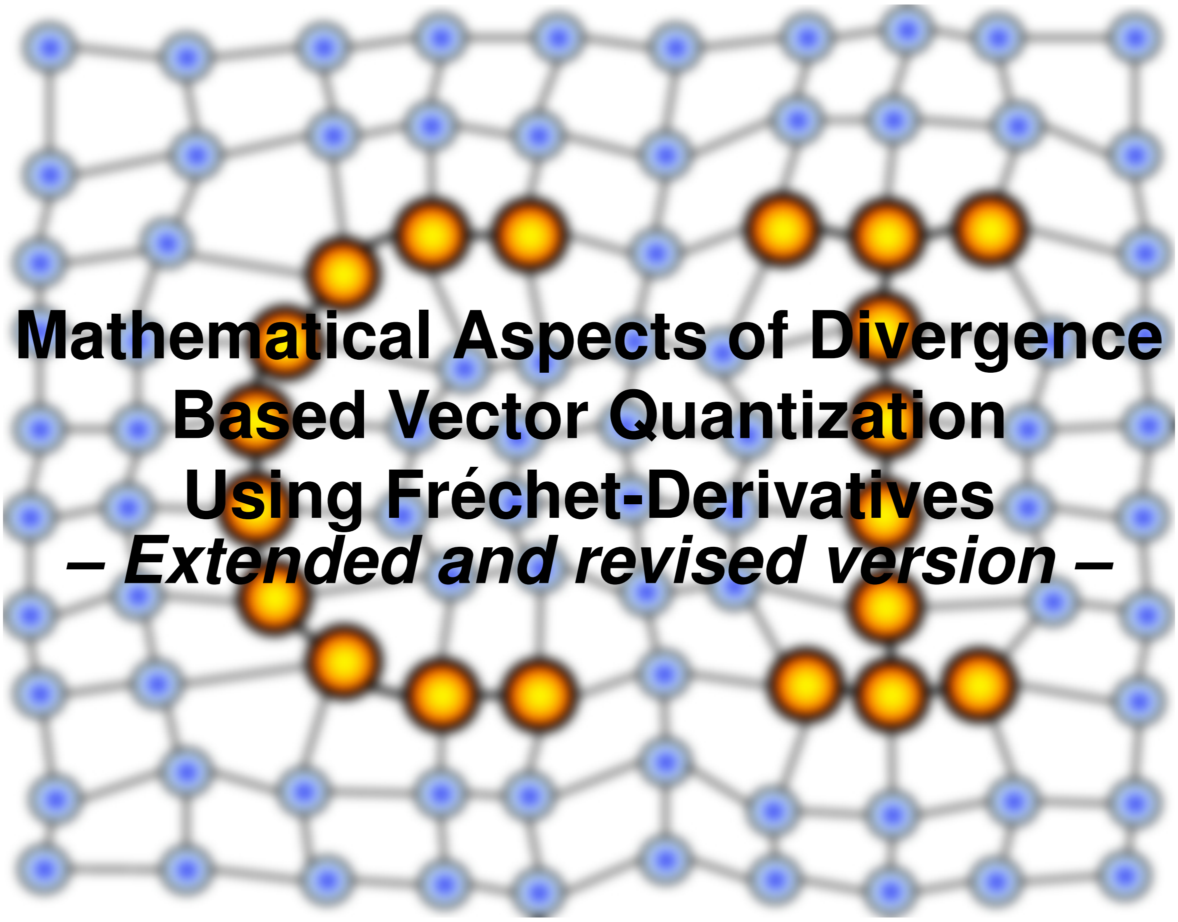
MACHINE LEARNING REPORTS



Mathematical Aspects of Divergence Based Vector Quantization Using Fréchet-Derivatives

– Extended and revised version –

Report 01/2010
Submitted: 21.01.2010
Published: 22.01.2010

T. Villmann and S. Haase
University of Applied Sciences Mittweida, Dep. of Mathematics/Natural Sciences, and Computer Sciences

Abstract

Supervised and unsupervised vector quantization methods for classification and clustering traditionally use dissimilarities, frequently taken as Euclidean distances. In this article we investigate the applicability of divergences instead. We deduce the mathematical fundamentals for its utilization in derivative based vector quantization algorithms. It bears on the generalized derivatives known as Fréchet-derivatives. We exemplary show the application of this methodology for widely applied supervised and unsupervised vector quantization schemes including self-organizing maps, neural gas, and learning vector quantization. Further we show principles for hyperparameter optimization and relevance learning for parametrized divergences in the case of supervised vector quantization to achieve an improved classification accuracy.

1 Introduction

Supervised and unsupervised vector quantization for classification and clustering is strongly associated with the concept of dissimilarity usually judged in terms of distances. The most common choice is the Euclidean metric. Yet, in the last years alternative dissimilarity measures became attractive for advanced data processing. Examples are functional metrics like *Sobolev-distances* or kernel based dissimilarity measures [62],[35]. These metrics take the functional structure of the data into account [34],[47],[50],[58].

Recently, information theory based approaches are proposed considering *divergences* for clustering [3], [26],[36],[20]. For other data processing methods like multi-dimensional scaling (MDS) [33], stochastic neighbor embedding [40], blind source separation [42] or non-negative matrix factorization [8], also divergence based approaches are introduced. In prototype based classification, first approaches utilizing information theoretic approaches were recently proposed [13],[56],[61].

Yet, a systematic analysis of prototype based clustering and classification relying on divergences is not given so far. Further, the respective existing approaches usually are carry out in the so-called batch mode for optimization but *are not available for online learning*. The latter method requires the calculation of the derivatives of the underlying metrics, i.e. divergences here.

In the present contribution we offer a systematic approach for divergence based vector quantization using divergence derivatives. For this purpose, important but general classes of divergences are identified, widely following and extending the scheme introduced by CICHOCKI ET AL. in [9]. The mathematical framework for functional derivatives of divergences is given by the functional-analytic generalization of usual derivatives – the concept of *Fréchet-derivatives* [16],[29].

After characterization of the different classes of divergences and a short introduction of Fréchet-derivatives we apply this framework to the several divergences obtaining generalized derivatives, which can be used for online learning of divergence based methods for supervised and unsupervised vector quantization as well as other gradient based approaches. We explore explicitly for prominent examples the respective derivatives. Thereafter, we exemplarily consider some of the most prominent approaches for unsupervised as well as supervised prototype based vector quantization in the light of divergence based online learning using Fréchet-derivatives. For the latter approaches we additionally provide a gradient learning scheme, called hyperparameter adaptation, for optimization of parameters occurring in parametrized divergences.

2 Characterization of divergences

In a general mean, divergences are functionals designed for determination of similarity between non-negative integrable measure functions p and ρ. We denote such functions as *positve measures*. Normalized such measures are denoted as density functions. Divergences $D(p||\rho)$ are defined as functionals which have to be non-negative and zero iff $p \equiv \rho$ except on a zero-measure set. Further, $D(p||\rho)$ is required to be covex with respect to the first argument. Yet, divergences are neither necessarily symmmetric nor have to fulfill the triangle inequality as it is supposed for metrics. According to the classification given in CICHOCKI ET AL. [9] one can dis-

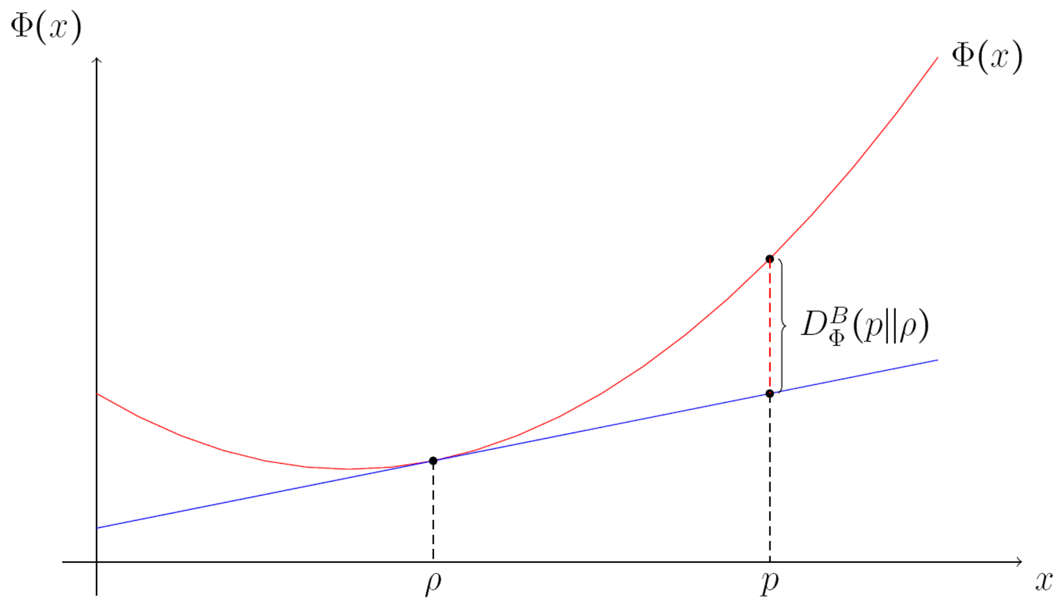


Figure 1: Illustration of the Bregman divergence $D^B_\Phi(p||\rho)$ as a vertical distance between p and the tangential hyperplane to the graph of Φ at point ρ taking p and ρ as points in a functional space.

tinguish at least *three* main classes of divergences, the *Bregman*-divergences, the *Csiszár's* f-divergences and the γ-divergences emphasizing different properties.

We remark at this point explicitly that we generally assume that p and ρ are positive measure (densities), not necessarily be normalized. In case of (normalized) densities we explicitly refer to these as probability densities.

2.1 Bregman divergences

Bregman divergences are defined by generating convex functions Φ in the following way [7]:

Let Φ be a strictly convex real-valued function with the domain $\mathcal{L}$ (the Lebesgue-integrable functions). Further, Φ is assumed to be twice continuously Fréchet-differentiable [29]. A Bregman divergence is defined as $D^B_\Phi : \mathcal{L} \times \mathcal{L} \longrightarrow \mathbb{R}^+$ with

$$D^B_\Phi(p||\rho) = \Phi(p) - \Phi(\rho) - \frac{\delta\Phi(\rho)}{\delta\rho}(p-\rho) \tag{2.1}$$

whereby $\frac{\delta\Phi(\rho)}{\delta\rho}$ is the Fréchet-derivative of Φ with respect to ρ (see sec. 3.1.1).

The Bregman divergence $D^B_\Phi(p||\rho)$ can be interpreted as a measure of convexity of the generating function Φ. Taking p and ρ as points in a functional space $D^B_\Phi(p||\rho)$ plays the role of the vertical distance between p and the tangential hyperplane to the graph of Φ at point ρ, which is illustrated in Fig.2.1.

Bregman divergences are linear according to the generating function Φ:

$$D^B_{\Phi_1+\lambda\Phi_2}(p||\rho) = D^B_{\Phi_1}(p||\rho) + \lambda \cdot D^B_{\Phi_2}(p||\rho).$$

Further, $D^B_\Phi(p||\rho)$ is invariant under affine transforms $\Gamma(q) = \Phi(q) + \Psi_g[q] + \xi$ for positive measures g and q with

$$\Psi_g[q] = \frac{\delta\Gamma(g)}{\delta g}\cdot q - \frac{\delta\Phi(g)}{\delta g}\cdot q$$

and Ψ_g is a linear operator independent from q [15]. ξ is a scalar. In that case

$$D_{\Gamma}^{B}(p||\rho) = D_{\Phi}^{B}(p||\rho)$$

is valid. Further, the *generalized Pythagorean theorem* holds for any triple p, ρ, τ of non-negative integrable functions:

$$D_{\Phi}^{B}(p||\tau) = D_{\Phi}^{B}(p||\rho) + D_{\Phi}^{B}(\rho||\tau) + \frac{\delta\Phi(\rho)}{\delta\rho}(p-\rho) - \frac{\delta\Phi(\tau)}{\delta\tau}(p-\rho)$$

The *sensitivity* of a Bregman divergence at p is defined as

$$\begin{aligned} s(p,\tau) &= \frac{\partial^2 D_{\Phi}^{B}(p||p+\alpha\tau)}{\partial\alpha^2}|_{\alpha=0} \\ &= -\tau\frac{\delta^2\Phi(p)}{\delta p^2}\tau \end{aligned} \tag{2.2}$$

with $\tau \in \mathcal{L}$ and the restriction that $\int \tau(x)\,dx = 0$ [51]. Note that $\frac{\delta^2\Phi(p)}{\delta p^2}$ is the Hessian of the generating function. The sensitivity $s(p,\tau)$ measures the velocity of change of the divergence at the point p in the direction of τ.

A last property mentioned here is an optimality one, stated in [3]: Given a set S of positive measures p with mean $\mu = E[S]$ and $\mu \in S$. Then, the unique minimizer of $E_{p\in S}\left[D_{\Phi}^{B}(p||\rho)\right]$ is $\rho = \mu$. The inverse direction of this statement is also true, i.e if $E_{p\in S}\left[D_{F}^{B}(p||\rho)\right]$ is minimum for $\rho = \mu$ then $D_{F}^{B}(p||\rho)$ is a Bregman divergence. This property predestinates Bregman divergences for clustering problems.

Finally, we give some important examples:

1. *generalized Kullback-Leibler-divergence* for non-normalized p and ρ [9]:

$$D_{GKL}(p||\rho) = \int p(\mathbf{x})\log\left(\frac{p(\mathbf{x})}{\rho(\mathbf{x})}\right)d\mathbf{x} - \int p(\mathbf{x}) - \rho(\mathbf{x})\,d\mathbf{x} \tag{2.3}$$

with the generating function

$$\Phi(f) = \int f\cdot\log f - f d\mathbf{x}\ .$$

If p and ρ are normalized densities (probability densities) $D_{GKL}(p||\rho)$ is reduced to the usual Kullback-Leibler-divergence [32],[30]:

$$D_{KL}(p||\rho) = \int p(\mathbf{x})\log\left(\frac{p(\mathbf{x})}{\rho(\mathbf{x})}\right)d\mathbf{x}\ , \tag{2.4}$$

which is related to the *Shannon-entropy* [54]

$$H^S(p) = -\int p(\mathbf{x})\log(p(\mathbf{x}))\,d\mathbf{x} \tag{2.5}$$

via

$$D_{KL}(p||\rho) = V_S(p,\rho) - H^S(p)$$

where

$$V_S(p,\rho) = -\int p(\mathbf{x})\log(\rho(\mathbf{x}))\,d\mathbf{x}$$

is *Shannon's cross entropy*.

