Gradient Based Learning in Vector Quantization Using Differentiable Kernels

Thomas Villmann*, Sven Haase, and Marika Kästner

Computational Intelligence Group,
University of Applied Sciences Mittweida, 09648 Mittweida, Germany
thomas.villmann@hs-mittweida.de, {villmann,haase,kaestner}@hs-mittweida.de

Abstract. Supervised and unsupervised prototype based vector quantization frequently are proceeded in the Euclidean space. In the last years, also non-standard metrics became popular. For classification by support vector machines, Hilbert space representations are very successful based on so-called kernel metrics. In this paper we give the mathematical justification that gradient based learning in prototype-based vector quantization is possible by means of kernel metrics instead of the standard Euclidean distance. We will show that an appropriate handling requires *differentiable universal kernels* defining the kernel metric. This allows a prototype adaptation in the original data space but equipped with a metric determined by the kernel. This approach avoids the Hilbert space representation as known for support vector machines. Moreover, we give prominent examples for differentiable universal kernels based on information theoretic concepts and show exemplary applications.

1 Introduction

Prototype based vector quantization is an ongoing topic of research with applications in unsupervised and supervised data modeling. Famous unsupervised models applied in data clustering or visualization are the self-organizing map (SOM,[21]), neural gas (NG, [26]) or respective fuzzy variants like fuzzy-c-means (FCM, [3,4]). Supervised approaches comprise the family of learning vector quantizers (LVQ, [21]) as well as support vector machines (SVM,[41]). LVQ models generate class typical prototypes whereas SVMs determine prototypes (support vectors) defining the class borders. Both paradigms are margin classifiers [11]. Recent developments in the field address the utilization of non-standard metrics to improve the model performance for domain specific problems like processing of functional data, e.g. spectra, time series, etc. [20,29,47], or better interpretability of the adapted models (relevance/matrix learning, [16,42]).

One of the most challenging ideas in classification learning is the kernel trick realized in SVMs. According to this idea, the data as well as the prototypes are implicitly mapped into a high-dimensional (infinite) feature mapping Hilbert space (FMHS) uniquely determined by the kernel, but the dissimilarities still

* Corresponding author.

P.A. Estévez et al. (Eds.): *Advances in Self-Organizing Maps*, AISC 198, pp. 193–204.
DOI: 10.1007/978-3-642-35230-0_20

are calculated using the original data whereas model adaptation is processed in the dual space of the FMHS. This implicit mapping frequently offers a great flexibility and good separation possibility. This advantage, however, makes it more difficult to interpret the model because the prototypes in these models are given as infinite-dimensional representations in the FMHS. Moreover, the support vectors are not typical representatives of the classes, as mentioned before. Several variants of LVQ were established also integrating the kernel mapping concept in those models to keep the idea of class-typical prototypes (Kernel GLVQ, KGLVQ) [35,34]. Yet, these models also have to the infinity of the mapping space. Usually, it is approximated by a finite one using the Nystrøm-approximation technique [40], which obviously leads to an information loss in general.

In this paper we offer an alternative for the integration of kernels in prototype based vector quantization. For this purpose, we consider *differentiable* universal kernels determining a new differentiable metric in the data space to be used in the vector quantization model. Thus gradient based learning becomes available whereby the topological structure of the new metric space is isomorphic to the FMHS.

The paper is structured as follows: First we briefly review the idea and justification of kernel mapping into FHMS. Thereafter, we present the theoretical justification of the differentiable kernel online vector quantization approach. Subsequently, we present information theoretic kernels. Sample applications and concluding remarks complete the contribution.

2 Reproducing Kernels for Hilbert Spaces and Kernel Mapping

We start with a brief review of the kernel theory. For that we assume the data space as a compact metric space (V, d_V), i.e. a vector space V equipped with a metric d_V. A function

$$\kappa_\Phi : V \times V \to \mathbb{C} \tag{1}$$

is a kernel, if there exists a *Hilbert space* $\mathcal{H}$ and a map

$$\Phi : V \ni \mathbf{v} \longmapsto \Phi(\mathbf{v}) \in \mathcal{H} \tag{2}$$

with

$$\kappa_\Phi(\mathbf{v}, \mathbf{w}) = \langle \Phi(\mathbf{v}), \Phi(\mathbf{w}) \rangle_\mathcal{H} \tag{3}$$

for all $\mathbf{v}, \mathbf{w} \in V$ and $\langle \cdot, \cdot \rangle_\mathcal{H}$ is the inner product of this Hilbert space. As a consequence the kernel is Hermitian, i.e. $\kappa_\Phi(\mathbf{v}, \mathbf{w}) = \overline{\kappa_\Phi(\mathbf{w}, \mathbf{v})}$ and, therefore, sesquilinear. The mapping Φ is called feature map and $\mathcal{H}$ the feature space of V. Without further restrictions on the kernel κ_Φ both $\mathcal{H}$ and Φ are not unique. A function $f : V \longrightarrow \mathbb{C}$ is *induced by* κ_Φ if there exists an element $g \in \mathcal{H}$ with $f(\mathbf{w}) = \langle g, \Phi(\mathbf{w}) \rangle_\mathcal{H}$. The following important Lemma is shown in [46]:

Lemma 1. *Let κ_Φ be a kernel of a metric space (V, d_V) and Φ a corresponding feature map into a Hilbert space $\mathcal{H}$. Then κ_Φ is continuous iff Φ does. In this case*

$$d_{\kappa_\Phi}(\mathbf{v}, \mathbf{w}) = \|\Phi(\mathbf{v}) - \Phi(\mathbf{w})\|_{\mathcal{H}} \tag{4}$$

defines a semi-metric[1] *on V and the identity map Ψ between the different metric spaces over the vector space V*

$$\Psi : (V, d_V) \longrightarrow (V, d_{\kappa_\Phi}) \tag{5}$$

is continuous. If the feature map Φ is injective d_{κ_Φ} is even a metric.

We have to to state the following important remark:

Remark 1. In the proof of this lemma the inner product property (3) of the kernel is never used. Only the norm properties of Hilbert spaces and their completeness are required. Hence, the lemma is also valid if Φ would map into a Banach space $\mathcal{B}$ with metric d_{κ_Φ}.

To ensure the separability of the feature map Φ the kernel has to be *universal* [46]. Further, STEINWART has also proofed that continuous universal kernel imply the injectivity of the corresponding feature map Φ. Again, we have to emphasize that the proof of this theorem does not utilize the inner product property (3) of the kernel. Only, the semi-metric properties of the corresponding metric are needed, which would remain valid also regarding Banach spaces instead of Hilbert spaces.

An important role in feature mapping play positive definite kernels, which *uniquely* correspond to Hilbert spaces $\mathcal{H}$ in a canonical manner according to the Mercer-theorem [1,27]. The kernel κ_Φ is said to be (strictly) positive definite if for all finite subsets $V_n \subseteq V$ with cardinality $\#V_n = n$, the Gram-Matrix

$$\mathbf{G}_n = [\kappa(\mathbf{v}_i, \mathbf{v}_j) : i, j = 1 \ldots n] \tag{6}$$

is (strictly) positive semi-definite [1]. In that case, the Hilbert space $\mathcal{H}$ is a so-called *reproducing kernel Hilbert space* (RKHS), i.e. the kernel function $\kappa_\Phi(\mathbf{v}, \cdot) \in \mathcal{H}$ and for each $\mathbf{v} \in V$ and all $f \in \mathcal{H}$ and $\mathbf{w} \in V$ the relation $f(\mathbf{w}) = \langle f, \kappa_\Phi(\mathbf{w}, \cdot)\rangle_{\mathcal{H}}$ is valid according to the Riesz representation theorem [1,22]. Here, κ_Φ is denoted as a *reproducing kernel* obviously being symmetric, real and, hence, bi-linear. The space $\mathcal{I}_{\kappa_\Phi}$ of kernel induced functions is given as the set

$$\mathcal{I}_{\kappa_\Phi} = \{\kappa_\Phi(\mathbf{w}, \cdot) | \mathbf{w} \in V\} \tag{7}$$

with $\mathcal{I}_{\kappa_\Phi} \subseteq \mathcal{H}$. For positive kernels the associated inner product implies a norm $\|\Phi(\mathbf{v})\|_{\mathcal{H}} = \sqrt{\langle \Phi(\mathbf{v}), \Phi(\mathbf{v})\rangle_{\mathcal{H}}}$ and, hence, also a metric $d_{\mathcal{H}}(\Phi(\mathbf{v}), \Phi(\mathbf{w})) = \|\Phi(\mathbf{v}) - \Phi(\mathbf{w})\|_{\mathcal{H}}$. Hence, the positive semi-definiteness of the kernel ensures the metric properties in comparison to the semi-metric (4) obtained for general kernels. Because κ_Φ is a kernel, the metric $d_{\mathcal{H}}(\Phi(\mathbf{v}), \Phi(\mathbf{w}))$ can be rewritten as

$$d_{\mathcal{H}}(\Phi(\mathbf{v}), \Phi(\mathbf{w})) = \sqrt{\kappa_\Phi(\mathbf{v}, \mathbf{v}) - 2\kappa_\Phi(\mathbf{v}, \mathbf{w}) + \kappa_\Phi(\mathbf{w}, \mathbf{w})} \tag{8}$$

using the bi-linearity and the symmetry of the positive kernel.

[1] Note, for a semi-metric the triangle inequality does not hold [32].

Remark 2. Obviously, the semi-metric d_{κ_Φ} from (4) coincides with $d_{\mathcal{H}}$ on $\mathcal{I}_{\kappa_\Phi}$ for positive kernels.

