On Conformal Divergences and their Population Minimizers

Richard Nock, Frank Nielsen, Shun-ichi Amari

Abstract—Total Bregman divergences are a recent tweak of ordinary Bregman divergences originally motivated by applications that required invariance by rotations. They have displayed superior results compared to ordinary Bregman divergences on several clustering, computer vision, medical imaging and machine learning tasks. These preliminary results raise two important problems : First, report a complete characterization of the left and right population minimizers for this class of total Bregman divergences. Second, characterize a principled superset of total and ordinary Bregman divergences with good clustering properties, from which one could tailor the choice of a divergence to a particular application. In this paper, we provide and study one such superset with interesting geometric features, that we call conformal divergences, and focus on their left and right population minimizers. Our results are obtained in a recently coined (u, v)-geometric structure that is a generalization of the dually flat affine connections in information geometry. We characterize both analytically and geometrically the population minimizers. We prove that conformal divergences (resp. total Bregman divergences) are essentially exhaustive for their left (resp. right) population minimizers. We further report new results and extend previous results on the robustness to outliers of the left and right population minimizers, and discuss the role of the (u, v)-geometric structure in clustering. Additional results are also given.

Index Terms—Ordinary Bregman divergences, total Bregman divergences, (u, v)-geometric structure.

1 Introduction

Loosely defined in its most general form, the clustering problem is related to the grouping of a data sample according to *unknown* classes or clusters [1]. One of the most popular and well-posed approaches to clustering is centroid-based: it seeks to summarize data into a fixed set of cluster centers — or *population minimizers* — that best describe the sample, where "best" is understood with respect to an expected measure of distortion to the whole sample [2], [3]. Because of their convexity properties and links to likelihoods in exponential families, ordinary Bregman divergences are often used to compute these distortions in clustering algorithms, such as in k-means and EM [4], [2], [3]. Their left and right population minimizers are respectively a cluster's f-mean [5], [6] and the cluster's average (ordinary Bregman divergences are in general not symmetric). It has been shown that modulo technical assumptions, ordinary Bregman divergences are exhaustive for their right population minimizer: any divergence whose population minimizer is the sample average is a Bregman divergence [7], [8]. 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. .

There has been a recent burst of interest in a new class of divergences, built from Bregman divergences, known as total Bregman divergences [9], [10], [11], [12], [13], [14], [15], [16], [17], [18]. 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 in domains like DTI interpolation and segmentation [9]. These results exploit the fact that the left population minimiz-

- *R. Nock (**contact author**) is with NICTA, Australia. E-mail:richard.nock@nicta.com.au.*
F. Nielsen is with Ecole Polytechnique, France. E-mail:nielsen@lix.polytechnique.fr.
S.-i. Amari is with Riken Brain Science Institute, Japan. E-mail:amari@brain.riken.jp.

ers of total Bregman divergences are weighted generalized f-means [9]. In the general context of clustering, and also for reasons related to statistics and maximum likelihood estimation [19], 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 the context of clustering, the results of [9] also contribute to the advocacy that clustering is in fact a domain dependent method [1], thereby raising the question of how we may generalize further the set of candidate (total or ordinary) Bregman divergences, while keeping good properties, from which one may select the best candidates to solve a particular problem.

In this paper, we address these questions in a setting which generalizes in two ways total and ordinary Bregman divergences. First, we consider a superset of total Bregman divergences and ordinary Bregman divergences that we define as conformal divergences. Second, we consider a coordinate system which is not the usual dually flat affine coordinate system of (total, ordinary) Bregman divergences, but a generalization in information geometry studied by Zhang and Amari, defined as the (u, v)-geometric structure [20], [21], [22], in which two coordinate mappings u and v define the gradient (and its reciprocal inverse) 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 [9].

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) [19].

The paper is structured as follows. The following Section gives definitions. Section 3 compares the various notions of divergences we consider. Section 4 is devoted to left population minimizers of conformal divergences in the (u, v)-geometric structure. Section 5 does the same for right population minimizers. Section 6 studies the robustness of the population minimizers and Section 7 discusses our results. A last section concludes. In order not to laden the paper's body, some proofs are given in an Appendix in Section 10. Other proofs have been included in a techreport with proofs [23].

2 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 $x^1, x^2, ..., x^d$ for vector $\boldsymbol{x} \in \mathbb{R}^d$, where $d \geq 1$.

A (right-sided) *conformal divergence*, $D_{\varphi,g}$, is parameterized by two real-valued functions φ and g with $\mathrm{img} \subseteq (0, +\infty)$, whose domains are a compact convex of $\mathbb{R}^d$. The expression of $D_{\varphi,g}$ is:

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

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

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

$\nabla\varphi$ denotes the gradient of φ. g admits continuous directional derivatives: function $\mathrm{D}_{\boldsymbol{z}} g(\boldsymbol{x}) \doteq \lim_{t\rightarrow 0} \mathrm{D}_{t,\boldsymbol{z}} g(\boldsymbol{x})$, defining directional derivatives, is continuous and exist for any

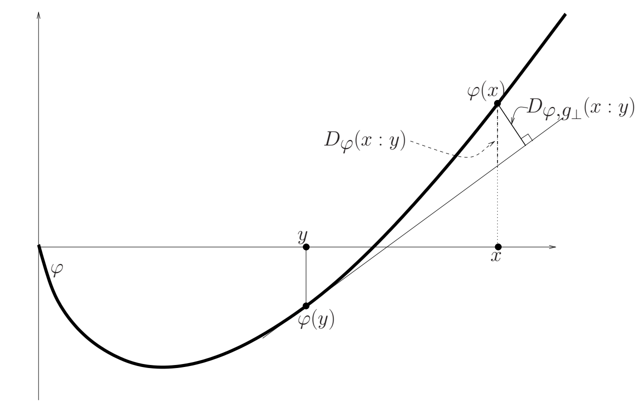

Fig. 1. Depiction of $D_\varphi(x:y)$ and $D_{\varphi,g_\perp}(x:y)$ when $\varphi(x) = x\ln(x) - x$.

valid direction $\boldsymbol{z}$ such that $\mathrm{D}_{t,\boldsymbol{z}}g(\boldsymbol{x})$ is defined in a neighborhood of 0 (with respect to t). We give:

$$\mathrm{D}_{t,\boldsymbol{z}}g(\boldsymbol{x}) \doteq \frac{g(\boldsymbol{x}+t\boldsymbol{z}) - g(\boldsymbol{x})}{t} .$$

Ordinary 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 $g = Kg_\perp$ for some constant $K > 0$ and :

$$g_\perp(\boldsymbol{y}) \doteq \frac{1}{\sqrt{1+\|\nabla\varphi(\boldsymbol{y})\|_2^2}} , \quad (3)$$

which defines total Bregman divergences, that are invariant to rotations of the coordinate axes [10], [12], [13], [17], [14], [15], [16] (among others). Table 2 presents some examples of total Bregman divergences (with $K = 1$). Remark that $g_\perp(\boldsymbol{y})$ is of the form $f_\perp(\nabla\varphi(\boldsymbol{y}))$, with

$$f_\perp(\boldsymbol{x}) \doteq \frac{1}{\sqrt{1+\|\boldsymbol{x}\|_2^2}} . \quad (4)$$

Figure 1 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, $\forall p \geq 1$:

$$g_p(\boldsymbol{y}) \doteq f_p(\nabla\varphi(\boldsymbol{y})) ; \quad (5)$$

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

The p-norm of $\boldsymbol{x}$ is $\|\boldsymbol{x}\|_p \doteq (\sum_i |x^i|^p)^{\frac{1}{p}}$.

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 $D^v_{\varphi,g}$ as:

$$D^v_{\varphi,g}(\boldsymbol{x}:\boldsymbol{y}) \doteq g(\boldsymbol{y})D_\varphi(v(\boldsymbol{x}):v(\boldsymbol{y})) . \quad (7)$$

v-conformal divergences are inspired by divergences in the (u,v) geometric structure [20], [21] (see also Section 7). They generalize conformal divergences for which $v = \mathrm{Id}$. 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 ordinary Bregman divergences [24]. Function g in v-conformal divergences depends on the right parameter of the divergence. We shall see (Lemma 7) that when g is not constant, the v-conformal divergence cannot be symmetric: $g(\boldsymbol{y})D_\varphi(v(\boldsymbol{x}):v(\boldsymbol{y})) \neq g(\boldsymbol{y})D_\varphi(v(\boldsymbol{y}):v(\boldsymbol{x}))$. However, our results extend at little cost to left-sided conformal divergences, *i.e.* whose regularization factor g depends on the left parameter of the divergence. Indeed, calling to convex conjugates, we obtain:

$$\begin{aligned} & D^v_{\varphi,g}(\boldsymbol{x}:\boldsymbol{y}) \\ &= g(\boldsymbol{y})D_\varphi(v(\boldsymbol{x}):v(\boldsymbol{y})) \\ &= g(\boldsymbol{y})D_{\varphi^\star}((\nabla\varphi\circ v)(\boldsymbol{y}):(\nabla\varphi\circ v)(\boldsymbol{x})) \\ &= g(\boldsymbol{y})D_{\varphi^\star}(u(\boldsymbol{y}):u(\boldsymbol{x})) , \end{aligned}$$

where $u \doteq \nabla\varphi\circ v$ also defines a coordinate mapping and $\varphi^\star$ is the convex conjugate of φ. Any such coordinate mappings u and v such that $u\circ v^{-1}$ defines the gradient of a strictly convex differentiable function φ is called an (u,v)-geometric structure [20], that we write $(u,v)_\varphi$ from now on, to make explicit the reference to φ.

We now define population minimizers for conformal divergences.

Definition 1: (left- and right-population minimizers) Let $\mathcal{S} \doteq \{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n\}$, with $\boldsymbol{x}_i \in \mathbb{R}^d, \forall i = 1, 2, ..., n$. Let $D : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be the shorthand for some ordinary Bregman (resp. conformal, resp. v-conformal) divergence D_φ (resp. $D_{\varphi,g}$, resp. $D^v_{\varphi,g}$). A *left population*

minimizer for D on $\mathcal{S}$ is any $\boldsymbol{\mu}$ such that $\sum_i D(\boldsymbol{\mu} : \boldsymbol{x}_i) = \min_{\boldsymbol{x}} \sum_i D(\boldsymbol{x} : \boldsymbol{x}_i)$. A *right population minimizer* for D on $\mathcal{S}$ is any $\boldsymbol{\mu}$ such that $\sum_i D(\boldsymbol{x}_i : \boldsymbol{\mu}) = \min_{\boldsymbol{x}} \sum_i D(\boldsymbol{x}_i : \boldsymbol{x})$.
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.

