$ and $g$ with $\text{Im} g \subseteq \mathbb{R}_+$, whose domains are compact convex of $\mathbb{R}^d$. The expression of $D_{\varphi, g}$ is:

$$D_{\varphi, g}(x : y) = g(y) D_{\varphi}(x : y). \quad (1)$$

$\varphi$ is real-valued strictly convex twice differentiable, and $D_{\varphi}(x : y)$ is the Bregman divergence with generator $\varphi$:

$$D_{\varphi}(x : y) = \varphi(x) - \varphi(y) - (x - y)^\top \nabla \varphi(y) \quad ; \quad (2)$$

$\nabla \varphi$ denotes the gradient of $\varphi$. $g$ admits continuous directional derivatives: function $D_{t, z} g(x)$.

$$= \lim_{t \to 0} \frac{g(x + tz) - g(x)}{t}$$

Bregman divergences match the subset of conformal divergences for which $g(.) = K$, a constant. The most popular recent example of conformal divergences is obtained for $g = g_{\perp}$:

$$g_{\perp}(y) = \sqrt{1 + \|\nabla \varphi(y)\|^2} \quad ; \quad (3)$$

which defines total Bregman divergences, that are invariant to rotations of the coordinate axes [7], [8], [9], [16], [10], [11], [12] (among others). Table I presents some examples of total Bregman divergences. Remark that $g_{\perp}(y)$ is of the form $f_{\perp}(\nabla \varphi(y))$, with

$$f_{\perp}(x) = \frac{1}{\sqrt{1 + \|x\|^2}} \quad ; \quad (4)$$

Figure I depicts $D_{\varphi}(x : y)$ and $D_{\varphi, g_{\perp}}(x : y)$ on a simple example.

We also investigate the generalization of (3) to $p$-norms, and define, $\forall p \geq 1$:

$$g_p(y) = f_p(\nabla \varphi(y)) \quad ; \quad (5)$$

$$f_p(x) = \frac{1}{(1 + \|x\|^p)^{\frac{1}{p}}} \quad . \quad (6)$$

and the $p$-norm of $x$ is $\|x\|_p = (\sum_i |x^i|^p)^{\frac{1}{p}}$.

A coordinate mapping $v$ is a $C^1$, bijective function $v : \mathbb{R}^d \to \mathbb{R}^d$. For any coordinate mapping $v$, we define the $v$-conformal divergence $D_{\varphi, g}^v$ as:

$$D_{\varphi, g}^v(x : y) = g(y) D_{\varphi}(v(x) : v(y)) \quad . \quad (7)$$

$v$-conformal divergences are inspired by divergences in the $(u, v)$ geometric structure [14], [15] (see also Section VII). They generalize conformal divergences for which $v = \text{Id}$. We shall investigate several interesting cases of
$v$-conformal divergences, including those where $g$ is a function of $\nabla \varphi$, and those where $g$ is a function of coordinate mapping $u$ in the $(u, v)$ geometric structure.

Let us now motivate the $(u, v)$ geometric structure in the context of the dual coordinate systems of Bregman divergences [17]. Function $g$ in $v$-conformal divergences depends on the right parameter of the divergence. We shall see (Lemma [10]) that when $g$ is not constant, the $v$-conformal divergence cannot be symmetric: $g(y)D_{\varphi}(v(x) : v(y)) \neq g(y)D_{\varphi}(v(y) : v(x))$. However, our results extend at no cost to left-sided conformal divergences, i.e. whose regularization $g$ depends on the left parameter of the divergence. Indeed, calling to convex conjugates, we obtain:

$$D_{\varphi,g}^v(x : y) = g(y)D_{\varphi}(v(x) : v(y))$$
$$= g(y)D_{\varphi^*}((\nabla \varphi \circ v)(y) : (\nabla \varphi \circ v)(x))$$
$$= g(y)D_{\varphi^*}(u(y) : u(x)),$$

where $u = \nabla \varphi \circ v$ also defines a coordinate mapping and $\varphi^*$ is the convex conjugate of $\varphi$. Any such coordinate mappings $u$ and $v$ such that $u \circ v^{-1}$ defines the gradient of a strictly convex differentiable function $\varphi$ is called an $(u, v)$ geometric structure [14], that we may also write $(u, v)_\varphi$ to make explicit the reference to $\varphi$.

We now define population minimizers for conformal divergences.

**Definition 1:** (left- and right-population minimizers) Let $S = \{x_1, x_2, ..., x_n\}$, with $x_i \in \mathbb{R}^d, \forall i = 1, 2, ..., n$. Let
$D : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be the shorthand for some Bregman (resp. conformal, resp. $v$-conformal) divergence $D_\varphi$ (resp. $D_{\varphi,g}$, resp. $D_{\varphi,g}^v$). A left population minimizer for $F$ on $\mathcal{S}$ is any $\mu$ such that $\sum_i D(\mu : x_i) = \min_x \sum_i D(x : x_i)$. A right population minimizer for $D$ on $\mathcal{S}$ is any $\mu$ such that $\sum_i D(x_i : \mu) = \min_x \sum_i D(x_i : x)$.

This definition, as well as the results in this paper, can be extended to non uniform distributions over $\mathcal{S}$, and to population minimizers in the continuous case.

III. LEFT POPULATION MINIMIZERS OF $v$-CONFORMAL DIVERGENCES

We are interested in this Section in characterizing the left population minimizers of general $v$-conformal divergences. We build on results known from [16], [10] for the elicitation of the left population minimizer when $v = \text{Id}$, and the well known results from [5] for the elicitation of the divergences having the arithmetic average as right population minimizer. Technicalities are simpler than for the right population minimizers because function $g$ does not depend on the left parameter of $D_\varphi$. We first show that the left population minimizer of some $v$-conformal divergence $D_{\varphi,g}^v$ is a weighted $u$-mean, where $(u, v)$ is a geometric structure.

**Lemma 1:** The left population minimizer $\mu$ of any $v$-conformal divergence $D_{\varphi,g}^v$ on $\mathcal{S}$ is unique and equals:

$$\mu = u^{-1} \left( \frac{1}{\sum_i g(x_i)} \sum_i g(x_i) u(x_i) \right),$$

where $(u, v)_\varphi$ is a geometric structure.

**Proof** Any left population minimizer $\mu$ satisfies $\nabla \sum_i D_{\varphi,g}^v(\mu : x_i) = 0$, and so, after simplification, we obtain:

$$J_v(\mu) \sum_i g(x_i) (\nabla \varphi (v(\mu)) - \nabla \varphi (v(x_i))) = 0,$$

where $J_v$ is the Jacobian of $v$. Since $v$ is bijective, the null space of $J_v$ is reduced to $\{0\}$, and so the sigma in (9) must be the null vector. After simplification, we obtain that:

$$\mu = v^{-1} \nabla \varphi^{-1} \left( \frac{1}{\sum_i g(x_i)} \sum_i g(x_i) \nabla \varphi (v(x_i)) \right).$$

There remains to use the $u$-coordinate mapping to obtain (8).

We now show that the characterization of left population minimizers for $v$-conformal divergences is exhaustive, as any distortion function admitting a weighted $u$-mean as left population minimizer equals a $v$-conformal divergence $D_{\varphi,g}^v$, for some $(u, v)_\varphi$ geometric structure.

**Lemma 2:** Let $\mu = u^{-1} (\sum_i w_i u(x_i))$ be the unique solution to $\min_x \sum_i D(x : x_i)$, where:

1) $D : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is non-negative, twice continuously differentiable and such that $D(x : x) = 0, \forall x$;
2) $u : \mathbb{R}^d \to \mathbb{R}^d$ is a coordinate mapping;
3) $\sum_i w_i = 1$ and $w_i > 0, \forall i$.

Then there exist a function $g : \mathbb{R}^d \to \mathbb{R}$ admitting continuous directional derivatives and a geometric structure $(u, v)_\varphi$ such that

$$D(x : y) = D_{\varphi,g}^v(x : y).$$

(11)
Proof Let \( x_u \doteq u(x), \forall x \) and let \( S_u \doteq \{(x_i)_u, i = 1, 2, \ldots, n\}. \) We have \( \mu_u = \sum_i w_i (x_i)_u \) and
\[
D(\mu : x_i) = D(u^{-1}(\mu_u) : u^{-1}((x_i)_u)) \\
= D_2(\mu_u : (x_i)_u). \tag{12}
\]
From (12), the assumptions on \( D \) and the properties of \( u \), it follows that \( D_2 \) is non-negative, differentiable, satisfies \( D_2(x : x) = 0, \forall x \), and its population minimizer over \( S_u \) is the weighted arithmetic average \( \mu_u \): it is thus a Bregman divergence \([4]\) with:
\[
D_2(\mu_u : (x_i)_u) = w_i D_\phi((x_i)_u : \mu_u), \tag{13}
\]
for some strictly convex differentiable \( \phi \). Calling to convex conjugates, we obtain:
\[
D_\phi((x_i)_u : \mu_u) = D_{\phi^*}(\nabla \phi(\mu_u) : \nabla \phi((x_i)_u)) \\
= D_{\phi^*}((\nabla \phi \circ u)(\mu) : (\nabla \phi \circ u)(x_i)) \\
= D_\phi(v(\mu) : v(x_i)), \tag{14}
\]
with \( \phi \doteq \phi^* \) and \( (v, u)_{\phi} \) geometric structure. \( (u, v)_{\phi} \) is thus a geometric structure and merging (12 — 14), we obtain:
\[
D(\mu : x_i) = w_i D_\phi((x_i)_u : \mu_u) \\
= g(x_i) D_\phi(v(\mu) : v(x_i)), \tag{15}
\]
for some \( g \) admitting continuous directional derivatives which meets \( g(x_i) = w_i, \forall i = 1, 2, \ldots, n \) (we can pick e.g. a degree-\( n \) polynomial). We obtain (11), as claimed.

**IV. RIGHT POPULATION MINIMIZERS FOR CONFORMAL DIVERGENCES**

We now derive the right population minimizers for a general conformal divergence \( D_{\phi,g} \), thus considering \( v \)-conformal divergences with \( v = \text{Id} \). Because \( g \) admits continuous directional derivatives, so does \( D_{\phi,g}(x : y) \) for its both arguments. Let us define:
\[
D_{t,z}D_{\phi,g}(x : y) \doteq \frac{D_{\phi,g}(x : y + tz) - D_{\phi,g}(x : y)}{t}, \tag{16}
\]
so that the directional derivative in the right parameter \( D_zD_{\phi,g}(x : y) \doteq \lim_{t \to 0} D_{t,z}D_{\phi,g}(x : y) \) exists, for any valid direction \( z \). Define from any \( S \) the following averages:
\[
\varphi \doteq \frac{1}{n} \sum_i \varphi(x_i), \tag{17}
\]
\[
\pi \doteq \frac{1}{n} \sum_i x_i. \tag{18}
\]
Let us define the following vectors $\mathbf{x}^+, \mu^+, \delta^+, z^+ \in \mathbb{R}^{d+1}$:

\[
\mathbf{x}^+ = \begin{bmatrix} \mathbf{x} \\ \varphi \end{bmatrix},
\]

\[
\mu^+ = \begin{bmatrix} \mu \\ \varphi(\mu) \end{bmatrix},
\]

\[
\delta^+ = \mathbf{x}^+ - \mu^+,
\]

\[
z^+ = \begin{bmatrix} D_z(g(\mu) \nabla \varphi(\mu)) \\ -D_z g(\mu) \end{bmatrix},
\]

from which we define the following sets:

\[
\mathbb{P}_{S,\varphi,g} = \{ \mu \in \text{dom}(D_{\varphi,g}) : \delta^+ \perp z^+, \forall z \text{ valid} \},
\]

and $\mathbb{B}_{\varphi,g}$, the eventually empty set of non-differentiable boundary points of the intersection of the domains of $\varphi$ and $g$. In the following Lemmata, we let $\mathcal{P}(D_{\varphi,g}; S)$ denote the set of right population minimizers for conformal divergence $D_{\varphi,g}$ on set $S$.

**Lemma 3:** $\mathcal{P}(D_{\varphi,g}; S) \subseteq \mathbb{P}_{S,\varphi,g} \cup \mathbb{B}_{\varphi,g}$. Furthermore, $\forall \mu \in \mathbb{P}_{S,\varphi,g} \setminus \{x\}$, the average distortion is a weighted Mahalanobis distortion to the population average:

\[
\frac{1}{n} \sum_i D_{\varphi,g}(x_i : \mu) = \rho_g \times (\mathbf{x} - \mu)^T H_{\varphi}(\mu) (\mathbf{x} - \mu),
\]

where $\rho_g \equiv g^2(\mu)/D(\mathbf{x} - \mu)g(\mu) > 0$.

