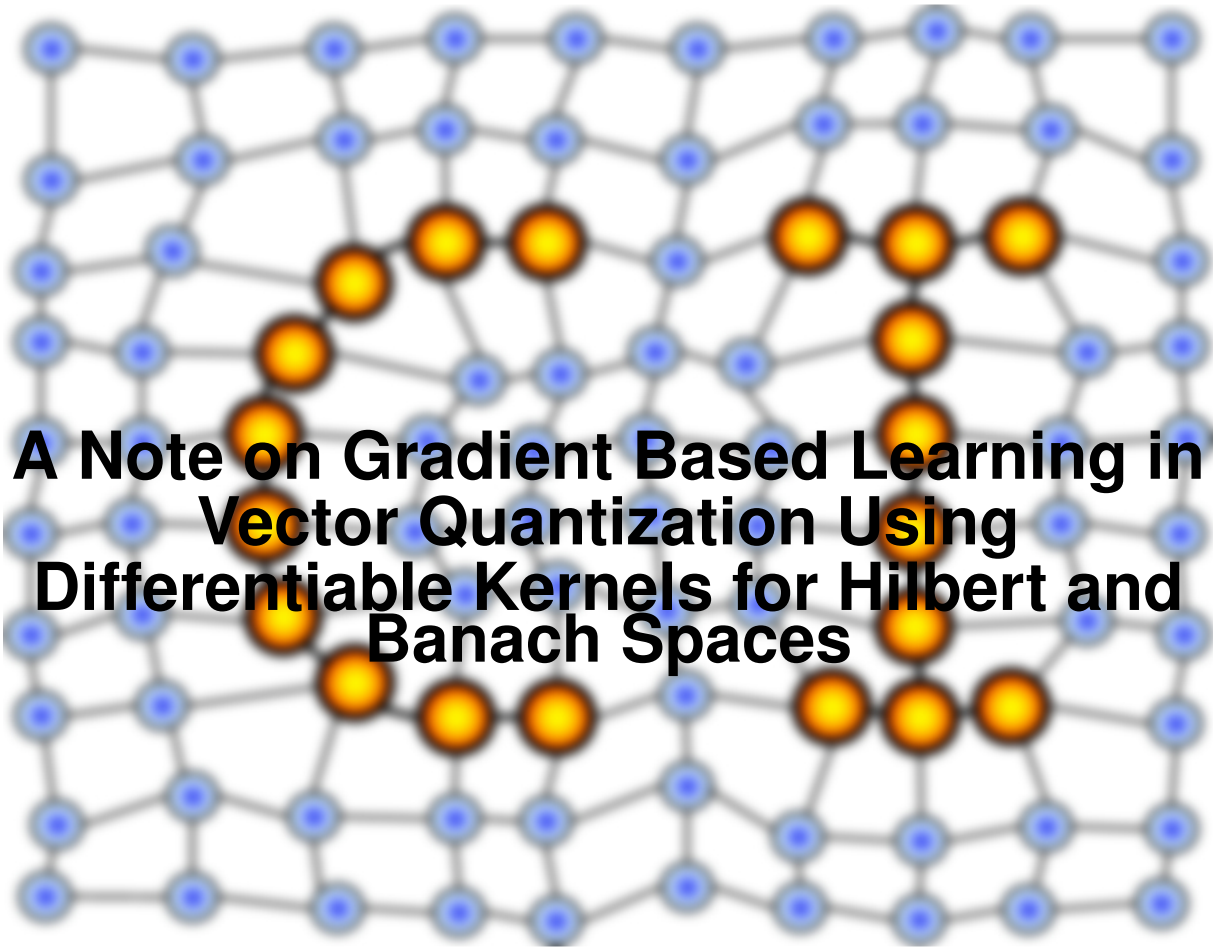
MACHINE LEARNING REPORTS



A Note on Gradient Based Learning in Vector Quantization Using Differentiable Kernels for Hilbert and Banach Spaces

Report 02/2012
Submitted: 17.04.2012
Published: 24.04.2012

T. Villmann*, S. Haase
(1) Computational Intelligence Group, University of Applied Sciences, Mittweida,
Technikumplatz 17, 09648 Mittweida, Germany
* corresponding author, *email: thomas.villmann@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 or Banach 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 an 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.

A Note on Gradient Based Learning in Vector Quantization Using Differentiable Kernels for Hilbert and Banach Spaces

Thomas Villmann* and Sven Haase

Computational Intelligence Group,
University of Applied Sciences Mittweida,
Technikumplatz 17, 09648 Mittweida, Germany,

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 or Banach 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 an 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.

* *corresponding author*, email: thomas.villmann@hs-mittweida.de

1 Introduction

Vector quantization by prototypes is one of the key methods in unsupervised and supervised machine learning. Prominent examples for unsupervised models applied in data clustering or visualization are the self-organizing map (SOM,[19]), neural gas (NG, [25]) as a robust version of the k-means or respective fuzzy variants like fuzzy-c-means (FCM, [3, 4]) and alternatives thereof. Supervised prototype based approaches are mainly influenced by the learning vector quantization models (LVQ, [19]) and support vector machines (SVM,[42]). Whereas LVQ models generate class typical prototypes SVMs determine prototypes (support vectors) defining the class borders. Both paradigms are margin classifiers [8]. During the last years application of non-standard metrics for these models became popular to improve the classifier performance for domain specific problems like processing of functional data, e.g. spectra, time series, ...,[18, 28, 48] or better interpretability of the adapted models (relevance and matrix learning, [13, 43]).

One key idea remaining powerful in classification is the idea of kernel mapping realized in SVMs. According to this idea, the data as well as the prototypes are described and handled in a high-dimensional (infinite) feature mapping Hilbert space (FMHS) uniquely determined by the kernel, which offers frequently a great flexibility and good separation possibility. Yet, this processing is done only implicitly in the mapping space. 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 SVM prototypes are not typical representers of the classes, as mentioned before. Several variants of LVQ were established integrating the kernel mapping concept in those models to keep the idea of class-typical prototypes (Kernel GLVQ, KGLVQ) [41, 36, 35]. However, these models also have to deal with the problem of the infinite representation of prototypes. Usually, the infinite representation is approximated by a finite one using the Nystrøm-approximation approach, which obviously leads to an information loss in general.

In this paper we provide a way to overcome this circumstance: we want to have in the new model the topological richness of the FMHS to keep the high classification ability and data separability while avoiding the infinite data and prototype representation or its necessary approximation. For this purpose we suggest the utilization of universal differentiable kernels in vector quantization models defining a new metric in the data space. Now, the differentiability ensures that the prototype adaptation can be processed in this new metric space without any approximation requirements or other Hilbert space representations. Further, we show that this new metric space is topologically equivalent to the FMHS associated to the universal kernel, such that the demanded topological richness is kept. More specifically, we show that both spaces are isometric. Additionally, we demonstrate that this framework can also be applied for a recently proposed variant of kernel feature mapping, where the feature mapping space is a certain type of Banach spaces with weaker assumptions than a Hilbert space [49].

2 Reproducing Kernels for Hilbert Spaces

2.1 General Kernels for Hilbert Spaces

In the following we assume a compact metric space (V, d_V) with the vector space V equipped with a metric d_V. A function κ on V is a kernel

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

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 the 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 [47]:

Lemma 2.1 *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*

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

[1] Note, for a semi-metric the triangle inequality does not hold [33].

