A scaled Bregman theorem with applications

Richard Nock Aditya Krishna Menon Cheng Soon Ong
Data61 and the Australian National University
{richard.nock, aditya.menon, chengsoon.ong}@data61.csiro.au

Abstract

Bregman divergences play a central role in the design and analysis of a range of machine learning algorithms. This paper explores the use of Bregman divergences to establish *reductions* between such algorithms and their analyses. We present a new *scaled isodistortion theorem involving Bregman divergences* (scaled Bregman theorem for short) which shows that certain "Bregman distortions" (employing a potentially *non-convex* generator) may be *exactly* re-written as a scaled Bregman divergence computed over transformed data. Admissible distortions include geodesic distances on curved manifolds and projections or gauge-normalisation, while admissible data include scalars, vectors and matrices.

Our theorem allows one to leverage to the wealth and convenience of Bregman divergences when analysing algorithms relying on the aforementioned Bregman distortions. We illustrate this with three novel applications of our theorem: a reduction from multi-class density ratio to class-probability estimation, a new adaptive projection free yet norm-enforcing dual norm mirror descent algorithm, and a reduction from clustering on flat manifolds to clustering on curved manifolds. Experiments on each of these domains validate the analyses and suggest that the scaled Bregman theorem might be a worthy addition to the popular handful of Bregman divergence properties that have been pervasive in machine learning.

1 Introduction: Bregman divergences as a reduction tool

Bregman divergences play a central role in the design and analysis of a range of machine learning algorithms. In recent years, Bregman divergences have arisen in procedures for convex optimisation [Beck and Teboulle, 2003], online learning [Cesa-Bianchi and Lugosi, 2006, Chapter 11] clustering [Banerjee et al., 2005], matrix approximation [Dhillon and Tropp, 2008], class-probability estimation [Buja et al., 2005, Nock and Nielsen, 2009, Reid and Williamson, 2010, 2011], density ratio estimation [Sugiyama et al., 2012], boosting [Collins et al., 2002], variational inference [Hernández-Lobato et al., 2016], and computational geometry [Boissonnat et al., 2010]. Despite these being very different applications, many of these algorithms and their analyses basically rely on three beautiful analytic properties of Bregman divergences, properties that we summarize for differentiable scalar convex functions φ with derivative φ', conjugate $\varphi^\star$, and divergence D_φ:

- the triangle equality: $D_\varphi(x\|y) + D_\varphi(y\|z) - D_\varphi(x\|z) = (\varphi'(z) - \varphi'(y))(x - y)$;
- the dual symmetry property: $D_\varphi(x\|y) = D_{\varphi^\star}(\varphi'(y)\|\varphi'(x))$;
- the right-centroid (population minimizer) is the average: $\arg\min_\mu \mathbb{E}[D_\varphi(\mathsf{X}\|\mu)] = \mathbb{E}[\mathsf{X}]$.

Casting a problem as a Bregman minimisation allows one to employ these properties to simplify analysis; for example, by interpreting mirror descent as applying a particular Bregman regulariser, Beck and Teboulle [2003] relied on the triangle equality above to simplify its proof of convergence.

Another intriguing possibility is that one may derive *reductions* amongst learning problems by connecting their underlying Bregman minimisations. Menon and Ong [2016] recently established how (binary) density ratio estimation (DRE) can be exactly reduced to class-probability estimation (CPE). This was facilitated by interpreting CPE as a Bregman minimisation [Buja et al., 2005, Section 19], and a new property of Bregman divergences —

Problem A	Problem B that Theorem 1 reduces A to	Reference
Multiclass density-ratio estimation	Multiclass class-probability estimation	§3, Lemma 2
Online optimisation on L_q ball	Convex unconstrained online learning	§4, Lemma 4
Clustering on curved manifolds	Clustering on flat manifolds	§5, Lemma 5

Table 1: Applications of our scaled Bregman Theorem (Theorem 1) — "Reduction" encompasses shortcuts on algorithms *and* on analyses (algorithm/proof A uses algorithm/proof B as subroutine).

Menon and Ong [2016, Lemma 2] showed that for *any* twice differentiable scalar convex φ, for $g(x) = 1 + x$ and $\varphi^\dagger(x) \doteq g(x) \cdot \varphi(x/g(x))$,

$$g(x) \cdot D_\varphi(x/g(x) \| y/g(y)) = D_{\varphi^\dagger}(x \| y). \tag{1}$$

Since the binary class-probability function $\eta(\boldsymbol{x}) = \Pr(\mathsf{Y} = 1 | \mathsf{X} = \boldsymbol{x})$ is related to the class-conditional density ratio $r(\boldsymbol{x}) = \Pr(\mathsf{X} = \boldsymbol{x} | \mathsf{Y} = 1) / \Pr(\mathsf{X} = \boldsymbol{x} | \mathsf{Y} = -1)$ via Bayes' rule as $\eta(\boldsymbol{x}) = r(\boldsymbol{x}) / g(r(\boldsymbol{x}))$, any $\hat{\eta}$ with small $D_\varphi(\eta \| \hat{\eta})$ implicitly produces an $\hat{r}$ with low $D_{\varphi^\dagger}(r \| \hat{r})$ i.e. a good estimate of the density ratio. The Bregman property of Equation 1 thus establishes a reduction from DRE to CPE. Two natural questions arise from this analysis: can we generalise Equation 1 to other $g(\cdot)$, and if so, can we similarly relate *other* problems to each other?

This paper presents a new Bregman identity (Theorem 1), the *scaled Bregman theorem*, a significant generalisation of Menon and Ong [2016, Lemma 2]. It shows that general *distortions* $D_{\varphi^\dagger}$ – which are not necessarily convex, positive, bounded or symmetric – may be re-expressed as a Bregman divergence D_φ computed over transformed data, where this transformation can be as simple as a projection or normalisation by a gauge, or more involved like the exponential map on lifted coordinates for a curved manifold. Interestingly, candidate distortions include geodesic distances on curved manifolds. Equivalently, Theorem 1 shows various distortions can be "reverse engineered" as Bregman divergences (despite appearing *prima facie* to be a very different object), and thus inherit their good properties. Hence, Bregman divergences can embed several distances in a different — and arguably less involved — way than the transformations known to date [Acharyya et al., 2013].

As with the aforementioned key properties of Bregman divergences, Theorem 1 has potentially wide applicability. We present three such novel applications (see Table 1) to vastly different problems:

- a reduction of multiple density ratio estimation to multiclass-probability estimation (§3), generalising the results of Menon and Ong [2016] for the binary label case,
- a *projection-free* yet norm-enforcing mirror gradient algorithm (enforced norms are those of mirrored vectors *and* of the offset) with guarantees for adaptive filtering (§4), and
- a seeding approach for clustering on positively or negatively (constant) curved manifolds based on a popular seeding for flat manifolds and with the same approximation guarantees (§5).

Experiments on each of these domains (§6) validate our analysis. The Supplementary Material details the proofs of all results, provides the experimental results *in extenso* and some additional (nascent) applications of the scaled Bregman theorem to exponential families and computational geometry.

2 Main result: the scaled Bregman theorem

In the remaining, $[k] \doteq \{0, 1, ..., k\}$ and $[k]_* \doteq \{1, 2, ..., k\}$ for $k \in \mathbb{N}$. For any differentiable (but not necessarily convex) $\varphi : \mathcal{X} \to \mathbb{R}$, we define the Bregman "distortion" D_φ as

$$D_\varphi(\boldsymbol{x} \| \boldsymbol{y}) \quad \doteq \quad \varphi(\boldsymbol{x}) - \varphi(\boldsymbol{y}) - (\boldsymbol{x} - \boldsymbol{y})^\top \nabla \varphi(\boldsymbol{y}) \ . \tag{2}$$

When φ is convex, D_φ is the familiar Bregman divergence with generator φ.

Without further ado, we present our main result.

Theorem 1 *Let, $\varphi : \mathcal{X} \to \mathbb{R}$ be convex differentiable, and $g : \mathcal{X} \to \mathbb{R}_*$ be differentiable. Then,*

$$g(\boldsymbol{x}) \cdot D_{\varphi}\left(\boldsymbol{x}/g(\boldsymbol{x}) \,\|\, \boldsymbol{y}/g(\boldsymbol{y})\right) = D_{\varphi^\dagger}\left(\boldsymbol{x} \,\|\, \boldsymbol{y}\right) \ , \forall \boldsymbol{x}, \boldsymbol{y} \in \mathcal{X} \ , \tag{3}$$

$$\text{where } \varphi^\dagger(\boldsymbol{x}) \doteq g(\boldsymbol{x}) \cdot \varphi\left(\boldsymbol{x}/g(\boldsymbol{x})\right) \ , \tag{4}$$

if and only if (i) g is affine on $\mathcal{X}$, and/or (ii) for every $\boldsymbol{z} \in \mathcal{X}_g \doteq \{(1/g(\boldsymbol{x})) \cdot \boldsymbol{x} : \boldsymbol{x} \in \mathcal{X}\}$,

$$\varphi\left(\boldsymbol{z}\right) = \boldsymbol{z}^\top \nabla \varphi(\boldsymbol{z}) \ . \tag{5}$$

Table 2 presents some examples of (sometimes involved) triplets $(D_\varphi, D_{\varphi^\dagger}, g)$ for which Equation 3 holds; related proofs are in Appendix C. If we fold g into D_φ in the left hand-side of eq. (3), then Theorem 1 states a scaled isodistortion (*sometimes* it turns out to be equivalently an *adaptive* scaled isometry, see Appendix I) property between $\mathcal{X}$ and $\mathcal{X}_g$. Because D_φ is such an important object, we do not perform this folding and refer to Theorem 1 as the *scaled Bregman theorem* for short.

Remark. If $\mathcal{X}_g$ is a vector space, φ satisfies Equation 5 if and only if it is positive homogeneous of degree 1 on $\mathcal{X}_g$ (i.e. $\varphi(\alpha \boldsymbol{z}) = \alpha \cdot \varphi(\boldsymbol{z})$ for any $\alpha > 0$) from Euler's homogenous function theorem. When $\mathcal{X}_g$ is not a vector space, this only holds for α such that $\alpha \boldsymbol{z} \in \mathcal{X}_g$ as well. We thus call the gradient condition of Equation 5 "restricted positive homogeneity" for simplicity. ■

Remark. Appendix D gives a "deep composition" extension of Theorem 1. ■

For the special case where $\mathcal{X} = \mathbb{R}$, and $g(x) = 1 + x$, Theorem 1 is exactly Menon and Ong [2016, Lemma 2] (c.f. Equation 1). We wish to highlight a few points with regard to our more general result. First, the "distortion" generator $\varphi^\dagger$ may be[1] *non-convex*, as the following illustrates.

Example. Suppose $\varphi(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{x}\|_2^2$ corresponds to the generator for squared Euclidean distance. Then, for $g(\boldsymbol{x}) = 1 + \mathbf{1}^\top \boldsymbol{x}$, we have $\varphi^\dagger(\boldsymbol{x}) = \frac{1}{2} \cdot \frac{\|\boldsymbol{x}\|_2^2}{1 + \mathbf{1}^\top \boldsymbol{x}}$, which is non-convex on $\mathcal{X} = \mathbb{R}^d$. ■

When $\varphi^\dagger$ is non-convex, the right hand side in Equation 3 is an object that ostensibly bears only a superficial similarity to a Bregman divergence; it is somewhat remarkable that Theorem 1 shows this general "distortion" between a pair $(\boldsymbol{x}, \boldsymbol{y})$ to be entirely equivalent to a (scaling of a) Bregman divergence between some transformation of the points. Second, when g is linear, Equation 3 holds for *any* convex φ. (This was the case considered in Menon and Ong [2016].) When g is non-linear, however, φ must be chosen carefully so that (φ, g) satisfies the restricted homogeneity conditon[2] of Equation 5. In general, given a convex φ, one can "reverse engineer" a suitable g to guarantee this conditon, as illustrated by the following example.

Example. Suppose[3] $\varphi(\boldsymbol{x}) = (1 + \|\boldsymbol{x}\|_2^2)/2$. Then, Equation 5 requires that $\|\boldsymbol{x}\|_2^2 = 1$ for every $\boldsymbol{x} \in \mathcal{X}_g$, i.e. $\mathcal{X}_g$ is (a subset of) the unit sphere. This is afforded by the choice $g(\boldsymbol{x}) = \|\boldsymbol{x}\|_2$. ■

Third, Theorem 1 is not merely a mathematical curiosity: we now show that it facilitates novel results in three very different domains, namely estimating multiclass density ratios, constrained online optimisation, and clustering data on a manifold with non-zero curvature. We discuss nascent applications to exponential families and computational geometry in Appendices E and F.

3 Multiclass density-ratio estimation via class-probability estimation

Given samples from a number of densities, density ratio estimation concerns estimating the ratio between each density and some reference density. This has applications in the covariate shift problem wherein the train and test distributions over instances differ [Shimodaira, 2000]. Our first application of Theorem 1 is to show how density ratio estimation can be reduced to class-probability estimation [Buja et al., 2005, Reid and Williamson, 2010].

[1] Evidently, $\varphi^\dagger$ is convex iff g is non-negative, by Equation (3) and the fact that a function is convex iff its Bregman "distortion" is nonnegative [Boyd and Vandenberghe, 2004, Section 3.1.3].

[2] We stress that this condition only needs to hold on $\mathcal{X}_g \subseteq \mathcal{X}$.; it would not be really interesting in general for φ to be homogeneous *everywhere* in its domain, since we would basically have $\varphi^\dagger = \varphi$.

[3] The constant $1/2$ added in φ does *not* change D_φ, since a Bregman divergence is invariant to affine terms; removing this however would make the divergences D_φ and $D_{\varphi^\dagger}$ differ by a constant.

$\mathcal{X}$	$D_\varphi\left(\boldsymbol{x}\middle\|\boldsymbol{y}\right)$	$D_{\varphi^\dagger}\left(\boldsymbol{x}\middle\|\boldsymbol{y}\right)$	$g(\boldsymbol{x})$
$\mathbb{R}^d$	$\frac{1}{2}\cdot\left\|\boldsymbol{x}-\boldsymbol{y}\right\|_2^2$	$\left\|\boldsymbol{x}\right\|_2\cdot(1-\cos\angle\boldsymbol{x},\boldsymbol{y})$	$\left\|\boldsymbol{x}\right\|_2$
$\mathbb{R}^d$	$\frac{1}{2}\cdot(\left\|\boldsymbol{x}\right\|_q^2-\left\|\boldsymbol{y}\right\|_q^2)-\sum_i\frac{(x_i-y_i)\cdot\text{sign}(y_i)\cdot\left\|y_i\right\|^{q-1}}{\left\|\boldsymbol{y}\right\|_q^{q-2}}$	$W\cdot\left\|\boldsymbol{x}\right\|_q-W\cdot\sum_i\frac{x_i\cdot\text{sign}(y_i)\cdot\left\|y_i\right\|^{q-1}}{\left\|\boldsymbol{y}\right\|_q^{q-1}}$	$\left\|\boldsymbol{x}\right\|_q/W$
$\mathbb{R}^d\times\mathbb{R}$	$\frac{1}{2}\cdot\left\|\boldsymbol{x}^S-\boldsymbol{y}^S\right\|_2^2$	$\frac{\left\|\boldsymbol{x}\right\|_2}{\sin\left\|\boldsymbol{x}\right\|_2}\cdot(1-\cos D_G(\boldsymbol{x},\boldsymbol{y}))$	$\left\|\boldsymbol{x}\right\|_2/\sin\left\|\boldsymbol{x}\right\|_2$
$\mathbb{R}^d\times\mathbb{C}$	$\frac{1}{2}\cdot\left\|\boldsymbol{x}^H-\boldsymbol{y}^H\right\|_2^2$	$-\frac{\left\|\boldsymbol{x}\right\|_2}{\sinh\left\|\boldsymbol{x}\right\|_2}\cdot(\cosh D_G(\boldsymbol{x},\boldsymbol{y})-1)$	$-\left\|\boldsymbol{x}\right\|_2/\sinh\left\|\boldsymbol{x}\right\|_2$
$\mathbb{R}_+^d$	$\sum_i x_i\log\frac{x_i}{y_i}-\mathbf{1}^\top(\boldsymbol{x}-\boldsymbol{y})$	$\sum_i x_i\log\frac{x_i}{y_i}-d\cdot\mathbb{E}[\mathsf{X}]\cdot\log\frac{\mathbb{E}[\mathsf{X}]}{\mathbb{E}[\mathsf{Y}]}$	$\mathbf{1}^\top\boldsymbol{x}$
$\mathbb{R}_+^d$	$\sum_i\frac{x_i}{y_i}-\sum_i\log\frac{x_i}{y_i}-d$	$\sum_i\frac{x_i(\prod_j y_j)^{1/d}}{y_i}-d(\prod_j x_j)^{1/d}$	$\prod_i x_i^{1/d}$
$\mathbf{S}(d)$	$\text{tr}\left(\mathrm{X}\log\mathrm{X}-\mathrm{X}\log\mathrm{Y}\right)-\text{tr}\left(\mathrm{X}\right)+\text{tr}\left(\mathrm{Y}\right)$	$\text{tr}\left(\mathrm{X}\log\mathrm{X}-\mathrm{X}\log\mathrm{Y}\right)-\text{tr}\left(\mathrm{X}\right)\cdot\log\frac{\text{tr}(\mathrm{X})}{\text{tr}(\mathrm{Y})}$	$\text{tr}\left(\mathrm{X}\right)$
$\mathbf{S}(d)$	$\text{tr}\left(\mathrm{X}\mathrm{Y}^{-1}\right)-\log\det(\mathrm{X}\mathrm{Y}^{-1})-d$	$\det(\mathrm{Y}^{1/d})\text{tr}\left(\mathrm{X}\mathrm{Y}^{-1}\right)-d\cdot\det(\mathrm{X}^{1/d})$	$\det(\mathrm{X}^{1/d})$

Table 2: Examples of $(D_\varphi, D_{\varphi^\dagger}, g)$ for which Equation 3 holds. Function $\boldsymbol{x}^S \doteq f(\boldsymbol{x}) : \mathbb{R}^d \to \mathbb{R}^{d+1}$ and $\boldsymbol{x}^H \doteq f(\boldsymbol{x}) : \mathbb{R}^d \to \mathbb{R}^d \times \mathbb{C}$ are the Sphere and Hyperbolic lifting maps defined in Equation 67, 78. $W > 0$ is a constant. D_G denotes the Geodesic distance on the sphere (for $\boldsymbol{x}^S$) or the hyperboloid (for $\boldsymbol{x}^H$). $\mathbf{S}(d)$ is the set of symmetric real matrices. Related proofs are in Section C.

To proceed, we fix notation. For some integer $C \geq 1$, consider a distribution $\mathbb{P}(\mathsf{X}, \mathsf{Y})$ over an (instance, label) space $\mathcal{X} \times [C]$. Let $(\{P_c\}_{c=1}^C, \boldsymbol{\pi})$ be densities giving $\mathbb{P}(\mathsf{X}|\mathsf{Y} = c)$ and $\mathbb{P}(\mathsf{Y} = c)$ respectively, and M giving $\mathbb{P}(\mathsf{X})$ accordingly. Fix $c^* \in [C]$ a reference class, and suppose for simplicity that $c^* = C$. Let $\tilde{\boldsymbol{\pi}} \in \triangle^{C-1}$ such that $\tilde{\pi}_c \doteq \pi_c/(1-\pi_C)$. *Density ratio estimation* [Sugiyama et al., 2012] concerns inferring the vector $\boldsymbol{r}(\boldsymbol{x}) \in \mathbb{R}^{C-1}$ of density ratios relative to C, with $r_c(\boldsymbol{x}) \doteq \mathbb{P}(\mathsf{X} = \boldsymbol{x}|\mathsf{Y} = c)/\mathbb{P}(\mathsf{X} = \boldsymbol{x}|\mathsf{Y} = C)$, while *class-probability estimation* [Buja et al., 2005] concerns inferring the vector $\boldsymbol{\eta}(\boldsymbol{x}) \in \mathbb{R}^{C-1}$ of class-probabilities, with $\eta_c(\boldsymbol{x}) \doteq \mathbb{P}(\mathsf{Y} = c|\mathsf{X} = \boldsymbol{x})/\tilde{\pi}_c$. In both cases, we estimate the respective quantities given an iid sample $\mathsf{S} \sim \mathbb{P}(\mathsf{X}, \mathsf{Y})^N$.

The genesis of the reduction from density ratio to class-probability estimation is the fact that $\boldsymbol{r}(\boldsymbol{x}) = (\pi_C/(1-\pi_C)) \cdot \boldsymbol{\eta}(\boldsymbol{x})/\eta_C(\boldsymbol{x})$. In practice one will only have an estimate $\hat{\boldsymbol{\eta}}$, typically derived by minimising a suitable loss on the given S [Williamson et al., 2014], with a canonical example being multiclass logistic regression. Given $\hat{\boldsymbol{\eta}}$, it is natural to estimate the density ratio via:

$$\hat{\boldsymbol{r}}(\boldsymbol{x}) \quad = \quad \frac{\hat{\boldsymbol{\eta}}(\boldsymbol{x})}{\hat{\eta}_C(\boldsymbol{x})} \ . \tag{6}$$

While this estimate is intuitive, to establish a formal reduction we must relate the quality of $\hat{\boldsymbol{r}}$ to that of $\hat{\boldsymbol{\eta}}$. Since the minimisation of a suitable loss for class-probability estimation is equivalent to a Bregman minimisation [Buja et al., 2005, Section 19], [Williamson et al., 2014, Proposition 7], this is however immediate by Theorem 1, as shown below.

Lemma 2 *Given a class-probability estimator $\hat{\eta}\colon \mathcal{X} \to [0,1]^{C-1}$, let the density ratio estimator $\hat{r}$ be as per Equation 6. Then for any convex differentiable $\varphi\colon [0,1]^{C-1} \to \mathbb{R}$,*

$$\mathbb{E}_{\mathsf{X}\sim M}[D_\varphi(\boldsymbol{\eta}(\mathsf{X})\|\hat{\boldsymbol{\eta}}(\mathsf{X}))] \quad = \quad (1-\pi_C)\cdot\mathbb{E}_{\mathsf{X}\sim P_C}\left[D_{\varphi^\dagger}(\boldsymbol{r}(\mathsf{X})\|\hat{\boldsymbol{r}}(\mathsf{X}))\right] \tag{7}$$

where $\varphi^\dagger$ is as per Equation 4 with $g(\boldsymbol{x}) \doteq \pi_C/(1-\pi_C) + \tilde{\boldsymbol{\pi}}^\top\boldsymbol{x}$.

Lemma 2 generalises Menon and Ong [2016, Proposition 3], which focussed on the binary case with $\pi = 1/2$. (See Appendix G for a review of that result.) Unpacking the Lemma, the LHS in Equation 7 represents the object minimised by some suitable loss for class-probability estimation. Since g is affine, we can use *any* convex, differentiable φ, and so can use *any* suitable class-probability loss to estimate $\hat{\boldsymbol{\eta}}$. Lemma 2 thus implies that producing $\hat{\boldsymbol{\eta}}$ by minimising any class-probability loss *equivalently* produces an $\hat{\boldsymbol{r}}$ as per Equation 6 that minimises a Bregman divergence to the true $\boldsymbol{r}$. Thus, Theorem 1 provides a reduction from density ratio to multiclass probability estimation.

We now detail two applications where $g(\cdot)$ is no longer affine, and φ must be chosen more carefully.

4 Dual norm mirror descent: projection-free online learning on L_p balls

A substantial amount of work in the intersection of machine learning and convex optimisation has focused on constrained optimisation within a ball [Shalev-Shwartz et al., 2007, Duchi et al., 2008]. This optimisation is typically via projection operators that can be expensive to compute [Hazan and Kale, 2012, Jaggi, 2013]. We now show that *gauge functions* can be used as an inexpensive alternative, and that Theorem 1 easily yields guarantees for this procedure in online learning.

We consider the adaptive filtering problem, closely related to the online least squares problem with linear predictors [Cesa-Bianchi and Lugosi, 2006, Chapter 11]. Here, over a sequence of T rounds, we observe some $\boldsymbol{x}_t \in \mathcal{X}$. We must the predict a target value $\hat{y}_t = \boldsymbol{w}_{t-1}^\top \boldsymbol{x}_t$ using our current weight vector $\boldsymbol{w}_{t-1}$. The true target $y_t = \boldsymbol{u}^\top \boldsymbol{x}_t + \epsilon_t$ is then revealed, where ϵ_t is some unknown noise, and we may update our weight to $\boldsymbol{w}_t$. Our goal is to minimise the regret of the sequence $\{\boldsymbol{w}_t\}_{t=0}^T$,

$$R(\boldsymbol{w}_{1:T}|\boldsymbol{u}) \doteq \sum_{t=1}^T \left(\boldsymbol{u}^\top \boldsymbol{x}_t - \boldsymbol{w}_{t-1}^\top \boldsymbol{x}_t\right)^2 - \sum_{t=1}^T \left(\boldsymbol{u}^\top \boldsymbol{x}_t - y_t\right)^2 \quad . \tag{8}$$

Let $q \in (1, 2]$ and p be such that $1/p + 1/q = 1$. For $\varphi \doteq \frac{1}{2} \cdot \|\boldsymbol{x}\|_q^2$ and loss $\ell_t(\boldsymbol{w}) = \frac{1}{2} \cdot (y_t - \boldsymbol{w}^\top \boldsymbol{x}_t)^2$, the p-LMS algorithm [Kivinen et al., 2006] employs the stochastic mirror gradient updates

$$\boldsymbol{w}_t \doteq \operatorname*{argmin}_{\boldsymbol{w}} \eta_t \cdot \ell_t(\boldsymbol{w}) + D_\varphi(\boldsymbol{w}\|\boldsymbol{w}_{t-1}) = (\nabla\varphi)^{-1}\left(\nabla\varphi(\boldsymbol{w}_{t-1}) - \eta_t \cdot \nabla\ell_t\right), \tag{9}$$

where η_t is a learning rate to be specified by the user. Kivinen et al. [2006, Theorem 2] shows that for appropriate η_t, one has $R(\boldsymbol{w}_{1:T}|\boldsymbol{u}) \leq (p-1) \cdot \max_{\boldsymbol{x}\in\mathcal{X}} \|\boldsymbol{x}\|_p^2 \cdot \|\boldsymbol{u}\|_q^2$.