**Proof** The first part of the proof is standard, and shows that $\mathbb{P}_{S,\varphi,g}$ is the set of critical points for the right parameter $[18]$. Assume $\mu$ is a population minimizer, and define $D_{t,z} D_{\varphi,g}(S : y) = \sum_i D_{t,z} D_{\varphi,g}(x_i : y)$. Fix any valid direction $z$. Because $\mu$ is a right population minimizer, it comes $D_{t,z} D_{\varphi,g}(S : \mu) \leq 0$ for $t \leq 0$, and $D_{t,z} D_{\varphi,g}(S : \mu) \geq 0$ for $t \geq 0$. Since directional derivatives are defined in direction $z$, we obtain $0 \leq \lim_{t \rightarrow 0} D_{t,z} D_{\varphi,g}(S : \mu) = D_z D_{\varphi,g}(S : \mu) = \lim_{t \rightarrow 0} D_{t,z} D_{\varphi,g}(S : \mu) \leq 0$ and $\mu$ is a solution of:

\[
\lim_{t \rightarrow 0} D_{t,z} D_{\varphi,g}(S : \mu) = D_z \sum_i D_{\varphi,g}(x_i : \mu) = 0.
\]

We plug in [25] the expression of $D_{\varphi,g}$ and obtain that for any right population minimizer $\mu$, the following holds:

\[
\frac{1}{n} D_z \sum_i D_{\varphi,g}(x_i : \mu) = D_z g(\mu) \times \frac{1}{n} \sum_i D_{\varphi}(x_i : \mu) + g(\mu)(\mu - \mathbf{x})^T H_{\varphi}(\mu) z = 0.
\]
Rewriting, we thus need:
\[
D_zg(\mu) \times (\overline{\varphi} - \varphi(\mu))
\]
\[
= (\overline{\varphi} - \mu)^\top (g(\mu)H\varphi(\mu)z + D_zg(\mu)\nabla \varphi(\mu))
\]
\[
= (\overline{\varphi} - \mu)^\top D_z(g(\mu)\nabla \varphi(\mu)) ,
\]
which implies \( \mu \in \mathbb{P}_{S,\varphi,g} \). If a population minimizer does not belong to \( \mathbb{P}_{S,\varphi,g} \), it is in the non differentiable part of the boundary, that is, in \( \mathbb{B}_{\varphi,g} \). Eq. (27) brings:
\[
D_zg(\mu) \times \frac{1}{n} \sum_i D_{\varphi}(x_i : \mu) = g(\mu) \times (\overline{\varphi} - \mu)^\top H\varphi(\mu)z ,
\]
and so:
\[
\frac{1}{n} \sum_i D_{\varphi,g}(x_i : \mu) = \frac{(\overline{\varphi} - \mu)^\top H\varphi(\mu)z}{D_zg(\mu)} g^2(\mu) ,
\]
a quantity which does not depend on the direction \( z \neq 0 \). Fixing \( z = \overline{\varphi} - \mu \) yields the statement of (24). □

We now investigate a particular relevant case where \( g \) depends on the gradient of \( \varphi \).

**Lemma 4:** Suppose \( g(x) = f(\nabla \varphi(x)) \), with \( \varphi \) strictly convex twice differentiable and \( f \) differentiable. Then \( \mathbb{P}_{S,\varphi,g} \) in (23) is defined with \( \delta^+ \) as in (21) and:
\[
z^+ = \left[ f(\nabla \varphi(\mu)) \times z + \nabla f(\nabla \varphi(\mu))^\top z \times \nabla \varphi(\mu) \right],
\]
for any \( z \). We also have \( \forall \mu \in \mathbb{P}_{S,\varphi,g} \):
\[
\frac{1}{n} \sum_i D_{\varphi,g}(x_i : \mu) = \frac{(\overline{\varphi} - \mu)^\top \nabla f(\nabla \varphi(\mu))}{\| \nabla f(\nabla \varphi(\mu)) \|_2^2} f^2(\nabla \varphi(\mu)) .
\]

**Proof** The chain rule gives \( D_zg(\mu) = D_z(f(\nabla \varphi(\mu))) = D_{\varphi,z}f(\nabla \varphi) = D_{H\varphi(\mu)z}f(\nabla \varphi) = z^\top (H\varphi(\mu))^\top \nabla f(\nabla \varphi(\mu)) = z^\top H\varphi(\mu)\nabla f(\nabla \varphi(\mu)) \), so that (27) becomes:
\[
z^\top H\varphi(\mu) ((\overline{\varphi} - \varphi(\mu)) \times \nabla f(\nabla \varphi(\mu)))
\]
\[
= z^\top H\varphi(\mu) (g(\mu) \times (\overline{\varphi} - \mu))
\]
\[
+ z^\top H\varphi(\mu) (\nabla \varphi(\mu)^\top (\overline{\varphi} - \mu) \times \nabla f(\nabla \varphi(\mu)))
\]
\[
= z^\top H\varphi(\mu) \left( g(\mu) \times (\overline{\varphi} - \mu) \ight.
\]
\[
\left.+ \nabla \varphi(\mu)^\top (\overline{\varphi} - \mu) \times \nabla f(\nabla \varphi(\mu)) \right) .
\]
Eq. (30) is of the form \( z^\top H\varphi(\mu) a = z^\top H\varphi(\mu) b \) which implies \( a = b \) as otherwise picking \( z = b - a \neq 0 \) would contradict the positive definiteness of the Hessian \( H\varphi \). After reordering, we get:
\[
\overline{\varphi} - \mu = \frac{1}{f(\nabla \varphi(\mu))} \left( \frac{1}{n} \sum_i D_{\varphi}(x_i : \mu) \right)
\]
\[
\times \nabla f(\nabla \varphi(\mu)) .
\]
(31)
This is a vector equality. Making the inner product with some \( z \in \mathbb{R}^d \) and reordering yields \((\varphi - \varphi(\mu)) \times \nabla f(\nabla \varphi(\mu)) \cdot z = 0\), and so \((\delta^+)^\top z^+ = 0\), which is the statement of the Lemma. There remains to make the inner product of both sides of (31) with \((1/\|\nabla f(\nabla \varphi(\mu))\|_2^2)\nabla f(\nabla \varphi(\mu))\) and reorganize to obtain the average divergence in the Lemma’s statement.

Lemma 5: Consider \( g = Kg_\perp \) for some constant \( K > 0 \). The following holds true:

\[ \delta^+ \perp \begin{bmatrix} \nabla \varphi(\mu) \\ \|\nabla \varphi(\mu)\|_2^2 \end{bmatrix}, \forall \mu \in \mathbb{P}_{S,\varphi,g_\perp}; \tag{32} \]

\( i.e., \) the orthogonal projection of \((\overline{x}, \varphi)\) on the tangent hyperplane \( T_\varphi(\mu) \) to \( \varphi \) at \( \mu \) is point \((\mu, \varphi(\mu))\). Furthermore, assuming \( B_{\varphi,g} = \emptyset \), we have:

\[ \mathcal{P}(D_{\varphi,Kg_\perp};S) = \arg\min_{\mu} \|\overline{x}^+ - \mu^+\|_2, \tag{33} \]

with \( \overline{x}^+ \) and \( \mu^+ \) defined in (19) and (20).

Remark: Figure 2 displays how to find \( \mu \) which meets condition (32). Notice that, by construction, the right population minimizer for \( D_{\varphi,g_\perp} \) is invariant by rotation of the axes. Figure 3 depicts the construction of the population minimizer in a simple 1D case.

Proof (of Lemma 5) We apply Lemma 4 with \( g = Kg_\perp \in (3) \). We have \( f(x) = Kf_\perp(x) = K/\sqrt{1 + \|x\|_2^2} \), and so

\[ \nabla f(x) = -\frac{K}{(1 + \|x\|_2^2)^{3/2}} \times x \tag{34} \]

\( \forall \mu \in \mathbb{P}_{S,\varphi,g_\perp}, z^+ \) in (29) is:

\[ z^+ = \frac{K}{(1 + \|\nabla \varphi(\mu)\|_2^2)^{3/2}} \times \left[ \frac{1 + \|\nabla \varphi(\mu)\|_2^2}{\|\nabla \varphi(\mu)\|_2^2} \times z - \nabla \varphi(\mu)^\top z \times \nabla \varphi(\mu) \right], \]

Picking \( z = \nabla \varphi(\mu) \) brings \( z^+ = \nabla \varphi(\mu)^+ \) with:

\[ \nabla \varphi(\mu)^+ = \frac{K}{(1 + \|\nabla \varphi(\mu)\|_2^2)^{3/2}} \begin{bmatrix} \nabla \varphi(\mu) \\ \nabla \varphi(\mu)^\top \nabla \varphi(\mu) \end{bmatrix} \times \begin{bmatrix} \nabla \varphi(\mu) \\ \|\nabla \varphi(\mu)\|_2^2 \end{bmatrix}, \tag{35} \]

and we thus get (32) and the interpretation of (32).
To prove (33), we need two identities. The first is obtained by making the inner product of (31) with \((x - \mu)\) and using (34), which yields:

\[
\|x - \mu\|_2^2 = -\frac{(x - \mu)\top \nabla \varphi(\mu)}{1 + \|\nabla \varphi(\mu)\|_2^2} \times \frac{1}{n} \sum_i D_\varphi(x_i : \mu) .
\] (36)

To state the second, we start from eq. (32), which states equivalently:

\[
\varphi - \varphi(\mu) = -\frac{(x - \mu)\top \nabla \varphi(\mu)}{\|\nabla \varphi(\mu)\|_2^2} ,
\] (37)

and after plugging in \(f = f_\perp\), the average divergence in Lemma 4 yields:

\[
\frac{1}{n \sum_i D_\varphi(x_i : \mu)} = -\frac{(x - \mu)\top \nabla \varphi(\mu)}{\|\nabla \varphi(\mu)\|_2^2} \times (1 + \|\nabla \varphi(\mu)\|_2^2) .
\] (38)
Fig. 3. Computation of the (unique) population minimizer on a simple 1D example with $\varphi(x) = x \ln x - x$, following Lemma 5. Remark that $\mu > \pi$ in this case.

Eqs (37) and (38) yield the identity:

$$
\overline{\varphi} - \varphi(\mu) = \frac{1}{1 + \|\nabla \varphi(\mu)\|_2^2} \times \frac{1}{n} \sum_i D_{\varphi}(x_i : \mu).
$$

(39)

We now prove (33), and let $m$ denote for short $(1/n) \sum_i D_{\varphi}(x_i : \mu)$. For any $\mu \in \mathbb{P}_{S,\varphi,\varphi_L}$, we get:

$$
\|\overline{x}^\top - \mu^\top\|_2^2
= (\overline{\varphi} - \varphi(\mu))^2 + \|\overline{x} - \mu\|_2^2
= (\overline{\varphi} - \varphi(\mu))^2 - \frac{(\overline{x} - \mu)^\top \nabla \varphi(\mu)}{1 + \|\nabla \varphi(\mu)\|_2^2} \times m
= (\overline{\varphi} - \varphi(\mu)) \times \left( \frac{1}{1 + \|\nabla \varphi(\mu)\|_2^2} \times m \right)
- \frac{(\overline{x} - \mu)^\top \nabla \varphi(\mu)}{1 + \|\nabla \varphi(\mu)\|_2^2} \times m
= \frac{m^2}{1 + \|\nabla \varphi(\mu)\|_2^2}
= \frac{1}{K^2} \left( \frac{1}{n} \sum_i D_{\varphi,K g_\perp}(x_i : \mu) \right)^2.
$$

(43)
We have made use of the definition of $\overline{x}^+$ and $\mu^+$ in (40); we have used (36) in (41) and (39) in (42). This ends the proof of Lemma 5.

The following Theorem states a generalization of Lemma 5 to $p$-norms.

**Theorem 1:** Pick $g = Kg_p$ as in (5) with $K > 0$ a constant, $p = 2k/(2k - 1)$ and $k \in \mathbb{N}_+$. Then, assuming $B_{\varphi,g} = \emptyset$, the right population minimizer(s) for $D_{\varphi,g_p}$ on $S$ match the set:

$$P(D_{\varphi,Kg_p};S) = \arg\min_{\mu} \|\overline{x}^+ - \mu^+\|_q,$$

with $\overline{x}^+$ and $\mu^+$ defined in (19) and (20), and $q = 2k \in \mathbb{N}$ is the Hölder conjugate of $p$.