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

Remark 2.2 *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 maps into a Banach space $\mathcal{B}$.*

It turns out from this lemma that for each function f induced by a continuous kernel κ_Φ is continuous itself. This property is needed for the definition of an *universal kernel:*

Definition 2.3 *A continuous kernel κ_Φ on a compact metric space (V, d_V) is called universal if the space $\mathcal{I}_{\kappa_\Phi}$ of all functions induced by κ_Φ is dense in the space of continuous functions $\mathcal{C}(V)$ over V, i.e. for all $g \in \mathcal{C}(V)$ and $\varepsilon > 0$ exists a function $f \in \mathcal{I}_{\kappa_\Phi}$ with $\|f - g\|_\infty \leq \varepsilon$.*

Following the explanations from I. Steinwart in [47] we can conclude first that every universal kernel separates all compact subsets. Second, this statement leads us to the most important result of that publication with respect to the aim of our paper:

Theorem 2.4 *Every feature map Φ of an universal kernel κ_Φ is injective.*

Remark 2.5 *Here we have again to emphasize an important observation: In the proof of this theorem, again, the inner product property (3) of the kernel is never used. Only its corresponding semi-metric properties are needed, which remain valid also regarding Banach spaces instead of Hilbert spaces.*

2.2 Positive and Universal Kernels for Hilbert Spaces

An important role in feature mapping play the positive definite kernels. The kernel κ_Φ is said to be positive definite if for all finite subsets $V_n \subseteq V$ with cardinality $\# V_n = n$, the Gram-Matrix

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

is positive semi-definite [1]. The kernel is strictly positive definite if the Gram-matrices $\mathbf{G}_n$ are strictly positive definite. These positive kernels are of special interest because they *uniquely* correspond to Hilbert spaces $\mathcal{H}$ in a canonical manner according to the Mercer-theorem: For each feature map Φ (2) there exists a canonical, unique positive kernel

$$\kappa_\Phi : V \times V \to \mathbb{R} \tag{7}$$

satisfying (3) and, conversely, each positive kernel κ_Φ defines uniquely a Hilbert space $\mathcal{H}$ and a corresponding mapping Φ such that the equation (3) is valid [1, 26]. In that case, the space $\mathcal{H}$ is a Hilbert space of functions on V for which point evaluations are always continuous linear functions. In particular, it is a so-called *reproducing kernel Hilbert space* (RKHS) i.e. $\kappa_\Phi(\mathbf{v}, \cdot) \in \mathcal{H}$ such that for each $\mathbf{v} \in V$ and all $f \in \mathcal{H}$ and $\mathbf{w} \in V$

$$f(\mathbf{w}) = \langle f, \kappa_\Phi(\mathbf{w}, \cdot)\rangle_{\mathcal{H}}$$

is valid according to the Riesz representation theorem [1, 20]. For this case, κ_Φ is denoted as a *reproducing kernel.* Reproducing kernels obviously are symmetric, real and, hence, bi-linear. The space $\mathcal{I}_{\kappa_\Phi}$ of induced functions is now given as the set

$$\mathcal{I}_{\kappa_\Phi} = \{\kappa_\Phi(\mathbf{w}, \cdot) | \mathbf{w} \in V\} \tag{8}$$

with $\mathcal{I}_{\kappa_\Phi} \subseteq \mathcal{H}$. For positive kernels the associated inner product implies a norm

$$\|\Phi(\mathbf{v})\|_{\mathcal{H}} = \sqrt{\langle \Phi(\mathbf{v}), \Phi(\mathbf{v})\rangle_{\mathcal{H}}} \tag{9}$$

and, hence, also a metric

$$d_{\mathcal{H}}(\Phi(\mathbf{v}), \Phi(\mathbf{w})) = \|\Phi(\mathbf{v}) - \Phi(\mathbf{w})\|_{\mathcal{H}} = \sqrt{\langle (\Phi(\mathbf{v}) - \Phi(\mathbf{w})), (\Phi(\mathbf{v}) - \Phi(\mathbf{w}))\rangle_{\mathcal{H}}} \tag{10}$$

in the Hilbert space $\mathcal{H}$, i.e. the positive semi-definiteness ensures the metric properties in comparison to the 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

$$d_{\mathcal{H}}(\Phi(\mathbf{v}), \Phi(\mathbf{w})) = \sqrt{\kappa_\Phi(\mathbf{v}, \mathbf{v}) - 2\kappa_\Phi(\mathbf{v}, \mathbf{w}) + \kappa_\Phi(\mathbf{w}, \mathbf{w})} \tag{11}$$

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

Remark 2.6 *Obviously, for positive kernels the semi-metric d_{κ_Φ} from (4) coincides with $d_{\mathcal{H}}$ on $\mathcal{I}_{\kappa_\Phi}$.*

In conclusion we explicitly state the following lemma:

$$\begin{tikzcd}
(V, d_V) \arrow[r, "\Phi, \kappa_\Phi"] \arrow[d, "\Psi"'] & \mathcal{I}_{\kappa_\Phi} \subseteq (\mathcal{H}, d_\mathcal{H}) \\
(V, d_\mathcal{H}) \arrow[ur, leftrightarrow, "\Phi \circ \Psi^{-1}"'] &
\end{tikzcd}$$

Figure 1: Visualization of the statement of Lemma 2.7: 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 $\Phi \circ \Psi^{-1}$.

Lemma 2.7 *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 (11). 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.*

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 Theorem 2.4. 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 2.1 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 2.6, also isometric. ■

The result of the Lemma 2.7 is visualized in Fig.2.2

We now give some examples of universal kernels taken from [44, 47] and [27].

Example 2.8 *The following kernels are universal:*

1. *The Gaussian kernel* $\kappa_\Phi(\mathbf{u},\mathbf{v}) = \exp\left(\frac{-||\mathbf{u}-\mathbf{v}||_E^2}{2\sigma^2}\right)$ *is universal on every compact subset of* $\mathbb{R}^n$ *whereby* $||\cdot||_E$ *is the stnadard Euclidean distance.*
2. *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* $\alpha, \beta > 0$ *is universal on every compact subset of* $\mathbb{R}^n$.
3. *The exponential kernel* $\kappa_\Phi(\mathbf{u},\mathbf{v}) = \exp\left(\langle\mathbf{u},\mathbf{v}\rangle_E\right)$ *is universal on every compact subset of* $\mathbb{R}^n$ *with* $\langle\cdot,\cdot\rangle_E$ *being the standard Euclidean inner product.*
4. *Let* $V_1 = \{\mathbf{v}\in\mathbb{R}^n : ||\mathbf{v}||_E < 1\}$ *the open unit ball and* $\alpha > 0$. *Then the so-called* infinite polynomial kernel $\kappa_\Phi(\mathbf{u},\mathbf{v}) = \left(\beta - \langle\mathbf{u},\mathbf{v}\rangle_E\right)^{-\alpha}$ *is universal on each compact subset of* V_1 *for an arbitrary constant* $\beta > 0$.
5. *Let* $P(x) = \sum_{k\in\mathbb{Z}_+} a_k x^k$ *be power series with convergence radius* $r < \infty$ *and all coefficients* a_k *are positive. Then the so-called* dot product kernel $\kappa_\Phi(\mathbf{u},\mathbf{v}) = P\left(\langle\mathbf{u},\mathbf{v}\rangle_E\right)$ *with* $\mathbf{u},\mathbf{v}\in\mathbb{C}^n$ *is universal on each compact subset of* $\mathbb{C}^n$.

At his point we remark that the above kernels are differentiable, which becomes important in Sect. 4. Another class of kernels are *information theoretic kernels* based on divergences [24, 34]. This class is investigated in the light of universality in the next subsection. The relation of universal kernels to *characteristic kernels* is adressed in [46].

