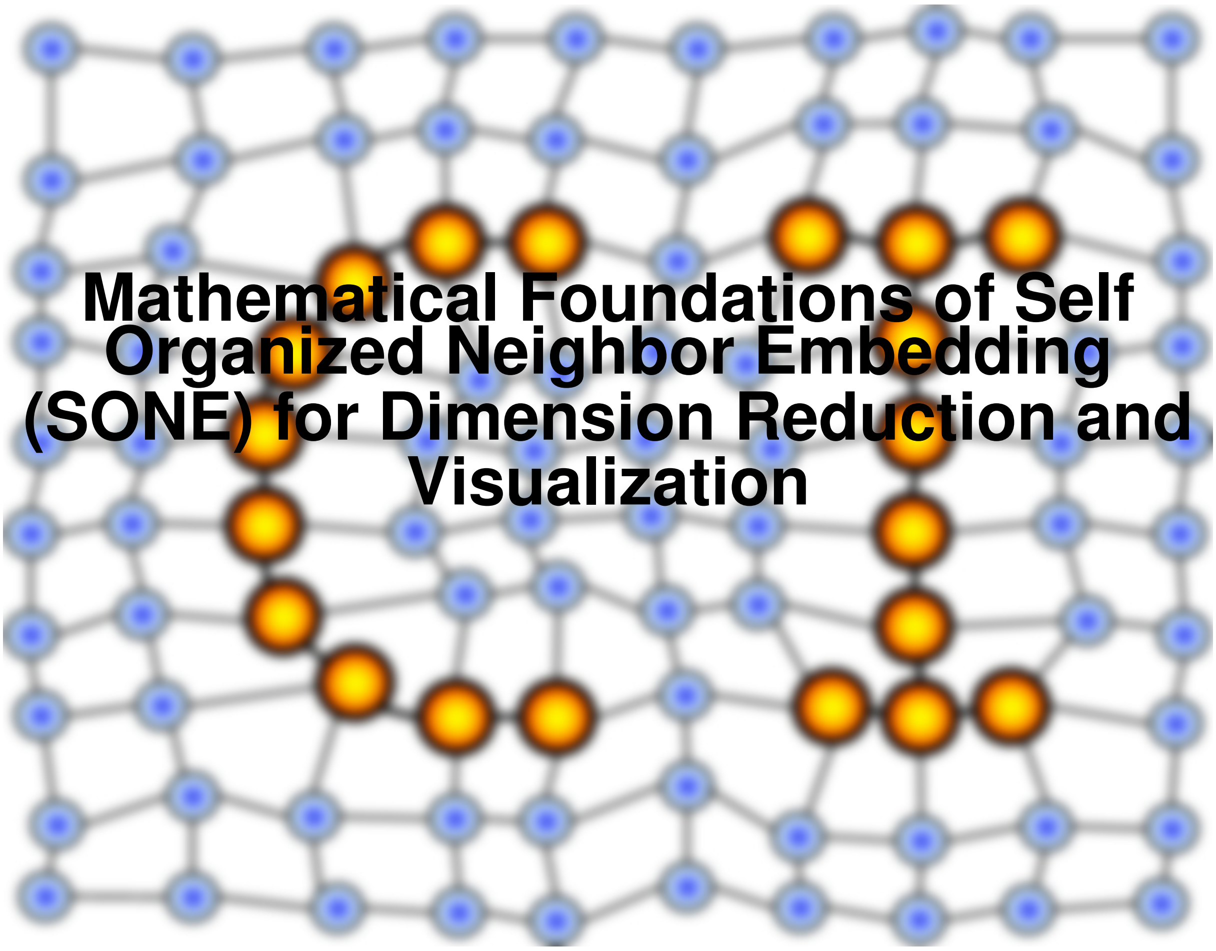
MACHINE LEARNING REPORTS



Mathematical Foundations of Self Organized Neighbor Embedding (SONE) for Dimension Reduction and Visualization

Report 03/2010
Submitted: 25.07.2010
Published:28.07.2010

Kerstin Bunte[1], Sven Haase[2], Michael Biehl[1] and Thomas Villmann[2]
(1) University of Groningen, Institute for Mathematics and Computing Science
P.O. Box 407, 9700 AK Groningen - The Netherlands
(2) University of Applied Sciences Mittweida, Department of MPI
Technikumplatz 17, 09648 Mittweida - Germany

Machine Learning Reports,Research group on Computational Intelligence,
http://www.mni.hs-mittweida.de/

Abstract

In this paper we propose the generalization of the recently introduced Neighbor Embedding Exploratory Observation Machine (NE-XOM) for dimension reduction and visualization. We provide a general mathematical framework called Self Organized Neighbor Embedding (SONE). It treats the components, like data similarity measures and neighborhood functions, independently and easily changeable. And it enables the utilization of different divergences, based on the theory of Fréchet derivatives. In this way we propose a new dimension reduction and visualization algorithm, which can easily adapted to the user specific request and the actual problem.

1 Introduction

Various dimension reduction techniques have been introduced based on different properties of the original data to be preserved. The spectrum ranges from linear projections of original data, such as in Principal Component Analysis (PCA) or classical Multidimensional Scaling (MDS) [29] to a wide range of locally linear and non-linear approaches, such as Isomap [41, 17], Locally Linear Embedding (LLE) [39], Local Linear Coordination (LLC) [43], or charting [8, 40]. Stochastic Neighbor Embedding (SNE) [24] approximates the probability distribution in the high-dimensional space, defined by neighboring points, with their probability distribution in a lower-dimensional space. In [42] the authors proposed a technique called t-SNE, which is a variation of SNE considering another statistical model assumption for data distributions.

Other methods aim at the preservation of the classification accuracy in lower dimensions and incorporate the available label information for the embedding, e. g. Linear Discriminant Analysis (LDA) [22] and generalizations thereof [6], extensions of the Self Organizing Map (SOM) [28] incorporating class labels [47], and Limited Rank Matrix Learning Vector Quantization (LiRaM LVQ) [13, 12]. For a comprehensive review on nonlinear dimensionality reduction methods, we refer to [31].

Recently, the idea of fast and efficient online learning was combined with the high-quality of divergence based optimization, resulting in a new dimension reduction algorithm called Neighbor Embedding XOM (NE-XOM) [10]. The authors connected a computational approach to topology learning, the Exploration Observation Machine (XOM) as intruduced by Wismüller [48, 49], with the divergence optimization of SNE.

In this contribution, we extend the approach proposed in [10], with a mathematical foundation for the generalization of the principle to arbitrary divergences based on Fréchet derivatives. This generalized framework is called Self Organized Neighbor Embedding (SONE) in the following. In this way we propose a new dimension reduction and visualization algorithm, which can easily adapted to the user specific request and the actual problem.

We will describe the XOM algorithm and its NE-XOM extension in section 2.1 and section 2.2, describe the new generalized framework SONE in section 3, show the extension for some famous families of divergences and conclude in section 5.

2 The basic Algorithms

In this chapter we briefly explain the recently introduced combination of direct divergence optimization, inspired by Stochastic Neighbor Embedding (SNE) [24], and fast online learning, using the Exploration Observation Machine (XOM) [48, 49, 50]. The SNE is a recently proposed powerful method, which yields high quality embeddings measured by, e. g., trustworthiness and continuity. It aims at minimizing differences of the pairwise probability distribution of points in the data and embedding space measured by the Kullback-Leibler (KL) Divergence. Like many other dimension reduction techniques, SNE has a computational and memory complexity, which is quadratic in the number of points. The complexity of the XOM algorithm on the other hand can be easily controlled by the structure hypothesis and its complexity is linear with the number of points. It embeds low-dimensional image vectors driven by the topology of the data points in the high-dimensional space. In the following we will briefly review the XOM and its combination with the ideas of SNE, which results in the new algorithm called Neighbor Embedding XOM (NE-XOM) introduced in [11].

2.1 The Exploration Observation Machine (XOM)

XOM maps a finite number of high-dimensional data points $\boldsymbol{x}^i \in \mathcal{X}$ in the observation space $\mathcal{X}$ to low-dimensional image vectors $\boldsymbol{y}^i \in \mathcal{E}$ in the embedding space $\mathcal{E}$. The embedding space is associated with a structure hypothesis, given by a number of sampling vectors $\boldsymbol{s} \in \mathcal{E}$, which corresponds to the final structure in which the data is embedded. These can be seen as a generalization of the prototypes as included in the Self Organizing Map (SOM). Reasonable choices for the sampling vectors $\boldsymbol{s}$ are: the location on a regular lattice structure in $\mathcal{E}$, discrete positions in $\mathcal{E}$ as representation of a finite number of class centers, drawn from a mixture of Gaussian to represent a finite number of clusters, or uniformly sampled in a region of $\mathcal{E}$ to indicate that the visualization of the data should occupy the full projection space. Unlike SOM, XOM does not project the sampling vectors $\boldsymbol{s}$ to the data space, rather it projects the data to the embedding space:

The image vectors $\boldsymbol{y}$ can be initialized arbitrarily, e. g. randomly or by means of a PCA. The XOM algorithm follows than three main steps:

1. Present a sampling vector $\boldsymbol{s}$ from the structure hypothesis,
2. find the best matching input vector

$$\Psi(\boldsymbol{s}) = \boldsymbol{x}^i \text{ for which } d_{\mathcal{E}}(\boldsymbol{s}, \boldsymbol{y}^i) \text{ is minimum.} \tag{2.1}$$

   Here, $d_{\mathcal{E}}(\boldsymbol{s}, \boldsymbol{y}) : \mathcal{E} \times \mathcal{E} \to \mathrm{I\!R}$ denotes an arbitrary dissimilarity measure used in the embedding space.

3. Perform the update of all image vectors with the adaptation rule:

$$\boldsymbol{y}^k \to \boldsymbol{y}^k - \tau \cdot h_\sigma(d_{\mathcal{X}}(\Psi(\boldsymbol{s}), \boldsymbol{x}^k)) \frac{\partial d_{\mathcal{E}}(\boldsymbol{s}, \boldsymbol{y}^k)}{\partial \boldsymbol{y}^k} \quad , \tag{2.2}$$

   where $d_{\mathcal{X}}(\boldsymbol{x}^j, \boldsymbol{x}^l) : \mathcal{X} \times \mathcal{X} \to \mathrm{I\!R}$ denotes the dissimilarity measure used in the observation space $\mathcal{X}$, τ defines a learning rate with $0 < \tau \leq 1$ and $h_\sigma(d_{\mathcal{X}}(\boldsymbol{x}^i, \boldsymbol{x}^j))$ or for short h_σ^{ij} defines the neighborhood cooperation in the observation space. It constitutes the topology of the data, which is tried to be preserved also in the low-dimensional space $\mathcal{E}$. It might be chosen according to a distribution with variance σ, e. g. a Gaussian:

$$h_\sigma^{ij} = \exp\left(\frac{-d_{\mathcal{X}}(\boldsymbol{x}^i, \boldsymbol{x}^j)}{2\sigma^2}\right) \text{ with } \sigma > 0 \ . \tag{2.3}$$

The steps 1-3 are repeated until a stopping criteria is met, e. g. the maximal number of iterations is reached.

