The Mathematics of Divergence Based Online Learning in Vector Quantization

Thomas Villmann[1,⋆], Sven Haase[1], Frank-Michael Schleif[2], Barbara Hammer[2], and Michael Biehl[3]

[1] Department of Mathematics/Natural Sciences/Informatics, University of Applied Sciences Mittweida, 09648 Mittweida, Germany
thomas.villmann@hs-mittweida.de
[2] Clausthal University of Technology, Institute of Computer Science, Clausthal-Zellerfeld, Germany
[3] Rijksuniversity Groningen, Johann Bernoulli Inst. for Mathematics and Computer Science, The Netherlands

Abstract. We propose the utilization of divergences in gradient descent learning of supervised and unsupervised vector quantization as an alternative for the squared Euclidean distance. The approach is based on the determination of the Fréchet-derivatives for the divergences, wich can be immediately plugged into the online-learning rules. We provide the mathematical foundation of the respective framework. This framework includes usual gradient descent learning of prototypes as well as parameter optimization and relevance learning for improvement of the performance.

Keywords: vector quantization, divergence based learning, information theory, clustering, classification.

1 Introduction

The utilization of non-standard metrics in unsupervised and supervised vector quantization is a challenging topic which has an increasing importance for data processing. Prototype based vector quantization for clustering and classification usually is based on the Euclidean distance like the prominent k-means [18], the self-organizing map (SOM,[15]) or the neural gas (NG,[19]) for unsupervised data modeling and learning vector quantization schemes (LVQ,[15]) or support vector machines (SVM,[29]) in case of supervised learning.

However, the standard Euclidean metric may be not appropriate for faithful data processing [25],. Therefore, recent developments extend the standard approaches by incorporating advanced dissimilarity measures for the data modelling. Examples are in the area of functional data processing and visualization [17],[22],[33] or more generally – kernelized metrics [21],[12], bilinear forms for dissimilarities [28] or general dissimilarities [3],[5]. These dissimilarity measures take into account the structure of the data and, therefore, realize a data adequate processing, which may lead to better results.

In this paper we concentrate on a special data type – positive measures $p(\mathbf{x})$. Positive measures are supposed to be positive functions $p(\mathbf{x})$ for the support

⋆ Corresponding author.

F. Schwenker and N. El Gayar (Eds.): ANNPR 2010, LNAI 5998, pp. 108–119, 2010.

$\mathbf{x} \in \Omega$. If further $\int_\Omega p(\mathbf{x})\, d\mathbf{x} = 1$ holds, p is called a density measure, or simply density for short. Density data play an important role in many research areas: For example, spectral data occurring in mass-spectrometry or remote sensing usually are positive measures or densities [34], [32]. The dissimilarity between densities (positive measures) is naturally judged by (generalized) divergences. First vector quantization approaches using divergences apply the batch mode of learning [2],[13] by means of the expectation-maximization methodology. In this scheme, all data have to be available at hand, which is not assumed in the online learning mode of the respective algorithms. Gradient descent leaning usually is realized as stochastic gradient descent optimization. However, this learning mode requires the calculation of the derivatives, which determine the adaptation rule for the prototypes. Thus, we concentrate in this paper, how divergences can be incorporated into gradient based supervised and unsupervised prototype-based learning schemes. For this purpose, we have to investigate the derivatives of divergences, which turn out to be *functional derivatives* mathematically known as *Fréchet-derivatives*.

The paper is organized as follows: We first briefly reconsider SOM/NG and generalized LVQ (GLVQ,[26]) as widely used representatives for the families of gradient based unsupervised and supervised vector quantization algorithms to explain, how the derivatives of the underlying dissimilarity measure come into play. Thereafter we give the *Fréchet-derivatives* for several divergence families, which then can immediately plugged in. Further, we explain some parameter optimization strategies for parametrized divergences, which are related to hyperparameter optimization [27] and relevance learning [10], respectively.

2 Prototype Based Vector Quantization

2.1 Unsupervised Vector Quantization

Prototype based vector quantization (VQ) is a mapping of data $\mathbf{v} \in V \subseteq \mathbb{R}^n$, distributed according to the data density P, onto a set $\mathbf{W} = \{\mathbf{w_r} \in \mathbb{R}^n\}_{\mathbf{r}\in A}$ of prototypes. The set A is an appropriate index set, D is the input dimension and $N = \#A$ the number of prototypes.