This last remark allows an important conclusion regarding the mapping Ψ from (5) in relation to a given positive continuous kernel κ_Φ:

Lemma 2. *Let (V, d_V) be a compact metric space, $\kappa_\Phi : V \times V \to \mathbb{R}$ a continuous positive kernel with the feature map $\Phi : V \longrightarrow \mathcal{H}$, and the kernel determining a metric $d_{\mathcal{H}}$ in $\mathcal{H}$ by (8). If the space of the induced functions $\mathcal{I}_{\kappa_\Phi}$ is dense in the space of continuous functions $\mathcal{C}(V)$, then the metric space $(V, d_{\mathcal{H}})$ is topologically equivalent to induced space $\mathcal{I}_{\kappa_\Phi} \subseteq \mathcal{H}$ with the metric $d_{\mathcal{H}}$. Moreover, both spaces are isometric, and, hence, $(V, d_{\mathcal{H}})$ is a Hilbert space, too. In consequence, the generally non-linear mapping Ψ from (5) is an bijective, separable and continuous mapping. The result of the Lemma 2 is visualized in Fig.1.*

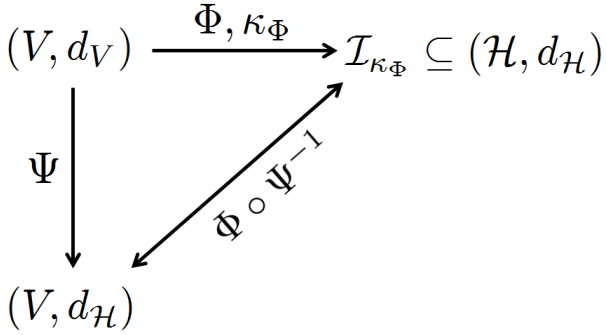

Fig. 1. Visualization of Lemma 2: For universal kernels κ_Φ the metric spaces $(V, d_{\mathcal{H}})$ and $\left(\mathcal{I}_{\kappa_\oplus}, d_{\mathcal{H}}\right)$ are topologically equivalent and isometric by means of the continuous bijective mapping $\Phi \circ \Psi^{-1}$

Proof. The kernel κ_Φ is assumed to be positive, continuous and generating a space of induced functions $\mathcal{I}_{\kappa_\Phi}$, which is dense in the space of continuous functions $\mathcal{C}(V)$. Hence, κ_Φ is universal and, therefore, the uniquely corresponding feature map $\Phi : V \longrightarrow \mathcal{H}$ is injective according to [46]. Hence, it is bijective for $\Phi : V \longrightarrow \mathcal{I}_{\kappa_\Phi} \subseteq \mathcal{H}$, whereby $\mathcal{H}$ is equipped with the Hilbert space metric $d_{\mathcal{H}}$. Because (V, d_V) is compact and the bijective mapping Φ is continuous, it follows immediately that $\mathcal{I}_{\kappa_\Phi}$ is a subspace of $\mathcal{H}$ and, therefore, a Hilbert space itself. Moreover, it follows from Lemma 1 that Φ is also continuous as well as the obviously bijective identity map $\Psi : (V, d_V) \longrightarrow (V, d_{\mathcal{H}})$ from (5). Hence, the map $\Phi\left(\Psi^{-1}(\mathbf{v})\right) = \Phi \circ \Psi^{-1}(\mathbf{v})$ with $\mathbf{v} \in (V, d_{\mathcal{H}})$ is bijective and continuous. Therefore, $(V, d_{\mathcal{H}})$ and $\mathcal{I}_{\kappa_\Phi}$ are isomorphic and, according to the Remark 2, also isometric. The separability of Ψ follows immediately from the separability property of Φ. ∎

It is well known that the Gaussian kernel $\kappa_\Phi(\mathbf{u}, \mathbf{v}) = \exp\left(\frac{-||\mathbf{u}-\mathbf{v}||_E^2}{2\sigma^2}\right)$, the Student-type Gaussian kernel $\kappa_\Phi(\mathbf{u}, \mathbf{v}) = \left(\beta + \frac{||\mathbf{u}-\mathbf{v}||_E^2}{\sigma^2}\right)^{-\alpha}$ with $\alpha, \beta > 0$ and

the exponential kernel $\kappa_\Phi(\mathbf{u},\mathbf{v}) = \exp(\langle \mathbf{u},\mathbf{v}\rangle_E)$ are universal on every compact subset of $\mathbb{R}^n$. Other universal kernel can be found in [28,43,46]. At this point we remark that these kernels are also differentiable, which becomes important in Sect. 4.

Another class of kernels are *information theoretic kernels* based on divergences [25,33]. This class is investigated in the light of universality in the next subsection. The relation of universal kernels to *characteristic kernels* is addressed in [45].