3 On divergences: ordinary, total and conformal

There is a need to generalize the source of divergences from which efficient centroid-based clustering algorithms may be derived. The comparison between ordinary Bregman and total Bregman divergences is enlightening from that standpoint: ordinary Bregman divergences D_φ are *axiomatically* characterized as the unique family of divergences (under mild conditions) that yield their right population minimizers matching the *sample average* [8]. Hence, the right population minimizer is the simplest to compute, *but* having fixed the data sample, regardless of the generator of the divergence φ, it is always the *same*. From a clustering standpoint, it may be more intuitive that since changing the generator changes the geometry of the problem, it should possibly change this population minimizer as well. Also, this invariance is not convenient to further optimize the population minimizer by tuning the divergence at hand.

This problem does not appear anymore with total Bregman divergences. Initially, total Bregman divergences [9] $D_{\varphi,g_\perp}$ have been geometrically designed to enforce invariance by rotations in the parameter space [9], thus mimicking the ordinary/total least squares relationships. Rotation invariance is a very desirable property in medical imaging [16] and computer vision [18]. Thus, total Bregman divergences have been specifically engineered to solve a particular geometric problem, which has led to improved results on several key applications related to clustering. Besides, total Bregman divergences have also proven *experimentally* superior in boosting [13] and tensor-based graph matching [11], etc., just to name a few. One theoretical argument that explains the superiority of total Bregman divergences was detailed in [17], where it was proved that total Bregman divergences are robust compared to ordinary Bregman divergences, by studying the impact of outliers via the influence function.

The difference between total and ordinary Bregman divergences can also be captured from a statistical standpoint. It is well-known that regular exponential families $p(\boldsymbol{x};\boldsymbol{\theta}) = h(\boldsymbol{x})\exp(\boldsymbol{\theta}^\top \boldsymbol{t}(\boldsymbol{x}) - \varphi(\boldsymbol{\theta}))$ are in bijection with (regular) Bregman divergences [4], $p(\boldsymbol{x};\boldsymbol{\theta}) = h(\boldsymbol{x})\exp(-D_{\varphi^\star}(\boldsymbol{t}(\boldsymbol{x}) : \boldsymbol{\eta}(\boldsymbol{\theta})))$ where $\boldsymbol{\eta}(\boldsymbol{\theta}) = \nabla\varphi(\boldsymbol{\theta})$ is the dual moment parameter, and that the Maximum Likelihood Estimator (MLE) for n identically and independently distributed observations of an exponential family coincides with the so-called *observed point* in information geometry [24]: $\bar{\boldsymbol{t}} = (1/n)\sum_i \boldsymbol{t}(\boldsymbol{x}_i)$, where $\boldsymbol{t}(.)$ denotes the vector of sufficient statistics of the exponential families under consideration. That is, the MLE $\hat{\boldsymbol{\eta}}$ expressed in the $\boldsymbol{\eta}$-parameter matches the centroid of sufficient statistics: $\hat{\boldsymbol{\eta}} = \bar{\boldsymbol{t}}$. Since there is also a bijection between ordinary and total Bregman divergences, we deduce by transitivity with the ordinary Bregman-exponential family bijection that we can associate an exponential family $p(\boldsymbol{x};\boldsymbol{\theta}) = h(\boldsymbol{x})\exp(-D_{\varphi^\star,g_\perp}(\boldsymbol{t}(\boldsymbol{x}) : \boldsymbol{\eta})\sqrt{1+\|\nabla\varphi^\star(\boldsymbol{\eta})\|^2})$ to any total Bregman divergence $D_{\varphi^\star,g_\perp}$. This statistical distribution $p(\boldsymbol{x};\boldsymbol{\theta})$ corresponds also to a *lifted exponential family* $\tilde{p}(\tilde{\boldsymbol{x}};\tilde{\boldsymbol{\theta}})$ in disguise, as exemplified in [17] with $\tilde{\boldsymbol{\theta}} = (\boldsymbol{\theta},\varphi(\boldsymbol{\theta}))$ and $\tilde{\boldsymbol{x}} = (1/\sqrt{1+\|\nabla\varphi(\boldsymbol{x})\|^2})\cdot(\boldsymbol{x},1)$. In other words, the ordinary exponential family is lifted to the space having one extra dimension and embedded as a hypersurface. Now, it can be proved that the *total (left) observed point* is the Bayesian MAP estimator in the lifted exponential family with prior distribution $\pi(\boldsymbol{\theta}) = \exp(-n\tilde{\varphi}(\boldsymbol{\theta}))$, where $\tilde{\varphi}(\boldsymbol{\theta})$ is the normalization factor. Table 1 compares properties of ordinary and total Bregman divergences. "Information" relates to the sample divergence to the right population minimizer [4].

Our definition of conformal divergences is inspired by information geometry. In information geometry [24], a *divergence* [22], [25] (also called a contrast function or yoke) is a measure of dissimilarity $D(\boldsymbol{p} : \boldsymbol{q})$ (with $D(\boldsymbol{p} : \boldsymbol{q}) \geq 0$

	ordinary Bregman divergence	total Bregman divergence
Right population minimizer	centroid	this paper (Corollary 1)
Left population minimizer	$\nabla\varphi$-mean, weights $w_i = 1/n$	$\nabla\varphi$-mean, weights $w_i = 1/\sqrt{1+\\|\nabla\varphi(\boldsymbol{x}_i)\\|^2}$
Robustness of pop. minimizers	no	yes
Information	Bregman information	this paper (Lemma 3)
Exhaustiveness	right pop. minimizer = sample average	this paper (Theorem 2)
Bijection	exponential families	lifted exponential families
Inference	MLE = observed point ($\hat{\boldsymbol{\eta}}$=right population minimizer)	Bayesian MAP = total observed point (left population minimizer)

TABLE 1
Properties of ordinary Bregman vs total Bregman divergences.

with equality iff $p = q$, but not necessarily symmetric nor satisfying the triangle inequality) that further needs to satisfy some smoothness conditions [22] to induce properly a metric tensor and a cubic form (for the coefficients of the connections). In Riemannian geometry, a *conformal metric* g' of a metric g is expressed by $g' = \varrho g$, where $\varrho > 0$ (hence also written as $g' = e^{\varrho}g$). The uniformization theorem states that Riemannian surfaces are conformally equivalent to either the spherical, planar or hyperbolic manifolds — all of constant curvatures. In our definition of conformal divergences in eq. (7), factor $g(.)$ plays the role of the conformal factor; we shall see in Lemma 3 below that it indeed defines a conformal factor. In the particular case of total Bregman divergences, $\varrho \propto 1/\sqrt{1+\|\nabla\varphi(.)\|^2}$ plays the role of the conformal factor. In information geometry induced by a generalized logarithm function, a conformal flattening [26] allows to obtain a dually flat structure. Conformal mappings also explain the role of *escort distributions*, and yield efficient algorithms for Voronoi diagrams induced by conformal divergences, a geometric structure particularly relevant to clustering [26], [27].

4 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 [17], [14] for the elicitation of the left population minimizer when $v = \mathrm{Id}$, and the well known results from [8] 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_φ. We first show that the left population minimizer of some v-conformal divergence $D^v_{\varphi,g}$ is a weighted u-mean, where $(u,v)_\varphi$ is a geometric structure.

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

$$\boldsymbol{\mu} \;=\; u^{-1}\left(\frac{1}{\sum_i g(\boldsymbol{x}_i)}\sum_i g(\boldsymbol{x}_i)u(\boldsymbol{x}_i)\right) \;, \quad (8)$$

where $(u,v)_\varphi$ is a geometric structure.
(Proof in Appendix, Subsection 10.1) 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 $D^v_{\varphi,g}$, for some $(u,v)_\varphi$-geometric structure.

Lemma 2: Let $\mu \doteq u^{-1}(\sum_i w_i u(\boldsymbol{x}_i))$ be the unique solution to $\min_{\boldsymbol{x}} \sum_i D(\boldsymbol{x} : \boldsymbol{x}_i)$, where:

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

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

$$D(\boldsymbol{x} : \boldsymbol{y}) \;=\; D^v_{\varphi,g}(\boldsymbol{x} : \boldsymbol{y}) \;. \quad (9)$$

(Proof in Appendix, Subsection 10.2)

5 Right population minimizers of v-conformal divergences

5.1 Case $v = \mathrm{Id}$

We now derive the right population minimizers for a conformal divergence $D_{\varphi,g}$, thus considering v-conformal divergences with $v = \mathrm{Id}$. Because g admits continuous directional derivatives, so does $D_{\varphi,g}(\boldsymbol{x} : \boldsymbol{y})$ for both its arguments. Let us define:

$$\begin{aligned}&\mathrm{D}_{t,\boldsymbol{z}} D_{\varphi,g}(\boldsymbol{x} : \boldsymbol{y}) \\ &\quad \doteq \frac{D_{\varphi,g}(\boldsymbol{x} : \boldsymbol{y} + t\boldsymbol{z}) - D_{\varphi,g}(\boldsymbol{x} : \boldsymbol{y})}{t} \ , \quad (10)\end{aligned}$$

so that the directional derivative in the right parameter $\mathrm{D}_{\boldsymbol{z}} D_{\varphi,g}(\boldsymbol{x} : \boldsymbol{y}) \doteq \lim_{t \rightarrow 0} \mathrm{D}_{t,\boldsymbol{z}} D_{\varphi,g}(\boldsymbol{x} : \boldsymbol{y})$ exists, for any valid direction $\boldsymbol{z}$. Define from any $\mathcal{S}$ the following averages:

$$\overline{\varphi} \doteq \frac{1}{n} \sum_i \varphi(\boldsymbol{x}_i) \ , \quad (11)$$

$$\overline{\boldsymbol{x}} \doteq \frac{1}{n} \sum_i \boldsymbol{x}_i \ . \quad (12)$$