The p-LMS updates do not provide any explicit control on $\|\boldsymbol{w}_t\|$, i.e. there is no regularisation. Experiments (§6) suggest that leaving $\|\boldsymbol{w}_t\|$ uncontrolled may not be a good idea as the increase of the norm sometimes prevents (significant) updates (9). Also, the wide success of regularisation in machine learning calls for regularised variants that *retain* the regret guarantees and computational efficiency of p-LMS. (Adding a projection step to Equation 9 would not achieve both.) We now do just this. For fixed $W > 0$, let $\varphi \doteq (1/2)(W^2 + \|\boldsymbol{x}\|_q^2)$, a translation of that used in p-LMS. Invoking Theorem 1 with the admissible $g_q(\boldsymbol{x}) = ||\boldsymbol{x}||_q/W$ yields $\varphi^\dagger \doteq \varphi_q^\dagger = W\|\boldsymbol{x}\|_q$ (see Table 2). Using the fact that L_p and L_q norms are dual of each other, we replace Equation 9 by:

$$\boldsymbol{w}_t \doteq \nabla\varphi_p^\dagger\left(\nabla\varphi_q^\dagger(\boldsymbol{w}_{t-1}) - \eta_t \cdot \nabla\ell_t\right) \quad . \tag{10}$$

See Lemma 6 of the Appendix for the simple forms of $\nabla\varphi_{\{p,q\}}^\dagger$. We call update (10) the *dual norm p-LMS (DN-pLMS) algorithm*, noting that the dual refers to the polar transform of the norm, and g stems from a gauge normalization for $\mathcal{B}_q(W)$, the closed L_q ball with radius $W > 0$. Namely, we have $\gamma_{\text{GAU}}(\boldsymbol{x}) = W/\|\boldsymbol{x}\|_q = g(\boldsymbol{x})^{-1}$ for the gauge $\gamma_{\text{GAU}}(\boldsymbol{x}) \doteq \sup\{z \geq 0 : z \cdot \boldsymbol{x} \in \mathcal{B}_q(W)\}$, so that $\varphi_q^\dagger$ implicitly performs gauge normalisation of the data. This update is no more computationally expensive than Equation 9 — we simply need to compute the p- and q-norms of appropriate terms — but, crucially, automatically constrains the norms of $\boldsymbol{w}_t$ and its image by $\nabla\varphi_q^\dagger$.

Lemma 3 *For the update in Equation 10, $\|\boldsymbol{w}_t\|_q = \|\nabla\varphi_q^\dagger(\boldsymbol{w}_t)\|_p = W, \forall t > 0$.*

Lemma 3 is remarkable, since *nowhere in Equation 10 do we project onto the L_q ball*. Nonetheless, for the DN-pLMS updates to be principled, we need a similar regret guarantee to the original p-LMS. Fortunately, this may be done using Theorem 1 to exploit the original proof of Kivinen et al. [2006]. For any $\boldsymbol{u} \in \mathbb{R}^d$, define the *$q$-normalised regret* of $\{\boldsymbol{w}_t\}_{t=0}^T$ by

$$R_q(\boldsymbol{w}_{1:T}|\boldsymbol{u}) \doteq \sum_{t=1}^T \left((1/g_q(\boldsymbol{u})) \cdot \boldsymbol{u}^\top \boldsymbol{x}_t - \boldsymbol{w}_{t-1}^\top \boldsymbol{x}_t\right)^2 - \sum_{t=1}^T \left((1/g_q(\boldsymbol{u})) \cdot \boldsymbol{u}^\top \boldsymbol{x}_t - y_t\right)^2 \quad . \tag{11}$$

We have the following bound on R_q for the DN-pLMS updates. (We cannot expect a bound on the unnormalised $R(\cdot)$ of Equation 8, since by Lemma 3 we can only compete against norm W vectors.)

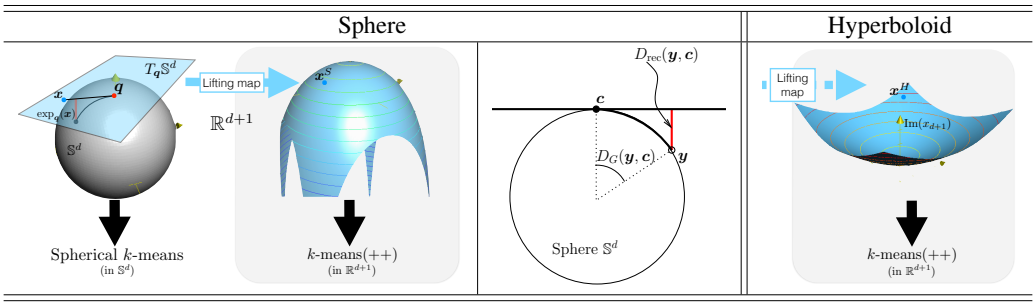

Figure 1: (L) Lifting map into $\mathbb{R}^d \times \mathbb{R}$ for clustering on the sphere with k-means++. (M) D_{rec} in Eq. (15) in vertical thick red line. (R) Lifting map into $\mathbb{R}^d \times \mathbb{C}$ for the hyperboloid.

Lemma 4 *Pick any $\boldsymbol{u} \in \mathbb{R}^d$, p, q satisfying $1/p + 1/q = 1$ and $p > 2$, and $W > 0$. Suppose $\|\boldsymbol{x}_t\|_p \leq X_p$ and $|y_t| \leq Y, \forall t \leq T$. Let $\{\boldsymbol{w}_t\}$ be as per Equation 10, using learning rate*

$$\eta_t \doteq \gamma_t \cdot \frac{W}{4(p-1)\max\{W, X_p\}X_p W + |y_t - \boldsymbol{w}_{t-1}^\top \boldsymbol{x}_t| X_p} , \tag{12}$$

*for **any** desired $\gamma_t \in [1/2, 1]$. Then,*

$$R_q(\boldsymbol{w}_{1:T}|\boldsymbol{u}) \leq 4(p-1)X_p^2 W^2 + (16p-8)\max\{W, X_p\}X_p^2 W + 8YX_p^2 . \tag{13}$$

Several remarks can be made. First, the bound depends on the maximal signal value Y, but this is the maximal signal in the observed sequence, so it may not be very large in practice; if it is comparable to W, then our bound is looser than Kivinen et al. [2006] by just a constant factor. Second, the learning rate is adaptive in the sense that its choice depends on the last mistake made. There is a nice way to represent the "offset" vector $\eta_t \cdot \nabla \ell_t$ in eq. (10), since we have, for $Q'' \doteq 4(p-1)\max\{W, X_p\}X_p W$,

$$\eta_t \cdot \nabla \ell_t = W \cdot \frac{|y_t - \boldsymbol{w}_{t-1}^\top \boldsymbol{x}_t| X_p}{Q'' + |y_t - \boldsymbol{w}_{t-1}^\top \boldsymbol{x}_t| X_p} \cdot \text{sign}(y_t - \boldsymbol{w}_{t-1}^\top \boldsymbol{x}_t) \cdot \left(\frac{1}{X_p} \cdot \boldsymbol{x}\right) , \tag{14}$$

so the L_p norm of the offset is actually equal to $W \cdot Q$, where $Q \in [0, 1]$ is all the smaller as the vector $\boldsymbol{w}_\cdot$ gets better. Hence, the update in eq. (10) controls in fact *all* norms (that of $\boldsymbol{w}_\cdot$, its image by $\nabla \varphi_q^\dagger$ and the offset). Third, because of the normalisation of $\boldsymbol{u}$, the bound actually does not depend on $\boldsymbol{u}$, but on the radius W chosen for the L_q ball.

5 Clustering on a manifold via data transformation

Our final application can be related to two problems that have received a steadily growing interest over the past decade in unsupervised machine learning: clustering on a non-linear manifold [Dhillon and Modha, 2001], and subspace custering [Vidal, 2011]. We consider two fundamental manifolds investigated by Galperin [1993] to compute centers of mass from relativistic theory: the sphere $\mathbb{S}^d$ and the hyperboloid $\mathbb{H}^d$, the former being of positive curvature, and the latter of negative curvature. Applications involving these specific manifolds are numerous in text processing, computer vision, geometric modelling, computer graphics, to name a few [Buss and Fillmore, 2001, Dhillon and Modha, 2001, Endo and Miyamoto, 2015, Kuang et al., 2014, Rong et al., 2010, Shahani et al., 2015, Straub et al., 2015a,b,c, 2014]. We emphasize the fact that the clustering problem has significant practical impact for d as small as 2 in computer vision [Straub et al., 2014].

The problem is non-trivial for two separate reasons. First, the ambient space, *i.e.* the space of registration of the input data, is often implicitly Euclidean and therefore *not* the manifold [Dhillon and Modha, 2001]: if the mapping to the manifold is not carefully done, then geodesic distances measured on the manifold may be inconsistent with respect to the ambient space. Second, the fact that the manifold has non-zero curvature essentially prevents the direct use of Euclidean optimization algorithms [Zhang and Sra, 2016] — put simply, the average of two points

(Sphere) Sk-means++$(\mathcal{S}, k)$	(Hyperboloid) Hk-means++$(\mathcal{S}, k)$
Input: dataset $\mathcal{S} \subset T_{\boldsymbol{q}}\mathbb{S}^d, k \in \mathbb{N}_*$;	**Input:** dataset $\mathcal{S} \subset T_{\boldsymbol{q}}\mathbb{H}^d, k \in \mathbb{N}_*$;
Step 1: $\mathcal{S}^+ \leftarrow \{g_S^{-1}(\boldsymbol{x}^S) \cdot \boldsymbol{x}^S : \boldsymbol{x}^S \in \text{lift}(\mathcal{S})\}$;	Step 1: $\mathcal{S}^+ \leftarrow \{g_H^{-1}(\boldsymbol{x}^H) \cdot \boldsymbol{x}^H : \boldsymbol{x}^H \in \text{lift}(\mathcal{S})\}$;
Step 2: $\mathcal{C}^+ \leftarrow k$-means++_seeding$(\mathcal{S}^+, k)$;	Step 2: $\mathcal{C}^+ \leftarrow k$-means++_seeding$(\mathcal{S}^+, k)$;
Step 3: $\mathcal{C} \leftarrow \exp_{\boldsymbol{q}}^{-1}(\mathcal{C}^+)$;	Step 3: $\mathcal{C} \leftarrow \exp_{\boldsymbol{q}}^{-1}(\mathcal{C}^+)$;
Output: Cluster centers $\mathcal{C} \in T_{\boldsymbol{q}}\mathbb{S}^d$;	**Output:** Cluster centers $\mathcal{C} \in T_{\boldsymbol{q}}\mathbb{H}^d$;
$\boldsymbol{x}^S \doteq [x_1 \quad x_2 \quad \cdots \quad x_d \quad \|\boldsymbol{x}\|_2 \cot \|\boldsymbol{x}\|_2]$	$\boldsymbol{x}^H \doteq [x_1 \quad x_2 \quad \cdots \quad x_d \quad i\|\boldsymbol{x}\|_2 \coth \|\boldsymbol{x}\|_2]$
$g_S(\boldsymbol{x}^S) \doteq \|\boldsymbol{x}\|_2 / \sin \|\boldsymbol{x}\|_2$	$g_H(\boldsymbol{x}^H) \doteq -\|\boldsymbol{x}\|_2 / \sinh \|\boldsymbol{x}\|_2$

Table 3: How to use k-means++ to cluster points on the sphere (left) or the hyperboloid (right).

that belong to a manifold does not necessarily belong to the manifold, so we have to be careful on how to compute centroids for hard clustering [Galperin, 1993, Nock et al., 2016, Rong et al., 2010, Schwander and Nielsen, 2013].

What we show now is that Riemannian manifolds with constant sectional curvature may be clustered with the k-means++ seeding for flat manifolds [Arthur and Vassilvitskii, 2007], *without even touching a line of the algorithm*. To formalise the problem, we need three key components of Riemannian geometry: tangent planes, exponential map and geodesics [Amari and Nagaoka, 2000]. We assume that the ambient space is a tangent plane to the manifold $\mathcal{M}$, which conveniently makes it look Euclidean (see Figure 1). The point of tangency is called $\boldsymbol{q}$, and the tangent plane $T_{\boldsymbol{q}}\mathcal{M}$. The exponential map, $\exp_{\boldsymbol{q}} : T_{\boldsymbol{q}}\mathcal{M} \to \mathcal{M}$, performs a distance preserving mapping: the geodesic length between $\boldsymbol{q}$ and $\exp_{\boldsymbol{q}}(\boldsymbol{x})$ in $\mathcal{M}$ is the same as the Euclidean length between $\boldsymbol{q}$ and $\boldsymbol{x}$ in $T_{\boldsymbol{q}}\mathcal{M}$. Our clustering objective is to find $\mathcal{C} \doteq \{\boldsymbol{c}_1, \boldsymbol{c}_2, ...\boldsymbol{c}_k\} \subset \mathcal{M}$ such that $D_{\text{rec}}(\mathcal{S} : \mathcal{C}) = \inf_{\mathcal{C}' \subset \mathcal{M}, |\mathcal{C}'|=k} D_{\text{rec}}(\mathcal{S}, \mathcal{C}')$, with

$$D_{\text{rec}}(\mathcal{S}, \mathcal{C}) \quad \doteq \sum_{i \in [m]_*} \min_{j \in [k]_*} D_{\text{rec}}(\exp_{\boldsymbol{q}}(\boldsymbol{x}_i), \boldsymbol{c}_j) \ , \tag{15}$$

where D_{rec} is a *rec*onstruction loss, a function of the geodesic distance between $\exp_{\boldsymbol{q}}(\boldsymbol{x}_i)$ and $\boldsymbol{c}_j$. We use two loss functions defined from Galperin [1993] and used in machine learning for more than a decade [Dhillon and Modha, 2001]:

$$\mathbb{R}_+ \ni D_{\text{rec}}(\boldsymbol{y}, \boldsymbol{c}) \quad \doteq \quad \left\{ \begin{array}{ccc} 1 - \cos D_G(\boldsymbol{y}, \boldsymbol{c}) & \text{for} & \mathcal{M} = \mathbb{S}^d \\ \cosh D_G(\boldsymbol{y}, \boldsymbol{c}) - 1 & \text{for} & \mathcal{M} = \mathbb{H}^d \end{array} \right. . \tag{16}$$

Here, $D_G(\boldsymbol{y}, \boldsymbol{c})$ is the corresponding geodesic distance of $\mathcal{M}$ between $\boldsymbol{y}$ and $\boldsymbol{c}$. Figure 1 shows that $D_{\text{rec}}(\boldsymbol{y}, \boldsymbol{c})$ is the orthogonal distance between $T_{\boldsymbol{c}}\mathcal{M}$ and $\boldsymbol{y}$ when $\mathcal{M} = \mathbb{S}^d$. The solution to the clustering problem in eq. (15) is therefore the one that minimizes the error between tangent planes defined at the centroids, and points on the manifold.

It turns out that both distances in 16 can be engineered as Bregman divergences via Theorem 1, as seen in Table 2. Furthermore, they imply the same φ, which is just the generator of Mahalanobis distortion, but a different g. The construction involves a third party, a *lifting map* (lift$(.)$) that increases the dimension by one. The *Sphere* lifting map $\mathbb{R}^d \ni \boldsymbol{x} \mapsto \boldsymbol{x}^S \in \mathbb{R}^{d+1}$ is indicated in Table 3 (left). The new coordinate depends on the norm of $\boldsymbol{x}$. The *Hyperbolic* lifting map, $\mathbb{R}^d \ni \boldsymbol{x} \mapsto \boldsymbol{x}^H \in \mathbb{R}^d \times \mathbb{C}$, involves a pure imaginary additional coordinate, is indicated in in Table 3 (right, with a slight abuse of notation) and Figure 1. Both $\boldsymbol{x}^S$ and $\boldsymbol{x}^H$ live on a d-dimensional manifold, depicted in Figure 1. When they are scaled by the corresponding $g_.(.)$, they happen to be mapped to $\mathbb{S}^d$ or $\mathbb{H}^d$, respectively, by what happens to be the manifold's exponential map for the original $\boldsymbol{x}$ (see Appendix C).

Theorem 1 is interesting in this case because φ corresponds to a Mahalanobis distortion: this shows that k-means++ seeding [Arthur and Vassilvitskii, 2007, Nock et al., 2008] can be used directly on the scaled coordinates $(g_{\{S,H\}}^{-1}(\boldsymbol{x}^{\{S,H\}}) \cdot \boldsymbol{x}^{\{S,H\}})$ to pick centroids that yield an approximation of the global optimum for the clustering problem on the manifold which is just *as good as* the original Euclidean approximation bound [Arthur and Vassilvitskii, 2007].

Lemma 5 *The expected potential of Sk-means++ seeding over the random choices of $\mathcal{C}^+$ satisfies:*

$$\mathbb{E}[D_{\text{rec}}(\mathcal{S} : \mathcal{C})] \quad \leq \quad 8(2 + \log k) \cdot \inf_{\mathcal{C}' \in \mathbb{S}^d} D_{\text{rec}}(\mathcal{S} : \mathcal{C}') \ . \tag{17}$$

The same approximation bounds holds for Hk-means++ seeding on the hyperboloid ($\mathcal{C}', \mathcal{C}^+ \in \mathbb{H}^d$).

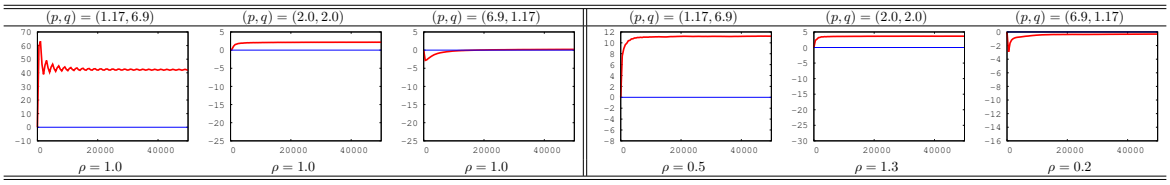

Table 4: Summary of the experiments displaying (y) the error of p-LMS minus error of DN-p-LMS (when > 0, DN-p-LMS beats p-LMS) as a function of t, in the setting of Kivinen et al. [2006], for various values of (p, q) (columns). Left panel: (D)ense target; Right panel: (S)parse target.

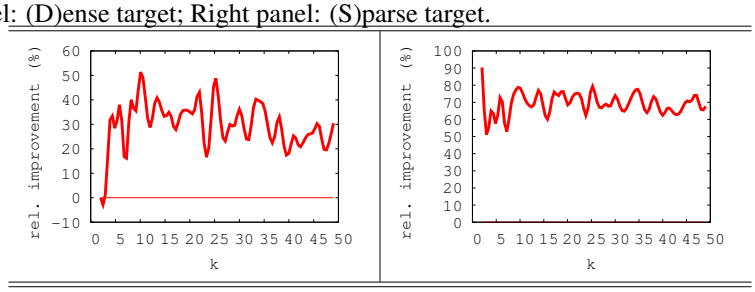

Table 5: (L) Relative improvement (decrease) in k-means potential of SKM$\circ$Sk-means++ compared to SKM alone. (R) Relative improvement of Sk-means++ over Forgy initialization on the sphere.

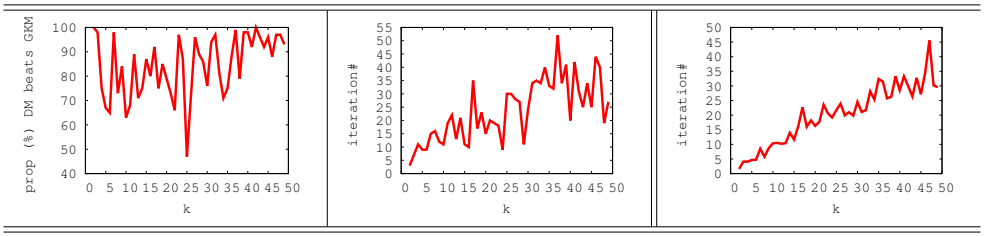

Table 6: (L) % of the number of runs of SKM whose output (*when it has converged*) is better than Sk-means++. (C) Maximal # of iterations for SKM after which it beats Sk-means++ (ignoring runs of SKM that do not beat Sk-means++). (R) Average # of iterations for SKM to converge.

Lemma 5 is notable since it was only recently shown that such a bound is possible for the sphere [Endo and Miyamoto, 2015], and to our knowledge, no such approximation quality is known for clustering on the hyperboloid [Rong et al., 2010, Schwander and Nielsen, 2013]. Notice that Lloyd iterations on non-linear manifolds would require repetitive renormalizations to keep centers on the manifold [Dhillon and Modha, 2001], an additional disadvantage compared to clustering on flat manifolds that $\{\mathrm{G}, \mathrm{K}\}$-means++ seedings do not bear.

6 Experimental validation

We present some experiments validating our theoretical analysis for the applications above.

Multiple density ratio estimation. See Appendix H.1 for experiments in this domain.

Dual norm p-LMS (DN-p-LMS). We ran p-LMS and the DN-pLMS of §4 on the experimental setting of Kivinen et al. [2006]. We refer to that paper for an exhaustive description of the experimental setting, which we briefly summarize: it is a noisy signal processing setting, involving a dense or a sparse target. We compute, over the signal received, the error of our predictor on the signal. We keep all parameters as they are in [Kivinen et al., 2006], except for one: we make sure that data are scaled to fit in a L_p ball of prescribed radius, to test the assumption related in [Kivinen et al., 2006] that fixing the learning rate η_t is not straightforward in p-LMS. Knowing the true value of X_p, we then scale it by a misestimation factor ρ, typically in $[0.1, 1.7]$. We use the same misestimation in DN-p-LMS. Thus, both algorithms suffer the same source of uncertainty. Also, we periodically change the signal (each 1000 iterations), to assess the performances of the algorithms in tracking changes in the signal.

Experiments, given *in extenso* in Appendix H.2, are sumarized in Table 4. The following trends emerge: in the mid to long run, DN-p-LMS is never beaten by p-LMS by more than a fraction of percent. On the other hand, DN-p-LM can beat p-LMS by very significant differences (exceeding 40%), in particular when $p < 2$, *i.e.* when we are outside the regime of the proof of Kivinen et al. [2006]. This indicates that significantly stronger and more general results than the one of Lemma 4 may be expected. Also, it seems that the problem of p-LMS lies in an "exploding" norm problem: in various cases, we observe that $\|\boldsymbol{w}_t\|$ (in any norm) blows up with t, and this correlates with a very significant degradation of its performances. Clearly, DN-p-LMS does not have this problem since all relevant norms are under tight control. Finally, even when the norm does not explode, DN-p-LMS can still beat p-LMS, by less important differences though. Of course, the output of p-LMS can repeatedly be normalised, but the normalisation would escape the theory of Kivinen et al. [2006] and it is not clear which kind of normalisation would bring the best results.

Clustering on the sphere. For $k \in [50]_*$, we simulate on $T_{\mathbf{0}}\mathbb{S}^2$ a mixture of spherical Gaussian and uniform densities with $2k$ components. We run three algorithms: (i) SKM [Dhillon and Modha, 2001] on the data embedded on $\mathbb{S}^2$ with random (Forgy) initialization, (ii), Sk-means++ and (iii) SKM with Sk-means++ initialisation. Results are averaged over the algorithms' runs.

Table 5 (left) displays that using Sk-means++ as initialization for SKM brings a very significant leverage over SKM alone, since we almost divide the k-means potential by a factor 2 on some runs. The right plot of Table 5 shows that S-k-means++ consistently reduces the k-means potential by at least a factor 2 over Forgy. The left plot in Table 6 displays that even when it has converged, SKM does *not* necessarily beat Sk-means++. Finally, the center+right plots in Table 6 display that even when it does beat Sk-means++ when it has converged, the iteration number after which SKM beats Sk-means++ increases with k, and in the worst case may *exceed* the average number of iterations needed for SKM to converge (we stopped SKM if relative improvement is not above $1‰$).

7 Conclusion

We presented a new scaled Bregman identity (Theorem 1), and used it to derive novel results in multiple density ratio estimation, adaptive filtering, and clustering on curved manifolds. We believe that, like other established properties of Bregman divergences, there is potential for several other applications of the result; Appendix E, F present preliminary thoughts in this direction.

A Additional helper lemmas

We begin with some helper lemmas that will be used in some of the proofs. In what follows, let

$$\varphi_q(\boldsymbol{w}) = (1/2)(W^2 + \|\boldsymbol{w}\|_q^2)$$
$$\varphi_q^\dagger(\boldsymbol{w}) = W \cdot \|\boldsymbol{w}\|_q$$

for some $W > 0$ and $p, q \in (1, \infty)$ such that $1/p + 1/q = 1$.

A.1 Properties of φ_q and $\varphi_q^\dagger$

We use the following properties of $\varphi, \varphi^\dagger$.

Lemma 6 *For any $\boldsymbol{w}$,*

$$\nabla\varphi_q(\boldsymbol{w}) = \|\boldsymbol{w}\|_q^{2-q} \cdot \operatorname{sign}(\boldsymbol{w}) \otimes |\boldsymbol{w}|^{q-1}$$
$$\nabla\varphi_q^\dagger(\boldsymbol{w}) = W \cdot \|\boldsymbol{w}\|_q^{1-q} \cdot \operatorname{sign}(\boldsymbol{w}) \otimes |\boldsymbol{w}|^{q-1},$$

where $\otimes$ denotes Hadamard product.

Proof The first identity was shown in Kivinen et al. [2006, Example 1]. The second identity follows from a simple calculation. ■

This implies the follows useful relation between the gradients of φ_q and $\varphi_q^\dagger$.

Corollary 7 *For any $\boldsymbol{w}$,*

$$\nabla\varphi_q(\boldsymbol{w}) = (\|\boldsymbol{w}\|_q / W) \cdot \nabla\varphi_q^\dagger(\boldsymbol{w})$$
$$\|\nabla\varphi_q^\dagger(\boldsymbol{w})\|_p = W$$
$$\|\nabla\varphi_q(\boldsymbol{w})\|_p = \|\boldsymbol{w}\|_q.$$

Proof [Proof of Corollary 7] The proof follows by direct application of Lemma 6 and the definition of p, q. Note the third identity was shown in Kivinen et al. [2006, Appendix I]. ■

As a consequence, we conclude that the gradients of φ and $\varphi^\dagger$ coincide when considering vectors on the W-sphere.

Lemma 8 *For any $\|\boldsymbol{w}\|_q = W$,*

$$\nabla\varphi_q(\boldsymbol{w}) = \nabla\varphi_q^\dagger(\boldsymbol{w}).$$

Proof This follows from the relation between $\nabla\varphi_q$ and $\nabla\varphi_q^\dagger$ from Lemma 6. ■

Finally, we have the following result about the composition of gradients.