In this way the projections $\boldsymbol{y}$ are arranged around the a priori chosen structure elements $\boldsymbol{s}$, such that image vectors are close to the same sampling vector if their corresponding data points $\boldsymbol{x}$ are neighbors in the data space.
The XOM algorithm in its original form does not correspond to a cost function. However, a variation following Heskes [23] by replacing the best match input data vector by

$$\Psi(\boldsymbol{s}) = \boldsymbol{x}^i \text{ where } \sum_j h_\sigma(d_{\mathcal{X}}(\boldsymbol{x}^i, \boldsymbol{x}^j))\; d_{\mathcal{E}}(\boldsymbol{s}, \boldsymbol{y}^j) \text{ is minimum,} \tag{2.4}$$

leads to the cost function:

$$E_{\text{XOM}} \sim \int \sum_i \delta_{\Psi(\boldsymbol{s}),\boldsymbol{x}^i} \cdot \sum_j h_\sigma(d_{\mathcal{X}}(\boldsymbol{x}^i, \boldsymbol{x}^j))\; d_{\mathcal{E}}(\boldsymbol{s}, \boldsymbol{y}^j)\; p(\boldsymbol{s}) d\boldsymbol{s} \ . \tag{2.5}$$

The XOM learning rule corresponds to a stochastic gradient descent procedure with respect to this cost function.

2.2 The Neighbor Embedding XOM (NE-XOM)

In this section we review the combination of direct divergence inspired by SNE with fast sequential online learning resulting in a new algorithm called Neighbor Embedding XOM (NE-XOM) introduced in [11].

Let $h_\sigma(d_{\mathcal{X}}(\Psi_{\text{GKL}}(\boldsymbol{s}), \boldsymbol{x}^k))$ and $g_\varsigma(d_{\mathcal{E}}(\boldsymbol{s}, \boldsymbol{y}^k))$ $\Big($abbreviated by $h_\sigma^{\Psi_{\text{GKL}}(\boldsymbol{s})}(k)$ and $g_\varsigma^{\boldsymbol{s}}(k)\Big)$ be any positive integrable measures denoting the neighborhood cooperation in the observation and the embedding space respectively. Following the ideas of SNE, NE-XOM tries to minimize the difference between these two neighborhood functions measured by the Kullback-Leibler (KL) divergence. Note, that in contrast to SNE, which is originally defined for probability densities $p(r)$ with scalar r, the constraint $\int p(r)\ dr = 1$ is not imposed here.

The neighborhood function $h_\sigma^{\Psi_{\text{GKL}}(\boldsymbol{s})}$ of the observation space $\mathcal{X}$ might be a Gaussian like h_σ in Eq. (2.3). Depending on the choice for the neighborhood cooperation g_ς in the embedding space with variance ς the learning rule and thus the final embedding may vary a lot. We will provide in the following the learning rules for the case of a Gaussian neighborhood cooperation:

$$g_\varsigma^{\boldsymbol{s}}(k) = \exp\left(\frac{-d_{\mathcal{E}}(\boldsymbol{s}, \boldsymbol{y}^k)}{2\varsigma^2}\right) \tag{2.6}$$

$$\text{with the derivative} \quad \frac{\partial g_\varsigma^{\boldsymbol{s}}(k)}{\partial \boldsymbol{y}^k} = \left(-\frac{g_\varsigma^{\boldsymbol{s}}(k)}{2\varsigma^2}\right)\frac{\partial d_{\mathcal{E}}(\boldsymbol{s}, \boldsymbol{y}^k)}{\partial \boldsymbol{y}^k} \tag{2.7}$$

and a t-Distribution-like cooperation function:

$$g_\varsigma^{\boldsymbol{s}}(k) = (1 + d_{\mathcal{E}}(\boldsymbol{s}, \boldsymbol{y}^k)/\varsigma)^{\left(-\frac{\varsigma+1}{2}\right)} \tag{2.8}$$

$$\text{with the derivative} \quad \frac{\partial g_\varsigma^{\boldsymbol{s}}(k)}{\partial \boldsymbol{y}^k} = \left(-\frac{\varsigma+1}{2\varsigma}\right)\frac{g_\varsigma^{\boldsymbol{s}}(k)}{(1 + d_{\mathcal{E}}(\boldsymbol{s}, \boldsymbol{y}^k)/\varsigma)}\frac{\partial d_{\mathcal{E}}(\boldsymbol{s}, \boldsymbol{y}^k)}{\partial \boldsymbol{y}^k}\ . \tag{2.9}$$

For positive measures p and q the Generalized KL (GKL) divergence:

$$\mathrm{D}_{\text{GKL}}(p\|q) = \int p(r)\ln\left(\frac{p(r)}{q(r)}\right)\ dr - \int [p(r) - q(r)]\ dr \tag{2.10}$$

is used. In analogy with the XOM cost function E_{XOM} Eq. (2.5) we are able to define a cost function using the neighborhood functions from the original and the embedding space and the GKL divergence D_{GKL} Eq. (2.10):

$$E_{\text{NEXOM}} \sim \int \sum_i \delta_{\Psi_{\text{GKL}}(\boldsymbol{s}), \boldsymbol{x}^i} \cdot \sum_j \left[h_\sigma^{\Psi_{\text{GKL}}(\boldsymbol{s})}(j)\ln\left(\frac{h_\sigma^{\Psi_{\text{GKL}}(\boldsymbol{s})}(j)}{g_\varsigma^{\boldsymbol{s}}(j)}\right) - h_\sigma^{\Psi_{\text{GKL}}(\boldsymbol{s})}(j) + g_\varsigma^{\boldsymbol{s}}(j)\right]\ p(\boldsymbol{s})d\boldsymbol{s}\ , \tag{2.11}$$

where $h_\sigma^{\Psi_{\text{GKL}}(\boldsymbol{s})} = h_\sigma(d_{\mathcal{X}}(\Psi_{\text{GKL}}(\boldsymbol{s}), \boldsymbol{x}^k))$ and the best match data point $\Psi_{\text{GKL}}(\boldsymbol{s})$ for a given sampling vector $\boldsymbol{s}$ is given by

$$\Psi_{\text{GKL}}(\boldsymbol{s}) = \boldsymbol{x}^i \text{ such that } \sum_j \left[h_\sigma^{\Psi_{\text{GKL}}(\boldsymbol{s})}(j)\ln\left(\frac{h_\sigma^{\Psi_{\text{GKL}}(\boldsymbol{s})}(j)}{g_\varsigma^{\boldsymbol{s}}(j)}\right) - h_\sigma^{\Psi_{\text{GKL}}(\boldsymbol{s})}(j) + g_\varsigma^{\boldsymbol{s}}(j)\right] \text{ is minimum.} \tag{2.12}$$

This results in the learning rule for the NE-XOM:

$$\boldsymbol{y}^k = \boldsymbol{y}^k - \tau\frac{\partial g_\varsigma^{\boldsymbol{s}}(k)}{\partial \boldsymbol{y}_k}\left(1 - \frac{h_\sigma^{\Psi_{\text{GKL}}(\boldsymbol{s})}(k)}{g_\varsigma^{\boldsymbol{s}}(k)}\right)\ , \tag{2.13}$$

In case of a Gaussian g_ς^k the learning rule reads:

$$\boldsymbol{y}^k = \boldsymbol{y}^k - \frac{\tau}{2\varsigma^2}\left(h_\sigma^{\Psi_{\mathsf{GKL}}(\boldsymbol{s})}(k) - g_\varsigma^{\boldsymbol{s}}(k)\right)\frac{\partial d_{\mathcal{E}}(\boldsymbol{s},\boldsymbol{y}^k)}{\partial \boldsymbol{y}^k} \quad , \tag{2.14}$$

and with a t-distributed g_ς^k defined in Eq. (2.8) it leads to:

$$\boldsymbol{y}^k = \boldsymbol{y}^k - \frac{\varsigma+1}{2\varsigma}\frac{\tau}{(1+d_{\mathcal{E}}(\boldsymbol{s},\boldsymbol{y}^k)/\varsigma)}\left(h_\sigma^{\Psi_{\mathsf{GKL}}(\boldsymbol{s})}(k) - g_\varsigma^{\boldsymbol{s}}(k)\right)\frac{\partial d_{\mathcal{E}}(\boldsymbol{s},\boldsymbol{y}^k)}{\partial \boldsymbol{y}^k} \quad . \tag{2.15}$$

In the following sections we will generalize this concept for arbitrary divergences.

3 A Generalized Framework for Dimension Reduction

In [45] the authors discuss the use of divergences in different supervised and unsupervised Vector Quantization schemes. They show that divergences can be an alternative to the most frequently used Euclidean distance and may lead to improved classification accuracy. Furthermore divergences can be applied in the field of dimension reduction: for example in Stochastic Neighbor Embedding (SNE), t-distributed SNE (t-SNE) and Multidimensional Scaling (MDS) [24, 42, 30]. In [46] the mathematical foundation to extend SNE and t-SNE for use with arbitrary divergences is given. We will use this concept to generalize the algorithm explained in section 2.2. In the following we will briefly review the concept of divergences and Fréchet derivatives and we will define the mathematical framework for Self Organized Neighbor embedding (SONE) using arbitrary divergences.

3.1 Divergences

Divergences are functionals $\mathrm{D}(p\|q)$ designed as dissimilarity measures between two nonnegative integrable functions p and q [14]. In practice, usually p corresponds to the observed data and q denotes the estimated or expected data. We call p and q positive measures defined on r in the domain V. The weight of the functional p is defined as

$$W(p) = \int_V p(r)\; dr \quad . \tag{3.1}$$

Positive measures with the additional constraint $W(p) = 1$ are denoted as probability density functions. Generally speaking, divergences measure a quasi-distance or directed difference, while we are mostly interested in separable measures, which satisfy the condition

$$\mathrm{D}(p\|q)\begin{cases} > 0 \text{ for } p \neq q \\ = 0 \text{ iff } p \equiv q \quad . \end{cases} \tag{3.2}$$

In contrast to a metric, a divergence must not be symmetric in the sense $\mathrm{D}(p\|q) = \mathrm{D}(q\|p)$ and does not necessarily satisfy the triangular inequality $\mathrm{D}(p\|q) \leq \mathrm{D}(p\|z) + \mathrm{D}(z\|q)$. Note, that the definition of the considered divergences for non-normalized positive measures has an important property. It allows the analysis of patterns of different size to be weighted differently, e. g. images with different size or documents of variable length. Following [14] one can distinguish at least three main families of divergences with the same consistent properties: Bregman-divergences, Csiszár's f-divergences and γ-divergences. Note that all these families contain the Kullback-Leibler (KL) divergence as special case, so the KL-divergence can be seen as the non empty intersection between these sets of divergences.