2. *Itakura-Saito*-divergence [25]:

$$D_{IS}(p||\rho) = \int \left[\frac{p}{\rho} - \log\left(\frac{p}{\rho}\right) - 1\right] d\mathbf{x} \tag{2.6}$$

   based on the *Burg entropy*

$$H^B(p) = -\int \log(p)\, d\mathbf{x}$$

   which also serves as the generating function

$$\Phi(f) = H^B(f)\ .$$

   The Itakura-Saito-divergence is also known as negative *cross Burg entropy* and is frequently applied in image processing and sound processing.

3. The Euclidean distance in terms of a Bregman-divergence is obtained by the generating function

$$\Phi(f) = f^2$$

   We extend this definition and introduce the parametrized version

$$\Phi_\eta(f) = f^\eta$$

   defining the $\eta-$*divergence* also to be known as norm-like divergence [44]:

$$D_\eta(p||\rho) = \int p^\eta + (\eta - 1)\cdot\rho^\eta - \eta\cdot p\cdot\rho^{(\eta-1)} d\mathbf{x} \tag{2.7}$$

   which obviously converges to the Euclidean distance for $\eta \to 2$.

If we assume that p and ρ are positive measures, then an important subset of Bregman divergences belong to the class of $\beta-$divergences [12], which are defined, following CICHOCKI ET AL. [9], as

$$D_\beta(p||\rho) = \int p\cdot\frac{p^{\beta-1} - \rho^{\beta-1}}{\beta - 1} d\mathbf{x} - \int \frac{p^\beta - \rho^\beta}{\beta} d\mathbf{x} \tag{2.8}$$

$$= \int p^\beta\left(\frac{1}{\beta-1} - \frac{1}{\beta}\right) - \rho^{\beta-1}\left(\frac{p}{\beta-1} + \frac{\rho}{\beta}\right) d\mathbf{x} \tag{2.9}$$

with $\beta \neq 1$ and $\beta \neq 0$. In the limit $\beta \to 1$ the divergence $D_\beta(p,\rho)$ becomes the generalized Kullback-Leibler-divergence (2.3)[1]. The limit $\beta \to 0$ gives the Itakura-Saito-divergence (2.6). Further, $\beta-$divergences are related to the density power divergences $\hat{D}_\beta$ introduced in [5] by

$$\hat{D}_\beta(p||\rho) = \frac{1}{(1+\beta)} D_\beta(p||\rho)\ .$$

[1]We remark here that the relations $\frac{p^\gamma - \rho^\gamma}{\gamma} \underset{\gamma\to 0}{\longrightarrow} \log\frac{p}{\rho}$ and $\frac{p^\gamma - 1}{\gamma} \underset{\gamma\to 0}{\longrightarrow} \log p$ hold.

2.2 Csiszár's f-divergences

Csiszár's f-divergences are defined for *convex* functions $f \in \mathcal{F}$ with $f(1) = 0$ (without loss of generality) whereby

$$\mathcal{F} = \{g | g : [0, \infty) \to \mathbb{R}, \ g \text{ - convex}\} \, .$$

The f-divergences D_f for positive measures p and ρ are given by

$$D_f(p||\rho) = \int \rho(\mathbf{x}) \cdot f\left(\frac{p(\mathbf{x})}{\rho(\mathbf{x})}\right) d\mathbf{x} \tag{2.10}$$

with the definitions $0 \cdot f\left(\frac{0}{0}\right) = 0$, $0 \cdot f\left(\frac{a}{0}\right) = \lim_{x\to 0} x \cdot f\left(\frac{a}{x}\right) = \lim_{u\to\infty} a \cdot \frac{f(u)}{u}$ [11],[38],[55]. f is called *determining function* for $D_f(p||\rho)$. It corresponds to a *generalized* f-entropy [9] of the form

$$H_f(p) = -\int f(p) \, d\mathbf{x} \, . \tag{2.11}$$

The f-divergence D_f can be interpreted as an average (with respect to ρ) of the likelihood ratio $\frac{p}{\rho}$ describing the change rate of p with respect to ρ weighted by the determining function f. $D_f(p||\rho)$ is jointly convex in both p and ρ. Further, f defines an equivalence class in the sense that $D_f(p||\rho) = D_{\hat{f}}(p||\rho)$ iff $f(x) = \hat{f}(x) + c \cdot (x - 1)$ for $c \in \mathbb{R}$, i.e. $D_f(p||\rho)$ is invariant according to a linear shift regarding the determining function f. For f-divergences a certain kind of symmetry can be stated: Let $f, f^* \in \mathcal{F}$ and f^* is the *conjugate function* of f, i.e.$f^*(x) = x \cdot f\left(\frac{1}{x}\right)$ for $x \in (0, \infty)$. Then the relation $D_f(p||\rho) = D_{f^*}(\rho||p)$ is valid iff the conjugate differs from the original by a linear shift as above: $f(x) = f^*(x) + c \cdot (x - 1)$. A symmetric divergence can be obtained for an arbitrary convex function $g \in \mathcal{F}$ using its conjugate g^* for the definition $f = g + g^*$ as determing function.

An important and characterizing property is the monotonicity with respect to the coarse-graining of the underlying domain $\mathcal{D}$ of the positive measures p and ρ, which is similary to the monotonicity of the Fisher metric [2]: Let $\mathcal{K} = \{\kappa(\mathbf{y}|\mathbf{x}) \geq 0, \mathbf{x} \in \mathcal{D}, \mathbf{y} \in \mathcal{D}_{\mathbf{y}}\}$ with $\mathcal{D}_{\mathbf{y}}$ being the range of $\mathbf{y}$. κ describes a transition probability density, i.e. $\int \kappa(\mathbf{y}|\mathbf{x}) \, d\mathbf{y} = 1$ holds $\forall \mathbf{x} \in \mathcal{D}$. Denoting the positive measures of $\mathbf{y}$ derived from $p(\mathbf{x})$ and $\rho(\mathbf{x})$ by $p_\kappa(\mathbf{y})$ and $\rho_\kappa(\mathbf{y})$ the monotonicity is expressed by $D_f(p||\rho) \geq D_f(p_\kappa||\rho_\kappa)$.[2] Further, an isomorphy can be stated for f-divergences in the following way: Let

$$h : \mathbf{x} \longmapsto \mathbf{y} \tag{2.12}$$

be an invertible function transforming positive measures $p_1(\mathbf{x})$ and $\rho_1(\mathbf{x})$ to $p_2(\mathbf{y})$ and $\rho_2(\mathbf{y})$. Then $D_f(p_1||\rho_1) = D_f(p_2||\rho_2)$ holds and the pairs (p_1, ρ_1) and (p_2, ρ_2) are called isomporph [37]. Conversely, if a measure $D(p||\rho) = \int \rho(\mathbf{x}) \cdot G(p(\mathbf{x}), \rho(\mathbf{x})) \, d\mathbf{x}$ for an integrable function G is invariant according to invertible transformations h, then D is a f-divergence [46]. This isomorphy as well as the monotonicity employ f-divergences for application in speech, signal and pattern recognition [4],[46].

CICHOCKI ET AL. suggested a generalization of the f-divergences D_f [9]. In that divergence, f has not longer to be convex. It is proposed to be

$$D_f^G(p||\rho) = c_f \int p - \rho d\mathbf{x} + \int \rho(\mathbf{x}) \cdot f\left(\frac{p(\mathbf{x})}{\rho(\mathbf{x})}\right) d\mathbf{x} \tag{2.13}$$

[2]The equality holds iff the conditional densities $p_\kappa(\mathbf{x}|\mathbf{y}) = \frac{p(\mathbf{x}) \cdot \kappa(\mathbf{y}|\mathbf{x})}{p_\kappa(\mathbf{y})}$ and $\rho_\kappa(\mathbf{x}|\mathbf{y}) = \frac{\rho(\mathbf{x}) \cdot \kappa(\mathbf{y}|\mathbf{x})}{\rho_\kappa(\mathbf{y})}$ are identic, see [2], p. 56.

with $c_f = f'(1) \neq 0$ and denoted as *generalized* f-divergence. As consequence of this relaxation of the convexity condition, in case of p and ρ being probability densities, the first term vanishes, such that the usual form of f-divergences is obtained. Thus, as a famous example the *Hellinger divergence* [38],[55]:

$$D_H(p||\rho) = \int (\sqrt{p} - \sqrt{\rho})^2 \, d\mathbf{x} \tag{2.14}$$

with the generating function $f(u) = (\sqrt{u} - 1)^2$ with $u = \frac{p}{\rho}$. We remark, $D_H(p||\rho)$ is not a $f-$divergence for general densities p and ρ because f is not convex in that case, whereas for probability densities it is a $f-$divergence according to the *Cichocki-f*-divergence properties [9].

As the $\beta-$divergences in case of Bregman divergences, one can identify here an important subset of the $f-$divergences, the so-called $\alpha-$divergences according to the definition given in [9]:

$$\begin{aligned} D_\alpha(p||\rho) &= \frac{1}{\alpha(\alpha-1)} \int \left[p^\alpha \rho^{1-\alpha} - \alpha \cdot p + (\alpha - 1)\rho\right] d\mathbf{x} &(2.15)\\ &= \frac{1}{\alpha(\alpha-1)} \int \left[\rho\left(\left(\frac{p}{\rho}\right)^\alpha + (\alpha-1)\right) - \alpha \cdot p\right] d\mathbf{x} &(2.16)\end{aligned}$$

with the generating $f-$function

$$f(u) = u\frac{(u^{\alpha-1} - 1)}{\alpha^2 - \alpha} + \frac{1-u}{\alpha}$$

and $u = \frac{\rho}{p}$. In the limit $\alpha \to 1$ the generalized Kullback-Leibler-divergence D_{GKL} (2.3) is obtained. Further, in [9] is stated that the β-divergences can be generated from the α-divergences applying the non-linear transforms

$$p \to p^{\beta+2} \text{ and } \rho \to \rho^{\beta+2} .$$

The *Tsallis-divergence* is a widely applied divergence related to $\alpha-$divergence (2.15), however, only defined for probability densities. It is defined as

$$D_\alpha^T(p||\rho) = \int p \cdot \log_\alpha\left(\frac{\rho}{p}\right) d\mathbf{x} \tag{2.17}$$

with the convention

$$\log_\alpha(z) = \frac{z^{1-\alpha} - 1}{1-\alpha} \tag{2.18}$$

such that

$$D_\alpha^T(p||\rho) = \frac{1}{1-\alpha}\left(\int p^\alpha \rho^{1-\alpha} d\mathbf{x} - 1\right) \tag{2.19}$$

and $\alpha \neq 1$. The Tsallis-divergence is based on the *Tsallis-entropy*

$$\begin{aligned} H_\alpha^T(p) &= -\frac{1}{\alpha-1}\left(\int p^\alpha d\mathbf{x} - 1\right) &(2.20)\\ &= \int p \log_\alpha\left(\frac{1}{p}\right) d\mathbf{x} &(2.21)\end{aligned}$$

with $\log_\alpha(p)$ as defined in (2.18). In the limit $\alpha \to 1$ for $H_\alpha^T(p)$ becomes the Shannon-entropy (2.5) and the divergence $D_\alpha^T(p||\rho)$ converges to the Kullback-Leibler-divergence (2.4).

Further, the $\alpha-$divergences are closely related to the generalized *Rényi-divergences* [1],[9]:

$$D_\alpha^{GR}(p||\rho) = \frac{1}{\alpha-1}\log\left(\int\left[p^\alpha\rho^{1-\alpha} - \alpha\cdot p + (\alpha-1)\rho + 1\right]d\mathbf{x}\right) \tag{2.22}$$

for non-normalized ρ and p, whereas for probability densities the usual *Rényi-divergence* [48],[49]

$$D_\alpha^R(p||\rho) = \frac{1}{\alpha-1}\log\left(\int p^\alpha\rho^{1-\alpha}d\mathbf{x}\right) \tag{2.23}$$

is obtained[3]. The divergence $D_\alpha^R(p||\rho)$ is based on the *Rényi-entropy*

$$H_\alpha^R(p) = -\frac{1}{\alpha-1}\log\left(\int p^\alpha d\mathbf{x}\right) \ . \tag{2.24}$$

The Rényi-entropy fulfills the additivity property for independent propabilities p and q:

$$H_\alpha^R(p\times q) = H_\alpha^R(p) + H_\alpha^R(q) \ .$$

Further, the entropy $H_\alpha^R(p)$ is related to the Tsallis-entropy (2.21) by

$$H_\alpha^R(p) = -\frac{1}{\alpha-1}\log\left(1+(1-\alpha)\cdot H_\alpha^T\right) \ ,$$

which, however, has in consequence a different sub-additivity property:

$$H_\alpha^T(p\times q) = H_\alpha^T(p) + H_\alpha^T(q) + (1-\alpha)\cdot H_\alpha^T(p)\cdot H_\alpha^T(q)$$

for $\alpha \neq 1$.

2.3 γ-divergences

A class of very robust divergences with respect to outliers has been proposed by FUJISAWA&EGUCHI [17][4]. It is called γ*-divergences* defined for positive measures ρ and p as

$$\begin{aligned} D_\gamma(p||\rho) &= \log\left[\frac{\left(\int p^{\gamma+1}d\mathbf{x}\right)^{\frac{1}{\gamma(\gamma+1)}}\cdot\left(\int\rho^{\gamma+1}d\mathbf{x}\right)^{\frac{1}{\gamma+1}}}{\left(\int p\cdot\rho^\gamma d\mathbf{x}\right)^{\frac{1}{\gamma}}}\right] &(2.25)\\ &= \frac{1}{\gamma+1}\log\left[\left(\int p^{\gamma+1}d\mathbf{x}\right)^{\frac{1}{\gamma}}\cdot\left(\int\rho^{\gamma+1}d\mathbf{x}\right)\right] &(2.26)\\ &\quad -\log\left[\left(\int p\cdot\rho^\gamma d\mathbf{x}\right)^{\frac{1}{\gamma}}\right] \ . \end{aligned}$$

[3] Notify that a careful transformation of the parameter α is required for exact transformations between both divergences. For details see [1] p.84ff and [9] p.104.

[4] The divergence $D_\gamma(p||\rho)$ is proposed to be robust for $\gamma\in[0,1]$ with existence of $D_{\gamma=0}$ in the limit $\gamma\to 0$. A detailed analysis of robustness is given in [17].

The divergence $D_\gamma(p||\rho)$ is invariant under scalar multiplication with positive constants c_1 and c_2:

$$D_\gamma(p||\rho) = D_\gamma(c_1 \cdot p||c_2 \cdot \rho) \ .$$

The equation $D_\gamma(p||\rho) = 0$ only holds if $p = c \cdot \rho$ ($c > 0$) in case of positive measures. Yet, for probability densities $c = 1$ is required. In contradiction to the f-divergences, here an isomorphy can be stated for h-transformations (2.12) which more strictly assumed to be affine.