(proof in Subsection IX-A of the Appendix)

We now study to what extent total Bregman divergences are exhaustive for the construction of the right population minimizer depicted in Lemma 5. It has been shown that Bregman divergences are exhaustive for the expectation as right population minimizer, i.e. if the expectation is the right population minimum of a loss $D(x : y)$, then under mild conditions this loss is a Bregman divergence [4], [5]. It turns out that total Bregman divergence are also exhaustive for their right population minimizer. For the sake of simplicity, we are going to show the result in the one-dimensional setting ($d = 1$). For this objective, we let:

$$P_{S,\varphi} = \left\{ \mu \in \mathbb{R} : \begin{bmatrix} \overline{x} - \mu \\ \overline{\varphi} - \varphi(\mu) \end{bmatrix} \perp \begin{bmatrix} 1 \\ \varphi'(\mu) \end{bmatrix} \right\}.$$  \hspace{1cm} (45)

When $\mu \neq \overline{x}$, the condition is equivalent to

$$\varphi_S'(\mu)\varphi'(\mu) = -1,$$  \hspace{1cm} (46)

with

$$\varphi_S'(z) = \frac{\overline{\varphi} - \varphi(z)}{\overline{x} - z}, \forall z \in \text{dom}(\varphi) \setminus \{\overline{x}\}.$$  \hspace{1cm} (47)

**Theorem 2:** Let $D : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ be a function differentiable such that $D(x : y)$ is twice continuously differentiable in $x$, and satisfies the following assumptions: (i) $D(x : x) = 0, \forall x$, (ii) $D(x : y) > 0, \forall y \neq x$, (iii) $D$ is invariant by rotation of the axes, (iv) the right population minimizer of $D$ on $S = \{x_1, x_2, ..., x_n\}$ is unique and satisfies $\{\mu\} \in P_{S,\varphi}$ for some strictly convex twice differentiable $\varphi$. Then

$$D(x : y) = D_{\varphi,Kg_\perp}(x : y),$$  \hspace{1cm} (48)

where $K > 0$ is a constant and $g_\perp$ is defined in eq. (3).

(proof in Subsection IX-B of the Appendix)

\(^{1}\)Two reals $p, q \geq 1$ are Hölder conjugates when $(1/p) + (1/q) = 1$. 
V. EXTENSION TO $v$-CONFORMAL DIVERGENCES

We now concentrate on $v$-conformal divergences with $(u,v)_\varphi$ a geometric structure. In order not to laden this Section and its notations, we make the simplifying assumption that $B_{\varphi,g} = \emptyset$. This is not restrictive: even in the multidimensional extension of the total Bregman divergences of Table I the cardinal of $B_{\varphi,g} = \emptyset$ would be at most one, so the main structural and algorithmic issues to characterize the right population minimizers essentially lie in the characterization of $P_{S,\varphi}$. Define from $S$ the following averages:

$$\overline{\varphi}_v = \frac{1}{n} \sum_i \varphi(v(x_i))$$
$$\overline{u} = \frac{1}{n} \sum_i u(x_i)^\top v(x_i) - \bar{\varphi}^*_u ;$$
$$\overline{\varphi}_v = \frac{1}{n} \sum_i v(x_i) .$$  

**Lemma 6:** Let $(u,v)_\varphi$ be a geometric structure. Any right population minimizer $\mu$ for the $v$-conformal divergence $D_{\varphi,g}^v$ satisfies:

$$D_z g(\mu) \times (\overline{\varphi}_v - \varphi(v(\mu)))$$
$$= (\overline{\varphi}_v - v(\mu))^\top D_z (g(\mu) \times u(\mu)) ,$$

for any valid direction $z$.

**Proof** We start from (26) which brings, in the case of $v$-conformal divergences (again, directional derivatives apply to the right argument in divergences):

$$\frac{1}{n} D_z \sum_i D^{v}_{\varphi,g}(x_i : \mu)$$
$$= D_z g(\mu) \times \frac{1}{n} \sum_i D_{\varphi}(v(\mu) : v(x_i))$$
$$+ g(\mu) \times \frac{1}{n} D_z \sum_i D_{\varphi}(v(x_i) : v(\mu))$$
$$= 0 .$$

We have $(1/n)D_z \sum_i D_{\varphi}(v(x_i) : v(\mu)) = (v(x_i) - v(\mu))^\top J_{\varphi}(\mu) H_{\varphi}(v(\mu)) z = (v(x_i) - v(\mu))^\top D_z \nabla \varphi(v(\mu))$, so that reordering (52) yields $D_z g(\mu) \times (\overline{\varphi}_v - \varphi(v(\mu))) = (\overline{\varphi}_v - v(\mu))^\top D_z (g(\mu) \times \nabla \varphi(v(\mu)))$. There remains to use the $(u,v)_\varphi$ geometric structure to get the statement of the Lemma.

The Theorem to follow extends Lemma 4 from the geometric structure $(\nabla \varphi, \text{Id})_\varphi$ to a general geometric structure $(u,v)_\varphi$. It is interesting to notice that in the definition of $P_{S,\varphi,g}$, $\delta^+$ is formulated in the $v$ coordinate mapping while $z^+$ is formulated in the $u$ coordinate mapping.

**Theorem 3:** Let $(u,v)_\varphi$ be a geometric structure. Suppose $g(x) = f(u(x))$, with $f$ differentiable. For any $S$ and
any \( \mu, z \in \mathbb{R}^d \), define \( \delta^+_v, z^+_u \in \mathbb{R}^{d+1} \) with:

\[
\delta^+_v = \begin{bmatrix}
\mathbf{x}_v - \phi(v(\mu)) \\
\mathbf{v} - \phi(v(\mu)) \\
\end{bmatrix}
= \begin{bmatrix}
\mathbf{x}_v \\
\mathbf{v} \\
\end{bmatrix} - \begin{bmatrix}
v(\mu) \\
\phi(v(\mu)) \\
\mu^+_v \\
\end{bmatrix},
\]

(53)

\[
z^+_u = \begin{bmatrix}
f(u(\mu)) \times z + \nabla f(u(\mu))^\top z \times u(\mu) \\
- \nabla f(u(\mu))^\top z \\
\end{bmatrix}.
\]

(54)

Then any right population minimizer \( \mu \) for the \( v \)-conformal divergence \( \mathcal{D}^v_{\phi, g} \) satisfies \( \delta^+_v \perp z^+_u \), for any valid direction \( z \).

**Proof** The left-hand side of (51) is \((\mathbf{x}_v - \phi(v(\mu))) \times z^\top J_u(\mu) \nabla f(u(\mu))\), while the directional derivative in the right-hand side is:

\[
\mathcal{D}_z(g(\mu) \times u(\mu)) = \begin{bmatrix}
\mathbf{x}_v - \phi(v(\mu)) \\
\mathbf{v} - \phi(v(\mu)) \\
\end{bmatrix} \nabla f(u(\mu)) \times u(\mu) + g(\mu) \times \mathcal{D}_z u(\mu)
= \begin{bmatrix}
\mathbf{x}_v - \phi(v(\mu)) \\
\mathbf{v} - \phi(v(\mu)) \\
\end{bmatrix} \nabla f(u(\mu)) \times u(\mu) + g(\mu) \times J^\top_u(\mu) z.
\]

So, the right-hand side is:

\[
(\mathbf{x}_v - \phi(v(\mu))) \times z^\top J_u(\mu) \nabla f(u(\mu))
= (\mathbf{x}_v - \phi(v(\mu))) \times \begin{bmatrix}
z^\top J_u(\mu) \nabla f(u(\mu)) \times u(\mu) + g(\mu) \times J^\top_u(\mu) z
\end{bmatrix}.
\]

Finally, (51) becomes \( z^\top J_u(\mu) a = z^\top J_u(\mu)(b + c) \) with:

\[
a \doteq (\mathbf{x}_v - \phi(v(\mu))) \times \nabla f(u(\mu)),
\]

(55)

\[
b \doteq (\mathbf{x}_v - \phi(v(\mu)))^\top u(\mu) \times \nabla f(u(\mu)),
\]

(56)

\[
c \doteq g(\mu) \times (\mathbf{x}_v - \phi(v(\mu))).
\]

(57)

As in Lemma 4 and for the same reasons, we obtain the vector identity \( a = b + c \), which, after making the inner product with vector \( z \) yields this time to the identity:

\[
(\mathbf{x}_v - \phi(v(\mu))) \times - \nabla f(u(\mu))^\top z
+ (\mathbf{x}_v - \phi(v(\mu)))^\top u(\mu) \times \nabla f(u(\mu))^\top z
+ g(\mu) \times (\mathbf{x}_v - \phi(v(\mu))^\top z
= 0,
\]

which states the orthogonality of \( \delta^+_v \) and \( z^+_u \) as stated.

We now build on the proof of Theorem 3 to generalize Lemma 5 to arbitrary geometric structures.
Theorem 4: Let \((u, v)\) be a geometric structure. Pick \(g(\mu) = Kg^u(\mu) = Kf_1(u(\mu))\) for any constant \(K > 0\). Any right population minimizer \(\mu\) for the \(v\)-conformal divergence \(D_{\phi, Kg^u}^{\mu, v}\) satisfies:

\[
\begin{bmatrix}
\n \mathbf{\overline{v}_v} - v(\mu) \\
\n \mathbf{\overline{v}_v} - \varphi(v(\mu))
\end{bmatrix}
\perp
\begin{bmatrix}
\n u(\mu) \\
\n \|u(\mu)\|_2^2
\end{bmatrix}.
\]  

(58)

Furthermore, we have:

\[
\mathcal{P}(D_{\phi, Kg^u}^{\mu, v}; S) = \arg\min_{\mu} \|\mathbf{\overline{v}_v}^+ - \mu^+_v\|_2,
\]

(59)

with \(\mathbf{\overline{v}_v}^+\) and \(\mu^+_v\) defined as follows:

\[
\mathbf{\overline{v}_v}^+ = \begin{bmatrix}
\n \mathbf{\overline{v}_v} \\
\n \mathbf{\overline{v}_v}
\end{bmatrix} = \begin{bmatrix}
\n \frac{1}{n} \sum_i \mathbf{x}_i \mathbf{x}_i^\top v(x_i) - \varphi^*(u(\mu)) \\
\n \frac{1}{n} \sum_i \mathbf{x}_i \mathbf{x}_i^\top v(x_i) - \varphi^*(u(\mu))
\end{bmatrix},
\]

(60)

\[
\mu^+_v = \begin{bmatrix}
\n v(\mu) \\
\n \varphi(v(\mu))
\end{bmatrix} = \begin{bmatrix}
\n u(\mu)^\top v(\mu) - \varphi^*(u(\mu)) \\
\n u(\mu)^\top v(\mu) - \varphi^*(u(\mu))
\end{bmatrix}.
\]

(61)

Proof We follow the proof of Lemma 5, and obtain this time that:

\[
\mathbf{z}^+ = \frac{K}{(1 + \|u(\mu)\|_2^3)^{3/2}} \times \begin{bmatrix}
\n (1 + \|u(\mu)\|_2^3) \times z - u(\mu)^\top z \times u(\mu) \\
\n u(\mu)^\top z
\end{bmatrix}.
\]

Now, if we pick \(z = u(\mu)\), then we have:

\[
\begin{aligned}
u(\mu) &= (1 + \|u(\mu)\|_2^3) \times z - u(\mu)^\top z \times u(\mu),
\end{aligned}
\]

with which Theorem 3 achieves the proof of (58). To prove (59), we first remark that (58) brings equivalently:

\[
\mathbf{\overline{v}_v} - \varphi(v(\mu)) = -\frac{(\mathbf{\overline{v}_v} - v(\mu))^\top \nabla \varphi(v(\mu))}{\|u(\mu)\|_2^2}.
\]

(62)

The average divergence in Lemma 4 also yields:

\[
\frac{1}{n} \sum_i \mathbf{D}_\phi(v(x_i) : v(\mu))
\]

(63)

\[
= -\frac{(\mathbf{\overline{v}_v} - v(\mu))^\top \nabla \varphi(v(\mu))}{\|u(\mu)\|_2^2} \times (1 + \|u(\mu)\|_2^3),
\]

(64)

which brings with (62):

\[
\mathbf{\overline{v}_v} - \varphi(v(\mu)) = \frac{1}{1 + \|u(\mu)\|_2^3} \times \frac{1}{n} \sum_i \mathbf{D}_\phi(v(x_i) : v(\mu)).
\]

(65)
The identity \( a = b + c \) in (53) — (57), read \( a - b = c \), brings:

\[
\mathcal{F}_v - v(\mu) = \frac{1}{f(u(\mu))} (\mathcal{F}_v - \varphi(v(\mu)) - (\mathcal{F}_v - v(\mu))^\top u(\mu))
\]

\[
\times \nabla f(u(\mu)) = \frac{1}{f(u(\mu))} \left( \frac{1}{n} \sum_i D_\varphi(v(x_i) : v(\mu)) \right) \times \nabla f(u(\mu)) ,
\]

because \( u = \nabla \varphi \circ v \). Making the inner product with \( \mathcal{F}_v - v(\mu) \) and using the expression of \( f = f_\perp \), we obtain:

\[
\| \mathcal{F}_v - v(\mu) \|_2^2 = \frac{- \left( \mathcal{F}_v - v(\mu) \right)^\top u(\mu)}{1 + \| u(\mu) \|^2_2} \left( \frac{1}{n} \sum_i D_\varphi(v(x_i) : v(\mu)) \right) .
\]

Assembling this last identity and (65) yields that for any \( \mu \) that meets (58), we have:

\[
(\mathcal{F}_v - \varphi(v(\mu)))^2 + \| \mathcal{F}_v - v(\mu) \|_2^2
\]

\[
= \frac{(\mathcal{F}_v - \varphi(v(\mu)))^2}{1 + \| u(\mu) \|^2_2} \times \frac{1}{n} \sum_i D_\varphi(v(x_i) : v(\mu))
\]

\[
- \frac{(\mathcal{F}_v - \varphi(v(\mu)))^2}{1 + \| u(\mu) \|^2_2} \left( \frac{1}{n} \sum_i D_\varphi(v(x_i) : v(\mu)) \right)
\]

\[
= \frac{1}{1 + \| u(\mu) \|^2_2} \left( \frac{1}{n} \sum_i D_\varphi(v(x_i) : v(\mu)) \right)
\]

\[
\times (\mathcal{F}_v - \varphi(v(\mu)) - (\mathcal{F}_v - v(\mu))^\top u(\mu))
\]

\[
= \frac{1}{K^2} \left( \frac{1}{n} \sum_i \frac{K}{\sqrt{1 + \| u(\mu) \|^2_2}} \times D_\varphi(v(x_i) : v(\mu)) \right)^2 .
\]

We finally have to remark that \( \| \mathcal{F}_v^+ - \mu_v^+ \|^2_2 = (\mathcal{F}_v - \varphi(v(\mu)))^2 + \| \mathcal{F}_v - v(\mu) \|^2_2 \) for (67) to bring the statement of (59).

we can also state a generalization of Theorem 1 to arbitrary geometric structures \((u, v)_\varphi \). Its proof, omitted, combines the arguments of the proofs of Theorems 1 and 4.