3 Universal Kernels Based on Divergences

Information theoretic kernels based on divergences are considered in many applications [8,23,25,33]. Here we relate them to universal differentiable kernels, such that the diagram in Fig.1 holds also for those kernels. For this purpose, we introduce the class of *radial kernels* $\kappa_r : \mathbb{R}^m \times \mathbb{R}^m \longrightarrow \mathbb{R}$ [19,41,43]. These kernels are defined as

$$\kappa_r(\mathbf{u},\mathbf{v}) = g(d(\mathbf{u},\mathbf{v})) \tag{9}$$

where $d(\mathbf{u},\mathbf{v})$ is a metric and g is a function on $\mathbb{R}_0^+ = \{x \in \mathbb{R} | x \geq 0\}$. Equivalently, $d(\mathbf{u},\mathbf{v})$ could be a norm of the difference $(\mathbf{u}-\mathbf{v})$. One important point to be emphasized here is that the argument of a radial kernel is required to be a metric or, equivalently, a norm. Radial kernels stand out due to its close relation to universal kernels. The following lemma holds for radial kernels [45]:

Lemma 3. *If the radial kernel is strictly positive definite then it is also universal.*

If we want to obtain a differentiable universal kernel based on divergences, we have, hence, to ensure that the divergence is differentiable, metric, and that the corresponding radial kernel is positive definite. Generally, divergences are not symmetric and, therefore, cannot serve as a metric [9,10,14]. Yet, there exist some special divergences for vectorized data, which are metrics at the same time under the assumption that the data vectors represent probability densities or at least positive functions [47]. For example, the Euclidean distance is a so-called η-divergence belonging to the class of Bregman-divergences with parameter $\eta = 2$ [30]. Österreicher and Vajda considered a subset of Csiszár's f-divergences to be metric [31,47]. To this class belongs the subclass of f_β-divergences, a prominent member of which is the squared *Hellinger distance*

$$D_H(\mathbf{u}\|\mathbf{v}) = \sum_{i=1}^{m} (\sqrt{u_i} - \sqrt{v_i})^2 \tag{10}$$

obtained for the value $\beta = \frac{1}{2}$. Another example is the *Jensen-Shannon-divergence*

$$D_{JS}(\mathbf{u}\|\mathbf{v}) = \frac{D_{KL}(\mathbf{u}\|\mathbf{w}) + D_{KL}(\mathbf{v}\|\mathbf{w})}{2} \tag{11}$$

obtained for $\beta = 1$ with $\mathbf{w} = \frac{\mathbf{u}+\mathbf{v}}{2}$ and

$$D_{KL}(\mathbf{u}\|\mathbf{w}) = \sum_{i=1}^{m} u_i \log \frac{u_i}{v_i} \tag{12}$$

being the *Kullback-Leibler-divergence* [24]. It can be calculated based on the *Shannon-entropy*

$$H(\mathbf{v}) = -\sum_{i=1}^{m} v_i \log v_i \tag{13}$$

as

$$D_{JS}(\mathbf{u}\|\mathbf{v}) = H\left(\frac{\mathbf{u}+\mathbf{v}}{2}\right) - \left(\frac{H(\mathbf{u}) + H(\mathbf{v})}{2}\right) \tag{14}$$

as shown in [25,44].

An analog divergence can be installed using the *Rényis α-entropy*

$$H_\alpha(\mathbf{v}) = \frac{1}{1-\alpha} \log\left(\sum_{i=1}^{m} (v_i)^\alpha\right) \tag{15}$$

defined for $\alpha > 0$ [36,37]. In the limit $\alpha \to 1$ $H_\alpha(\mathbf{v})$ converges to the Shannon-entropy $H(\mathbf{v})$ from (13). Based on the Rényi-entropy (15) the *Jensen-Rényi-α-divergence* is defined as

$$D_{JR}^{\alpha}(\mathbf{u}\|\mathbf{v}) = H_\alpha\left(\frac{\mathbf{u}+\mathbf{v}}{2}\right) - \left(\frac{H_\alpha(\mathbf{u}) + H_\alpha(\mathbf{v})}{2}\right) \tag{16}$$

in complete analogy to (14) [2]. It turns out that both, $\sqrt{D_{JS}(\mathbf{u}\|\mathbf{v})}$ and $\sqrt{D_{JR}^{\alpha}(\mathbf{u}\|\mathbf{v})}$, are metrics [25] or, more precisely, they are Hilbertian metrics [17]. Moreover it is shown in the paper [25] by Martin et al. that the following lemma holds:

Lemma 4. *The kernels*

1. $\kappa_{JS}^{1}(\mathbf{u},\mathbf{v}) = \exp(-t \cdot D_{JS}(\mathbf{u}\|\mathbf{v}))$, $t > 0$,
2. $\kappa_{JR}^{1}(\mathbf{u},\mathbf{v},\alpha) = \exp(-t \cdot D_{JR}^{\alpha}(\mathbf{u}\|\mathbf{v}))$, $t > 0$,
3. $\kappa_{JS}^{2}(\mathbf{u},\mathbf{v}) = (t + D_{JS}(\mathbf{u}\|\mathbf{v}))^{-1}$, $t > 0$ *and*
4. $\kappa_{JR}^{2}(\mathbf{u},\mathbf{v},\alpha) = (t + D_{JR}^{\alpha}(\mathbf{u}\|\mathbf{v}))^{-1}$, $t > 0$

are strictly positive definite. For the kernels κ_{JR}^{1} and κ_{JR}^{2} the additional condition of $\alpha \in [0,1]$ has to be fulfilled for positive definiteness.