As for Bregman divergences a modified Pythagorean relation between positive measures can be stated for special choices of positive measures p, ρ, τ: Let p be a distortion of ρ defined as convex combination with a positve distortion measure δ:

$$p_\varepsilon(\mathbf{x}) = (1-\varepsilon) \cdot \rho(\mathbf{x}) + \varepsilon \cdot \delta(\mathbf{x})$$

Further a positve measure g is denoted as δ-*consistant* if

$$\nu_g = \left(\int \delta(\mathbf{x})\, g(\mathbf{x})^\alpha \, d\mathbf{x} \right)^{\frac{1}{\alpha}}$$

is sufficiently small for large $\alpha > 0$. If two positive measures ρ and τ are δ-consistant according to a distortion measure δ, then the Pythagorean relation approximately holds for ρ, τ and the distortion p_ε of ρ

$$\triangle(p_\varepsilon, \rho, \tau) = D_\gamma(p_\varepsilon||\tau) - D_\gamma(p_\varepsilon||\rho) - D_\gamma(\rho||\tau) = \mathcal{O}(\varepsilon \nu^\gamma)$$

with $\nu = \max\{\nu_\rho, \nu_\tau\}$. This property implies the robustness of γ-divergences according to distortions.

In the limit $\gamma \to 0$ $D_\gamma(\rho||p)$ becomes the usual Kullback-Leibler-divergence (2.4) $D_{KL}(\hat{\rho}||\hat{p})$ with normalized densities

$$\hat{\rho} = \frac{\rho}{\int \rho d\mathbf{x}} \text{ and } \hat{p} = \frac{p}{\int p d\mathbf{x}} \ .$$

For $\gamma = 1$ the γ-divergence becomes the *Cauchy-Schwarz*-divergence

$$D_{CS}(p||\rho) = \frac{1}{2} \log \left(\int \rho^2(\mathbf{x})\, d\mathbf{x} \cdot \int p^2(\mathbf{x})\, d\mathbf{x} \right) - \log(V(p,\rho)) \qquad (2.27)$$

with

$$V(p,\rho) = \int p(\mathbf{x}) \cdot \rho(\mathbf{x})\, d\mathbf{x} \qquad (2.28)$$

being the *cross correlation potential*. The *Cauchy-Schwarz*-divergence $D_{CS}(p||\rho)$ was introduced by J. PRINCIPE considering the Cauchy-Schwarz-inequality for norms [45]. It is based on the quadratic Rényi-entropy $H_2^R(p)$ from (2.24) [27]. Obviously, $D_{CS}(p||\rho)$ is symmetric. It is frequently applied for Parzen window estimation and particularly suitable for spectral clustering as well as for related graph cut problems [28].

3 Derivatives of divergences – a functional analytic approach

In this section we provide the mathematical formalism of generalized derivatives for functionals p and ρ. It is known as *Fréchet-derivatives* or *functional derivatives*.

First, we briefly reconsider the theory of functional derivatives including Fréchet- and Gâteaux-derivatives and its relation to directional derivatives. Thereafter we investigate the above divergence classes within this framework. In particular, we explain their Fréchet-derivatives.

3.1 Functional derivatives

3.1.1 Fréchet-derivatives and Fréchet-derivatives of a functional

Suppose, X and Y are *Banach spaces*, $U \subset X$ is open and $F : X \to Y$. F is called *Fréchet differentiable* at $x \in X$, if there exists a *bounded linear* operator $\frac{F[x]}{\delta x} : X \to Y$, such that for $h \in X$ the limit

$$\lim_{h\to 0} \frac{\left\| F\left(u+h\right) - F\left(u\right) - \frac{F[u]}{\delta u}\left[h\right] \right\|_Y}{\left\|h\right\|_X} = 0 \ .$$

This general definition can be focussed for functional mapping: Let L be a functional mapping from a linear, functional *Banach-space* B to $\mathbb{R}$. Further, let B be equipped with a norm $\|\cdot\|$, and $f, h \in B$ are two functionals. The *Fréchet-derivative* of L at point f is formally defined as

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

The existence and continuity of the limes is equivalent to the existence and continuity of the derivative. For a detailed introduction we refer to [29].

Yet, we recall two main properties of the Fréchet-derivative for functionals, which are important in studying divergences: First, if L is linear then

$$L\left[f+\varepsilon h\right] - L\left[f\right] = \varepsilon L\left[h\right]$$

and, hence, $\frac{L[f]}{\delta f}\left[h\right] = L$. Further, an analogon of the chain rule known from usual differential calculus can be stated:

Let $F : \mathbb{R} \to \mathbb{R}$ be a continuously-differentiable mapping. We consider the functional

$$L\left[f\right] = \int F\left(f\left(x\right)\right) dx$$

Then the Fréchet-derivative is obtained as $\frac{L[f]}{\delta f} = F'\left(f\right)$, which can be seen from

$$\begin{aligned} \frac{1}{\varepsilon}\left(L\left[f+\varepsilon h\right] - L\left[f\right]\right) &= \frac{1}{\varepsilon}\int F\left(f\left(x\right)+\varepsilon h\left(x\right)\right) - F\left(f\left(x\right)\right) dx \\ &= \frac{1}{\varepsilon}\int F'\left(f\left(x\right)\right)\cdot \varepsilon h\left(x\right) + \mathcal{O}\left(\varepsilon^2 h\left(x\right)^2\right) dx \\ &\underset{\varepsilon\to 0}{\longrightarrow} \int F'\left(f\left(x\right)\right)\cdot h\left(x\right) dx \end{aligned}$$

and utilization of the linear property of the integral operator.

Last but not least we state the following important remark: The Fréchet derivative in finite-dimensional spaces is the usual derivative. In particular, it is represented in coordinates by the *Jacobi matrix*. Thus, the Fréchet derivative is a generalization of the *directional derivatives*.

3.1.2 Gâteaux-derivatives

Gâteaux-derivatives are also a generalization of the concept of directional derivatives and can be seen in the middle between Fréchet-derivatives and usual derivatives.

Suppose, X and Y are locally convex topological vector spaces (for example, *Banach spaces)*, $U \subset X$ is open and $F : X \to Y$. The Gâteaux-differential of F at $u \in U$ is *in the direction* $v \in X$ defined as

$$dF(u; v) = \lim_{\tau \to 0} \frac{F(u + \tau \cdot v) - F(u)}{\tau} = T_u(v) \ ,$$

if the limit exists, and the operator $T_u(v) : X \to Y$ is bounded, $\tau \in \mathbb{R}$. The value

$$T_u(v) = \frac{d}{d\tau} F(u + \tau \cdot v) \,|_{\tau=0}$$

is denoted as *Gâteaux-derivative* at u, if the limit exists for all $v \in X$, and one says that F is Gâteaux differentiable at u.

If F is Fréchet differentiable, then it is also Gâteaux differentiable, and its Fréchet and Gâteaux derivatives are identical $\frac{F[u]}{\delta u}[v] = T_u(v)$ and, hence, $T_u(v)$ is linear. The converse is clearly not true, since the Gâteaux derivative may fail to be linear or continuous.[5]

3.2 Fréchet-derivatives for the different divergence classes

We are now ready to investigate functional derivatives of divergences. In particular we focus on Fréchet-derivatives.

3.2.1 Bregman-divergences

We investigate the Fréchet-derivative for the Bregman-divergences (2.1) and formally obtain

$$\frac{\delta D_\Phi^B(p||\rho)}{\delta \rho} = \frac{\Phi(p)}{\delta \rho} - \frac{\Phi(\rho)}{\delta \rho} - \frac{\delta \left[\frac{\delta \Phi(\rho)}{\delta \rho}(p - \rho)\right]}{\delta \rho} \tag{3.1}$$

with

$$\frac{\delta \left[\frac{\delta \Phi(\rho)}{\delta \rho}(p - \rho)\right]}{\delta \rho} = \frac{\delta^2 [\Phi(\rho)]}{\delta \rho^2}(p - \rho) - \frac{\delta \Phi(\rho)}{\delta \rho} \ .$$

In case of the generalized Kullback-Leibler-divergence (2.3) this reads as

$$\frac{\delta D_{GKL}(p||\rho)}{\delta \rho} = -\frac{p}{\rho} + 1 \tag{3.2}$$

whereas for the usual Kullback-Leibler-divergence (2.4)

$$\frac{\delta D_{KL}(p||\rho)}{\delta \rho} = -\frac{p}{\rho} \tag{3.3}$$

[5]In fact, it is even possible for the Gâteaux derivative to be linear and continuous but for the Fréchet derivative to fail to exist.

is obtained.

The η-divergence (2.7) leads to

$$\frac{\delta D_{\eta}\left(p||\rho\right)}{\delta\rho}=\rho^{\eta-2}\cdot\left(1-\eta\right)\cdot\eta\cdot\left(p-\rho\right) \tag{3.4}$$

which reduces in case of $\eta=2$ to the derivative of the Euclidean distance $-2\left(p-\rho\right)$, commonly used in many vector quantization algorithms including the online variant of k-means, SOMs, NG and so on.

Further for the subset of β-divergences (2.8) we have

$$\begin{aligned}\frac{\delta D_{\beta}\left(p||\rho\right)}{\delta\rho} &= -p\cdot\rho^{\beta-2}+\rho^{\beta-1} && (3.5)\\ &= \rho^{\beta-2}\left(\rho-p\right)\ . && (3.6)\end{aligned}$$

3.2.2 f-divergences

For f-divergences (2.10) the Fréchet-derivative is

$$\begin{aligned}\frac{\delta D_{f}\left(p||\rho\right)}{\delta\rho} &= f\left(\frac{p\left(\mathbf{x}\right)}{\rho\left(\mathbf{x}\right)}\right)+\rho\left(\mathbf{x}\right)\frac{\partial f\left(u\right)}{\partial u}\frac{\delta u}{\delta\rho}\\ &= f\left(\frac{p\left(\mathbf{x}\right)}{\rho\left(\mathbf{x}\right)}\right)+\rho\left(\mathbf{x}\right)\frac{\partial f\left(u\right)}{\partial u}\cdot\frac{-p}{\rho^{2}} && (3.7)\end{aligned}$$

with $u=\frac{p}{\rho}$. As a famous example we get for the Hellinger divergence (2.14)

$$\frac{\delta D_{H}\left(p||\rho\right)}{\delta\rho}=1-\sqrt{\frac{p}{\rho}}\ . \tag{3.8}$$

The subset of α-divergences (2.15) can be handled by

$$\frac{\delta D_{\alpha}\left(p||\rho\right)}{\delta\rho}=-\frac{1}{\alpha}\left(p^{\alpha}\rho^{-\alpha}-1\right)\ . \tag{3.9}$$

The related Tsallis-divergence D_{α}^{T} (2.19) leads to the derivative

$$\frac{\delta D_{\alpha}^{T}\left(p||\rho\right)}{\delta\rho}=\left(\frac{p}{\rho}\right)^{\alpha} \tag{3.10}$$

depending on the parameter α. The generalized Rényi-divergences (2.22) are treated according to

$$\begin{aligned}\frac{\delta D_{\alpha}^{GR}\left(p||\rho\right)}{\delta\rho} &= -\frac{p^{\alpha}\rho^{-\alpha}-1}{\int\left[p^{\alpha}\rho^{1-\alpha}-\alpha\cdot p+\left(\alpha-1\right)\rho+1\right]d\mathbf{x}}\\ &= \frac{\alpha}{\int\left[p^{\alpha}\rho^{1-\alpha}-\alpha\cdot p+\left(\alpha-1\right)\rho+1\right]d\mathbf{x}}\frac{\delta D_{\alpha}\left(p||\rho\right)}{\delta\rho}\ , && (3.11)\end{aligned}$$

which is reduced to

$$\frac{\delta D_{\alpha}^{R}\left(p||\rho\right)}{\delta\rho}=-\frac{p^{\alpha}\rho^{-\alpha}}{\int p^{\alpha}\rho^{1-\alpha}d\mathbf{x}} \tag{3.12}$$

in case of the usual Rényi-divergences (2.23).

3.2.3 γ-divergences

For the γ-divergences we rewrite (2.25) in the form

$$D_\gamma(p||\rho) = \frac{1}{\gamma+1}\ln F_1 - \ln F_2$$

with $F_1 = \left(\int p^{\gamma+1} d\mathbf{x}\right)^{\frac{1}{\gamma}} \cdot \left(\int \rho^{\gamma+1} d\mathbf{x}\right)$ and $F_2 = \left(\int p \cdot \rho^{\gamma} d\mathbf{x}\right)^{\frac{1}{\gamma}}$. Then we get

$$\frac{\delta D_\gamma(p||\rho)}{\delta\rho} = \frac{1}{\gamma+1}\frac{1}{F_1}\frac{\delta F_1}{\delta\rho} - \frac{1}{F_2}\frac{\delta F_2}{\delta\rho}$$

with

$$\begin{aligned}\frac{\delta F_1}{\delta\rho} &= \left(\int p^{\gamma+1} d\mathbf{x}\right)^{\frac{1}{\gamma}} \frac{\left(\int \rho^{\gamma+1} d\mathbf{x}\right)}{\delta\rho}\\ &= \left(\int p^{\gamma+1} d\mathbf{x}\right)^{\frac{1}{\gamma}} (\gamma+1)\,\rho^{\gamma}\end{aligned}$$

and

$$\begin{aligned}\frac{\delta F_2}{\delta\rho} &= \frac{1}{\gamma}\left(\int p\cdot\rho^{\gamma} d\mathbf{x}\right)^{\frac{1}{\gamma}-1} \frac{p\cdot\rho^{\gamma}}{\delta\rho}\\ &= \left(\int p\cdot\rho^{\gamma} d\mathbf{x}\right)^{\frac{1}{\gamma}-1} p\rho^{\gamma-1}\ ,\end{aligned}$$

such that $\frac{\delta D_\gamma(p||\rho)}{\delta\rho}$ finally yields

$$\frac{\delta D_\gamma(p||\rho)}{\delta\rho} = \frac{\rho^{\gamma}}{\left(\int \rho^{\gamma+1} d\mathbf{x}\right)} - \frac{p\rho^{\gamma-1}}{\left(\int p\cdot\rho^{\gamma} d\mathbf{x}\right)} \tag{3.13}$$

$$= \rho^{\gamma-1}\left[\frac{\rho}{\left(\int \rho^{\gamma+1} d\mathbf{x}\right)} - \frac{p}{\left(\int p\cdot\rho^{\gamma} d\mathbf{x}\right)}\right]\ . \tag{3.14}$$

Considering the important special case $\gamma = 1$, the Fréchet-derivative of the Cauchy-Schwarz-divergence (2.27) is derived:

$$\frac{\delta D_{CS}(p||\rho)}{\delta\rho} = \frac{\rho}{\left(\int \rho^2 d\mathbf{x}\right)} - \frac{p}{V(p,\rho)}\ . \tag{3.15}$$

4 Divergence based Vector Quantization using Fréchet-derivatives

Supervised and unsupervised vector quantization frequently are described in terms of dissimilarities or distances. Suppose, data are given as data vectors $\mathbf{v} \in \mathbb{R}^n$. We further assume that the vectors are discrete representations of continuous positive valued functions $p(x)$ with $v_i = p(x_i)$, $i = 1\ldots n$.