**Theorem 5**: Let \((u, v)_\varphi \) be a geometric structure. Pick \( g = Kg_p^u(\mu) \) with \( g_p^u(\mu) \equiv f_p(u(\mu)) \), \( f_p \) is defined in (6). \( K > 0 \) is a constant, and \( p = 2k/(2k - 1) \) with \( k \in \mathbb{N}^\ast \). Then, the right population minimizer(s) for the \( v \)-conformal divergence \( D_{\varphi,Kg_p^u} \) on \( S \) match the set:

\[
\mathcal{P}(D_{\varphi,Kg_p^u} : S) = \arg \min_{\mu} \| \mathcal{F}_v^+ - \mu_v^+ \|_q ,
\]

with \( \mathcal{F}_v^+ \) and \( \mu_v^+ \) defined in (60) and (61), and \( q = 2k \in \mathbb{N} \) is the Hölder conjugate of \( p \).

**VI. ROBUSTNESS OF THE POPULATION MINIMIZERS**

Suppose we add an outlier element \( x_\ast \) with small weight \( 0 < \epsilon < 1 \) to \( S \). The population minimizer (left or right) of \( S \), \( \mu \), eventually drifts to a new population minimizer \( \mu_\ast = \mu + \epsilon \delta_\mu \) of \( S \cup \{ x_\ast \} \). \( \delta_\mu \) is called the influence
function of $x_*$. A population minimizer is robust to outliers iff the magnitude of $\delta_\mu$ is bounded, as explained in the following definition where $0 < \tau < 1$ is any small constant.

**Definition 2:** The population minimizer of some divergence $D$ is robust to outliers when, for any outlier $x_*$ and any weight $0 < \epsilon < 1 - \tau$, $\|\delta_\mu\|_2 \leq C$, where $C$ does not depend upon $x_*$ nor $\epsilon$.

Robustness according to Definition 2 is stronger than in the model of [6], [12] as our robustness strictly implies theirs (which relies on very small weights $\epsilon$). So the Lemma to follow is a twofolds generalization of the results of [6], [12], not only from the standpoint of the divergences, but also from the model’s.

**Lemma 7:** Let $(u, v)_\varphi$ be a geometric structure. Suppose the following assumptions are verified: (i) $g(x) = O(1), \forall x$, (ii) $\|u(x)\|_2 = O(1/g(x)), \forall x$, (iii) the minimal eigenvalue of $J_u^T J_u$ is $\lambda > 0$, $J_u$ being the Jacobian of $u$. Then under assumptions (i-iii), the left population minimizer of $v$-conformal divergence $D^{v, g}_{\varphi, g}$ is robust to outliers.

**Proof** We use Lemma 1 and we fix $S$ satisfying assumption (i). The left population minimizers $\mu$ and $\mu_*$ satisfy:

$$u(\mu) = \frac{1}{\Sigma} \sum_i g(x_i) u(x_i),$$

$$u(\mu_*) = \frac{1 - \epsilon}{\Sigma_*} \sum_i g(x_i) u(x_i) + \frac{\epsilon g(x_*)}{\Sigma_*} u(x_*)$$

$$= \frac{(1 - \epsilon) \Sigma}{\Sigma_*} u(\mu) + \frac{\epsilon g(x_*)}{\Sigma_*} u(x_*)$$

$$= u(\mu) + \frac{\epsilon g(x_*)}{\Sigma_*} (u(x_*) - u(\mu)),$$

where $\Sigma = \sum_i g(x_i)$ and $\Sigma_* = (1 - \epsilon) \Sigma + \epsilon g(x_*)$. Now, a Taylor expansion of $u^{-1}$ brings:

$$\mu_* = u^{-1}(u(\mu) + \epsilon \delta) = \mu + \epsilon J_u^{-1}(\mu_*) \delta,$$

for some $\mu_0 = u(\mu) + \alpha \epsilon \delta$ with $0 < \alpha < 1$. We also have $J_u^{-1}(\mu_0) = J_u^{-1}(u^{-1}(\mu_0))$, which, since $\mu_0 - \mu = \epsilon \delta_\mu$, yields the influence function of $x_*$:

$$\delta_\mu = \frac{g(x_*)}{\Sigma_*} \times J_u^{-1}(u^{-1}(\mu_0)) (u(x_*) - u(\mu)).$$

(69)

Let $J_u$ denote $J_u(u^{-1}(\mu_0))$ for short. Eq. (69) brings

$$\|\delta_\mu\|_2^2 = \delta^T (J_u^{-1})^T J_u^{-1} \delta \leq \|\delta\|_2^2 \lambda_{\text{max}},$$

(70)

for some upperbound $\lambda_{\text{max}}$ on the eigenvalues of $(J_u^{-1})^T J_u^{-1}$. We also have

$$\|\delta\|_2^2 = \frac{g(x_*)}{\Sigma_*} \times \|u(x_*) - u(\mu)\|_2^2$$

$$\leq \frac{2g(x_*)}{\Sigma_*} \times (\|u(x_*)\|_2^2 + \|u(\mu)\|_2^2)$$

$$= \frac{2g(x_*)}{\Sigma_*} \|u(x_*)\|_2^2 + \frac{2g(x_*)}{\Sigma_*} \|u(\mu)\|_2^2.$$
Because $\epsilon < 1 - \tau$, we have $\Sigma_\ast \geq \tau g(x_1) = K_1$ for some $K_1 > 0$ which does not depend upon $x_\ast$ or $\epsilon$. Because of assumption (i), $g(x_\ast) \leq K_2$ for some constant $K_2 > 0$ and so $b = O(1)$ (the tilda meaning that the function does not depend upon $x_\ast$ or $\epsilon$). Because of assumption (ii), $g(x_\ast)\|u(x_\ast)\|_2 \leq K_3$ for some constant $K_3 > 0$ and so $a = O(1)$. Hence, $\|\delta\|_2^2 = O(1)$. There remains to plug this into (70), and remark that $\lambda_{\text{max}} \leq 1/\lambda$ from assumption (iii), to conclude.

Lemma 7 generalizes the robustness of the left population centers of total Bregman divergences (Theorem III.2 in [12]), for which $g = g_{\perp}, v = \text{Id}$, $u = \nabla \varphi$ (the Jacobian of $u$ being the Hessian of $\varphi$, it satisfies assumption (iii) since $\varphi$ is strictly convex).

The right population minimizer is unfortunately not robust to outliers for any $g$ according to Definition 2, yet it satisfies in a general setting of $v$-conformal divergences, a weaker notion of robustness which says that the influence function must be properly bounded by a divergence between $x_\ast$ and $\mu$ — this divergence ignoring $g, u, v, \varphi$ —, as long as $x_\ast$ does not deviate too much from $\mu$ in the $v$-coordinate mapping. This last notion exploits the fact that convex function are locally Lipschitz.

**Definition 3:** Let $(u, v)\varphi$ be a geometric structure. The population minimizer of some $v$-conformal divergence $D^v_{\varphi,g}$ is $K$-weakly robust to outliers when for any outlier $x_\ast$ and any weight $0 < \epsilon < 1$:

$$|\varphi(v(x_\ast)) - \varphi(v(\mu))| \leq L\|v(x_\ast) - v(\mu)\|_2 \Rightarrow \|\delta_\mu\|_2 \leq K\ell(L)\|x_\ast - \mu\|_2,$$

where $K \geq 0$ is not a function of $x_\ast$ or $\epsilon$, and $\ell(L)$ is a linear function in $L$.

We now show that the right population minimizer is $K$-weakly robust to outliers, for a $K$ which depends solely on the coordinate mapping $v$. We assume in the Lemma that $0 \in \text{im} u$, which is a mild assumption as it postulates in the $(u, v)\varphi$ geometric structure that the gradient $\nabla \varphi$ has a root in coordinate mapping $v$. We exploit the fact that any matrix $A \in \mathbb{R}^{d \times d}$ satisfies $A^\top A \succeq 0$, where “$\succeq$” means positive semi-definite.

**Lemma 8:** Let $(u, v)\varphi$ be a geometric structure, and let $f = g \circ u^{-1}$. We make the following assumptions: (i) $0 \in \text{im} u$, (ii) $f(z) \neq 0, \forall z$, (iii) the ratio of the maximal to the minimal eigenvalue of $J_v^\top J_v$, noted $\lambda_v$, is finite, where $J_v$ is the Jacobian of $v$. Then the right population minimizer of $v$-conformal divergence $D^v_{\varphi,g}$ is $\sqrt{\lambda_v}$-weakly robust to outliers.

(Proof in Subsection IX-C of the Appendix)

**VII. DISCUSSION**

In this Section, we discuss several aspects of population minimizers in the setting of conformal divergences; in particular, we discuss further the geometric structure relation, the approximation of the right population minimizers in the 1D setting, the existence of symmetric conformal divergences, and the uniqueness of the right population minimizer.
a) The nature of the \((u, v)\) geometric structure relation: The objective of this subsection is to provide an information geometric primer on the \((u, v)\) geometric structure which shows that any \((u, v)\) geometric structure shapes a dually flat manifold with a divergence which is a conformal divergence, and then provide a property on the relation which might be useful in the context of clustering. The \((u, v)\) geometric structure has been introduced in the context of information geometry to provide a way to compute and analyze the dually flat \((\eta, \theta)\) coordinate system arising e.g. in exponential families and Bregman divergences, through a single source parameter which is originally a distribution \([14]\).

Let us state some basic results for the \((u, v)\) geometric structure in the 1D setting, that can be found in \([14]\).

Consider two strictly monotonous differentiable functions \(u(\xi)\) and \(v(\xi)\) with \(u(0) = v(0) = 0\). Consider the positive measures on \(\mathbb{R}^{d+1}\), and denote by \(m(x, \xi) = \sum_{i=0}^{n} \xi_i \mathbb{1}_{x=x_i}\) a positive distribution computed from \(S = \{x_0, x_2, ..., x_n\}\), where \(\mathbb{1}\) is the indicator variable. \(\xi = [\xi_0 \xi_2 ... \xi_n]^T\) defines a coordinate system from which we may define two coordinate systems \(\eta, \theta\) of \(\mathbb{R}^{d+1}\) with \(\theta^i = u(\xi_i)\) and \(\eta^i = v(\xi_i)\). These coordinate systems have the following interesting information-geometric properties.

Theorem 6: [14] The \((u, v)\)-geometric structure is dually flat, with the following two potential functions:

\[
\psi(\theta) = \sum \int \int \frac{v'(u^{-1}(\theta^i))}{u'(u^{-1}(\theta^i))} (d\theta^i)^2,
\]

\[
\phi(\eta) = \sum \int \int \frac{u'(v^{-1}(\eta^i))}{v'(v^{-1}(\eta^i))} (d\eta^i)^2;
\]

the divergence between two \(p\) and \(q\) is given by:

\[
D(p, q) = \psi(\theta_p) + \phi(\eta_q) - \theta_p \cdot \eta_q,
\]

the metric in the \(\theta\) coordinate system is:

\[
g_{ij}(\theta) = \frac{v'(\xi_i)}{u'(\xi_i)} \delta_{ij},
\]

and finally the third order tensor is:

\[
T_{ijk}(\theta) = \sum \psi'''(\theta^i) \delta_{ijk}.
\]

\(\psi, \phi\) are also called respectively the geometrical free energy and the negative entropy of the dually flat manifold \([19]\). Their role permute as the \((u, v)\)-geometric structure relation is symmetric (Lemma 9 below). One important example of \((u, v)\)-geometric structure is Amari’s \((\alpha, \beta)\) structure:

\[
u(\xi) = \xi^\alpha,
\]

\[
v(\xi) = \xi^\beta,
\]

and it follows \(\theta^i = (\xi_i)^\alpha\), \(\eta^i = (\xi_i)^\beta\). The potentials are:

\[
\psi(\theta) = \frac{\alpha}{\alpha + \beta} \sum \theta_i^{\alpha+\beta},
\]

\[
\phi(\eta) = \frac{\beta}{\alpha + \beta} \sum \eta_i^{\alpha+\beta},
\]
the metric is
\[ g_{ij}(\theta) = \frac{\beta}{\alpha} \xi_i^{\beta-\alpha} \delta_{ij}, \]
and the third order tensor is:
\[ T_{ijk}(\theta) = \frac{\beta(\beta - \alpha)}{\alpha^2} \xi_i^{\beta-2\alpha} \delta_{ijk}. \]
Finally, the divergence is the \( \alpha - \beta \)-divergence:
\[
D(p : q) = \sum_i \left\{ \frac{\alpha}{\alpha + \beta} (\theta_p^i)^{\alpha+\beta} + \frac{\beta}{\alpha + \beta} (\eta_q^i)^{\alpha+\beta} \right\} - \theta_p \cdot \eta_q
\]
where we have let for short: \( \theta_p^i = (\xi_i)^\alpha \), \( \eta_q^i = (\xi_i')^\beta \).