3.2 The Fréchet Derivative

Suppose V and W are Banach spaces and $U \subset V$ is an open subset of V. The function $f : U \to W$ is called Fréchet differentiable at $x \in U$, if there exists a bounded linear operator $A_x : V \to W$, such that for $h \in U$

$$\lim_{h \to 0} \frac{\|f(x+h) - f(x) - A_x(h)\|_W}{\|h\|_V} = 0 \ . \tag{3.3}$$

This general definition can be used for functions $L : B \to \mathrm{I\!R}$, defined as mappings from a functional Banach space B to $\mathrm{I\!R}$. Further let B be equipped with a norm $\|\cdot\|$ and $f, h \in B$ are two functionals. The Fréchet derivative $\frac{\delta L[f]}{\delta f}$ of L at point f (i. e. in a function f) in the direction h is formally defined as:

$$\lim_{\epsilon \to 0} \frac{1}{\epsilon} \left(L[f + \epsilon h] - L[f] \right) =: \frac{\delta L[f]}{\delta f}[h] \ . \tag{3.4}$$

This concept will be used for the generalization of the definition given in Eq. (2.13).

3.3 Self Organized Neighbor Embedding (SONE)

We define a cost function for arbitrary Divergences $\mathrm{D}(p\|q)$:

$$E_{\text{SONE}} = \int \sum_i \delta_{\Psi^{\mathrm{D}}(\boldsymbol{s}),\boldsymbol{x}^i} \cdot \sum_j \mathrm{D}\left(h_\sigma^{\Psi_{\mathrm{D}}(\boldsymbol{s})}(j) \middle\| g_\varsigma^{\boldsymbol{s}}(j) \right) \ p(\boldsymbol{s}) d\boldsymbol{s} \ \ , \tag{3.5}$$

where the best matching data point $\Psi^{\mathrm{D}}(\boldsymbol{s})$ for $\boldsymbol{s}$ is defined as:

$$\Psi_{\mathrm{D}}(\boldsymbol{s}) = \boldsymbol{x}^i \text{ such that } \sum_j \mathrm{D}\left(h_\sigma^{\Psi_{\mathrm{D}}(\boldsymbol{s})}(j) \middle\| g_\varsigma^{\boldsymbol{s}}(j) \right) \text{ is minimum.} \tag{3.6}$$

Let V be a Banach space and $U \subset V$ an open subset of V. The divergence $\mathrm{D} : U \to \mathrm{I\!R}$ is defined as a mapping from U to $\mathrm{I\!R}$. Further D uses a bounded linear operator: the integral $\int : V \to \mathrm{I\!R}$. So the derivative of the cost function (3.5) with respect to the image vectors $\boldsymbol{y}^k$ can be done using the Fréchet derivative Eq. (3.4):

$$\begin{aligned}
\frac{\partial E_{\text{SONE}}}{\partial \boldsymbol{y}^k} &= \int \left[\frac{\delta \mathrm{D}\left(h_\sigma^{\Psi_{\mathrm{D}}(\boldsymbol{s})} \middle\| g_\varsigma^{\boldsymbol{s}} \right)}{\delta g_\varsigma^{\boldsymbol{s}}}[l] \cdot \frac{\partial g_\varsigma^{\boldsymbol{s}}}{\partial \boldsymbol{y}^k} \right] \ dl && (3.7) \\
&= \int \left[\left. \frac{\delta \mathrm{D}\left(h_\sigma^{\Psi_{\mathrm{D}}(\boldsymbol{s})} \middle\| g_\varsigma^{\boldsymbol{s}} \right)}{\delta g_\varsigma^{\boldsymbol{s}}} \right|_l \cdot \delta_{l,k} \cdot \frac{\partial g_\varsigma^{\boldsymbol{s}}}{\partial \boldsymbol{y}^k} \right] \ dl && (3.8) \\
&= \left. \frac{\delta \mathrm{D}\left(h_\sigma^{\Psi_{\mathrm{D}}(\boldsymbol{s})} \middle\| g_\varsigma^{\boldsymbol{s}} \right)}{\delta g_\varsigma^{\boldsymbol{s}}} \right|_k \cdot \frac{\partial g_\varsigma^{\boldsymbol{s}}(k)}{\partial \boldsymbol{y}^k} \ . && (3.9)
\end{aligned}$$

This yields the online learning update rule for a given sampling vector $\boldsymbol{s}$ and learning rate τ:

$$\begin{aligned} \boldsymbol{y}^k &= \boldsymbol{y}^k - \tau \Delta \boldsymbol{y}^k \\ \Delta \boldsymbol{y}^k &= \left. \frac{\delta \mathrm{D}\left(h_\sigma^{\Psi_{\mathrm{D}}(\boldsymbol{s})} \| g_\varsigma^{\boldsymbol{s}}\right)}{\delta g_\varsigma^{\boldsymbol{s}}} \right|_k \cdot \frac{\partial g_\varsigma^{\boldsymbol{s}}(k)}{\partial \boldsymbol{y}^k} \end{aligned} \tag{3.10}$$

4 SONE Gradients for Positive Measures

In the following we formulate the learning rules for general positive measures p and q, for which the constraint $\int p(r)\, dr = 1$ is not imposed. Thus we demand the neighborhood function to fulfill $0 \leq h_\sigma, g_\varsigma \leq 1$.

4.1 Bregman divergences

A very famous class of divergences are the Bregman divergences, which are widely used in optimization and clustering [3, 9, 18, 19, 36]. A Bregman divergence is defined as a pseudo-distance between two positive measures p and q: $D_{\mathsf{B}}(p\|q) : \mathcal{L} \times \mathcal{L} \to \mathrm{I\!R}^+$. Let ϕ be a strictly convex real-valued function with the domain of the Lebesgue-integrable functions $\mathcal{L}$ and twice continuously Fréchet-differentiable [33]. Then the Bregman divergence can be defined by

$$D_{\mathsf{B}}^{\phi}(p\|q) = \phi(p) - \phi(q) - \frac{\delta \phi(q)}{\delta q}[p - q] \ , \tag{4.1}$$

where $\frac{\delta \phi(q)}{\delta q}$ is the Fréchet derivative of ϕ with respect to q [45].

The Bregman divergence includes many prominent dissimilarity measures like the Euclidean distance (with generating function $\phi(p) = p^2$), the generalized Kullback-Leibler (or I-) divergence, the Itakura-Saito divergence and the β-divergence [14, 45, 20].

Well known fundamental properties of the Bregman divergences are [14]:

1. **Convexity:** A Bregman divergence is always convex in its first argument but not necessarily in its second.

2. **Non-negativity:** $D_{\mathsf{B}}^{\phi}(p\|q) \geq 0$ and $D_{\mathsf{B}}^{\phi}(p\|q) = 0$ iff $p \equiv q$

3. **Linearity:** Any positive linear combination of Bregman divergences is also a Bregman divergence:

$$\mathrm{D}_{\mathsf{B}}^{c_1\phi_1 + c_2\phi_2}(p\|q) = c_1 \mathrm{D}_{\mathsf{B}}^{\phi_1}(p\|q) + c_2 \mathrm{D}_{\mathsf{B}}^{\phi_2}(p\|q) \quad \text{with } c_1, c_2 > 0 \tag{4.2}$$

4. **Invariance:** A Bregman divergence is invariant under affine transformations. Thus, $\mathrm{D}_{\mathsf{B}}^{\Gamma}(p\|q) = \mathrm{D}_{\mathsf{B}}^{\phi}(p\|q)$ is valid for any affine transformation

$$\Gamma(q) = \phi(q) + \Psi_p[q] + c \tag{4.3}$$

$$\text{with linear operator } \Psi_p[q] = \frac{\delta \Gamma(p)}{\delta p} \cdot q - \frac{\delta \phi(p)}{\delta p} \cdot q \tag{4.4}$$

for positive measures p and q and scalar c.

5. **Three-point property:** For any triple p, q, ρ of positive measures the property holds:
$$D^{\phi}_{\mathsf{B}}(p\|\rho) = D^{\phi}_{\mathsf{B}}(p\|q) + D^{\phi}_{\mathsf{B}}(q\|\rho) + (p-q)\left(\frac{\delta\phi(q)}{\delta q} - \frac{\delta\phi(\rho)}{\delta\rho}\right) \tag{4.5}$$

6. **Generalized Pythagorean theorem:** Let $P_{\Omega}(q) = \underset{\omega\in\Omega}{\arg\min}\mathrm{D}^{\phi}_{\mathsf{B}}(\omega\|q)$ be the Bregman projection onto the convex set Ω and $p \in \Omega$. The inequality
$$\mathrm{D}^{\phi}_{\mathsf{B}}(p\|q) \geq \mathrm{D}^{\phi}_{\mathsf{B}}(p\|P_{\Omega}(q)) + \mathrm{D}^{\phi}_{\mathsf{B}}(P_{\Omega}(q)\|q) \tag{4.6}$$
is known as generalized Pythagorean theorem. If Ω is an affine set it holds with equality.

7. **Optimality:** In [4] an optimality property is stated. Given a set S of positive measures p with mean $\mu = E[S]$ and $\mu \in S$ the unique minimizer $E_{p\in S}[\mathrm{D}(p\|q)]$ is minimum for $q = \mu$ if D is a Bregman divergence. This property favors the Bregman divergences for clustering problems.

The Fréchet-derivative of $\mathrm{D}^{\phi}_{\mathsf{B}}$ with respect to q is formally given by
$$\frac{\delta \mathrm{D}^{\phi}_{\mathsf{B}}(p\|q)}{\delta q} = \frac{\delta\phi(p)}{\delta q} - \frac{\delta\phi(q)}{\delta q} - \frac{\delta\left[\frac{\delta\phi(q)}{\delta q}(p-q)\right]}{\delta q} \tag{4.7}$$

with
$$\frac{\delta\left[\frac{\delta\phi(q)}{\delta q}(p-q)\right]}{\delta q} = \frac{\delta^2[\phi(q)]}{\delta q^2}(p-q) - \frac{\delta\phi(q)}{\delta q} \quad .$$

In the following we will provide detailed information and the Fréchet derivatives for some special cases and subsets of the Bregman divergences.