We now focus on prototype based vector quantization, i.e. data processing (clustering or classification) is realized using prototypes $\mathbf{w} \in \mathbb{R}^n$ as representatives, whereby the dissimilarity between data points as well as between data and prototypes are determined by dissimilarity measures ξ (not necessarily fulfilling triangle inequality or symmetry restrictions).

Frequently, such algorithms optimize a somewhat cost function E depending on the dissimilarity between the data points and the prototypes, i.e. usually one has $E = E(\xi(\mathbf{v}_i, \mathbf{w}_k))$ and $i = 1 \ldots N$ the number of data and $k = 1 \ldots C$ the number of prototypes. This cost function can be a variant of the usual classification error in supervised learning or modified mean squared error of the dissimilarities $\xi(\mathbf{v}_i, \mathbf{w}_k)$.

If $E = E(\xi(\mathbf{v}_i, \mathbf{w}_k))$ is differentiable with respect to ξ, and ξ differentiable with respect to the prototype $\mathbf{w}$, then a stochastic gradient minimization is a widely used optimization scheme for E. This methodology implies the calculation of the derivatives $\frac{\partial \xi}{\partial \mathbf{w}_k}$, which has now to be considered in the light of the above functional analytic investigations for divergence measures.

If we identify the prototypes as discrete realizations of a function $\rho(x)$ and further require that p and ρ are positive functions (measures), the dissimilarity measure ξ can be chosen as a discrete variant of a divergence. The derivative $\frac{\partial \xi}{\partial \mathbf{w}}$ has to be replaced in this scenario by the Fréchet-derivative $\frac{\delta \xi}{\delta \rho}$ in the continuous case, which reduces to usual derivatives in the discrete case (see remark in sec. 3.1.1). This is formally achieved by replacing p and ρ by their vectorial counterparts $\mathbf{v}$ and $\mathbf{w}$ in the formulae of the divergences provided in sec. 3.2 and further translating integrals into sums.

In the following we give prominent examples of unsupervised and supervised vector quantization, which can be optimized by gradient methods using the above introduced frame work.

4.1 Unsupervised Vector Quantization

4.1.1 Basic Vector Quantization

Unsupervised vector quantization is a class of algorithm for distributing prototypes $W = \{\mathbf{w}_k\}_A$, $\mathbf{w}_k \in \mathbb{R}^n$ such that data points $\mathbf{v} \in V \subseteq \mathbb{R}^n$ are faithfully represented in terms of a dissimilarity measure ξ. Thereby, $C = card(A)$ is the cardinality of the index set A. More formally, the data point $\mathbf{v}$ is represented by this prototype $\mathbf{w}_{s(\mathbf{v})}$ minimizing the dissimilarity $\xi(\mathbf{v}, \mathbf{w}_k)$, i.e.

$$\mathbf{v} \mapsto s(\mathbf{v}) = \operatorname*{argmin}_{k \in A} \xi(\mathbf{v}, \mathbf{w}_k) \ . \tag{4.1}$$

The aim of the algorithm is to distribute the prototypes in such a way that the quantization error

$$E_{\mathrm{VQ}} = \frac{1}{2} \int P(\mathbf{v})\, \xi\left(\mathbf{v}, \mathbf{w}_{s(\mathbf{v})}\right) d\mathbf{v} \tag{4.2}$$

is minimized. In its simplest form basic vector quantization (VQ) leads to a (stochastic) gradient descent on E_{VQ} with

$$\triangle\, \mathbf{w}_{s(\mathbf{v})} = -\varepsilon \cdot \frac{\partial \xi\left(\mathbf{v}, \mathbf{w}_{s(\mathbf{v})}\right)}{\partial \mathbf{w}_{s(\mathbf{v})}} \tag{4.3}$$

for prototype update of the winning prototype $\mathbf{w}_{s(\mathbf{v})}$ according to (4.1), also known as the online variant of LBG-algorithm (C–means) [39],[65]. Here, ε is a small positive value called learning rate. As we see, the update (4.3) take into account the derivative of the dissimilarity measure ξ with respect to the prototype. Beside the common choice of ξ being the squared Euclidean distance, the choice is given to the user with the restriction of differentiability. Hence, we are here allowed to apply divergences using its derivatives in the sense of Fréchet-derivatives.

4.1.2 Self-Organizing Maps and Neural Gas

There exist several variants of the basic vector quantization scheme to avoid local minima or to realize a projective mapping. For example, the latter can be obtained introducing an topological structure in A, usually a regular grid structure. The resulting vector quantization scheme is the Self-Organizing *Map (SOM)* introduced by T. KOHONEN [31]. The respective cost function (in the variant of Heskes, [21]) is

$$E_{\text{SOM}} = \frac{1}{2K(\sigma)} \int P(\mathbf{v}) \sum_{\mathbf{r}\in A} \delta_{\mathbf{r}}^{s(\mathbf{v})} \sum_{\mathbf{r}'\in A} h_{\sigma}^{SOM}(\mathbf{r},\mathbf{r}')\xi(\ \mathbf{v},\mathbf{w}_{\mathbf{r}'})d\mathbf{v} \tag{4.4}$$

with the so-called neighborhood function

$$h_{\sigma}^{SOM}(\mathbf{r},\mathbf{r}') = \exp\left(-\frac{\|\mathbf{r}-\mathbf{r}'\|_A}{2\sigma^2}\right)$$

and $\|\mathbf{r}-\mathbf{r}'\|_A$ is the distance in A according to the topological structure. $K(\sigma)$ is a normalization constant depending on the neighborhood range σ. For this SOM, the mapping rule (4.1) is modified to

$$\mathbf{v} \mapsto s\left(\mathbf{v}\right) = \underset{\mathbf{r}\in A}{\operatorname{argmin}} \sum_{\mathbf{r}'\in A} h_{\sigma}^{SOM}(\mathbf{r},\mathbf{r}') \cdot \xi\left(\mathbf{v},\mathbf{w}_{\mathbf{r}'}\right) \tag{4.5}$$

which yields in the limit $\sigma \to 0$ the original mapping (4.1). The prototype update for *all* prototypes then is given as [21]:

$$\triangle\,\mathbf{w}_{\mathbf{r}} = -\varepsilon h_{\sigma}^{SOM}\left(\mathbf{r}, s(\mathbf{v})\right) \frac{\partial \xi\left(\mathbf{v},\mathbf{w}_{\mathbf{r}}\right)}{\partial \mathbf{w}_{\mathbf{r}}}. \tag{4.6}$$

As above, the utilization of divergence based update is straightforward also for SOM.

If the aspect of projective mapping can be ignored while keeping the neighborhood cooperativeness aspect to avoid local minima in vector quantization, then the Neural Gas algorithm (NG) is an alternative to SOM presented by T. MARTINETZ [41]. The cost function of NG to be minimized writes as

$$E_{NG} = \frac{1}{2C\left(\sigma\right)} \sum_{j\in A} \int P\left(\mathbf{v}\right) h_{\sigma}^{NG}\left(\mathbf{v}, W, j\right) \xi\left(\mathbf{v},\mathbf{w}_j\right) d\mathbf{v} \tag{4.7}$$

with

$$h_{\sigma}^{NG}\left(\mathbf{v}, W, i\right) = \exp\left(-\frac{k_i\left(\mathbf{v}, W\right)}{\sigma}\right), \tag{4.8}$$

with the rank function

$$k_i\left(\mathbf{v}, W\right) = \sum_j \theta\left(\xi\left(\mathbf{v},\mathbf{w}_i\right) - \xi\left(\mathbf{v},\mathbf{w}_j\right)\right)\ . \tag{4.9}$$

The mapping is realized as in basic VQ (4.1) and the prototype update for all prototypes is similar to that of SOM

$$\triangle\,\mathbf{w}_i = -\varepsilon h_{\sigma}^{NG}\left(\mathbf{v}, W, i\right) \frac{\partial \xi\left(\mathbf{v},\mathbf{w}_i\right)}{\partial \mathbf{w}_i}. \tag{4.10}$$

Again, the incorporation of divergences is obvious also for NG.

4.1.3 Further vector quantization approaches

There exist a long list of other vector quantization approaches, like kernelized SOMs [22],[24],[23], Generative Topographic Mapping (GTM) [6], Soft Topographic Mapping [18] etc. to name just a few. Most of them utilize the Euclidean metric and the respective derivatives for adaptation. Thus, the idea of divergence based processing can be transferred to these in a similar manner.

Yet, recently, a somewhat reverse SOM has been proposed for embedding data into an embedding space $\mathcal{S}$, called *Exploration Machine (XOM)* [64]. This XOM can be seen as a projective structure preserving mapping of the input data into the embedding space and shows, therefore, similarities to MDS. In the XOM approach the data points $\mathbf{v}_k \in V \subseteq \mathbb{R}^n$, $k = 1, \ldots, N$ are uniquely associated with prototypes $\mathbf{w}_k \in \mathcal{S}$ in the embedding space $\mathcal{S}$ and $W = \{\mathbf{w}_k\}_{k=1}^N$. The dissimilarity $\xi_{\mathcal{S}}$ in the embedding space usually is chosen to be the quadratic Euclidean metric. Further, a *hypothesis* about the topological structure of the data $\mathbf{v}_k$ to be embedded is formulated for the embedding space $\mathcal{S}$ by defining a probability distribution $P_{\mathcal{S}}(\mathbf{s})$ for so-called *sampling vectors* $\mathbf{s} \in \mathcal{S}$. The cost function of XOM is defined as

$$E_{\mathrm{XOM}} = \frac{1}{2K(\sigma)} \int_{\mathcal{S}} P_{\mathcal{S}}(\mathbf{s}) \sum_{k=1}^{N} \cdot \delta_k^{k^*(\mathbf{s})} \sum_{j=1}^{N} h_\sigma^{XOM}(\mathbf{v}_k, \mathbf{v}_j) \cdot \xi_{\mathcal{S}}(\,\mathbf{s}, \mathbf{w}_j) d\mathbf{s} \tag{4.11}$$

with the mapping rule

$$k^*(\mathbf{s}) = \underset{i=1,\ldots,N}{\operatorname{argmin}} \sum_{j=1}^{N} h_\sigma^{XOM}(\mathbf{v}_k, \mathbf{v}_j) \cdot \xi_{\mathcal{S}}(\,\mathbf{s}, \mathbf{w}_j). \tag{4.12}$$

As in usual SOMs, the neighborhood cooperativeness is given by a Gaussian

$$h_\sigma^{XOM}(\mathbf{v}_k, \mathbf{v}_j) = \exp\left(-\frac{\xi_V\left(\mathbf{v}_k, \mathbf{v}_j\right)}{2\sigma^2}\right)$$

with the data dissimilarity $\xi_V\left(\mathbf{v}_k, \mathbf{v}_j\right)$ defined as Euclidean distance in the original XOM. The update of the prototypes in the embedding space is obtained in complete analogy to SOM as

$$\triangle\, \mathbf{w}_i = -\varepsilon h_\sigma^{XOM}\left(\mathbf{v}_i, \mathbf{v}_{k^*(\mathbf{s})}\right) \frac{\partial \xi_{\mathcal{S}}\left(\mathbf{s}, \mathbf{w}_i\right)}{\partial \mathbf{w}_i}. \tag{4.13}$$

As one can see, we can apply divergences to both ξ_V and $\xi_{\mathcal{S}}$. In case of the latter one, the prototype update (4.13) has to be changed accordingly using the respective Fréchet-derivatives.

4.2 Learning Vector Quantization

Learning Vector Quantization (LVQ) is the supervised counterpart of basic VQ. Now the data $\mathbf{v} \in V \subseteq \mathbb{R}^n$ to be learned are equipped with class information $\mathbf{c}_{\mathbf{v}}$. Suppose, we have K classes, we define $\mathbf{c}_{\mathbf{v}} \in [0,1]^K$. If $\sum_{k=1}^K c_i = 1$ the labeling is *probabilistic* and *possibilistic* otherwise. In case of a probabilistic labeling with $\mathbf{c}_{\mathbf{v}} \in \{0,1\}^K$ the labeling is called *crisp*.

We now briefly explore, how divergences can be used also for supervised learning. Again we start with the widely applied basic LVQ-approaches and outline afterwards the procedure for some more sophisticated methods without any claim of completeness.

4.2.1 Basic LVQ algorithms

The basic LVQ-schemes are invented by T. KOHONEN [31]. For standard LVQ a crisp data labeling is assumed. Further, the prototypes $\mathbf{w}_j$ with labels y_j correspond to the K classes in such a way that at least one prototype is assigned to each class. For simplicity, we take exactly one prototype for each class now. The task is to distribute the prototypes in such a manner that the classification error is reduced. The respective algorithms LVQ1...LVQ3 are heuristically motivated.

As in the unsupervised vector quantization, the similarity between data and prototypes for LVQs are judged by a dissimilarity measure $\xi\left(\mathbf{v},\mathbf{w}_j\right)$. Beside some small modifications the basic LVQ-schemes LVQ1...LVQ3 mainly consist in determination of the most proximate prototype(s) $\mathbf{w}_{s(\mathbf{v})}$ for given $\mathbf{v}$ according to the mapping rule (4.1) and subsequent adaptation. Depending on the agreement of $\mathbf{c}_\mathbf{v}$ and $y_{s(\mathbf{v})}$ the adaptation of the prototype(s) takes place according to

$$\triangle\,\mathbf{w}_{s(\mathbf{v})}=\alpha\cdot\varepsilon\cdot\frac{\partial\xi\left(\mathbf{v},\mathbf{w}_{s(\mathbf{v})}\right)}{\partial\mathbf{w}_{s(\mathbf{v})}}\tag{4.14}$$

and $\alpha=1$ iff $\mathbf{c}_\mathbf{v}=y_{s(\mathbf{v})}$, and $\alpha=-1$ otherwise.

A popular generalization of these standard algorithms is the *generalized LVQ (GLVQ)* introduced by SATO&YAMADA [52]. In GLVQ the classification error is replaced by a dissimilarity based cost function, which is, of course, closely related to the classification error but not identical.