Lemma 9 *For any $\boldsymbol{w}$,*

$$\nabla\varphi_q \circ \nabla\varphi_p^\dagger(\boldsymbol{w}) = \nabla\varphi_q^\dagger \circ \nabla\varphi_p^\dagger(\boldsymbol{w}) = \frac{W}{\|\boldsymbol{w}\|_p} \cdot \boldsymbol{w}.$$

Proof For the first identity, applying Lemma 6 twice,

$$\begin{aligned}\nabla\varphi_q \circ \nabla\varphi_p^\dagger(\boldsymbol{w}) &= \frac{1}{\|\nabla\varphi_p^\dagger(\boldsymbol{w})\|_q^{q-2}} \cdot \operatorname{sign}(\nabla\varphi_p^\dagger(\boldsymbol{w})) \otimes |\nabla\varphi_p^\dagger(\boldsymbol{w})|^{q-1} \\ &= \frac{1}{W_q^{q-2}} \cdot \operatorname{sign}(\boldsymbol{w}) \otimes \frac{W^{q-1}}{\|\boldsymbol{w}\|_p^{(p-1)(q-1)}} \cdot |\boldsymbol{w}|^{(p-1)(q-1)} \\ &= \frac{W}{\|\boldsymbol{w}\|_p} \cdot \boldsymbol{w} \ . \end{aligned} \tag{18}$$

For the second identity, use Corollary 7 to conclude that

$$\begin{aligned}\nabla\varphi_q^\dagger \circ \nabla\varphi_p^\dagger(\boldsymbol{w}) &= \frac{W}{\|\nabla\varphi_p^\dagger(\boldsymbol{w})\|_q} \cdot \nabla\varphi_q(\nabla\varphi_p^\dagger(\boldsymbol{w})) \\ &= W \cdot \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|_p}.\end{aligned}$$

■

A.2 Bound on successive iterate divergence

The following Lemma extends [Kivinen et al., 2006, Appendix I] to $\varphi^\dagger$.

Lemma 10 *For any $\boldsymbol{w}$ and $\boldsymbol{\delta}$,*

$$\begin{aligned}&D_{\varphi_q^\dagger}\left(\boldsymbol{w} \| \nabla\varphi_p^\dagger\left(\nabla\varphi_q^\dagger(\boldsymbol{w}) + \boldsymbol{\delta}\right)\right) \\ &\leq \frac{(p-1)\|\boldsymbol{w}\|_q W}{2} \cdot \left\| \frac{1}{\|\nabla\varphi_q^\dagger(\boldsymbol{w}) + \boldsymbol{\delta}\|_p} \cdot \left(\nabla\varphi_q^\dagger(\boldsymbol{w}) + \boldsymbol{\delta}\right) - \frac{1}{W} \cdot \nabla\varphi_q^\dagger(\boldsymbol{w}) \right\|_p^2 .\end{aligned} \tag{19}$$

Proof [Proof of Lemma 10] In this proof, $\circ$ denotes composition and $\otimes$ is Hadamard product. The key step in the proof is the use of Theorem 1 to "branch" on the proof of [Kivinen et al., 2006, Appendix I] on the first following identity (letting $\varphi_q(\boldsymbol{w}) \doteq (1/2) \cdot (W^2 + \|\boldsymbol{w}\|_q^2)$). We also make use of the dual symmetry of Bregman divergences and we obtain third identity of:

$$\begin{aligned}&D_{\varphi_q^\dagger}\left(\boldsymbol{w} \| \nabla\varphi_p^\dagger\left(\nabla\varphi_q^\dagger(\boldsymbol{w}) + \boldsymbol{\delta}\right)\right) \\ &= \frac{\|\boldsymbol{w}\|_q}{W} \cdot D_{\varphi_q}\left(\frac{W}{\|\boldsymbol{w}\|_q} \cdot \boldsymbol{w} \left\| \frac{W}{\|\nabla\varphi_p^\dagger(\nabla\varphi_q^\dagger(\boldsymbol{w}) + \boldsymbol{\delta})\|_q} \cdot \nabla\varphi_p^\dagger\left(\nabla\varphi_q^\dagger(\boldsymbol{w}) + \boldsymbol{\delta}\right)\right.\right) \\ &= \frac{\|\boldsymbol{w}\|_q}{W} \cdot D_{\varphi_q}\left(\frac{W}{\|\boldsymbol{w}\|_q} \cdot \boldsymbol{w} \left\| \nabla\varphi_p^\dagger\left(\nabla\varphi_q^\dagger(\boldsymbol{w}) + \boldsymbol{\delta}\right)\right.\right) \qquad (20) \\ &= \frac{\|\boldsymbol{w}\|_q}{W} \cdot D_{\varphi_p}\left(\nabla\varphi_q \circ \nabla\varphi_p^\dagger\left(\nabla\varphi_q^\dagger(\boldsymbol{w}) + \boldsymbol{\delta}\right) \left\| \nabla\varphi_q\left(\frac{W}{\|\boldsymbol{w}\|_q} \cdot \boldsymbol{w}\right)\right.\right) \text{ by dual symmetry} \\ &= \frac{\|\boldsymbol{w}\|_q}{W} \cdot D_{\varphi_p}\left(\frac{W}{\|\nabla\varphi_q^\dagger(\boldsymbol{w}) + \boldsymbol{\delta}\|_p} \cdot \left(\nabla\varphi_q^\dagger(\boldsymbol{w}) + \boldsymbol{\delta}\right) \left\| \frac{W}{\|\boldsymbol{w}\|_q} \cdot \nabla\varphi_q(\boldsymbol{w})\right.\right) \qquad (21) \\ &= \frac{\|\boldsymbol{w}\|_q}{W} \cdot D_{\varphi_p}\left(\frac{W}{\|\nabla\varphi_q^\dagger(\boldsymbol{w}) + \boldsymbol{\delta}\|_p} \cdot \left(\nabla\varphi_q^\dagger(\boldsymbol{w}) + \boldsymbol{\delta}\right) \left\| \nabla\varphi_q^\dagger(\boldsymbol{w})\right.\right) \qquad (22) \\ &= \|\boldsymbol{w}\|_q W \cdot D_{\varphi_p}\left(\frac{1}{\|\nabla\varphi_q^\dagger(\boldsymbol{w}) + \boldsymbol{\delta}\|_p} \cdot \left(\nabla\varphi_q^\dagger(\boldsymbol{w}) + \boldsymbol{\delta}\right) \left\| \frac{1}{W} \cdot \nabla\varphi_q^\dagger(\boldsymbol{w})\right.\right) . \qquad (23)\end{aligned}$$

Equations (20) – (22) hold because of Corollary 7. We now use Appendix I[4] in Kivinen et al. [2006] on Equation (23) and obtain

$$
\begin{aligned}
& D_{\varphi_q^\dagger}\left(\boldsymbol{w} \| \nabla \varphi_p^\dagger\left(\nabla \varphi_q^\dagger(\boldsymbol{w})+\boldsymbol{\delta}\right)\right) \\
& \quad \leq \quad \frac{(p-1)\|\boldsymbol{w}\|_q W}{2} \cdot \left\| \frac{1}{\|\nabla \varphi_q^\dagger(\boldsymbol{w})+\boldsymbol{\delta}\|_p} \cdot \left(\nabla \varphi_q^\dagger(\boldsymbol{w})+\boldsymbol{\delta}\right) - \frac{1}{W} \cdot \nabla \varphi_q^\dagger(\boldsymbol{w}) \right\|_p^2 ,
\end{aligned}
$$

as claimed. ■

A.3 Bound on successive iterate divergence to target

In what follows, we write the DN-pLMS updates as $\boldsymbol{w}_t = \nabla \varphi_p^\dagger(\boldsymbol{\theta}_t)$, where

$$\boldsymbol{\theta}_t \doteq \nabla \varphi_q^\dagger(\boldsymbol{w}_{t-1}) - \Delta_t$$

for $\Delta_t = \eta_t \cdot (\boldsymbol{w}_{t-1}^\top \boldsymbol{x}_t - y_t) \cdot \boldsymbol{x}_t$. Further, for notational ease, we write

$$\bar{\boldsymbol{u}} \doteq \frac{\boldsymbol{u}}{g_q(\boldsymbol{u})}$$

and

$$\bar{\boldsymbol{\theta}}_t \doteq \frac{\boldsymbol{\theta}_t}{\|\boldsymbol{\theta}_t\|_p}.$$

We have the following preliminary bound on the distance from iterates of DN-pLMS to the (normalised) target.

Lemma 11 *Fix any learning rate sequence $\{\eta_t\}_{t=1}^T$. Pick any $\mathbf{u}$, and consider iterates $\{\boldsymbol{w}_t\}_{t=0}^T$ as per Equation 10. Denote $s_t \doteq (\bar{\boldsymbol{u}} - \boldsymbol{w}_{t-1})^\top \boldsymbol{x}_t, r_t \doteq \bar{\boldsymbol{u}}^\top \boldsymbol{x}_t - y_t$, and $\alpha_t \doteq \frac{W}{\|\boldsymbol{\theta}_t\|_p}$. Suppose $\|\boldsymbol{x}_t\|_p \leq X_p$. Then,*

$$D_{\varphi_q}(\bar{\boldsymbol{u}} \,\|\boldsymbol{w}_{t-1}) - D_{\varphi_q}(\bar{\boldsymbol{u}} \,\|\boldsymbol{w}_t) \quad \geq \quad Q + R + S + T \ ,$$

with

$$
\begin{aligned}
Q \quad &\doteq \quad \frac{\alpha_t}{2} \eta_t (s_t^2 - r_t^2), \\
R \quad &\doteq \quad (1-\alpha_t) \cdot \underbrace{\left(W^2 - \bar{\boldsymbol{u}}^\top \nabla \varphi_q^\dagger(\boldsymbol{w}_{t-1})\right)}_{\in [0, 2W^2]} , \\
S \quad &\doteq \quad \frac{p-1}{2} \cdot \underbrace{\left(2\alpha_t^2 \eta_t^2 (s_t - r_t)^2 X_p^2 - \left\|(s_t - r_t)\eta_t \alpha_t \cdot \boldsymbol{x}_t - (1-\alpha_t) \cdot \nabla \varphi_q^\dagger(\boldsymbol{w}_{t-1})\right\|_p^2\right)}_{\geq -2(1-\alpha_t)^2 W^2} , \\
T \quad &\doteq \quad \frac{\alpha_t}{2} \eta_t (s_t - r_t)^2 \left(1 - 2(p-1)\eta_t \alpha_t X_p^2\right) \ .
\end{aligned}
$$

Proof [Proof of Lemma 11] The Bregman triangle equality (also called the three points property) [Boissonnat et al., 2010, Property 5], [Cesa-Bianchi and Lugosi, 2006, Lemma 11.1] brings:

$$
\begin{aligned}
& D_{\varphi_q}(\bar{\boldsymbol{u}} \,\|\boldsymbol{w}_{t-1}) - D_{\varphi_q}(\bar{\boldsymbol{u}} \,\|\boldsymbol{w}_t) \\
& \quad = \quad (\bar{\boldsymbol{u}} - \boldsymbol{w}_{t-1})^\top \left(\nabla \varphi_q(\boldsymbol{w}_t) - \nabla \varphi_q(\boldsymbol{w}_{t-1})\right) - D_{\varphi_q}(\boldsymbol{w}_{t-1} \,\|\boldsymbol{w}_t) \\
& \quad = \quad (\bar{\boldsymbol{u}} - \boldsymbol{w}_{t-1})^\top \left(\nabla \varphi_q^\dagger(\boldsymbol{w}_t) - \nabla \varphi_q^\dagger(\boldsymbol{w}_{t-1})\right) - D_{\varphi_q^\dagger}(\boldsymbol{w}_{t-1} \,\|\boldsymbol{w}_t) \ \text{ by Lemmas 3, 8} \quad .
\end{aligned}
$$

[4] This result is stated as a bound on $D_{\varphi_q}(\boldsymbol{w} \| (\nabla \varphi_q)^{-1}(\nabla \varphi_q(\boldsymbol{w}) + \boldsymbol{\delta}))$, which by the Bregman dual symmetry property is equivalent to a bound on $D_{\varphi_p}(\nabla \varphi_q(\boldsymbol{w}) + \boldsymbol{\delta} \| \nabla \varphi_q(\boldsymbol{w}))$.

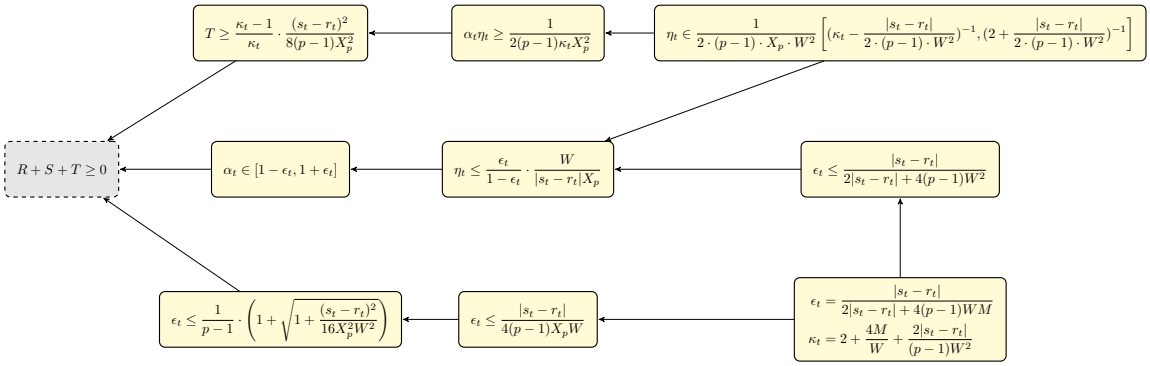

Figure 2: Schematic of proof of Lemma 12. Arrows from equation A to B indicate that $A \implies B$.

We now have

$$\nabla \varphi_q^\dagger (\boldsymbol{w}_t) = \nabla \varphi_q^\dagger \circ \nabla \varphi_p^\dagger (\boldsymbol{\theta}_t) = W \cdot \bar{\boldsymbol{\theta}}_t$$

by Corollary 7. We get

$$\begin{aligned}
& D_{\varphi_q} (\bar{\boldsymbol{u}} \,\|\boldsymbol{w}_{t-1}) - D_{\varphi_q} (\bar{\boldsymbol{u}} \,\|\boldsymbol{w}_t) \\
&\geq \; (\bar{\boldsymbol{u}} - \boldsymbol{w}_{t-1})^\top \left(W \cdot \bar{\boldsymbol{\theta}}_t - \nabla \varphi_q^\dagger (\boldsymbol{w}_{t-1}) \right) - \frac{(p-1)W^2}{2} \cdot \left\| \bar{\boldsymbol{\theta}}_t - \frac{1}{W} \cdot \nabla \varphi_q^\dagger (\boldsymbol{w}_{t-1}) \right\|_p^2 \\
&= \; (\bar{\boldsymbol{u}} - \boldsymbol{w}_{t-1})^\top \left(W \cdot \bar{\boldsymbol{\theta}}_t - \nabla \varphi_q^\dagger (\boldsymbol{w}_{t-1}) \right) - \frac{p-1}{2} \cdot \left\| W \cdot \bar{\boldsymbol{\theta}}_t - \nabla \varphi_q^\dagger (\boldsymbol{w}_{t-1}) \right\|_p^2 .
\end{aligned}$$

Now, note that

$$\boldsymbol{\theta}_t = \nabla \varphi_q^\dagger (\boldsymbol{w}_{t-1}) + \eta_t \cdot (s_t - r_t) \cdot \boldsymbol{x}_t .$$

We can thus rewrite the above as

$$\begin{aligned}
& D_{\varphi_q} (\bar{\boldsymbol{u}} \,\|\boldsymbol{w}_{t-1}) - D_{\varphi_q} (\bar{\boldsymbol{u}} \,\|\boldsymbol{w}_t) \\
&\geq \; s_t (s_t - r_t) \eta_t \alpha_t + (1 - \alpha_t) \left(\boldsymbol{w}_{t-1}^\top \nabla \varphi_q^\dagger (\boldsymbol{w}_{t-1}) - \bar{\boldsymbol{u}}^\top \nabla \varphi_q^\dagger (\boldsymbol{w}_{t-1}) \right) \\
&\quad - \frac{p-1}{2} \cdot \left\| (s_t - r_t) \eta_t \alpha_t \cdot \boldsymbol{x}_t - (1 - \alpha_t) \cdot \nabla \varphi_q^\dagger (\boldsymbol{w}_{t-1}) \right\|_p^2 \\
&= \; s_t (s_t - r_t) \eta_t \alpha_t + (1 - \alpha_t) \left(W^2 - \bar{\boldsymbol{u}}^\top \nabla \varphi_q^\dagger (\boldsymbol{w}_{t-1}) \right) \\
&\quad - \frac{p-1}{2} \cdot \left\| (s_t - r_t) \eta_t \alpha_t \cdot \boldsymbol{x}_t - (1 - \alpha_t) \cdot \nabla \varphi_q^\dagger (\boldsymbol{w}_{t-1}) \right\|_p^2 \text{ by definition of } \nabla \varphi_q^\dagger \\
&= \; Q + R + S + T \; .
\end{aligned}$$

■

We can show that the sum $R + S + T \geq 0$. This proof involves chaining together multiple simple inequalities. We give a high level overview in Figure 2.

Lemma 12 *Let R, S, T be as per Lemma 11. Suppose we fix*

$$\eta_t \;=\; \gamma \cdot \frac{W}{4(p-1)MX_pW + |y_t - \boldsymbol{w}_{t-1}^\top \boldsymbol{x}_t| X_p} \;, \tag{24}$$

for any $\gamma \in [1/2, 1]$, and $M \doteq \max\{W, X_p\}$. Then, $T + R + S \geq 0$.

Proof The triangle inequality and the fact that $\|\nabla\varphi_q^\dagger(\boldsymbol{w}_{t-1})\|_p = W$ brings

$$\begin{aligned}\alpha_t \quad &\in \quad \left[\frac{W}{\|\nabla\varphi_q^\dagger(\boldsymbol{w}_{t-1})\|_p + \eta_t|s_t - r_t| \cdot \|\boldsymbol{x}_t\|_p}, \frac{W}{\|\nabla\varphi_q^\dagger(\boldsymbol{w}_{t-1})\|_p - \eta_t|s_t - r_t| \cdot \|\boldsymbol{x}_t\|_p}\right] \\ &\subseteq \left[\frac{W}{W + \eta_t|s_t - r_t|X_p}, \frac{W}{W - \eta_t|s_t - r_t| \cdot X_p}\right], \end{aligned} \tag{25}$$

assuming that η_t is chosen so that

$$\eta_t \quad \leq \quad \frac{W}{|s_t - r_t| \cdot X_p} \ , \tag{26}$$

so that the right bound is non negative. To indeed ensure this, suppose that for some $0 < \epsilon_t \leq 1/2$, we fix

$$\eta_t \quad \leq \quad \frac{\epsilon_t}{1 - \epsilon_t} \cdot \frac{W}{|s_t - r_t|X_p} \ . \tag{27}$$

We would in addition obtain from Equation (25) that $\alpha_t \in [1 - \epsilon_t, 1 + \epsilon_t]$. Suppose η_t is also fixed to ensure

$$\eta_t \quad \in \quad \left[\frac{W}{2(p-1)\kappa_t X_p W^2 - |s_t - r_t|X_p}, \frac{W}{4(p-1)X_p W^2 + |s_t - r_t|X_p}\right] \ , \tag{28}$$

for some κ_t such that

$$\kappa_t \quad \geq \quad 2 + \frac{|s_t - r_t|}{(p-1)W^2} \ . \tag{29}$$

Notice that constraint on κ_t makes the interval non empty and its left bound strictly positive. Assuming (28) holds, we would have

$$\alpha_t\eta_t \quad \in \quad \left[\frac{1}{2(p-1)\kappa_t X_p^2}, \frac{1}{4(p-1)X_p^2}\right] \ . \tag{30}$$

The left bound of (30) holds because

$$\begin{aligned}\alpha_t\eta_t \quad &\geq \quad \eta_t \cdot \frac{W}{W + \eta_t|s_t - r_t|X_p} \\ &\geq \quad \frac{1}{2(p-1)\kappa_t X_p^2} \ . \end{aligned} \tag{31}$$

The first inequality holds because of (25) and the second one holds because of (28). The right bound of (30) holds because of (25), and so

$$\begin{aligned}\alpha_t\eta_t \quad &\leq \quad \eta_t \cdot \frac{W}{W - \eta_t|s_t - r_t|X_p} \\ &\leq \quad \frac{1}{4(p-1)X_p^2} \ , \end{aligned} \tag{32}$$

where the last inequality is due to (28).

Equation (30) makes that $T(\eta_t\alpha_t)$ is at least its value when $\alpha_t\eta_t$ attains the lower bound of (31), that is,

$$T(\eta_t\alpha_t) \quad \geq \quad \frac{\kappa_t - 1}{\kappa_t} \cdot \frac{(s_t - r_t)^2}{8(p-1)X_p^2} \ . \tag{33}$$

Now, to guarantee $\alpha_t \in [1 - \epsilon_t, 1 + \epsilon_t]$, it is sufficient that the right-hand side of inequality (27) belongs to interval (28) *and* we pick η_t within the interval [left bound (28), right-hand side (27)]. To guarantee that the right-hand side of inequality (27) falls in interval (28), we need first,

$$\frac{W}{2(p-1)\kappa_t X_p W^2 - |s_t - r_t|X_p} \quad \leq \quad \frac{\epsilon_t}{1 - \epsilon_t} \cdot \frac{W}{|s_t - r_t|X_p} \ , \tag{34}$$

that is,

$$\kappa_t \;\geq\; \frac{1}{\epsilon_t}\cdot\frac{|s_t-r_t|}{2(p-1)W^2}\;. \tag{35}$$

To guarantee that the right-hand side of inequality (27) falls in interval (28) we need then

$$\frac{W}{4(p-1)X_pW^2+|s_t-r_t|X_p} \;\geq\; \frac{\epsilon_t}{1-\epsilon_t}\cdot\frac{W}{|s_t-r_t|X_p}\;, \tag{36}$$

that is,

$$\epsilon_t \;\leq\; \frac{|s_t-r_t|}{2|s_t-r_t|+4(p-1)W^2}\;. \tag{37}$$

To summarize, if we pick any strictly positive ϵ_t following inequality (37) (note $\epsilon_t<1$) and

$$\kappa_t \;\doteq\; 2+\frac{1}{\epsilon_t}\cdot\frac{|s_t-r_t|}{(p-1)W^2}\;, \tag{38}$$

then we shall have both $\alpha_t\in[1-\epsilon_t,1+\epsilon_t]$ and inequality (33) holds as well. In this case, we shall have

$$\begin{aligned} T+R+S \;&\geq\; \left(1-\frac{1}{2+\frac{1}{\epsilon_t}\cdot\frac{|s_t-r_t|}{(p-1)W^2}}\right)\cdot\frac{(s_t-r_t)^2}{8(p-1)X_p^2}-2\epsilon_tW^2-(p-1)\epsilon_t^2W^2\\ &\geq\; \left(1-\frac{1}{2}\right)\cdot\frac{(s_t-r_t)^2}{8(p-1)X_p^2}-2\epsilon_tW^2-(p-1)\epsilon_t^2W^2\\ &=\; \frac{(s_t-r_t)^2}{16(p-1)X_p^2}-2\epsilon_tW^2-(p-1)\epsilon_t^2W^2\;. \end{aligned} \tag{39}$$

To finish up, we want to solve for ϵ_t the right-hand side such that it is non negative, and we find that ϵ_t has to satisfy

$$\epsilon_t \;\leq\; \frac{1}{p-1}\cdot\left(1+\sqrt{1+\frac{(s_t-r_t)^2}{16X_p^2W^2}}\right)\;. \tag{40}$$

Since $\sqrt{1+x}\geq\sqrt{x}$, a sufficient condition is

$$\epsilon_t \;\leq\; \frac{|s_t-r_t|}{4(p-1)X_pW}\;. \tag{41}$$

To ensure this and inequality (37), it is sufficient that we fix

$$\epsilon_t \;\doteq\; \frac{|s_t-r_t|}{2|s_t-r_t|+4(p-1)WM}\;, \tag{42}$$

where $M\doteq\max\{W,X_p\}$. With this expression for ϵ_t, we get from (38),

$$\kappa_t \;\doteq\; 2+\frac{4M}{W}+\frac{2|s_t-r_t|}{(p-1)W^2}\;. \tag{43}$$

For these choices, Lemma 13 implies that the given η_t is feasible. ■

Lemma 13 *Suppose ϵ_t satisfies (42) and κ_t satisfies (43). Then, a sufficient condition for η_t to satisfy both (27) and (28) is*

$$\eta_t \;=\; \gamma\cdot\frac{W}{4(p-1)MX_pW+|y_t-\boldsymbol{w}_{t-1}^{\top}\boldsymbol{x}_t|X_p}\;,$$

for any $\gamma\in[1/2,1]$.

Proof [Proof of Lemma 13] Notice the range of values authorized for η_t:

$$\begin{aligned}
\eta_t &\in \left[\frac{W}{2(p-1)\kappa_t X_p W^2 - |s_t - r_t| X_p}, \frac{\epsilon_t}{1-\epsilon_t} \cdot \frac{W}{|s_t - r_t| X_p}\right] \\
&= \left[\frac{W}{2(p-1)\left(2 + \frac{4M}{W} + \frac{2|s_t - r_t|}{(p-1)W^2}\right) X_p W^2 - |s_t - r_t| X_p}, \frac{W}{4(p-1)MX_pW + |s_t - r_t| X_p}\right] \\
&= \left[\frac{W}{2(2(p-1)W^2 + 4M(p-1)W + 2|s_t - r_t|)X_p - |s_t - r_t| X_p}, \frac{W}{4(p-1)MX_pW + |s_t - r_t| X_p}\right] \\
&= \left[\frac{W}{4(p-1)X_pW^2 + 8(p-1)MX_pW + 3|s_t - r_t| X_p}, \frac{W}{4(p-1)MX_pW + |s_t - r_t| X_p}\right] \\
&\supset \left[\frac{W}{8(p-1)MX_pW + 2|s_t - r_t| X_p}, \frac{W}{4(p-1)MX_pW + |s_t - r_t| X_p}\right] . \qquad (44)
\end{aligned}$$

A sufficient condition for η_t to fall in interval (44) is

$$\eta_t = \gamma \cdot \frac{W}{4(p-1)MX_pW + |y_t - \boldsymbol{w}_{t-1}^\top \boldsymbol{x}_t| X_p} ,$$

for any $\gamma \in [1/2, 1]$. ■

Lemma 14 *Suppose we fix the learning rate as per (24). Pick any $\boldsymbol{u}$, and consider iterates $\{\boldsymbol{w}_t\}_{t=0}^T$ as per Equation 10. Suppose $\|\boldsymbol{x}_t\|_p \leq X_p$ and $|y_t| \leq Y, \forall t \leq T$. Then, for any t,*

$$D_{\varphi_q}(\bar{\boldsymbol{u}} \| \boldsymbol{w}_{t-1}) - D_{\varphi_q}(\bar{\boldsymbol{u}} \| \boldsymbol{w}_t) \geq \frac{1}{4(p-1)\left(2 + \frac{4M}{W} + \frac{2(Y + X_pW)}{(p-1)W^2}\right) X_p^2} \cdot (s_t^2 - r_t^2)$$

where $s_t \doteq (\bar{\boldsymbol{u}} - \boldsymbol{w}_{t-1})^\top \boldsymbol{x}_t$, $r_t \doteq \bar{\boldsymbol{u}}^\top \boldsymbol{x}_t - y_t$.