2.3 Universal Kernels Based on Divergences

Information theoretic kernels based on divergences are considered in many applications [5, 21, 24, 34]. Here we relate them to universal differentiable kernels, such that the diagram in Fig.2.2 holds also for those kernels. For this purpose, we introduce the class of *radial kernels* $\kappa_r : \mathbb{R}^m \times \mathbb{R}^m \longrightarrow \mathbb{R}$ [16, 42, 44]. These kernels are defined as

$$\kappa_r(\mathbf{u},\mathbf{v}) = g\left(d(\mathbf{u},\mathbf{v})\right) \tag{12}$$

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 [46]:

Lemma 2.9 *If the radial kernel is strictly positive definite then it is also universal.*

Following this lemma, if we want to obtain a differentiable universal kernel based on divergences, we have to ensure that the divergence is

- differentiable
- metric
- and that the corresponding radial kernel is positive definite.

Generally, divergences are not symmetric and, therefore, not serving as a metric [7, 6, 12]. 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 [48]. For example, the Euclidean distance is a so-called η-divergence belonging to the class of Bregman-divergences with parameter $\eta = 2$ [29]. Österreicher and Vajda considered a subset of Csiszár's f-divergences to be metric [32, 48]. To this class belongs the subclass of f_β-divergences. A prominent member of this subclass is the squared *Hellinger distance*

$$D_H(\mathbf{u}\|\mathbf{v}) = \sum_{i=1}^{m} \left(\sqrt{u_i} - \sqrt{v_i}\right)^2 \tag{13}$$

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

$$D_{JS}(\mathbf{u}\|\mathbf{v}) = \frac{D_{KL}(\mathbf{u}\|\mathbf{w}) + D_{KL}(\mathbf{v}\|\mathbf{w})}{2} \tag{14}$$

with $\mathbf{w} = \frac{\mathbf{u}+\mathbf{v}}{2}$ and

$$D_{KL}(\mathbf{u}\|\mathbf{w}) = \sum_{i=1}^{m} u_i \log \frac{u_i}{v_i} \tag{15}$$

being the *Kullback-Leibler-divergence* [22]. It can be calculated based on the *Shannon-entropy [45]*

$$H(\mathbf{v}) = -\sum_{i=1}^{m} v_i \log v_i \tag{16}$$

as

$$D_{JS}(\mathbf{u}\|\mathbf{v}) = H\left(\frac{\mathbf{u}+\mathbf{v}}{2}\right) - \left(\frac{H(\mathbf{u}) + H(\mathbf{v})}{2}\right) \tag{17}$$

as shown in [24].

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

$$H_{\alpha}(\mathbf{v})=\frac{1}{1-\alpha}\log\left(\sum_{i=1}^{m}(v_i)^{\alpha}\right) \tag{18}$$

defined for $\alpha>0$ [37, 38]. In the limit $\alpha\rightarrow 1$ $H_{\alpha}(\mathbf{v})$ converges to the Shannon-entropy $H(\mathbf{v})$ from (16). Based on the Rényi-entropy (18) the *Jensen-Rényi-α-divergence* is defined as

$$D_{JR}^{\alpha}(\mathbf{u}\|\mathbf{v})=H_{\alpha}\left(\frac{\mathbf{u}+\mathbf{v}}{2}\right)-\left(\frac{H_{\alpha}(\mathbf{u})+H_{\alpha}(\mathbf{v})}{2}\right) \tag{19}$$

in complete analogy to (17) [2]. It turns out that both, $\sqrt{D_{JS}(\mathbf{u}\|\mathbf{v})}$ and $\sqrt{D_{JR}^{\alpha}(\mathbf{u}\|\mathbf{v})}$, are metric [24] or, more precisely, they are Hilbertian metrics [14]. Moreover it is shown in the paper [24] by Martin et al. that the following lemma holds:

Lemma 2.10 *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. $\kappa_{JR}^{1}(\mathbf{u},\mathbf{v},\alpha)=(t+D_{JR}^{\alpha}(\mathbf{u}\|\mathbf{v}))^{-1}$, $t>0$

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

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

Corollary 2.11 *The kernels given in Lemma 2.10 based on the Jensen-Shannon-divergence (17) and the Jensen-Rényi-α-divergence (19) are universal.*

Proof. This property follows immediately from Lemma 2.10 together with the Lemma 2.9. ■

Last but not least we remark again that the kernels defined in Lemma 2.10 are differentiable [48], which relates them to the considerations in Sect. 4.

3 Reproducing Kernels for Banach Spaces

Banach spaces are the generalization of Hilbert spaces in such a way that the existence of an inner product is not assumed. Therefore, a straight forward definition of reproducing kernels as for Hilbert spaces is not possible. However, under certain assumptions an analog approach can be established. We adopt the following explanations from Zhang et al. [49].

3.1 Semi-inner Products for Vector Spaces

We need some preliminary definitions, facts and notations in the beginning. We start with the fact that to each vector space V exists a dual space V^* of linear real functions, which itself is again a vector space. A normed vector space V taken as a vector space is called *reflexive* if $(V^*)^* = V$. Further, the normed vector space V is said to be *Gâteaux-differentiable*, if for all elements $\mathbf{v}, \mathbf{w} \in V \setminus \{\mathbf{0}\}$ and $t \in \mathbb{R}$ the Fréchet derivative

$$\partial_F (\mathbf{v}, \mathbf{w}) = \lim_{t \to 0} \frac{\|\mathbf{v} + t \cdot \mathbf{w}\|_V - \|\mathbf{v}\|_V}{t} \tag{20}$$

exists [17]. The space V is denoted as uniformly differentiable or *uniformly Fréchet-differentiable*, if the limit is approached uniformly on $S(V) \times S(V)$ with $S(V) = \{\mathbf{v} \in V, \|\mathbf{v}\|_V = 1\}$ is the unit sphere. Additionally, we need the definition of the concept of uniform convexity: A normed vector space V is *uniformly convex* if for all $\varepsilon > 0$ there exists a $\delta > 0$ such that $\|\mathbf{v} + \mathbf{w}\|_X \leq 2 - \delta$ is valid for all $\mathbf{v}, \mathbf{w} \in S(V)$ with $\|\mathbf{v} - \mathbf{w}\|_V \geq \varepsilon$. The uniform convexity is closely related to the Fréchet differentiability: A normed vector space is uniformly Fréchet differentiable iff its dual space is uniformly convex [9]. Just a last definition is required in advance:

Definition 3.1 *Let V be a vector space and $[\cdot, \cdot]_V : V \times V \longrightarrow \mathbb{C}$ a function such that for all $\mathbf{u}, \mathbf{v}, \mathbf{w} \in V$ and $\alpha \in \mathbb{C}$ the conditions*

1. $[\mathbf{u} + \mathbf{v}, \mathbf{w}]_V = [\mathbf{u}, \mathbf{w}]_V + [\mathbf{v}, \mathbf{w}]_V$
2. $[\alpha \mathbf{u}, \mathbf{v}]_V = \alpha [\mathbf{u}, \mathbf{v}]_V$ *and* $[\mathbf{u}, \alpha \mathbf{v}]_V = \bar{\alpha} [\mathbf{u}, \mathbf{v}]_V$
3. $[\mathbf{v}, \mathbf{v}]_V > 0$ *for* $\mathbf{v} \neq \mathbf{0}$
4. $| [\mathbf{u}, \mathbf{v}]_V |^2 \leq [\mathbf{u}, \mathbf{u}]_V \cdot [\mathbf{v}, \mathbf{v}]_V$ *(Cauchy-Schwarz inequality)*

are valid. Then this function is called a semi-inner product (s.i.p.).