For a given data point $\mathbf{v}$ with class label $\mathbf{c}_\mathbf{v}$ the two best matching prototypes with respect to the data metric ξ, usually the quadratic Euclidian, are determined: $\mathbf{w}_{s^+(\mathbf{v})}$ has minimum distance $\xi^+=\xi\left(\mathbf{v},\mathbf{w}_{s^+(\mathbf{v})}\right)$ under the constraint that the class labels are identically: $\mathbf{y}_{s^+(\mathbf{v})}=\mathbf{c}_\mathbf{v}$. The other best prototype $\mathbf{w}_{s^-(\mathbf{v})}$ has has minimum distance $\xi^-=\xi\left(\mathbf{v},\mathbf{w}_{s^-(\mathbf{v})}\right)$ supposing the class labels are different: $\mathbf{y}_{s^-(\mathbf{v})}\neq\mathbf{c}_\mathbf{v}$. Then the classifier function $\mu\left(\mathbf{v}\right)$ is defined as

$$\mu\left(\mathbf{v}\right)=\frac{\xi^+-\xi^-}{\xi^++\xi^-}\tag{4.15}$$

being negative in case of a correct classification. The value $\xi^+-\xi^-$ yields the hypothesis margin of the classifier [10]. Then the *generalized* LVQ (GLVQ) is derived as gradient descent on the cost function

$$E_{\mathrm{GLVQ}}=\sum_{\mathbf{v}}\mu\left(\mathbf{v}\right)\tag{4.16}$$

with respect to the prototypes. In each learning step, for a given data point, both $\mathbf{w}_{s^+(\mathbf{v})}$ and $\mathbf{w}_{s^-(\mathbf{v})}$ are adapted in parallel. Taking the derivatives $\frac{\partial E_{\mathrm{GLVQ}}}{\partial\mathbf{w}_{s^+(\mathbf{v})}}$ and $\frac{\partial E_{\mathrm{GLVQ}}}{\partial\mathbf{w}_{s^-(\mathbf{v})}}$ we get for the updates

$$\triangle\mathbf{w}_{s^+(\mathbf{v})}=\epsilon^+\cdot\theta^+\cdot\frac{\partial\xi\left(\mathbf{v},\mathbf{w}_{s^+(\mathbf{v})}\right)}{\partial\mathbf{w}_{s^+(\mathbf{v})}}\tag{4.17}$$

and

$$\triangle\mathbf{w}_{s^-(\mathbf{v})}=-\epsilon^-\cdot\theta^-\cdot\frac{\partial\xi\left(\mathbf{v},\mathbf{w}_{s^-(\mathbf{v})}\right)}{\partial\mathbf{w}_{s^-(\mathbf{v})}}\tag{4.18}$$

with the scaling factors

$$\theta^{+}=\frac{2\cdot\xi^{-}}{\left(\xi^{+}+\xi^{-}\right)^{2}} \text{ and } \theta^{-}=\frac{2\cdot\xi^{+}}{\left(\xi^{+}+\xi^{-}\right)^{2}}\ . \tag{4.19}$$

The values ϵ^{+} and $\epsilon^{-}\in(0,1)$ are the learning rates.

Obviously, the distance measure ξ could be replaced for all these LVQ schemes by one of the introduced divergences. This offers a new possibility for information theoretic learning in classification schemes, which differs from the previous approaches significantly. These earlier approaches stress the information optimum class representation whereas here the expected information loss in terms of the applied divergence measure is optimized [57],[56],[61].

4.2.2 Advanced Learning Vector Quantization

Apart from the basic LVQ schemes, many more sophisticated prototype based learning schemes are proposed for classification learning. Here we will only restrict ourself to such approaches which can deal with probabilistic or possibilistic labeled training data (uncertain decisions), and which are additionally related to the basic unsupervised and supervised vector quantization algorithms mentioned in this paper so far.

In particular, we focus on the Fuzzy-labeled SOM (FLSOM) and the very similar Fuzzy-labeled NG (FLNG) [63],[60]. Both approaches extend the cost function of its unsupervised counterpart in the following shorthand manner

$$E_{\mathrm{FLSOM/FLNG}}=(1-\beta)\,E_{\mathrm{SOM/NG}}+\beta E_{\mathrm{FL}}$$

where E_{FL} measures the classification accuracy . The factor $\beta\in[0,1]$ is a factor balancing unsupervised and supervised learning. The classification accuracy term E_{FL} is defined as

$$E_{\mathrm{FL}}=\frac{1}{2}\int P\left(\mathbf{v}\right)\sum_{\mathbf{r}}g_{\gamma}\left(\mathbf{v},\mathbf{w_r}\right)\psi\left(\mathbf{c_v},\mathbf{y_r}\right)d\mathbf{v} \tag{4.20}$$

where $g_{\gamma}\left(\mathbf{v},\mathbf{w_r}\right)$ is a Gaussian kernel describing a neighborhood range in the data space

$$g_{\gamma}\left(\mathbf{v},\mathbf{w_r}\right)=\exp\left(-\frac{\xi\left(\mathbf{v},\mathbf{w_r}\right)}{2\gamma^{2}}\right)\ . \tag{4.21}$$

using the dissimilarity $\xi\left(\mathbf{v},\mathbf{w_r}\right)$ in the data space. $\psi\left(\mathbf{c_v},\mathbf{y_r}\right)$ judges the dissimilarities between label vectors of data and prototypes. $\psi\left(\mathbf{c_v},\mathbf{y_r}\right)$ is originally suggested to be the quadratic Euclidean distance.

Note that E_{FL} depends on the dissimilarity in the data space $\xi\left(\mathbf{v},\mathbf{w_r}\right)$ via $g_{\gamma}\left(\mathbf{v},\mathbf{w_r}\right)$. Hence, prototype adaptation in FLSOM/FLNG is influenced by the classification accuracy:

$$\frac{\partial E_{\mathrm{FLSOM/NG}}}{\partial\mathbf{w_r}}=\frac{\partial E_{\mathrm{SOM/NG}}}{\partial\mathbf{w_r}}+\frac{\partial E_{\mathrm{FL}}}{\partial\mathbf{w_r}} \tag{4.22}$$

which yields

$$\begin{aligned}\triangle\,\mathbf{w_r} &= -\epsilon(1-\beta)\cdot h_{\sigma}^{SOM/NG}\left(\mathbf{r},s(\mathbf{v})\right)\frac{\partial\xi\left(\mathbf{v},\mathbf{w_r}\right)}{\partial\mathbf{w_r}} \\ &\quad +\epsilon\beta\frac{1}{4\gamma^{2}}\cdot g_{\gamma}\left(\mathbf{v},\mathbf{w_r}\right)\frac{\partial\xi\left(\mathbf{v},\mathbf{w_r}\right)}{\partial\mathbf{w_r}}\psi\left(\mathbf{c_v},\mathbf{y_r}\right)\ .\end{aligned} \tag{4.23}$$

The label adaptation is only influenced by the second part E_{FL}. The derivative $\frac{\partial E_{\mathrm{FL}}}{\partial \mathbf{y_r}}$ yields

$$\triangle\, \mathbf{y_r} = \epsilon_l \beta \cdot g_\gamma\left(\mathbf{v}, \mathbf{w_r}\right) \frac{\partial \psi\left(\mathbf{c_v}, \mathbf{y_r}\right)}{\partial \mathbf{y_r}} \tag{4.24}$$

with learning rate $\epsilon_l > 0$ [63],[60]. This label learning leads to a weighted average $\mathbf{y_r}$ of the fuzzy labels $\mathbf{c_v}$ of those data $\mathbf{v}$, which are close to the associated prototypes according to $\xi\left(\mathbf{v}, \mathbf{w_r}\right)$.

It should be noted at this point that a similar approach can easily be installed also for XOM in an analog manner yielding FLXOM.

Clearly, beside the possibility of choosing a divergence measure for $\xi\left(\mathbf{v}, \mathbf{w_r}\right)$ as in the unsupervised case, there is no contradiction to do so also for the label dissimilarity $\psi\left(\mathbf{c_v}, \mathbf{y_r}\right)$ in these FL-methods. As before, the simple plug-in of the respective discrete divergence variants and their Fréchet-derivatives modifies the algorithms such that a semi-supervised learning can be proceeded relying on divergences for both variants.

5 Simulations for $\alpha-$, $\beta-$, $\gamma-$, and $\eta-$ divergences

In this chapter we demonstrate the prototype based unsupervised vector quantization learning using the derivatives for several parametrized divergences to illustrate the influence of the parameter setting. For this purpose, we consider a Heskes-SOM according to (4.4) using a chain of 100 prototypes $\mathbf{w_r} \in \mathbb{R}^2$. The data are 10^7 data points $\mathbf{v} = (v_1, v_2) \in [0,1]^2$ with the additionally constraint that $v_1 + v_2 = 1$, i.e. the data $\mathbf{v}$ can be taken as 'probability densities' in $\mathbb{R}^2$. Further, the first argument v_1 is chosen randomly according to the data density $P_1(v_1) = 2 \cdot v_1$. The learning rate ε as well as the neighborhood range σ converged during the learning to the final values $\varepsilon_{final} = 0.0001$ and $\sigma_{final} = 1$,respectively, and appropriately chosen initial values.

We trained SOM-networks for each of the parametrized $\alpha-$, $\beta-$, $\gamma-$, and $\eta-$ divergences as introduced in sect. 2 using the Fréchet-derivatives deduced in sec. 3.1.1 with different parameter values.

For the η-divergences (belonging to the Bregman-divergences) the results are depicted in Fig. 5

One can observe that the influence of the parameter η is only marginal. Yet, small variations can be detected. For the special choice $\eta = 2$ Euclidean learning is realized.

For the β-divergences the influence of the parameter value β is stronger than parameter effect for η-divergences, see Fig. 5.

In particular, significant deviations can be observed for higher prototype w_1-values, giving the hint for a better discrimination property for this probability range.

The α-divergence based learning shows a similar behaviour as known from η-divergences with respect to the parameter variation, i.e. the divergence is relativily robust with respect to the control parameter α, see Fig. 5.

The γ-divergences show the most sensitive behaviour of all parametrized divergences investigated here, Fig. 5. In particular, the choice of the control parameter γ influence both ranges of probability the low and the high one with approximately the same sensitivity, see Fig.5.

Thus, it differs from the sensitivity observed for β-divergences. This behaviour offers the possibility to tune precisely the divergence depending on the specific vec-

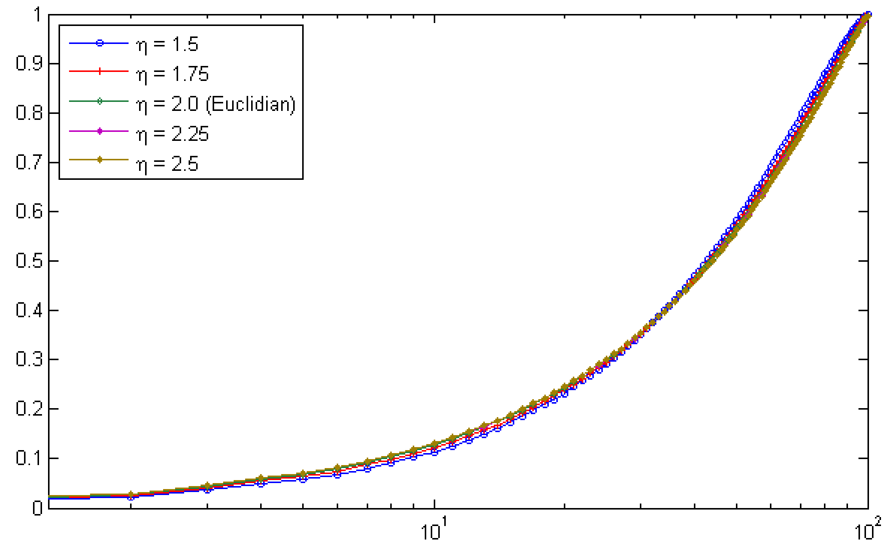


Figure 2: Prototype distribution for η-divergence based SOM for different η-values. Horizontal axis: logarithmic value of the one-dimensional prototype index. Vertical axis: first component w_1 of the prototypes $\mathbf{w} = (w_1, w_2)$.

tor quantization task. Together with stated robustness of the γ-divergence at all, as mentioned in [17], this adaptive specifity could provide high potential for a wide range of application. This is underlyed by the applications in supervised and unsupervised vector quantization based on the Cauchy-Scharz- divergence ($\gamma = 1$) [28],[43],[45],[59].

6 Extensions for the basic adaptation scheme – hyperparameter and relevance learning

6.1 Hyperparameter learning for $\alpha-$, $\beta-$, $\gamma-$, and $\eta-$ divergences

Considering the parametrized divergence families of γ-, α-, β-, and η-divergences, one could further think about the optimal choice of the so-called hyperparameters γ, α, β, η as suggested in a similar manner for other parametrized LVQ-algorithms [53]. In case of supervised learning schemes for classification based on differentiable cost functions, the optimization can be handled as an object of a gradient descent based adaptation procedure. Thus, the parameter is optimized in dependence of the classification task at hand.

Suppose, the classification accuracy for a certain approach is given as

$$E = E\left(\xi_\theta, W\right)$$

depending on a *parametrized divergence* ξ_θ with parameter θ. If E and ξ_θ are both differentiable with respect to θ according to

$$\frac{\partial E\left(\xi_\theta, W\right)}{\partial \theta} = \frac{\partial E}{\partial \xi_\theta} \cdot \frac{\partial \xi_\theta}{\partial \theta},$$

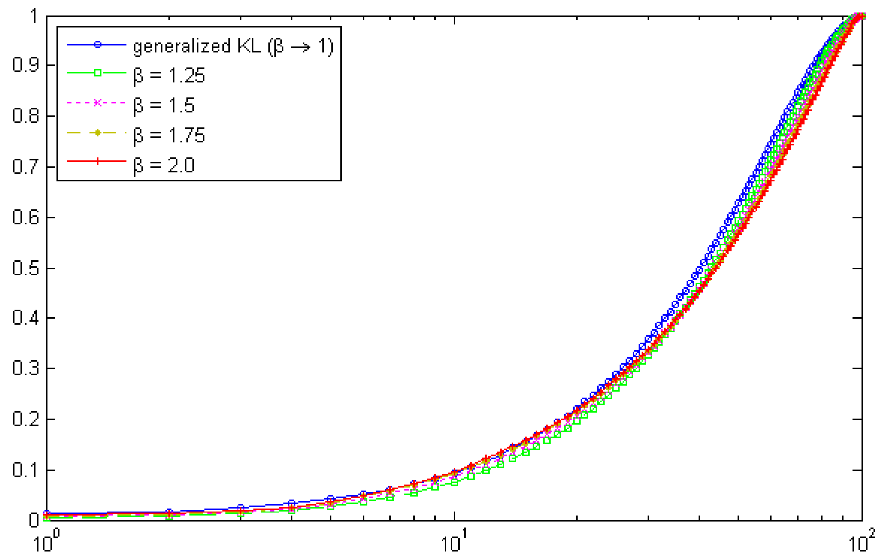


Figure 3: Prototype distribution for β-divergence based SOM for different β-values. Horizontal axis: logarithmic value of the one-dimensional prototype index. Vertical axis: first component w_1 of the prototypes $\mathbf{w} = (w_1, w_2)$.

a gradient based optimization is derived by

$$\triangle\theta = -\varepsilon \frac{\partial E\left(\xi_\theta, W\right)}{\partial\theta} = -\varepsilon \frac{\partial E}{\partial \xi_\theta} \cdot \frac{\partial \xi_\theta}{\partial \theta}$$

depending on the derivative $\frac{\partial \xi_\theta}{\partial \theta}$ for a certain choice of the divergence ξ_θ.