Proof [Proof of Lemma 14] We start from the bound of Lemma 11:

$$\begin{aligned}
D_{\varphi_q}(\bar{\boldsymbol{u}} \| \boldsymbol{w}_{t-1}) - D_{\varphi_q}(\bar{\boldsymbol{u}} \| \boldsymbol{w}_t) &\geq Q + R + S + T \\
&\geq Q \text{ by Lemma 12} \\
&= \frac{\alpha_t}{2} \eta_t (s_t^2 - r_t^2) \text{ by definition} \\
&\geq \frac{1}{4(p-1)\kappa_t X_p^2} \cdot (s_t^2 - r_t^2) \\
&\geq \frac{1}{4(p-1)\left(2 + \frac{4M}{W} + \frac{2 \max_t |y_t - \boldsymbol{w}_{t-1}^\top \boldsymbol{x}_t|}{(p-1)W^2}\right) X_p^2} \cdot (s_t^2 - r_t^2) \\
&\geq \frac{1}{4(p-1)\left(2 + \frac{4M}{W} + \frac{2(Y + X_pW)}{(p-1)W^2}\right) X_p^2} \cdot (s_t^2 - r_t^2) . \qquad (45)
\end{aligned}$$

The last constraint to check for this bound to be valid is our ϵ_t in (42) has to be $< 1/2$ from inequality (26), which trivially holds since $4(p-1)WM \geq 0$.

We conclude by noting Lemma 13 provides a feasible value of η_t. ■

A.4 Gauge normalisation

The following lemma about the gauge of $\boldsymbol{x}$ will be useful.

Lemma 15 *Let $g_q(\boldsymbol{x}) = \|\boldsymbol{x}\|_q / W$ for some $W > 0$. Then, for the iterates $\{\boldsymbol{w}_t\}$ as per Equation 9, $g_q(\boldsymbol{w}_t) = 1$.*

Proof We have

$$\begin{aligned} g_q(\boldsymbol{w}_t) &= \frac{\|\boldsymbol{w}_t\|_q}{W} \\ &= \frac{W}{W} \text{ by Lemma 3} \\ &= 1 \ , \forall t \geq 1 \ . \end{aligned} \tag{46}$$

■

B Proofs of results in main body

We present proofs of all results in the main body.

Proof [Proof of Theorem 1] Let $\mathrm{J} : \mathcal{X} \to \mathcal{X}_g$ denote the Jacobian of $h\colon \boldsymbol{x} \mapsto (1/g(\boldsymbol{x})) \cdot \boldsymbol{x}$. By an elementary calculation,

$$g(\boldsymbol{x}) \cdot \mathrm{J} = \mathrm{I}_d - (1/g(\boldsymbol{x})) \cdot \boldsymbol{x} \nabla g(\boldsymbol{x})^\top,$$

which by the chain rule brings the following expression for the gradient of $\varphi^\dagger(\boldsymbol{y}) = g(\boldsymbol{y}) \cdot (\varphi \circ h)(\boldsymbol{y})$:

$$\begin{aligned}
\nabla \varphi^\dagger(\boldsymbol{y}) &= \nabla g(\boldsymbol{y}) \cdot (\varphi \circ h)(\boldsymbol{y}) + g(\boldsymbol{y}) \cdot \nabla (\varphi \circ h)(\boldsymbol{y}) \\
&= \nabla g(\boldsymbol{y}) \cdot (\varphi \circ h)(\boldsymbol{y}) + g(\boldsymbol{y}) \cdot \mathrm{J}^\top \nabla \varphi(h(\boldsymbol{y})) \\
&= \nabla g(\boldsymbol{y}) \cdot (\varphi \circ h)(\boldsymbol{y}) + \nabla \varphi(h(\boldsymbol{y})) - (1/g(\boldsymbol{y})) \cdot \nabla g(\boldsymbol{y}) \boldsymbol{y}^\top \nabla \varphi(h(\boldsymbol{y})) \\
&= \nabla \varphi\left(\frac{1}{g(\boldsymbol{y})} \cdot \boldsymbol{y}\right) + \left(\varphi\left(\frac{1}{g(\boldsymbol{y})} \cdot \boldsymbol{y}\right) - \frac{1}{g(\boldsymbol{y})} \cdot \boldsymbol{y}^\top \nabla \varphi\left(\frac{1}{g(\boldsymbol{y})} \cdot \boldsymbol{y}\right)\right) \cdot \nabla g(\boldsymbol{y}) \ . \qquad (47)
\end{aligned}$$

For simplicity, let $\boldsymbol{u} = \boldsymbol{x}/g(\boldsymbol{x})$ and $\boldsymbol{v} = \boldsymbol{y}/g(\boldsymbol{y})$, so that $\varphi^\dagger(\boldsymbol{x}) = g(\boldsymbol{x}) \cdot \varphi(\boldsymbol{u})$ and $\varphi^\dagger(\boldsymbol{y}) = g(\boldsymbol{y}) \cdot \varphi(\boldsymbol{v})$. The above then reads

$$\nabla \varphi^\dagger(\boldsymbol{y}) = \nabla \varphi(\boldsymbol{v}) + \left(\varphi(\boldsymbol{v}) - \boldsymbol{v}^\top \nabla \varphi(\boldsymbol{v})\right) \cdot \nabla g(\boldsymbol{y}). \qquad (48)$$

Now, the LHS of Equation (3) is

$$\begin{aligned}
g(\boldsymbol{x}) \cdot D_\varphi\left(\frac{1}{g(\boldsymbol{x})} \cdot \boldsymbol{x} \,\middle\|\, \frac{1}{g(\boldsymbol{y})} \cdot \boldsymbol{y}\right) &= g(\boldsymbol{x}) \cdot D_\varphi(\boldsymbol{u} \| \boldsymbol{v}) \\
&= g(\boldsymbol{x}) \cdot \varphi(\boldsymbol{u}) - g(\boldsymbol{x}) \cdot \varphi(\boldsymbol{v}) - g(\boldsymbol{x}) \cdot \nabla \varphi(\boldsymbol{v})^\top (\boldsymbol{u} - \boldsymbol{v}) \\
&= \varphi^\dagger(\boldsymbol{x}) - g(\boldsymbol{x}) \cdot \varphi(\boldsymbol{v}) - \nabla \varphi(\boldsymbol{v})^\top (\boldsymbol{x} - g(\boldsymbol{x}) \cdot \boldsymbol{v}) \\
&= \varphi^\dagger(\boldsymbol{x}) - g(\boldsymbol{x}) \cdot (\varphi(\boldsymbol{v}) - \nabla \varphi(\boldsymbol{v})^\top \boldsymbol{v}) - \nabla \varphi(\boldsymbol{v})^\top \boldsymbol{x},
\end{aligned}$$

while the RHS is

$$\begin{aligned}
D_{\varphi^\dagger}\left(\boldsymbol{x} \,\middle\|\, \boldsymbol{y}\right) &= \varphi^\dagger(\boldsymbol{x}) - \varphi^\dagger(\boldsymbol{y}) - \nabla \varphi^\dagger(\boldsymbol{y})^\top (\boldsymbol{x} - \boldsymbol{y}) \\
&= \varphi^\dagger(\boldsymbol{x}) - g(\boldsymbol{y}) \cdot \varphi(\boldsymbol{v}) - \nabla \varphi(\boldsymbol{v})^\top (\boldsymbol{x} - \boldsymbol{y}) - \left(\varphi(\boldsymbol{v}) - \boldsymbol{v}^\top \nabla \varphi(\boldsymbol{v})\right) \cdot \nabla g(\boldsymbol{y})^\top (\boldsymbol{x} - \boldsymbol{y}).
\end{aligned}$$

Cancelling the common $\varphi^\dagger(\boldsymbol{x})$ and $\nabla \varphi(\boldsymbol{v})^\top \boldsymbol{y}$ terms, the difference $\Delta = \mathrm{RHS} - \mathrm{LHS}$ is

$$\begin{aligned}
\Delta &= g(\boldsymbol{x}) \cdot (\varphi(\boldsymbol{v}) - \nabla \varphi(\boldsymbol{v})^\top \boldsymbol{v}) - g(\boldsymbol{y}) \cdot \varphi(\boldsymbol{v}) + \nabla \varphi(\boldsymbol{v})^\top \boldsymbol{y} - \left(\varphi(\boldsymbol{v}) - \boldsymbol{v}^\top \nabla \varphi(\boldsymbol{v})\right) \cdot \nabla g(\boldsymbol{y})^\top (\boldsymbol{x} - \boldsymbol{y}) \\
&= g(\boldsymbol{x}) \cdot (\varphi(\boldsymbol{v}) - \nabla \varphi(\boldsymbol{v})^\top \boldsymbol{v}) - g(\boldsymbol{y}) \cdot \varphi(\boldsymbol{v}) + g(\boldsymbol{y}) \cdot \nabla \varphi(\boldsymbol{v})^\top \boldsymbol{v} - \left(\varphi(\boldsymbol{v}) - \boldsymbol{v}^\top \nabla \varphi(\boldsymbol{v})\right) \cdot \nabla g(\boldsymbol{y})^\top (\boldsymbol{x} - \boldsymbol{y}) \\
&= g(\boldsymbol{x}) \cdot (\varphi(\boldsymbol{v}) - \nabla \varphi(\boldsymbol{v})^\top \boldsymbol{v}) - g(\boldsymbol{y}) \cdot \left(\varphi(\boldsymbol{v}) - \nabla \varphi(\boldsymbol{v})^\top \boldsymbol{v}\right) - \left(\varphi(\boldsymbol{v}) - \boldsymbol{v}^\top \nabla \varphi(\boldsymbol{v})\right) \cdot \nabla g(\boldsymbol{y})^\top (\boldsymbol{x} - \boldsymbol{y}) \\
&= (\varphi(\boldsymbol{v}) - \nabla \varphi(\boldsymbol{v})^\top \boldsymbol{v}) \cdot (g(\boldsymbol{x}) - g(\boldsymbol{y}) - \nabla g(\boldsymbol{y})^\top (\boldsymbol{x} - \boldsymbol{y})) \\
&= (\varphi(\boldsymbol{v}) - \nabla \varphi(\boldsymbol{v})^\top \boldsymbol{v}) \cdot B_g(\boldsymbol{x} \| \boldsymbol{y}).
\end{aligned}$$

Thus, the identity holds, if and only if either $\varphi(\boldsymbol{v}) = \nabla \varphi(\boldsymbol{v})^\top \boldsymbol{v}$ for every $\boldsymbol{v} \in \mathcal{X}_g$, or $B_g(\boldsymbol{x} \| \boldsymbol{y}) = 0$. The latter is true if and only if g is affine from Equation 2. The result follows. ■

It is easy to check that Theorem 1 in fact holds for separable (matrix) trace divergences [Kulis et al., 2009] of the form

$$D_\varphi(\mathrm{X} \| \mathrm{Y}) \ \doteq \ \varphi(\mathrm{X}) - \varphi(\mathrm{Y}) - \mathrm{tr}\left(\nabla \varphi(\mathrm{Y})^\top (\mathrm{X} - \mathrm{Y})\right) \ , \qquad (49)$$

with $\varphi, g : \mathbf{S}(d) \to \mathbb{R}$ (for $\mathbf{S}(d)$ the set of symmetric real matrices), with φ convex. In this case, the restricted positive homogeneity property becomes

$$\varphi(\mathrm{U}) \ = \ \mathrm{tr}\left(\nabla \varphi(\mathrm{U})^\top \mathrm{U}\right) \ , \forall \mathrm{U} \in \mathcal{X}_g \ . \qquad (50)$$

Proof [Proof of Lemma 2] Note that by construction, $g(\boldsymbol{r}(\boldsymbol{x})) = \mathbb{P}(\mathsf{X}=\boldsymbol{x})/((1-\pi_C)\cdot\mathbb{P}(\mathsf{X}=\boldsymbol{x}|\mathsf{Y}=C))$, and so

$$\begin{aligned}\left(\frac{1}{g(\boldsymbol{r}(\boldsymbol{x}))}\cdot\boldsymbol{r}(\boldsymbol{x})\right)_c &= \frac{(1-\pi_C)\cdot\mathbb{P}(\mathsf{X}=\boldsymbol{x}|\mathsf{Y}=C)}{\mathbb{P}(\mathsf{X}=\boldsymbol{x})}\cdot\frac{\mathbb{P}(\mathsf{X}=\boldsymbol{x}|\mathsf{Y}=c)}{\mathbb{P}(\mathsf{X}=\boldsymbol{x}|\mathsf{Y}=C)}\\ &= \frac{(1-\pi_C)}{\pi_c}\cdot\frac{\pi_c\mathbb{P}(\mathsf{X}=\boldsymbol{x}|\mathsf{Y}=c)}{\mathbb{P}(\mathsf{X}=\boldsymbol{x})}\\ &= \eta(\boldsymbol{x})\ . \end{aligned} \tag{51}$$

Furthermore,

$$\begin{aligned}\mathbb{P}(\mathsf{X}=\boldsymbol{x}) &= \sum_{c=1}^{C}\pi_c\mathbb{P}(\mathsf{X}=\boldsymbol{x}|\mathsf{Y}=c)\\ &= (1-\pi_C)\cdot\left(\frac{\pi_C}{1-\pi_C}+\sum_{c<C}\frac{\pi_c}{1-\pi_C}\cdot\frac{\mathbb{P}(\mathsf{X}=\boldsymbol{x}|\mathsf{Y}=c)}{\mathbb{P}(\mathsf{X}=\boldsymbol{x}|\mathsf{Y}=C)}\right)\cdot\mathbb{P}(\mathsf{X}=\boldsymbol{x}|\mathsf{Y}=C)\\ &= (1-\pi_C)\cdot g(\boldsymbol{r}(\boldsymbol{x}))\cdot\mathbb{P}(\mathsf{X}=\boldsymbol{x}|\mathsf{Y}=C)\ . \end{aligned} \tag{52}$$

Now let

$$\hat{\boldsymbol{r}}(\boldsymbol{x}) = \frac{1}{\hat{\eta}_C(\boldsymbol{x})}\cdot\hat{\boldsymbol{\eta}}(\boldsymbol{x}).$$

It then comes

$$\begin{aligned}&\mathbb{E}_M[D_\varphi(\boldsymbol{\eta}(\mathsf{X})\|\hat{\boldsymbol{\eta}}(\mathsf{X}))]\\ &= (1-\pi_C)\cdot\mathbb{E}_{P_C}\left[g(\boldsymbol{r}(\boldsymbol{x}))\cdot D_\varphi(\boldsymbol{\eta}(\mathsf{X})\|\hat{\boldsymbol{\eta}}(\mathsf{X}))\right]\\ &= (1-\pi_C)\cdot\mathbb{E}_{P_C}\left[g(\boldsymbol{r}(\boldsymbol{x}))\cdot D_\varphi\left(\frac{1}{g(\boldsymbol{r}(\boldsymbol{x}))}\cdot\boldsymbol{r}(\boldsymbol{x})\middle\|\hat{\boldsymbol{\eta}}(\mathsf{X})\right)\right]\\ &= (1-\pi_C)\cdot\mathbb{E}_{P_C}\left[g(\boldsymbol{r}(\boldsymbol{x}))\cdot D_\varphi\left(\frac{1}{g(\boldsymbol{r}(\boldsymbol{x}))}\cdot\boldsymbol{r}(\boldsymbol{x})\middle\|\frac{1}{g(\hat{\boldsymbol{r}}(\mathsf{X}))}\cdot\hat{\boldsymbol{r}}(\mathsf{X})\right)\right]\\ &= (1-\pi_C)\cdot\mathbb{E}_{P_C}\left[D_{\varphi^\dagger}(\boldsymbol{r}(\mathsf{X})\|\hat{\boldsymbol{r}}(\mathsf{X}))\right]\ , \end{aligned}$$

as claimed. ■

Proof [Proof of Lemma 3] For any $\boldsymbol{x}$, $\|\nabla\varphi_p^\dagger(\boldsymbol{x})\|_q = W$ by Corollary 7. Since $\boldsymbol{w}_t = \nabla\varphi_p^\dagger(\boldsymbol{\theta}_{t-1})$ for suitable $\boldsymbol{\theta}_{t-1}$, the result follows. The result for $\|\nabla\varphi_q^\dagger(\boldsymbol{w}_t)\|_p$ follows similarly by Corollary 7.

Note that while $\|\boldsymbol{w}_t\|_q = \|\nabla\varphi(\boldsymbol{w}_t)\|_p$ for the standard p-LMS update [Kivinen et al., 2006, Appendix I], these norms may vary with each iteration i.e. $\boldsymbol{w}_t$ may not lie in the L_q ball. ■

Proof [Proof of Lemma 4] Similarly to the proof of Lemma 10, a key to the proof of Lemma 4 relies on branching on Kivinen et al. [2006] through the use of Theorem 1. We first note that $D_{\varphi_q^\dagger}(\boldsymbol{u}\|\boldsymbol{w}_0) = W\cdot\|\boldsymbol{u}\|_q$ since $\boldsymbol{w}_0 = \boldsymbol{0}$, and $D_{\varphi_q^\dagger}(\boldsymbol{u}\|\boldsymbol{w}_{T+1}) \geq 0$, and so

$$\begin{aligned}W\cdot\|\boldsymbol{u}\|_q &\geq D_{\varphi_q^\dagger}(\boldsymbol{u}\|\boldsymbol{w}_0) - D_{\varphi_q^\dagger}(\boldsymbol{u}\|\boldsymbol{w}_{T+1})\\ &= \sum_{t=1}^{T}\left\{D_{\varphi_q^\dagger}(\boldsymbol{u}\|\boldsymbol{w}_{t-1}) - D_{\varphi_q^\dagger}(\boldsymbol{u}\|\boldsymbol{w}_t)\right\}\ \text{by telescoping property}\\ &= g_q(\boldsymbol{u})\cdot\sum_{t=1}^{T}\left\{D_{\varphi_q}\left(\frac{\boldsymbol{u}}{g_q(\boldsymbol{u})}\middle\|\frac{\boldsymbol{w}_{t-1}}{g_q(\boldsymbol{w}_{t-1})}\right) - D_{\varphi_q}\left(\frac{\boldsymbol{u}}{g_q(\boldsymbol{u})}\middle\|\frac{\boldsymbol{w}_t}{g_q(\boldsymbol{w}_t)}\right)\right\}\ \text{by Theorem 1}\\ &= g_q(\boldsymbol{u})\cdot\sum_{t=1}^{T}\left\{D_{\varphi_q}\left(\frac{\boldsymbol{u}}{g_q(\boldsymbol{u})}\|\boldsymbol{w}_{t-1}\right) - D_{\varphi_q}\left(\frac{\boldsymbol{u}}{g_q(\boldsymbol{u})}\|\boldsymbol{w}_t\right)\right\}\ \text{by Lemma 15}\ . \end{aligned} \tag{53}$$

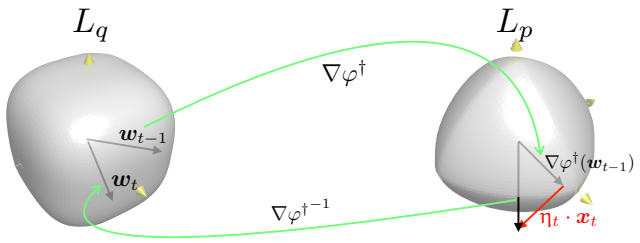

Figure 3: Illustration of the case $W = 1$ for the $\mathcal{B}_q(W)$-update: all classifiers and image via $\nabla\varphi^\dagger$ belong to a ball of radius 1 (here, $q = 3$, $p = 3/2$).

Recall from Lemma 14 that

$$D_{\varphi_q}\left(\frac{\boldsymbol{u}}{g_q(\boldsymbol{u})} \| \boldsymbol{w}_{t-1}\right) - D_{\varphi_q}\left(\frac{\boldsymbol{u}}{g_q(\boldsymbol{u})} \| \boldsymbol{w}_t\right) \geq \frac{1}{4(p-1)\left(2 + \frac{4M}{W} + \frac{2(Y+X_pW)}{(p-1)W^2}\right)X_p^2} \cdot (s_t^2 - r_t^2)$$

where

$$\begin{aligned} s_t &\doteq ((1/g_q(\boldsymbol{u})) \cdot \boldsymbol{u} - \boldsymbol{w}_{t-1})^\top \boldsymbol{x}_t \\ r_t &\doteq (1/g_q(\boldsymbol{u})) \cdot \boldsymbol{u}^\top \boldsymbol{x}_t - y_t. \end{aligned}$$

Note that $R_q(\boldsymbol{w}_{1:T}|\boldsymbol{u}) = \sum_{t=1}^T (s_t^2 - r_t^2)$ by definition. Summing the above for $t = 1, 2, ..., T$ and telescoping sums yields

$$\begin{aligned} R_q(\boldsymbol{w}_{1:T}|\boldsymbol{u}) &\leq 4(p-1)\left(2 + \frac{4M}{W} + \frac{2(Y+X_pW)}{(p-1)W^2}\right)X_p^2W^2 \\ &= 4(p-1)X_p^2W^2 + 16(p-1)MX_p^2W + 8(Y+X_pW)X_p^2 \\ &\leq 4(p-1)X_p^2W^2 + (16p-8)MX_p^2W + 8YX_p^2 \ . \end{aligned} \tag{54}$$

See Figure 3 for some geometric intuition about the updates. ■

Proof [Proof of Lemma 5] We start by the sphere. Let $\varphi(\boldsymbol{x}) \doteq (1/2) \cdot \|\boldsymbol{x}\|_2^2$. Since a Bregman divergence is invariant to linear transformation, it comes from Table 7 that

$$D_\varphi\left(\frac{\boldsymbol{x}^S}{g_S(\boldsymbol{x}^S)} \| \frac{\boldsymbol{c}^S}{g_S(\boldsymbol{c}^S)}\right) = \frac{1}{g_S(\boldsymbol{c}^S)} \cdot D_{\varphi^\dagger}(\boldsymbol{x}\|\boldsymbol{c}) = 1 - \cos D_G(\boldsymbol{x}, \boldsymbol{c}),$$

where we recall that D_G denotes the geodesic distance on the sphere (see Figure 1 and Appendix C). Equivalently,

$$\left\| \frac{1}{g_S(\boldsymbol{x}^S)} \cdot \boldsymbol{x}^S - \frac{1}{g_S(\boldsymbol{c}^S)} \cdot \boldsymbol{c}^S \right\|_2^2 = 1 - \cos D_G(\boldsymbol{x}, \boldsymbol{c}) \ . \tag{55}$$

This equality allows us to use k-means++ using the LHS of (55) to compute the distribution that picks a center. The key to using the approximation property of k-means++ relies on the existence of a coordinate system on the sphere for which the cluster centroid is just the average of the cluster points (polar coordinates), an average that eventually has to be rescaled if the coordinate system is not that one [Dhillon and Modha, 2001, Endo and Miyamoto, 2015]. The existence of this coordinate system makes that the proof of Arthur and Vassilvitskii [2007] (and in particular the key Lemmata 3.2 and 3.3) can be carried out without modification to yield the same approximation ratio as that of Arthur and Vassilvitskii [2007] *if* the distortion at hand is the squared Euclidean distance, which turns out to be $D_{\rm rec}(.:.)$ from eq. (55).
The case of the hyperboloid follows the exact same path, but starts from the fact that Table 7 now brings

$$D_\varphi\left(\frac{\boldsymbol{x}^H}{g_H(\boldsymbol{x}^H)} \| \frac{\boldsymbol{c}^H}{g_H(\boldsymbol{c}^H)}\right) = \cosh D_G(\boldsymbol{y}, \boldsymbol{c}) - 1 = \left\| \frac{1}{g_H(\boldsymbol{x}^H)} \cdot \boldsymbol{x}^H - \frac{1}{g_H(\boldsymbol{c}^H)} \cdot \boldsymbol{c}^H \right\|_2^2 \ .$$

To finish, in the same way as for the Sphere, we just need the existence of a coordinate system for which the centroid is an average of the cluster points, which can be obtained from hyperbolic barycentric coordinates [Ungar, 2014, Section 18]. ■

C Working out examples of Table 7

We fill in the details justifying each of the examples of Equation 3 provided in Table 2. We also provide the form of the corresponding divergences D_φ and distortions $D_{\varphi^\dagger}$ in the augmented Table 7.