The aim of *unsupervised vector quantization* during learning is to distribute the prototypes in the data space such that they represent the data as good as possible. This property is judged by quantization error

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

based on the dissimilarity measure ξ and

$$\mathbf{s}(\mathbf{v}) = \underset{\mathbf{r}\in A}{\operatorname{argmin}} \left[\xi(\mathbf{v}, \mathbf{w_r})\right] \tag{2}$$

being the best matching unit (winner). Hence, the quantization error can be seen as the *expectation value* for the mapping error in the winner determination. Robust approximators for optimum unsupervised vector quantizers are the NG and SOM.

For the NG the above cost function E_{VQ} is modified to

$$E_{\mathrm{NG}} = \frac{1}{2C(\lambda)} \sum_{\mathbf{r}} \int P(\mathbf{v})\, h_\sigma(\mathbf{r})\, \xi(\mathbf{v}, \mathbf{w}_\mathbf{r})\, d\mathbf{v}$$

with the so-called neighborhood function $h_\sigma(\mathbf{r}) = \exp\left(\frac{-rank(\mathbf{r})}{2\sigma^2}\right)$ and is the rank function counting the number of prototypes $\mathbf{r}'$ for which $\xi(\mathbf{v}, \mathbf{w}_{\mathbf{r}'}) \leq \xi(\mathbf{v}, \mathbf{w}_\mathbf{r})$ holds [19]. For SOM a cost function can be defined by

$$E_{\mathrm{SOM}} = \int P(\mathbf{v}) \sum_{\mathbf{r}} \delta_\mathbf{r}^{\mathbf{s}(\mathbf{v})} \cdot le(\mathbf{v}, \mathbf{r})\, d\mathbf{v}$$

with local errors $le(\mathbf{v}, \mathbf{r}) = \sum_{\mathbf{r}'} h_\sigma(\mathbf{r}, \mathbf{r}')\xi(\mathbf{v}, \mathbf{w}_{\mathbf{r}'})$ and $\delta_\mathbf{r}^{\mathbf{s}(\mathbf{v})}$ is the Kronecker-symbol using Heskes' variant [11]. Here, the neighborhood function $h_\sigma(\mathbf{r}, \mathbf{r}') = \exp\left(\frac{-\xi_A(\mathbf{r},\mathbf{r}')}{2\sigma^2}\right)$ is the distance measured in the index set A.

For SOMs, the index set A is equipped with a topological order usually taken as regular low-dimensional grid. However, compared with standard SOM the winning rule in Heskes-SOM is slightly modified:

$$\mathbf{s}(\mathbf{v}) = \underset{\mathbf{r} \in A}{\operatorname{argmin}} \left[le(\mathbf{v}, \mathbf{r})\right]. \tag{3}$$

For both algorithms learning is realized as a stochastic gradient with respect to the prototypes $\mathbf{w}_\mathbf{r}$:

$$\triangle\, \mathbf{w}_\mathbf{r} = -\varepsilon \frac{\partial E_{\mathrm{NG/SOM}}}{\partial \mathbf{w}_\mathbf{r}} \tag{4}$$

which contains as an essential ingredients the derivative $\frac{\partial \xi(\mathbf{v}, \mathbf{w}_{\mathbf{r}'})}{\partial \mathbf{w}_\mathbf{r}}$.

2.2 Supervised Vector Quantization

The goal of *supervised learning vector quantization* (LVQ) is the optimization of the classification accuracy for given data $\mathbf{v} \in V \subseteq \mathbb{R}^n$ equipped with class labels $\mathbf{c}_\mathbf{v}$. Further, a class label $\mathbf{y}_\mathbf{r}$ is attached to each prototype. Again, the data $\mathbf{v}$ are mapped onto the winning prototype according to the mapping rule (2). If $\mathbf{c}_\mathbf{v} \neq \mathbf{y}_{\mathbf{s}(\mathbf{v})}$ a classification error is detected. The overall classification error cannot be optimized directly by gradient descent learning, because it is not differentiable. Therefore, it has to be replaced by an differentiable cost function reflecting essential properties of the classification accuracy. For this purpose the generalized learning vector quantization (GLVQ) scheme was developed [26]. The cost function of GLVQ is given by