Therefore, we can finally state the following corollary for divergence based kernels:

Corollary 1. *The kernels given in Lemma 4 based on the Jensen-Shannon-divergence (14) and the Jensen-Rényi-α-divergence (16) are universal.*

Proof. This property follows immediately from Lemma 4 together with the Lemma 3. ∎

Last but not least we remark again that the kernels defined in Lemma 4 are differentiable [47], which relates them to the considerations in Sect. 4.

4 Differentiable Kernel and Gradient Based Vector Quantization

Vector quantization can be distinguished into unsupervised and supervised approaches. The main task for unsupervised models is to minimize some reconstruction error E for a given data set $V \subseteq \mathbb{R}^n$ of vectors $\mathbf{v}$ with respect to set of prototypes $W = \{\mathbf{w}_k\}_{k \in A}$, where A is a finite index set. Prominent examples are the self-organizing map (SOM,[21]), neural gas (NG, [26]), whereby for the SOM the variant of Heskes is taken [18]. For those models, the reconstruction error is given in terms of the dissimilarity measure $d(\mathbf{v}, \mathbf{w}_k)$ between data and prototypes, which is assumed to be differentiable. Adaptation for these models is frequently realized as a stochastic gradient descent. In that case, the gradient $\partial E / \partial \mathbf{w}_k$ contains the derivative $\partial d(\mathbf{v}, \mathbf{w}_k) / \partial \mathbf{w}_k$ originating from the chain rule of differentiation. For example, the cost function of the Heskes variant of SOM is

$$E_{\mathrm{SOM}} = \int P(\mathbf{v}) \sum_{\mathbf{r} \in A} \delta_{\mathbf{r}}^{s(\mathbf{v})} \sum_{\mathbf{r}' \in A} \frac{h_\sigma^{SOM}(\mathbf{r}, \mathbf{r}')}{2K(\sigma)} d(\mathbf{v}, \mathbf{w}_{\mathbf{r}'}) d\mathbf{v} \tag{17}$$

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 the SOM-lattice A according to its topological structure [18]. Further, $P($**v** is the data density and the Kronegger symbol $\delta_{\mathbf{r}}^{s(\mathbf{v})}$ assigns a data vector $\mathbf{v}$ to the winning unit $s(\mathbf{v})$. $K(\sigma)$ is a normalization constant depending on the neighborhood range σ. Then the stochastic gradient prototype update for all prototypes is given as [18]:

$$\triangle \mathbf{w}_{\mathbf{r}} = -\varepsilon h_\sigma^{SOM}(\mathbf{r}, s(\mathbf{v})) \frac{\partial d(\mathbf{v}, \mathbf{w}_{\mathbf{r}})}{\partial \mathbf{w}_{\mathbf{r}}}. \tag{18}$$

depending on the derivatives of the used dissimilarity measure d, which allows the application of differentiable kernel metrics.

Prototype based classification in the context of learning vector quantization models (LVQ, [21]) was renewed by the idea of Sato&Yamada to approximate the non-differentiable classification error C by a differentiable function E_C referred as *Generalized* LVQ (GLVQ,[39,38]). As in unsupervised vector quantization, E_C depends on the underlying dissimilarity measure $d(\mathbf{v}, \mathbf{w}_k)$ according to

$$E_C(W) = \frac{1}{2} \sum_{\mathbf{v} \in V} f(\mu(\mathbf{v})) \text{ with } \mu(\mathbf{v}) = \frac{d^+(\mathbf{v}) - d^-(\mathbf{v})}{d^+(\mathbf{v}) + d^-(\mathbf{v})} . \tag{19}$$

with $d^+(\mathbf{v}) = d(\mathbf{v}, \mathbf{w}^+)$ denoting the distance between the data point $\mathbf{v}$ and the nearest prototype $\mathbf{w}^+$, belonging to the same class as the presented data point $\mathbf{v}$. Analogously, $d^-(\mathbf{v})$ is defined as the distance to the best matching prototype of all other classes. The function $\mu(\mathbf{v})$ is the classifier function. Like in SOMs, $d(\mathbf{v}, \mathbf{w})$ in (19) is required to be some differentiable dissimilarity measure with respect to $\mathbf{w}$. Then the cost function can be minimized by gradient descent learning based on the (stochastic) derivatives

$$\frac{\partial_s E}{\partial \mathbf{w}^+} = \frac{2d^- \cdot f'(\mu(\mathbf{v}))}{(d^+ + d^-)^2} \frac{\partial d^+}{\partial \mathbf{w}^+}, \quad \frac{\partial_s E}{\partial \mathbf{w}^-} = -\frac{2d^+ \cdot f'(\mu(\mathbf{v}))}{(d^+ + d^-)^2} \frac{\partial d^-}{\partial \mathbf{w}^-} \quad , \tag{20}$$

where we used the abbreviations d^+ for $d^+(\mathbf{v})$ for simplicity and d^+, analogously.