This definition was introduced by G. Lumer in [23] and nourished by J.R. Giles in [11]. The semi-inner product differs from an usual inner product in that way that one can always find $\mathbf{u}, \mathbf{v}, \mathbf{w} \in V$ such that

$$[\mathbf{u}, \mathbf{v}+\mathbf{w}]_V \neq [\mathbf{u}, \mathbf{v}]_V + [\mathbf{u}, \mathbf{w}]_V$$

which is equivalent to the property $[\mathbf{u}, \mathbf{v}]_V \neq \overline{[\mathbf{v}, \mathbf{u}]_V}$ of the conjugate asymmetry. The following lemma was presented in [49]:

Lemma 3.2 *A semi-inner product on a complex vector space V is an inner product iff for all $\mathbf{u}, \mathbf{v}, \mathbf{w} \in V$*

$$[\mathbf{u}, \mathbf{v}+\mathbf{w}]_V = [\mathbf{u}, \mathbf{v}]_V + [\mathbf{u}, \mathbf{w}]_V$$

holds.

Although not being an inner product, the semi-inner product induces a norm for a vector space. The following theorem was proofed in [11] and [23]:

Theorem 3.3 *A vector space V equipped with a semi-inner product $[\cdot, \cdot]_V$ is a normed space with the induced norm*

$$\|\mathbf{v}\|_V = \sqrt{[\mathbf{v}, \mathbf{v}]_V} \tag{21}$$

and, conversely, for a normed vector space always a semi-inner product can be defined, which induces the norm via (21).

According to this theorem we denote a vector space V with a semi-inner product $[\cdot, \cdot]$ a *s.i.p. space*. Obviously, it defines a metric, too. Yet, the determination of a semi-inner product for a given norm is in general not unique. The uniqueness is ensured if the vector space is Fréchet differentiable.

Theorem 3.4 *If a s.i.p. space V is Gâteaux differentiable, then the semi-inner product is uniquely defined for all $\mathbf{v}, \mathbf{w} \in V$ and $\mathbf{v} \neq \mathbf{0}$ by*

$$\operatorname{Re}\left([\mathbf{w}, \mathbf{v}]_V\right) = \|\mathbf{v}\|_V \cdot \partial_F\left(\mathbf{v}, \mathbf{w}\right)$$

with the Fréchet derivative $\partial_F\left(\mathbf{v}, \mathbf{w}\right)$ from (20).

The proof of this theorem can be found in [49].

3.2 Banach Spaces as Semi-inner Product Spaces

With the above preliminary definitions and results we are now able to characterize Banach spaces more precisely reflecting semi-inner product properties for these spaces. In particular, we are able to determine kernel functions and respective feature maps for so-called reproducing kernel Banach space (RKBS) comparable to those known from RKHS. Again we follow the argumentation by Zhang et al. [49].

The first statement is that an uniformly convex Banach space $\mathcal{B}$ is reflexive, i.e. $\mathcal{B} = (\mathcal{B}^*)^*$. Together with the property of differentiability a Riesz-representation-theorem can be stated [11]:

Theorem 3.5 *Let $\mathcal{B}$ be uniformly convex and uniformly Fréchet-differentiable Banach space with its dual $\mathcal{B}^*$. Then there exists for each $f \in \mathcal{B}^*$ an unique $g \in \mathcal{B}$ such that $f^* = g$ with*

$$f(h) = [h, g]_{\mathcal{B}} \tag{22}$$

for all $h \in \mathcal{B}$. Moreover, $\|f\|_{\mathcal{B}^*} = \|g\|_{\mathcal{B}}$

By means of the last theorem a norm preserving bijection $f \to f^*$ is established between uniformly convex, uniformly Fréchet-differentiable Banach space $\mathcal{B}$ and its dual $\mathcal{B}^*$. The object f^* is called the dual element of f. This duality mapping, however, is in general non-linear. Otherwise, because $\mathcal{B}^*$ is uniformly Fréchet-differentiable, by means of Lemma 3.2 it is equipped with an unique semi-inner product

$$[h^*, g^*]_{\mathcal{B}^*} = [h, g]_{\mathcal{B}} \tag{23}$$

induced by that of the Banach space $\mathcal{B}$.

Definition 3.6 *A reflexive Banach space $\mathcal{B}$ of functions on a vector pace V for which $\mathcal{B}^*$ is isometric to a Banach space $\mathcal{B}^{\#}$ of functions on V and the point evaluation is continuous on both $\mathcal{B}^*$ and $\mathcal{B}^{\#}$ is denotes as a reproducing kernel Banach space (RKBS) on V.*

For RKBS now it is possible to identify a *reproducing kernel* [49]:

Theorem 3.7 *Suppose $\mathcal{B}$ to be a RKBS on V. Let further be $(\cdot, \cdot)_{\mathcal{B}}$ be a bi-linear form on $(\mathcal{B} \times \mathcal{B}^*)$. Then there exists an unique function $\kappa : V \times V \longrightarrow \mathbb{C}$ satisfying the following conditions:*

1. *for each* $\mathbf{v} \in V$ *the function* $\kappa(\cdot, \mathbf{v}) \in \mathcal{B}^*$ *and*

$$f(\mathbf{v}) = (f, \kappa(\cdot, \mathbf{v}))_{\mathcal{B}} \text{ for all } f \in \mathcal{B} \tag{24}$$

2. *for each* $\mathbf{v} \in V$ *the function* $\kappa(\mathbf{v}, \cdot) \in \mathcal{B}$ *and*

$$f^*(\mathbf{v}) = (\kappa(\mathbf{v}, \cdot), f^*)_{\mathcal{B}} \text{ for all } f^* \in \mathcal{B}^* \tag{25}$$

3. *the linear span of* $\mathcal{B}_\kappa = \{\kappa(\mathbf{v}, \cdot) : \mathbf{v} \in V\}$ *is dense in* $\mathcal{B}$*, i.e.*

$$\overline{\mathrm{span}\mathcal{B}_\kappa} = \mathcal{B} \tag{26}$$

4. *the linear span of* $\mathcal{B}^*_\kappa = \{\kappa(\cdot, \mathbf{v}) : \mathbf{v} \in V\}$ *is dense in* $\mathcal{B}^*$*, i.e.*

$$\overline{\mathrm{span}\mathcal{B}^*_\kappa} = \mathcal{B}^* \tag{27}$$

5. *for all* $\mathbf{u}, \mathbf{v} \in V$

$$\kappa(\mathbf{u}, \mathbf{v}) = (\kappa(\mathbf{u}, \cdot), \kappa(\cdot, \mathbf{v}))_{\mathcal{B}} \tag{28}$$

The introduced function κ is called *the reproducing kernel* for the RKBS $\mathcal{B}$. It is unique for a given RKBS but it turns out that several RKBS may have the same reproducing kernel. By the following theorem it is possible to generate a reproducing kernel and their corresponding RKBS using the concept of feature maps:

Theorem 3.8 *Let* $\mathcal{W}$ *be a reflexive Banach space on* V *with its dual space* $\mathcal{W}^*$*. Assume that there exist feature mappings* $\Phi : V \longrightarrow \mathcal{W}$ *and* $\Phi^* : V \longrightarrow \mathcal{W}^*$ *such that*

$$\overline{span\Phi(V)} = \mathcal{W} \text{ and } \overline{span\Phi^*(V)} = \mathcal{W}^* \tag{29}$$

holds. Let further be $(\cdot, \cdot)_{\mathcal{W}}$ *be a bi-linear form on* $(\mathcal{W} \times \mathcal{W}^*)$*. Then* $\mathcal{B} = \{(\mathbf{w}, \Phi^*(\cdot))_{\mathcal{W}} | \mathbf{w} \in \mathcal{W}\}$ *with the norm*

$$\| (\mathbf{w}, \Phi^*(\cdot))_{\mathcal{W}} \|_{\mathcal{B}} = \| \mathbf{w} \|_{\mathcal{W}} \tag{30}$$

is a RKBS on V *with the dual space* $\mathcal{B}^* = \{(\Phi(\cdot), \mathbf{w}^*)_{\mathcal{W}} | \mathbf{w}^* \in \mathcal{W}^*\}$ *equipped with the norm*

$$\| (\Phi(\cdot), \mathbf{w}^*)_{\mathcal{W}} \|_{\mathcal{B}^*} = \| \mathbf{w}^* \|_{\mathcal{W}^*} \tag{31}$$

and the bi-linear form

$$((\mathbf{w}, \Phi^*(\cdot))_{\mathcal{W}}, (\Phi(\cdot), \mathbf{w}^*)_{\mathcal{W}})_{\mathcal{B}} = (\mathbf{w}, \mathbf{w}^*)_{\mathcal{W}} \tag{32}$$

with $\mathbf{w} \in \mathcal{W}$ *and* $\mathbf{w}^* \in \mathcal{W}^*$. *Further, for the unique reproducing kernel* $\kappa_\Phi : V \times V \longrightarrow \mathbb{C}$ *on* $\mathcal{B}$ *corresponding to the feature map* Φ *the relation*

$$\kappa_\Phi(\mathbf{u}, \mathbf{v}) = (\Phi(\mathbf{u}), \Phi^*(\mathbf{v}))_{\mathcal{W}} \tag{33}$$

is valid for $\mathbf{u}, \mathbf{v} \in V$.

Remark 3.9 *It turns out that for a reflexive Banach space* $\mathcal{W}$ *on* V *and a function* $\kappa_\Phi : V \times V \longrightarrow \mathbb{C}$ *it is necessary and sufficient to be a reproducing kernel that* κ_Φ *is of the form (33) and the mappings* $\Phi : V \longrightarrow \mathcal{W}$ *and* $\Phi^* : V \longrightarrow \mathcal{W}^*$ *satisfy (29), see [49].*

In the next step we relate RKBS to semi-inner product spaces. We denote a uniformly convex, uniformly Fréchet-differentiable RKBS $\mathcal{B}$ on a vector space V a *s.i.p. reproducing kernel Banach space* (s.i.p. RKBS). As a consequence of Lemma 3.2, we immediately have that a RKHS is a s.i.p. RKBS. Obviously, also the dual $\mathcal{B}^*$ of a s.i.p. RKBS $\mathcal{B}$ is a s.i.p. RKBS itself. Hence, the unique s.i.p. $[\cdot,\cdot]_{\mathcal{B}^*}$ characterizes the relation between the s.i.p. RKBS $\mathcal{B}$ and its dual $\mathcal{B}^*$ according to the Riesz-Theorem 3.5. This observation leads to the following more specific representer theorem presented by Zhang et al. in [49]:

Theorem 3.10 *Let* $\mathcal{B}$ *be a s.i.p RKBS on a vector space* V *and* κ_Φ *its reproducing kernel determined by the feature map* $\Phi : V \longrightarrow \mathcal{B}$. *Then there exist an unique function* $\gamma : V \times V \longrightarrow \mathbb{C}$ *such that* $\{\gamma(\mathbf{v}, \cdot) : \mathbf{v} \in V\} \subseteq \mathcal{B}$ *and*

$$f(\mathbf{u}) = [f, \gamma(\mathbf{u}, \cdot)]_{\mathcal{B}} \tag{34}$$

for all $f \in \mathcal{B}$ *and* $\mathbf{u} \in V$. *The function* γ *is denoted as s.i.p. kernel, which is related to the reproducing kernel by*

$$\kappa_\Phi(\cdot, \mathbf{v}) = (\gamma(\mathbf{v}, \cdot))^* \tag{35}$$

and

$$f^*(\mathbf{v}) = [\kappa_\Phi(\mathbf{v}, \cdot), f]_{\mathcal{B}} \tag{36}$$

for all $f \in \mathcal{B}$ *and* $\mathbf{v} \in V$.

For RKHS the s.i.p. kernel is identical with the reproducing kernel. In general, if for a s.i.p. RKBS $\kappa_\Phi \equiv \gamma$ holds, we call it a *s.i.p. reproducing kernel* denoted by γ_Φ to keep in mind its connection to the feature map Φ. From (34) it becomes clear that

$$\gamma_\Phi(\mathbf{u},\mathbf{v}) = [\gamma_\Phi(\mathbf{u},\cdot), \gamma_\Phi(\mathbf{v},\cdot)]_{\mathcal{B}} \tag{37}$$

which shows the formal equivalence to reproducing kernels for RKHS.

Analogously to the Theorem 3.8 and the Remark 3.9 the following theorem is valid for s.i.p. reproducing kernels:

Theorem 3.11 *Let $\mathcal{W}$ be an uniformly convex, uniformly Fréchet-differentiable Banach space on V and Φ a map $\Phi : V \longrightarrow \mathcal{W}$ such that*

$$\overline{span\Phi(V)} = \mathcal{W} \text{ and } \overline{span\Phi^*(V)} = \mathcal{W}^* \tag{38}$$

holds. Then $\mathcal{B} = \{[\mathbf{w}, \Phi(\cdot)]_{\mathcal{W}} | \mathbf{w} \in \mathcal{W}\}$ with the semi-inner product

$$[[\mathbf{w}, \Phi(\cdot)]_{\mathcal{W}}, [\mathbf{z}, \Phi(\cdot)]_{\mathcal{W}}]_{\mathcal{B}} := [\mathbf{w}, \mathbf{z}]_{\mathcal{W}} \tag{39}$$

and $\mathcal{B}^* = \{[\Phi(\cdot), \mathbf{w}]_{\mathcal{W}} | \mathbf{w} \in \mathcal{W}\}$ equipped with the semi-inner product

$$[[\Phi(\cdot), \mathbf{w}]_{\mathcal{W}}, [\Phi(\cdot), \mathbf{z}]_{\mathcal{W}}]_{\mathcal{B}^*} := [\mathbf{z}, \mathbf{w}]_{\mathcal{W}} \tag{40}$$

are uniformly convex and uniformly Fréchet-differentiable Banach spaces. The space $\mathcal{B}^*$ is the dual of $\mathcal{B}$ with the bi-linear form

$$([\mathbf{w}, \Phi(\cdot)]_{\mathcal{W}}, [\Phi(\cdot), \mathbf{z}]_{\mathcal{W}})_{\mathcal{B}} := [\mathbf{z}, \mathbf{w}]_{\mathcal{W}} \text{ for } \mathbf{z}, \mathbf{w} \in \mathcal{W}. \tag{41}$$

Further, the unique s.i.p. reproducing kernel $\gamma_\Phi : V \times V \longrightarrow \mathbb{C}$ of $\mathcal{B}$ is given by

$$\gamma_\Phi(\mathbf{u},\mathbf{v}) = [\Phi(\mathbf{u}), \Phi(\mathbf{v})]_{\mathcal{W}} \tag{42}$$

with $\mathbf{u},\mathbf{v} \in V$, i.e. the s.i.p. reproducing kernel coincides with the reproducing kernel κ_Φ under this conditions, which legitimates the notation γ_Φ instead of simple γ.

Remark 3.12 *We observe that*

$$||\mathbf{v}||_{\mathcal{W}} = \sqrt{\gamma_\Phi(\mathbf{v},\mathbf{v})} \tag{43}$$

defines a metric $d_{\mathcal{W}}$ according to the Theorem 3.3.