$$E_{\mathrm{GLVQ}} = \sum_{\mathbf{v}} \mu(\mathbf{v}) \tag{5}$$

defining the classifier function $\mu(\mathbf{v})$

$$\mu(\mathbf{v}) = \frac{\xi^+ - \xi^-}{\xi^+ + \xi^-}\,. \tag{6}$$

with $\xi^+ = \xi\left(\mathbf{v}, \mathbf{w}_{s^+(\mathbf{v})}\right)$. The value $s^+(\mathbf{v})$ is the winning prototype with the additional constraint that $\mathbf{c_v} = \mathbf{y_{s^+(v)}}$ holds. In analogy, $\mathbf{w}_{s^-(\mathbf{v})}$ has has minimum distance $\xi^- = \xi\left(\mathbf{v}, \mathbf{w}_{s^-(\mathbf{v})}\right)$ for all prototypes $\mathbf{w_r}$ with class labels different to $\mathbf{c_v}$, i.e. $\mathbf{y_r} \neq \mathbf{c_v}$. Then the *generalized* LVQ (GLVQ) is derived as gradient descent on the cost function E_{GLVQ} (5) 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 $\frac{\partial E_{\mathrm{GLVQ}}}{\partial \mathbf{w}_{s^-(\mathbf{v})}}$:

$$\triangle \mathbf{w}_{s^+(\mathbf{v})} = \epsilon^+ \cdot \theta^+ \cdot \frac{\partial \xi\left(\mathbf{v}, \mathbf{w}_{s^+(\mathbf{v})}\right)}{\partial \mathbf{w}_{s^+(\mathbf{v})}} \text{ and } \triangle \mathbf{w}_{s^-(\mathbf{v})} = -\epsilon^- \cdot \theta^- \cdot \frac{\partial \xi\left(\mathbf{v}, \mathbf{w}_{s^-(\mathbf{v})}\right)}{\partial \mathbf{w}_{s^-(\mathbf{v})}} \quad (7)$$

with the scaling factors

$$\theta^+ = \frac{2 \cdot \xi^-}{(\xi^+ + \xi^-)^2} \text{ and } \theta^- = \frac{2 \cdot \xi^+}{(\xi^+ + \xi^-)^2} \,. \quad (8)$$

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

3 Divergences as Dissimilarities and Derivatives Thereof

As mentioned in the introduction, frequently the quadratic Euclidean norm is used for the dissimilarity measure ξ in both supervised and unsupervised vector quantization. In the following we show how it can be replaced by divergence measures. Yet, the strategy is straight forward: If the derivative of a divergence is determined it can be plugged into each gradient based vector quantization scheme including the above examples SOMs, NG or GLVQ.

Divergences estimate the dissimilarity between density functions or positive measures. In information theory they are related mutual information [16]. According to the classification given in Cichocki et al. [4], one can distinguish at least *three* main classes of divergences, the *Bregman*-divergences, the *Csiszár's* f-divergences and the γ-divergences [4]. If a divergence $D(p||\rho)$ is given, the mathematical framework for the functional derivative with respect to ρ is the concept of *Fréchet-derivatives* or *functional derivatives* $\frac{\delta D(p||\rho)}{\delta \rho}$ [8],[14]. In the following we will explain the functional derivatives for these divergence classes. Thereby we assume that p and ρ are positive measures in $\mathbf{x} \in \Omega$ and integrals are taken according to support Ω.

3.1 Basic Divergences

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 [14]. Bregman divergences are defined as $D_\Phi^B : \mathcal{L} \times \mathcal{L} \longrightarrow \mathbb{R}^+$ with

$$D_\Phi^B(p||\rho) = \Phi(p) - \Phi(\rho) - \frac{\delta \Phi(\rho)}{\delta \rho}(p - \rho) \quad (9)$$

whereby $\frac{\delta\Phi(\rho)}{\delta\rho}$ is the Fréchet-derivative of Φ with respect to ρ. For the choice $\Phi(f) = f^2$, the Euclidean distance is obtained. The Fréchet-derivative is

$$\frac{\delta D_\Phi^B(p||\rho)}{\delta\rho} = \frac{\Phi(p)}{\delta\rho} - \frac{\Phi(\rho)}{\delta\rho} - \frac{\delta\left[\frac{\delta\Phi(\rho)}{\delta\rho}(p-\rho)\right]}{\delta\rho} \tag{10}$$

An important subset of Bregman divergences are the β-divergences

$$D_\beta(p||\rho) = \int p \cdot \frac{p^{\beta-1} - \rho^{\beta-1}}{\beta-1} d\mathbf{x} - \int \frac{p^\beta - \rho^\beta}{\beta} d\mathbf{x} \tag{11}$$

with $\beta \neq 1$ and $\beta \neq 0$. The Fréchet-derivative is

$$\frac{\delta D_\beta(p||\rho)}{\delta\rho} = -p \cdot \rho^{\beta-2} + \rho^{\beta-1} . \tag{12}$$