4.1.1 Generalized Kullback-Leibler Divergence (I-Divergence):

A famous example for Bregman divergence [45] is the generalized Kullback-Leibler divergence:
$$\mathrm{D}_{\mathsf{GKL}}(p\|q) = \int p(r)\ln\left(\frac{p(r)}{q(r)}\right)\,dr - \int [p(r) - q(r)]\,dr \tag{4.8}$$

with the generating function
$$\phi(p) = \int p\ln p - p\,dr \quad . \tag{4.9}$$

The Fréchet-derivative of $\mathrm{D}_{\mathsf{GKL}}$ with respect to q is given by
$$\frac{\delta \mathrm{D}_{\mathsf{GKL}}(p\|q)}{\delta q} = 1 - \frac{p}{q} \quad . \tag{4.10}$$

So the learning rule for SONE with the generalized KL divergence equals the NE-XOM given by Eq. (2.13).

4.1.2 Itakura-Saito:

The Itakura-Saito (IS) divergence:

$$\mathrm{D}_{\mathsf{IS}}(p\|q) = \int \left(\frac{p}{q} - \ln\left(\frac{p}{q}\right) - 1 \right) \, dr \tag{4.11}$$

was derived in 1968 from the maximum likelihood (ML) estimation of short-time speech spectra [25]. This divergence is a Bregman divergence based on the Burg entropy

$$H_{\mathsf{B}}(p) = -\int \ln(p) \, dr \tag{4.12}$$

which also serves as the generating function

$$\phi(p) = H_{\mathsf{B}}(p) \ . \tag{4.13}$$

Due to the good perceptual properties of the reconstructed signals this divergence became a standard measure in speech processing. Besides the properties that the IS divergence inherits from the Bregman divergences, it is scale invariant $\mathrm{D}_{\mathsf{IS}}(c \cdot p \| c \cdot q) = \mathrm{D}_{\mathsf{IS}}(p\|q)$, meaning that low energy components of p bear the same relative importance as high energy ones. As a consequence the Itakura-Saito divergence is frequently applied in image and sound processing [7].

The Fréchet-derivative of D_{IS} with respect to q is given by

$$\frac{\delta \mathrm{D}_{\mathsf{IS}}(p\|q)}{\delta q} = \frac{1}{q} - \frac{p}{q^2} \ . \tag{4.14}$$

The learning rules for SONE with Itakura-Saito divergence are

in case of a Gaussian g:

$$\boldsymbol{y}^k = \boldsymbol{y}^k - \tau \frac{1}{2\varsigma^2} \left(\frac{h_\sigma^{\Psi_{\mathsf{IS}}(\boldsymbol{s})}(k)}{g_\varsigma^{\boldsymbol{s}}(k)} - 1 \right) \frac{\partial d_{\mathcal{E}}(\boldsymbol{s}, \boldsymbol{y}^k)}{\partial \boldsymbol{y}^k} \tag{4.15}$$

and in case of a t-Distribution g:

$$\boldsymbol{y}^k = \boldsymbol{y}^k - \tau \frac{\varsigma + 1}{2\varsigma} \frac{1}{(1 + d_{\mathcal{E}}(\boldsymbol{s}, \boldsymbol{y}^k)/\varsigma)} \left(\frac{h_\sigma^{\Psi_{\mathsf{IS}}(\boldsymbol{s})}(k)}{g_\varsigma^{\boldsymbol{s}}(k)} - 1 \right) \frac{\partial d_{\mathcal{E}}(\boldsymbol{s}, \boldsymbol{y}^k)}{\partial \boldsymbol{y}^k} \tag{4.16}$$

4.1.3 Beta-Divergences:

The β-divergence was introduced as density power divergences by Basu [5], Mihoko and Eguchi [20, 34]. It is not invariant under a change of the dominating measure and not invariance monotone for summarization, except for the special case $\beta = 1$ which gives the KL-divergence. The β-divergence for positive measures p and q is defined as:

$$\mathrm{D}_\beta(p\|q) = \int p \cdot \frac{p^{\beta-1} - q^{\beta-1}}{\beta - 1} \, dr - \int \frac{p^\beta - q^\beta}{\beta} \, dr \quad \text{with} \begin{cases} \beta \neq 1 \\ \beta \neq 0 \end{cases} \tag{4.17}$$

$$= \int p^\beta \left(\frac{1}{\beta - 1} - \frac{1}{\beta} \right) - q^{\beta-1} \left(\frac{p}{\beta - 1} - \frac{q}{\beta} \right) \, dr \ . \tag{4.18}$$

In the case of $\beta = 2$ we obtain the standard squared Euclidean distance, while the limit $\beta \to 1$ leads to the generalized Kullback-Leibler divergence (I-divergence) and the limit $\beta \to 0$ gives the Itakura-Saito distance.
The Fréchet-derivative of D_β with respect to q is given by

$$\frac{\delta \mathrm{D}_\beta(p\|q)}{\delta q} = -p \cdot q^{\beta-2} + q^{\beta-1} = q^{\beta-2}(q-p) \ . \tag{4.19}$$

So the learning rules for SONE with the β-divergence are

in case of a Gaussian g:

$$\boldsymbol{y}^k = \boldsymbol{y}^k - \frac{\tau}{2\varsigma^2} \cdot g_\varsigma^{\boldsymbol{s}}(k)^{(\beta-1)} \left(h_\sigma^{\Psi_\beta(\boldsymbol{s})}(k) - g_\varsigma^{\boldsymbol{s}}(k) \right) \frac{\partial d_{\mathcal{E}}(\boldsymbol{s}, \boldsymbol{y}^k)}{\partial \boldsymbol{y}^k} \tag{4.20}$$

and in case of a t-Distribution g:

$$\boldsymbol{y}^k = \boldsymbol{y}^k - \frac{\varsigma+1}{2\varsigma} \frac{\tau}{(1 + d_{\mathcal{E}}(\boldsymbol{s}, \boldsymbol{y}^k)/\varsigma)} \cdot g_\varsigma^{\boldsymbol{s}}(k)^{(\beta-1)} \left(h_\sigma^{\Psi_\beta(\boldsymbol{s})}(k) - g_\varsigma^{\boldsymbol{s}}(k) \right) \frac{\partial d_{\mathcal{E}}(\boldsymbol{s}, \boldsymbol{y}^k)}{\partial \boldsymbol{y}^k} \tag{4.21}$$

4.2 Generalized Csizár f-divergence:

We denote by $\mathcal{F}$ the class of convex functions satisfying $f(1) = 1, f'(1) = 0,$ and $f''(1) = 1$. For a convex function $f \in \mathcal{F}$ the Csizár f-divergence is given by:

$$\mathrm{D}_f(p\|q) = \int q \ f\left(\frac{p}{q}\right) \ dr \tag{4.22}$$

with the definitions $0 \cdot f\left(\frac{0}{0}\right) = 0$ and $0 \cdot f\left(\frac{a}{0}\right) = \lim_{x\to 0} x \cdot f(\frac{a}{x}) = \lim_{u\to\infty} a \cdot \frac{f(u)}{u}$ [15, 16, 2]. The f-divergence can be interpreted as an average of the likelihood ratio $\frac{p}{q}$ describing the change rate of p with respect to q weighted by the determining function f. For a general f, which does not have to be convex, with $f'(1) = c_f \neq 0$, this form is not invariant and we need to use

$$\mathrm{D}_f(p\|q) = c_f \int (p-q) \ dr + \int q \ f\left(\frac{p}{q}\right) \ dr \ . \tag{4.23}$$

Some basic properties of the Csiszár f-divergence are [37, 14]:

1. **Non-negativity:** $\mathrm{D}_f(p\|q) \geq 0$ and equals zero iff $p \equiv q$ follows from the Jensens inequality,

2. **Generalized entropy:** It corresponds to a generalized f-entropy if the form

$$H_f(p) = -\int f(p(r)) \ dr \ . \tag{4.24}$$

3. **Strict convexity:** The f-divergence is convex in both arguments p and q:

$$\mathrm{D}_f(tp_1 + (1-t)p_2 \| tq_1 + (1-t)q_2) \leq t\mathrm{D}_f(\tilde{p_1}\|\tilde{q_1}) + (1-t)\mathrm{D}_f(p_2\|q_2) \quad \forall t \in [0,1] \tag{4.25}$$

4. **Scalability:** $c\mathrm{D}_f(p\|q) = \mathrm{D}_{cf}(p\|q)$ for any positive constant $c > 0$.

5. **Invariance:** $\mathrm{D}_f(p\|q)$ is invariant with respect to a linear shift regarding the function f: e. g. $\mathrm{D}_f(p\|q) = \mathrm{D}_{\tilde{f}}(p\|q)$ iff $\tilde{f}(u) = f(u) + c \cdot (u-1)$ for any constant $c \in \mathbb{R}$.

6. **Symmetry:** For $f, f^* \in \mathcal{F}$, where $f^*(u) = u \cdot f(\frac{1}{u})$ denotes the conjugate function of f, the relation $\mathrm{D}_f(p\|q) = \mathrm{D}_{f^*}(q\|p)$ is valid. It is possible to construct a symmetric Csizár f-divergence with $f_{\text{sym}}(u) = f(u) + f^*(u)$ as determining function.

7. **Upper bound:** The f-divergence is bounded by

$$0 \leq \mathrm{D}_f(p\|q) \leq \lim_{u \to 0^+} \{f(u) + f^*(u)\} \text{ with } u = \frac{p}{q} \ . \tag{4.26}$$

   The existence of this limit for probability densities p and q was shown by Liese and Vajda in [32]. Villmann and Haase showed that these bound still holds for positive measures p and q [44].

8. **Monotonicity:** The f-divergence is monotonic with respect to the coarse-graining of the underlying domain $\mathcal{D}$ of the positive measures p and q, which is similar to the monotonicity of the Fisher metric [2].