To finish up with the basic results on the \((u,v)\) geometric structure, the submanifold of probability measures \( S^d \) with \( \sum p_i = 1 \) yields the constraints:
\[
\sum_i u^{-1}(\xi_i) = 1, \\
\sum_i v^{-1}(\xi_i) = 1,
\]
that is not dually flat except when both \( u \) and \( v \) are linear functions. Finally, the \((u,v)\) geometric structure is not confined to vectors: it has also been extended to matrices [15]. Any \((u,v)\) geometric structure thus completely defines a dually flat manifold with good information geometric properties, and a divergence which is a special case of conformal divergence.

To study the nature of the \((u,v)\) geometric structure, let us define a tolerance relation [20] as a binary relation which is reflexive and symmetric but not necessarily transitive. An equivalence relation is reflexive, symmetric and transitive. We study the “geometric structure” binary relation, \((u,v)\), which holds when there exists some \( \phi \) such that \( (u,v)_\phi \) is a geometric structure.

**Lemma 9:** The “geometric structure” relation is a tolerance relation. It is an equivalence relation in the subset of functions \( S_\phi \) indexed by some strictly convex differentiable \( \phi \) and defined by: \( S_\phi = \{ \phi : H\phi(\nabla\phi) = H\phi P_\phi D_\phi^{-1} D\phi P_\phi^T \} \), where \( P_\phi, D_\phi \) are the eigenspace and eigenvalues matrix of \( H\phi \) and \( D > 0 \) is diagonal.

**Proof** \((u,u)\) is a geometric structure for \( \phi \doteq (1/2) \sum_i (x^i)^2 \) so the relation is reflexive. If \((u,v)\) is a geometric structure for \( \phi \), then \((v,u)\) is a geometric structure for \( \phi^* \), so the relation is symmetric. Let \((u,v)_\phi \) and \((v,w)_\phi \) be two geometric structures. We have \( u \circ w^{-1} = \nabla\phi \circ \nabla\phi \), and so \( J_{uw^{-1}} = H\phi(\nabla\phi)H\phi \), that we want to be symmetric positive definite for the “geometric structure” relation to be transitive. Both \( H\phi \) and \( H\phi \) are symmetric positive definite. Since (i) the product of two positive definite matrices is positive definite iff their product is normal, and (ii) the product of two symmetric matrices is symmetric iff their have the same eigenspace, \( H\phi(\nabla\phi)H\phi \succ 0 \)
iff we have the diagonalizations $H\varphi(\nabla \phi) = PD_1P^\top$ and $H\phi = PD_2P^\top$, with $P$ unitary and $D_1, D_2 > 0$. This finishes the proof of Lemma 9.

Hence, for example, the geometric structure relation is an equivalence relation on any subset of positive definite quadratic forms that have the same eigenspace. The compactness and convexity of some of these subgroups may be interesting from a clustering standpoint to learn the $(u, v)_\phi$ geometric structure when $\phi$ is fixed (see Section V).

b) Simple right population minimizers: the following corollary is a safe-check of Lemma 3 which states when the right population minimizer has simple forms.

Corollary 1: Suppose $S$ contains at least two distinct elements. The right population minimizer of $D_{\phi, g}$ on set $S$ is:

1) always the arithmetic average (i.e. $\mu = \overline{x}$) iff $g(y)$ is constant;
2) always the $\varphi$-mean (i.e. $\varphi(\mu) = \overline{\varphi}$) iff $\varphi(x) = K \int 1/h(u^\top x) + K'$, with (i) $K, K'$ and vector $u$, constants,
   (ii) $g(x) = h(u^\top x)$ for some function $h : \mathbb{R} \to \mathbb{R}$ strictly monotonous with derivative sign opposite to that of $K$.

Proof (Of point 1) ($\Rightarrow$) $\mu = \overline{x}$ zeroes the right-hand side of (27), which, since $\overline{\varphi} \neq \varphi(\mu)$, implies $D_z g(\mu) = 0$, for any $z \neq 0$, and so $g(\mu)$ is constant. ($\Leftarrow$) is a property of Bregman divergences.

(Of point 2) ($\Rightarrow$) This time, $\varphi(\mu) = \overline{\varphi}$ zeroes the left-hand side of (27). Because $\varphi$ is strictly convex, $\overline{x} \neq \mu$ and so (27) brings $D_z(g(\mu)\nabla \varphi(\mu)) = 0, \forall z$, and so $\nabla \varphi(\mu) = (K/g(\mu))u$ for some constants $K$ and vector $u$. The hessian coordinates are $H_{ij}\varphi(\mu) = -(Ku^i/g^2(\mu))\partial g(\mu)/\partial u_j$. Because the Hessian is symmetric, we obtain $u^j\partial g(\mu)/\partial u_i = u^i\partial g(\mu)/\partial u_j$, and so $g(\mu)$ can be expressed as $g(\mu) = h(u^\top \mu)$ for some function $h : \mathbb{R} \to \mathbb{R}$. We get $x^\top H\varphi x = -(K' h'(u^\top \mu)/h^2(u^\top \mu))||u \cdot x||^2_2$, with "\cdot" denoting Hadamard product, and since we want $x^\top H\varphi x > 0$ when $x \neq 0$, $h'$ has to be of a different sign than $K$. ($\Leftarrow$) is immediate.

c) One symmetric conformal divergence: We show that there exists a single 1D symmetric conformal divergence in the $(u, v)$-structure, the square loss, $D_{\phi, g}(v(x) : v(y)) \propto (v(x) - v(y))^2$. As a corollary, it shows that there is no symmetric total Bregman divergence. The proof is made in the 1D case, that is, when the domain and image of $u$ and $v$ is $\mathbb{R}$, and it can be extended at no cost to dD separable conformal divergences, for which $g(u(y))D_\phi(v(x) : v(y)) = \sum_i g(u(y^i))D_\phi(v(x^i) : v(y^i))$.

Lemma 10: Let $(u, v)_\phi$ be a geometric structure and $D_{\phi, g}$ a conformal divergence for some strictly convex twice differentiable $\phi$. Suppose that $\forall x, y$:

$$g(u(y))D_\phi(v(x) : v(y)) = g(u(x))D_\phi(v(y) : v(x)) .$$

Then (i) $g(.) = K_1 > 0$, (ii) $u = v$, (iii) $\varphi(x) = K_2 x^2 + K_3 x + K_4$ for some constants $K_1, K_2, K_3, K_4$.(proof in Subsection IX-D of the Appendix) Thus, conformal divergence are not metrics, yet they can be used to craft metrics by symmetrization, as there exists $0 < \alpha < 1$ such that $(g(u(y))D_\phi(v(x) : v(y)) + g(u(x))D_\phi(v(y) : v(x)))^\alpha$ is a metric.
\(\varphi(x)\) Name or expression for \(D_{\varphi,g,l}(x : y)\) \(\varphi\)-mean Name of \(\varphi\)-mean Location

| Expression | Name | Location |
|------------|------|----------|
| \(-\log(x)\) | Total Itakura-Saito \(\prod_i x_i^{\perp}\) | Geometric mean | \(\mathcal{I}_{\varphi}, \mathcal{I}\) |
| \(1/x\) | Total inverse \(\frac{n}{\sum_i x_i^{1-p}}\) | Harmonic mean | \(\mathcal{I}_{\varphi}, \mathcal{I}\) |
| \(x^2\) | Total square loss \(\pm \sqrt{\frac{1}{n} \sum_i x_i^2}\) | Root mean square | \(\mathcal{I}, \mathcal{I}_{\varphi}\) or \(\mathcal{I}_{\varphi}, \mathcal{I}\) |
| \(x^p, p \geq 2\) | Total power loss \(\sum_i x_i^{p}\) | Power mean | \(\mathcal{I}_{\varphi}, \mathcal{I}\) |
| \(\exp(x)\) | Total exp divergence \(\log \sum_i \exp x_i\) | None | \(\mathcal{I}_{\varphi}, \mathcal{I}\) |
| \(x \log(x)\) | Total KL \(\frac{1}{2} \sum_i x_i \log x_i\) | None | \(\mathcal{I}, \mathcal{I}_{\varphi}\) |
| \(W(x)\) | \(x(\exp(x) - \exp(y)) + y(\exp(x) - \exp(y))\) \(W(\frac{1}{n} \sum_i W^{-1}(x_i))\) | None | \(\mathcal{I}, \mathcal{I}_{\varphi}\) |

**Table 1** Examples of total Bregman divergences \(D_{\varphi,g,l}(x : y)\), and location of the candidate population minimizer according to Lemma 11. \(W\) is Lambert \(W\) function and \(W^{-1}(x)\) is a shorthand for \(x \exp(x)\).

d) **Fast approximation of right population minimizers:** we now show that under mild assumptions on \(\varphi\), candidates for right population minimizer may be easily located and approximated in the 1D setting. Assume wlog that \(S\) is ordered, that is \(x_1 \leq x_2 \leq \ldots \leq x_n\). Whenever \(\varphi\) is bijective over \([x_1, x_n]\), we define the \(\varphi\)-mean:

\[
\mathcal{I}_{\varphi} = \varphi^{-1}(\mathcal{I}).
\]

Let us denote a candidate right population minimizer as a real which is solution of \(\varphi\). Candidate population minimizers are critical points for the right parameter of the average divergence \(\varphi\).

**Lemma 11:** Suppose \(g(y) = g_\perp(y)\), and assume that \(\varphi'\) has constant sign on \([x_1, x_n]\). Then there exists a candidate right population minimizer \(\mu\) in \([x_1, x_n]\). Furthermore, \(\mu \in [\mathcal{I}, \mathcal{I}_{\varphi}]\) if sign = –, and \(\mu \in [\mathcal{I}_{\varphi}, \mathcal{I}]\) if sign = +. Here, “sign” denotes the sign of \(\varphi'\) over \([x_1, x_n]\).

**Proof** The proof relies on the study in \([x_1, x_n]\) of function \(\varphi_\perp(x) = -1/\varphi_\parallel(x)\) (see eq. (47)), which is the slope of the line orthogonal to the segment which links \((\mathcal{I}, \mathcal{I})\) to \((\mathcal{I}_{\varphi}, \mathcal{I})\).

Suppose \(\varphi'(x) < 0\) on \([x_1, x_n]\), which implies \(\varphi(x_1) \geq \mathcal{I}\), and so \(\mathcal{I}_{\varphi} \in [x_1, \mathcal{I}]\), and satisfies \(\varphi(\mathcal{I}_{\varphi}) = \mathcal{I}\). It comes \(\varphi_\perp(x) \leq 0\) on \((\mathcal{I}_{\varphi}, \mathcal{I})\), with \(\lim_{x \to \mathcal{I}_{\varphi}} \varphi_\perp(x) = -\infty\) and \(\varphi_\parallel(\mathcal{I}) = 0\). Because \(\varphi_\parallel(x)\) is continuous, so is \(\varphi_\parallel(x)\), and so must be \(\mu \in [\mathcal{I}, \mathcal{I}_{\varphi}]\) such that \(\varphi_\parallel(\mu) = \varphi'(x)\). This \(\mu\) is a candidate right population minimizer.

Suppose now that \(\varphi'(x) > 0\) on \([x_1, x_n]\), which implies \(\varphi(x_n) \geq \mathcal{I}\), and so \(\mathcal{I}_{\varphi} \in [\mathcal{I}_{\varphi}, \mathcal{I}]\), and satisfies \(\varphi(\mathcal{I}_{\varphi}) = \mathcal{I}\). This time, \(\varphi_\perp(\mathcal{I}) = 0\) and \(\lim_{x \to \mathcal{I}_{\varphi}} \varphi_\perp(x) = +\infty\), so there must be \(\mu \in [\mathcal{I}, \mathcal{I}_{\varphi}]\) such that \(\varphi_\parallel(\mu) = \varphi'(x)\). This \(\mu\) is a candidate right population minimizer. This ends the proof of Lemma 11.