In the limit $\beta \to 1$ the divergence $D_\beta(p,\rho)$ becomes the generalized Kullback-Leibler-divergence

$$D_{GKL}(p||\rho) = \int p \log\left(\frac{p}{\rho}\right) d\mathbf{x} - \int p - \rho d\mathbf{x}. \tag{13}$$

with the Fréchet-derivative

$$\frac{\delta D_{GKL}(p||\rho)}{\delta\rho} = -\frac{p}{\rho} + 1 \tag{14}$$

Csiszár's f-divergences are generated by a *convex* function $f : [0,\infty) \to \mathbb{R}$ with $f(1) = 0$ (without loss of generality) as

$$D_f(p||\rho) = \int \rho \cdot f\left(\frac{p}{\rho}\right) d\mathbf{x} \tag{15}$$

with the definitions $0 \cdot f\left(\frac{0}{0}\right) = 0$, $0 \cdot f\left(\frac{a}{0}\right) = \lim_{x\to 0} x \cdot f\left(\frac{a}{x}\right) = \lim_{u\to\infty} a \cdot \frac{f(u)}{u}$ [6] with the famous Hellinger divergence in case of densities p and ρ [30]:

$$D_H(p||\rho) = \int \left(\sqrt{p} - \sqrt{\rho}\right)^2 d\mathbf{x} \tag{16}$$

with the generating function $f(u) = \left(\sqrt{u} - 1\right)^2$ with $u = \frac{p}{\rho}$. The Fréchet-derivative of $D_f(p||\rho)$ writes as

$$\frac{\delta D_f(p||\rho)}{\delta\rho} = f\left(\frac{p}{\rho}\right) + \rho\frac{\partial f(u)}{\partial u} \cdot \frac{-p}{\rho^2} \tag{17}$$

with $u = \frac{p}{\rho}$ which yields $\frac{\delta D_H(p||\rho)}{\delta\rho} = 1 - \sqrt{\frac{p}{\rho}}$. We can identify also an important subset of f-divergences – the so-called $\alpha-$divergences [4]:

$$D_\alpha(p||\rho) = \frac{1}{\alpha(\alpha-1)} \int \left[p^\alpha \rho^{1-\alpha} - \alpha \cdot p + (\alpha-1)\rho\right] d\mathbf{x} \tag{18}$$

with the generating $f-$function

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

and $u = \frac{\rho}{p}$. In the limit $\alpha \to 1$ the generalized Kullback-Leibler-divergence D_{GKL} (13) is obtained. The Fréchet-derivative is calculated as

$$\frac{\delta D_\alpha(p||\rho)}{\delta\rho} = -\frac{1}{\alpha}\left(p^\alpha\rho^{-\alpha}-1\right) . \tag{19}$$

The $\alpha-$divergences are closely related to the generalized *Rényi-divergences* [1],[23],[24]:

$$D_\alpha^{GR}(p||\rho) = \frac{1}{\alpha-1}\log\left(\int\left[p^\alpha\rho^{1-\alpha} - \alpha\cdot p + (\alpha-1)\rho + 1\right]d\mathbf{x}\right) \tag{20}$$

with the Fréchet-derivative

$$\frac{\delta D_\alpha^{GR}(p||\rho)}{\delta\rho} = -\frac{\alpha}{\int\left[p^\alpha\rho^{1-\alpha} - \alpha\cdot p + (\alpha-1)\rho + 1\right]d\mathbf{x}}\frac{\delta D_\alpha(p||\rho)}{\delta\rho} . \tag{21}$$

The very outlier-robust γ-divergence class is defined according to

$$D_\gamma(p||\rho) = \frac{1}{\gamma+1}\log\left[\left(\int p^{\gamma+1}d\mathbf{x}\right)^{\frac{1}{\gamma}}\cdot\left(\int\rho^{\gamma+1}d\mathbf{x}\right)\right] - \log\left[\left(\int p\cdot\rho^\gamma d\mathbf{x}\right)^{\frac{1}{\gamma}}\right] \tag{22}$$

proposed by Fujisawa&Eguchi [9]. In the limit $\gamma \to 0$ $D_\gamma(\rho||p)$ becomes the usual Kullback-Leibler-divergence for normalized densities. For $\gamma = 1$ the *Cauchy-Schwarz*-divergence