We assume in the following that the (positive) measures p and ρ are continuously differentiable. Than, considering derivatives of parametrized divergences $\frac{\partial \xi_\theta}{\partial \theta}$ with respect to the parameter θ, it is allowed to interchange integration and differentiation, if the resulting integral exists [14]. Hence, we can differentiate parametrized divergences with respect to their hyperparameter in that case. For the several α-, β-, γ-, and η-divergences, characterized in sec. 2, we obtain after some elementary calculations:

- η-divergence $D_\eta\left(p||\rho\right)$ from (2.7)

$$\frac{\partial D_\eta\left(p||\rho\right)}{\partial \eta} = \int p^\eta \ln p + \rho^{\eta-1} \cdot \left(\rho - p + \left(\eta\rho - \rho - \eta p\right) \cdot \ln \rho\right) d\mathbf{x}$$

- β-divergence $D_\beta\left(p||\rho\right)$ from (2.8) – see Appendix:

$$\begin{aligned}\frac{\partial D_\beta\left(p||\rho\right)}{\partial \beta} &= \frac{1}{\beta-1}\int p\left(p^{\beta-1}\ln p - \rho^{\beta-1}\ln\rho - \frac{\left(p^{\beta-1}-\rho^{\beta-1}\right)}{\left(\beta-1\right)}\right) d\mathbf{x} \\ &\quad - \int \left(p^\beta \ln p - \rho^\beta \ln \rho\right)\frac{1}{\beta} - \frac{1}{\beta^2}\left(p^\beta - \rho^\beta\right) d\mathbf{x}\end{aligned}$$

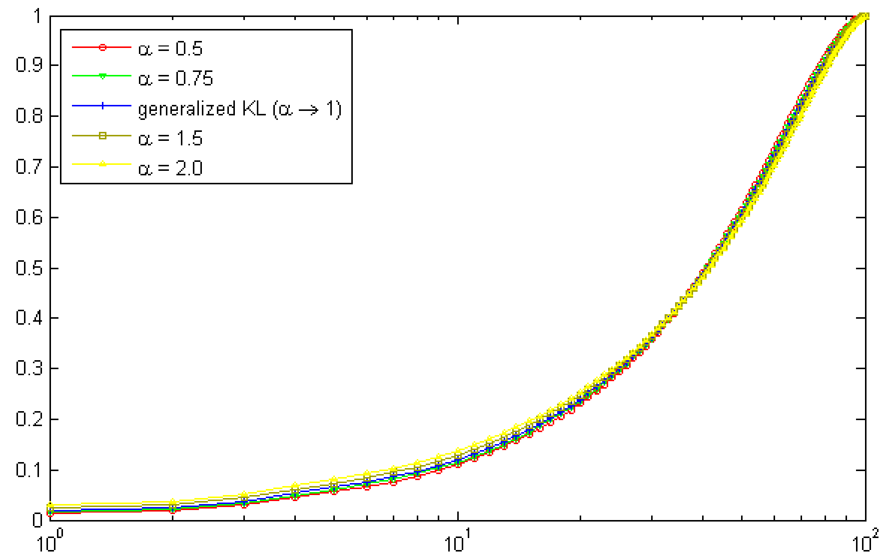


Figure 4: Prototype distribution for α-divergence based SOM for different α-values. Horizontal axis: logarithmic value of the one-dimensional prototype index. Vertical axis: first component w_1 of the prototypes $\mathbf{w} = (w_1, w_2)$.

- α-divergence $D_\alpha(p||\rho)$ from (2.15) – see Appendix:

$$\begin{aligned}\frac{\partial D_\alpha(p||\rho)}{\partial\alpha} &= -\frac{(2\alpha-1)}{\alpha^2(\alpha-1)^2}\int\left[p^\alpha\rho^{1-\alpha}-\alpha\cdot p+(\alpha-1)\rho\right]d\mathbf{x}\\ &+\frac{1}{\alpha(\alpha-1)}\int p^\alpha\rho^{1-\alpha}(\ln p-\ln\rho)-p+\rho d\mathbf{x}\end{aligned}$$

- Tsallis-divergence $D_\alpha^T(p||\rho)$ from (2.19)

$$\begin{aligned}\frac{\partial D_\alpha^T(p||\rho)}{\partial\alpha} &= \frac{1}{(1-\alpha)^2}\left(\int p^\alpha\rho^{1-\alpha}d\mathbf{x}-1\right)\\ &+\frac{1}{(1-\alpha)}\int p^\alpha\rho^{1-\alpha}(\ln p-\ln\rho)\,d\mathbf{x}\end{aligned}$$

- generalized Rényi-divergence $D_\alpha^{GR}(p||\rho)$ from (2.22) – see Appendix:

$$\begin{aligned}\frac{\partial D_\alpha^{GR}(p||\rho)}{\partial\alpha} &= -\frac{1}{(\alpha-1)^2}\log\left(\int\left[p^\alpha\rho^{1-\alpha}-\alpha\cdot p+(\alpha-1)\rho+1\right]d\mathbf{x}\right)\\ &+\frac{1}{\alpha-1}\frac{\int p^\alpha\rho^{1-\alpha}(\ln p-\ln\rho)-p+\rho d\mathbf{x}}{\int\left[p^\alpha\rho^{1-\alpha}-\alpha\cdot p+(\alpha-1)\rho+1\right]d\mathbf{x}}\end{aligned}$$

- Rényi-divergence $D_\alpha^R(p||\rho)$ from (2.23)

$$\begin{aligned}\frac{\partial D_\alpha^{GR}(p||\rho)}{\partial\alpha} &= -\frac{1}{(\alpha-1)^2}\log\left(\int p^\alpha\rho^{1-\alpha}d\mathbf{x}\right)\\ &+\frac{1}{\alpha-1}\frac{\int p^\alpha\rho^{1-\alpha}(\ln p-\ln\rho)\,d\mathbf{x}}{\int p^\alpha\rho^{1-\alpha}d\mathbf{x}}\end{aligned}$$

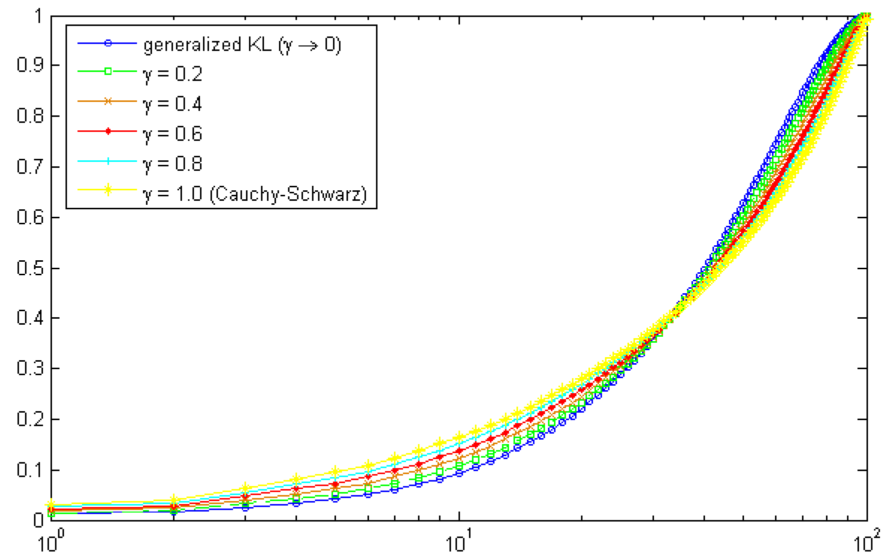


Figure 5: Prototype distribution for γ-divergence based SOM for different γ-values. Horizontal axis: logarithmic value of the one-dimensional prototype index. Vertical axis: first component w_1 of the prototypes $\mathbf{w} = (w_1, w_2)$.

- γ-divergence $D_\gamma(p||\rho)$ from (2.25) – see Appendix:

$$\begin{aligned}\frac{\partial D_\gamma(p||\rho)}{\partial\gamma} &= -\frac{(2\gamma+1)}{\gamma^2(\gamma+1)^2}\ln\left(\int p^{\gamma+1}d\mathbf{x}\right)+\frac{\int p^{\gamma+1}\ln p d\mathbf{x}}{(\gamma+1)\gamma\int p^{\gamma+1}d\mathbf{x}}\\ &\quad -\frac{1}{(\gamma+1)^2}\ln\left(\int \rho^{\gamma+1}d\mathbf{x}\right)+\frac{\int \rho^{\gamma+1}\ln\rho d\mathbf{x}}{(\gamma+1)\int \rho^{\gamma+1}d\mathbf{x}}\\ &\quad +\frac{1}{\gamma^2}\ln\left(\int p\cdot\rho^{\gamma}d\mathbf{x}\right)-\frac{\int p\rho^{\gamma}\ln\rho d\mathbf{x}}{\gamma\int p\cdot\rho^{\gamma}d\mathbf{x}}\end{aligned}$$

6.2 Relevance learning for divergences

Density functions are required to fulfill the normalization condition whereas positive measure are more flexible. This offers the possibility to transfer the idea of relevance learning also to divergence based learning vector quantization. *Relevance learning* in learning vector quantization is weighting the input data dimensions such that classification accuracy is improved [19].

In the framework of divergence based gradient descent learning we multiplicatively weight a positive measure $q(x)$ by $\lambda(x)$ with $0 \leq \lambda(\mathbf{x}) < \infty$ and the regularization condition $\int\lambda(\mathbf{x})\,dx = 1$. Incorporating this idea into the above approaches we have to replace in the divergences p by $p\cdot\lambda$ and ρ by $\rho\cdot\lambda$. Doing so we can optimize $\lambda(x)$ during learning for better performance by gradient descent optimization as it is known from vectorial relevance learning. This leads here, again, to Fréchet-derivatives of the divergences but now with respect to the weighting function $\lambda(\mathbf{x})$.

In particular we obtain for the Bregman divergence

$$\frac{\delta D_\Phi^B(\lambda\cdot p||\lambda\cdot\rho)}{\delta\lambda}=\frac{\Phi(\lambda\cdot p)}{\delta\lambda}-\frac{\Phi(\lambda\cdot\rho)}{\delta\lambda}-\frac{\delta\left[\frac{\delta\Phi(\lambda\cdot\rho)}{\delta\rho}\lambda(p-\rho)\right]}{\delta\lambda} \tag{6.1}$$

with

$$\frac{\delta\left[\frac{\delta\Phi(\lambda\cdot\rho)}{\delta\rho}\lambda\left(p-\rho\right)\right]}{\delta\lambda}=\left(p-\rho\right)\left(\frac{\delta^{2}\left[\Phi\left(\lambda\cdot\rho\right)\right]}{\delta\rho\,\delta\lambda}\lambda+\frac{\delta\Phi\left(\lambda\cdot\rho\right)}{\delta\rho}\right)\,.$$

This yields for the *generalized* Kullback-Leibler-divergence

$$\frac{\delta D_{GKL}\left(\lambda\cdot p||\lambda\cdot\rho\right)}{\delta\lambda}=p\cdot\log\left(\frac{p}{\rho}\right)-p+\rho\,. \tag{6.2}$$

In case of the η-divergences (2.7) we calculate

$$\frac{\delta D_{\eta}\left(\lambda\cdot p||\lambda\cdot\rho\right)}{\delta\lambda}=\lambda^{\eta-1}\eta\left(p^{\eta}-\rho^{\eta-1}\left(p\eta+\left(1-\eta\right)\rho\right)\right) \tag{6.3}$$

which reduces for the choice $\eta=2$ (Euclidean distance) to

$$\frac{\delta D_{\eta}\left(\lambda\cdot p||\lambda\cdot\rho\right)}{\delta\lambda}=2\lambda\left(p-\rho\right)^{2}$$

as it known from [19]. Further, for the β-divergences (2.8), also belonging to the Bregman divergence class, we have

$$\frac{\delta D_{\beta}\left(\lambda\cdot p||\lambda\cdot\rho\right)}{\delta\lambda}=\frac{\rho\cdot\left(\lambda\cdot p\right)^{\beta}+\left(\rho\cdot\left(\beta-1\right)-p\cdot\beta\right)\cdot\left(\lambda\cdot\rho\right)^{\beta}}{\lambda\rho\left(\beta-1\right)}\,. \tag{6.4}$$

For the class of f-divergences (2.10) we consider

$$\begin{aligned}\frac{\delta D_{f}\left(\lambda\cdot p||\lambda\cdot\rho\right)}{\delta\lambda} &= \rho\cdot f\left(\frac{p}{\rho}\right)+\lambda\cdot\rho\frac{\partial f\left(u\right)}{\partial u}\frac{\delta u}{\delta\lambda}\\ &= \rho\cdot f\left(\frac{p}{\rho}\right)\end{aligned} \tag{6.5}$$

with $u=\frac{p}{\rho}$ using the fact that $\frac{\delta u}{\delta\lambda}=0$. The relevance learning of the subclass of α-divergences (2.15) follows

$$\frac{\delta D_{\alpha}\left(\lambda\cdot p||\lambda\cdot\rho\right)}{\delta\lambda}=\frac{1}{\alpha\left(\alpha-1\right)}\left[\rho\cdot\left(\left(\frac{p}{\rho}\right)^{\alpha}+\alpha-1\right)-p\cdot\alpha\right]\,, \tag{6.6}$$

whereas the respective gradient of generalized Rényi-divergences (2.22) can be derived from this as

$$\frac{\delta D_{\alpha}^{GR}\left(\lambda\cdot p||\lambda\cdot\rho\right)}{\delta\lambda}=\frac{\alpha}{\int\left[\lambda\cdot\left(\rho\cdot\left(\frac{p}{\rho}\right)^{\alpha}-\alpha\cdot p+\left(\alpha-1\right)\cdot\rho\right)+1\right]d\mathbf{x}}\frac{\delta D_{\alpha}\left(\lambda\cdot p||\lambda\cdot\rho\right)}{\delta\lambda}\,. \tag{6.7}$$

The subset of Tsallis-divergences is treated by

$$\frac{\delta D_{\alpha}^{T}\left(\lambda\cdot p||\lambda\cdot\rho\right)}{\delta\lambda}=\frac{1}{1-\alpha}p^{\alpha}\rho^{1-\alpha}\,. \tag{6.8}$$

The γ-divergence classes finally yield

$$\frac{\delta D_{\gamma}\left(\lambda\cdot p||\lambda\cdot\rho\right)}{\delta\lambda}=\frac{p\left(\lambda\cdot p\right)^{\gamma}}{\gamma\int\left(\lambda\cdot p\right)^{\gamma+1}d\mathbf{x}}+\frac{\rho\left(\lambda\cdot\rho\right)^{\gamma}}{\int\left(\lambda\cdot\rho\right)^{\gamma+1}d\mathbf{x}}-\frac{p\cdot\left(\gamma+1\right)\cdot\left(\lambda\cdot\rho\right)^{\gamma}}{\gamma\int\left(\lambda\cdot p\right)\cdot\left(\lambda\cdot\rho\right)^{\gamma}d\mathbf{x}}\,.$$

Again the important special case $\gamma=1$ is considered: the relevance learning scheme for the Cauchy-Schwarz divergence (2.27) is derived as

$$\frac{\delta D_{CS}\left(\lambda\cdot p||\lambda\cdot\rho\right)}{\delta\lambda}=\frac{p\cdot\lambda\cdot p}{\int\left(\lambda\cdot p\right)^{2}d\mathbf{x}}+\frac{\rho\cdot\lambda\cdot\rho}{\int\left(\lambda\cdot\rho\right)^{2}d\mathbf{x}}-\frac{2\cdot p\cdot\lambda\cdot\rho}{\int\lambda^{2}\cdot p\cdot\rho d\mathbf{x}}\,. \tag{6.9}$$

7 Conclusion

In this article we provide the mathematical foundation for divergence based supervised and unsupervised vector quantization bearing on the derivatives of the applied divergences. For this purpose, we first characterized the main sub-classes of divergences, Bregman-, α-, β-, γ-, and f-divergences following [9]. The mathematical framework of Fréchet-derivatives is then used to calculate the functional divergence derivatives.