Table 1 presents some applications of Lemma 11 (the domain considered for \(\varphi(x) = 1/x\) is \(\mathbb{R}_{++}\)). Approximating the candidate right population minimizer \(\mu\) in the interval may be done by fitting the roots of equations of the form \(f(\mu, \mathcal{I}, \mathcal{I}_{\varphi}) = 0\), some of which are given below as examples:

\[
\log(\mathcal{I}_{\varphi}) - \log(\mu) + \mu^2 - \mathcal{I} \mu = 0 \tag{73}
\]
for total Itakura Saito divergence,
\[ 2\mu^3 - (2\varphi^2 - 1)\mu - \overline{\varphi} = 0 \] (74)
for total square loss divergence (notice that a closed-form expression for \( \mu \) is available in that case),
\[ p\mu^{2p-1} - p\varphi^p\mu^{p-1} + \mu - \overline{\varphi} = 0 \] (75)
for total power loss divergence,
\[ \exp(2\mu) - \exp(\varphi + \mu) + \mu - \overline{\varphi} = 0 \] (76)
for total exp divergence, and finally
\[ \left( \mu \log \mu - \frac{\varphi}{W(\varphi)} \log \frac{\varphi}{W(\varphi)} \right) (1 + \log \mu) + \mu - \overline{\varphi} = 0 \] (77)
for total KL divergence.

e) Non-uniqueness and existence of the right population minimizers: The left population minimizer of any \( v \)-conformal divergence is unique (Lemma 1). This is not always the case for the right population minimizer. In very seldom but typical pathological cases, the population minimizers may even span the complete domain of \( \varphi \), as displayed in Figure 4.

We also notice that the compactness of \( \text{dom}(D_{\varphi,g}) \) appears necessary for the right population minimizers to exist, as otherwise one may build pathological Cauchy sequences for the right divergence parameter that converge to a right population minimizer not in \( \text{dom}(D_{\varphi,g}) \).

f) Extension to scaled Bregman divergences: a new generalization of Bregman divergences has been recently coined [13], called scaled Bregman divergences. A scaled Bregman divergence is a particular case, for \( g = \text{Id} \), of what we call a scaled conformal divergence, defined as:
\[ D_{\varphi,g}(x : y; w) \equiv wD_{\varphi,g}(x/w : y/w), \] (78)
for $w > 0$. A conformal divergence is obtained when $w = 1$. Scaled Bregman divergences generalize other important classes of divergences such as Csiszár’s $f$-divergences, and they yield explicit formulas for the scaled divergence of exponential families, which means they have a significant potential for applications in clustering \cite{13}. It is thus important to characterize their population minimizers. Though it is out of the scope of our paper to extent further our results to scaled divergences, we can give some insights into the similarities and differences with the case $w = 1$. Population minimizers are now sought with respect to some sets $\mathcal{S} = \{x_1, x_2, \ldots, x_n\}$ and $\mathcal{W} = \{w_1, w_2, \ldots, w_n\}$, such that a left population minimizer of the ordered pair $(\mathcal{S}, \mathcal{W})$ for $D_{\varphi, g}$ is defined as $\mu$ that minimizes $\sum_i D_{\varphi, g}(\mu : x_i; w_i)$. The following Lemma shows that, despite the left population minimizer is not always available in closed form in general (unlike $\upsilon$-conformal divergences), it is in between the minimal and maximal values of $\mathcal{S}$ (like $\upsilon$-conformal divergences).

**Lemma 12:** The left population minimizer of $D_{\varphi, g}$ over $(\mathcal{S}, \mathcal{W})$ is unique and in $[\min_i x_i, \max_i x_i]$.

**Proof** We build upon \cite{9}. Any left population minimizer is a solution of:

$$0 = \frac{d}{d\mu} \sum_i w_i D_{\varphi, g}(\mu/w_i : x_i/w_i)$$

$$= \sum_i g \left( \frac{x_i}{w_i} \right) v' \left( \frac{\mu}{w_i} \right) \left( u \left( \frac{\mu}{w_i} \right) - u \left( \frac{x_i}{w_i} \right) \right)$$

$$= \sum_i g \left( \frac{x_i}{w_i} \right) v' \left( \frac{\mu}{w_i} \right) \left( \frac{\mu - x_i}{w_i} \right) u' \left( \frac{\mu}{w_i} \right)$$

where $\mu_i = \mu + \alpha_i (x_i - \mu)$ for some $0 < \alpha_i < 1$. Eq \cite{80} is obtained after $n$ Taylor expansions of $u$. We also have $(v \circ u^{-1})' = (v' \circ u^{-1})/(u' \circ u^{-1}) \equiv (\varphi^*)''$, and so, since $\varphi^*$ is strictly convex, $v'(x)$ and $u'(x)$ have the same sign. Since $u$ is strictly monotonous, $u'$ does not change sign over its domain, and so the product $\pi_i = g \left( \frac{x_i}{w_i} \right) v' \left( \frac{\mu}{w_i} \right) u' \left( \frac{\mu}{w_i} \right)$ is non negative, $\forall i$. We can summarize \cite{80} as $h(\mu) = \sum_i \pi_i (\mu - x_i)/w_i = 0$: since all $w_i > 0$, we get $h(\min_i x_i) \leq 0$ and $h(\max_i x_i) \geq 0$. Since each summand in $h$ is the product of continuous functions, there must be $\mu \in [\min_i x_i, \max_i x_i]$ such that \cite{80} holds, and since $h$ is strictly increasing, there is only one such point. Since $\sum_i D_{\varphi, g}(\mu : x_i; w_i)$ is strictly convex in $\mu$, this is the left population minimizer.

This Lemma can be extended to separable divergences in $\mathbb{R}^d$, to show that the left population minimizer of scaled conformal divergences lies in $\prod_j [\min_j x'_j, \max_j x'_j]$.

**VIII. Conclusion**

We have studied the left and right population minimizers of conformal divergences, a superset of Bregman divergences and total Bregman divergences, in the $(u, v)$ geometric structure of Amari which generalizes dually flat affine connections. We have characterized analytically and geometrically the population minimizers, shown the exhaustivity property of conformal divergences for the left population minimizer, and the exhaustivity of total Bregman divergences for the right population minimizers. We do believe that these results, as well as additional results we provide on the robustness of the population minimizers, the nature of the $(u, v)$ geometric structure relation, and the simple approximation of 1D population minimizers, shall be useful to improve further and widen
the scope of new and successful clustering algorithms relying on broad classes of distortions that escape the conventional framework of Bregman divergences, as e.g. recently initiated with total Bregman divergences.

REFERENCES

[1] A. Banerjee, S. Merugu, I. Dhillon, and J. Ghosh, “Clustering with bregman divergences,” in Proc. of the 4th SIAM International Conference on Data Mining., 2004, pp. 234–245.
[2] A. P. Dempster, N. M. Laird, and D. B. Rubin, “Maximum likelihood from incomplete data via the EM algorithm,” J. of the Royal Stat. Soc. B, vol. 39, pp. 1–38, 1977.
[3] J. McQueen, “Some methods for classification and analysis of multivariate observations,” in Proc. of the 5th Berkeley symposium on mathematical statistics and probability, 1967, pp. 281–297.
[4] J. Abernethy and R.-M. Frongillo, “A characterization of proper scoring rules for linear properties,” in Proc. of the 25th COLT, 2012, pp. 1–14.
[5] A. Banerjee, X. Guo, and H. Wang, “On the optimality of conditional expectation as a bregman predictor,” IEEE Trans. on Information Theory, vol. 51, pp. 2664–2669, 2005.
[6] M. Liu, “Total Bregman divergence, a robust divergence measure, and its applications,” Ph.D. dissertation, University of Florida, 2011.
[7] F. Escolano, E.-R. Hancock, M. Liu, and M.-A. Lozano, “Information-theoretic dissimilarities for graphs,” in Similarity-Based Pattern Recognition, ser. Lecture Notes in Computer Science, 2013, vol. 7953, pp. 90–105.
[8] M. Liu, L. Lu, X. Ye, S. Yu, and H. Huang, “Coarse-to-fine classification via parametric and nonparametric models for computer-aided diagnosis,” in Proc. of the 20th ACM International Conference on Information and Knowledge Management, 2011, pp. 2509–2512.
[9] M. Liu and B.-C. Vemuri, “Robust and efficient regularized boosting using total bregman divergence,” in Proc. of the 24th IEEE CVPR, 2011, pp. 2897–2902.
[10] M. Liu, B.-C. Vemuri, S.-I. Amari, and F. Nielsen, “total Bregman divergence and its applications to shape retrieval,” in Proc. of the 23rd IEEE CVPR, 2010, pp. 3463–3468.
[11] M. Liu, B.-C. Vemuri, and R. Deriche, “A robust variational approach for simultaneous smoothing and estimation of DTI,” NeuroImage, vol. 67, pp. 33 – 41, 2013.
[12] B.-C. Vemuri, M. Liu, S.-I. Amari, and F. Nielsen, “Total bregman divergence and its applications to DTI analysis,” IEEE Transactions on Medical Imaging, vol. 30, no. 2, pp. 475–483, 2011.
[13] W. Stummer and I. Vajda, “On Bregman distances and divergences of probability measures,” IEEE Trans. on Information Theory, vol. 58, pp. 1277–1288, 2012.
[14] S.-I. Amari, “New developments of information geometry (17): Tsallis q-entropy, escort geometry, conformal geometry,” in Mathematical Sciences (saurikagaku). Science Company, October 2012, no. 592, pp. 73–82, in japanese.
[15] S.-I. Amari, “New developments of information geometry (26): Information geometry of convex programming and game theory,” in Mathematical Sciences (saurikagaku). Science Company, November 2013, no. 605, pp. 65–74, in japanese.
[16] M. Liu, B. C. Vemuri, S.-I. Amari, and F. Nielsen, “Shape retrieval using hierarchical total bregman soft clustering,” IEEE T. PAMI, vol. 34, no. 12, pp. 2407–2419, 2012.
[17] S.-I. Amari and H. Nagaoka, Methods of Information Geometry. Oxford University Press, 2000.
[18] F.-H. Clarke, Optimization and Nonsmooth Analysis. Wiley, 1989.
[19] S.-I. Amari, A. Ohara, and H. Matsuzoe, “Geometry of deformed exponential families: invariant, dually flat and conformal geometries,” Physica A, vol. 391, pp. 4308–4319, 2012.
[20] A.-B. Sossinsky, “Tolerance space theory and some applications,” Acta Applicandae Mathematicae, vol. 5, pp. 137–167, 1986.

IX. APPENDIX: PROOFS

A. Proof of Theorem 1

Let us fix \( f(x) = f_p(x) = 1/(1 + \|x\|^p_p)^{1/p} \). We have:

\[
\nabla f(\nabla \varphi(\mu)) = - \frac{1}{(1 + \|\nabla \varphi(\mu)\|^p_p)^{1+1/p}} \nabla_p \varphi(\mu),
\]

(81)
where \( \nabla_p \varphi(\mu) \) is the vector whose \( j^{th} \) coordinate is \( \text{sign}(\nabla^j)|\nabla^j|^{p-1} \), where \( \nabla^j \) is coordinate \( j \) of \( \nabla \varphi(\mu) \). This definition brings the following relationship:

\[
\nabla \varphi(\mu)^\top \nabla_p \varphi(\mu) = \| \nabla \varphi(\mu) \|_p^p.
\]  

(82)

We now use (31) with \( f = f_p \) and obtain:

\[
\mathbf{x} - \mu = -\frac{1}{1 + \| \nabla \varphi(\mu) \|_p^p} \left( \frac{1}{n} \sum_i D \varphi(x_i : \mu) \right) \times \nabla_p \varphi(\mu) .
\]  

(83)

Coordinate \( j \) in \( \mathbf{x} - \mu \), \( (\mathbf{x} - \mu)^j \), satisfies:

\[
((\mathbf{x} - \mu)^j)^{q-1} = \left( \frac{m}{1 + \| \nabla \varphi(\mu) \|_p^p} \right)^{q-1} \times (\text{sign}(\nabla^j)|\nabla^j|^{p-1})^{q-1}
\]

\[
= - \left( \frac{m}{1 + \| \nabla \varphi(\mu) \|_p^p} \right)^{q-1} \times \nabla^j ,
\]  

(84)

where \( m = (1/m) \sum_i D \varphi(x_i : \mu) \geq 0 \). Eq. (84) holds because \((p-1)(q-1) = 1\) and \( q = 2k \in \mathbb{N} \) is even. So, we may write:

\[
\| \mathbf{x} - \mu \|_q^q
\]

\[
= \sum_j ((\mathbf{x} - \mu)^j)^{q-1}
\]

\[
= \sum_j ((\mathbf{x} - \mu)^j)^{q-1} \times (\mathbf{x} - \mu)^j
\]

\[
= - \left( \frac{m}{1 + \| \nabla \varphi(\mu) \|_p^p} \right)^{q-1} \times (\mathbf{x} - \mu)^\top \nabla \varphi(\mu) .
\]  

(85)