Thus, stochastic gradient learning in supervised and unsupervised vector quantization can be seen as a gradient descent learning of an error function in the metric space $(V, d(\mathbf{v}, \mathbf{w}_k))$. Obviously, under gentle conditions on V (continuous, local convex, ...) it can be assumed that $\partial d(\mathbf{v}, \mathbf{w}_k)/\partial \mathbf{w}_k \in V$ is valid. Yet, the choice of the metric is free except the necessary differentiability. Hence, metrics determined by differentiable kernel are applicable [15]. Obviously, the kernels presented in Sec.2 and 3 are differentiable (for the latter kernels, see [47] for differentiability of the respective divergences). If such a metric is obtained from an universal kernel κ_Φ for RKHS, respectively, the Lemma 2 ensures the topological and isometric equivalence to the respective FMHS. Hence, the algorithm operates in the same structural space as SVMs do and, therefore, can profit from its richness in shape, which frequently delivers excellent performance. At this point we empasize the following essential drawn from the Lemma 2:

Remark 3. The take home message of the Lemma 2 in context of gradient based online learning is: Assume a set of prototypes W', which has to be learned in the induced image space $\mathcal{I}_{\kappa_\Phi} \subseteq \mathcal{H}$. Because $\mathcal{I}_{\kappa_\Phi}$ is a subspace of $\mathcal{H}$ any linear combinations of prototypes belongs to $\mathcal{H}$. Further, if the corresponding universal kernel κ_Φ is continuous and differentiable, it is sufficient to train prototypes W by gradient descent learning in the isomorph-isometric space $(V, d_\mathcal{H})$ induced by the mapping Ψ. Lemma 2 ensures the unique equivalence. An analogous statement obviously holds also for the Banach space problem.

More properties of differentiable Mercer-like kernels and their reproducing properties can be found in [12,48].

5 Exemplary Applications

In this section we briefly give results from exemplary applications for classification problems. We compare the GLVQ with differentiable kernels (DK-GLVQ) with several state-of-the-art prototype based classification algorithms including SVMs using Gaussian kernels based on an Extreme Learning Kernel (ELM,[13]) and improved GLVQ variants. For the latter we consider standard GLVQ with Euclidean metric, and the powerful variant based on matrix learning (GMLVQ, [42]) as a generalization of the relevance learning approach [16]. The GMLVQ uses the distance $d(\mathbf{v}, \mathbf{w}) = (\Omega(\mathbf{v} - \mathbf{w}))^2$ with a here squared matrix Ω, which is automatically adapted during learning for optimal classification performance. Moreover, we include the recently proposed kernel GLVQ (KGLVQ) based on a Nystrøm-approximation [40]. For the DK-GLVQ we applied two variants: The first one used a Gaussian kernel with self-adapting kernel-with σ. The second one uses the kernel $\kappa_\Phi(\mathbf{v}, \mathbf{w}) = \exp\left(-(\Omega(\mathbf{v} - \mathbf{w}))^2\right)$ with a self-adapting non-degenerating (squared) matrix for comparison with GMLVQ. We refer to this

variant as DK-GMLVQ. This variant is much more stable than the σ-adjusting variant which may be addressed to the regularizing properties of matrix learning known from GMLVQ [5,6].

We applied these algorithms to two standard benchmark data sets taken from UCI repository [7]. Both data sets are two-class problems to to establish compatibility with SVMs. The first one is breast cancer data set (WDBC) consisting of 569 samples with 32 dimensions. The second data set is a diabetes study (PIMA) with 768 eight-dimensional samples. All experiments were performed by three-fold cross-validation. For the GLVQ variants we used one prototype for each class. The SVM resulted in 512 and 691 support vectors for both problems, respectively. The results are depicted in Tab. 1

Table 1. Classification accuracies in % together with their variances for the several algorithms and datasets (PIMA and WDBC). Results are obtained by three-fold cross-validation.

Dataset	GLVQ	KGLVQ	DK-GLVQ	GMLVQ	DK-GMLVQ	SVM-ELM
PIMA	75.1(±0.062)	71.1 (±0.031)	76.2 (±0.031)	77.7 (±0.016)	**78.3** (±0.025)	76.4 (±0.042)
WDBC	93.49(±0.016)	92.3(±0.034)	92.2(±0.009)	94.7 (±0.020)	95.4 (±0.025)	**97.7** (±0.014)

We observe a good performance of both kernel GLVQ variants using differentiable kernels. These are significantly improved compared to KGLVQ, which uses approximation techniques. Hence, we can conclude that the Nyström-approximation leads to a significant loss in accuracy. Further, comparison to GLVQ and GMLVQ also shows clear improvements, although standard GMLVQ achieved high performance. Last but not least, comparison to the SVM demonstrates that differentiable kernel are an excellent alternative to SVM. In particular we emphasize the drastically reduced model complexity taking only two prototypes compared to hundreds of support vectors while achieving similar accuracies.