	φ	$D_\varphi\left(\boldsymbol{x}\|\boldsymbol{y}\right)$	g	$\varphi^\dagger$	$D_{\varphi^\dagger}\left(\boldsymbol{x}\|\boldsymbol{y}\right)$
I	$\frac{1}{2}\cdot(1+\|\boldsymbol{x}\|_2^2)$	$(1/2)\cdot\|\boldsymbol{x}-\boldsymbol{y}\|_2^2$	$\|\boldsymbol{x}\|_2$	$\|\boldsymbol{x}\|_2$	$\|\boldsymbol{x}\|_2\cdot(1-\cos\angle\boldsymbol{x},\boldsymbol{y})$
II	$\frac{1}{2}\cdot(W+\|\boldsymbol{x}\|_q^2)$	$(1/2)\cdot(\|\boldsymbol{x}\|_q^2-\|\boldsymbol{y}\|_q^2)$ $-\sum_i\frac{(x_i-y_i)\cdot\mathrm{sign}(y_i)\cdot\|y_i\|^{q-1}}{\|\boldsymbol{y}\|_q^{q-2}}$	$\frac{\|\boldsymbol{x}\|_q}{W}$	$W\cdot\|\boldsymbol{x}\|_q$	$W\cdot\|\boldsymbol{x}\|_q-W\cdot\sum_i\frac{x_i\cdot\mathrm{sign}(y_i)\cdot\|y_i\|^{q-1}}{\|\boldsymbol{y}\|_q^{q-1}}$
III	$\frac{1}{2}\cdot(u^2+\|\boldsymbol{x}^S\|_2^2)$	$(1/2)\cdot\|\boldsymbol{x}^S-\boldsymbol{y}^S\|_2^2$	$\frac{\|\boldsymbol{x}\|_2}{\sin\|\boldsymbol{x}\|_2}$	$\|\boldsymbol{x}^S\|_2$	$\frac{\|\boldsymbol{x}\|_2}{\sin\|\boldsymbol{x}\|_2}\cdot(1-\cos D_G(\boldsymbol{x},\boldsymbol{y}))$
IV	$\frac{1}{2}\cdot(u^2+\|\boldsymbol{x}^H\|_2^2)$	$(1/2)\cdot\|\boldsymbol{x}^H-\boldsymbol{y}^H\|_2^2$	$-\frac{\|\boldsymbol{x}\|_2}{\sinh\|\boldsymbol{x}\|_2}$	$\|\boldsymbol{x}^H\|_2$	$-\frac{\|\boldsymbol{x}\|_2}{\sinh\|\boldsymbol{x}\|_2}\cdot(\cosh D_G(\boldsymbol{x},\boldsymbol{y})-1)$
V	$\sum_i x_i\log x_i-x_i$	$\sum_i x_i\log\frac{x_i}{y_i}$ $-\boldsymbol{1}^\top(\boldsymbol{x}-\boldsymbol{y})$	$\boldsymbol{1}^\top\boldsymbol{x}$	$\sum_i x_i\log x_i-\boldsymbol{1}^\top\boldsymbol{x}$ $-(\boldsymbol{1}^\top\boldsymbol{x})\log(\boldsymbol{1}^\top\boldsymbol{x})$	$\sum_i x_i\log\frac{x_i}{y_i}$ $-d\cdot\mathbb{E}[\mathsf{X}]\cdot\log\frac{\mathbb{E}[\mathsf{X}]}{\mathbb{E}[\mathsf{Y}]}$
VI	$-d-\sum_i\log x_i$	$\sum_i\frac{x_i}{y_i}$ $-\sum_i\log\frac{x_i}{y_i}-d$	$\prod_i x_i^{1/d}$	$-d\cdot\prod_i x_i^{1/d}$	$\sum_i\frac{x_i(\pi_{\boldsymbol{y}})^{1/d}}{y_i}-d(\pi_{\boldsymbol{x}})^{1/d}$
VII	$\mathrm{tr}\left(\mathsf{X}\log\mathsf{X}-\mathsf{X}\right)$	$\mathrm{tr}\left(\mathsf{X}\log\mathsf{X}-\mathsf{X}\log\mathsf{Y}\right)$ $-\mathrm{tr}\left(\mathsf{X}\right)+\mathrm{tr}\left(\mathsf{Y}\right)$	$\mathrm{tr}\left(\mathsf{X}\right)$	$\mathrm{tr}\left(\mathsf{X}\log\mathsf{X}-\mathsf{X}\right)$ $-\mathrm{tr}\left(\mathsf{X}\right)\log\mathrm{tr}\left(\mathsf{X}\right)$	$\mathrm{tr}\left(\mathsf{X}\log\mathsf{X}-\mathsf{X}\log\mathsf{Y}\right)$ $-\mathrm{tr}\left(\mathsf{X}\right)\cdot\log\frac{\mathrm{tr}(\mathsf{X})}{\mathrm{tr}(\mathsf{Y})}$
VIII	$-d-\log\det(\mathsf{X})$	$\mathrm{tr}\left(\mathsf{X}\mathsf{Y}^{-1}\right)$ $-\log\det(\mathsf{X}\mathsf{Y}^{-1})-d$	$\det(\mathsf{X}^{1/d})$	$-d\cdot\det(\mathsf{X}^{1/d})$	$\det(\mathsf{Y}^{1/d})\mathrm{tr}\left(\mathsf{X}\mathsf{Y}^{-1}\right)-d\cdot\det(\mathsf{X}^{1/d})$

Table 7: Example of distortions (right columns) that can be "reverse engineered" as Bregman divergences involving a particular, non necessary linear g. Function $\boldsymbol{x}^S\doteq f(\boldsymbol{x}):\mathbb{R}^d\to\mathbb{R}^{d+1}$ is the (S)phere lifting map defined in (67), and $\boldsymbol{x}^H$ is the (H)yperboloid lifting map defined in (78). $D_G(.,.)$ is the geodesic distance between the exponential map of $\boldsymbol{x}$ and $\boldsymbol{y}$ on their respective manifold (sphere or hyperboloid). Related proofs are in Section C. Expectation $\mathbb{E}[\mathsf{X}]$ is a shorthand for $(1/d)\cdot\sum_i x_i$. $W\in\mathbb{R}_{+*}$ and $u\in\mathbb{R}$ are constants.

Row I — for $\mathcal{X}=\mathbb{R}^d$, consider $\varphi(\boldsymbol{x})=(1+\|\boldsymbol{x}\|_2^2)/2$ and $g(\boldsymbol{x})=\|\boldsymbol{x}\|_2$ (we project on the Euclidean sphere). It comes

$$\varphi^\dagger(\boldsymbol{x}) \quad = \quad \|\boldsymbol{x}\|_2\cdot\left(\frac{1+\left\|\frac{1}{\|\boldsymbol{x}\|_2}\cdot\boldsymbol{x}\right\|_2^2}{2}\right)=\|\boldsymbol{x}\|_2\ . \tag{56}$$

g is not linear (but it is homogeneous of degree 1), but we have

$$\varphi(\boldsymbol{x})=1=\boldsymbol{x}^\top\nabla\varphi(\boldsymbol{x})\ ,\forall\boldsymbol{x}:\|\boldsymbol{x}\|_2=1\ , \tag{57}$$

so φ is 1-homogeneous on the Euclidean sphere, and we can apply Theorem 1. We have

$$\begin{aligned} g(\boldsymbol{x})\cdot D_\varphi\left(\frac{1}{g(\boldsymbol{x})}\cdot\boldsymbol{x}\|\frac{1}{g(\boldsymbol{y})}\cdot\boldsymbol{y}\right) &= \frac{\|\boldsymbol{x}\|_2}{2}\cdot\left\|\frac{1}{\|\boldsymbol{x}\|_2}\cdot\boldsymbol{x}-\frac{1}{\|\boldsymbol{y}\|_2}\cdot\boldsymbol{y}\right\|_2^2 \\ &= \|\boldsymbol{x}\|_2\cdot\left(1-\frac{\boldsymbol{x}^\top\boldsymbol{y}}{\|\boldsymbol{x}\|_2\|\boldsymbol{y}\|_2}\right)=\|\boldsymbol{x}\|_2\cdot(1-\cos(\boldsymbol{x},\boldsymbol{y}))\ , \end{aligned} \tag{58}$$

and we also have

$$\begin{aligned} D_{\varphi^\dagger}\left(\boldsymbol{x}\|\boldsymbol{y}\right) &= \|\boldsymbol{x}\|_2-\|\boldsymbol{y}\|_2-\frac{1}{\|\boldsymbol{y}\|_2}\cdot(\boldsymbol{x}-\boldsymbol{y})^\top\boldsymbol{y} \\ &= \|\boldsymbol{x}\|_2-\|\boldsymbol{y}\|_2-\frac{\boldsymbol{x}^\top\boldsymbol{y}}{\|\boldsymbol{y}\|_2}+\|\boldsymbol{y}\|_2 \qquad (59) \\ &= \|\boldsymbol{x}\|_2\cdot\left(1-\frac{\boldsymbol{x}^\top\boldsymbol{y}}{\|\boldsymbol{x}\|_2\|\boldsymbol{y}\|_2}\right)=\|\boldsymbol{x}\|_2\cdot(1-\cos(\boldsymbol{x},\boldsymbol{y}))\ , \qquad (60) \end{aligned}$$

which is equal to Equation (58), so we check that Theorem 1 applies in this case. $D_{\varphi^\dagger}$ has some properties. One is a weak form of triangle inequality.

Lemma 16 *$D_{\varphi^\dagger}(\boldsymbol{x}\|\boldsymbol{y}) + D_{\varphi^\dagger}(\boldsymbol{y}\|\boldsymbol{z}) \leq D_{\varphi^\dagger}(\boldsymbol{x}\|\boldsymbol{z})$, $\forall \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}$ such that $\|\boldsymbol{y}\|_2 \leq \|\boldsymbol{x}\|_2$.*

Proof

$$\begin{aligned}
& D_{\varphi^\dagger}(\boldsymbol{x}\|\boldsymbol{y}) + D_{\varphi^\dagger}(\boldsymbol{y}\|\boldsymbol{z}) \\
&= \|\boldsymbol{x}\|_2 \cdot (1 - \cos(\boldsymbol{x}, \boldsymbol{y})) + \|\boldsymbol{y}\|_2 \cdot (1 - \cos(\boldsymbol{y}, \boldsymbol{z})) \\
&= \|\boldsymbol{x}\|_2 \cdot ((1 - \cos(\boldsymbol{x}, \boldsymbol{y})) + (1 - \cos(\boldsymbol{y}, \boldsymbol{z}))) + (\|\boldsymbol{y}\|_2 - \|\boldsymbol{x}\|_2) \cdot (1 - \cos(\boldsymbol{y}, \boldsymbol{z})) \\
&\leq \|\boldsymbol{x}\|_2 \cdot (1 - \cos(\boldsymbol{x}, \boldsymbol{z})) + (\|\boldsymbol{y}\|_2 - \|\boldsymbol{x}\|_2) \cdot (1 - \cos(\boldsymbol{y}, \boldsymbol{z})) \\
&\leq D_{\varphi^\dagger}(\boldsymbol{x}\|\boldsymbol{z}) + (\|\boldsymbol{y}\|_2 - \|\boldsymbol{x}\|_2) \cdot (1 - \cos(\boldsymbol{y}, \boldsymbol{z})) \\
&\leq D_{\varphi^\dagger}(\boldsymbol{x}\|\boldsymbol{z}) \ , \qquad (61)
\end{aligned}$$

since $\|\boldsymbol{y}\|_2 \leq \|\boldsymbol{x}\|_2$. We have used the fact that $(1 - \cos(\boldsymbol{x}, \boldsymbol{y}))$ is half the Euclidean distance between unit-normalized vectors. ■

It turns out that $D_{\varphi^\dagger}(\boldsymbol{x}\|\boldsymbol{\mu})$ can be related to the log-likelihood of a von Mises-Fisher distribution with expected direction $\boldsymbol{\mu}$, which is useful in text analysis [Reisinger et al., 2010].

Row II — Let $\varphi(\boldsymbol{x}) \doteq (1/2) \cdot (u^2 + \|\boldsymbol{x}\|_q^2)$, for $q > 1$ [Kivinen et al., 2006]. We have

$$\varphi\left(\frac{1}{g(\boldsymbol{x})} \cdot \boldsymbol{x}\right) = \frac{u^2}{2} + \frac{1}{2} \cdot \left\|\frac{1}{g(\boldsymbol{x})} \cdot \boldsymbol{x}\right\|_q^2 = \frac{u^2}{2} + \frac{1}{2g^2(\boldsymbol{x})} \cdot \|\boldsymbol{x}\|_q^2 \ . \qquad (62)$$

We also have

$$\begin{aligned}
\left(\frac{1}{g(\boldsymbol{x})} \cdot \boldsymbol{x}\right)^\top \nabla\varphi\left(\frac{1}{g(\boldsymbol{x})} \cdot \boldsymbol{x}\right) &= \frac{1}{g(\boldsymbol{x})} \cdot \sum_i \frac{x_i \cdot \operatorname{sign}\left(\frac{1}{g(\boldsymbol{x})} \cdot x_i\right) \left|\frac{1}{g(\boldsymbol{x})} \cdot x_i\right|^{q-1}}{\left\|\frac{1}{g(\boldsymbol{x})} \cdot \boldsymbol{x}\right\|_q^{q-2}} \\
&= \sum_i \frac{\left|\frac{1}{g(\boldsymbol{x})} \cdot x_i\right|^q}{\left\|\frac{1}{g(\boldsymbol{x})} \cdot \boldsymbol{x}\right\|_q^{q-2}} \\
&= \frac{1}{g^2(\boldsymbol{x})} \cdot \|\boldsymbol{x}\|_q^2 \ . \qquad (63)
\end{aligned}$$

To have the condition of Theorem 1 satisfied, we therefore need

$$\|\boldsymbol{x}\|_q = ug(\boldsymbol{x}) \ , \qquad (64)$$

So we use $g(\boldsymbol{x}) = \|\boldsymbol{x}\|_q / W$ and $u = W$, observing that φ is 1-homogeneous on the L_p sphere. We check that

$$\varphi^\dagger(\boldsymbol{x}) = W \cdot \|\boldsymbol{x}\|_q \ . \qquad (65)$$

and we obtain

$$D_{\varphi^\dagger}(\boldsymbol{w}\|\boldsymbol{w}') = W \cdot \|\boldsymbol{w}\|_q - W \cdot \sum_i \frac{w_i \cdot \operatorname{sign}(w_i') \cdot |w_i'|^{q-1}}{\|\boldsymbol{w}'\|_q^{q-1}} \ . \qquad (66)$$

Row III — As in Buss and Fillmore [2001], we assume $\|\boldsymbol{x}\|_2 \leq \pi$, or we renormalize or change the radius of the ball) We first lift the data points using the *Sphere* lifting map $\mathbb{R}^d \ni \boldsymbol{x} \mapsto \boldsymbol{x}^S \in \mathbb{R}^{d+1}$:

$$\boldsymbol{x}^S \doteq [x_1 \quad x_2 \quad \cdots \quad x_d \quad r_{\boldsymbol{x}} \cot r_{\boldsymbol{x}}]^\top \ , \qquad (67)$$

where $r_{\boldsymbol{x}} \doteq \|\boldsymbol{x}\|_2$ is the Euclidean norm of $\boldsymbol{x}$. Notice that the last coordinate is a coordinate of the Hessian of the geodesic distance to the origin on the sphere [Buss and Fillmore, 2001]. We then let $g(\boldsymbol{x}^S) \doteq r_{\boldsymbol{x}} / \sin r_{\boldsymbol{x}}$ (notice that g is computed uding the first d coordinates). Finally, for $\mathcal{X} = \mathbb{R}^{d+1}$ and $u > 1$, consider $\varphi(\boldsymbol{x}^S) = (u^2 + \|\boldsymbol{x}^S\|_2^2)/2$. The set of points for which $\varphi(\boldsymbol{x}^S) = (\boldsymbol{x}^S)^\top \nabla \varphi(\boldsymbol{x}^S)$ is equivalently the subset $\mathcal{X}_g \subseteq \mathbb{R}^{d+1}$ such that

$$\mathcal{X}_g \doteq \{\boldsymbol{x}^S : g^2(\boldsymbol{x}^S) = u^2\} \ . \tag{68}$$

So φ satisfies the restricted positive homogeneity of degree 1 on $\mathcal{X}_g$ and we can apply Theorem 1. We first remark that:

$$\begin{aligned} \|\boldsymbol{x}^S\|_2^2 &= r_{\boldsymbol{x}}^2 + r_{\boldsymbol{x}}^2 \cot^2 r_{\boldsymbol{x}} \\ &= \frac{r_{\boldsymbol{x}}^2}{\sin r_{\boldsymbol{x}}^2} = g^2(\boldsymbol{x}^S) \ , \end{aligned} \tag{69}$$

and

$$\varphi^\dagger(\boldsymbol{x}^S) = \frac{r_{\boldsymbol{x}}}{\sin r_{\boldsymbol{x}}} \cdot \varphi\left(\frac{\sin r_{\boldsymbol{x}}}{r_{\boldsymbol{x}}} \cdot \boldsymbol{x}^S\right) = \frac{r_{\boldsymbol{x}}}{\sin r_{\boldsymbol{x}}} \cdot \left(\frac{\sin r_{\boldsymbol{x}}}{r_{\boldsymbol{x}}}\right)^2 \cdot \|\boldsymbol{x}^S\|_2^2 = \|\boldsymbol{x}^S\|_2 \ , \tag{70}$$

and finally, because of the spherical law of cosines,

$$\sin r_{\boldsymbol{x}} \sin r_{\boldsymbol{y}} \cos(\boldsymbol{x}, \boldsymbol{y}) + \cos r_{\boldsymbol{x}} \cos r_{\boldsymbol{y}} = \cos D_G(\boldsymbol{x}, \boldsymbol{y}) \ , \tag{71}$$

where we recall from eq. (16) that $D_G(\boldsymbol{x}, \boldsymbol{y})$ is the geodesic distance between the image of the exponential maps of $\boldsymbol{x}$ and $\boldsymbol{y}$ on the sphere.

We then derive

$$\begin{aligned} & g(\boldsymbol{x}^S) \cdot D_\varphi\left(\frac{1}{g(\boldsymbol{x}^S)} \cdot \boldsymbol{x}^S \| \frac{1}{g(\boldsymbol{y}^S)} \cdot \boldsymbol{y}^S\right) \\ &= \frac{r_{\boldsymbol{x}}}{2 \sin r_{\boldsymbol{x}}} \cdot \left\| \frac{\sin r_{\boldsymbol{x}}}{r_{\boldsymbol{x}}} \cdot \boldsymbol{x}^S - \frac{\sin r_{\boldsymbol{y}}}{r_{\boldsymbol{y}}} \cdot \boldsymbol{y}^S \right\|_2^2 \\ &= \frac{r_{\boldsymbol{x}}}{2 \sin r_{\boldsymbol{x}}} \cdot \left(\frac{\sin^2 r_{\boldsymbol{x}}}{\|\boldsymbol{x}\|_2^2} \cdot \|\boldsymbol{x}^S\|_2^2 + \frac{\sin^2 r_{\boldsymbol{y}}}{\|\boldsymbol{y}\|_2^2} \cdot \|\boldsymbol{y}^S\|_2^2 - 2 \cdot \frac{\sin r_{\boldsymbol{x}}}{r_{\boldsymbol{x}}} \cdot \frac{\sin r_{\boldsymbol{y}}}{r_{\boldsymbol{y}}} \cdot (\boldsymbol{x}^S)^\top \boldsymbol{y}^S \right) \\ &= \frac{r_{\boldsymbol{x}}}{\sin r_{\boldsymbol{x}}} \cdot \left(1 - \frac{\sin r_{\boldsymbol{x}}}{r_{\boldsymbol{x}}} \cdot \frac{\sin r_{\boldsymbol{y}}}{r_{\boldsymbol{y}}} \cdot (\boldsymbol{x}^S)^\top \boldsymbol{y}^S \right) \qquad (72) \\ &= \frac{r_{\boldsymbol{x}}}{\sin r_{\boldsymbol{x}}} \cdot \left(1 - \frac{\sin r_{\boldsymbol{x}}}{r_{\boldsymbol{x}}} \cdot \frac{\sin r_{\boldsymbol{y}}}{r_{\boldsymbol{y}}} \cdot \left(\boldsymbol{x}^\top \boldsymbol{y} + r_{\boldsymbol{x}} r_{\boldsymbol{y}} \cot r_{\boldsymbol{x}} \cot r_{\boldsymbol{y}}\right) \right) \qquad (73) \\ &= \frac{r_{\boldsymbol{x}}}{\sin r_{\boldsymbol{x}}} \cdot (1 - \sin r_{\boldsymbol{x}} \sin r_{\boldsymbol{y}} \cdot (\cos(\boldsymbol{x}, \boldsymbol{y}) + \cot r_{\boldsymbol{x}} \cot r_{\boldsymbol{y}})) \\ &= \frac{r_{\boldsymbol{x}}}{\sin r_{\boldsymbol{x}}} \cdot (1 - (\sin r_{\boldsymbol{x}} \sin r_{\boldsymbol{y}} \cos(\boldsymbol{x}, \boldsymbol{y}) + \cos r_{\boldsymbol{x}} \cos r_{\boldsymbol{y}})) \\ &= \frac{r_{\boldsymbol{x}}}{\sin r_{\boldsymbol{x}}} \cdot (1 - \cos D_G(\boldsymbol{x}, \boldsymbol{y})) \ . \qquad (74) \end{aligned}$$

In Equation (72), we use Equation (69), and we use Equation (71) in Equation (74). We also check

$$\begin{aligned} D_{\varphi^\dagger}(\boldsymbol{x}^S \| \boldsymbol{y}^S) &= \|\boldsymbol{x}^S\|_2 - \|\boldsymbol{y}^S\|_2 - \frac{1}{\|\boldsymbol{y}^S\|_2} \cdot (\boldsymbol{x}^S - \boldsymbol{y}^S)^\top \boldsymbol{y}^S \\ &= \|\boldsymbol{x}^S\|_2 - \frac{1}{\|\boldsymbol{y}^S\|_2} \cdot (\boldsymbol{x}^S)^\top \boldsymbol{y}^S \\ &= \|\boldsymbol{x}^S\|_2 \cdot \left(1 - \frac{(\boldsymbol{x}^S)^\top \boldsymbol{y}^S}{\|\boldsymbol{x}^S\|_2 \|\boldsymbol{y}^S\|_2}\right) \qquad (75) \\ &= \frac{r_{\boldsymbol{x}}}{\sin r_{\boldsymbol{x}}} \cdot (1 - \cos D_G(\boldsymbol{x}, \boldsymbol{y})) \ . \qquad (76) \end{aligned}$$

To obtain (76), we use the fact that

$$\frac{(\boldsymbol{x}^S)^\top \boldsymbol{y}^S}{\|\boldsymbol{x}^S\|_2 \|\boldsymbol{y}^S\|_2} = \frac{\sin r_{\boldsymbol{x}}}{r_{\boldsymbol{x}}} \cdot \frac{\sin r_{\boldsymbol{y}}}{r_{\boldsymbol{y}}} \cdot \left(\boldsymbol{x}^\top \boldsymbol{y} + r_{\boldsymbol{x}} r_{\boldsymbol{y}} \cot r_{\boldsymbol{x}} \cot r_{\boldsymbol{y}}\right) \ , \tag{77}$$

and then plug it into Equation (76), which yields the identity between Equation (73) (and thus (74)) and (76). So Theorem 1 holds in this case as well.

We remark that $(1/g(\boldsymbol{x}^S)) \cdot \boldsymbol{x}^S = \exp_{\mathbf{0}}(\boldsymbol{x})$ is the exponential map for the sphere [Buss and Fillmore, 2001].

Row IV — In the same way as we did for row IV, we first create a lifting map, but this time *complex* valued, the Hyperboloid lifting map $H\colon \mathbb{R}^d \ni \boldsymbol{x} \mapsto \boldsymbol{x}^H \in \mathbb{R}^d \times \mathbb{C}$. With an abuse of notation, it is given by

$$\boldsymbol{x}^H \doteq [x_1 \quad x_2 \quad \cdots \quad x_d \quad i r_{\boldsymbol{x}} \coth r_{\boldsymbol{x}}]^\top \ , \tag{78}$$

and we let $g(\boldsymbol{x}^H) \doteq -r_{\boldsymbol{x}}/\sinh r_{\boldsymbol{x}}$, with $\coth$ and $\sinh$ defining respectively the hyperbolic cotangent and hyperbolic sine. We let $0 \coth 0 = 0/\sinh 0 = 1$. Notice that the complex number is pure imaginary and so H defines a d dimensional manifold that lives in $\mathbb{R}^{d+1}$ assuming that the last coordinate is the imaginary axis. Let $\exp_{\boldsymbol{q}}(\boldsymbol{x}) \doteq (1/g(\boldsymbol{x}^H) \cdot \boldsymbol{x}^H$. Notice that

$$\begin{aligned} \left\|\exp_{\boldsymbol{q}}(\boldsymbol{x})\right\|_2^2 &= \frac{\sinh^2 r_{\boldsymbol{x}}}{r_{\boldsymbol{x}}^2} \cdot \left(r_{\boldsymbol{x}}^2 + i^2 r_{\boldsymbol{x}}^2 \coth^2 r_{\boldsymbol{x}}\right) \\ &= \sinh^2 r_{\boldsymbol{x}} + i^2 \cosh^2 r_{\boldsymbol{x}} \\ &= \sinh^2 r_{\boldsymbol{x}} - \cosh^2 r_{\boldsymbol{x}} = -1 \ , \end{aligned} \tag{79}$$

so $\exp_{\boldsymbol{q}}(\boldsymbol{x})$ defines a lifting map from $\mathbb{R}^d$ to the hyperboloid model $\mathbb{H}^d$ of hyperbolic geomety Galperin [1993]. In fact, it defines the exponential map for the plane $T_{\boldsymbol{q}}\mathbb{H}^d$ tangent to $\mathbb{H}^d$ in point $\boldsymbol{q} \doteq [0 \quad 0 \quad \cdots \quad 0 \quad i] = \mathbf{0}^H$. To see this, remark that we can express the geodesic distance D_G with the hyperbolic metric between $\boldsymbol{x}^H$ and $\boldsymbol{y}^H$ as

$$D_G(\boldsymbol{x}^H, \boldsymbol{y}^H) \doteq \cosh^{-1}(-(\boldsymbol{x}^H)^\top \boldsymbol{y}^H) \ , \tag{80}$$

where $\cosh^{-1}$ is the inverse hyperbolic cosine. So, for any $\boldsymbol{x} \in T_{\boldsymbol{q}}\mathbb{H}^d$, since $r_{\boldsymbol{x}} = \|\boldsymbol{x} - \mathbf{0}\|_2$, we have

$$\begin{aligned} D_G(\exp_{\boldsymbol{q}}(\boldsymbol{x}), \boldsymbol{q}) &= \cosh^{-1}(-(\boldsymbol{x}^H)^\top \mathbf{0}^H) \\ &= \cosh^{-1}(-i^2 \cosh r_{\boldsymbol{x}}) \\ &= r_{\boldsymbol{x}} = \|\boldsymbol{x} - \mathbf{0}\|_2 \ , \end{aligned} \tag{81}$$

and $\exp_{\boldsymbol{q}}(\boldsymbol{x})$ is indeed the exponential map for $T_{\boldsymbol{q}}\mathbb{H}^d$. Now, remark that

$$\begin{aligned} \exp_{\boldsymbol{q}}(\boldsymbol{x})^\top \exp_{\boldsymbol{q}}(\boldsymbol{y}) &= \frac{\sinh r_{\boldsymbol{x}}}{r_{\boldsymbol{x}}} \cdot \frac{\sinh r_{\boldsymbol{y}}}{r_{\boldsymbol{y}}} \cdot (\boldsymbol{x}^H)^\top \boldsymbol{y}^H \\ &= \frac{\sinh r_{\boldsymbol{x}}}{r_{\boldsymbol{x}}} \cdot \frac{\sinh r_{\boldsymbol{y}}}{r_{\boldsymbol{y}}} \cdot \left(\boldsymbol{x}^\top \boldsymbol{y} + i^2 r_{\boldsymbol{x}} r_{\boldsymbol{y}} \coth r_{\boldsymbol{x}} \coth r_{\boldsymbol{y}}\right) \\ &= \sinh r_{\boldsymbol{x}} \sinh r_{\boldsymbol{y}} \cdot (\cos(\boldsymbol{x}, \boldsymbol{y}) - \coth r_{\boldsymbol{x}} \coth r_{\boldsymbol{y}}) \\ &= \sinh r_{\boldsymbol{x}} \sinh r_{\boldsymbol{y}} \cos(\boldsymbol{x}, \boldsymbol{y}) - \cosh r_{\boldsymbol{x}} \cosh r_{\boldsymbol{y}} \\ &= -\cosh D_G(\boldsymbol{x}^H, \boldsymbol{y}^H) \ . \end{aligned} \tag{82}$$

Eq. (82) holds by the hyperbolic law of cosines. Now, we let $\varphi(\boldsymbol{x}^H) = (u^2 + \|\boldsymbol{x}^H\|_2^2)/2$ and

$$\mathcal{X}_g \doteq \{\boldsymbol{x}^H : \|\boldsymbol{x}^H\|_2^2 = u^2\} \ . \tag{83}$$

We check that $\varphi(\boldsymbol{x}^H) = u^2 = (\boldsymbol{x}^H)^\top \nabla\varphi(\boldsymbol{x}^H)$ for any $\boldsymbol{x}^H \in \mathcal{X}_g$, so we can apply Theorem 1. We then use eqs. (79) and (82) and derive

$$\begin{aligned} & g(\boldsymbol{x}^H) \cdot D_\varphi\left(\frac{1}{g(\boldsymbol{x}^H)} \cdot \boldsymbol{x}^H \| \frac{1}{g(\boldsymbol{y}^H)} \cdot \boldsymbol{y}^H\right) \\ = & -\frac{r_{\boldsymbol{x}}}{2\sinh r_{\boldsymbol{x}}} \cdot \left\|\exp_{\boldsymbol{q}}(\boldsymbol{x}) - \exp_{\boldsymbol{q}}(\boldsymbol{y})\right\|_2^2 \\ = & -\frac{r_{\boldsymbol{x}}}{2\sinh r_{\boldsymbol{x}}} \cdot \left(\left\|\exp_{\boldsymbol{q}}(\boldsymbol{x})\right\|_2^2 + \left\|\exp_{\boldsymbol{q}}(\boldsymbol{y})\right\|_2^2 - 2\exp_{\boldsymbol{q}}(\boldsymbol{x})^\top \exp_{\boldsymbol{q}}(\boldsymbol{y})\right) \\ = & -\frac{r_{\boldsymbol{x}}}{\sinh r_{\boldsymbol{x}}} \cdot \left(\cosh D_G(\boldsymbol{x}^H, \boldsymbol{y}^H) - 1\right) \ . \end{aligned} \tag{84}$$

Note that eq. (84) is a negative-valued and concave distortion.