Let us define the following vectors $\overline{\boldsymbol{x}}^+, \boldsymbol{\mu}^+, \boldsymbol{\delta}^+, \boldsymbol{z}^+ \in \mathbb{R}^{d+1}$:

$$\overline{\boldsymbol{x}}^+ \doteq \left[\begin{array}{c} \overline{\boldsymbol{x}} \\ \overline{\varphi} \end{array}\right] \ , \quad (13)$$

$$\boldsymbol{\mu}^+ \doteq \left[\begin{array}{c} \boldsymbol{\mu} \\ \varphi(\boldsymbol{\mu}) \end{array}\right] \ , \quad (14)$$

$$\boldsymbol{\delta}^+ \doteq \overline{\boldsymbol{x}}^+ - \boldsymbol{\mu}^+ \ , \quad (15)$$

$$\boldsymbol{z}^+ \doteq \left[\begin{array}{c} \mathrm{D}_{\boldsymbol{z}}(g(\boldsymbol{\mu})\nabla\varphi(\boldsymbol{\mu})) \\ -\mathrm{D}_{\boldsymbol{z}} g(\boldsymbol{\mu}) \end{array}\right] \ , \quad (16)$$

from which we define the following sets:

$$\mathbb{P}_{\mathcal{S},\varphi,g} \doteq \left\{\boldsymbol{\mu} \in \mathrm{dom}(D_{\varphi,g}) : \boldsymbol{\delta}^+ \perp \boldsymbol{z}^+, \forall \boldsymbol{z}\right\} \ , \quad (17)$$

where the directions $\boldsymbol{z}$ have to be valid, *i.e.* such that the directional derivative of g is defined in a neighborhood of 0 (with respect to t). We also define $\mathbb{B}_{\varphi,g}$, the eventually empty set of non-differentiable boundary points of the intersection of the domains of φ and g. In the following, we let $\mathcal{P}(D_{\varphi,g}; \mathcal{S})$ denote the set of right population minimizers for conformal divergence $D_{\varphi,g}$ on set $\mathcal{S}$.

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

$$\begin{aligned}&\frac{1}{n} \sum_i D_{\varphi,g}(\boldsymbol{x}_i : \boldsymbol{\mu}) \\ &\quad = \varrho_g \times (\overline{\boldsymbol{x}} - \boldsymbol{\mu})^\top \mathrm{H}\varphi(\boldsymbol{\mu})(\overline{\boldsymbol{x}} - \boldsymbol{\mu}) \ , \quad (18)\end{aligned}$$

with $\varrho_g \doteq g^2(\boldsymbol{\mu}) / \mathrm{D}_{(\overline{\boldsymbol{x}} - \boldsymbol{\mu})} g(\boldsymbol{\mu}) > 0$.
(Proof in Appendix, Subsection 10.3) With respect to the discussion on conformal divergences in Section 3, we see that ϱ_g defines a conformal factor, and so all points in $\mathbb{P}_{\mathcal{S},\varphi,g} \backslash \{\overline{\boldsymbol{x}}\}$ are points for which the conformal divergence reduces to a (square) distance on metric $\mathrm{H}\varphi$ conformally transformed. We shall see that set $\mathbb{P}_{\mathcal{S},\varphi,g} \backslash \{\overline{\boldsymbol{x}}\}$ also contains population minimizers. Finally, Lemma 3 shows that the conformal Bregman information generalizes the Bregman information [4] to *weighted* square Mahalanobis distance — because Bregman divergences can be formulated as square Mahalanobis distance over particular metrics, Bregman information can also be expressed using square Mahalanobis distance.

We are now ready to state a first Theorem that provides the right population minimizers for a subset of conformal divergences which encompasses total Bregman divergences.

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

$$\mathcal{P}(D_{\varphi,K g_p}; \mathcal{S}) = \arg\min_{\boldsymbol{\mu}} \|\overline{\boldsymbol{x}}^+ - \boldsymbol{\mu}^+\|_q \ , \quad (19)$$

with $\overline{\boldsymbol{x}}^+$ and $\boldsymbol{\mu}^+$ defined in (13) and (14), and $q = 2k \in \mathbb{N}$ is the Hölder conjugate of p[1].
(Proof in Appendix, Subsection 10.4)

Fixing $k = 1$ allows to retrieve the right population minimizers for total Bregman divergences. Because of their importance, we state their characterization as a separate corollary.

Corollary 1: Consider $g = K g_\perp$ for some constant $K > 0$. The following holds true:

$$\boldsymbol{\delta}^+ \perp \left[\begin{array}{c} \nabla\varphi(\boldsymbol{\mu}) \\ \|\nabla\varphi(\boldsymbol{\mu})\|_2^2 \end{array}\right] \ , \forall \boldsymbol{\mu} \in \mathbb{P}_{\mathcal{S},\varphi,g_\perp} \ ; \quad (20)$$

i.e., the orthogonal projection of $(\overline{\boldsymbol{x}}, \overline{\varphi})$ on the tangent hyperplane $T_\varphi(\boldsymbol{\mu})$ to φ at $\boldsymbol{\mu}$ is point

1. Two reals $p, q \geq 1$ are Hölder conjugates when $(1/p) + (1/q) = 1$.

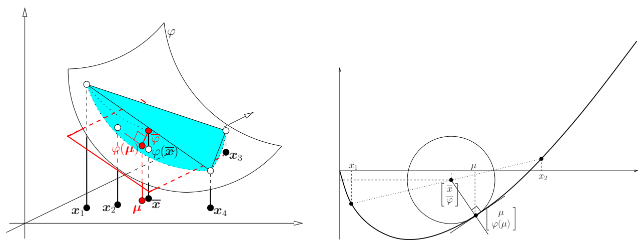

Fig. 2. Left: how to find $\boldsymbol{\mu}$ as stated in Corollary 1: the orthogonal projection of point $(\overline{\boldsymbol{x}}, \overline{\varphi})$ on the hyperplane $T_\varphi(\boldsymbol{\mu})$ tangent to φ at $\boldsymbol{\mu}$ coincides with point $(\boldsymbol{\mu}, \varphi(\boldsymbol{\mu}))$. The blue region depicts the subset of the epigraph of φ which is below the image by φ of the convex envelope of $\mathcal{S}$. Right: computation of the (unique) population minimizer on a simple 1D example with $\varphi(x) = x \ln x - x$, following Corollary 1. Remark that $\mu > \overline{x}$ in this case.

$(\boldsymbol{\mu}, \varphi(\boldsymbol{\mu}))$. Furthermore, assuming $\mathbb{B}_{\varphi,g} = \emptyset$, we have:

$$\mathcal{P}(D_{\varphi,Kg_\perp}; \mathcal{S}) = \arg\min_{\boldsymbol{\mu}} \|\overline{\boldsymbol{x}}^+ - \boldsymbol{\mu}^+\|_2 \ , \quad (21)$$

with $\overline{\boldsymbol{x}}^+$ and $\boldsymbol{\mu}^+$ defined in (13) and (14).
Figure 2 (left) displays how to find $\boldsymbol{\mu}$ which meets condition (20). Notice that, by construction, the right population minimizer for $D_{\varphi,Kg_\perp}$ is invariant by rotation of the axes. Figure 2 (right) depicts the construction of the population minimizer in a simple 1D case.

We now study to what extent total Bregman divergences are exhaustive for the construction of the right population minimizer depicted in Corollary 1. It has been shown that ordinary 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 an ordinary Bregman divergence [7], [8]. 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:

$$\mathbb{P}_{\mathcal{S},\varphi} \doteq \left\{ \mu \in \mathbb{R} : \left[\begin{array}{c} \overline{x} - \mu \\ \overline{\varphi} - \varphi(\mu) \end{array} \right] \perp \left[\begin{array}{c} 1 \\ \varphi'(\mu) \end{array} \right] \right\} \quad (22)$$

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

$$\tilde{\varphi}'_{\mathcal{S}}(\mu)\varphi'(\mu) = -1 \ , \quad (23)$$

with

$$\tilde{\varphi}'_{\mathcal{S}}(z) \doteq \frac{\overline{\varphi} - \varphi(z)}{\overline{x} - z} \ , \forall z \in \mathrm{dom}(\varphi) \backslash \{\overline{x}\} \ . \quad (24)$$

Theorem 2: Let $D : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$ 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 $\mathcal{S} = \{x_1, x_2, ..., x_n\}$ is unique and satisfies $\{\mu\} \in \mathbb{P}_{\mathcal{S},\varphi}$ for some strictly convex twice differentiable φ. Then

$$D(x : y) = D_{\varphi,Kg_\perp}(x : y) \ , \quad (25)$$

where $K > 0$ is a constant and $g_\perp$ is defined in eq. (3).
(proof in [23])

5.2 Case v arbitrary

We now focus on general 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 $\mathbb{B}_{\varphi,g} = \emptyset$. This is not restrictive: even in the multidimensional extension of the total Bregman divergences of Table 2, the cardinal of $\mathbb{B}_{\varphi,g} = \emptyset$ would be at most one, so the main structural and algorithmic issues to characterize the right population minimizers essentially lie in the characterization of $\mathbb{P}_{\mathcal{S},\varphi}$. Define from $\mathcal{S}$ the following averages:

$$\overline{\varphi}_v \doteq \frac{1}{n} \sum_i \varphi(v(\boldsymbol{x}_i)) \quad (26)$$

$$= \frac{1}{n} \sum_i u(\boldsymbol{x}_i)^\top v(\boldsymbol{x}_i) - \overline{\varphi^\star}_u \ ;$$

$$\overline{\boldsymbol{x}}_v \doteq \frac{1}{n} \sum_i v(\boldsymbol{x}_i) \ . \quad (27)$$

We first state a generalization of Theorem 1 to arbitrary $(u, v)_\varphi$-geometric structures.