Again, we can state the following characterization:

Remark 3.13 *A function γ_Φ on $V \times V$ is a s.i.p. reproducing kernel iff it is of the form (42) with the feature map $\Phi : V \longrightarrow \mathcal{W}$ from a vector space V to a uniformly convex, uniformly Fréchet-differentiable Banach $\mathcal{W}$ satisfying (38), see [49]. The s=\left\{ \mathbf{}\right\} pace $\mathcal{W}$ is also denoted as the feature space.*

It follows from the duality relationship (35) and the density condition (27) that for a s.i.p. kernel γ of a s.i.p. RKBS $\mathcal{B}$ on V the equivalence

$$\overline{span}\left\{(\gamma_\Phi(\mathbf{v},\cdot))^* : \mathbf{v} \in V\right\} = \mathcal{B}^* \tag{44}$$

is valid. According to the above Remark 3.13 the relation $\kappa_\Phi \equiv \gamma_\Phi$ between the reproducing and the s.i.p. reproducing kernel only holds iff

$$\overline{span}\left\{\gamma_\Phi(\mathbf{v},\cdot) : \mathbf{v} \in V\right\} = \mathcal{B} \tag{45}$$

and the duality mapping from $\mathcal{B}$ to $\mathcal{B}^*$ become non-linear if $\mathcal{B}$ is not a Hilbert space, i.e. (44) does not always implies (45).

3.3 Some Properties of S.i.p. Reproducing Kernels and their S.i.p. RKBS

In this section we will consider some properties of s.i.p. reproducing kernels, which are interesting in the context of machine learning.

Let $\gamma_\Phi : V \times V \longrightarrow \mathbb{C}$ be a s.i.p. reproducing kernel such that (38) and (42) are satisfied. From the Definition 3.1 properties 3 and 4 it follows that $\gamma_\Phi(\mathbf{v},\mathbf{v}) \geq 0$ for all $\mathbf{v} \in V$ and the *s.i.p.-Cauchy-Schwarz-inequality*

$$\gamma_\Phi(\mathbf{u},\mathbf{v}) \leq |\gamma_\Phi(\mathbf{u},\mathbf{u})| \cdot |\gamma_\Phi(\mathbf{v},\mathbf{v})| \tag{46}$$

for all $\mathbf{u}, \mathbf{v} \in V$ is still valid. However, we can not generally assume a complex s.i.p. kernel to be positive definite. For an example we refer to [49] and Example 3.18.

Let the sequence of $f_n \in \mathcal{B}$ converge to f in s.i.p. RKBS $\mathcal{B}$ over the vector space V with the s.i.p. kernel γ_Φ. As a consequence of the s.i.p.-Cauchy-Schwarz-inequality (46) for all $\mathbf{v} \in V$ the limes

$$f_n(\mathbf{v}) \longrightarrow f(\mathbf{v})$$

is valid and the limit is uniform if $\gamma_\Phi(\mathbf{u},\mathbf{v})$ is bounded. This property is called *point-wise convergence.*

In analogy to the universality of kernels for RKHS, we now characterize the concept universality for s.i.p. kernels.

Definition 3.14 *Suppose, (V, d) is a compact metric space and $\gamma_\Phi : V \times V \longrightarrow \mathbb{C}$ is a s.i.p. reproducing kernel on V. The s.i.p. kernel γ_Φ is called weakly universal if it is continuous and bounded and the space of induced functions*

$$\mathcal{I}_{\gamma_\Phi} = \{\gamma_\Phi(\mathbf{v}, \cdot) | \mathbf{v} \in V\} \tag{47}$$

is dense in $\mathcal{C}(V)$.

We can state the following proposition [49]:

Proposition 3.15 *Let (V, d) be a compact metric space and Φ be a feature map from V to a Banach space $\mathcal{W}$ such that both $\Phi : V \longrightarrow \mathcal{W}$ and $\Phi^* : V \longrightarrow \mathcal{W}^*$ are continuous. Then the s.i.p. kernel $\gamma_\Phi : V \times V \longrightarrow \mathbb{C}$ defined by (42) is continuous and, there holds in $\mathcal{C}(V)$ the equality*

$$\overline{span}\left(\mathcal{I}_{\gamma_\Phi}\right) = \overline{span}\left\{[\mathbf{w}, \Phi(\cdot)]_{\mathcal{W}} | \mathbf{w} \in \mathcal{W}\right\}.$$

Consequently, the s.i.p. kernel γ_Φ is weakly universal iff

$$\overline{span}\left\{[\mathbf{w}, \Phi(\cdot)]_{\mathcal{W}} | \mathbf{w} \in \mathcal{W}\right\} = \mathcal{C}(V).$$

Remark 3.16 *Obviously, for weakly universal kernels the metric d_{γ_Φ} from (4) coincides with $d_{\mathcal{W}}$ defined in Remark 3.12 on $\mathcal{I}_{\kappa_\Phi}$.*

In conclusion we explicitly state the following lemma which is the complement of the Lemma 2.7 for RKHS:

Lemma 3.17 *Let (V, d_V) be a compact metric space, $\gamma_\Phi : V \times V \to \mathbb{R}$ a continuous weakly universal s.i.p. kernel with the feature map $\Phi : V \longrightarrow \mathcal{B}$ and $\mathcal{B}$ being an uniformly convex, uniformly Fréchet-differentiable Banach space. Let $d_{\mathcal{B}}$ be the metric determined by the norm via (42) induced by the kernel γ_Φ. If the space of induced functions $\mathcal{I}_{\gamma_\Phi}$ defined in (47) is dense in the space of continuous functions $\mathcal{C}(V)$, then the metric space $(V, d_{\mathcal{B}})$ is topologically equivalent to induced space $\mathcal{I}_{\gamma_\Phi}$ with the metric $d_{\mathcal{B}}$. Moreover, both spaces are isometric.*

Proof. The kernel γ_Φ is assumed to be continuous and weakly uniform. Hence, the space of induced functions $\mathcal{I}_{\gamma_\Phi}$ is dense in the space of continuous functions $\mathcal{C}(V)$ with the metric $d_{\mathcal{B}}$ determined by the norm (43). According to the Remark 2.5 we can apply the Steinwart-Theorem 2.4 for universal kernels in RKHS although we have only the weak universality of a s.i.p. kernel. Hence, the uniquely

$$\begin{tikzcd}
(V, d_V) \arrow[r, "{\Phi, \gamma_\Phi}"] \arrow[d, "\Psi"'] & \mathcal{I}_{\gamma_\Phi} \subseteq (\mathcal{B}, d_\mathcal{B}) \\
(V, d_\mathcal{B}) \arrow[ur, leftrightarrow, "\Phi \circ \Psi^{-1}"'] &
\end{tikzcd}$$

Figure 2: Visualization of the statement of Lemma 2.7: For s.i.p.-universal kernels γ_Φ the metric spaces $(V, d_\mathcal{B})$ and $(\mathcal{I}_{\gamma_\Phi}, d_\mathcal{B})$ are topologically equivalent and isometric by means of the continuous bijective mapping $\Phi \circ \Psi^{-1}$.

corresponding feature map $\Phi : V \longrightarrow \mathcal{B}$ is injective and, together with the continuity ensured by the Remark 2.2, it is bijective for $\Phi : V \longrightarrow \mathcal{I}_{\gamma_\Phi} \subseteq \mathcal{B}$, whereby $\mathcal{B}$ is equipped with the Banach space metric $d_\mathcal{B}$. Moreover, it follows from Lemma 2.1 again together with the Remark 2.2 that the identity map

$$\Psi : (V, d_V) \longrightarrow (V, d_\mathcal{B})$$

is also continuous and, therefore, bijective. Hence, the map $\Phi\left(\Psi^{-1}(\mathbf{v})\right) = \Phi \circ \Psi^{-1}(\mathbf{v})$ with $\mathbf{v} \in (V, d_\mathcal{B})$ is bijective and continuous. Therefore, $(V, d_\mathcal{B})$ and $\mathcal{I}_{\gamma_\Phi}$ are isomorphic and, according to Remark 3.16, also isometric. ■