Row V — for $\mathcal{X} = \mathbb{R}^d_{+*}$, consider $\varphi(\boldsymbol{x}) = \sum_i x_i \log x_i - x_i$ and $g(\boldsymbol{x}) = \mathbf{1}^\top \boldsymbol{x}$ (we normalize on the simplex). Since g is linear, we do not need to check for the homogeneity of φ, and we directly obtain:

$$\begin{aligned} g(\boldsymbol{x}) \cdot D_\varphi\left(\frac{1}{g(\boldsymbol{x})} \cdot \boldsymbol{x} \| \frac{1}{g(\boldsymbol{y})} \cdot \boldsymbol{y}\right) = & \sum_i x_i \log x_i - (\mathbf{1}^\top \boldsymbol{x}) \log(\mathbf{1}^\top \boldsymbol{x}) - \mathbf{1}^\top \boldsymbol{x} \\ & -\frac{\mathbf{1}^\top \boldsymbol{x}}{\mathbf{1}^\top \boldsymbol{y}} \cdot \sum_i y_i \log y_i - (\mathbf{1}^\top \boldsymbol{x}) \log(\mathbf{1}^\top \boldsymbol{y}) + \mathbf{1}^\top \boldsymbol{x} \\ & -(\mathbf{1}^\top \boldsymbol{x}) \cdot \sum_i \left(\frac{x_i}{\mathbf{1}^\top \boldsymbol{x}} - \frac{y_i}{\mathbf{1}^\top \boldsymbol{y}}\right) \cdot \log \frac{y_i}{\mathbf{1}^\top \boldsymbol{y}} \\ = & \sum_i x_i \log \frac{x_i}{y_i} - (\mathbf{1}^\top \boldsymbol{x}) \cdot \log \frac{\mathbf{1}^\top \boldsymbol{x}}{\mathbf{1}^\top \boldsymbol{y}} \ . \end{aligned} \tag{85}$$

Furthermore,

$$\varphi^\dagger(\boldsymbol{x}) = \mathbf{1}^\top \boldsymbol{x} \cdot \left(\sum_i \frac{x_i}{\mathbf{1}^\top \boldsymbol{x}} \cdot \log \frac{x_i}{\mathbf{1}^\top \boldsymbol{x}} - 1\right) = \sum_i x_i \log x_i - (\mathbf{1}^\top \boldsymbol{x}) \log(\mathbf{1}^\top \boldsymbol{x}) - \mathbf{1}^\top \boldsymbol{x} \ . \tag{86}$$

Noting that $\varphi^\dagger(\boldsymbol{x})$ is the sum of three terms, one of which is linear and can be removed for the divergence, so the divergence is just the sum of the two divergences with the two generators, which is found to be Equation (85) as well. Remark that while the KL divergence is convex in its both arguments, $D_{\varphi^\dagger}(\boldsymbol{x}\|\boldsymbol{y})$ may not be (jointly) convex. Indeed, its Hessian in $\boldsymbol{y}$ equals:

$$\mathrm{H}_{\boldsymbol{y}}(D_{\varphi^\dagger}) = \mathrm{Diag}(\{x_i/y_i^2\}_i) - \frac{\mathbf{1}^\top \boldsymbol{x}}{(\mathbf{1}^\top \boldsymbol{y})^2} \cdot \mathbf{1}\mathbf{1}^\top \ , \tag{87}$$

which may be indefinite.

Row VI — for $\mathcal{X} = \mathbb{R}^d_{+*}$, consider $\varphi(\boldsymbol{x}) = -d - \sum_i \log x_i$ and $g(\boldsymbol{x}) = (\pi_{\boldsymbol{x}})^{1/d}$, where we let $\pi_{\boldsymbol{x}} \doteq \prod_i x_i$ (we normalize with the geometric average). It comes

$$\varphi^\dagger(\boldsymbol{x}) = (\pi_{\boldsymbol{x}})^{1/d} \cdot \left(-d - \sum_i \log \frac{x_i}{(\pi_{\boldsymbol{x}})^{1/d}}\right) = -d \cdot (\pi_{\boldsymbol{x}})^{1/d} \ . \tag{88}$$

g is not linear (but it is homogeneous of degree 1), and we have

$$\varphi(\boldsymbol{x}) = -d = \boldsymbol{x}^\top \nabla\varphi(\boldsymbol{x}) \ , \forall \boldsymbol{x} : \prod_i x_i = 1 \ , \tag{89}$$

so φ is 1-homogeneous on $\mathcal{X}_g$, and we can apply Theorem 1. We have

$$\begin{aligned} & g(\boldsymbol{x}) \cdot D_\varphi \left(\frac{1}{g(\boldsymbol{x})} \cdot \boldsymbol{x} \| \frac{1}{g(\boldsymbol{y})} \cdot \boldsymbol{y} \right) \\ = \; & (\pi_{\boldsymbol{x}})^{1/d} \cdot \sum_i \left(\frac{x_i (\pi_{\boldsymbol{y}})^{1/d}}{y_i (\pi_{\boldsymbol{x}})^{1/d}} - \log \frac{x_i (\pi_{\boldsymbol{y}})^{1/d}}{y_i (\pi_{\boldsymbol{x}})^{1/d}} \right) - d(\pi_{\boldsymbol{x}})^{1/d} \\ = \; & \sum_i \frac{x_i (\pi_{\boldsymbol{y}})^{1/d}}{y_i} - d(\pi_{\boldsymbol{x}})^{1/d} \log (\pi_{\boldsymbol{x}})^{1/d} - (\pi_{\boldsymbol{x}})^{1/d} \log \pi_{\boldsymbol{y}} + (\pi_{\boldsymbol{x}})^{1/d} \log \pi_{\boldsymbol{y}} \\ & + d(\pi_{\boldsymbol{x}})^{1/d} \log (\pi_{\boldsymbol{x}})^{1/d} - d(\pi_{\boldsymbol{x}})^{1/d} \\ = \; & \sum_i \frac{x_i (\pi_{\boldsymbol{y}})^{1/d}}{y_i} - d(\pi_{\boldsymbol{x}})^{1/d} \; . \end{aligned} \tag{90}$$

We also have

$$\frac{\partial}{\partial x_i} \varphi^\dagger(\boldsymbol{x}) \quad = \quad -(1/x_i) \cdot (\pi_{\boldsymbol{x}})^{1/d} \; , \tag{91}$$

and so

$$\begin{aligned} D_{\varphi^\dagger}(\boldsymbol{x} \| \boldsymbol{y}) \quad &= \quad -d(\pi_{\boldsymbol{x}})^{1/d} + d(\pi_{\boldsymbol{y}})^{1/d} + \cdot \sum_i (x_i - y_i) \cdot \frac{(\pi_{\boldsymbol{y}})^{1/d}}{y_i} \\ &= \quad -d(\pi_{\boldsymbol{x}})^{1/d} + d(\pi_{\boldsymbol{y}})^{1/d} + \sum_i \frac{x_i (\pi_{\boldsymbol{y}})^{1/d}}{y_i} - d(\pi_{\boldsymbol{y}})^{1/d} \\ &= \quad \sum_i \frac{x_i (\pi_{\boldsymbol{y}})^{1/d}}{y_i} - d(\pi_{\boldsymbol{x}})^{1/d} \; , \end{aligned} \tag{92}$$

which is equal to Equation (90), so we check that Theorem 1 applies in this case.

Row VII — We use the following fact Kulis et al. [2009]. Let $\mathrm{X} = \mathrm{U}\mathrm{L}\mathrm{U}^\top$ and $\mathrm{Y} = \mathrm{V}\mathrm{T}\mathrm{V}^\top$ be the eigendecomposition of symmetric positive definite matrices X and Y, with $\mathrm{L} \doteq \mathrm{Diag}(\boldsymbol{l})$, $\mathrm{T} \doteq \mathrm{Diag}(\boldsymbol{t})$, and $\mathrm{U} \doteq [\boldsymbol{u}_1 | \boldsymbol{u}_2 | \cdots | \boldsymbol{u}_d]$, $\mathrm{V} \doteq [\boldsymbol{v}_1 | \boldsymbol{v}_2 | \cdots | \boldsymbol{v}_d]$ orthonormal; let $\varphi = \mathrm{tr}\,(\mathrm{X} \log \mathrm{X} - \mathrm{X})$. Then we have

$$D_\varphi(\mathrm{X} \| \mathrm{Y}) \quad = \quad \sum_{i,j} (\boldsymbol{u}_i^\top \boldsymbol{v}_j)^2 \cdot D_{\varphi_2}(l_i \| t_j) \; , \tag{93}$$

with $\varphi_2(x) = x \log x - x$. We pick $g(\mathrm{X}) = \mathrm{tr}\,(\mathrm{X}) = \sum_i l_i$, which brings from Equation (85)

$$\begin{aligned} & g(\mathrm{X}) \cdot D_\varphi \left(\frac{1}{g(\mathrm{X})} \cdot \mathrm{X} \| \frac{1}{g(\mathrm{Y})} \cdot \mathrm{Y} \right) \\ = \; & \sum_{i,j} (\boldsymbol{u}_i^\top \boldsymbol{v}_j)^2 \cdot \mathrm{tr}\,(\mathrm{X}) \cdot D_{\varphi_2} \left(\frac{l_i}{\mathrm{tr}\,(\mathrm{X})} \| \frac{t_j}{\mathrm{tr}\,(\mathrm{Y})} \right) \\ = \; & \sum_{i,j} (\boldsymbol{u}_i^\top \boldsymbol{v}_j)^2 \cdot \mathrm{tr}\,(\mathrm{X}) \cdot \left(\frac{l_i}{\mathrm{tr}\,(\mathrm{X})} \cdot \log \frac{l_i \cdot \mathrm{tr}\,(\mathrm{Y})}{t_j \cdot \mathrm{tr}\,(\mathrm{X})} - \frac{l_i}{\mathrm{tr}\,(\mathrm{X})} + \frac{t_j}{\mathrm{tr}\,(\mathrm{Y})} \right) \\ = \; & \sum_{i,j} (\boldsymbol{u}_i^\top \boldsymbol{v}_j)^2 \cdot \left(l_i \log \frac{l_i}{t_j} - l_i + t_j \right) + \log \left(\frac{\mathrm{tr}\,(\mathrm{Y})}{\mathrm{tr}\,(\mathrm{X})} \right) \cdot \sum_{i,j} (\boldsymbol{u}_i^\top \boldsymbol{v}_j)^2 \cdot l_i \\ & + \frac{\mathrm{tr}\,(\mathrm{X})}{\mathrm{tr}\,(\mathrm{Y})} \cdot \sum_{i,j} (\boldsymbol{u}_i^\top \boldsymbol{v}_j)^2 \cdot t_j - \sum_{i,j} (\boldsymbol{u}_i^\top \boldsymbol{v}_j)^2 \cdot t_j \; . \end{aligned} \tag{94}$$

Because U, V are orthonormal, we also get $\sum_{i,j}(\boldsymbol{u}_i^\top \boldsymbol{v}_j)^2 \cdot l_i = \sum_i l_i \sum_j \cos^2(\boldsymbol{u}_i, \boldsymbol{v}_j) = \sum_i l_i = \mathrm{tr}\,(\mathrm{X})$ and $\sum_{i,j}(\boldsymbol{u}_i^\top \boldsymbol{v}_j)^2 \cdot t_j = \mathrm{tr}\,(y)$, and so Equation (94) becomes

$$\begin{aligned} & g(\mathrm{X}) \cdot D_\varphi\left(\frac{1}{g(\mathrm{X})} \cdot \mathrm{X} \| \frac{1}{g(\mathrm{Y})} \cdot \mathrm{Y}\right) \\ &= \mathrm{tr}\,(\mathrm{X}\log\mathrm{X} - \mathrm{X}\log\mathrm{Y}) - \mathrm{tr}\,(\mathrm{X}) + \mathrm{tr}\,(\mathrm{Y}) + \mathrm{tr}\,(\mathrm{X}) \cdot \log\left(\frac{\mathrm{tr}\,(\mathrm{Y})}{\mathrm{tr}\,(\mathrm{X})}\right) + \mathrm{tr}\,(\mathrm{X}) - \mathrm{tr}\,(\mathrm{Y}) \\ &= \mathrm{tr}\,(\mathrm{X}\log\mathrm{X} - \mathrm{X}\log\mathrm{Y}) - \mathrm{tr}\,(\mathrm{X}) \cdot \log\left(\frac{\mathrm{tr}\,(\mathrm{X})}{\mathrm{tr}\,(\mathrm{Y})}\right) \quad . \end{aligned} \tag{95}$$

We also check that

$$\begin{aligned} \varphi^\dagger(\mathrm{X}) &= \mathrm{tr}\,(\mathrm{X}) \cdot \mathrm{tr}\left(\frac{1}{\mathrm{tr}\,(\mathrm{X})} \cdot \mathrm{X}\log\left(\frac{1}{\mathrm{tr}\,(\mathrm{X})} \cdot \mathrm{X}\right) - \frac{1}{\mathrm{tr}\,(\mathrm{X})} \cdot \mathrm{X}\right) \\ &= \mathrm{tr}\left(\mathrm{X}\log\left(\frac{1}{\mathrm{tr}\,(\mathrm{X})} \cdot \mathrm{X}\right)\right) - \mathrm{tr}\,(\mathrm{X}) \quad , \end{aligned} \tag{96}$$

and

$$\begin{aligned} \mathrm{X}\log\left(\frac{1}{\mathrm{tr}\,(\mathrm{X})} \cdot \mathrm{X}\right) &= \mathrm{U}\mathrm{L}\mathrm{U}^\top \mathrm{U}\log\left(\frac{1}{\mathbf{1}^\top \boldsymbol{l}} \cdot \mathrm{L}\right)\mathrm{U}^\top \\ &= \mathrm{U}\mathrm{L}\log\left(\frac{1}{\mathbf{1}^\top \boldsymbol{l}} \cdot \mathrm{L}\right)\mathrm{U}^\top \qquad (97) \\ &= \mathrm{U}\mathrm{L}\log\mathrm{L}\mathrm{U}^\top - \log\mathrm{tr}\,(\mathrm{X}) \cdot \mathrm{U}\mathrm{L}\mathrm{U}^\top \quad , \qquad (98) \end{aligned}$$

so that $\varphi^\dagger(\mathrm{X}) = \mathrm{tr}\,(\mathrm{X}\log\mathrm{X} - \mathrm{X}) - \mathrm{tr}\,(\mathrm{X}) \cdot \log\mathrm{tr}\,(\mathrm{X})$. Let $\varphi_3(\mathrm{X}) \doteq \mathrm{tr}\,(\mathrm{X}) \cdot \log\mathrm{tr}\,(\mathrm{X})$. We have $\nabla\varphi_3(\mathrm{X}) = (1 + \log\mathrm{tr}\,(\mathrm{X})) \cdot \mathrm{I}$. Since a (Bregman) divergence involving a sum of generators is the sum of (Bregman) divergences, we get

$$\begin{aligned} D_{\varphi^\dagger}(\mathrm{X}\|\mathrm{Y}) &= \mathrm{tr}\,(\mathrm{X}\log\mathrm{X} - \mathrm{X}\log\mathrm{Y} - \mathrm{X} + \mathrm{Y}) - \mathrm{tr}\,(\mathrm{X}) \cdot \log\mathrm{tr}\,(\mathrm{X}) + \mathrm{tr}\,(\mathrm{Y}) \cdot \log\mathrm{tr}\,(\mathrm{Y}) \\ &\quad + (1 + \log\mathrm{tr}\,(\mathrm{Y})) \cdot \mathrm{tr}\,(\mathrm{X} - \mathrm{Y}) \\ &= \mathrm{tr}\,(\mathrm{X}\log\mathrm{X} - \mathrm{X}\log\mathrm{Y}) - \mathrm{tr}\,(\mathrm{X}) \cdot \log\mathrm{tr}\,(\mathrm{X}) + \mathrm{tr}\,(\mathrm{X}) \cdot \log\mathrm{tr}\,(\mathrm{Y}) \\ &= \mathrm{tr}\,(\mathrm{X}\log\mathrm{X} - \mathrm{X}\log\mathrm{Y}) - \mathrm{tr}\,(\mathrm{X}) \cdot \log\left(\frac{\mathrm{tr}\,(\mathrm{X})}{\mathrm{tr}\,(\mathrm{Y})}\right) \quad , \end{aligned} \tag{99}$$

which is Equation (95).

Row VIII — We have the same property as for Row V, but this time with $\varphi_2 = -d - \log x$ [Kulis et al., 2009]. We check that whenever $\det(\mathrm{X}) = 1$, we have

$$\begin{aligned} \varphi(\mathrm{X}) = -d - \log\det(\mathrm{X}) &= -d \\ &= -\det(\mathrm{X})\mathrm{tr}\,(\mathrm{I}) \\ &= \mathrm{tr}\left(\det(\mathrm{X})\mathrm{X}^{-1}\mathrm{X}\right) = \mathrm{tr}\left(\nabla\varphi(\mathrm{X})^\top \mathrm{X}\right) \quad . \end{aligned} \tag{100}$$

For $g(\mathrm{X}) \doteq \det \mathrm{X}^{1/d}$, we get:

$$\begin{aligned} \varphi^\dagger(\mathrm{X}) &= \det \mathrm{X}^{1/d} \cdot \left(-d - \log\det\left(\frac{1}{\det \mathrm{X}^{1/d}} \cdot \mathrm{X}\right)\right) \\ &= \det \mathrm{X}^{1/d} \cdot \left(-d - \log\frac{1}{\det \mathrm{X}} \cdot \det \mathrm{X}\right) = -d \cdot \det \mathrm{X}^{1/d} \quad , \end{aligned} \tag{101}$$

and furthermore

$$\begin{aligned} \nabla\varphi^\dagger(\mathrm{X}) &= -d \cdot \nabla(\det \mathrm{X}^{1/d})(\mathrm{X}) \\ &= -\det(\mathrm{X}^{1/d}) \cdot \mathrm{X}^{-1} \end{aligned} \tag{102}$$

So,

$$\begin{aligned} D_{\varphi^\dagger}(\mathrm{X}\|\mathrm{Y}) &= -d\cdot\det \mathrm{X}^{1/d} + d\cdot\det \mathrm{Y}^{1/d} + \mathrm{tr}\left(\det(\mathrm{Y}^{1/d})\cdot \mathrm{Y}^{-1}(\mathrm{X}-\mathrm{Y})\right) \\ &= -d\cdot\det \mathrm{X}^{1/d} + d\cdot\det \mathrm{Y}^{1/d} + \det(\mathrm{Y}^{1/d})\mathrm{tr}\left(\mathrm{X}\mathrm{Y}^{-1}\right) - d\cdot\det \mathrm{Y}^{1/d} \\ &= \det(\mathrm{Y}^{1/d})\mathrm{tr}\left(\mathrm{X}\mathrm{Y}^{-1}\right) - d\cdot\det \mathrm{X}^{1/d} \ . \end{aligned} \tag{103}$$

We check that it is equal to:

$$\begin{aligned} & g(\mathrm{X})\cdot D_\varphi\left(\frac{1}{g(\mathrm{X})}\cdot\mathrm{X}\|\frac{1}{g(\mathrm{Y})}\cdot\mathrm{Y}\right) \\ &= \det \mathrm{X}^{1/d}\cdot\sum_{i,j}(\boldsymbol{u}_i^\top\boldsymbol{v}_j)^2\cdot\left(\frac{l_i\det \mathrm{Y}^{1/d}}{t_j\det \mathrm{X}^{1/d}} - \log\frac{l_i\det \mathrm{Y}^{1/d}}{t_j\det \mathrm{X}^{1/d}} - d\right) \ . \end{aligned} \tag{104}$$

To check it, we use the fact that, since U and V are orthonormal,

$$\begin{aligned} & \sum_{i,j}(\boldsymbol{u}_i^\top\boldsymbol{v}_j)^2\cdot\log\frac{l_i\det \mathrm{Y}^{1/d}}{t_j\det \mathrm{X}^{1/d}} \\ &= \sum_{i,j}(\boldsymbol{u}_i^\top\boldsymbol{v}_j)^2\cdot\log l_i - \sum_{i,j}(\boldsymbol{u}_i^\top\boldsymbol{v}_j)^2\cdot\log\det \mathrm{X}^{1/d} \\ &\quad + \sum_{i,j}(\boldsymbol{u}_i^\top\boldsymbol{v}_j)^2\cdot\log\det \mathrm{Y}^{1/d} - \sum_{i,j}(\boldsymbol{u}_i^\top\boldsymbol{v}_j)^2\cdot\log t_j \\ &= \underbrace{\sum_i \log l_i - d\cdot\log\det \mathrm{X}^{1/d}}_{=0} + \underbrace{d\cdot\log\det \mathrm{Y}^{1/d} - \sum_j\log l_j}_{=0} = 0 \ , \end{aligned} \tag{105}$$

which yields

$$\begin{aligned} g(\mathrm{X})\cdot D_\varphi\left(\frac{1}{g(\mathrm{X})}\cdot\mathrm{X}\|\frac{1}{g(\mathrm{Y})}\cdot\mathrm{Y}\right) &= \det \mathrm{X}^{1/d}\cdot\sum_{i,j}(\boldsymbol{u}_i^\top\boldsymbol{v}_j)^2\cdot\left(\frac{l_i\det \mathrm{Y}^{1/d}}{t_j\det \mathrm{X}^{1/d}} - d\right) \\ &= \det \mathrm{Y}^{1/d}\cdot\sum_{i,j}(\boldsymbol{u}_i^\top\boldsymbol{v}_j)^2\cdot\frac{l_i}{t_j} - d\cdot\det \mathrm{X}^{1/d} \\ &= \det \mathrm{Y}^{1/d}\cdot\mathrm{tr}\left(\mathrm{X}\mathrm{Y}^{-1}\right) - d\cdot\det \mathrm{X}^{1/d} \ , \end{aligned} \tag{106}$$

which is equal to Equation (103).

D Going deep: higher-order identities

We can generalize Theorem 1 to higher order identities. For this, consider $k > 0$ an integer, and let $g_1, g_2, ..., g_k : \mathcal{X} \to \mathbb{R}_*$ be a sequence of differentiable functions. For any $\ell, \ell' \in [k]_*$ such that $\ell \leq \ell'$, we let $\tilde{g}_{\ell,\ell'}$ be defined recursively as:

$$\tilde{g}_{\ell,\ell'}(\boldsymbol{x}) \doteq \begin{cases} \tilde{g}_{\ell-1,\ell'}(\boldsymbol{x}) \cdot g_{\ell'-(\ell-1)}\left(\frac{1}{\tilde{g}_{\ell-1,\ell'}(\boldsymbol{x})} \cdot \boldsymbol{x}\right) & \text{if} \quad 1 < \ell \leq \ell' \ , \\ g_{\ell'}(\boldsymbol{x}) & \text{if} \quad \ell = 1 \ , \end{cases} \tag{107}$$

and, for any $\ell \in [k]$,

$$\varphi^{\dagger(\ell)}(\boldsymbol{x}) \doteq \begin{cases} g_\ell(\boldsymbol{x}) \cdot \varphi^{\dagger(\ell-1)}\left(\frac{1}{g_\ell(\boldsymbol{x})} \cdot \boldsymbol{x}\right) & \text{if} \quad 0 < \ell \leq k \ , \\ \varphi(\boldsymbol{x}) & \text{if} \quad \ell = 0 \ . \end{cases} \tag{108}$$

Notice that even when all $g_\cdot$ are linear, this does not guarantee that some $\tilde{g}_{\ell,\ell'}$ for $\ell \neq 1$ is going to be linear. However, if for example $g_{\ell'}$ is linear and all "preceeding" g_ℓ ($\ell \leq \ell'$) are homogeneous of degree 1, then all $\tilde{g}_{\ell,\ell'}$ ($\forall \ell \leq \ell'$) are linear.

Corollary 17 *For any $k \in \mathbb{N}_*$, let $\varphi : \mathcal{X} \to \mathbb{R}$ be convex differentiable, and $g_\ell : \mathcal{X} \to \mathbb{R}_*$ ($\ell \in [k]$) a sequence of k differentiable functions. Then the following relationship holds, for any $\ell, \ell' \in [k]_*$ with $\ell \leq \ell'$:*

$$\tilde{g}_{\ell,\ell'}(\boldsymbol{x}) \cdot D_{\varphi^{\dagger(\ell'-\ell)}}\left(\frac{1}{\tilde{g}_{\ell,\ell'}(\boldsymbol{x})} \cdot \boldsymbol{x} \| \frac{1}{\tilde{g}_{\ell,\ell'}(\boldsymbol{y})} \cdot \boldsymbol{y}\right) = D_{\varphi^{\dagger(\ell')}}(\boldsymbol{x}\|\boldsymbol{y}) \ , \forall \boldsymbol{x}, \boldsymbol{y} \in \mathcal{X} \ , \tag{109}$$

with $\tilde{g}_{\ell,\ell'}$ defined as in Equation (107) and $\varphi^{\dagger(\ell')}$ defined as in Equation (108), if and only if at least one of the two following conditions hold:

(i) $\tilde{g}_{\ell,\ell'}$ is linear on $\mathcal{X}$;

(ii) $\varphi^{\dagger(\ell'-\ell)}$ is positive homogeneous of degree 1 on $\mathcal{X}_{\ell,\ell'} \doteq \{(1/\tilde{g}_{\ell,\ell'}(\boldsymbol{x})) \cdot \boldsymbol{x} : \boldsymbol{x} \in \mathcal{X}\}$.