Theorem 3: Let $(u, v)_\varphi$ be a geometric structure. Pick $g = Kg^u_p(\boldsymbol{\mu})$ with $g^u_p(\boldsymbol{\mu}) \doteq f_p(u(\boldsymbol{\mu}))$, f_p is defined in (6), $K > 0$ is a constant, and $p = 2k/(2k - 1)$ with $k \in \mathbb{N}_*$. Then, the right population minimizer(s) for the v-conformal divergence D^v_{φ,Kg^u_p} on $\mathcal{S}$ match the set:

$$\mathcal{P}(D^v_{\varphi,Kg^u_p};\mathcal{S}) \;=\; \arg\min_{\boldsymbol{\mu}} \|\overline{\boldsymbol{x}}^+_v - \boldsymbol{\mu}^+_v\|_q \ , \quad (28)$$

with $\overline{\boldsymbol{x}}^+_v$ and $\boldsymbol{\mu}^+_v$ defined in (31) and (32), and $q = 2k \in \mathbb{N}$ is the Hölder conjugate of p.

(Proof omitted) We also provide the following generalization of Corollary 1, which stands as a Corollary to Theorem 3.

Corollary 2: Let $(u,v)_\varphi$ be a geometric structure. Pick $g(\boldsymbol{\mu}) = Kg^u_\perp(\boldsymbol{\mu}) \doteq Kf_\perp(u(\boldsymbol{\mu}))$ for any constant $K > 0$. Any right population minimizer $\boldsymbol{\mu}$ for the v-conformal divergence $D^v_{\varphi,Kg^u_\perp}$ satisfies:

$$\left[\begin{array}{c} \overline{\boldsymbol{x}}_v - v(\boldsymbol{\mu}) \\ \overline{\varphi}_v - \varphi(v(\boldsymbol{\mu}))\end{array}\right] \;\perp\; \left[\begin{array}{c} u(\boldsymbol{\mu}) \\ \|u(\boldsymbol{\mu})\|_2^2\end{array}\right] \ . \quad (29)$$

Furthermore, we have:

$$\mathcal{P}(D^v_{\varphi,Kg^u_\perp};\mathcal{S}) \;=\; \arg\min_{\boldsymbol{\mu}} \|\overline{\boldsymbol{x}}^+_v - \boldsymbol{\mu}^+_v\|_2 \ , \quad (30)$$

with $\overline{\boldsymbol{x}}^+_v$ and $\boldsymbol{\mu}^+_v$ defined as follows:

$$\overline{\boldsymbol{x}}^+_v \doteq \left[\begin{array}{c} \overline{\boldsymbol{x}}_v \\ \overline{\varphi}_v \end{array}\right] = \left[\begin{array}{c} \overline{\boldsymbol{x}}_v \\ \frac{1}{n}\sum_i u(\boldsymbol{x}_i)^\top v(\boldsymbol{x}_i) - \overline{\varphi^\star}_u \end{array}\right] \quad (31)$$

$$\boldsymbol{\mu}^+_v \doteq \left[\begin{array}{c} v(\boldsymbol{\mu}) \\ \varphi(v(\boldsymbol{\mu})) \end{array}\right] = \left[\begin{array}{c} v(\boldsymbol{\mu}) \\ u(\boldsymbol{\mu})^\top v(\boldsymbol{\mu}) - \varphi^\star(u(\boldsymbol{\mu})) \end{array}\right] \quad (32)$$

(Proof omitted) Finally, we provide a general characterization of the population minimizers for a general $g = f \circ u$. This is a generalization of the orthogonality property in (29), which is interesting since $\boldsymbol{\delta}^+$ is formulated in the v coordinate mapping while $\boldsymbol{z}^+$ is formulated in the u coordinate mapping.

Theorem 4: Let $(u,v)_\varphi$ be a geometric structure. Suppose $g(\boldsymbol{x}) = f(u(\boldsymbol{x}))$, with f differentiable. For any $\mathcal{S}$ and any $\boldsymbol{\mu}, \boldsymbol{z} \in \mathbb{R}^d$, define $\boldsymbol{\delta}^+_v, \boldsymbol{z}^+_u \in \mathbb{R}^{d+1}$ with:

$$\boldsymbol{\delta}^+_v \doteq \left[\begin{array}{c} \overline{\boldsymbol{x}}_v - v(\boldsymbol{\mu}) \\ \overline{\varphi}_v - \varphi(v(\boldsymbol{\mu}))\end{array}\right] = \underbrace{\left[\begin{array}{c} \overline{\boldsymbol{x}}_v \\ \overline{\varphi}_v \end{array}\right]}_{\boldsymbol{x}^+_v} - \underbrace{\left[\begin{array}{c} v(\boldsymbol{\mu}) \\ \varphi(v(\boldsymbol{\mu})) \end{array}\right]}_{\boldsymbol{\mu}^+_v} \ , \quad (33)$$

$$\boldsymbol{z}^+_u \doteq \left[\begin{array}{c} f(u(\boldsymbol{\mu})) \times \boldsymbol{z} + \nabla f(u(\boldsymbol{\mu}))^\top \boldsymbol{z} \times u(\boldsymbol{\mu}) \\ -\nabla f(u(\boldsymbol{\mu}))^\top \boldsymbol{z} \end{array}\right] \quad (34)$$

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

(Proof omitted)

6 Robustness of the population minimizers

Suppose we add an outlier element $\boldsymbol{x}_*$ with small weight $0 < \epsilon < 1$ to $\mathcal{S}$. The population minimizer (left or right) of $\mathcal{S}$, $\boldsymbol{\mu}$, eventually drifts to a new population minimizer $\boldsymbol{\mu}_* = \boldsymbol{\mu} + \epsilon\boldsymbol{\delta}_{\boldsymbol{\mu}}$ of $\mathcal{S} \cup \{\boldsymbol{x}_*\}$. $\boldsymbol{\delta}_{\boldsymbol{\mu}}$ is called the influence function of $\boldsymbol{x}_*$ [9]. A population minimizer is robust to outliers iff the magnitude of $\boldsymbol{\delta}_{\boldsymbol{\mu}}$ 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 $\boldsymbol{x}_*$ and any weight $0 < \epsilon < 1 - \tau$, $\|\boldsymbol{\delta}_{\boldsymbol{\mu}}\|_2 \leq C$, where C does not depend upon $\boldsymbol{x}_*$ nor ϵ.

Robustness according to Definition 2 is stronger than in the model of [9], [16] as our robustness strictly implies theirs (which relies on very small weights ϵ). So the Lemma to follow is a twofolds generalization of the results of [9], [16], not only from the standpoint of the divergences, but also from the model's.

Lemma 4: Let $(u,v)_\varphi$ be a geometric structure. Suppose the following assumptions are verified: (i) $g(\boldsymbol{x}) = O(1), \forall \boldsymbol{x}$, (ii) $\|u(\boldsymbol{x})\|_2 = O(1/g(\boldsymbol{x})), \forall \boldsymbol{x}$, (iii) the minimal eigenvalue of $J_u^\top J_u$ is $\lambda > 0$, J_u being the Jacobian of u. Then under assumptions (i-iii), the left population minimizer of v-conformal divergence $D^v_{\varphi,g}$ is robust to outliers.

(proof in [23]) Lemma 4 generalizes the robustness of the left population centers of total Bregman divergences (Theorem III.2 in [16]), for which $g = Kg_\perp, v = \mathrm{Id}, u = \nabla\varphi$ (the Jacobian of u being the Hessian of φ, it satisfies assumption (iii) since φ 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 $\boldsymbol{x}_*$ and $\boldsymbol{\mu}$, as long as $\boldsymbol{x}_*$ does not deviate too much from $\boldsymbol{\mu}$ 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 $\boldsymbol{x}_*$ and any weight $0 < \epsilon < 1$:

$$|\varphi(v(\boldsymbol{x}_*)) - \varphi(v(\boldsymbol{\mu}))| \leq L\|v(\boldsymbol{x}_*) - v(\boldsymbol{\mu})\|_2 \Rightarrow \|\boldsymbol{\delta_\mu}\|_2 \leq K\ell(L)\|\boldsymbol{x}_* - \boldsymbol{\mu}\|_2 \ , \quad (35)$$

where $K \geq 0$ is not a function of $\boldsymbol{x}_*$ or ϵ, 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 $\mathbf{0} \in \mathrm{im} u$, which is a mild assumption as it postulates in the $(u, v)_\varphi$-geometric structure that the gradient ∇_φ 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 5: Let $(u, v)_\varphi$ be a geometric structure, and let $f \doteq g \circ u^{-1}$. We make the following assumptions: (i) $\mathbf{0} \in \mathrm{im} u$, (ii) $f(\boldsymbol{z}) \neq 0, \forall \boldsymbol{z}$, (iii) the ratio of the maximal to the minimal eigenvalue of $J_v^\top J_v$, noted λ_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 [23])

7 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.

The nature of the $(u, v)_\varphi$-geometric structure relation — The $(u, v)_\varphi$-geometric structure has been introduced in the context of information geometry to provide a way to compute and analyze the dually flat $(\boldsymbol{\eta}, \boldsymbol{\theta})$ coordinate system arising *e.g.* in exponential families and ordinary Bregman divergences, through a single source parameter which is originally a distribution [20]. To state the key result about the $(u, v)_\varphi$-geometric structure, we 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 $m(x, \xi) = \sum_{i=1}^{d+1} \xi_i \mathbb{1}_{x=x_i}$ a positive distribution computed from $\mathcal{S} = \{x_1, x_2, ..., x_{d+1}\}$, where $\mathbb{1}$ is the indicator variable. $\boldsymbol{\xi} \doteq [\xi_1 \ \xi_2 \ ... \ \xi_{d+1}]^\top$ defines a coordinate system from which we may define two coordinate systems $\boldsymbol{\eta}, \boldsymbol{\theta}$ of $\mathbb{R}^{d+1}$ with $\theta^i \doteq u(\xi_i)$ and $\eta^i \doteq v(\xi_i)$. These coordinate systems have the following interesting information-geometric properties.