6 Conclusion

In this paper we considered the theoretical framework of differentiable kernels for application in gradient based learning in supervised and unsupervised prototype based vector quantization. We show that utilization of a data metric determined by universal kernels as known from support vector machines leads to an optimization space equivalent and isometric to a reproducing kernel Hilbert space. Hence, gradient based vector quantization schemes with differentiable universal kernels can benefit from this property. The main results of topological and isometric equivalence is the Lemma 2. An extension of this theory for reproducing kernel Banach spaces can be found in [48], which assume weaker restrictions and, therefore, offer greater flexibility [49]. Last but not least we provide some examples of differentiable universal kernels based on divergences as fundamental information theoretic concepts. Further, we demonstrated abilities of GLVQ

using differentiable kernel for exemplary datasets, which show high performance also compared to SVMs but with lower model complexity.

Otherwise, the presented approach cannot deal with arbitrary kernels such as structure kernels. So the method trades increased efficiency by reduced flexibility in kernel choice.

References

1. Aronszajn, N.: Theory of reproducing kernels. Transactions of the American Mathematical Society 68, 337–404 (1950)
2. Ben-Hamza, A., Krim, H.: Jensen-Rényi divergence measure: theoretical and computational perspectives. In: Proceedings of the IEEE International Symposium on Information Theory, pp. 257–257 (2003)
3. Bezdek, J.: A convergence theorem for the fuzyy ISODATA clustering algorithm. IEEE Transactions on Pattern Analysis and Machine Intelligence 2(1), 1–8 (1980)
4. Bezdek, J.: Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum, New York (1981)
5. Biehl, M., Bunte, K., Schleif, F.-M., Schneider, P., Villmann, T.: Large margin discriminative visualization by matrix relevance learning. In: Abbass, H., Essam, D., Sarker, R. (eds.) Proc. of the International Joint Conference on Neural Networks (IJCNN), Brisbane, pp. 1873–1880. IEEE Computer Society Press, Los Alamitos (2012)
6. Biehl, M., Hammer, B., Schleif, F.-M., Schneider, P., Villmann, T.: Stationarity of matrix relevance learning vector quantization. Machine Learning Reports 3(MLR-01-2009), 1–17 (2009) ISSN:1865-3960, `http://www.uni-leipzig.de/~compint/mlr/mlr_01_2009.pdf`
7. Blake, C., Merz, C.: UCI repository of machine learning databases. Dep. of Information and Computer Science, University of California, Irvine (1998), `http://www.ics.edu/mlearn/MLRepository.html`
8. Chan, A., Vasconcelos, N., Moreno, P.: A family of probabilistic kernels based on information divergence. Technical Report SVCL-TR 2004/01, Statistical Visual Computing Laboratory (SVCL) at Universit of California, San Diego (2004)
9. Cichocki, A., Amari, S.-I.: Families of alpha- beta- and gamma- divergences: Flexible and robust measures of similarities. Entropy 12, 1532–1568 (2010)
10. Cichocki, A., Cruces, S., Amari, S.-I.: Generalized alpha-beta divergences and their application to robust nonnegative matrix factorization. Entropy 13, 134–170 (2011)
11. Crammer, K., Gilad-Bachrach, R., Navot, A., Tishby, A.: Margin analysis of the LVQ algorithm. In: Becker, S., Thrun, S., Obermayer, K. (eds.) Advances in Neural Information Processing (Proc. NIPS 2002), vol. 15, pp. 462–469. MIT Press, Cambridge (2003)
12. Ferreira, J., Menegatto, V.: Reproducing properties of differentiable Mercer-like kernels. Mathematische Nachrichten 285 (in press, 2012)
13. Frénay, B., Verleysen, M.: Parameter-free kernel in extreme learning for non-linear support vector regression. Neurocomputing 74(16), 2526–2531 (2011)
14. Gretton, A., Herbrich, R., Smola, A., Bousquet, O., Schölkopf, B.: Kernel methods for measuring independence. Journal of Machine Learning Research 6, 2075–2129 (2005)
15. Hammer, B., Strickert, M., Villmann, T.: Supervised neural gas with general similarity measure. Neural Processing Letters 21(1), 21–44 (2005)