$$D_{CS}(p||\rho) = \frac{1}{2}\log\left(\int\rho^2(\mathbf{x})\,d\mathbf{x}\cdot\int p^2(\mathbf{x})\,d\mathbf{x}\right) - \log\left(\int p(\mathbf{x})\cdot\rho(\mathbf{x})\,d\mathbf{x}\right) \tag{23}$$

is obtained, which was suggested for information theoretic learning by J. Principe investigating the Cauchy-Schwarz-inequality for norms [20]. The Fréchet-derivative of $D_\gamma(p||\rho)$ becomes

$$\frac{\delta D_\gamma(p||\rho)}{\delta\rho} = \frac{\rho^\gamma}{\left(\int\rho^{\gamma+1}d\mathbf{x}\right)} - \frac{p\rho^{\gamma-1}}{\left(\int p\cdot\rho^\gamma d\mathbf{x}\right)} \tag{24}$$

Due to the lack of space, the derivation of these results can be found in [31].

If we now identify $\mathbf{v}$ with a vectorial representation of p and the prototypes $\mathbf{w}$ as the respective ρ representation, the obtained derivative can be immediately plugged into gradient learning schemes as above outlined.

In an example application we consider the data vectors $\mathbf{v} \in \mathbb{R}^2$ with $\|\mathbf{v}\| = 1$ and v_1 distributed in $[0,1]$ according to the density $q(v_1) = 2v_1$. We learned a one-dimensional SOM for α-, β- and γ-divergences with different parameter setting. The resulted prototype distributions are depicted in Fig. 1. Obviously, the influence of the parameter variations is detectable. In particular, the limits to the Kullback-Leibler-divergence setting are clearly observable.

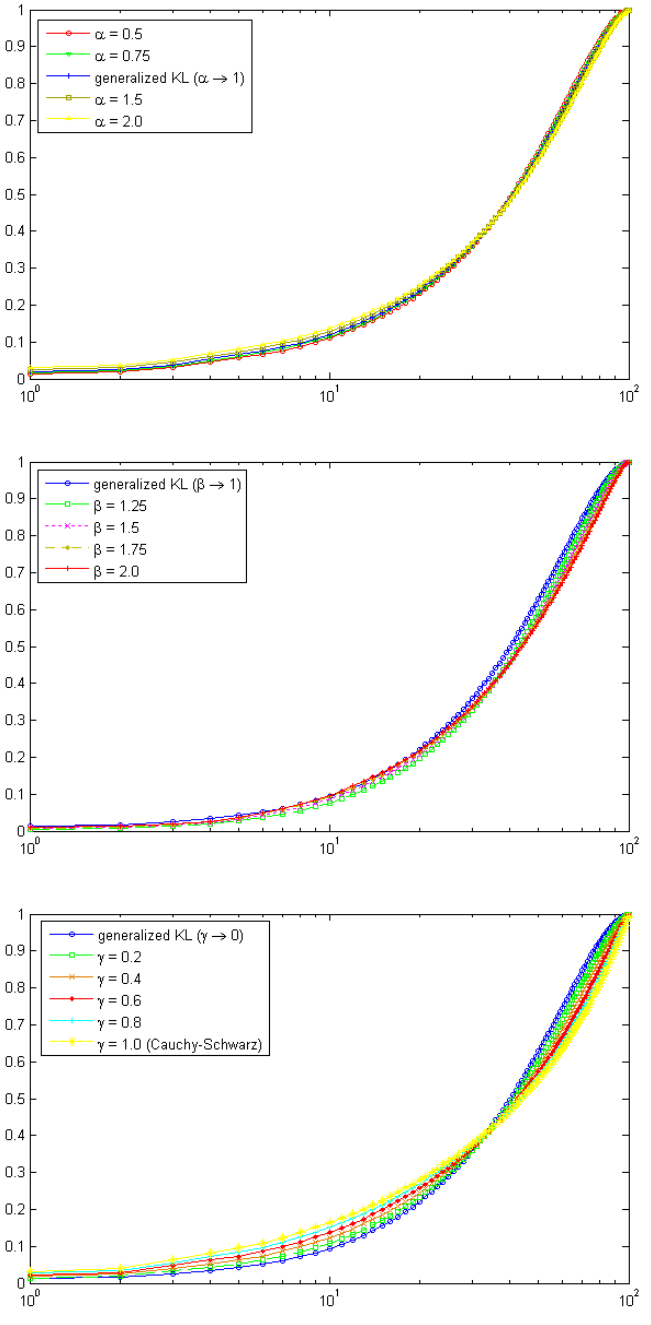

Fig. 1. Illustration of divergence based learning. The w_1-components of the prototypes are depicted for learning α-, β-, γ-divergences (from top to bottom). The horizontal axis is the prottype number. The data distribution was according to $q(v_1) = 2v_1$ with $v_1 \in [0, 1]$, randomly, and $v_2 = 1 - v_1$.