We check that whenever φ is convex and all $g_\cdot$ are non-negative, then all $\varphi^{\dagger(\ell)}$ are convex ($\forall \ell \in [k]$). To prove this, we choose $\ell' = \ell$ and rewrite Equation (3), which brings, since $\varphi^{\dagger(\ell'-\ell)} = \varphi^{\dagger(0)} = \varphi$,

$$\tilde{g}_{\ell,\ell}(\boldsymbol{x}) \cdot D_{\varphi}\left(\frac{1}{\tilde{g}_{\ell,\ell}(\boldsymbol{x})} \cdot \boldsymbol{x} \| \frac{1}{\tilde{g}_{\ell,\ell}(\boldsymbol{y})} \cdot \boldsymbol{y}\right) = D_{\varphi^{\dagger(\ell)}}(\boldsymbol{x}\|\boldsymbol{y}) \ , \forall \boldsymbol{x}, \boldsymbol{y} \in \mathcal{X} \ . \tag{110}$$

Since φ is convex, a sufficient condition to prove our result is to show that $\tilde{g}_{\ell,\ell}$ is non-negative — which will prove that the right hand side of (110) is non-negative, and therefore $\varphi^{\dagger(\ell)}$ is convex —. This can easily be proven by induction from the expression of $\tilde{g}_{\ell,\ell'}$ in (107) and the fact that all $g_\cdot$ are non-negative.

One interesting candidate for simplification is when all $g_\cdot$ are the same linear function, say $g_\ell(\boldsymbol{x}) = \boldsymbol{a}^\top \boldsymbol{x} + b, \forall \ell \in [k]$. In this case, we have indeed:

$$\begin{aligned} \tilde{g}_{\ell,\ell'}(\boldsymbol{x}) &= \boldsymbol{a}^\top \boldsymbol{x} + b \cdot \tilde{g}_{\ell-1,\ell'}(\boldsymbol{x}) \\ &= b^\ell + \boldsymbol{a}^\top \boldsymbol{x} \cdot \sum_{j=1}^{\ell-1} b^j \ , \end{aligned} \tag{111}$$

$$\varphi^{\dagger(\ell')}(\boldsymbol{x}) = \left(b^{\ell'} + \boldsymbol{a}^\top \boldsymbol{x} \cdot \sum_{j=1}^{\ell'-1} b^j\right) \cdot \varphi\left(\frac{1}{b^{\ell'} + \boldsymbol{a}^\top \boldsymbol{x} \cdot \sum_{j=1}^{\ell'-1} b^j} \cdot \boldsymbol{x}\right) \ . \tag{112}$$

Proof [Proof of Corollary 17] To check eq. (4), we first remark (ℓ' being fixed) that it holds for $\ell = 1$ (this is eq. (4)), and then proceed by an induction from the induction base hypothesis that, for some $\ell \leq \ell'$,

$$\varphi^{\dagger(\ell')}(\boldsymbol{x}) = \tilde{g}_{\ell,\ell'}(\boldsymbol{x}) \cdot \varphi^{\dagger(\ell'-\ell)}\left(\frac{1}{\tilde{g}_{\ell,\ell'}(\boldsymbol{x})} \cdot \boldsymbol{x}\right) \ . \tag{113}$$

We now have

$$
\begin{aligned}
&\varphi^{\dagger(\ell')}(\boldsymbol{x}) \\
&= \frac{\tilde{g}_{\ell+1,\ell'}(\boldsymbol{x})}{g_{\ell'-\ell}\left(\frac{1}{\tilde{g}_{\ell,\ell'}(\boldsymbol{x})}\cdot\boldsymbol{x}\right)}\cdot\varphi^{\dagger(\ell'-\ell)}\left(\frac{1}{\tilde{g}_{\ell,\ell'}(\boldsymbol{x})}\cdot\boldsymbol{x}\right) && (114)\\
&= \frac{\tilde{g}_{\ell+1,\ell'}(\boldsymbol{x})}{g_{\ell'-\ell}\left(\frac{1}{\tilde{g}_{\ell,\ell'}(\boldsymbol{x})}\cdot\boldsymbol{x}\right)}\cdot g_{\ell'-\ell}\left(\frac{1}{\tilde{g}_{\ell,\ell'}(\boldsymbol{x})}\cdot\boldsymbol{x}\right)\cdot\varphi^{\dagger(\ell'-(\ell+1))}\left(\frac{1}{\tilde{g}_{\ell,\ell'}(\boldsymbol{x})g_{\ell'-\ell}\left(\frac{1}{\tilde{g}_{\ell,\ell'}(\boldsymbol{x})}\cdot\boldsymbol{x}\right)}\cdot\boldsymbol{x}\right) && (115)\\
&= \tilde{g}_{\ell+1,\ell'}(\boldsymbol{x})\cdot\varphi^{\dagger(\ell'-(\ell+1))}\left(\frac{1}{\tilde{g}_{\ell,\ell'}(\boldsymbol{x})g_{\ell'-\ell}\left(\frac{1}{\tilde{g}_{\ell,\ell'}(\boldsymbol{x})}\cdot\boldsymbol{x}\right)}\cdot\boldsymbol{x}\right)\\
&= \tilde{g}_{\ell+1,\ell'}(\boldsymbol{x})\cdot\varphi^{\dagger(\ell'-(\ell+1))}\left(\frac{1}{\tilde{g}_{\ell+1,\ell'}(\boldsymbol{x})}\cdot\boldsymbol{x}\right)\ . && (116)
\end{aligned}
$$

Eq. (114) comes from eq. (113) and the definition of $\tilde{g}_\ell$ in (107), eq. (115) comes from the definition of $\varphi^{\dagger(\ell'-\ell)}$ in (108), eq. (116) is a second use of the definition of $\tilde{g}_\ell$ in (107). ■

Notice the eventual high non-linearities introduced by the composition in eqs (107,108), which justifies the "deep" characterization.

E Additional application: exponential families

Let φ be the cumulant function of a regular φ-exponential family with pdf $p_\varphi(.|\boldsymbol{\theta})$, where $\boldsymbol{\theta} \in \mathcal{X}$ is its natural parameter. Let $\Omega(.)$ be a norm on $\mathcal{X}$. Let $\mathcal{X}_\Omega$ be the image of the application from $\mathcal{X}$ onto the Ω-ball of unit norm defined by $\boldsymbol{x} \mapsto (1/\Omega(\boldsymbol{x})) \cdot \boldsymbol{x}$. Let $\boldsymbol{\theta}_\Omega$ be the image of $\boldsymbol{\theta} \in \mathcal{X}$. For any two $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathcal{X}$, let

$$\mathrm{KL}_\varphi(\boldsymbol{\theta} \| \boldsymbol{\theta}') \quad \doteq \quad \int p_\varphi(\boldsymbol{x}|\boldsymbol{\theta}) \log \frac{p_\varphi(\boldsymbol{x}|\boldsymbol{\theta})}{p_\varphi(\boldsymbol{x}|\boldsymbol{\theta}')} \mathrm{d}\boldsymbol{x} \tag{117}$$

be the KL divergence between the two densities $p_\varphi(.|\boldsymbol{\theta})$ and $p_\varphi(.|\boldsymbol{\theta}')$.

Lemma 18 *For any convex φ which is restricted positive 1-homogeneous on $\mathcal{X}_\Omega$, the KL-divergence between two members of the same φ-exponential family satisfies:*

$$\mathrm{KL}_\varphi(\boldsymbol{\theta}_\Omega \| \boldsymbol{\theta}'_\Omega) \quad = \quad \frac{1}{\Omega(\boldsymbol{\theta})} \cdot D_{\varphi^\dagger}(\boldsymbol{\theta}' \| \boldsymbol{\theta}) \ . \tag{118}$$

Proof We know that $\mathrm{KL}(\boldsymbol{\theta} \| \boldsymbol{\theta}') = D_\varphi(\boldsymbol{\theta}' \| \boldsymbol{\theta})$ Boissonnat et al. [2010]. Hence,

$$\begin{aligned} D_{\varphi^\dagger}(\boldsymbol{\theta}' \| \boldsymbol{\theta}) \quad &= \quad \Omega(\boldsymbol{\theta}) \cdot D_\varphi \left(\frac{1}{\Omega(\boldsymbol{\theta})} \cdot \boldsymbol{\theta}' \| \frac{1}{\Omega(\boldsymbol{\theta}')} \cdot \boldsymbol{\theta} \right) \\ &= \quad \Omega(\boldsymbol{\theta}) \cdot D_\varphi(\boldsymbol{\theta}'_\Omega \| \boldsymbol{\theta}_\Omega) \\ &= \quad \Omega(\boldsymbol{\theta}) \cdot \mathrm{KL}_\varphi(\boldsymbol{\theta}_\Omega \| \boldsymbol{\theta}'_\Omega) \ , \end{aligned} \tag{119}$$

as claimed. ■

The interest in Lemma 18 is to provide an integral-free expression of the KL-divergence when natural parameters are scaled by non-trivial transformations. Furthermore, even when $D_{\varphi^\dagger}$ may not be a Bregman divergence, it still bears the same analytical form, which still may be useful for formal derivations.

F Additional application: computational geometry, nearest neighbor rules

Two important objects of central importance in (computational) geometry are balls and Voronoi diagrams induced by a distortion, with which we can characterize the topological and computational aspects of major structures (Voronoi diagrams, triangulations, nearest neighbor topologies, etc.) [Boissonnat et al., 2010].

Since a Bregman divergence is not necessarily symmetric, there are two types of (dual) balls that can be defined, the first or second types, where the variable $\boldsymbol{x}$ is respectively placed in the left or right position. The first type Bregman balls are convex while the second type are not necessarily convex. A Bregman ball of the second type (with center $\boldsymbol{c}$ and "radius" r) is defined as:

$$B'(\boldsymbol{c}, r|\mathcal{X}, \varphi) \quad \doteq \quad \{\boldsymbol{x} \in \mathcal{X} : D_\varphi(\boldsymbol{c}\|\boldsymbol{x}) \leq r\} \ . \tag{120}$$

It turns out that any divergence $D_{\varphi^\dagger}$ induces a ball of the second type, which is not necessarily analytically a Bregman ball (when $\varphi^\dagger$ is not convex), *but* turns out to define the *same* ball as a Bregman ball over transformed coordinates. In other words and to be a little bit more specific,

*"any $\boldsymbol{x}$ belongs to the ball of the second type induced by $D_{\varphi^\dagger}$ over $\mathcal{X}$ **iff** $(1/g(\boldsymbol{x})) \cdot \boldsymbol{x}$ $(\in \mathcal{X}_g)$ belongs to the Bregman ball of the second type induced by D_φ over $\mathcal{X}_g$."*

Theorem 19 *Let $(\varphi, g, \varphi^\dagger)$ satisfy the conditions of Theorem 1, with g non negative. Then*

$$B'(\boldsymbol{c}, r|\varphi^\dagger, \mathcal{X}) \quad = \quad B'\left(\boldsymbol{c}, \frac{r}{g(\boldsymbol{c})}\middle\| \varphi, \mathcal{X}_g\right) \ . \tag{121}$$

Proof From Theorem 1, we have

$$D_{\varphi^\dagger}(\boldsymbol{c}\|\boldsymbol{x}) \quad \leq \quad r \tag{122}$$

iff

$$D_\varphi\left(\frac{1}{g(\boldsymbol{c})} \cdot \boldsymbol{c}\middle\|\frac{1}{g(\boldsymbol{c})} \cdot \boldsymbol{x}\right) \quad \leq \quad \frac{1}{g(\boldsymbol{c})} \cdot r \ . \tag{123}$$

Hence,

$$\begin{aligned} B'(\boldsymbol{c}, r|\varphi^\dagger, \mathcal{X}) \quad &= \quad \{\boldsymbol{x} \in \mathcal{X}_g : D_\varphi(\boldsymbol{c}/g(\boldsymbol{c})\|\boldsymbol{x}/g(\boldsymbol{x})) \leq r/g(\boldsymbol{c})\} \\ &= \quad B'(\boldsymbol{c}, r/g(\boldsymbol{c})|\varphi, \mathcal{X}_g) \ , \end{aligned} \tag{124}$$

as claimed. ■

This property is not true for balls of the first type. What Theorem 19 says is that the topology induced by $D_{\varphi^\dagger}$ over $\mathcal{X}$ is just *no different* from that induced by D_φ over $\mathcal{X}_g$.

Let us now investigate Bregman Voronoi diagrams. In the same way as there exists two types of Bregman balls, we can define two types of Bregman Voronoi diagrams that depend on the equation of the *Bregman bisector* [Boissonnat et al., 2010]. Of particular interest is the Bregman bisector of the *first* type:

$$BB_\varphi(\boldsymbol{x}, \boldsymbol{y}|\mathcal{X}) \quad = \quad \{\boldsymbol{z} \in \mathcal{X} : D_\varphi(\boldsymbol{z}\|\boldsymbol{x}) = D_\varphi(\boldsymbol{z}\|\boldsymbol{y})\} \ . \tag{125}$$

It turns out that any divergence $D_{\varphi^\dagger}$ induces a bisector of the first type which is not necessarily analytically a Bregman bisector (when $\varphi^\dagger$ is not convex), *but* turns out to define the *same* bisector as a Bregman bisector over transformed coordinates. Again, we get more precisely

*"any $\boldsymbol{x}$ belongs to a Bregman bisector of the first type induced by $D_{\varphi^\dagger}$ over $\mathcal{X}$ **iff** $(1/g(\boldsymbol{x})) \cdot \boldsymbol{x}$ $(\in \mathcal{X}_g)$ belongs to the corresponding Bregman bisector of the first type induced by D_φ over $\mathcal{X}_g$."*

Theorem 20 *Let* $(\varphi, g, \varphi^\dagger)$ *satisfy the conditions of Theorem 1. Then*

$$BB_{\varphi^\dagger}(\boldsymbol{x}, \boldsymbol{y}|\mathcal{X}) \quad = \quad BB_{\varphi}(\boldsymbol{x}, \boldsymbol{y}|\mathcal{X}_g) \ . \tag{126}$$

(proof similar to Theorem 19) This property is not true for Bregman bisectors of the second type. Theorems 19, 20 have several important algorithmic consequences, some of which are listed now:

- the Voronoi diagram (resp. Delaunay triangulation) of the first type associated to $\varphi^\dagger$ can be constructed via the Voronoi diagram (resp. Delaunay triangulation) of the first type associated to φ [Boissonnat et al., 2010];
- range search using ball trees on $D_{\varphi^\dagger}$ can be efficiently implemented using Bregman divergence D_φ on $\mathcal{X}_g$ [Cayton, 2009];
- the minimum enclosing ball problem, the one-class clustering problem (an important problem in machine learning), with balls of the second type on $D_{\varphi^\dagger}$ can be solved via the minimum Bregman enclosing ball problem on D_φ [Nock and Nielsen, 2005].

G Review: binary density ratio estimation

For completeness, we quickly review the central result of Menon and Ong [2016, Proposition 3]. Let (P, Q, π) be densities giving $\mathbb{P}(\mathsf{X}|\mathsf{Y}=1), \mathbb{P}(\mathsf{X}=\boldsymbol{x}|\mathsf{Y}=-1), \mathbb{P}(\mathsf{Y}=1)$ respectively, and M giving $\mathbb{P}(\mathsf{X}=\boldsymbol{x})$ accordingly. Let $r(\boldsymbol{x}) \doteq \mathbb{P}(\mathsf{X}=\boldsymbol{x}|\mathsf{Y}=1)/\mathbb{P}(\mathsf{X}=\boldsymbol{x}|\mathsf{Y}=-1)$ be the density ratio of the class-conditional densities, and $\eta(\boldsymbol{x}) \doteq \mathbb{P}[\mathsf{Y}=1|\mathsf{X}=\boldsymbol{x}]$ be the class-probability function. Then, we have the following, which extends [Menon and Ong, 2016, Proposition 6] for the case $\pi \neq \frac{1}{2}$.

Lemma 21 *Given a class-probability estimator $\hat{\eta}\colon \mathcal{X} \to [0,1]$, let the density ratio estimator $\hat{r}$ be*

$$\hat{r}(\boldsymbol{x}) \quad = \quad \frac{1-\pi}{\pi} \cdot \frac{\hat{\eta}(\boldsymbol{x})}{1-\hat{\eta}(\boldsymbol{x})} \ . \tag{127}$$

Then for any convex differentiable $\varphi\colon [0,1] \to \mathbb{R}$,

$$\mathbb{E}_{\mathsf{X}\sim M}[D_{\varphi}(\eta(\mathsf{X})\|\hat{\eta}(\mathsf{X}))] \quad = \quad \pi \cdot \mathbb{E}_{\mathsf{X}\sim Q}\left[D_{\varphi^{\dagger}}(r(\mathsf{X})\|\hat{r}(\mathsf{X}))\right] \ . \tag{128}$$

where $\varphi^{\dagger}$ is as per Equation 4 with $g(z) \doteq \frac{1-\pi}{\pi} + z$.

Proof [Proof of Lemma 21] Note that

$$\begin{aligned} \frac{1}{g(r(\boldsymbol{x}))} \cdot r(\boldsymbol{x}) \quad &= \quad \frac{\pi\mathbb{P}(\mathsf{X}=\boldsymbol{x}|\mathsf{Y}=-1)}{\mathbb{P}(\mathsf{X}=\boldsymbol{x})} \cdot \frac{\mathbb{P}(\mathsf{X}=\boldsymbol{x}|\mathsf{Y}=1)}{\mathbb{P}(\mathsf{X}=\boldsymbol{x}|\mathsf{Y}=-1)} \\ &= \quad \frac{\pi\mathbb{P}(\mathsf{X}=\boldsymbol{x}|\mathsf{Y}=1)}{\mathbb{P}(\mathsf{X}=\boldsymbol{x})} \\ &= \quad \eta(\boldsymbol{x}) \ , \end{aligned} \tag{129}$$

and furthermore

$$\begin{aligned} \mathbb{P}(\mathsf{X}=\boldsymbol{x}) \quad &= \quad (1-\pi)\mathbb{P}(\mathsf{X}=\boldsymbol{x}|\mathsf{Y}=-1) + \pi\mathbb{P}(\mathsf{X}=\boldsymbol{x}|\mathsf{Y}=1) \\ &= \quad \pi \cdot \left(\frac{1-\pi}{\pi} + \frac{\mathbb{P}(\mathsf{X}=\boldsymbol{x}|\mathsf{Y}=1)}{\mathbb{P}(\mathsf{X}=\boldsymbol{x}|\mathsf{Y}=-1)}\right) \cdot \mathbb{P}(\mathsf{X}=\boldsymbol{x}|\mathsf{Y}=-1) \\ &= \quad \pi \cdot g(r(\boldsymbol{x})) \cdot \mathbb{P}(\mathsf{X}=\boldsymbol{x}|\mathsf{Y}=-1) \ . \end{aligned} \tag{130}$$

So,

$$\mathbb{E}_{M}[D_{\varphi}(\eta(\mathsf{X})\|\hat{\eta}(\mathsf{X}))] \quad = \quad \pi \cdot \mathbb{E}_{Q}\left[g(r(\mathsf{X})) \cdot D_{\varphi}(\eta(\mathsf{X})\|\hat{\eta}(\mathsf{X}))\right] \tag{131}$$

$$\quad = \quad \pi \cdot \mathbb{E}_{Q}\left[g(r(\mathsf{X})) \cdot D_{\varphi}\left(\frac{1}{g(r(\mathsf{X}))} \cdot r(\mathsf{X}) \middle\| \hat{\eta}(\mathsf{X})\right)\right] \tag{132}$$

$$\quad = \quad \pi \cdot \mathbb{E}_{Q}\left[g(r(\mathsf{X})) \cdot D_{\varphi}\left(\frac{1}{g(r(\mathsf{X}))} \cdot r(\mathsf{X}) \middle\| \frac{1}{g(\hat{r}(\mathsf{X}))} \cdot \hat{r}(\mathsf{X})\right)\right] \tag{133}$$

$$\quad = \quad \pi \cdot \mathbb{E}_{Q}\left[D_{\varphi^{\dagger}}(r(\mathsf{X})\|\hat{r}(\mathsf{X}))\right] \ , \tag{134}$$

as claimed. Equation (131) comes from (130), Equation (132) comes from (129), Equation (133) comes from (127) and the definition of g. Equation (134) comes from Theorem 1, noting that g is linear. ■

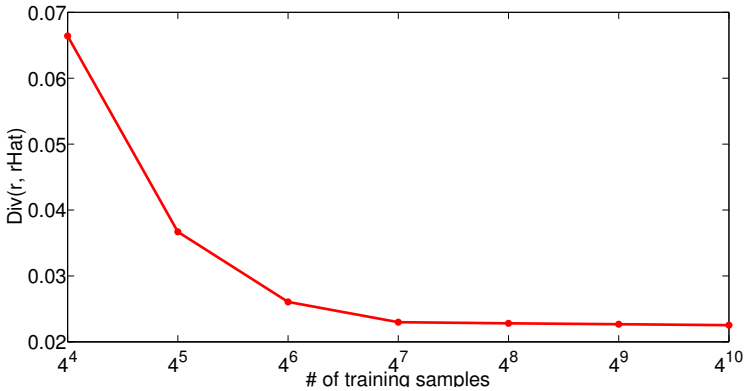

Figure 4: Density ratio estimate divergence $\mathbb{E}_{\mathsf{X}\sim P_C}\left[D_{\varphi^\dagger}(r(\mathsf{X}), \hat{r}(\mathsf{X}))\right]$ as a function of # of training samples.

H Additional experiments

H.1 Multiclass density ratio experiments

We consider a synthetic multiclass density ratio estimation problem. We fix $\mathcal{X} = \mathbb{R}^2$, and consider $C = 3$ classes. We consider a distribution where the class-conditionals $\Pr(\mathsf{X}|\mathsf{Y} = c)$ are multivariate Gaussians with means $\boldsymbol{\mu}_c$ and covariance $\sigma_c^2 \cdot \mathrm{Id}$. As the class-conditionals have a closed form, we can explicitly compute $\boldsymbol{\eta}$, as well the density ratio $\boldsymbol{r}$ to the reference class $c^* = C$.

For fixed class prior $\boldsymbol{\pi} = \Pr(\mathsf{Y} = c)$, we draw N_{Tr} samples from $\Pr(\mathsf{X}, \mathsf{Y})$. From this, we estimate the class-probability $\hat{\boldsymbol{\eta}}$ using multiclass logistic regression. This can be seen as minimising $\mathbb{E}_{\mathsf{X}\sim M}\left[D_\varphi(\boldsymbol{\eta}(\mathsf{X})\|\hat{\boldsymbol{\eta}}(\mathsf{X}))\right]$ where $\varphi(z) = \sum_i z_i \log z_i$ is the generator for the KL-divergence.

We then use Equation 6 to estimate the density ratios $\hat{\boldsymbol{r}}$ from $\hat{\boldsymbol{\eta}}$. On a fresh sample of N_{te} instances from $\Pr(\mathsf{X}, \mathsf{Y})$, we estimate the right hand side of Lemma 2, *viz.* $\mathbb{E}_{\mathsf{X}\sim P_C}\left[D_{\varphi^\dagger}(r(\mathsf{X})\|\hat{r}(\mathsf{X}))\right]$, where $\varphi^\dagger$ uses the g as specified in Lemma 2. From the result of Lemma 2, we expect this divergence to be small when $\hat{\boldsymbol{\eta}}$ is a good estimator of $\boldsymbol{\eta}$.

We perform the above for sample sizes $N \in \{4^4, 4^5, \ldots, 4^{10}\}$, with $N_{\text{Tr}} = 0.8N$ and $N_{\text{Te}} = 0.2N$. For each sample size, we perform $T = 25$ trials, where in each trial we randomly draw $\boldsymbol{\pi}$ uniformly over $(1/C)\mathbf{1} + (1 - 1/C) \cdot [0, 1]^C$, $\boldsymbol{\mu}_c$ from $0.1 \cdot \mathcal{N}(\mathbf{0}, 1)$, and σ_c uniformly from $[0.5, 1]$. Figure 4 summarises the mean divergence across the T trials for each sample size. We see that, as expected, with more training samples the divergence decreases in a monotone fashion.

H.2 Adaptive filtering experiments

Tables 8 – 13 present *in extenso* the experiments of p-LMS vs DN-p-LMS, as a function of (p, q), whether target $\boldsymbol{u}$ is sparse or not, and the misestimation factor ρ for X_p. We refer to Kivinen et al. [2006] for the formal definitions used for sparse / dense targets as well as for the experimental setting, which we have reproduced with the sole difference that the signal changes periodically each 1 000 iterations.

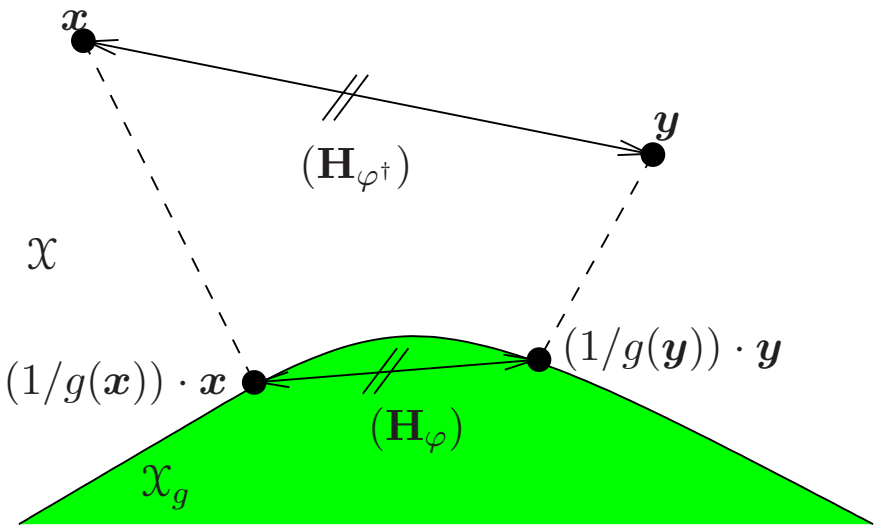

Figure 5: A depiction of the adaptive isometry that Theorem 1 provides.