The Fréchet derivative for the f-divergence is given by [45]:

$$\frac{\delta \mathrm{D}_f(p\|q)}{\delta q} = f\left(\frac{p}{q}\right) + q \frac{\partial f(u)}{\partial u} \frac{\partial u}{\partial q} \quad \text{with } u = \frac{p}{q} \tag{4.27}$$

$$= f\left(\frac{p}{q}\right) + q \frac{\partial f(u)}{\partial u} \cdot \frac{-p}{q^2} \tag{4.28}$$

4.2.1 Alpha-Divergence:

The α-divergences can be derived from the Csiszár f-divergences as well as from the Bregman divergence. It is a special case of Csizár f-divergences associated to any function $f(u)$ which is convex over $(0, \infty)$ and satisfies $f(1) = 0$. Indeed the generating function $f(u) = (u^\alpha - \alpha u + \alpha - 1)/(\alpha^2 - \alpha)$ inserted in $\mathrm{D}_f(p\|q)$ Eq. (4.22) yields the basic, asymmetric α-divergence:

$$\mathrm{D}_\alpha(p\|q) = \frac{1}{\alpha(\alpha-1)} \int \left[p^\alpha q^{1-\alpha} - \alpha p + (\alpha - 1)\, q\right] dr, \quad \alpha \in \mathbb{R} \tag{4.29}$$

$$= \frac{1}{\alpha(\alpha-1)} \int \left[q\left(\frac{p^\alpha}{q^\alpha} + (\alpha - 1)\right) - \alpha p\right] dr \ . \tag{4.30}$$

where p and q not need to be normalized. With the parameter α this divergence connects the I-divergence $\mathrm{D}_{\text{GKL}}(p\|q)$ (limiting case: $\lim_{\alpha \to 1} \mathrm{D}_\alpha(p\|q)$) with the dual I-divergence $\mathrm{D}_{\text{GKL}}(q\|p)$ (limiting case: $\lim_{\alpha \to 0} \mathrm{D}_\alpha(p\|q)$). Further β-divergences can be generated from the α-divergences, by applying a non-linear transformation: $p \to p^{\beta+2}$ and $q \to q^{\beta+2}$ with $\alpha = 1/(\beta + 1)$ [14]. Moreover, the α-divergences are closely related to the generalized Rényi-divergence [14, 1]:

$$\mathrm{D}^\alpha_{\text{GR}}(p\|q) = \frac{1}{\alpha - 1} \ln \left(\int \left[p^\alpha q^{1-\alpha} - \alpha p + (\alpha - 1)\, q\right] dr + 1 \right) \ . \tag{4.31}$$

The family of α-divergences can be converted to a wide range of divergences defined for probability densities, like the Hellinger, Pearson Chi-squared, Neyman Chi-squared (or inverse Pearson), Rényi and Tsallis divergence. Since we are dealing with positive measures here the interested reader is referred to [45, 14] for details. In practice a regularized version of the α-divergence with an additional penalty term on the right hand side:

$$\mathrm{D}_{\alpha}(p\|q) = \frac{1}{\alpha(\alpha-1)}\int_{p\neq 0}\left[p^{\alpha}q^{1-\alpha} - \alpha p + (\alpha-1)\,q\right]dr + \frac{1}{\alpha}\int_{p=0} q\,dr \tag{4.32}$$

often improves performance and robustness of estimators.

In addition to the general properties of the f-divergences stated above, the α-divergences exhibit specific characteristics:

1. **Continuity:** The α-divergence is a continuous function of the real variable α in the whole range including singularities.

2. **Duality:** $\mathrm{D}_{(\alpha)}(p\|q) = \mathrm{D}_{(1-\alpha)}(p\|q)$

3. **Properties depending on the Hyper-parameter:** [35]

$\alpha \to -\infty$: the estimation q may exclude modes of the target p. So the minimization of $\mathrm{D}_{\alpha}(p\|q)$ with respect to q will force the mass of q lying within p.

$\alpha \leq 0$: the estimation is zero-forcing, i. e. $p(r) = 0$ forces $q(r) = 0$.

$\alpha \geq 1$: the estimation is zero-avoiding, i. e. $q(r) > 0$ implies $q(r) > 0$.

$\alpha \to \infty$: the α-divergence is inclusive, i. e. q covers p completely.

The Fréchet derivatives of the subset of α-divergences and the generalized Rényi-divergence are [45]:

$$\frac{\delta \mathrm{D}_{\alpha}(p\|q)}{\delta q} = \frac{1}{\alpha}\left(1 - \frac{p^{\alpha}}{q^{\alpha}}\right) \tag{4.33}$$

$$\frac{\delta \mathrm{D}_{\mathsf{GR}}(p\|q)}{\delta q} = \frac{1 - p^{\alpha}q^{-\alpha}}{\int\left[p^{\alpha}q^{1-\alpha} - \alpha p + (\alpha-1)\,q + 1\right]dr}\ . \tag{4.34}$$

Hence the learning rules for SONE with the α-divergence are:

in case of a Gaussian g:

$$\boldsymbol{y}^{k} = \boldsymbol{y}^{k} - \tau\frac{1}{2\varsigma^{2}} \cdot \frac{g_{\varsigma}^{\boldsymbol{s}}(k)}{\alpha}\left(\left(\frac{h_{\sigma}^{\Psi_{\alpha}(\boldsymbol{s})}(k)}{g_{\varsigma}^{\boldsymbol{s}}(k)}\right)^{\alpha} - 1\right)\frac{\partial d_{\mathcal{E}}(\boldsymbol{s},\boldsymbol{y}^{k})}{\partial \boldsymbol{y}^{k}} \tag{4.35}$$

and in case of a t-Distribution g:

$$\boldsymbol{y}^{k} = \boldsymbol{y}^{k} - \tau\frac{\varsigma+1}{2\varsigma}\frac{1}{(1 + d_{\mathcal{E}}(\boldsymbol{s},\boldsymbol{y}^{k})/\varsigma)} \cdot \frac{g_{\varsigma}^{\boldsymbol{s}}(k)}{\alpha}\left(\left(\frac{h_{\sigma}^{\Psi_{\alpha}(\boldsymbol{s})}(k)}{g_{\varsigma}^{\boldsymbol{s}}(k)}\right)^{\alpha} - 1\right)\frac{\partial d_{\mathcal{E}}(\boldsymbol{s},\boldsymbol{y}^{k})}{\partial \boldsymbol{y}^{k}}\ . \tag{4.36}$$

4.3 Gamma-Divergence:

The γ-divergence is a very robust dissimilarity measure with respect to outliers [21] proposed by Fujisawa and Eguchi:

$$\mathrm{D}_\gamma(p\|q) = \ln\left[\frac{\left(\int p^{\gamma+1}\,dr\right)^{\frac{1}{\gamma(\gamma+1)}}\cdot\left(\int q^{\gamma+1}\,dr\right)^{\frac{1}{\gamma+1}}}{\left(\int p\cdot q^{\gamma}\,dr\right)^{\frac{1}{\gamma}}}\right] \tag{4.37}$$

It is robust for $\gamma\in[0,1]$ with existence of $\mathrm{D}_{\gamma=0}$ in the limit $\gamma\to 0$. In the limit $\gamma\to 0$ the γ-divergence becomes the Kullback-Leibler-divergence $\mathrm{D}_{\mathsf{KL}}(p\|q)$ for probability densities. And for $\gamma=1$ it becomes the Cauchy-Schwarz divergence $\mathrm{D}_{\mathsf{CS}}(p\|q)$, which is based on the quadratic Rényi-entropy and is frequently applied for Parzen window estimation, especially suitable for spectral clustering as well as related graph cut problems [38, 26, 27, 45].

The γ-divergence displays some nice properties [45]:

1. **Invariance:** $\mathrm{D}_\gamma(p\|q)$ is invariant under scalar multiplication with positive constants
$$\mathrm{D}_\gamma(p\|q) = \mathrm{D}_\gamma(c_1\cdot p\|c_2\cdot q)\quad \forall c_1,c_2>0\ . \tag{4.38}$$
In case of positive measures the equation $\mathrm{D}_\gamma(p\|q)=0$ holds only if $p=c\cdot q$ with $c>0$. For probability densities $c=1$ is required.

2. **Pythagorean relation:** As for Bregman divergences a modified Pythagorean relation between positive measures can be stated for special choices of p,q,ρ. Let p be a distortion of q defined as convex combination with a positive distortion measure $\phi(r)$
$$p_\varepsilon(r) = (1-\varepsilon)\cdot q(r)+\varepsilon\cdot\phi(r)\ . \tag{4.39}$$
A positive measure g is denoted as ϕ-consistant if $\nu_g=\left(\int\phi(r)g(r)^{\alpha}\,dr\right)^{\frac{1}{\alpha}}$ is sufficiently small for large $\alpha>0$. If two positive measures q and ρ are ϕ-consistant according to a distortion measure ϕ, then the Pythagorean relation approximately holds for q,ρ and the distortion p_ε of q:
$$\Delta(p_\varepsilon,q,\rho) = \mathrm{D}_\gamma(p_\varepsilon\|\rho)-\mathrm{D}_\gamma(p_\varepsilon\|q)-\mathrm{D}_\gamma(q\|\rho) = \mathcal{O}(\varepsilon\nu^{\gamma})\text{ with }\nu=\max\{\nu_q,\nu_\rho\}\ . \tag{4.40}$$
This property implies the robustness of D_γ according to distortions.

The Fréchet derivative of D_γ with respect to q yields [45]:

$$\frac{\delta\mathrm{D}_\gamma(p\|q)}{\delta q} = q^{\gamma-1}\left[\frac{q}{\int q^{\gamma+1}\,dr}-\frac{p}{\int p\cdot q^{\gamma}\,dr}\right] \tag{4.41}$$

The gradients for SONE with the γ-divergence and winner definition

$$\Psi_\gamma(\boldsymbol{s}) = \boldsymbol{x}^i\text{ such that }\sum_j \mathrm{D}_\gamma\left(h_\sigma^{\Psi_\gamma(\boldsymbol{s})}(j)\middle\|g_\varsigma^{\boldsymbol{s}}(j)\right)\text{ is minimum,} \tag{4.42}$$

leads the following learning rules. In case of a Gaussian neighborhood function g:

$$\Delta \boldsymbol{y}^k = \left[g_\varsigma^{\boldsymbol{s}}(k)\right]^{\gamma-1}\left[\frac{g_\varsigma^{\boldsymbol{s}}(k)}{\sum_j \left[g_\varsigma^{\boldsymbol{s}}(j)\right]^{\gamma+1}} - \frac{h_\sigma^{\Psi_\gamma(\boldsymbol{s})}(k)}{\sum_j h_\sigma^{\Psi_\gamma(\boldsymbol{s})}(j)\cdot\left[g_\varsigma^{\boldsymbol{s}}(j)\right]^{\gamma}}\right]\left(-\frac{g_\varsigma^{\boldsymbol{s}}(k)}{2\varsigma^2}\right)\frac{\partial d_{\mathcal{E}}(\boldsymbol{s},\boldsymbol{y}^k)}{\partial \boldsymbol{y}^k} \tag{4.43}$$

$$= -\frac{\left[g_\varsigma^{\boldsymbol{s}}(k)\right]^{\gamma}}{2\varsigma^2}\left[\frac{g_\varsigma^{\boldsymbol{s}}(k)}{\sum_j \left[g_\varsigma^{\boldsymbol{s}}(j)\right]^{\gamma+1}} - \frac{h_\sigma^{\Psi_\gamma(\boldsymbol{s})}(k)}{\sum_j h_\sigma^{\Psi_\gamma(\boldsymbol{s})}(j)\cdot\left[g_\varsigma^{\boldsymbol{s}}(j)\right]^{\gamma}}\right]\frac{\partial d_{\mathcal{E}}(\boldsymbol{s},\boldsymbol{y}^k)}{\partial \boldsymbol{y}^k} \tag{4.44}$$

and in case of t-Distribution:

$$\Delta \boldsymbol{y}^k = -\left(\frac{\varsigma+1}{2}\right)\frac{\left[g_\varsigma^{\boldsymbol{s}}(k)\right]^{\gamma}}{\varsigma + d_{\mathcal{E}}(\boldsymbol{s},\boldsymbol{y}^k)}\left[\frac{g_\varsigma^{\boldsymbol{s}}(k)}{\sum_j \left[g_\varsigma^{\boldsymbol{s}}(j)\right]^{\gamma+1}} - \frac{h_\sigma^{\Psi_\gamma(\boldsymbol{s})}(k)}{\sum_j h_\sigma^{\Psi_\gamma(\boldsymbol{s})}(j)\cdot\left[g_\varsigma^{\boldsymbol{s}}(j)\right]^{\gamma}}\right]\frac{\partial d_{\mathcal{E}}(\boldsymbol{s},\boldsymbol{y}^k)}{\partial \boldsymbol{y}^k}. \tag{4.45}$$