3.2 Parameter Adaptation in Divergence Based Learning

Considering the parametrized divergence families of γ-, α-, and β-divergences, one could further think about the optimal choice of the so-called hyperparameters γ, α, and β as suggested in a similar manner for other parametrized LVQ-algorithms [27]. 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

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

depending on a *parametrized divergence* ξ_η with parameter η and the set $W = \{\mathbf{w_r}\}$ of prototypes. If E and ξ_η are both differentiable with respect to η according to

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

a gradient based optimization is derived by

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

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

We assume in the following that the (positive) measures p and ρ represent the data $\mathbf{v}$ and prototypes $\mathbf{w}$, respectively. If the measures p and ρ are continuously differentiable, then , considering derivatives of parametrized divergences $\frac{\partial \xi_\eta}{\partial \eta}$ with respect to the parameter η, it is allowed to interchange integration and differentiation, under the assumption that the resulting integral exists [7]. Hence, we can differentiate parametrized divergences with respect to their hyperparameter in that case. For the several α-, β-, and γ-divergences characterized in sec. 3.1 we obtain after some elementary calculations [31]:

- β-divergence $D_\beta\left(p||\rho\right)$ from (11)

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

- α-divergence $D_\alpha\left(p||\rho\right)$ from (18)

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

– generalized Rényi-divergence $D_\alpha^{GR}(p||\rho)$ from (20)

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

– Rényi-divergence $D_\alpha^{R}(p||\rho)$ from (20) for normalized densities

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

– γ-divergence $D_\gamma(p||\rho)$ from (22)

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

3.3 Relevance Learning for Positive Measures

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 [10].

In the framework of divergence based gradient descent learning we multiplicatively weight a positive measure $q(x)$ by $\lambda(x)$ with $0 \le \lambda(\mathbf{x}) < \infty$ 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 of the GLVQ cost function (5) as it is known from vectorial relevance learning but paying now attention to the utilization of divergences. This leads here, again, to Fréchet-derivatives of the incorporated divergence D but now with respect to the weighting function $\lambda(\mathbf{x})$ – $\frac{\delta D(\lambda\cdot p||\lambda\cdot\rho)}{\delta\lambda}$.

In particular we obtain for the Bregman divergence

$$\frac{\delta D_\Phi^B(\lambda\cdot p||\lambda\cdot\rho)}{\delta\lambda} = \frac{\Phi(\lambda\cdot p)}{\delta\lambda} - \frac{\Phi(\lambda\cdot\rho)}{\delta\lambda} - \frac{\delta\left[\frac{\delta\Phi(\lambda\cdot\rho)}{\delta\rho}\lambda(p-\rho)\right]}{\delta\lambda} \tag{25}$$

with

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

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

$$\frac{\delta D_{GKL}\left(\lambda\cdot p||\lambda\cdot\rho\right)}{\delta\lambda}=p\cdot\log\left(\frac{p}{\rho}\right)-p+\rho\,. \tag{26}$$

Further, for the β-divergences (11) we have

$$\frac{\delta D_{\beta}\left(\lambda\cdot p||\lambda\cdot\rho\right)}{\delta\lambda}=\frac{\rho\cdot\left(\lambda\cdot p\right)^{\beta}+\left(\rho\cdot\left(\beta-1\right)-p\cdot\beta\right)\cdot\left(\lambda\cdot\rho\right)^{\beta}}{\lambda\rho\left(\beta-1\right)}\,. \tag{27}$$

For f-divergences (15) we consider with $u=\frac{p}{\rho}$

$$\begin{aligned}\frac{\delta D_{f}\left(\lambda\cdot p||\lambda\cdot\rho\right)}{\delta\lambda}&=\rho\cdot f\left(\frac{p}{\rho}\right)+\lambda\cdot\rho\frac{\partial f\left(u\right)}{\partial u}\frac{\delta u}{\delta\lambda}\\&=\rho\cdot f\left(\frac{p}{\rho}\right)\end{aligned} \tag{28}$$

because of $\frac{\delta u}{\delta\lambda}=0$. The relevance learning of α-divergences (18) follows