Theorem 5: [20] The $(u, v)_\varphi$-geometric structure is dually flat, with the following two potential functions:

$$\begin{aligned} \psi(\boldsymbol{\theta}) &\doteq \sum \int\int \frac{v'(u^{-1}(\theta^i))}{u'(u^{-1}(\theta^i))}(\mathrm{d}\theta^i)^2 \ , \\ \varphi(\boldsymbol{\eta}) &\doteq \sum \int\int \frac{u'(v^{-1}(\eta^i))}{v'(v^{-1}(\eta^i))}(\mathrm{d}\eta^i)^2 \ ; \end{aligned}$$

the divergence between two $\boldsymbol{p}$ and $\boldsymbol{q}$ is given by:

$$D(\boldsymbol{p}; \boldsymbol{q}) \doteq \psi(\boldsymbol{\theta_p}) + \varphi(\boldsymbol{\eta_q}) - \boldsymbol{\theta_p} \cdot \boldsymbol{\eta_q} \ ,$$

and the metric in the $\boldsymbol{\theta}$ coordinate system is:

$$g_{ij}(\boldsymbol{\theta}) \doteq \frac{v'(\xi_i)}{u'(\xi_i)}\delta_{ij} \ .$$

One may check that $u \circ v^{-1}$ defines $\nabla\varphi$, and that D is an ordinary Bregman divergence. One important example is Amari's (α, β) structure ($\alpha, \beta > 0$) for which $u(\xi_i) \doteq \xi_i^\alpha$ and $v(\xi_i) \doteq \xi_i^\beta$, which helps to see the usefulness of the $(u, v)_\varphi$-geometric structure in the context of clustering: assuming $\boldsymbol{\xi}$ is a source parameter recorded in data, one can jointly tune α, β to tune the coordinate system of the divergence without changing its generator φ as long as α/β remains a fixed constant (because $\varphi(\boldsymbol{\eta}) = (1 + \alpha/\beta)^{-1} \sum_i \eta_i^{1+\alpha/\beta}$, omitting the additive constant which does not change the divergence). Thus, we get new free parameters to tune that adapt the coordinate system from which the divergence is computed, which we may use to get improved clustering results.

We now show that, if we accept to change the generator, then we may have a significant freedom in picking and changing the coordinate mappings u and v. We study the nature of the $(u, v)_\varphi$-geometric structure, and define a tolerance relation [28] as a binary relation which is reflexive and symmetric but not necessarily transitive. An equivalence relation is reflexive, symmetric and transitive. We consider the "geometric structure" binary relation, (u, v) (without reference to φ), which holds when there exists some φ such that $(u, v)_\varphi$ is a geometric structure.

Lemma 6: The "geometric structure" relation is a tolerance relation. It is an equivalence relation in the subset of functions $\mathcal{S}_\phi$ indexed by some strictly convex differentiable ϕ and defined by: $\mathcal{S}_\phi \doteq \{\varphi : \mathrm{H}\varphi(\nabla\phi) = \mathrm{H}\phi P_\phi D_\phi^{-1} D P_\phi^\top\}$, where P_ϕ, D_ϕ are the eigenspace and eigenvalues matrix of $\mathrm{H}\phi$ and $D \succ 0$ is diagonal.

(Proof in Appendix, Subsection 10.5) 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 $\mathcal{S}_\phi$ may be interesting from the clustering standpoint to learn the $(u, v)_\varphi$-geometric structure (see Section 5).

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

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

(proof in [23])

Only one symmetric conformal divergence — We show that there exists a single 1D symmetric conformal divergence in the $(u, v)_\varphi$-geometric structure, the square loss, $D_{\varphi,g}(v(x) : v(y)) \propto (v(x) - v(y))^2$. 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 $g(u(\boldsymbol{y}))D_\varphi(v(\boldsymbol{x}) : v(\boldsymbol{y})) \doteq \sum_i g(u(y^i))D_\varphi(v(x^i) : v(y^i))$.

Lemma 7: Let $(u, v)_\varphi$ be a geometric structure and $D_{\varphi,g}$ a conformal divergence for some strictly convex twice differentiable φ. Suppose that $\forall x, y$:

$$\begin{aligned} & g(u(y))D_\varphi(v(x) : v(y)) \\ & \quad = \; g(u(x))D_\varphi(v(y) : v(x)) \; . \end{aligned} \tag{36}$$

Then **(i)** $g(.) = K_1$, **(ii)** $v = \ell(u)$, **(iii)** $\varphi(x) = K_2 x^2 + \ell(x)$ for some constants $K_1 > 0, K_2 > 0$, where $\ell(.)$ is a linear function in its argument. (proof in [23]) Thus, conformal divergence are not metrics, yet they can be used to craft metrics. To ensure that symmetry and triangle inequality are met without violating nonnegativity nor the identity of indiscernibles, we can search for the $\alpha \in (0, 1]$ with which $(g(u(\boldsymbol{y}))D_\varphi(v(\boldsymbol{x}) : v(\boldsymbol{y})) + g(u(\boldsymbol{x}))D_\varphi(v(\boldsymbol{y}) : v(\boldsymbol{x})))^\alpha$ meets the triangle inequality, or use [29]'s method.

Fast approximation of right population minimizers — We now show that under mild assumptions on φ, candidates for right population minimizer may be easily located and approximated in the 1D setting. Assume wlog that $\mathcal{S}$ is ordered, that is $x_1 \leq x_2 \leq ... \leq x_n$. Whenever φ is bijective over $[x_1, x_n]$, we define the φ-mean:

$$\overline{x}_\varphi \;\doteq\; \varphi^{-1}(\overline{\varphi}) \; .$$

Let us denote a *candidate* right population minimizer as a real which is solution of (23). Candidate population minimizers are critical points for the right parameter of the average divergence [30].

Lemma 8: Suppose $g(y) = K g_\perp(y)$, and assume that φ' has constant sign on $[x_1, x_n]$. Then there exists a candidate right population minimizer μ in $[x_1, x_n]$. Furthermore, $\mu \in [\overline{x}_\varphi, \overline{x}]$ if $\mathrm{sign} = -$, and $\mu \in [\overline{x}, \overline{x}_\varphi]$ if $\mathrm{sign} = +$. Here, "sign" denotes the sign of φ' over $[x_1, x_n]$.

(Proof in Appendix, Subsection 10.6) Table 2 presents some applications of Lemma 8 (the domain considered for $\varphi(x) = 1/x$ is $\mathbb{R}_{+*}$). Approximating the candidate right population minimizer μ in the interval may be done by fitting the roots of equations of the form $f(\mu, \overline{x}, \overline{x}_\varphi) = 0$, some of which are given below as examples:

$$\log(\overline{x}_\varphi) - \log(\mu) + \mu^2 - \overline{x}\mu \;=\; 0$$

$\varphi(x)$	Name or expression for $D_{\varphi,g_\perp}(x:y)$	$\overline{x}_\varphi$	Name of φ-mean	Location
$-\log(x)$	Total Itakura-Saito	$\prod_i x_i^{\frac{1}{n}}$	Geometric mean	$[\overline{x}_\varphi, \overline{x}]$
$1/x$	$\frac{1}{x} - \frac{2}{y} + \frac{x}{y^2}$	$n/\sum_i x_i^{-1}$	Harmonic mean	$[\overline{x}_\varphi, \overline{x}]$
x^2	Total square loss	$\pm\sqrt{\frac{1}{n}\sum_i x_i^2}$	$\pm$ Root mean square	$[\overline{x}, \overline{x}_\varphi]$ or $[\overline{x}_\varphi, \overline{x}]$
$x^p, p \geq 2$	Total power loss	$\left(\sum_i x_i^p\right)^{\frac{1}{p}}$	Power mean	$[\overline{x}, \overline{x}_\varphi]$
$\exp(x)$	Total exp divergence	$\log\sum_i \exp x_i$	None	$[\overline{x}_\varphi, \overline{x}]$
$x\log(x)$	Total KL	$\frac{\frac{1}{n}\sum_i x_i \log x_i}{W(\frac{1}{n}\sum_i x_i \log x_i)}$	None	$[\overline{x}, \overline{x}_\varphi]$
$W^{-1}(x)$	$x(\exp(x) - \exp(y))$ $+y(y-x)\exp(y)$	$W(\frac{1}{n}\sum_i W^{-1}(x_i))$	None	$[\overline{x}, \overline{x}_\varphi]$

TABLE 2

Examples of total Bregman divergences $D_{\varphi,g_\perp}(x:y)$, and Location of the candidate population minimizer according to Lemma 8. W is Lambert W function and $W^{-1}(x)$ is shorthand for $x\exp(x)$.

for total Itakura Saito divergence,

$$2\mu^3 - (2\overline{x}_\varphi^2 - 1)\mu - \overline{x} \quad = \quad 0$$

for total square loss divergence (notice that a closed-form expression for μ is available),

$$p\mu^{2p-1} - p\overline{x}_\varphi^p \mu^{p-1} + \mu - \overline{x} \quad = \quad 0$$

for total power loss divergence,

$$\exp(2\mu) - \exp(\overline{x}_\varphi + \mu) + \mu - \overline{x} \quad = \quad 0$$

for total exp divergence, and finally

$$\left(\mu\log\mu - \frac{\overline{x}_\varphi}{W(\overline{x}_\varphi)}\log\frac{\overline{x}_\varphi}{W(\overline{x}_\varphi)}\right)(1+\log\mu) + \mu - \overline{x} \quad = \quad 0$$

for total KL divergence.

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 φ, as displayed in Figure 3.

We also notice that the compactness of $\mathrm{dom}(D_{\varphi,g})$ 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 $\mathrm{dom}(D_{\varphi,g})$.

Extension to scaled Bregman divergences — A new generalization of ordinary Bregman divergences has been recently coined [19], called scaled Bregman divergences. A scaled

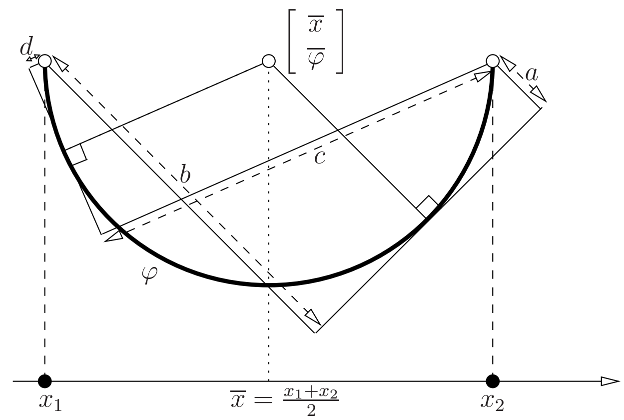

Fig. 3. φ is half-circle and $\mathcal{S} = \{x_1, x_2\}$. In this case, all points in $[x_1, x_2]$ are right population minimizers for $D_{\varphi,g_\perp}$ (the total Bregman divergence equals $a + b = c + d$).