We make the inner product of (83) with \( \nabla \varphi(\mu) \) and obtain because of (82):

\[
(\mathbf{x} - \mu)^\top \nabla \varphi(\mu)
\]

\[
= - \left( \frac{\| \nabla \varphi(\mu) \|_p^p}{1 + \| \nabla \varphi(\mu) \|_p^p} \right) \left( \frac{1}{n} \sum_i D \varphi(x_i : \mu) \right) ,
\]

\[
= -\alpha(\mathbf{x} - \varphi(\mu)) + \alpha(\mathbf{x} - \mu)^\top \nabla \varphi(\mu) ,
\]  

(86)

with \( \alpha = \| \nabla \varphi(\mu) \|_p^p/(1 + \| \nabla \varphi(\mu) \|_p^p) \). We obtain \(-(1 - \alpha)(\mathbf{x} - \mu)^\top \nabla \varphi(\mu) = \alpha(\mathbf{x} - \varphi(\mu)) \), that is, after adding \((1 - \alpha)(\mathbf{x} - \varphi(\mu)) \) on both sides:

\[
\mathbf{x} - \varphi(\mu)
\]

\[
= \frac{1}{1 + \| \nabla \varphi(\mu) \|_p^p} \left( \frac{1}{n} \sum_i D \varphi(x_i : \mu) \right) .
\]  

(87)
Fig. 5. We perform a rotation of angle $\theta$ on $\phi$ such that, after rotation, $\phi^{\text{rot}}(\mu^{\text{rot}}) = 0$, implying that $\phi^{\text{rot}}$ is a population minimizer in $P_{\mu^{\text{rot}},\phi^{\text{rot}}}$. We finally get from (85) and (87), using the shorthand $m = (1/n) \sum_i D_{\phi}(x_i : \mu)$:

$$
\|\phi^+ - \mu^+\|_q = \left(\|\phi - \phi(\mu)\|^q + \|\phi - \mu\|^q\right)^{\frac{1}{q}}
\]

$$
= \left(\|\phi - \phi(\mu)\|^{q-1} \times (\|\phi - \phi(\mu)\| + \|\phi - \mu\|^{\frac{1}{q}})\right)^{\frac{1}{q}}
\]

$$
= \left(\frac{m}{1 + \|\nabla \phi(\mu)\|^p}\right)^{\frac{q-1}{q}} \left(\|\phi - \phi(\mu)\| + \|\phi - \mu\|^q\right)^{\frac{1}{q}}
\]

$$
= \frac{m}{(1 + \|\nabla \phi(\mu)\|^p)^{\frac{1}{q}}} \times (\|\phi - \phi(\mu)\| + \|\phi - \mu\|^q)\left(m + \frac{1}{q}\right)^{\frac{1}{q}}
\]

$$
= \frac{1}{n} \sum_i D_{\phi,K}(x_i : \mu)
\]

$$
= \frac{1}{K} \times \left(\frac{1}{n} \sum_i D_{\phi,K}(x_i : \mu)\right),
\]

which yields the statement of the Theorem.
B. Proof of Theorem 2

We distinguish two cases, first assuming that \( \mathbf{x} \not\in \mathbb{F}_{\mathcal{S}, \varphi} \). As shown in Figure 5, we perform a rotation of angle \( \theta \) chosen so that
\[
\frac{1}{n} \sum_i M_\theta \mathbf{x}_i = M_\theta \left( \frac{1}{n} \sum_i \mathbf{x}_i \right) = M_\theta \mathbf{\mu}^+, \quad \text{with}
\]
\[
M_\theta = \begin{bmatrix}
\cos \theta & -\sin \theta \\
\sin \theta & \cos \theta
\end{bmatrix},
\]
and \( \mathbf{x}_i^+ \) and \( \mathbf{\mu}^+ \) are defined in (19) and (20). So, the population minimizer of \( \mathcal{S} \) after rotation is just the average, which implies, since the distortion is invariant to rotation after assumption (iii) and satisfies (i) and (ii), that the distortion equals a Bregman divergence computed after rotation, \( D_{\varphi^{\text{rot}}} \), for some convex differentiable \( \varphi^{\text{rot}} : \mathbb{R} \rightarrow \mathbb{R} \). We then get:
\[
D(x : \mu) = D((M_\theta \mathbf{x}^+)^0 : (M_\theta \mathbf{\mu}^+)^0) = D_{\varphi^{\text{rot}}}((M_\theta \mathbf{x}^+)^0 : (M_\theta \mathbf{\mu}^+)^0) = \varphi^{\text{rot}}((M_\theta \mathbf{x}^+)^0) - \varphi^{\text{rot}}((M_\theta \mathbf{\mu}^+)^0) \quad \text{and} \quad ((M_\theta \mathbf{x}^+)^0 - (M_\theta \mathbf{\mu}^+)^0) \varphi^{\text{rot}}((M_\theta \mathbf{\mu}^+)^0).
\]

Let us denote \( \varphi : \mathbb{R}^d \rightarrow \mathbb{R} \) the function obtained from \( \varphi^{\text{rot}} \) by rotation \( M_\theta \) of the curve. We have:
\[
\varphi^{\text{rot}}((M_\theta \mathbf{x}^+)^0) = (M_\theta \mathbf{x}^+)^1 = x \sin \theta + \phi(x) \cos \theta,
\]
\[
\varphi^{\text{rot}}((M_\theta \mathbf{\mu}^+)^0) = (M_\theta \mathbf{\mu}^+)^1 = \mu \sin \theta + \phi(\mu) \cos \theta,
\]
and
\[
((M_\theta \mathbf{x}^+)^0 - (M_\theta \mathbf{\mu}^+)^0) \varphi^{\text{rot}}((M_\theta \mathbf{\mu}^+)^0) = (x \cos \theta - \phi(x) \sin \theta - \mu \cos \theta + \phi(\mu) \sin \theta) \times \frac{\sin \theta + \phi'(\mu) \cos \theta}{\cos \theta - \phi'(\mu) \sin \theta}.
\]

Eq. (88) thus becomes
\[
D(x : \mu) = Q + R, \quad \text{with} \quad Q = (\phi(x) - \phi(\mu)) \cos \theta, \quad \text{and} \quad R = (x - \mu) \sin \theta - \left\{ x \cos \theta - \phi(x) \sin \theta \right\} \times \frac{\sin \theta + \phi'(\mu) \cos \theta}{\cos \theta - \phi'(\mu) \sin \theta}.
\]
We thus get:

\[
D(x : \mu) = \frac{\phi(x) - \phi(\mu) - (x - \mu)\phi'(\mu)}{\cos \theta - \phi'(\mu) \sin \theta}
\]

with

\[
g(\mu) = \sqrt{\frac{1 + \phi'(\mu)^2}{1 + \phi(\mu)\phi'(\mu)}},
\]

as indeed \(\cos \theta = 1/\sqrt{1 + \phi'(\mu)^2}\) and \(\sin \theta = -\phi'(\mu)/\sqrt{1 + \phi'(\mu)^2}\).

Eqs (89) and (90) are the consequences of assumptions (i-iv). On the other hand, assumption (iv) and (45) imply for right population minimizer \(\mu\) of set \(S\):

\[
-\phi'(\mu) (\overline{x} - \phi(\mu)) = (\overline{x} - \mu).
\]

Since \(\overline{x} \neq \mu\), we obtain \(\phi'(\mu) \neq 0\), so we can replace \(\phi'(\mu)\) in (90) by its expression from (91) and obtain:

\[
g(\mu) = \sqrt{\frac{(\overline{x} - \phi(\mu))^2}{1 - \phi(\mu) \times (\overline{x} - \mu)\phi'(\mu)}}
\]

\[
= \frac{\sqrt{(\overline{x} - \phi(\mu))^2 + (\overline{x} - \mu)^2}}{\overline{x} - \phi(\mu) - (\overline{x} - \mu)\phi'(\mu)}
\]

\[
= \frac{\|\overline{x}^+ - \mu^+\|_2}{\overline{x} - \phi(\mu) - (\overline{x} - \mu)\phi'(\mu)},
\]

where \(\overline{x}^+\) and \(\mu^+\) are defined in (19) and (20). For any \(S\) whose right population minimizer on \(D_{\phi,g}\) is \(\mu\), we get:

\[
\frac{1}{n} \sum_i D(x_i : \mu)
\]

\[
= \|\overline{x}^+ - \mu^+\|_2 \times \frac{\phi - \phi(\mu) - (\overline{x} - \mu)\phi'(\mu)}{\overline{x} - \phi(\mu) - (\overline{x} - \mu)\phi'(\mu)}
\]
Because of assumption (iii), we want (92) to be invariant to rotation of the axes. Only term $\|\mathbf{V} - \mu^+\|_2$ is invariant because of assumption (iv). Both the numerator and the denominator after the times in (92) are not invariant to rotation. To have their ratio invariant, it must therefore be independent from the choice of $S$, and thus constant, so we have:

$$\bar{\phi} - \phi(\mu) - (\mathbf{V} - \mu)\phi'(\mu) = K(\mathbf{V} - \varphi(\mu) - (\mathbf{V} - \mu)\phi'(\mu)),$$

Taking the derivative in some $x_i$ yields $\phi'(x_i) = K\phi'(x_i) - (K - 1)x_i\phi'(\mu)$, which implies that the right hand side is independent of $\mu$, and since $\phi'$ cannot always be zero, we obtain $K = 1$, and so:

$$\phi = \varphi + \text{constant}.$$

We obtain $D_{\phi,g} = D_{\varphi,g_{\perp}}$, and this completes the proof when $\mathbf{V} \not\in \mathbb{P}_{S,\varphi}$.

If $\mathbf{V} \in \mathbb{P}_{S,\varphi}$ is the population minimizer, then $D(x : \mu) = D(x : \mathbf{V})$ is a Bregman divergence [4], [5], say $D_{\phi}$ for some $\phi$ strictly convex twice differentiable. Because of assumption (iii), it comes in this case:

$$D(x : y) = \sqrt{\frac{1 + \phi'(x)^2}{1 + \phi'(y)^2}} D_{\phi}(x : y) \quad \text{for some } \mathbf{V} \not\in \mathbb{P}_{S,\varphi}.$$

Because of assumption (iv) and Lemma 5, $\phi' = \varphi'$ and so $\phi = \varphi + \text{constant}. This completes the proof in this second case, and completes the proof of Theorem 2.

C. Proof of Lemma 8

The proof of this Lemma relies on the following Taylor expansions:

$$v(\mu + \epsilon \delta_\mu) = v(\mu) + \epsilon J_v \delta_\mu,$$

for some value $J_v$ of the Jacobian of $v$ in between $\mu$ and $\mu + \epsilon \delta_\mu$,

$$v(x_\ast) = v(\mu) + J_v'(x_\ast - \mu),$$

for some value $J_v'$ of the Jacobian of $v$ in between $\mu$ and $x_\ast$, and

$$\varphi(v(\mu + \epsilon \delta_\mu))$$

$$\begin{align*}
&= \varphi(v(\mu)) + (v(\mu + \epsilon \delta_\mu) - v(\mu))^\top \nabla \varphi(v(\mu)) \\
&\quad + \frac{1}{2} (v(\mu + \epsilon \delta_\mu) - v(\mu))^\top H_1(v(\mu + \epsilon \delta_\mu) - v(\mu)) \\
&= \varphi(v(\mu)) + \epsilon J_v^\top \nabla \varphi(v(\mu)) \\
&\quad + \frac{\epsilon^2}{2} J_v^\top H_1 J_v \delta_\mu \\
&= \varphi(v(\mu)) + \epsilon J_v^\top u(\mu) + \frac{\epsilon^2}{2} \delta_\mu^\top H_1 J_v \delta_\mu,
\end{align*}$$

(97)
for some value $H_1$ of the Hessian of $H_\varphi$ in between $v(\mu)$ and $v(\mu+\epsilon\delta_\mu)$. We have made use of (94) in (96) and the fact that $\nabla \varphi \circ v = u$ in (97).