$$\frac{\delta D_{\alpha}\left(\lambda\cdot p||\lambda\cdot\rho\right)}{\delta\lambda}=\frac{1}{\alpha\left(\alpha-1\right)}\left[\rho\cdot\left(\left(\frac{p}{\rho}\right)^{\alpha}+\alpha-1\right)-p\cdot\alpha\right]\,, \tag{29}$$

whereas the respective gradient of generalized Rényi-divergences (20) can be derived from this as

$$\frac{\delta D_{\alpha}^{GR}\left(\lambda\cdot p||\lambda\cdot\rho\right)}{\delta\lambda}=\frac{\alpha}{\int\left[\lambda\cdot\left(\rho\cdot\left(\frac{p}{\rho}\right)^{\alpha}-\alpha\cdot p+\left(\alpha-1\right)\cdot\rho\right)+1\right]d\mathbf{x}}\frac{\delta D_{\alpha}\left(\lambda\cdot p||\lambda\cdot\rho\right)}{\delta\lambda}\,. \tag{30}$$

The γ-divergences finally yields

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

Again the important special case $\gamma=1$ is considered: the relevance learning scheme for the Cauchy-Schwarz divergence (23) is derived as

$$\frac{\delta D_{CS}\left(\lambda\cdot p||\lambda\cdot\rho\right)}{\delta\lambda}=\frac{p\cdot\lambda\cdot p}{\int\left(\lambda\cdot p\right)^{2}d\mathbf{x}}+\frac{\rho\cdot\lambda\cdot\rho}{\int\left(\lambda\cdot\rho\right)^{2}d\mathbf{x}}-\frac{2\cdot p\cdot\lambda\cdot\rho}{\int\lambda^{2}\cdot p\cdot\rho d\mathbf{x}}\,. \tag{31}$$

As before, if we identify p and ρ with the data $\mathbf{v}$ and the prototypes $\mathbf{w}$, the derivatives can be immediately put into a gradiend descent learning scheme.

4 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 subclasses of divergences, Bregman-, α-, β-, γ-, and f-divergences following [4]. The mathematical framework of Fréchet-derivatives is then used to calculate the functional divergence derivatives.

We exemplary explain the utilization of this methodology for famous examples of supervised and unsupervised vector quantization including SOM, NG, and GLVQ. Further, we discuss, how a parameter adaptation could be integrated in supervised learning to achieve improved classification results in case of the parametrized α-, β-, and γ-divergences. In the last step we considered a weighting function for generalized divergences based on positive measures. The optimization scheme for this weight function for a given classification task is again obtained by Fréchet derivatives, and one ends up with a relevance learning scheme analogously to relevance learning for usual (Eulidean) supervised learning vector quantization [10].

References

1. Amari, S.-I.: Differential-Geometrical Methods in Statistics. Springer, Heidelberg (1985)
2. Banerjee, A., Merugu, S., Dhillon, I., Ghosh, J.: Clustering with bregman divergences. Journal of Machine Learning Research 6, 1705–1749 (2005)
3. Bezdek, J., Hathaway, R., Windham, M.: Numerical comparison of RFCM and AP algorithms for clustering relational data. Pattern recognition 24, 783–791 (1991)
4. Cichocki, A., Zdunek, R., Phan, A., Amari, S.-I.: Nonnegative Matrix and Tensor Factorizations. Wiley, Chichester (2009)
5. Cottrell, M., Hammer, B., Hasenfuß, A., Villmann, T.: Batch and median neural gas. Neural Networks 19, 762–771 (2006)
6. Csiszár, I.: Information-type measures of differences of probability distributions and indirect observations. Studia Sci. Math. Hungaria 2, 299–318 (1967)
7. Fichtenholz, G.: Differential- und Integralrechnung, 9th edn., vol. II. Deutscher Verlag der Wissenschaften, Berlin (1964)
8. Frigyik, B.A., Srivastava, S., Gupta, M.: An introduction to functional derivatives. Technical Report UWEETR-2008-0001, Dept. of Electrical Engineering, University of Washington (2008)
9. Fujisawa, H., Eguchi, S.: Robust parameter estimation with a small bias against heavy contamination. Journal of Multivariate Analysis 99, 2053–2081 (2008)
10. Hammer, B., Villmann, T.: Generalized relevance learning vector quantization. Neural Networks 15(8-9), 1059–1068 (2002)
11. Heskes, T.: Energy functions for self-organizing maps. In: Oja, E., Kaski, S. (eds.) Kohonen Maps, pp. 303–316. Elsevier, Amsterdam (1999)
12. Hulle, M.M.V.: Kernel-based topographic map formation achieved with an information theoretic approach. Neural Networks 15, 1029–1039 (2002)
13. Jang, E., Fyfe, C., Ko, H.: Bregman divergences and the self organising map. In: Fyfe, C., Kim, D., Lee, S.-Y., Yin, H. (eds.) IDEAL 2008. LNCS, vol. 5326, pp. 452–458. Springer, Heidelberg (2008)
14. Kantorowitsch, I., Akilow, G.: Funktionalanalysis in normierten Räumen, 2nd revised edn. Akademie-Verlag, Berlin (1978)
15. Kohonen, T.: Self-Organizing Maps. Springer Series in Information Sciences, vol. 30. Springer, Heidelberg (1995) (2nd Extended edn. 1997)