Bregman divergence is a particular case, for $g = v = \mathrm{Id}$, of what we call a scaled conformal divergence, defined as:

$$D^v_{\varphi,g}(x:y;w) \quad \doteq \quad wD^v_{\varphi,g}(x/w:y/w) \ , \quad (37)$$

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 exponential families for scaled Bregman power divergences, which means they have a significant potential for applications in clustering [19]. 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 $\mathcal{S} = \{x_1, x_2, ..., x_n\}$ and $\mathcal{W} = \{w_1, w_2, ..., w_n\}$, such that a left population minimizer of the ordered pair $(\mathcal{S}, \mathcal{W})$ for $D^v_{\varphi,g}$ is defined as μ that minimizes $\sum_i D^v_{\varphi,g}(\mu : x_i; w_i)$. The following Lemma shows that, despite the left population minimizer is not always available in closed form in general (unlike v-conformal divergences), it is in between the minimal and maximal values of $\mathcal{S}$ (like v-conformal divergences).

Lemma 9: The left population minimizer of $D^v_{\varphi,g}$ over $(\mathcal{S}, \mathcal{W})$ is unique and in $[\min_i x_i, \max_i x_i]$.

(Proof in Appendix, Subsection 10.7) 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 $\prod_j [\min_i x^j_i, \max_i x^j_i]$.

8 Conclusion

We have studied the left and right population minimizers of conformal divergences, a superset of ordinary Bregman divergences and total Bregman divergences, in the $(u, v)_\varphi$-geometric structure [20], [21], [22], 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 $(u, v)_\varphi$ geometric structure relation, and the simple approximation of 1D population minimizers, shall be useful to widen the scope of existing clustering algorithms and/or develop algorithmically new clustering algorithms relying on broad classes of distortions that escape the conventional framework of ordinary Bregman divergences, as *e.g.* recently initiated with total Bregman divergences or scaled Bregman divergences.

9 Acknowledgments

The authors thank Meizhu Liu, and the reviewers for insightful comments that helped to significantly improve the paper. Work initiated while R. Nock was visiting Sony CS Labs Tokyo. NICTA is funded by the Australian Government through the Department of Communications and the Australian Research Council through the ICT Centre of Excellence Program.

References

[1] U. von Luxburg, R.-C. Williamson, and I. Guyon, "Clustering: science or art?" in *Workshop on Unsupervised and Transfer Learning*, 2012, pp. 65–79.
[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 5^{th} Berkeley symposium on mathematical statistics and probability*, 1967, pp. 281–297.
[4] A. Banerjee, S. Merugu, I. Dhillon, and J. Ghosh, "Clustering with bregman divergences," in *Proc. of the 4^{th} SIAM International Conference on Data Mining*, 2004, pp. 234–245.
[5] A.-N. Kolmogorov, "Sur la notion de moyenne," *Atti della R. Accademia Nazionale dei Lincei*, vol. 12, pp. 388–391, 1930.
[6] M. Nagumo, "Über eine klasse von mittelwerte," *Japanese Journal of Mathematics*, vol. 7, pp. 71–79, 1930.
[7] J. Abernethy and R.-M. Frongillo, "A characterization of proper scoring rules for linear properties," in *Proc. of the 25^{th} COLT*, 2012, pp. 1–14.
[8] 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.
[9] M. Liu, "Total Bregman divergence, a robust divergence measure, and its applications," Ph.D. dissertation, University of Florida, 2011.
[10] 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.
[11] F. Escolano, M. Liu, and E.-R. Hancock, "Tensor-based total bregman divergences between graphs," in *IEEE International Conference on Computer Vision Workshops*, 2011, pp. 1440–1447.
[12] 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 20^{th} ACM International Conference on Information and Knowledge Management*, 2011, pp. 2509–2512.
[13] M. Liu and B.-C. Vemuri, "Robust and efficient regularized boosting using total bregman divergence," in *Proc. of the 24^{th} IEEE CVPR*, 2011, pp. 2897–2902.
[14] M. Liu, B.-C. Vemuri, S.-I. Amari, and F. Nielsen, "total Bregman divergence and its applications to shape retrieval," in *Proc. of the 23^{rd} IEEE CVPR*, 2010, pp. 3463–3468.
[15] 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.

[16] 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.

[17] 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.

[18] A. R. M. y Terán, M. Gouiffès, and L. Lacassagne, "Total bregman divergence for multiple object tracking," in *Proc. of the 20 th IEEE International Conference on Image Processing*, 2013.

[19] W. Stummer and I. Vajda, "On Bregman distances and divergences of probability measures," *IEEE Trans. on Information Theory*, vol. 58, pp. 1277–1288, 2012.

[20] S.-I. Amari, "New developments of information geometry (17): Tsallis *q*-entropy, escort geometry, conformal geometry," in *Mathematical Sciences (suurikagaku)*. Science Company, October 2012, no. 592, pp. 73–82, in japanese.

[21] ——, "New developments of information geometry (26): Information geometry of convex programming and game theory," in *Mathematical Sciences (suurikagaku)*. Science Company, November 2013, no. 605, pp. 65–74, in japanese.

[22] J. Zhang, "Divergence function, duality, and convex analysis," *Neural Computation*, vol. 16, pp. 159–195, 2004.

[23] R. Nock, F. Nielsen, and S.-I. Amari, "On conformal divergences and their population minimizers," Computing Res. Repository, Tech. Rep. CoRR abs/1311.5125, 2013.

[24] S.-I. Amari and H. Nagaoka, *Methods of Information Geometry*. Oxford University Press, 2000.

[25] M. Basseville, "Divergence measures for statistical data processing," *Signal Processing*, vol. 93, pp. 621–633, 2013.

[26] A. Ohara, H. Matsuzoe, and S.-I. Amari, "A dually-flat structure on the space of escort distributions," *Journal of Physics : conference series*, vol. 201, no. 012012, 2010.

[27] J.-D. Boissonnat, F. Nielsen, and R. Nock, "Bregman Voronoi Diagrams," *DCG*, vol. 44, no. 2, pp. 281–307, 2010.

[28] A.-B. Sossinsky, "Tolerance space theory and some applications," *Acta Applicandae Mathematicae*, vol. 5, pp. 137–167, 1986.

[29] S. Acharyya, A. Banerjee, and D. Boley, "Bregman divergences and triangle inequality," in *Proc. of the 13^{th} SIAM International Conference on Data Mining*, 2013, pp. 476–484.

[30] F.-H. Clarke, *Optimization and Nonsmooth Analysis*. Wiley, 1989.

10 APPENDIX

10.1 Proof of Lemma 1

Any left population minimizer $\boldsymbol{\mu}$ satisfies $\nabla \sum_i D^v_{\varphi,g}(\boldsymbol{\mu} : \boldsymbol{x}_i) = \mathbf{0}$, and so, after simplification, we obtain:

$$J_v(\boldsymbol{\mu}) \sum_i g(\boldsymbol{x}_i) \left(\nabla\varphi(v(\boldsymbol{\mu})) - \nabla\varphi(v(\boldsymbol{x}_i))\right) = \mathbf{0} , \quad (38)$$

where J_v is the Jacobian of v. Since v is bijective, the null space of J_v is reduced to $\{\mathbf{0}\}$, and so we must have:

$$\sum_i g(\boldsymbol{x}_i) \left(\nabla\varphi(v(\boldsymbol{\mu})) - \nabla\varphi(v(\boldsymbol{x}_i))\right) = \mathbf{0} ,$$

which, after solving for $\boldsymbol{\mu}$, yields:

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

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

10.2 Proof of Lemma 2

Let $\boldsymbol{x}_u \doteq u(\boldsymbol{x}), \forall \boldsymbol{x}$ and let $\mathcal{S}_u \doteq \{(\boldsymbol{x}_i)_u, i = 1, 2, ..., n\}$. We have $\boldsymbol{\mu}_u = \sum_i w_i (\boldsymbol{x}_i)_u$ and

$$\begin{aligned} D(\boldsymbol{\mu} : \boldsymbol{x}_i) &= D(u^{-1}(\boldsymbol{\mu}_u) : u^{-1}((\boldsymbol{x}_i)_u)) \\ &\doteq D_2(\boldsymbol{\mu}_u : (\boldsymbol{x}_i)_u) . \end{aligned} \quad (39)$$

From (39), the assumptions on D and the properties of u, it follows that D_2 is non-negative, differentiable, satisfies $D_2(\boldsymbol{x} : \boldsymbol{x}) = 0, \forall \boldsymbol{x}$, and its left population minimizer over $\mathcal{S}_u$ is the weighted arithmetic average $\boldsymbol{\mu}_u$: it is thus an ordinary Bregman divergence [7] with:

$$D_2(\boldsymbol{\mu}_u : (\boldsymbol{x}_i)_u) = w_i D_\phi((\boldsymbol{x}_i)_u : \boldsymbol{\mu}_u) , \quad (40)$$

for some strictly convex differentiable ϕ. Calling to convex conjugates, we obtain:

$$\begin{aligned} & D_\phi((\boldsymbol{x}_i)_u : \boldsymbol{\mu}_u) \\ &= D_{\phi^\star}(\nabla\phi(\boldsymbol{\mu}_u) : \nabla\phi((\boldsymbol{x}_i)_u)) \\ &= D_{\phi^\star}((\nabla\phi \circ u)(\boldsymbol{\mu}) : (\nabla\phi \circ u)(\boldsymbol{x}_i)) \\ &\doteq D_\varphi(v(\boldsymbol{\mu}) : v(\boldsymbol{x}_i)) , \end{aligned} \quad (41)$$

with $\varphi \doteq \phi^\star$ and $(v, u)_\phi$-geometric structure. $(u, v)_\varphi$ is thus a geometric structure and merging (39 — 41), we obtain:

$$\begin{aligned} D(\boldsymbol{\mu} : \boldsymbol{x}_i) &= w_i D_\varphi(v(\boldsymbol{\mu}) : v(\boldsymbol{x}_i)) . \\ &= g(\boldsymbol{x}_i) D_\varphi(v(\boldsymbol{\mu}) : v(\boldsymbol{x}_i)) , \end{aligned} \quad (42)$$

for some g admitting continuous directional derivatives which meets $g(\boldsymbol{x}_i) = w_i, \forall i = 1, 2, ..., n$ (we can pick *e.g.* a degree-n polynomial). We obtain (9), as claimed. This ends the proof of Lemma 2.