We exemplary show the utilization of this methodology for famous examples of supervised and unsupervised vector quantization including SOM, NG, and GLVQ. In particular, we explained that the divergences can be taken as suitable dissimilarity measures for data, which leads to the usage of the respective Fréchet-derivatives in the online learning schemes. Further, we declare, how a parameter adaptation could be integrated in supervised learning to achieve improved classification results in case of the parametrized α-, β-, γ-, and η-divergences. In the last step we considered a weighting function for generalized divergences based on positive measure. The optimization scheme for this weight function is obtained by Fréchet derivatives again to obtain a relevance learning scheme in analogy to relevance learning in usual supervised learning vector quantization [19].

8 Appendix – Calculation of the derivatives of the parametrized divergences with respect to the hyperparameters

We assume for the differentiation of the divergences with respect to their hyperparameters that the (positive) measures p and ρ are continuously differentiable. Than, considering derivatives of divergences, it is allowed to interchange integration and differentiation, if the resulting integral exists [14].

8.1 β-divergence

The β-divergence is according to (2.8)

$$\begin{aligned} D_\beta(p||\rho) &= \int p \cdot \frac{p^{\beta-1}-\rho^{\beta-1}}{\beta-1} d\mathbf{x} - \int \frac{p^\beta - \rho^\beta}{\beta} d\mathbf{x} \\ &= I_1(\beta) - I_2(\beta) \end{aligned}$$

We treat both integrals independently.

$$\begin{aligned} \frac{\partial I_1(\beta)}{\partial\beta} &= \int \frac{\partial\left[p \cdot \frac{p^{\beta-1}-\rho^{\beta-1}}{\beta-1}\right]}{\partial\beta} d\mathbf{x} \\ &= \int p\left(\frac{\partial\left[p^{\beta-1}-\rho^{\beta-1}\right]}{\partial\beta}\frac{1}{\beta-1} - \frac{\left(p^{\beta-1}-\rho^{\beta-1}\right)}{(\beta-1)^2}\right) d\mathbf{x} \\ &= \frac{1}{\beta-1}\int p\left(p^{\beta-1}\ln p - \rho^{\beta-1}\ln\rho - \frac{\left(p^{\beta-1}-\rho^{\beta-1}\right)}{(\beta-1)}\right) d\mathbf{x} \end{aligned}$$

$$\begin{aligned} \frac{\partial I_2(\beta)}{\partial\beta} &= \int \frac{\partial\left[\frac{p^\beta-\rho^\beta}{\beta}\right]}{\partial\beta} d\mathbf{x} \\ &= \int \frac{\partial\left[p^\beta-\rho^\beta\right]}{\partial\beta}\frac{1}{\beta} - \frac{1}{\beta^2}\left(p^\beta-\rho^\beta\right) d\mathbf{x} \\ &= \int \left(p^\beta\ln p - \rho^\beta\ln\rho\right)\frac{1}{\beta} - \frac{1}{\beta^2}\left(p^\beta-\rho^\beta\right) d\mathbf{x} \end{aligned}$$

Thus

$$\begin{aligned} \frac{\partial D_\beta(p||\rho)}{\partial\beta} = & \frac{1}{\beta-1}\int p\left(p^{\beta-1}\ln p - \rho^{\beta-1}\ln\rho - \frac{\left(p^{\beta-1}-\rho^{\beta-1}\right)}{(\beta-1)}\right) d\mathbf{x} \\ & - \int \left(p^\beta\ln p - \rho^\beta\ln\rho\right)\frac{1}{\beta} - \frac{1}{\beta^2}\left(p^\beta-\rho^\beta\right) d\mathbf{x} \end{aligned}$$

if the integral exists for an appropriate choice of β.

8.2 α-divergences

We consider the α-divergence (2.15)

$$\begin{aligned} D_\alpha(p\|\rho) &= \frac{1}{\alpha(\alpha-1)}\int\left[p^\alpha\rho^{1-\alpha}-\alpha\cdot p+(\alpha-1)\rho\right]d\mathbf{x}. \\ &= \frac{1}{\alpha(\alpha-1)}I(\alpha) \end{aligned}$$

We have

$$\begin{aligned} \frac{\partial D_\alpha(p\|\rho)}{\partial\alpha} &= \frac{\partial\left[\frac{1}{\alpha(\alpha-1)}\right]}{\partial\alpha}I(\alpha)+\frac{1}{\alpha(\alpha-1)}\frac{\partial I(\alpha)}{\partial\alpha} \\ &= -\frac{(2\alpha-1)}{\alpha^2(\alpha-1)^2}I(\alpha)+\frac{1}{\alpha(\alpha-1)}\frac{\partial I(\alpha)}{\partial\alpha} \end{aligned}$$

The derivative $\frac{\partial I(\alpha)}{\partial\alpha}$ yields

$$\begin{aligned} \frac{\partial I(\alpha)}{\partial\alpha} &= \int\frac{\partial\left[p^\alpha\rho^{1-\alpha}-\alpha\cdot p+(\alpha-1)\rho\right]}{\partial\alpha}d\mathbf{x} \\ &= \int p^\alpha\rho^{1-\alpha}(\ln p-\ln\rho)-p+\rho d\mathbf{x} \end{aligned}$$

and, finally we get

$$\begin{aligned} \frac{\partial D_\alpha(p\|\rho)}{\partial\alpha} &= -\frac{(2\alpha-1)}{\alpha^2(\alpha-1)^2}\int\left[p^\alpha\rho^{1-\alpha}-\alpha\cdot p+(\alpha-1)\rho\right]d\mathbf{x} \\ &\quad+\frac{1}{\alpha(\alpha-1)}\int p^\alpha\rho^{1-\alpha}(\ln p-\ln\rho)-p+\rho d\mathbf{x} \end{aligned}$$

8.3 Rényi-divergences

Considering the *generalized* Rényi-divergence $D_\alpha^{GR}(p\|\rho)$ from (2.22)

$$\begin{aligned} D_\alpha^{GR}(p\|\rho) &= \frac{1}{\alpha-1}\log\left(\int\left[p^\alpha\rho^{1-\alpha}-\alpha\cdot p+(\alpha-1)\rho+1\right]d\mathbf{x}\right) \\ &= \frac{1}{\alpha-1}\log I(\alpha) \end{aligned}$$

we get:

$$\frac{\partial D_\alpha^{GR}(p\|\rho)}{\partial\alpha}=-\frac{1}{(\alpha-1)^2}\log I(\alpha)+\frac{1}{\alpha-1}\frac{1}{I(\alpha)}\frac{\partial I(\alpha)}{\partial\alpha}$$

with

$$\begin{aligned} \frac{\partial I(\alpha)}{\partial\alpha} &= \int\frac{\partial\left[p^\alpha\rho^{1-\alpha}-\alpha\cdot p+(\alpha-1)\rho+1\right]}{\partial\alpha}d\mathbf{x} \\ &= \int p^\alpha\rho^{1-\alpha}(\ln p-\ln\rho)-p+\rho d\mathbf{x} \end{aligned}$$

Summarizing the differentiation yields

$$\begin{aligned}\frac{\partial D_{\alpha}^{GR}(p||\rho)}{\partial\alpha} &= -\frac{1}{(\alpha-1)^2}\log\left(\int\left[p^{\alpha}\rho^{1-\alpha}-\alpha\cdot p+(\alpha-1)\rho+1\right]d\mathbf{x}\right)\\ &\quad+\frac{1}{\alpha-1}\frac{\int p^{\alpha}\rho^{1-\alpha}(\ln p-\ln\rho)-p+\rho d\mathbf{x}}{\int\left[p^{\alpha}\rho^{1-\alpha}-\alpha\cdot p+(\alpha-1)\rho+1\right]d\mathbf{x}}\end{aligned}$$

We now turn to the usual Rényi-divergence $D_{\alpha}^{R}(p||\rho)$ from (2.23)

$$D_{\alpha}^{GR}(p||\rho)=\frac{1}{\alpha-1}\log\left(\int p^{\alpha}\rho^{1-\alpha}d\mathbf{x}\right)$$

We analogously achieve

$$\begin{aligned}\frac{\partial D_{\alpha}^{GR}(p||\rho)}{\partial\alpha} &= -\frac{1}{(\alpha-1)^2}\log\left(\int p^{\alpha}\rho^{1-\alpha}d\mathbf{x}\right)\\ &\quad+\frac{1}{\alpha-1}\frac{\int p^{\alpha}\rho^{1-\alpha}(\ln p-\ln\rho)d\mathbf{x}}{\int p^{\alpha}\rho^{1-\alpha}d\mathbf{x}}\end{aligned}$$

8.4 γ-divergences

The remaining divergences are the γ-divergences (2.25):

$$\begin{aligned}D_{\gamma}(p||\rho) &= \frac{1}{\gamma+1}\ln\left[\left(\int p^{\gamma+1}d\mathbf{x}\right)^{\frac{1}{\gamma}}\cdot\left(\int\rho^{\gamma+1}d\mathbf{x}\right)\right]-\ln\left[\left(\int p\cdot\rho^{\gamma}d\mathbf{x}\right)^{\frac{1}{\gamma}}\right]\\ &= \frac{1}{\gamma+1}\ln\left[\left(\int p^{\gamma+1}d\mathbf{x}\right)^{\frac{1}{\gamma}}\right]+\frac{1}{\gamma+1}\ln\left[\left(\int\rho^{\gamma+1}d\mathbf{x}\right)\right]-\ln\left[\left(\int p\cdot\rho^{\gamma}d\mathbf{x}\right)^{\frac{1}{\gamma}}\right]\\ &= \frac{1}{(\gamma+1)\gamma}\ln I_1(\gamma)+\frac{1}{\gamma+1}\ln I_2(\gamma)-\frac{1}{\gamma}\ln I_3(\gamma)\end{aligned}$$

The derivative is obtained according to

$$\begin{aligned}\frac{\partial D_{\gamma}(p||\rho)}{\partial\gamma} &= -\frac{(2\gamma+1)}{\gamma^2(\gamma+1)^2}\ln I_1(\gamma)+\frac{1}{(\gamma+1)\gamma I_1(\gamma)}\frac{\partial I_1(\gamma)}{\partial\gamma}\\ &\quad-\frac{1}{(\gamma+1)^2}\ln I_2(\gamma)+\frac{1}{(\gamma+1)I_2(\gamma)}\frac{\partial I_2(\gamma)}{\partial\gamma}\\ &\quad+\frac{1}{\gamma^2}\ln I_3(\gamma)-\frac{1}{\gamma I_3(\gamma)}\frac{\partial I_3(\gamma)}{\partial\gamma}\end{aligned}$$

Next, we calculate the derivatives $\frac{\partial I_1(\gamma)}{\partial\gamma}$, $\frac{\partial I_2(\gamma)}{\partial\gamma}$ and $\frac{\partial I_3(\gamma)}{\partial\gamma}$:

$$\begin{aligned}\frac{\partial I_1(\gamma)}{\partial\gamma} &= \int\frac{\partial\left(p^{\gamma+1}\right)}{\partial\gamma}d\mathbf{x}\\ &= \int p^{\gamma+1}\ln p d\mathbf{x}\end{aligned}$$

$$\begin{aligned}\frac{\partial I_2(\gamma)}{\partial\gamma} &= \int\frac{\partial\left(\rho^{\gamma+1}\right)}{\partial\gamma}d\mathbf{x}\\ &= \int\rho^{\gamma+1}\ln\rho d\mathbf{x}\end{aligned}$$

$$
\begin{aligned}
\frac{\partial I_3(\gamma)}{\partial \gamma} &= \int \frac{\partial \left(p \cdot \rho^{\gamma}\right)}{\partial \gamma} d\mathbf{x} \\
&= \int p \rho^{\gamma} \ln \rho d\mathbf{x}
\end{aligned}
$$

Collecting all intermediate results we finally have

$$
\begin{aligned}
\frac{\partial D_{\gamma}(p||\rho)}{\partial \gamma} = & -\frac{(2\gamma+1)}{\gamma^2 (\gamma+1)^2} \ln\left(\int p^{\gamma+1} d\mathbf{x}\right) + \frac{\int p^{\gamma+1} \ln p d\mathbf{x}}{(\gamma+1)\gamma \int p^{\gamma+1} d\mathbf{x}} \\
& -\frac{1}{(\gamma+1)^2} \ln\left(\int \rho^{\gamma+1} d\mathbf{x}\right) + \frac{\int \rho^{\gamma+1} \ln \rho d\mathbf{x}}{(\gamma+1)\int \rho^{\gamma+1} d\mathbf{x}} \\
& +\frac{1}{\gamma^2} \ln\left(\int p \cdot \rho^{\gamma} d\mathbf{x}\right) - \frac{\int p \rho^{\gamma} \ln \rho d\mathbf{x}}{\gamma \int p \cdot \rho^{\gamma} d\mathbf{x}}.
\end{aligned}
$$

References

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

[2] S.-I. Amari and H. Nagaoka. *Methods of Information Geometry*, volume 191 of *Translations of Mathematical Monographs*. Oxford University Press, Oxford, 2000.