16. Kullback, S., Leibler, R.: On information and sufficiency. Annals of Mathematical Statistics 22, 79–86 (1951)
17. Lee, J., Verleysen, M.: Generalization of the l_p norm for time series and its application to self-organizing maps. In: Cottrell, M. (ed.) Proc. of Workshop on Self-Organizing Maps (WSOM) 2005, Paris, Sorbonne, pp. 733–740 (2005)
18. Linde, Y., Buzo, A., Gray, R.: An algorithm for vector quantizer design. IEEE Transactions on Communications 28, 84–95 (1980)
19. Martinetz, T.M., Berkovich, S.G., Schulten, K.J.: Neural-gas network for vector quantization and its application to time-series prediction. IEEE Trans. on Neural Networks 4(4), 558–569 (1993)
20. Principe, J.C., Fisher III, J., Xu, D.: Information theoretic learning. In: Haykin, S. (ed.) Unsupervised Adaptive Filtering. Wiley, New York (2000)
21. Qin, A., Suganthan, P.: A novel kernel prototype-based learning algorithm. In: Proceedings of the 17th International Conference on Pattern Recognition (ICPR 2004), vol. 4, pp. 621–624 (2004)
22. Ramsay, J., Silverman, B.: Functional Data Analysis, 2nd edn. Springer Science+Media, New York (2006)
23. Renyi, A.: On measures of entropy and information. In: Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability. University of California Press (1961)
24. Renyi, A.: Probability Theory. North-Holland Publishing Company, Amsterdam (1970)
25. Rossi, F., Delannay, N., Conan-Gueza, B., Verleysen, M.: Representation of functional data in neural networks. Neurocomputing 64, 183–210 (2005)
26. Sato, A., Yamada, K.: Generalized learning vector quantization. In: Touretzky, D.S., Mozer, M.C., Hasselmo, M.E. (eds.) Proceedings of the 1995 Conference on Advances in Neural Information Processing Systems, vol. 8, pp. 423–429. MIT Press, Cambridge (1996)
27. Schneider, P., Biehl, M., Hammer, B.: Hyperparameter learning in robust soft LVQ. In: Verleysen, M. (ed.) Proceedings of the European Symposium on Artificial Neural Networks ESANN, pp. 517–522. d-side publications (2009)
28. Schneider, P., Hammer, B., Biehl, M.: Adaptive relevance matrices in learning vector quantization. Neural Computation 21, 3532–3561 (2009)
29. Shawe-Taylor, J., Cristianini, N.: Kernel Methods for Pattern Analysis and Discovery. Cambridge University Press, Cambridge (2004)
30. Taneja, I., Kumar, P.: Relative information of type s, Csiszár's f -divergence, and information inequalities. Information Sciences 166, 105–125 (2004)
31. Villmann, T., Haase, S.: Mathematical aspects of divergence based vector quantization using fréchet-derivatives - extended and revised version. Machine Learning Reports 4(MLR-01-2010), 1–35 (2010), `http://www.uni-leipzig.de/~compint/mlr/mlr_01_2010.pdf`
32. Villmann, T., Merényi, E., Hammer, B.: Neural maps in remote sensing image analysis. Neural Networks 16(3-4), 389–403 (2003)
33. Villmann, T., Schleif, F.-M.: Functional vector quantization by neural maps. In: Chanussot, J. (ed.) Proceedings of First Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS 2009), pp. 1–4. IEEE Press, Los Alamitos (2009)
34. Villmann, T., Schleif, F.-M., Kostrzewa, M., Walch, A., Hammer, B.: Classification of mass-spectrometric data in clinical proteomics using learning vector quantization methods. Briefings in Bioinformatics 9(2), 129–143 (2008)