4.3.1 Cauchy-Schwarz-Divergence:

The Cauchy-Schwarz-divergence

$$\mathrm{D}_{\mathsf{CS}}(p\|q) = \frac{1}{2}\ln\left(\int q^2\, dr \cdot \int p^2\, dr\right) - \ln\left(\int p\cdot q\, dr\right) \tag{4.46}$$

was introduced by J. Principe considering the Cauchy-Schwarz-inequality for norms [38]. It follows as the special case of $\gamma = 1$ in the γ-divergence explained above and it is based on the quadratic Rényi-entropy. This divergence is frequently used for Parzen window estimation and particularly suitable for spectral clustering as well as for related graph cut problems [27].

The Fréchet-derivative of the Cauchy-Schwarz-divergence is derived [44]:

$$\frac{\delta \mathrm{D}_{\mathsf{CS}}(p\|q)}{\delta q} = \frac{q}{\int q^2\, dr} - \frac{p}{\int p\cdot q\, dr}\quad . \tag{4.47}$$

The gradients for SONE with the Cauchy-Schwarz-divergence and winner definition

$$\Psi_{\mathsf{CS}}(\boldsymbol{s}) = \boldsymbol{x}^i \text{ such that } \sum_j \mathrm{D}_{\mathsf{CS}}\left(h_\sigma^{\Psi_{\mathsf{CS}}(\boldsymbol{s})}(j)\middle\| g_\varsigma^{\boldsymbol{s}}(j)\right) \text{ is minimum,} \tag{4.48}$$

leads the following learning rules. In case of a Gaussian neighborhood function g:

$$\Delta \boldsymbol{y}^k = \left[\frac{g_\varsigma^{\boldsymbol{s}}(k)}{\sum_j \left[g_\varsigma^{\boldsymbol{s}}(j)\right]^{2}} - \frac{h_\sigma^{\Psi_{\mathsf{CS}}(\boldsymbol{s})}(k)}{\sum_j h_\sigma^{\Psi_{\mathsf{CS}}(\boldsymbol{s})}(j)\cdot g_\varsigma^{\boldsymbol{s}}(j)}\right]\left(-\frac{g_\varsigma^{\boldsymbol{s}}(k)}{2\varsigma^2}\right)\frac{\partial d_{\mathcal{E}}(\boldsymbol{s},\boldsymbol{y}^k)}{\partial \boldsymbol{y}^k} \tag{4.49}$$

and in case of t-Distribution:

$$\Delta \boldsymbol{y}^k = -\left(\frac{\varsigma+1}{2}\right)\frac{g_\varsigma^{\boldsymbol{s}}(k)}{\varsigma + d_{\mathcal{E}}(\boldsymbol{s},\boldsymbol{y}^k)}\left[\frac{g_\varsigma^{\boldsymbol{s}}(k)}{\sum_j \left[g_\varsigma^{\boldsymbol{s}}(j)\right]^{2}} - \frac{h_\sigma^{\Psi_{\mathsf{CS}}(\boldsymbol{s})}(k)}{\sum_j h_\sigma^{\Psi_{\mathsf{CS}}(\boldsymbol{s})}(j)\cdot g_\varsigma^{\boldsymbol{s}}(j)}\right]\frac{\partial d_{\mathcal{E}}(\boldsymbol{s},\boldsymbol{y}^k)}{\partial \boldsymbol{y}^k}. \tag{4.50}$$

4.4 Summary

This section is a summary and collection of equations for all the divergences, their derivatives and learning rules for the SONE algorithm. The generalized SONE learning rule is defined as:

$$\begin{aligned}\boldsymbol{y}^k &= \boldsymbol{y}^k - \tau \Delta \boldsymbol{y}^k \\ \Delta \boldsymbol{y}^k &= \left.\frac{\delta \mathrm{D}\left(h_\sigma^{\Psi_\mathrm{D}(\boldsymbol{s})} \| g_\varsigma^{\boldsymbol{s}}\right)}{\delta g_\varsigma^{\boldsymbol{s}}}\right|_k \cdot \frac{\partial g_\varsigma^{\boldsymbol{s}}(k)}{\partial \boldsymbol{y}^k}\end{aligned}$$

with the winner definition

$$\Psi_\mathrm{D}(\boldsymbol{s}) = \boldsymbol{x}^i \text{ such that } \sum_j \mathrm{D}\left(h_\sigma^{\Psi_\mathrm{D}(\boldsymbol{s})}(j) \| g_\varsigma^{\boldsymbol{s}}(j)\right) \text{ is minimum.}$$

The explicit formulas for the special learning rules in case of Gaussian (Eq. (2.6),(2.7)) and t-distribution (Eq. (2.8),(2.9)) and different divergences can be found in table 1.

Table 1: Summary of the divergences for positive measures and the SONE learning rules.

Divergence	Formula / Fréchet-derivative	Learning rules $\begin{cases}\text{Gaussian}\\\text{t-Distribution}\end{cases}$
Bregman	$D_{\mathsf{B}}^{\phi}(p\Vert q)=\phi(p)-\phi(q)-\frac{\delta\phi(q)}{\delta q}[p-q]$	
	$\frac{\delta \mathrm{D}_{\mathsf{B}}^{\phi}(p\Vert q)}{\delta q}=\frac{\delta\phi(p)}{\delta q}-\frac{\delta\phi(q)}{\delta q}-\frac{\delta^2[\phi(q)]}{\delta q^2}(p-q)+\frac{\delta\phi(q)}{\delta q}$	
Generalized	$\mathrm{D}_{\mathsf{GKL}}(p\Vert q)=\int p\ln\left(\frac{p}{q}\right)\,dr-\int[p-q]\,dr$	$\Delta\boldsymbol{y}^k=\frac{1}{2\varsigma^2}\left(h_\sigma^{\Psi_{\mathsf{GKL}}(\boldsymbol{s})}(k)-g_\varsigma^{\boldsymbol{s}}(k)\right)\frac{\partial d_{\mathcal{E}}(\boldsymbol{s},\boldsymbol{y}^k)}{\partial\boldsymbol{y}^k}$
Kullback-Leibler	$\frac{\delta \mathrm{D}_{\mathsf{GKL}}(p\Vert q)}{\delta q}=1-\frac{p}{q}$	$\Delta\boldsymbol{y}^k=\frac{\varsigma+1}{2\varsigma}\frac{1}{(1+d_{\mathcal{E}}(\boldsymbol{s},\boldsymbol{y}^k)/\varsigma)}\left(h_\sigma^{\Psi_{\mathsf{GKL}}(\boldsymbol{s})}(k)-g_\varsigma^{\boldsymbol{s}}(k)\right)\frac{\partial d_{\mathcal{E}}(\boldsymbol{s},\boldsymbol{y}^k)}{\partial\boldsymbol{y}^k}$
Itakura-Saito	$\mathrm{D}_{\mathsf{IS}}(p\Vert q)=\int\left(\frac{p}{q}-\ln\left(\frac{p}{q}\right)-1\right)$	$\Delta\boldsymbol{y}^k=\frac{1}{2\varsigma^2}\left(\frac{h_\sigma^{\Psi_{\mathsf{IS}}(\boldsymbol{s})}(k)}{g_\varsigma^{\boldsymbol{s}}(k)}-1\right)\frac{\partial d_{\mathcal{E}}(\boldsymbol{s},\boldsymbol{y}^k)}{\partial\boldsymbol{y}^k}$
	$\frac{\delta \mathrm{D}_{\mathsf{IS}}(p\Vert q)}{\delta q}=\frac{1}{q}-\frac{p}{q^2}$	$\Delta\boldsymbol{y}^k=\frac{\varsigma+1}{2\varsigma}\frac{1}{(1+d_{\mathcal{E}}(\boldsymbol{s},\boldsymbol{y}^k)/\varsigma)}\left(\frac{h_\sigma^{\Psi_{\mathsf{IS}}(\boldsymbol{s})}(k)}{g_\varsigma^{\boldsymbol{s}}(k)}-1\right)\frac{\partial d_{\mathcal{E}}(\boldsymbol{s},\boldsymbol{y}^k)}{\partial\boldsymbol{y}^k}$
Beta-divergence	$\mathrm{D}_\beta(p\Vert q)=\int p\cdot\frac{p^{\beta-1}-q^{\beta-1}}{\beta-1}\,dr-\int\frac{p^\beta-q^\beta}{\beta}\,dr$	$\Delta\boldsymbol{y}^k=\frac{1}{2\varsigma^2}\cdot g_\varsigma^{\boldsymbol{s}}(k)^{(\beta-1)}\left(h_\sigma^{\Psi_\beta(\boldsymbol{s})}(k)-g_\varsigma^{\boldsymbol{s}}(k)\right)\frac{\partial d_{\mathcal{E}}(\boldsymbol{s},\boldsymbol{y}^k)}{\partial\boldsymbol{y}^k}$
	$\frac{\delta \mathrm{D}_\beta(p\Vert q)}{\delta q}=q^{\beta-2}(q-p)$	$\Delta\boldsymbol{y}^k=\frac{\varsigma+1}{2}\frac{1}{\varsigma+d_{\mathcal{E}}(\boldsymbol{s},\boldsymbol{y}^k)}\cdot g_\varsigma^{\boldsymbol{s}}(k)^{(\beta-1)}\left(h_\sigma^{\Psi_\beta(\boldsymbol{s})}(k)-g_\varsigma^{\boldsymbol{s}}(k)\right)\frac{\partial d_{\mathcal{E}}(\boldsymbol{s},\boldsymbol{y}^k)}{\partial\boldsymbol{y}^k}$
Generalized	$\mathrm{D}_f(p\Vert q)=\int q\,f\left(\frac{p}{q}\right)\,dr$	
Csizár f	$\frac{\delta \mathrm{D}_f(p\Vert q)}{\delta q}=f\left(\frac{p}{q}\right)+q\frac{\partial f(u)}{\partial u}\cdot\frac{-p}{q^2}\,,\,u=\frac{p}{q}$	
Alpha	$\mathrm{D}_\alpha(p\Vert q)=\frac{1}{\alpha(\alpha-1)}\int[p^\alpha q^{1-\alpha}-\alpha p+(\alpha-1)\,q]\,dr$	$\Delta\boldsymbol{y}^k=\frac{1}{2\varsigma^2}\cdot\frac{g_\varsigma^{\boldsymbol{s}}(k)}{\alpha}\left(\left(\frac{h_\sigma^{\Psi_\alpha(\boldsymbol{s})}(k)}{g_\varsigma^{\boldsymbol{s}}(k)}\right)^\alpha-1\right)\frac{\partial d_{\mathcal{E}}(\boldsymbol{s},\boldsymbol{y}^k)}{\partial\boldsymbol{y}^k}$
-divergence	$\frac{\delta \mathrm{D}_\alpha(p\Vert q)}{\delta q}=\frac{1}{\alpha}\left(1-\frac{p^\alpha}{q^\alpha}\right)$	$\Delta\boldsymbol{y}^k=\frac{\varsigma+1}{2\varsigma}\frac{1}{(1+d_{\mathcal{E}}(\boldsymbol{s},\boldsymbol{y}^k)/\varsigma)}\cdot\frac{g_\varsigma^{\boldsymbol{s}}(k)}{\alpha}\left(\left(\frac{h_\sigma^{\Psi_\alpha(\boldsymbol{s})}(k)}{g_\varsigma^{\boldsymbol{s}}(k)}\right)^\alpha-1\right)\frac{\partial d_{\mathcal{E}}(\boldsymbol{s},\boldsymbol{y}^k)}{\partial\boldsymbol{y}^k}$
Generalized	$\mathrm{D}_{\mathsf{GR}}^\alpha(p\Vert q)=\frac{1}{\alpha-1}\ln(\int[p^\alpha q^{1-\alpha}-\alpha p+(\alpha-1)q]\,dr+1)$	
Rényi	$\frac{\delta \mathrm{D}_{\mathsf{GR}}(p\Vert q)}{\delta q}=\frac{1-p^\alpha q^{-\alpha}}{\int[p^\alpha q^{1-\alpha}-\alpha p+(\alpha-1)\;q+1]\;dr}$	