10.3 Proof of Lemma 3

The first part of the proof is standard, and shows that $\mathbb{P}_{\mathcal{S},\varphi,g}$ is the set of critical points for the right parameter [30]. Assume $\boldsymbol{\mu}$ is a population minimizer, and define $\mathrm{D}_{t,\boldsymbol{z}}D_{\varphi,g}(\mathcal{S}:\boldsymbol{y}) \doteq \sum_i \mathrm{D}_{t,\boldsymbol{z}}D_{\varphi,g}(\boldsymbol{x}_i:\boldsymbol{y})$. Fix any valid direction $\boldsymbol{z}$. Because $\boldsymbol{\mu}$ is a right population minimizer, it comes $\mathrm{D}_{t,\boldsymbol{z}}D_{\varphi,g}(\mathcal{S}:\boldsymbol{\mu}) \leq 0$ for $t \leq 0$, and $\mathrm{D}_{t,\boldsymbol{z}}D_{\varphi,g}(\mathcal{S}:\boldsymbol{\mu}) \geq 0$ for $t \geq 0$. Since directional derivatives are defined in direction $\boldsymbol{z}$, we obtain $0 \leq \lim_{t\downarrow 0} \mathrm{D}_{t,\boldsymbol{z}}D_{\varphi,g}(\mathcal{S}:\boldsymbol{\mu}) = \mathrm{D}_{\boldsymbol{z}}D_{\varphi,g}(\mathcal{S}:\boldsymbol{\mu}) = \lim_{t\uparrow 0} \mathrm{D}_{t,\boldsymbol{z}}D_{\varphi,g}(\mathcal{S}:\boldsymbol{\mu}) \leq 0$ and $\boldsymbol{\mu}$ is a solution of:

$$\lim_{t\to 0} \mathrm{D}_{t,\boldsymbol{z}}D_{\varphi,g}(\mathcal{S}:\boldsymbol{\mu}) = \mathrm{D}_{\boldsymbol{z}}\sum_i D_{\varphi,g}(\boldsymbol{x}_i:\boldsymbol{\mu}) = 0 \; . \quad (43)$$

We plug in (43) the expression of $D_{\varphi,g}$ and obtain that for any right population minimizer $\boldsymbol{\mu}$, the following holds:

$$\begin{aligned}
&\frac{1}{n}\mathrm{D}_{\boldsymbol{z}}\sum_i D_{\varphi,g}(\boldsymbol{x}_i:\boldsymbol{\mu}) \\
&= \mathrm{D}_{\boldsymbol{z}}g(\boldsymbol{\mu}) \times \frac{1}{n}\sum_i D_{\varphi}(\boldsymbol{x}_i:\boldsymbol{\mu}) \\
&\quad + g(\boldsymbol{\mu})(\boldsymbol{\mu}-\overline{\boldsymbol{x}})^\top \mathrm{H}\varphi(\boldsymbol{\mu})\boldsymbol{z} \\
&= 0 \; . \qquad (44)
\end{aligned}$$

Rewriting, we thus need:

$$\begin{aligned}
&\mathrm{D}_{\boldsymbol{z}}g(\boldsymbol{\mu}) \times (\overline{\varphi} - \varphi(\boldsymbol{\mu})) \\
&= (\overline{\boldsymbol{x}}-\boldsymbol{\mu})^\top (g(\boldsymbol{\mu})\mathrm{H}\varphi(\boldsymbol{\mu})\boldsymbol{z} \\
&\quad + \mathrm{D}_{\boldsymbol{z}}g(\boldsymbol{\mu})\nabla\varphi(\boldsymbol{\mu})) \qquad (45)\\
&= (\overline{\boldsymbol{x}}-\boldsymbol{\mu})^\top \mathrm{D}_{\boldsymbol{z}}(g(\boldsymbol{\mu})\nabla\varphi(\boldsymbol{\mu})) \; ,
\end{aligned}$$

which implies $\boldsymbol{\mu} \in \mathbb{P}_{\mathcal{S},\varphi,g}$. If a population minimizer does not belong to $\mathbb{P}_{\mathcal{S},\varphi,g}$, it is in the non differentiable part of the boundary, that is, in $\mathbb{B}_{\varphi,g}$. Eq. (45) brings:

$$\begin{aligned}
&\mathrm{D}_{\boldsymbol{z}}g(\boldsymbol{\mu}) \times \frac{1}{n}\sum_i D_{\varphi}(\boldsymbol{x}_i:\boldsymbol{\mu}) \\
&= g(\boldsymbol{\mu}) \times (\overline{\boldsymbol{x}}-\boldsymbol{\mu})^\top \mathrm{H}\varphi(\boldsymbol{\mu})\boldsymbol{z} \; ,
\end{aligned}$$

and so:

$$\begin{aligned}
&\frac{1}{n}\sum_i D_{\varphi,g}(\boldsymbol{x}_i:\boldsymbol{\mu}) \\
&= \frac{(\overline{\boldsymbol{x}}-\boldsymbol{\mu})^\top \mathrm{H}\varphi(\boldsymbol{\mu})\boldsymbol{z}}{\mathrm{D}_{\boldsymbol{z}}g(\boldsymbol{\mu})} g^2(\boldsymbol{\mu}) \; , \qquad (46)
\end{aligned}$$

a quantity which does not depend on the direction $\boldsymbol{z} \neq \boldsymbol{0}$. Fixing as direction $\boldsymbol{z} = \overline{\boldsymbol{x}} - \boldsymbol{\mu}$ yields the statement of (18).

10.4 Proof of Theorem 1

We first need the following Lemma.

Lemma 10: Suppose $g(\boldsymbol{x}) = f(\nabla\varphi(\boldsymbol{x}))$, with φ strictly convex twice differentiable and f differentiable. $\forall \boldsymbol{\mu} \in \mathbb{P}_{\mathcal{S},\varphi,g}$, we have:

$$\begin{aligned}
&\overline{\boldsymbol{x}} - \boldsymbol{\mu} \\
&= \frac{1}{f(\nabla\varphi(\boldsymbol{\mu}))}\left(\frac{1}{n}\sum_i D_{\varphi}(\boldsymbol{x}_i:\boldsymbol{\mu})\right) \\
&\quad \cdot \nabla f(\nabla\varphi(\boldsymbol{\mu})) \; . \qquad (47)
\end{aligned}$$

Proof The chain rule gives

$$\begin{aligned}
\mathrm{D}_{\boldsymbol{z}}g(\boldsymbol{\mu}) &= \mathrm{D}_{\boldsymbol{z}}(f(\nabla\varphi(\boldsymbol{\mu}))) \\
&= \mathrm{D}_{\mathrm{D}_{\boldsymbol{z}}\nabla\varphi(\boldsymbol{\mu})}f(\nabla\varphi) \\
&= \mathrm{D}_{\mathrm{H}\varphi(\boldsymbol{\mu})\boldsymbol{z}}f(\nabla\varphi) \\
&= \boldsymbol{z}^\top(\mathrm{H}\varphi(\boldsymbol{\mu}))^\top \nabla f(\nabla\varphi(\boldsymbol{\mu})) \\
&= \boldsymbol{z}^\top \mathrm{H}\varphi(\boldsymbol{\mu}) \nabla f(\nabla\varphi(\boldsymbol{\mu})) \; ,
\end{aligned}$$

so that (45) becomes:

$$\begin{aligned}
&\boldsymbol{z}^\top \mathrm{H}\varphi(\boldsymbol{\mu})\left((\overline{\varphi}-\varphi(\boldsymbol{\mu})) \times \nabla f(\nabla\varphi(\boldsymbol{\mu}))\right) \\
&= \boldsymbol{z}^\top \mathrm{H}\varphi(\boldsymbol{\mu})\left(g(\boldsymbol{\mu}) \times (\overline{\boldsymbol{x}}-\boldsymbol{\mu})\right) \\
&\quad + \boldsymbol{z}^\top \mathrm{H}\varphi(\boldsymbol{\mu})\left(\nabla\varphi(\boldsymbol{\mu})^\top(\overline{\boldsymbol{x}}-\boldsymbol{\mu}) \times \nabla f(\nabla\varphi(\boldsymbol{\mu}))\right) \\
&= \boldsymbol{z}^\top \mathrm{H}\varphi(\boldsymbol{\mu})\left(\begin{array}{c} g(\boldsymbol{\mu}) \times (\overline{\boldsymbol{x}}-\boldsymbol{\mu}) \\ + \\ \nabla\varphi(\boldsymbol{\mu})^\top(\overline{\boldsymbol{x}}-\boldsymbol{\mu}) \times \nabla f(\nabla\varphi(\boldsymbol{\mu})) \end{array}\right) \qquad (48)
\end{aligned}$$

Eq. (48) is of the form $\boldsymbol{z}^\top \mathrm{H}\varphi(\boldsymbol{\mu})\boldsymbol{a} = \boldsymbol{z}^\top \mathrm{H}\varphi(\boldsymbol{\mu})\boldsymbol{b}$ which implies $\boldsymbol{a} = \boldsymbol{b}$ as otherwise picking $\boldsymbol{z} = \boldsymbol{b} - \boldsymbol{a} \neq \boldsymbol{0}$ would contradict the positive definiteness of the Hessian $\mathrm{H}\varphi$. After reordering, we get eq. (47). This ends the proof of Lemma 10. ☐

(Continued proof of Theorem 1) Let us fix $f(\boldsymbol{x}) = f_p(\boldsymbol{x}) \doteq 1/(1+\|\boldsymbol{x}\|_p^p)^{1/p}$. We have:

$$\nabla f(\nabla\varphi(\boldsymbol{\mu})) = -\frac{1}{(1+\|\nabla\varphi(\boldsymbol{\mu})\|_p^p)^{1+\frac{1}{p}}}\nabla_p\varphi(\boldsymbol{\mu}) , \qquad (49)$$