The result of the Lemma 3.17 is visualized in Fig.3.3

We now give examples of universal s.i.p. kernels [49]:

Example 3.18 *We assume that $V \subseteq \mathbb{R}$.*

1. *Let be $V = \mathbb{R}$, then $\gamma_\Phi = \exp(-|u - v|)$ is an universal s.i.p. kernel.*

2. *Let be $V = (0, 1)$, then $\gamma_\Phi = \exp(-|u - v|)$ is an universal s.i.p. kernel.*

3. *Let be $V = [0, \infty)$, $1 < p < \infty$ and $\mathcal{B} = l_p(\mathbb{N}_2)$. Further, suppose $\Phi(x) = (1, v) : V \longrightarrow \mathcal{B}$. Then*

$$\Phi^*(v) = \frac{(1, v^{p-1})}{(1 + v^p)^{\frac{p-2}{p}}},$$

holds and

$$\gamma_{\Phi}\left(u,v\right)=\frac{\left(1+u\cdot v^{p-1}\right)}{\left(1+v^{p}\right)^{\frac{p-2}{p}}}$$

is an universal s.i.p. kernel. However, it is only positive definite iff $p=2$.

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 $W=\left\{\mathbf{w}_{k}\right\}_{k\in A}$, where A is a finite index set. Prominent examples are the self-organizing map (SOM,[19]), neural gas (NG, [25]), whereby for the SOM the variant of Heskes is taken [15]. For those models, the reconstruction error is given in terms of the dissimilarity measure $d\left(\mathbf{v},\mathbf{w}_{k}\right)$ between data and prototypes, which is assumed to be differentiable. In that case, the gradient $\partial E/\partial\mathbf{w}_{k}$ contains the derivative $\partial d\left(\mathbf{v},\mathbf{w}_{k}\right)/\partial\mathbf{w}_{k}$ originating from the chain rule of differentiation.

Prototype based classification in the context of learning vector quantization models (LVQ, [19]) was renewed by the idea of Sato&Yamada to approximate the non-differentiable classification error C by a differentiable function E_{C} [40, 39]. As in unsupervised vector quantization, E_{C} depends on the underlying dissimilarity measure $d\left(\mathbf{v},\mathbf{w}_{k}\right)$. Hence, gradient based classification learning also requires the term $\partial d\left(\mathbf{v},\mathbf{w}_{k}\right)/\partial\mathbf{w}_{k}$.

For example, the cost function of the unsupervised self-organizing maps (SOM) for vector quantization in the variant of T. Heskes is

$$E_{\text{SOM}}=\int P(\mathbf{v})\sum_{\mathbf{r}\in A}\delta_{\mathbf{r}}^{s(\mathbf{v})}\sum_{\mathbf{r}'\in A}\frac{h_{\sigma}^{SOM}(\mathbf{r},\mathbf{r}')}{2K(\sigma)}d(\mathbf{v},\mathbf{w}_{\mathbf{r}'})d\mathbf{v}\qquad(48)$$

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 [15]. $K(\sigma)$ is a normalization constant depending on the neighborhood range σ. Then the stochastic gradient prototype update for all prototypes is given as [15]:

$$\triangle \mathbf{w_r} = -\varepsilon h_{\sigma}^{SOM}\left(\mathbf{r}, s(\mathbf{v})\right) \frac{\partial d\left(\mathbf{v}, \mathbf{w_r}\right)}{\partial \mathbf{w_r}}. \tag{49}$$

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

Analogously the widely used *supervised* generalized learning vector quantization scheme (GLVQ) with the cost function

$$E(W) = \frac{1}{2}\sum_{\mathbf{v}\in V} f(\mu(\mathbf{v})) \text{ with } \mu(\mathbf{v}) = \frac{d^{+}(\mathbf{v}) - d^{-}(\mathbf{v})}{d^{+}(\mathbf{v}) + d^{-}(\mathbf{v})} \tag{50}$$

can be treated: the related gradient learning is based on the (stochastic) derivatives

$$\frac{\partial_s E}{\partial \mathbf{w}^{+}} = \frac{\partial_s E}{\partial d^{+}}\frac{\partial d^{+}}{\partial \mathbf{w}^{+}}, \quad \frac{\partial_s E}{\partial \mathbf{w}^{-}} = \frac{\partial_s E}{\partial d^{-}}\frac{\partial d^{-}}{\partial \mathbf{w}^{-}} \tag{51}$$

with $\frac{\partial_s}{\partial}$ and

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

where $\mu(\mathbf{v})$ is the classifier function with $d^{+}(\mathbf{v}) = d(\mathbf{v}, \mathbf{w}^{+})$ denotes the distance between the data point $\mathbf{v}$ and the nearest prototype $\mathbf{w}^{+}$, belonging to the same class as the presented data point. In the second equation the abbreviation d^{+} for $d^{+}(\mathbf{v})$ is used for simplicity. Again as in SOMs, $d(\mathbf{v}, \mathbf{w})$ in (50) is some differentiable dissimilarity measure with respect to $\mathbf{w}$. Hence, it could be replaced in (51) by a differentiable kernel metric. Analogously d^{-} is defined as the distance to the best prototype of all other classes.

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 $(V, d\left(\mathbf{v}, \mathbf{w}_k\right))$. Obviously, under gentle conditions on V (continuous, local convex, ...) it can be assumed that $\partial d\left(\mathbf{v}, \mathbf{w}_k\right)/\partial \mathbf{w}_k \in V$ is valid. Yet, the choice of the metric is free except the necessary differentiability. Hence, metrics determined by differentiable kernel are applicable. Obviously, the kernels presented in Example 2.8 as well as the information theroetic kernels in Lemma 2.10 are differentiable (for the latter kernels, see [48] for differentiability of the respective divergences). If such a metric is obtained from an universal kernel κ_Φ or γ_Φ for RKHS and RKBS, respectively, the Lemmata 2.7 and 3.17 ensure the topological and isometric equivalence to the respective Hilbert or Banach space. 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. More properties of differentiable Mercer-like kernels and their reproducing properties can be found in [10].

5 Conclusion

In this paper we considered the theoretical framework for applying differentiable kernels 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 or Banach space. Hence, gradient based vector quantization schemes with differentiable universal kernels can benefit from this property. The main results of topological and isometric equivalence are the Lemmata 2.7 and 3.17. Last but not least we provide some examples of differentiable universal kernels based on divergences as fundamental information theoretic concepts.

An important future task, which is just in progress, is the transfer of these ideas to non-Euclidean online principal component learning according to E. OJA's learning algorithms, which are based on the Euclidean inner product but could be replaced by a kernel [30],[31].

Acknowledgment: The authors would like to thank M. KÄSTNER for helpful discussions.

References

[1] N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68:337–404, 1950.

[2] A. Ben-Hamza and H. Krim. Jensen-Rényi divergence measure: theoretical and computational perspectives. In *Proceedings of the IEEE International Symposium on Information Theory*, pages 257–257, 2003.

[3] J. Bezdek. A convergence theorem for the fuzyy ISODATA clustering algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2(1):1–8, 1980.

[4] J. Bezdek. *Pattern Recognition with Fuzzy Objective Function Algorithms.* Plenum, New York, 1981.

[5] A. Chan, N. Vasconcelos, and P. Moreno. 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.

[6] A. Cichocki and S. C. S.-I. Amari. Generalized alpha-beta divergences and their application to robust nonnegative matrix factorization. *Entropy*, 13:134–170, 2011.