Gamma	$\mathrm{D}_{\gamma}(p\Vert q)=\ln\left[\frac{\left(\int p^{\gamma+1}\ dr\right)^{\frac{1}{\gamma(\gamma+1)}}\cdot\left(\int q^{\gamma+1}\ dr\right)^{\frac{1}{\gamma+1}}}{\left(\int p\cdot q^{\gamma}\ dr\right)^{\frac{1}{\gamma}}}\right]$	$\Delta\boldsymbol{y}^{k}=-\frac{\left[g_{\varsigma}^{\boldsymbol{s}}(k)\right]^{\gamma}}{2\varsigma^{2}}\left[\frac{g_{\varsigma}^{\boldsymbol{s}}(k)}{\sum_{j}\left[g_{\varsigma}^{\boldsymbol{s}}(j)\right]^{\gamma+1}}-\frac{h_{\sigma}^{\Psi_{\gamma}(\boldsymbol{s})}(k)}{\sum_{j}h_{\sigma}^{\Psi_{\gamma}(\boldsymbol{s})}(j)\cdot\left[g_{\varsigma}^{\boldsymbol{s}}(j)\right]^{\gamma}}\right]\frac{\partial d_{\mathcal{E}}(\boldsymbol{s},\boldsymbol{y}^{k})}{\partial\boldsymbol{y}^{k}}$
-divergence	$\frac{\delta\mathrm{D}_{\gamma}(p\Vert q)}{\delta q}=q^{\gamma-1}\left[\frac{q}{\int q^{\gamma+1}\ dr}-\frac{p}{\int p\cdot q^{\gamma}\ dr}\right]$	$\Delta\boldsymbol{y}^{k}=\frac{-(\varsigma+1)\cdot\left[g_{\varsigma}^{\boldsymbol{s}}(k)\right]^{\gamma}}{2(\varsigma+d_{\mathcal{E}}(\boldsymbol{s},\boldsymbol{y}^{k}))}\left[\frac{g_{\varsigma}^{\boldsymbol{s}}(k)}{\sum_{j}\left[g_{\varsigma}^{\boldsymbol{s}}(j)\right]^{\gamma+1}}-\frac{h_{\sigma}^{\Psi_{\gamma}(\boldsymbol{s})}(k)}{\sum_{j}h_{\sigma}^{\Psi_{\gamma}(\boldsymbol{s})}(j)\left[g_{\varsigma}^{\boldsymbol{s}}(j)\right]^{\gamma}}\right]\frac{\partial d_{\mathcal{E}}(\boldsymbol{s},\boldsymbol{y}^{k})}{\partial\boldsymbol{y}^{k}}$
Cauchy- Schwarz	$\mathrm{D}_{\mathsf{CS}}(p\Vert q)=\frac{1}{2}\ln\left(\int q^{2}\ dr\cdot\int p^{2}\ dr\right)-\ln\left(\int p\cdot q\ dr\right)$	$\Delta\boldsymbol{y}^{k}=\left[\frac{g_{\varsigma}^{\boldsymbol{s}}(k)}{\sum_{j}\left[g_{\varsigma}^{\boldsymbol{s}}(j)\right]^{2}}-\frac{h_{\sigma}^{\Psi_{\mathsf{CS}}(\boldsymbol{s})}(k)}{\sum_{j}h_{\sigma}^{\Psi_{\mathsf{CS}}(\boldsymbol{s})}(j)\cdot g_{\varsigma}^{\boldsymbol{s}}(j)}\right]\left(-\frac{g_{\varsigma}^{\boldsymbol{s}}(k)}{2\varsigma^{2}}\right)\frac{\partial d_{\mathcal{E}}(\boldsymbol{s},\boldsymbol{y}^{k})}{\partial\boldsymbol{y}^{k}}$
-divergence	$\frac{\delta\mathrm{D}_{\mathsf{CS}}(p\Vert q)}{\delta q}=\frac{q}{\int q^{2}\ dr}-\frac{p}{\int p\cdot q\ dr}$	$\Delta\boldsymbol{y}^{k}=-\frac{(\varsigma+1)\cdot g_{\varsigma}^{\boldsymbol{s}}(k)}{2(\varsigma+d_{\mathcal{E}}(\boldsymbol{s},\boldsymbol{y}^{k}))}\left[\frac{g_{\varsigma}^{\boldsymbol{s}}(k)}{\sum_{j}\left[g_{\varsigma}^{\boldsymbol{s}}(j)\right]^{2}}-\frac{h_{\sigma}^{\Psi_{\mathsf{CS}}(\boldsymbol{s})}(k)}{\sum_{j}h_{\sigma}^{\Psi_{\mathsf{CS}}(\boldsymbol{s})}(j)\cdot g_{\varsigma}^{\boldsymbol{s}}(j)}\right]\frac{\partial d_{\mathcal{E}}(\boldsymbol{s},\boldsymbol{y}^{k})}{\partial\boldsymbol{y}^{k}}$

5 Conclusion

In this article we provide the mathematical foundation for a generalization of Self Organized Neighbor Embedding (SONE) which can be applied in dimension reduction and visualization tasks. The framework allows for the use of a very broad class of divergences as costfunction. In this context, we first present a general formulation of SONE as a gradient based optimization scheme. The use of a particular dissimilarity measure requires the availability of its Fréchet-derivative, which we present for a wide class of divergences. These results are summarized in table 1.

In forthcoming publications we will provide experimental results, evaluation and comparison with other dimension reduction and visualization techniques. We will examine the role of the different divergence families and their advantages for some data domains.

References

[1] S. I. Amari. *Differential-Geometrical Methods in Statistics*. Springer, 1985.

[2] S. I. Amari and H. Nagaoka. Methods of information geometry. In *Translations of Methematical Monographs*, volume 191. Oxford University Press, New York, 2000.

[3] A. Banerjee, I. Dhillon, J. Ghosh, S. Merugu, and D. S. Modha. A generalized maximum entropy approach to bregman co-clustering and matrix approximation. In *KDD '04: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 509–514, New York, NY, USA, 2004. ACM.

[4] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh. Clustering with bregman divergences. *Journal of Machine Learning Research*, 6:1705–1749, 2005.

[5] A. Basu, N. L. H. Ian R. Harris, and M. C. Jones. Robust and efficient estimation by minimising a density power divergence. *Biometrika*, 85(3):549–559, 1998.

[6] G. Baudat and F. Anouar. Generalized discriminant analysis using a kernel approach. *Neural Computation*, 12(10):2385–2404, 2000.

[7] N. Bertin, C. Fevotte, and R. Badeau. A tempering approach for itakura-saito non-negative matrix factorization. with application to music transcription. In *ICASSP '09: Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1545–1548, Washington, DC, USA, 2009. IEEE Computer Society.

[8] M. Brand. Charting a manifold. Technical Report 15, Mitsubishi Electric Research Laboratories (MERL), 2003.

[9] L. M. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7:200–217, 1967.

[10] K. Bunte, B. Hammer, T. Villmann, M. Biehl, and A. Wismüller. Exploratory observation machine (XOM) with Kullback-Leibler divergence for dimensionality reduction and visualization. In M. Verleysen, editor, *18th European Symposium on Artificial Neural Networks (ESANN)*, pages 87–92, Bruges, Belgium, 2010.

[11] K. Bunte, B. Hammer, T. Villmann, M. Biehl, and A. Wismüller. Exploratory observation machine (XOM) with Kullback-Leibler Divergence for dimensionality reduction and visualization. In M. Verleysen, editor, *Proc. of European Symposium on Artificial Neural Networks (ESANN)*, pages 87–92, Bruges, Belgium, April 2010.

[12] K. Bunte, B. Hammer, A. Wismüller, and M. Biehl. Adaptive local dissimilarity measures for discriminative dimension reduction of labeled data. *Neurocomputing*, 73(7-9):1074–1092, 2010.

[13] K. Bunte, P. Schneider, B. Hammer, F.-M. Schleif, T. Villmann, and M. Biehl. Discriminative visualization by limited rank matrix learning. Technical Report MLR-03-2008, Leipzig University, 2008.