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

[4] M. Basseville. Distance measures for signal processing and pattern recognition. Technical Report 899, Institut National de Recherche en Informatique et en Automatique (INRIA), 1988.

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

[6] C. M. Bishop, M. Svensén, and C. K. I. Williams. GTM: The generative topographic mapping. *Neural Computation*, 10:215–234, 1998.

[7] L. 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(3):200–217, 1967.

[8] A. Cichocki, H. Lee, Y.-D. Kim, and S. Choi. Non-negative matrix factorization with α-divergence. *Pattern Recognition Letters*, 2008.

[9] A. Cichocki, R. Zdunek, A. Phan, and S.-I. Amari. *Nonnegative Matrix and Tensor Factorizations*. Wiley, Chichester, 2009.

[10] K. Crammer, R. Gilad-Bachrach, A.Navot, and A.Tishby. Margin analysis of the LVQ algorithm. In *Proc. NIPS 2002*, http://www-2.cs.cmu.edu/Groups/NIPS/NIPS2002/NIPS2002preproceedings/index.html, 2002.

[11] I. Csiszár. Information-type measures of differences of probability distributions and indirect observations. *Studia Sci. Math. Hungaria*, 2:299–318, 1967.

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

[13] D. Erdogmus. *Information theoretic Learning: Renyi's entropy and its application to adaptive systems training*. PhD thesis, University of Florida, 2002.

[14] G. Fichtenholz. *Differential- und Integralrechnung*, volume II. Deutscher Verlag der Wissenschaften, Berlin, 9th ed. edition, 1964.

[15] B. Frigyik, S. Srivastava, and M. Gupta. Functional Bregman divergence and Bayesian estimation of distributions. *IEEE Transaction on Information Theory*, 54(11):5130–5139, 2008.

[16] B. A. Frigyik, S. Srivastava, and M. Gupta. An introduction to functional derivatives. Technical Report UWEETR-2008-0001, Dept of Electrical Engineering, University of Washington, 2008.

[17] H. Fujisawa and S. Eguchi. Robust parameter estimation with a small bias against heavy contamination. *Journal of Multivariate Analysis*, 99:2053–2081, 2008.

[18] T. Graepel, M. Burger, and K. Obermayer. Self-organizing maps: generalizations and new optimization techniques. *Neurocomputing*, 21(1–3):173–90, 1998.

[19] B. Hammer and T. Villmann. Generalized relevance learning vector quantization. *Neural Networks*, 15(8-9):1059–1068, 2002.

[20] A. Hegde, D. Erdogmus, T. Lehn-Schioler, Y. Rao, and J. Principe. Vector quantization by density matching in the minimum Kullback-Leibler-divergence sense. In *Proc. of the International Joint Conference on Artificial Neural Networks (IJCNN) - Budapest*, pages 105–109, IEEE Press, 2004.

[21] T. Heskes. Energy functions for self-organizing maps. In E. Oja and S. Kaski, editors, *Kohonen Maps*, pages 303–316. Elsevier, Amsterdam, 1999.

[22] M. M. V. Hulle. *Faithful Representations and Topographic Maps*. Wiley Series and Adaptive Learning Systems for Signal Processing, Communications, and Control. Wiley & Sons, New York, 2000.

[23] M. M. V. Hulle. Joint entropy maximization in kernel-based topographic maps. *Neural Computation*, 14(8):1887–1906, 2002.

[24] M. M. V. Hulle. Kernel-based topographic map formation achieved with an information theoretic approach. *Neural Networks*, 15:1029–1039, 2002.

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

[26] E. Jang, C. Fyfe, and H. Ko. Bregman divergences and the self organising map. In C. Fyfe, D. Kim, S.-Y. Lee, and H. Yin, editors, *Intelligent Data Engineering and Automated Learning IDEAL 2008*, LNCS 5326, pages 452–458. Springer, 2008.

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

[28] R. Jenssen, J. 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.

[29] I. Kantorowitsch and G. Akilow. *Funktionalanalysis in normierten Räumen*. Akademie-Verlag, Berlin, 2nd, revised edition, 1978.

[30] J. Kapur. *Measures of Information and their Application*. Wiley, New Delhi, 1994.

[31] T. Kohonen. *Self-Organizing Maps*, volume 30 of *Springer Series in Information Sciences*. Springer, Berlin, Heidelberg, 1995. (Second Extended Edition 1997).

[32] S. Kullback and R. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22:79–86, 1951.

[33] P. Lai and C. Fyfe. Bregman divergences and multi-dimensional scaling. In M. Köppen, N. Kasabov, and G. Coghill, editors, *Proceedings of the International Conference on Information Processing 2008 (ICONIP)*, volume Part II of *LNCS 5507*, pages 935–942. Springer, 2009.

[34] J. Lee and M. Verleysen. Generalization of the l_p norm for time series and its application to self-organizing maps. In M. Cottrell, editor, *Proc. of Workshop on Self-Organizing Maps (WSOM) 2005*, pages 733–740, Paris, Sorbonne, 2005.

[35] J. Lee and M. Verleysen. *Nonlinear Dimensionality Reduction*. Information Sciences and Statistics. Springer Science+Business Media, New York, 2007.

[36] T. Lehn-Schiøler, A. Hegde, D. Erdogmus, and J. Principe. Vector quantization using information theoretic concepts. *Natural Computing*, 4(1):39–51, 2005.

[37] F. Liese and I. Vajda. *Convex Statistical Distances*, volume 95 of *Teubner-Texte zur Matheatik*. Teubner-Verlag, Leipzig, 1987.

[38] F. Liese and I. Vajda. On divergences and informations in statistics and information theory. *IEEE Transaction on Information Theory*, 52(10):4394–4412, 2005.

[39] Y. Linde, A. Buzo, and R. Gray. An algorithm for vector quantizer design. *IEEE Transactions on Communications*, 28:84–95, 1980.

[40] L. Maaten and G. Hinten. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

[41] T. M. Martinetz, S. G. Berkovich, and K. J. Schulten. 'Neural-gas' network for vector quantization and its application to time-series prediction. *IEEE Trans. on Neural Networks*, 4(4):558–569, 1993.

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

[43] E. Mwebaze, P. Schneider, F.-M. Schleif, S. Haase, T. Villmann, and M. Biehl. Divergence based learning vector quantization. In M. Verleysen, editor, *Proc. Of European Symposium on Artificial Neural Networks (ESANN'2010)*, page to appear, Evere, Belgium, 2010. d-side publications.

[44] F. Nielsen and R. Nock. Sided and symmetrized bregman centroids. *IEEE Transaction on Information Theory*, 55(6):2882–2903, 2009.

[45] J. C. Principe, J. F. III, and D. Xu. Information theoretic learning. In S. Haykin, editor, *Unsupervised Adaptive Filtering*. Wiley, New York, NY, 2000.

[46] Y. Qiao and N. Minematsu. f-divergence is a genearlized invariant measure between disributions. In *INTERSPEECH - Proc. of the Annual Conference of the Inetrnational Speech Communication Association*, pages 1349–1352, Brisbane, Australia, 2008.

[47] J. Ramsay and B. Silverman. *Functional Data Analysis*. Springer Science+Media, New York, 2nd edition, 2006.

[48] A. Renyi. On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*. University of California Press, 1961.

[49] A. Renyi. *Probability Theory*. North-Holland Publishing Company, Amsterdam, 1970.

[50] F. Rossi, N. Delannay, B. Conan-Gueza, and M. Verleysen. Representation of functional data in neural networks. *Neurocomputing*, 64:183–210, 2005.

[51] R. Santos-Rodríguez, A. Guerrero-Curieses, R. Alaiz-Rodríguez, and J. Cid-Sueiro. Cost-sensitive learning based on Bregman divergences. *Machine Learning*, 76(2-3):271–285, 2009.

[52] A. Sato and K. Yamada. Generalized learning vector quantization. In D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo, editors, *Advances in Neural Information Processing Systems 8. Proceedings of the 1995 Conference*, pages 423–9. MIT Press, Cambridge, MA, USA, 1996.

[53] P. Schneider, M. Biehl, and B. Hammer. Hyperparameter learning in robust soft LVQ. In M. Verleysen, editor, *Proceedings of the European Symposium on Artificial Neural Networks ESANN*, pages 517–522. d-side publications, 2009.

[54] C. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–432, 1948.

[55] I. Taneja and P. Kumar. Relative information of type s, Csiszár's f -divergence, and information inequalities. *Information Sciences*, 166:105–125, 2004.

[56] K. Torkkola. Feature extraction by non-parametric mutual information maximization. *Journal of Machine Learning Research*, 3:1415–1438, 2003.

[57] K. Torkkola and W. Campbell. Mutual information in learning feature transformations. In *Proc. Of International Conference on Machine Learning ICML'2000*, Stanford, CA, 2000.

[58] T. Villmann. Sobolev metrics for learning of functional data - mathematical and theoretical aspects. *Machine Learning Reports*, 1(MLR-03-2007):1–15, 2007. ISSN:1865-3960, http://www.uni-leipzig.de/˜compint/mlr/mlr_01_2007.pdf.

[59] T. Villmann, S. Haase, F.-M. Schleif, and B. Hammer. Divergence based online learning in vector quantization. In L. Rutkowski, W. Duch, J. Kaprzyk, J. Korbicz, and R. Tadeusiewicz, editors, *Proc. Of International Conference on Artifial Intelligence and Soft Computing (ICAISC'2010)*, LNAI, page submitted, Berlin, 2010. Springer.

[60] T. Villmann, B. Hammer, F.-M. Schleif, T. Geweniger, and W. Herrmann. Fuzzy classification by fuzzy labeled neural gas. *Neural Networks*, 19:772–779, 2006.

[61] T. Villmann, B. Hammer, F.-M. Schleif, W. Hermann, and M. Cottrell. Fuzzy classification using information theoretic learning vector quantization. *Neurocomputing*, 71:3070–3076, 2008.

[62] T. Villmann and F.-M. Schleif. Functional vector quantization by neural maps. In J. Chanussot, editor, *Proceedings of First Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS 2009)*, pages 1–4. IEEE Press, 2009. ISBN 978-1-4244-4948-4.

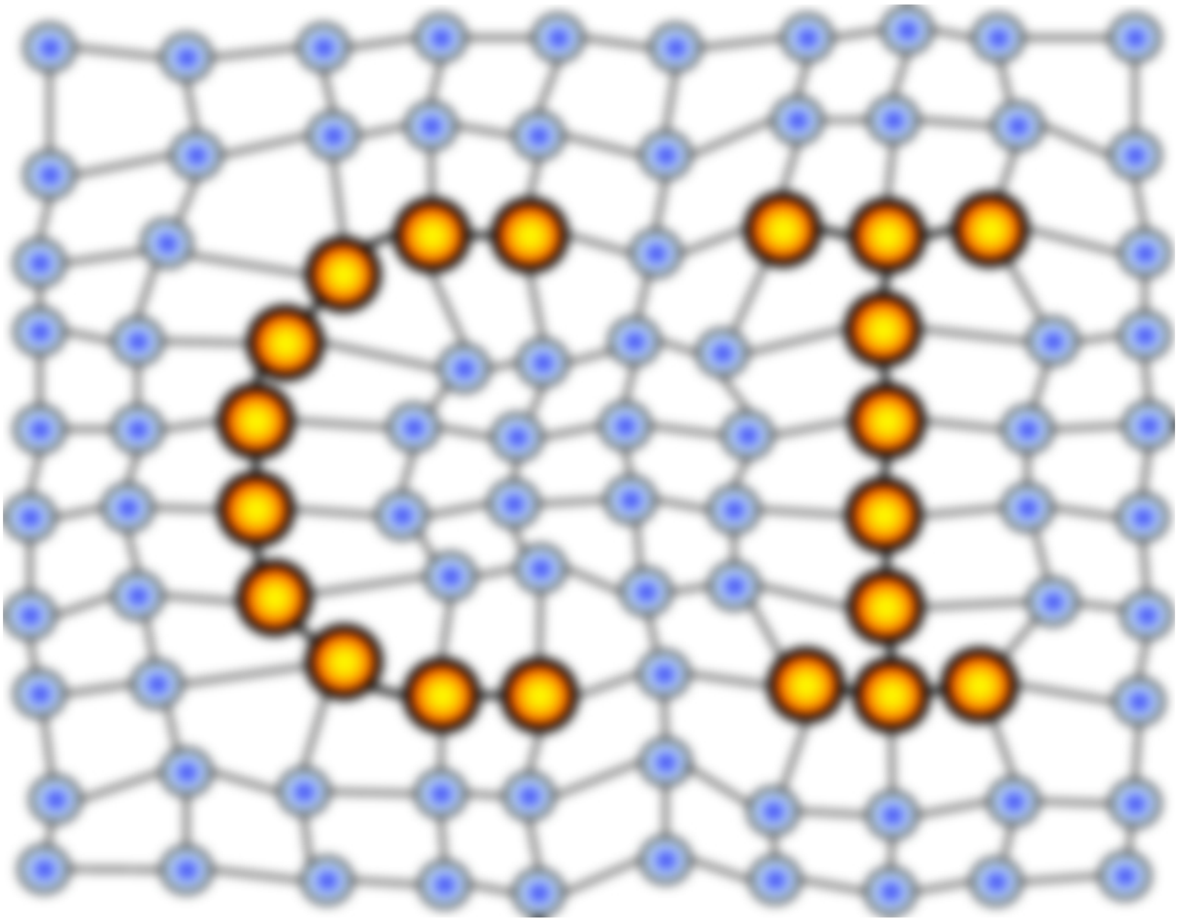
[63] T. Villmann, F.-M. Schleif, M. Kostrzewa, A. Walch, and B. Hammer. Classification of mass-spectrometric data in clinical proteomics using learning vector quantization methods. *Briefings in Bioinformatics*, 9(2):129–143, 2008.

[64] A. Wismüller. The exploration machine – a novel method for data visualization. In J. Principe and R. Miikkulainen, editors, *Advances in Self-Organizing Maps – Proceedings of the 7th InternationalWorkshop WSOM 2009, St. Augustine, FL, USA*, LNCS5629, pages 344–352, Berlin, 2009. Springer.

[65] P. L. Zador. Asymptotic quantization error of continuous signals and the quantization dimension. *IEEE Transaction on Information Theory*, (28):149–159, 1982.

MACHINE LEARNING REPORTS

Report 01/2010



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.