[7] A. Cichocki and S.-I. Amari. Families of alpha- beta- and gamma- divergences: Flexible and robust measures of similarities. *Entropy*, 12:1532–1568, 2010.

[8] K. Crammer, R. Gilad-Bachrach, A.Navot, and A.Tishby. Margin analysis of the LVQ algorithm. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing (Proc. NIPS 2002)*, volume 15, pages 462–469, Cambridge, MA, 2003. MIT Press.

[9] D. Cudia. On the localization and directionalization of uniform convexity. *Bulletin of the American Mathematical Society*, 69:265–267, 1963.

[10] J. Ferreira and V. Menegatto. Reproducing properties of differentiable Mercer-like kernels. *Mathematische Nachrichten*, 285:in press, 2012.

[11] J. Giles. Classes of semi-inner-product spaces. *Transactions of the American Mathematical Society*, 129:436–446, 1967.

[12] A. Gretton, R. Herbrich, A. Smola, O. Bousquet, and B. Schölkopf. Kernel methods for measuring independence. *Journal of Machine Learning Research*, 6:2075–2129, 2005.

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

[14] M. Hein and O. Bousquet. Hilbertian metrics and positive definite kernels on probability measures. Technical report, Max Planck Institute for Biological Cybernetics, 2004.

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

[16] T. Hoffmann, B. Schölkopf, and A. Smola. Kernel methods in machine learning. *The Annals of Statistics*, 36(3):1171–1220, 2008.

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

[18] M. Kästner, B. Hammer, M. Biehl, and T. Villmann. Functional relevance learning in generalized learning vector quantization. *Neurocomputing*, 75(9):in press, 2012.

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

[20] A. Kolmogorov and S. Fomin. *Reelle Funktionen und Funktionalanalysis.* VEB Deutscher Verlag der Wissenschaften, Berlin, 1975.

[21] B. Kulis, M. Sustik, and I. Dhillon. Low-rank kernel learning with Bregman matrix divergences. *Journal of Machine Learning Research*, 10:341–376, 2009.

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

[23] G. Lumer. Seni-inner-product spaces. *Transactions of the American Mathematical Society*, 100:29–43, 1961.

[24] A. Martin, N. Smith, E. Xing, P. Aguiar, and M. Figueiredo. Nonextensive information theoretic kernels on measures. *Journal of Machine Learning Research*, 10:935–975, 2009.

[25] 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.

[26] J. Mercer. 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.

[27] C. Micchelli, Y. Xu, and H. Zhang. Universal kernels. *Journal of Machine Learning Research*, 7:26051–2667, 2006.

[28] E. Mwebaze, P. Schneider, F.-M. Schleif, J. Aduwo, J. Quinn, S. Haase, T. Villmann, and M. Biehl. Divergence based classification in learning vector quantization. *Neurocomputing*, 74(9):1429–1435, 2011.

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

[30] E. Oja. Neural networks, principle components and suspaces. *International Journal of Neural Systems*, 1:61–68, 1989.

[31] E. Oja. Nonlinear pca: Algorithms and applications. In *Proc. Of the World Congress on Neural Networks Portland*, pages 396–400, Portland, 1993.

[32] F. Österreicher and I. Vajda. 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.

[33] E. Pekalska and R. Duin. *The Dissimilarity Representation for Pattern Recognition: Foundations and Applications.* World Scientific, 2006.

[34] J. Principe. *Information Theoretic Learning.* Springer, Heidelberg, 2010.

[35] A. Qin and P. Suganthan. A novel kernel prototype-based learning algorithm. In *Proceedings of the 17th International Conference on Pattern Recognition (ICPR'04)*, volume 4, pages 621–624, 2004.

[36] A. K. Qin and P. N. Suganthan. Kernel neural gas algorithms with application to cluster analysis. In *ICPR (4)*, pages 617–620, 2004.

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

[38] A. Rényi. *Probability Theory.* North-Holland Publishing Company, Amsterdam, 1970.

[39] 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.

[40] A. S. Sato and K. Yamada. Generalized learning vector quantization. In G. Tesauro, D. Touretzky, and T. Leen, editors, *Advances in Neural Information Processing Systems*, volume 7, pages 423–429. MIT Press, 1995.

[41] F.-M. Schleif, T. Villmann, B. Hammer, and P. Schneider. Efficient kernelized prototype based classification. *International Journal of Neural Systems*, 21(6):443–457, 2011.

[42] B. Schölkopf and A. Smola. *Learning with Kernels.* MIT Press, 2002.

[43] P. Schneider, B. Hammer, and M. Biehl. Adaptive relevance matrices in learning vector quantization. *Neural Computation*, 21:3532–3561, 2009.

[44] C. Scovel, D. Hush, I. Steinwart, and J. Theiler. Radial kernels and their reproducing kernel Hilbert spaces. *Journal of Complexity*, 26:641–660, 2010.

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

[46] B. Sriperumbudur, K. Fukumizu, and G. Lanckriet. Universality, characteristic kernels, and RKHS embedding of measures. *Journal of Machine Learning Research*, 12:2389–2410, 2011.

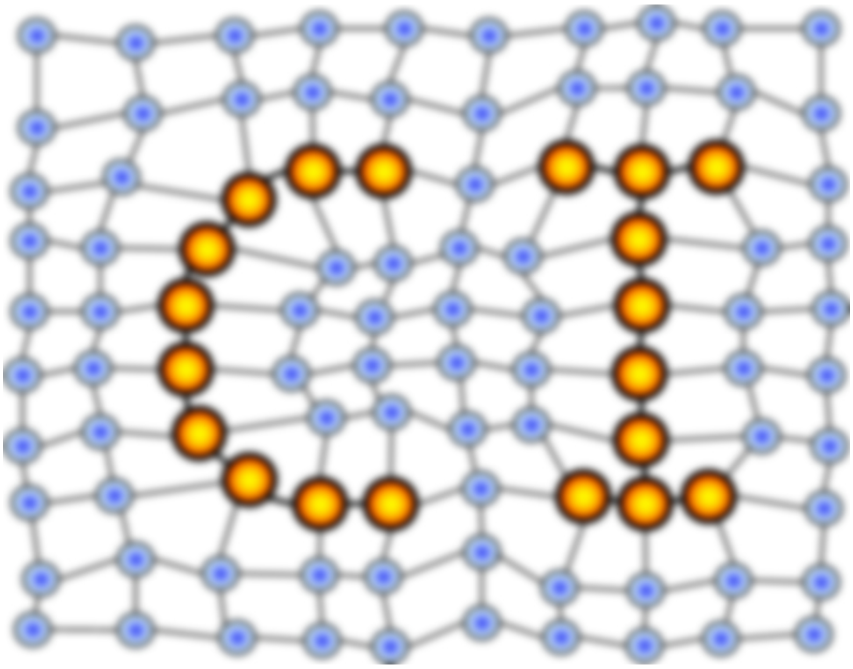
[47] I. Steinwart. On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Learning Research*, 2:67–93, 2001.

[48] T. Villmann and S. Haase. Divergence based vector quantization. *Neural Computation*, 23(5):1343–1392, 2011.

[49] H. Zhang, Y. Xu, and J. Zhang. Reproducing kernel banach spaces for machine learning. *Journal of Machine Learning Research*, 10:2741–2775, 2009.

MACHINE LEARNING REPORTS

Report 02/2012



Impressum

Machine Learning Reports ISSN: 1865-3960

▽ Publisher/Editors

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

Dr. rer. nat. Frank-Michael Schleif
University of Bielefeld
Universitätsstrasse 21-23, 33615 Bielefeld, Germany
• http://www.cit-ec.de/tcs/about

▽ Acknowledgments

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