According to Theorem 3, the population minimizers of $S$ to which we add $\mu$ with a weight of $\epsilon$ satisfy $\delta_{v,\epsilon}' = x_{v,\epsilon}' - \mu_{v,\epsilon}'$ and:

$$x_{v,\epsilon}' = (1 - \epsilon)x_v' + \epsilon \left[ \begin{array}{c} v(x_v') \\ \varphi(v(x_v')) \end{array} \right]$$

$$= (1 - \epsilon)x_v' + \epsilon x_{v,\epsilon}'$$

$$\mu_{v,\epsilon}' = \left[ \begin{array}{c} v(\mu + \epsilon\delta_\mu) \\ \varphi(v(\mu + \epsilon\delta_\mu)) \end{array} \right] = \mu_v' + \epsilon \delta_\mu'$$

$$\delta_{\mu}' = \left[ \begin{array}{c} J_v\delta_\mu \\ \delta_{\mu}' J_v^\top u(\mu) + \frac{1}{2} \delta_{\mu}' J_v^\top H_1 J_v \delta_\mu \end{array} \right].$$

(See (53) for the definitions of $x_{v,\epsilon}', \mu_{v,\epsilon}'$) In (100), we have used (94) and (97). We obtain:

$$0 = (\delta_{v,\epsilon}')^\top z_u^+$$

$$= (1 - \epsilon)(x_v' - \mu_{v,\epsilon}')^\top z_u^+ + \epsilon(x_{v,\epsilon}' - \mu_{v,\epsilon}')^\top z_u^+$$

$$= (1 - \epsilon)(\delta_{\mu}')^\top z_u^+ - \epsilon(1 - \epsilon)(\delta_{\mu}')^\top z_u^+$$

$$+ \epsilon(x_{v,\epsilon}' - \mu_{v}')^\top z_u^+ - \epsilon^2(\delta_{\mu}')^\top z_u^+$$

$$= \epsilon((x_{v,\epsilon}' - \mu_{v}')^\top z_u^+ - (\delta_{\mu}')^\top z_u^+) .$$

In (102), we have used the fact that $(\delta_{v}')^\top z_u^+ = 0$ since $\mu$ is a right population minimizer for the $v$-conformal divergence on $S$. Since $\epsilon \neq 0$, we obtain from (102) the equation which is central to the proof of Lemma 8

$$(\delta_{\mu}')^\top z_u^+ = (x_{v,\epsilon}' - \mu_{v}')^\top z_u^+ .$$

We now work on this equation. Looking at $z_u^+$ in (54), we observe that $\|z_u^+\|_2 \neq 0$, $\forall z \neq 0$. To see this, for $\|z_u^+\|_2 = 0$, we first need $\nabla f(u(\mu))^\top z = 0$, and this implies $f(u(\mu)) \times z = 0$, which implies $z = 0$. So, assuming that we pick $z \neq 0$, we can simplify (103) and obtain

$$\|\delta_{\mu}'\|_2^2 \leq \frac{1}{\cos^2(\delta_{\mu}' , z_u^+)} \times \|x_{v,\epsilon}' - \mu_{v}'\|_2^2 , \forall z \neq 0 .$$

We now find a $z \neq 0$ with which the inverse square cosine is small. To find this $z$, we use this intermediate result, (P):

(P) let $a, b, c \in \mathbb{R}^d$. The solution $z$ to the equation $z = a + c \times b$ is:

$$z = \begin{cases} a + \frac{c}{1 - c^2} b & \text{if } c \neq 0 \wedge b \neq ||c||_2 c \\ a & \text{if } c = 0 \end{cases}.$$
We use \((P)\) with the following vectors:

\[
\begin{align*}
a & \doteq \frac{1}{f(u(\mu))} \times J_v \delta \mu, \\
b & \doteq -\frac{1}{f(u(\mu))} \times u(\mu), \\
c & \doteq \nabla f(u(\mu)) .
\end{align*}
\]

When \(\|\nabla f(u(\mu))\|_2 \neq 0\), we need to check if it can be possible that \(b = \|c\|_2^2 c\). For this to happen from the definitions of \(b\) and \(c\), we need \(\nabla f = \alpha \text{Id}\) for some \(\alpha \neq 0\), implying \(f(z) = \alpha \|z\|_2^2 / 2\), which cannot be the case from assumptions (i) and (ii).

Let us analyze the two cases of \((P)\), starting from the case \(c \neq 0\). We have:

\[
z = \frac{1}{f(u(\mu))} \times \left( J_v \delta \mu - \frac{\nabla f(u(\mu))^T J_v \delta \mu}{f(u(\mu)) + \nabla f(u(\mu))^T u(\mu)} \times u(\mu) \right)
\]

which yields:

\[
z^+_u \doteq \begin{bmatrix} J_v \delta \mu \\ -\frac{\nabla f(u(\mu))^T J_v \delta \mu}{f(u(\mu)) + \nabla f(u(\mu))^T u(\mu)} \end{bmatrix} . \tag{105}
\]

Let us define:

\[
x \doteq \frac{1}{\|J_v \delta \mu\|_2} \times \frac{\nabla f(u(\mu))^T J_v \delta \mu}{f(u(\mu)) + \nabla f(u(\mu))^T u(\mu)} , \\
y \doteq \frac{1}{\|J_v \delta \mu\|_2} \times \left( \delta \mu^T J_v u(\mu) + \frac{\epsilon}{2} \delta \mu^T J_v^T H_1 J_v \delta \mu \right) .
\]

Plugging the expression of \(z^+_u\) in \(1 \cos^2(\delta^+_\mu, z^+_u)\), we obtain after simplification:

\[
\frac{1}{\cos^2(\delta^+_\mu, z^+_u)} = 1 + \frac{(x + y)^2}{1 + x^2 + y^2 + 2x^2y^2} \leq 2 . \tag{106}
\]

We obtain the following upperbound on \(\|\delta^+_\mu\|_2^2\) refined from \((103)\):

\[
\|\delta^+_\mu\|_2^2 \leq 2\|x^+_v - \mu^+_v\|_2^2 . \tag{107}
\]

Handling the second case for \((P)\) is simpler, as since \(c = \nabla f(u(\mu)) = 0\), picking \(z = (1 / f(u(\mu))) \times J_v \delta \mu\) yields \((106)\) with \(x = 0\), and \((107)\) is still valid. To finish up with the proof, we first upperbound the right-hand side of \((107)\), as:

\[
\|x^+_v - \mu^+_v\|_2^2 = \|v(x_*) - v(\mu)\|_2^2 + \varphi(v(x_*)) - \varphi(v(\mu)))^2 \leq \|v(x_*) - v(\mu)\|_2^2 (1 + L^2) \tag{108}
\]

\[
= (x_* - \mu)^T (J_v)^T J_v(x_* - \mu)(1 + L^2) \tag{109}
\]

\[
\leq \lambda_{\text{max}} (1 + L^2) \|x_* - \mu\|_2^2 . \tag{110}
\]
We then compute the partial derivatives in \( y \). We now use (112), letting 
\[
\frac{\partial}{\partial \mu} J_v^T J_v \delta \mu = \delta_v^T J_v^T J_v \delta \mu + \frac{\lambda}{2} \delta_v^T J_v^T H_1 J_v \delta \mu
\]
and so 
\[
\delta_v^T J_v^T J_v \delta \mu \geq \lambda_{\text{min}} \| \delta \mu \|^2_2 , \tag{111}
\]
for some non-zero lowerbound \( \lambda_{\text{min}} \) of the eigenvalues of \( J_v^T J_v \). We then obtain from (107), (110) and (111):
\[
\| \delta \mu \|^2_2 \leq \frac{2 \lambda_{\text{max}} (1 + L^2)}{\lambda_{\text{min}}} \times \| x_* - \mu \|^2_2 .
\]

using the definition of \( \lambda_v \). We obtain \( \| \delta \mu \|_2 \leq \sqrt{\lambda_v (L)} \| x_* - \mu \|_2 \) for \( \ell(L) = \sqrt{2} (1 + L) \geq \sqrt{2} (1 + L^2) \), as claimed.

D. Proof of Lemma 70

First, since \( \varphi' = v \circ u^{-1} \), we get 
\[
\varphi'' = \frac{v' \circ u^{-1}}{v' \circ u^{-1}} , \tag{112}
\]
and so (ii) would be a consequence of (iii). Let us compute the equality of partial derivatives in \( x \) of (72):
\[
g(u(y))v'(x)(u(x) - u(y)) = u'(x)g'(u(x))D_\varphi(v(y) : v(x)) - g(u(x))u'(x)(v(y) - v(x)) .
\]
We then compute the partial derivatives in \( y \) and reorganize:
\[
u'(x)v'(y)[g'(u(x))(u(y) - u(x)) - g(u(x))] = u'(y)v'(x)[g'(u(y))(u(x) - u(y)) - g(u(y))] . \tag{113}
\]
We now use (112), letting \( z = u(x) \) and \( \bar{r} = u(y) \), so that (113) becomes:
\[
\varphi''(r)[g'(z)(r - z) - g(z)] = \varphi''(z)[g'(r)(z - r) - g(r)] . \tag{114}
\]
Let us fix temporarily \( z \) to a constant, so that (114) is a function of \( r \), and thus reads:
\[
g'(z)(r - z) - g(z) = \varphi''(z) \times \varphi(r, z) , \tag{115}
\]
\[
\varphi(r, z) = \frac{g'(r)(z - r) - g(r)}{\varphi''(r)} = \frac{g'(r)z - g(r) + rg'(r)}{\varphi''(r)} . \tag{116}
\]
Because the left hand-side of (115) is linear in \( r \), so has to be \( \vartheta \) in (116), and so we get:

\[
\frac{g'(r)}{\varphi''(r)} = ar + b ,
\]

\[
\frac{g(r) + rg'(r)}{\varphi''(r)} = cr + d ,
\]

for some \( a, b, c, d \in \mathbb{R} \) that are constant since \( z \) is fixed; our objective is to prove that all but \( d \) are zero, so let us proceed by assuming that all are non zero. Substituting \( g'(r) \) from (117) in (118) yields:

\[
\varphi''(r) = g(r) f_1(r) ,
\]

\[
f_1(r) = -ar^2 + (c - b)r + d .
\]

Since \( a \neq 0 \), \( f_1 \) is the equation of a parabola. Using (117), we see that \( g \) is solution of the following homogeneous differential equation:

\[
(ar + b)g(r) - f_1(r)g'(r) = 0 ,
\]

whose solution is found to be, for any constant \( K_5 \):

\[
g(r) = K_5 \exp \left( \int \frac{ar + b}{f_1(r)} \right) = K_5 \sqrt{-f_1(r)} \left( \sqrt{\frac{\sqrt{K_6} + f_2(r)}{\sqrt{K_6} - f_2(r)}} \right)^{\frac{f_2(c/b)}{\sqrt{K_6}}} ;
\]

\[
K_6 = 4ad + (b - c)^2 ;
\]

\[
f_2(r) = 2ar + (b - c) .
\]

Eq. (122) implies \( K_6 \geq 0 \). For \( g \) as in (122) to exist, we have two more constraints to meet: (a) \(-f_1(r) > 0\) and (b) \((\sqrt{K_6} + f_2(r))/(\sqrt{K_6} - f_2(r)) \geq 0\). We distinguish two cases:

\((a > 0)\) To meet constraint (a), we need \( r > ((c - b) + \sqrt{K_6})/(2a) \) or \( r < ((c - b) - \sqrt{K_6})/(2a) \). In both cases, constraint (b) is violated as respectively the denominator or the numerator (only) of the fraction is strictly negative.

\((a < 0)\) To meet constraint (a), we need \(((c - b) - \sqrt{K_6})/(2a) < r < ((c - b) + \sqrt{K_6})/(2a) \). Again, constraint (b) is violated.

We end up with the conclusion that \( a = 0 \) so that \( f_1 \) is linear. Assume now that \( c \neq b \). The new solution to (121) is:

\[
g(r) = K_5 ((c - b)r + d)^{\frac{1}{c - b}} ,
\]

leading through (117) to:

\[
\varphi''(r) = K_5 ((c - b)r + d)^{\frac{1}{c - b} - 1} .
\]
This enforces $K_5 > 0$, but $\varphi''(-d/(c - b)) = 0$, which is not possible as $\varphi$ must be strictly convex. Hence $a = 0$ and $c = b$, so that $f_1(r) = d = f_1$ is a constant.

To finish up the proof, we consider the assumption $b \neq 0$. The new solution to (121) is:

$$g(r) = K_5 \exp(br/d),$$

leading through (117) to:

$$\varphi''(r) = \frac{K_5}{d} \exp(br/d),$$

enforcing this time $K_5/d > 0$.

Now, let us start back from (114), considering $b, c, d$ functions of $z$. We simplify (114) using (117), (118) and the expressions of $g$ and $\varphi''$ in (124) and (125), and obtain $b(r)r - (c(r)z + d(r)) = b(z)z - (c(z)r + d(z))$, that is, since $c = b$:

$$b(r)(r - z) - d(r) = b(z)(z - r) - d(z),$$

or, equivalently, for $z \neq r$,

$$\frac{d(z) - d(r)}{z - r} = b(z) + b(r).$$

This shows that $d$ is derivable, and its derivative satisfies $d'(z) = 2b(z)$, and so:

$$d(z) = 2 \int b(z) + K_7,$$

for any constant $K_7$. We put this expression in (126), differentiate in $r$ and obtain $b'(r)(r - z) + b(r) - 2b(r) = -b(z)$, that is, after reordering, $b(z) = b(r) + (z - r)b'(r)$, and so:

$$b(z) = K_8z + K_9,$$

for any constants $K_8$ and $K_9$. Plugging this in (126) using (128) yields the identity, valid for any $z$ and $r$:

$$(K_8r + K_9)(r - z) - (K_8r^2/2 + K_9r + K_2) = (K_8z + K_9)(z - r) - (K_8z^2/2 + K_9z + K_2).$$

Its simplification yields:

$$K_8(z - r)(z + r) = 2K_9(z - r), \forall z, r.$$  

This implies $K_8 = K_9 = 0$, and finally the solutions to (117) and (118) are $a = b = c = 0$ and $d = K_7$, constant. We obtain $g(r)$ constant as in (i) through (117), $\varphi''$ constant through (118) — and the expression of $\varphi$ as in (ii) —, and finally $u = v$ through (112), as claimed.