16. Hammer, B., Villmann, T.: Generalized relevance learning vector quantization. Neural Networks 15(8-9), 1059–1068 (2002)
17. Hein, M., Bousquet, O.: Hilbertian metrics and positive definite kernels on probability measures. Technical report, Max Planck Institute for Biological Cybernetics (2004)
18. Heskes, T.: Energy functions for self-organizing maps. In: Oja, E., Kaski, S. (eds.) Kohonen Maps, pp. 303–316. Elsevier, Amsterdam (1999)
19. Hoffmann, T., Schölkopf, B., Smola, A.: Kernel methods in machine learning. The Annals of Statistics 36(3), 1171–1220 (2008)
20. Kästner, M., Hammer, B., Biehl, M., Villmann, T.: Functional relevance learning in generalized learning vector quantization. Neurocomputing 90(9), 85–95 (2012)
21. Kohonen, T.: Self-Organizing Maps. Springer Series in Information Sciences, vol. 30. Springer, Heidelberg (1995) (Second Extended Edition 1997)
22. Kolmogorov, A., Fomin, S.: Reelle Funktionen und Funktionalanalysis. VEB Deutscher Verlag der Wissenschaften, Berlin (1975)
23. Kulis, B., Sustik, M., Dhillon, I.: Low-rank kernel learning with Bregman matrix divergences. Journal of Machine Learning Research 10, 341–376 (2009)
24. Kullback, S., Leibler, R.: On information and sufficiency. Annals of Mathematical Statistics 22, 79–86 (1951)
25. Martin, A., Smith, N., Xing, E., Aguiar, P., Figueiredo, M.: Nonextensive information theoretic kernels on measures. Journal of Machine Learning Research 10, 935–975 (2009)
26. 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)
27. Mercer, J.: Functions of positive and negative type and their connection with the theory of integral equations. Philosophical Transactions of the Royal Society, London, A 209, 415–446 (1909)
28. Micchelli, C., Xu, Y., Zhang, H.: Universal kernels. Journal of Machine Learning Research 7, 26051–22667 (2006)
29. Mwebaze, E., Schneider, P., Schleif, F.-M., Aduwo, J., Quinn, J., Haase, S., Villmann, T., Biehl, M.: Divergence based classification in learning vector quantization. Neurocomputing 74(9), 1429–1435 (2011)
30. Nielsen, F., Nock, R.: Sided and symmetrized Bregman centroids. IEEE Transaction on Information Theory 55(6), 2882–2903 (2009)
31. Österreicher, F., Vajda, I.: A new class of metric divergences on probability spaces and its applicability in statistics. Annals of the Institute of Statistical Mathematics 55(3), 639–653 (2003)
32. Pekalska, E., Duin, R.: The Dissimilarity Representation for Pattern Recognition: Foundations and Applications. World Scientific (2006)
33. Principe, J.: Information Theoretic Learning. Springer, Heidelberg (2010)
34. 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)
35. Qin, A.K., Suganthan, P.N.: Kernel neural gas algorithms with application to cluster analysis. In: ICPR (4), pp. 617–620 (2004)
36. Rényi, A.: On measures of entropy and information. In: Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability. University of California Press (1961)
37. Rényi, A.: Probability Theory. North-Holland Publishing Company, Amsterdam (1970)

38. Sato, A., Yamada, K.: Generalized learning vector quantization. In: Touretzky, D.S., Mozer, M.C., Hasselmo, M.E. (eds.) Proceedings of the 1995 Conference Advances in Neural Information Processing Systems, vol. 8, pp. 423–429. MIT Press, Cambridge (1996)
39. Sato, A.S., Yamada, K.: Generalized learning vector quantization. In: Tesauro, G., Touretzky, D., Leen, T. (eds.) Advances in Neural Information Processing Systems, vol. 7, pp. 423–429. MIT Press (1995)
40. Schleif, F.-M., Villmann, T., Hammer, B., Schneider, P.: Efficient kernelized prototype based classification. International Journal of Neural Systems 21(6), 443–457 (2011)
41. Schölkopf, B., Smola, A.: Learning with Kernels. MIT Press (2002)
42. Schneider, P., Hammer, B., Biehl, M.: Adaptive relevance matrices in learning vector quantization. Neural Computation 21, 3532–3561 (2009)
43. Scovel, C., Hush, D., Steinwart, I., Theiler, J.: Radial kernels and their reproducing kernel Hilbert spaces. Journal of Complexity 26, 641–660 (2010)
44. Shannon, C.: A mathematical theory of communication. Bell System Technical Journal 27, 379–432 (1948)
45. Sriperumbudur, B., Fukumizu, K., Lanckriet, G.: Universality, characteristic kernels, and RKHS embedding of measures. Journal of Machine Learning Research 12, 2389–2410 (2011)
46. Steinwart, I.: On the influence of the kernel on the consistency of support vector machines. Journal of Machine Learning Research 2, 67–93 (2001)
47. Villmann, T., Haase, S.: Divergence based vector quantization. Neural Computation 23(5), 1343–1392 (2011)
48. Villmann, T., Haase, S.: A note on gradient based learning in vector quantization using differentiable kernels for Hilbert and Banach spaces. Machine Learning Reports 6(MLR-02-2012), 1–29 (2012) ISSN:1865-3960, `http://www.techfak.uni-bielefeld.de/~fschleif/mlr/mlr$_$02$_$2012.pdf`,
49. Zhang, H., Xu, Y., Zhang, J.: Reproducing kernel banach spaces for machine learning. Journal of Machine Learning Research 10, 2741–2775 (2009)