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

$$\nabla\varphi(\boldsymbol{\mu})^\top \nabla_p\varphi(\boldsymbol{\mu}) = \|\nabla\varphi(\boldsymbol{\mu})\|_p^p \; . \qquad (50)$$

We now use (47) with $f = f_p$ and obtain:

$$\overline{\boldsymbol{x}} - \boldsymbol{\mu} = -\frac{1}{1+\|\nabla\varphi(\boldsymbol{\mu})\|_p^p}\left(\frac{1}{n}\sum_i D_\varphi(\boldsymbol{x}_i : \boldsymbol{\mu})\right) \times \nabla_p\varphi(\boldsymbol{\mu}) \ . \quad (51)$$

Coordinate j in $\overline{\boldsymbol{x}} - \boldsymbol{\mu}$, $(\overline{\boldsymbol{x}} - \boldsymbol{\mu})^j$, satisfies:

$$\begin{aligned}
&((\overline{\boldsymbol{x}} - \boldsymbol{\mu})^j)^{q-1} \\
&= \left(\frac{-m}{1+\|\nabla\varphi(\boldsymbol{\mu})\|_p^p}\right)^{q-1} \times (\mathrm{sign}(\nabla^j)|\nabla^j|^{p-1})^{q-1} \\
&= -\left(\frac{m}{1+\|\nabla\varphi(\boldsymbol{\mu})\|_p^p}\right)^{q-1} \times \nabla^j \ , \quad (52)
\end{aligned}$$

where $m \doteq (1/m)\sum_i D_\varphi(\boldsymbol{x}_i : \boldsymbol{\mu}) \geq 0$. Eq. (52) holds because $(p-1)(q-1) = 1$ and $q = 2k \in \mathbb{N}$ is even. So, we may write:

$$\begin{aligned}
&\|\overline{\boldsymbol{x}} - \boldsymbol{\mu}\|_q^q \\
&\doteq \sum_j ((\overline{\boldsymbol{x}} - \boldsymbol{\mu})^j)^q \\
&= \sum_j ((\overline{\boldsymbol{x}} - \boldsymbol{\mu})^j)^{q-1} \times (\overline{\boldsymbol{x}} - \boldsymbol{\mu})^j \\
&= -\left(\frac{m}{1+\|\nabla\varphi(\boldsymbol{\mu})\|_p^p}\right)^{q-1} \times (\overline{\boldsymbol{x}} - \boldsymbol{\mu})^\top \nabla\varphi(\boldsymbol{\mu}) \quad (53)
\end{aligned}$$

We make the inner product of (51) with $\nabla\varphi(\boldsymbol{\mu})$ and obtain because of (50):

$$\begin{aligned}
&(\overline{\boldsymbol{x}} - \boldsymbol{\mu})^\top \nabla\varphi(\boldsymbol{\mu}) \\
&= -\frac{\|\nabla\varphi(\boldsymbol{\mu})\|_p^p}{1+\|\nabla\varphi(\boldsymbol{\mu})\|_p^p}\left(\frac{1}{n}\sum_i D_\varphi(\boldsymbol{x}_i : \boldsymbol{\mu})\right) \ , \\
&= -\alpha(\overline{\varphi} - \varphi(\boldsymbol{\mu})) + \alpha(\overline{\boldsymbol{x}} - \boldsymbol{\mu})^\top \nabla\varphi(\boldsymbol{\mu}) \ , \quad (54)
\end{aligned}$$

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

$$\begin{aligned}
&\overline{\varphi} - \varphi(\boldsymbol{\mu}) \\
&= \frac{1}{1+\|\nabla\varphi(\boldsymbol{\mu})\|_p^p}\left(\frac{1}{n}\sum_i D_\varphi(\boldsymbol{x}_i : \boldsymbol{\mu})\right) \quad (55)
\end{aligned}$$

We finally get from (53) and (55), using the shorthand $m \doteq (1/n)\sum_i D_\varphi(\boldsymbol{x}_i : \boldsymbol{\mu})$:

$$\begin{aligned}
&\|\overline{\boldsymbol{x}}^+ - \boldsymbol{\mu}^+\|_q \\
&= \left(|\overline{\varphi} - \varphi(\boldsymbol{\mu})|^q + \|\overline{\boldsymbol{x}} - \boldsymbol{\mu}\|_q^q\right)^{\frac{1}{q}} \\
&= \left((\overline{\varphi} - \varphi(\boldsymbol{\mu}))^{q-1} \times (\overline{\varphi} - \varphi(\boldsymbol{\mu})) + \|\overline{\boldsymbol{x}} - \boldsymbol{\mu}\|_q^q\right)^{\frac{1}{q}} \\
&= \left(\begin{array}{c} \left(\frac{m}{1+\|\nabla\varphi(\boldsymbol{\mu})\|_p^p}\right)^{q-1}(\overline{\varphi} - \varphi(\boldsymbol{\mu})) \\ - \\ \left(\frac{m}{1+\|\nabla\varphi(\boldsymbol{\mu})\|_p^p}\right)^{q-1}(\overline{\boldsymbol{x}} - \boldsymbol{\mu})^\top \nabla\varphi(\boldsymbol{\mu}) \end{array}\right)^{\frac{1}{q}} \\
&= \frac{m^{\frac{1}{p}}}{(1+\|\nabla\varphi(\boldsymbol{\mu})\|_p^p)^{\frac{1}{p}}} \times \left(\overline{\varphi} - \varphi(\boldsymbol{\mu}) - (\overline{\boldsymbol{x}} - \boldsymbol{\mu})^\top \nabla\varphi(\boldsymbol{\mu})\right)^{\frac{1}{q}} \\
&= \frac{m^{\frac{1}{p}+\frac{1}{q}}}{(1+\|\nabla\varphi(\boldsymbol{\mu})\|_p^p)^{\frac{1}{p}}} \\
&= \frac{m}{(1+\|\nabla\varphi(\boldsymbol{\mu})\|_p^p)^{\frac{1}{p}}} \\
&= \frac{1}{n}\sum_i D_{\varphi, g_p}(\boldsymbol{x}_i : \boldsymbol{\mu}) \\
&= \frac{1}{K} \times \left(\frac{1}{n}\sum_i D_{\varphi, K g_p}(\boldsymbol{x}_i : \boldsymbol{\mu})\right) \ ,
\end{aligned}$$

which yields the statement of Theorem 1.

10.5 Proof of Lemma 6

Clearly, (u, u) holds since $(u, u)_\varphi$ is a geometric structure for $\varphi \doteq (1/2)\sum_i (x^i)^2$ so the relation is reflexive. If $(u, v)_\varphi$ is a geometric structure, then $(v, u)_{\varphi^\star}$ is a geometric structure, so the relation is symmetric, which completes the proof that (u, v) is a tolerance relation.

Let $(u, v)_\varphi$ and $(v, w)_\phi$ be two geometric structures. We have $u \circ w^{-1} = \nabla\varphi \circ \nabla\phi$, and so $J_{u \circ w^{-1}} = \mathrm{H}\varphi(\nabla\phi)\mathrm{H}\phi$, that we want to be symmetric positive definite for the "geometric structure" relation to be transitive. Both $\mathrm{H}\varphi$ and $\mathrm{H}\phi$ 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, it follows that $\mathrm{H}\varphi(\nabla\phi)\mathrm{H}\phi \succ 0$ iff we have the diagonalizations $\mathrm{H}\varphi(\nabla\phi) = PD_1P^\top$ and $\mathrm{H}\phi = PD_2P^\top$, with P unitary and $D_1, D_2 \succ 0$. This finishes the proof of Lemma 6.

10.6 Proof of Lemma 8

We suppose without loss of generality that $K = 1$. The proof relies on the study in $[x_1, x_n]$ of function $\tilde{\varphi}'_\perp(x) \doteq -1/\tilde{\varphi}'_{\mathcal{S}}(x)$ (see eq. (24)), which is the slope of the line orthogonal to the segment which links $(\overline{x}, \overline{\varphi})$ to $(\overline{x}, \varphi(\overline{x}))$.

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

Suppose now that $\varphi'(x)$ is > 0 on $[x_1, x_n]$, which implies $\varphi(x_n) \geq \overline{\varphi}$, and so $\overline{x}_\varphi \in [\overline{x}, x_n]$, and satisfies $\varphi(\overline{x}_\varphi) = \overline{\varphi}$. This time, $\tilde{\varphi}'_\perp(\overline{x}) = 0$ and $\lim_{x \uparrow \overline{x}_\varphi} \tilde{\varphi}'_\perp(x) = +\infty$, so there must be $\mu \in [\overline{x}, \overline{x}_\varphi)$ such that $\tilde{\varphi}'_\perp(\mu) = \varphi'(x)$. This μ is a candidate right population minimizer. This ends the proof of Lemma 8.

10.7 Proof of Lemma 9

We build upon eq. (38). Any left population minimizer is a solution of:

$$
\begin{aligned}
0 &= \frac{\mathrm{d}}{\mathrm{d}\mu} \sum_i w_i D^v_{\varphi,g}(\mu/w_i : x_i/w_i) \\
&= \sum_i g\left(\frac{x_i}{w_i}\right) v'\left(\frac{\mu}{w_i}\right) \left(u\left(\frac{\mu}{w_i}\right) - u\left(\frac{x_i}{w_i}\right)\right) \\
&= \sum_i g\left(\frac{x_i}{w_i}\right) v'\left(\frac{\mu}{w_i}\right) \left(\frac{\mu - x_i}{w_i}\right) u'\left(\frac{\mu_i}{w_i}\right) \quad (56)
\end{aligned}
$$

where $\mu_i \doteq \mu + \alpha_i(x_i - \mu)$ for some $0 < \alpha_i < 1$. Eq (56) is obtained after n Taylor expansions of u. We also have $(v \circ u^{-1})' = (v' \circ u^{-1})/(u' \circ u^{-1}) \doteq (\varphi^\star)''$, and so, since $\varphi^\star$ 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 $\pi_i \doteq g(x_i/w_i)\, v'(\mu/w_i)\, u'(\mu_i/w_i)$ is non negative, $\forall i$. We can summarize (56) as $h(\mu) \doteq \sum_i \pi_i(\mu - x_i)/w_i = 0$: 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 (56) 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 μ, this is the left population minimizer. This ends the proof of Lemma 9.