I Comment: Theorem 1 is a scaled isometry in disguise (sometimes)

Theorem 1 states in fact an isometry under some conditions, but an adaptive one in the sense that metrics involved rely on all parameters, and in particular on the points involved in the divergences (See Figure 5). Indeed, a simple Taylor expansion of the equation (2) (main file) shows that any such Bregman distortion with a twice differentiable generator can be expressed as:

$$D_\varphi(\boldsymbol{x}\|\boldsymbol{y}) \quad = \quad \frac{1}{2}\cdot(\boldsymbol{x}-\boldsymbol{y})^\top \mathbf{H}_\varphi(\boldsymbol{x}-\boldsymbol{y}) \ , \tag{135}$$

for *some* value of the Hessian $\mathbf{H}_\varphi$ depending on $\boldsymbol{x},\boldsymbol{y}$ (see for example [Kivinen et al., 2006, Appendix I], [Amari and Nagaoka, 2000]). Hence, under the constraint that both φ and $\varphi^\dagger$ are twice differentiable, eq. (3) becomes

$$g(\boldsymbol{x})\cdot\left(\frac{1}{g(\boldsymbol{x})}\cdot\boldsymbol{x}-\frac{1}{g(\boldsymbol{y})}\cdot\boldsymbol{y}\right)^\top \mathbf{H}_\varphi\left(\frac{1}{g(\boldsymbol{x})}\cdot\boldsymbol{x}-\frac{1}{g(\boldsymbol{y})}\cdot\boldsymbol{y}\right) \quad = \quad (\boldsymbol{x}-\boldsymbol{y})^\top \mathbf{H}_{\varphi^\dagger}(\boldsymbol{x}-\boldsymbol{y}) \ . \tag{136}$$

Assuming g non-negative (which, by the way, enforces the convexity of $\varphi^\dagger$), we get by taking square roots,

$$\sqrt{g(\boldsymbol{x})}\cdot\left\|\frac{1}{g(\boldsymbol{x})}\cdot\boldsymbol{x}-\frac{1}{g(\boldsymbol{y})}\cdot\boldsymbol{y}\right\|_{\mathbf{H}_\varphi} \quad = \quad \|\boldsymbol{x}-\boldsymbol{y}\|_{\mathbf{H}_{\varphi^\dagger}} \ , \tag{137}$$

which is a scaled isometry relationship between $\mathcal{X}_g$ (left) and $\mathcal{X}$ (right), but again the metrics involved depend on the arguments. Nevertheless, eq. (137) displays a sophisticated relationship between distances in $\mathcal{X}_g$ and in $\mathcal{X}$ which may prove useful in itself.

References

S. Acharyya, A. Banerjee, and D. Boley. Bregman divergences and triangle inequality. In *SDM*, 2013.

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

D. Arthur and S. Vassilvitskii. k-means++ : the advantages of careful seeding. In *19 th SODA*, 2007.

A. Banerjee, S. Merugu, I. Dhillon, and J. Ghosh. Clustering with Bregman divergences. *JMLR*, 6:1705–1749, 2005.

A. Beck and M. Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

J.-D. Boissonnat, F. Nielsen, and R. Nock. Bregman Voronoi diagrams. *DCG*, 44(2):281–307, 2010.

S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004. ISBN 0521833787.

A. Buja, W. Stuetzle, and Y. Shen. Loss functions for binary class probability estimation and classification: Structure and applications, 2005. Unpublished manuscript.

S.-R. Buss and J.-P. Fillmore. Spherical averages and applications to spherical splines and interpolation. *ACM Transactions on Graphics*, 20:95–126, 2001.

L. Cayton. Efficient bregman range search. In *NIPS*22*, pages 243–251, 2009.

N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning and Games*. Cambridge University Press, 2006.

M. Collins, R. Schapire, and Y. Singer. Logistic regression, AdaBoost and Bregman distances. *MLJ*, 2002.

I. Dhillon and D.-S. Modha. Concept decompositions for large sparse text data using clustering. *MLJ*, 42:143–175, 2001.

I.-S. Dhillon and J.-A. Tropp. Matrix nearness problems with Bregman divergences. *SIAM Journal on Matrix Analysis and Applications*, 29(4):1120–1146, 2008.

J. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra. Efficient projections onto the ℓ_1-ball for learning in high dimensions. In *ICML '08*, pages 272–279, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-205-4.

Y. Endo and S. Miyamoto. Spherical k-means++ clustering. In *Proc. of the 12th MDAI*, pages 103–114, 2015.

G.-A. Galperin. A concept of the mass center of a system of material points in the constant curvature spaces. *Communications in Mathematical Physics*, 154:63–84, 1993.

E. Hazan and S. Kale. Projection-free online learning. In John Langford and Joelle Pineau, editors, *ICML '12*, pages 521–528, New York, NY, USA, 2012. ACM.

M. Hernández-Lobato, Y. Li, M. Rowland, D. Hernández-Lobato, T. Bui, and R.-E. Turner. Black-box alpha-divergence minimization. In *33rd ICML*, 2016.

M. Jaggi. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In *30th ICML*, 2013.

J. Kivinen, M. Warmuth, and B. Hassibi. The p-norm generalization of the LMS algorithm for adaptive filtering. *IEEE Trans. SP*, 54:1782–1793, 2006.

D. Kuang, S. Yun, and H. Park. SymNMF: nonnegative low-rank approximation of a similarity matrix for graph clustering. *J. Global Optimization*, 62:545–574, 2014.

B. Kulis, M.-A. Sustik, and I.-S. Dhillon. Low-rank kernel learning with Bregman matrix divergences. *JMLR*, 10: 341–376, 2009.

A.-K. Menon and C.-S. Ong. Linking losses for class-probability and density ratio estimation. In *ICML*, 2016.

R. Nock and F. Nielsen. Fitting the Smallest Enclosing Bregman Ball. In *Proc. of the 16 th European Conference on Machine Learning*, pages 649–656. Springer-Verlag, 2005.

R. Nock and F. Nielsen. Bregman divergences and surrogates for learning. *IEEE PAMI*, 31:2048–2059, 2009.

R. Nock, P. Luosto, and J. Kivinen. Mixed Bregman clustering with approximation guarantees. In *ECML*, 2008.

R. Nock, F. Nielsen, and S.-I. Amari. On conformal divergences and their population minimizers. *IEEE Trans. IT*, 62:527–538, 2016.

M. Reid and R. Williamson. Information, divergence and risk for binary experiments. *JMLR*, 12:731–817, 2011.

M.-D. Reid and R.-C. Williamson. Composite binary losses. *JMLR*, 11:2387–2422, 2010.

J. Reisinger, A. Waters, B. Silverthorn, and R.-J. Mooney. Spherical topic models. In *27th ICML*, pages 903–910, 2010.

G. Rong, M. Jin, and X. Guo. Hyperbolic centroidal Voronoi tessellation. In *14 th ACM SPM*, 2010.

O. Schwander and F. Nielsen. *Matrix Information Geometry*, chapter Learning Mixtures by Simplifying Kernel Density Estimators, pages 403–426. Springer Berlin Heidelberg, 2013.

A.-J. Shahani, E.-B. Gulsoy, V.-J. Roussochatzakis, J.-W. Gibbs, J.-L. Fife, and P.-W. Voorhees. The dynamics of coarsening in highly anisotropic systems: Si particles in Al−Si liquids. *Acta Materialia*, 97:325 – 337, 2015.

S. Shalev-Shwartz, Y. Singer, and N. Srebro. Pegasos: Primal estimated sub-gradient solver for SVM. In *ICML '08*, page 807–814. ACM, 2007. ISBN 978-1-59593-793-3.

H. Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227 – 244, 2000. ISSN 0378-3758.

J. Straub, G. Rosman, O. Freifeld, J.-J. Leonard, and J.-W. Fisher III. A mixture of Manhattan frames: Beyond the Manhattan world. In *Proc. of the 27^{th} IEEE CVPR*, pages 3770–3777, 2014.

J. Straub, N. Bhandari, J.-J. Leonard, and J.-W. Fisher III. Real-time Manhattan world rotation estimation in 3d. In *Proc. of the 27^{th} IROS*, pages 1913–1920, 2015a.

J. Straub, T. Campbell, J.-P. How, and J.-W. Fisher III. Small-variance nonparametric clustering on the hypersphere. In *Proc. of the 28^{th} IEEE CVPR*, pages 334–342, 2015b.

J. Straub, J. Chang, O. Freifeld, and J.-W. Fisher III. A Dirichlet process mixture model for spherical data. In *Proc. of the 18^{th} AISTATS*, 2015c.

M. Sugiyama, T. Suzuki, and T. Kanamori. Density-ratio matching under the Bregman divergence: a unified framework of density-ratio estimation. *AISM*, 64(5):1009–1044, 2012. ISSN 0020-3157.

A.-A. Ungar. *Mathematics Without Boundaries: Surveys in Interdisciplinary Research*, chapter An Introduction to Hyperbolic Barycentric Coordinates and their Applications, pages 577–648. Springer New York, 2014.

R. Vidal. Subspace clustering. *IEEE Signal Processing Magazine*, 28:52–68, 2011.

R.-C. Williamson, E. Vernet, and M.-D. Reid. Composite multiclass losses, 2014. Unpublished manuscript.

H. Zhang and S. Sra. First-order methods for geodesically convex optimization. *CoRR*, abs/1602.06053, 2016.

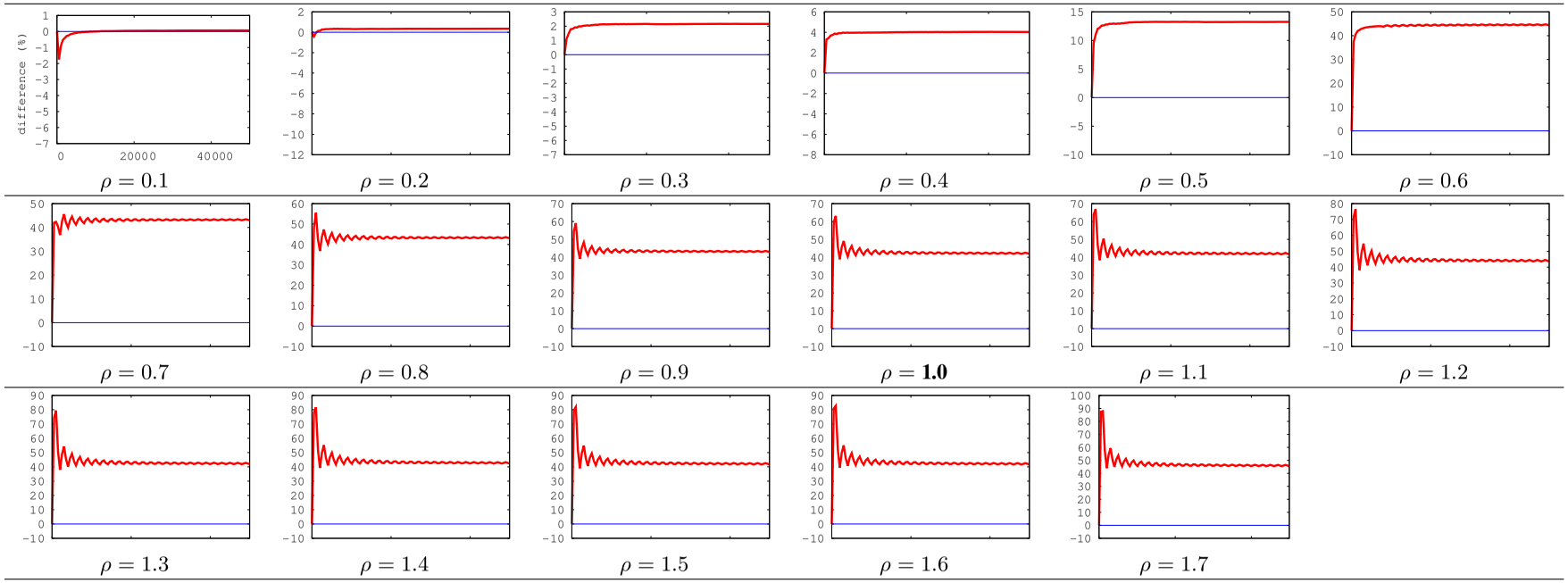

Table 8: Error(p-LMS) - Error(DN-p-LMS) as a function of t ($\in \{1, 2, ..., 50000\}$), $\boldsymbol{u}$ = dense, $(p, q) = (1.17, 6.9)$.

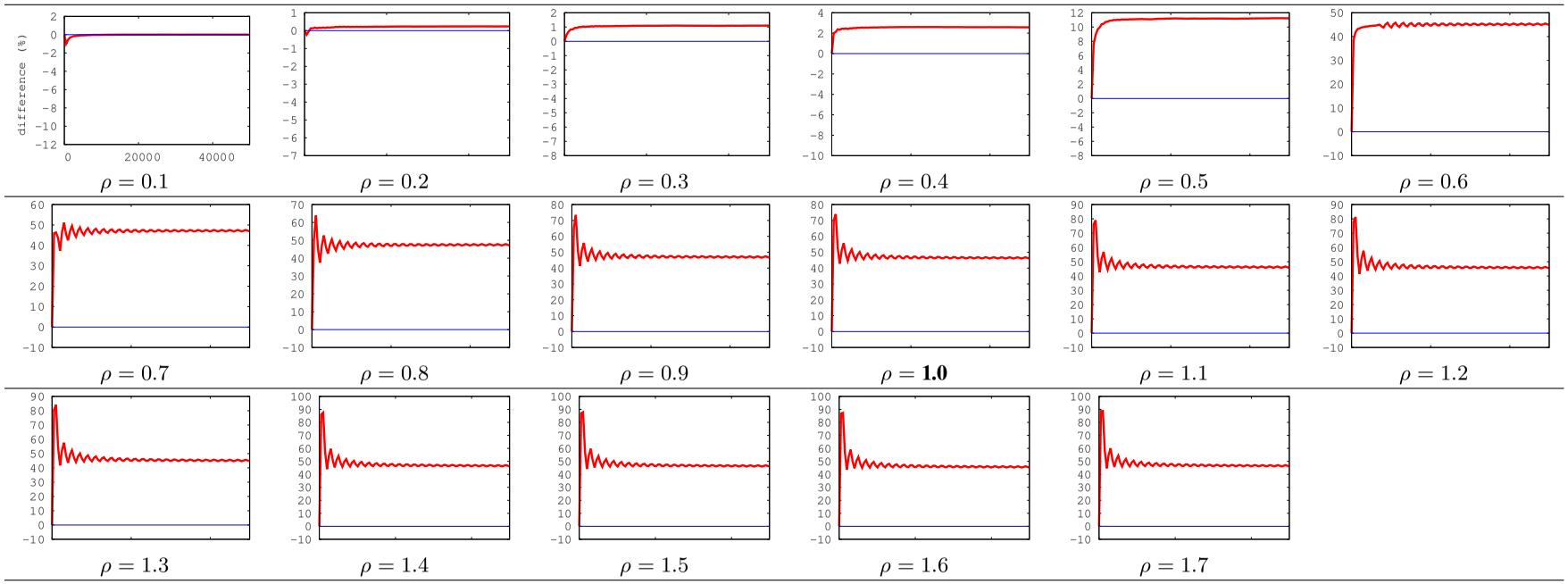

Table 9: Error(p-LMS) - Error(DN-p-LMS) as a function of t ($\in \{1, 2, ..., 50000\}$), $\boldsymbol{u}$ = sparse, $(p, q) = (1.17, 6.9)$.

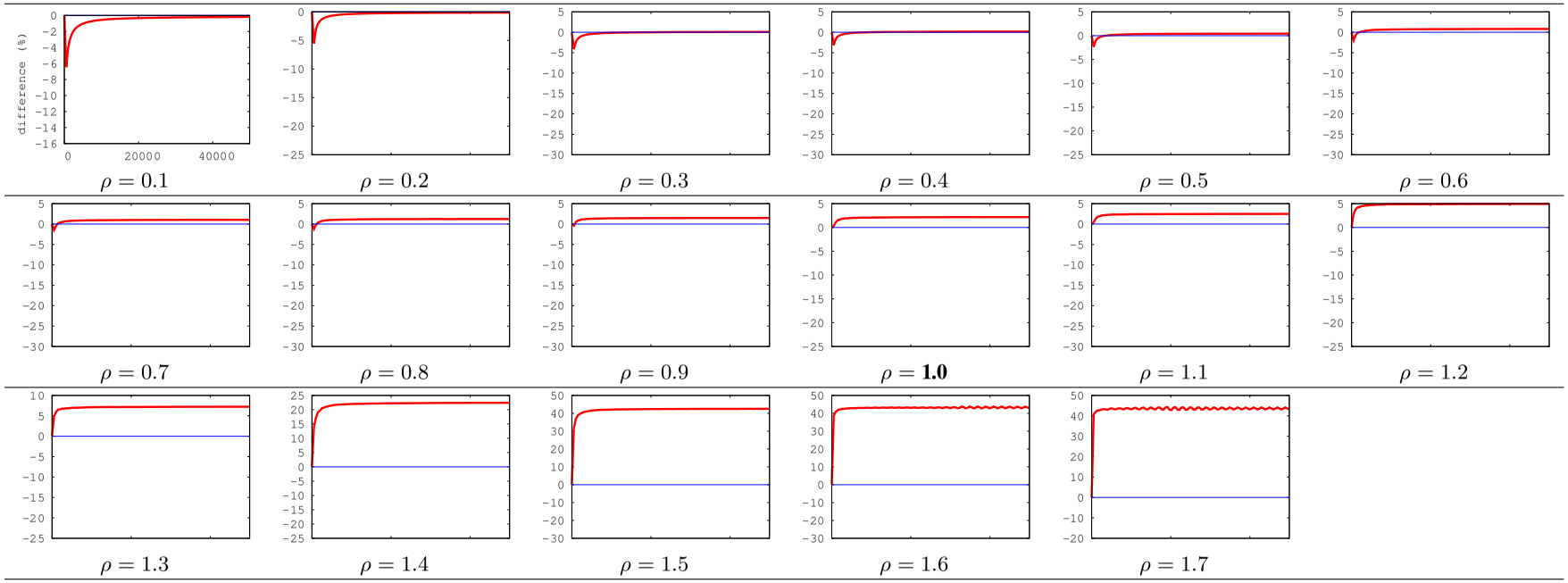

Table 10: Error(p-LMS) - Error(DN-p-LMS) as a function of t ($\in \{1, 2, ..., 50000\}$), $\boldsymbol{u}$ = dense, $(p, q) = (2.0, 2.0)$.

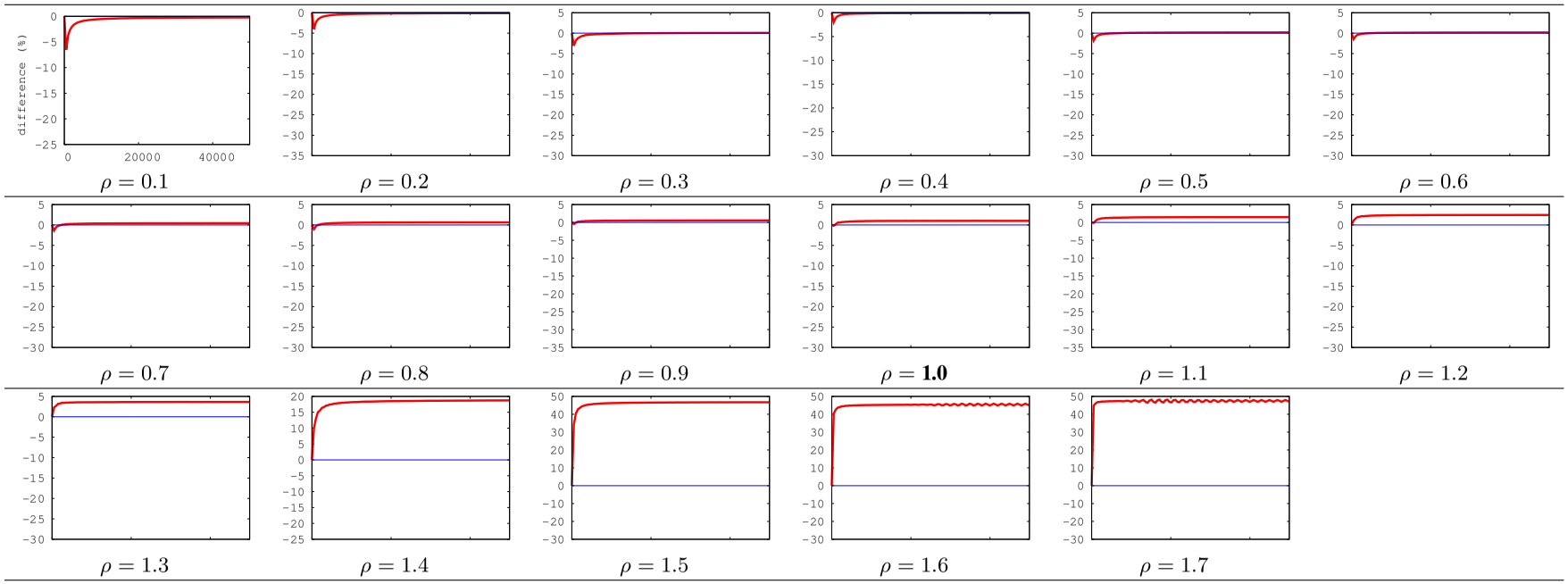

Table 11: Error(p-LMS) - Error(DN-p-LMS) as a function of t ($\in \{1, 2, ..., 50000\}$), $\boldsymbol{u}$ = sparse, $(p, q) = (2.0, 2.0)$.

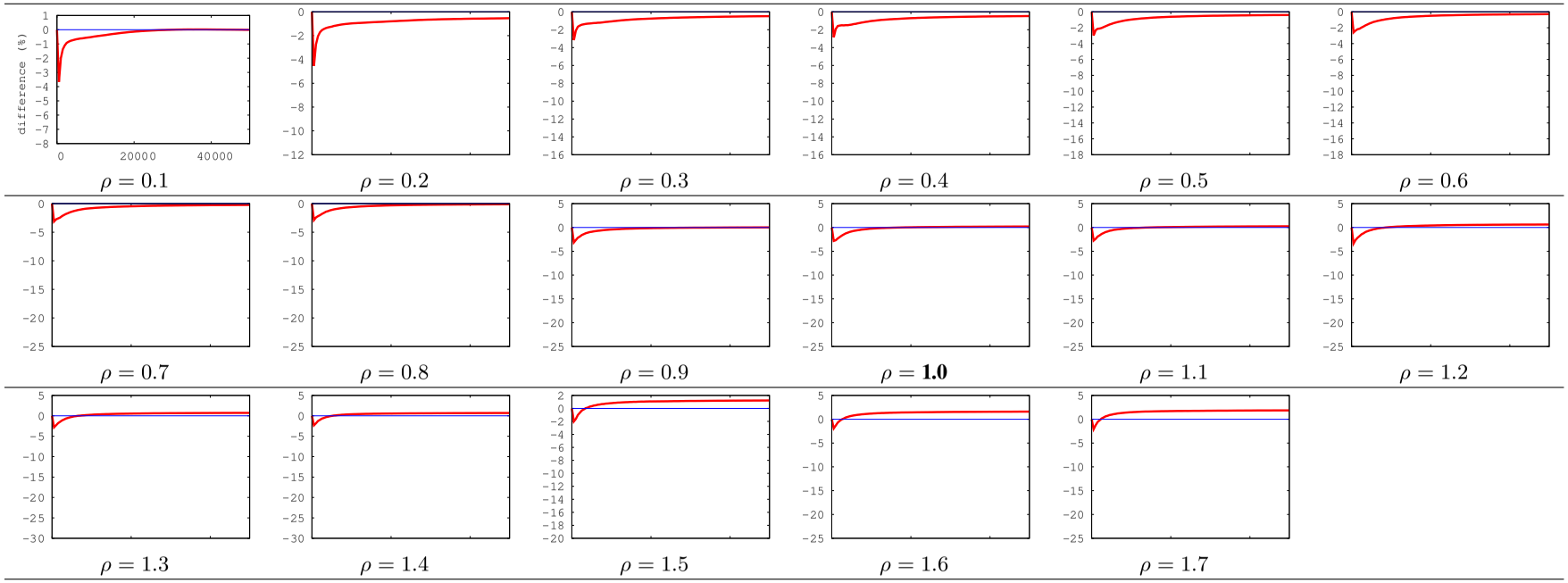

Table 12: Error(p-LMS) - Error(DN-p-LMS) as a function of t ($\in \{1, 2, ..., 50000\}$), $\boldsymbol{u}$ = dense, $(p, q) = (6.9, 1.17)$.

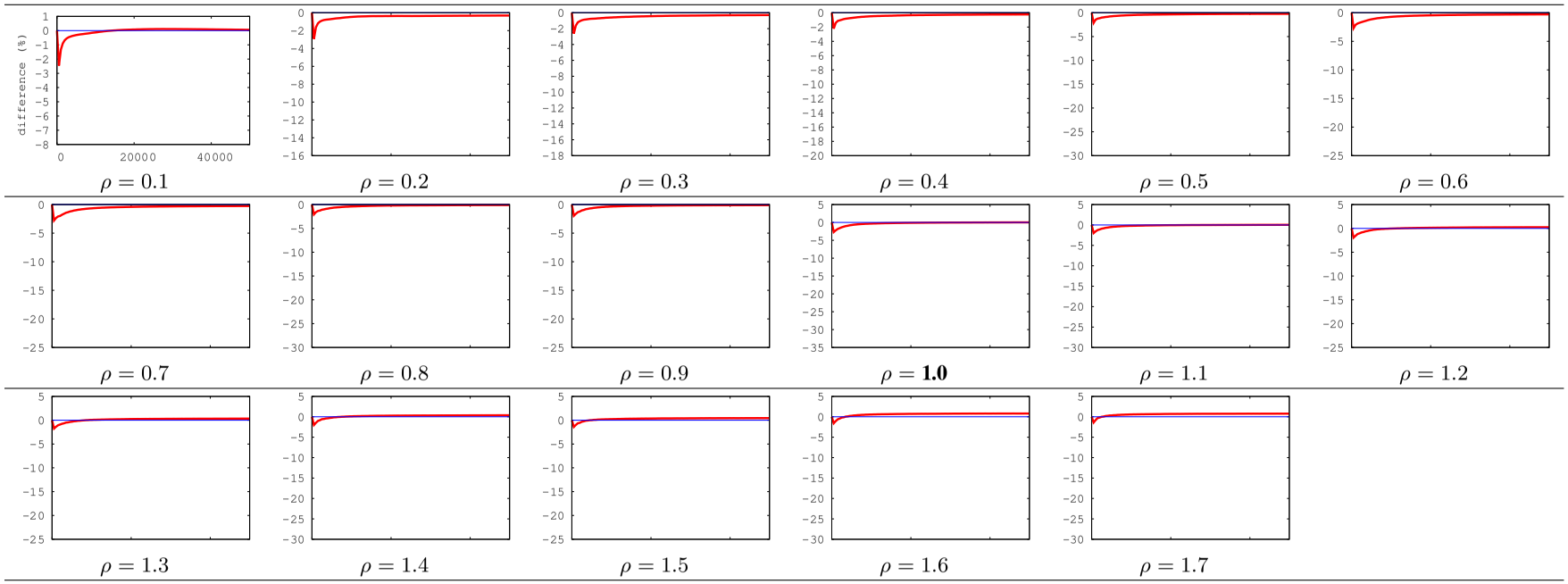

Table 13: Error(p-LMS) - Error(DN-p-LMS) as a function of t ($\in \{1, 2, ..., 50000\}$), $\boldsymbol{u}$ = sparse, $(p, q) = (6.9, 1.17)$.