[14] A. Cichocki, R. Zdunek, A. Phan, and S. I. Amari. *Non-negative matrix and tensor factorizations*. Wiley, Chichester, 2009.

[15] I. Csiszár. Information-type measures of difference of probability distributions and indirect observations. In *Studia Sci. Math. Hungar*, volume 2, pages 299–318, 1967.

[16] I. Csiszár. A class of measures of informativity of observation channels. In *Periodica Math. Hungar*, volume 2, pages 191–213, 1972.

[17] V. De Silva and J. B. Tenenbaum. Global versus local methods in nonlinear dimensionality reduction. In *Advances in Neural Information Processing Systems 15*, volume 15, pages 705–712, 2003.

[18] I. S. Dhillon and S. Sra. Generalized nonnegative matrix approximations with Bregman divergences. In *In: Neural Information Proc. Systems*, pages 283–290, Vancouver, Canada, December 2005.

[19] I. S. Dhillon and J. A. Tropp. Matrix nearness problems with Bregman divergences. *SIAM J. Matrix Anal. Appl.*, 29(4):1120–1146, 2007.

[20] S. Eguchi and Y. Kano. Robustifying maximum likelihood estimation. Technical Report 802, Tokyo-Institute of Statistical Mathematics, Tokyo, 2001.

[21] H. Fujisawa and S. Eguchi. Robust parameter estimation with a small bias against heavy contamination. *Multivariate Analalysis*, 99(9):2053–2081, 2008.

[22] K. Fukunaga. *Introduction to Statistical Pattern Recognition, 2nd edn. (Computer Science and Scientific Computing Series)*. Academic Press, September 1990.

[23] T. Heskes. Energy functions for self-organizing maps, 1999.

[24] G. Hinton and S. Roweis. Stochastic neighbor embedding. In *Advances in Neural Information Processing Systems 15*, pages 833–840. MIT Press, 2003.

[25] F. Itakura and S. Saito. Analysis synthesis telephony based on the maximum likelihood method. In *Proc. 6th of the International Congress on Acoustics*, volume C-5-5, pages 17–20, 1968.

[26] R. Jenssen. *An Information Theoretic Approach to Machine Learning*. PhD dissertation, University of Tromsø, Department of Physics, 2005.

[27] R. Jenssen, J. C. Principe, D. Erdogmus, and T. Eltoft. The cauchy-schwarz divergence and parzen windowing: Connections to graph theory and mercer kernels. *Journal of the Franklin Institute*, 343(6):614 – 629, 2006.

[28] T. Kohonen. *Self-Organizing Maps*. Springer, Berlin, Heidelberg, New York, 3rd edition, 2001.

[29] P. L. Lai and C. Fyfe. Bregman divergences and multi-dimensional scaling. In *15th International Conference on Neuro-Information Processing: (ICONIP), Revised Selected Papers, Part II*, pages 935–942. Springer-Verlag, Auckland, New Zealand, November 25-28 2008.

[30] P. L. Lai and C. Fyfe. Bregman divergences and multi-dimensional scaling. In *Advances in Neuro-Information Processing: 15th International Conference, ICONIP 2008, Auckland, New Zealand, November 25-28, 2008, Revised Selected Papers, Part II*, pages 935–942, Berlin, Heidelberg, 2009. Springer-Verlag.

[31] J. Lee and M. Verleysen. *Nonlinear dimensionality reduction*. Springer, 1st edition, 2007.

[32] F. Liese and I. Vajda. Convex statistical distances. In *Teubner-Texte zur Methematik*, volume 95. Teubner-Verlag, Leipzig, 1987.

[33] G. A. L.W. Kantorowitsch. *Funktionalanalysis in normierten Räumen*. Akademie Verlag, Berlin, second edition, 1978.

[34] M. Mihoko and S. Eguchi. Robust blind source separation by beta divergence. *Neural Computation*, 14(8):1859–1886, 2002.

[35] T. Minka. Divergence measures and message passing. Technical Report 173, Microsoft Research Ltd., Cambridge, UK, 2005.

[36] N. Murata, T. Takenouchi, and T. Kanamori. Information geometry of U-Boost and Bregman divergence. *Neural Computation*, 16:1437–1481, 2004.

[37] F. Österreicher. Csiszár f-divergences-basic properties. Technical report, Res. Report Collection, 2002.

[38] J. C. Principe, D. Xu, and J. W. Fisher III. Information-theoretic learning. In S. Haykin, editor, *Unsupervised Adaptive Filtering*, volume 1, chapter 7. Wiley, New York, second edition, 2000.

[39] S. T. Roweis and L. K. Saul. Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science*, 290(5500):2323–2326, 2000.

[40] Y. Teh and S. Roweis. Automatic alignment of local representations. *Advances in Neural Information Processing Systems*, 15:841–848, 2003.

[41] J. B. Tenenbaum, V. d. Silva, and J. C. Langford. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science*, 290(5500):2319–2323, 2000.

[42] L. van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, November 2008.

[43] L. J. P. van der Maaten, E. O. Postma, and H. J. van den Herik. Dimensionality reduction: A comparative review. Technical Report TiCC-TR 2009-005, Tilburg University, Oct 2009.

[44] T. Villmann and S. Haase. Divergence based Vector Quantization using Fréchet-derivatives. Submitted to Neural Computation, 2010.

[45] T. Villmann and S. Haase. Mathematical aspects of divergence based vector quantization using Fréchet-derivatives - extended and revised version -. Technical Report MLR-01-2010, Leipzig University, 2010.

[46] T. Villmann and S. Haase. Mathematical foundations of th generalization of t-SNE amd SNE for arbitrary divergences. Technical Report MLR-02-2010, Leipzig University, 2010.

[47] T. Villmann, B. Hammer, F.-M. Schleif, T. Geweniger, and W. Hermann. Fuzzy classification by fuzzy labeled neural gas. *Neural Networks*, 19(6-7):772–779, 2006.

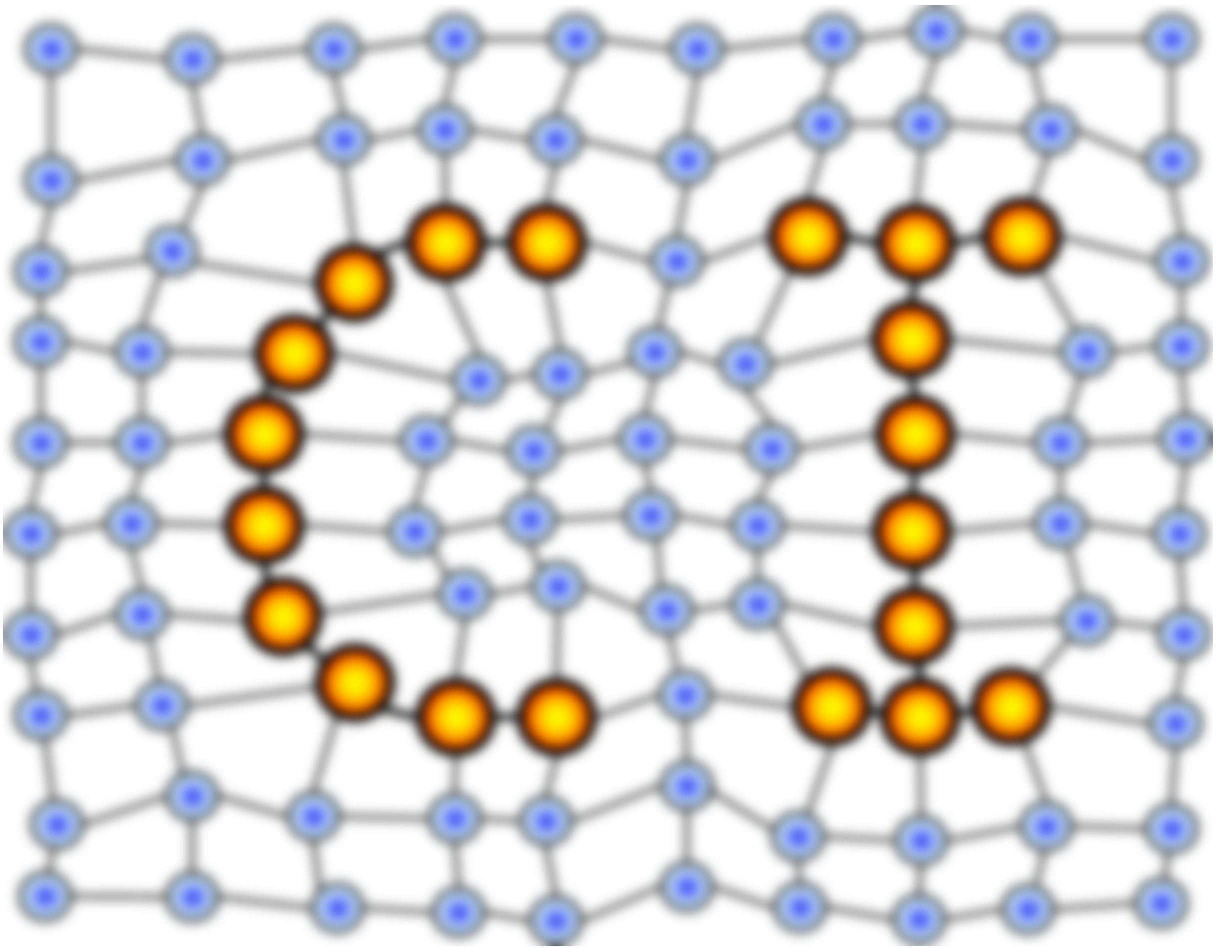
[48] A. Wismüller. Exploration-organized morphogenesis (XOM) – a general framework for learning by self-organization. In *Human and Machine Perception. Reports of the Institute for Phonetics and Speech Communication (FIPKM)*, volume 37, pages 205–239, 2001. ISSN 0342-782X.

[49] A. Wismüller. Exploratory Morphogenesis (XOM): A Novel Computational Framework for Self-Organization. Ph.D. thesis, Technical University of Munich, Department of Electrical and Computer Engineering, 2006.

[50] A. Wismüller. A computational framework for exploratory data analysis. In M. Verleysen, editor, *17th European Symposium on Artificial Neural Networks (ESANN)*, pages 547–552, Bruges, Belgium, 2009.

MACHINE LEARNING REPORTS

Report 03/2010



Impressum

Machine Learning Reports ISSN: 1865-3960

▽ Publisher/Editors

Prof. Dr. rer. nat. Thomas Villmann & Dr. rer. nat. Frank-Michael Schleif
University of Applied Sciences Mittweida
Technikumplatz 17, 09648 Mittweida - Germany•
http://www.mni.hs-mittweida.de/

▽ Acknowledgments

We would like to thank the reviewers for their time and patience.