About Approximation Completeness of Generalized Multilayer Perceptrons Consisting of Banach-like Perceptrons Based on Semi-Inner Products

T. Villmann[1], A. Engelsberger[1], J. Ravichandran[1], A. Villmann[2]

[1]University of Applied Sciences Mittweida – Saxon Institute for Computational Intelligence and Machine Learning (SICIM) and
[2]Berufliches Schulzentrum Döbeln-Mittweida

***Abstract*—The paper reconsiders multilayer perceptron networks for the case where the Euclidean inner product is replaced by a semi-inner product. This would be of interest, if the dissimilarity measure between data is given by a general norm such that the Euclidean inner product is not longer consistent to that situation. We prove mathematically that the universal approximation completeness is guaranteed also for those networks where the used semi-inner products are related either to uniformly convex or to reflexive Banach-spaces. Most famous examples of uniformly convex Banach spaces are the spaces L_p and l_p for $1 < p < \infty$. The result is valid for all discriminatory activation functions including the sigmoid and the $ReLU$ activation.**

I. Introduction and Motivation

Various types of multilayer perceptrons (MLP) including deep networks belong nowadays certainly to the standard neural networks in machine learning for classification and regression tasks [1], [8]. Biologically motivated by pyramid cells in brains the corresponding mathematical perceptron is the basis of those networks [24], see Fig. 1.

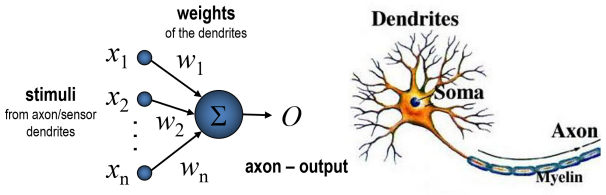

Figure 1. Schematic illustration of a mathematical perceptron (left) according to a pyramid cell (right). The input vector $\mathbf{x} = (x_1, \ldots, x_n)^T$ is weighted by the weight vector $\mathbf{w} = (w_1, \ldots, w_n)^T$ to generate the output O.

The capability for these networks is justified by Cybenko's theorem with states the universal approximation capability for MLP's with sigmoidal activation functions [5]. One key ingredient in the proof of the respective theorem is the Hilbert-space-property needed to ensure the application of the Riesz-Representation-Theorem (RRT). This property is given for each perceptron in the network, because perceptrons generate their output based on the Euclidean inner product (EIP) between the input and the weight vector. Thus the data space is implicitly assumed to be a Hilbert space equipped with the Euclidean norm, which is generated by the standard inner product. However, depending on the task, other than the Euclidean metric might be more appropriate, e.g. l_p-norms (metrics) with $p \neq 2$ [18] or kernel metrics [28]. However, those metrics relate to so-called semi-inner products (SIP, [20]) which show weaker requirements than inner products. Hence, a consistent approach for a perceptron network should make use of SIPs instead of the EIP. Consequently the question arises whether those networks remain universal approximators. The paper tackles exactly this problem and will provide respective proofs.

The remainder of the paper is as follows: First we provide the basic mathematical concepts and definitions needed for the mathematical analysis of the problem. Thereafter, we recapitulate the proof of Cybenko's theorem regarding the approximation completeness to identify the keypoints of this proof in the light of the given problem. For this purpose, we analyze the class of discriminatory activation functions regarding the Euclidean inner product (or general inner products) and show that both sigmoidal and ReLU activation function belong to that class. In the next step we provide the results for SIP-based perceptrons, which we also denote as Banach-like-perceptrons (BlP). For this purpose, we show that the class of discriminatory functions with respect to a given SIP can be appropriately defined and, again sigmoidal and ReLU activation belong to this class. Further, we show which parts of the original Cybenko-theorem have to be modified. In particular, we identify those SIPs (and respective Banach-spaces), which can be equipped with an RRT compared to that valid for Hilbert spaces. The technical structure of this paper follows closely the mathematical description of MLP's given in [10].

II. The Standard Multilayer Perceptron Revisited

The mathematical modeling of standard perceptrons assumes stimulus vectors $\mathbf{x} \in \mathbb{R}^n$ and a weight vector $\mathbf{w} \in \mathbb{R}^n$ to generate the output according to

$$O(\mathbf{w}, \mathbf{x}) = f(\langle \mathbf{w}, \mathbf{x} \rangle + b) \tag{1}$$

where $b \in \mathbb{R}$ is the bias and f is the so-called activation function. The quantity $\langle \mathbf{w}, \mathbf{x} \rangle = \sum_{k=1}^{n} x_k \cdot w_k$ is the (real) Euclidean inner product, which is motivated biologically by the weighted sum of inputs, see Fig. 1. The activation function f usually is a monotonically increasing function. Common choices are the the identity $\mathrm{id}(z) = z$ (linear perceptron), the

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

Heaviside function $H(z)$ (standard perceptron) or the sigmoid function

$$f_\theta(z) = \frac{1}{1+\exp(\theta z)} \tag{2}$$

as smooth (differentiable) approximation of $H(z)$. Nowadays, other activation functions became popular, rather motivated computationally than biologically [22]. Among them, the function

$$\mathrm{ReLU}(z) = \max(0, z) \tag{3}$$

known as *Rectified Linear Unit* has gained great focus because of its easy computation and derivative [8].

MLPs are directed graphs with mathematical perceptrons as nodes organized in layers [13]. Only the first layer (input layer) receives direct data inputs. The last layer is denoted as output layer and delivers the network response $\mathbf{o}$ for a given data vector $\mathbf{x}$. The stimulus vectors of perceptrons in all layers except the input layer are output vectors of previous layers. Mathematically speaking, MLPs realize a mapping

$$F_{W,B} : \mathbb{R}^n \ni \mathbf{x} \longmapsto \mathbf{o} \in \mathbb{R}^m \tag{4}$$

if m output units are available and W is the set of all weights $\mathbf{w}$ and B is the set of all biases in the network. It was shown by CYBENKO that under certain conditions MLP's are universal approximators [5]. We will consider the proof of this theorem in detail after giving useful definitions and theorems from mathematical analysis needed for an adequate problem description in the proof of the Cybenko-theorem.

A. Basic Concepts, Mathematical Definitions and Theorems

Definition 1. The function σ is n-discriminatory with respect to the inner product $\langle\cdot,\cdot\rangle$ if for a measure $\mu \in \mathcal{M}(I_n)$ of the closed (compact) subset $I_n = [0,1]^n \subset \mathbb{R}^n$ with the property

$$\int_{I_n} \sigma(\langle \mathbf{w}, \mathbf{x}\rangle + b)\, d\mu(\mathbf{x}) = 0$$

for all $\mathbf{w} \in \mathbb{R}^n$ and $b \in \mathbb{R}$ the implication $\mu \equiv 0$ follows. A function is said to be discriminatory with respect to the inner product $\langle\cdot,\cdot\rangle$ if it is n-discriminatory for all n.

A function σ is denoted as *sigmoidal* if

$$\sigma(z) \longrightarrow \begin{cases} 1 & \text{for} \quad z \to \infty \\ 0 & \text{for} \quad z \to -\infty \end{cases}$$

holds. Obviously, $f_\theta(z)$ from (2) is sigmoidal. Another example is

$$\lambda(z) = \begin{cases} 0 & \text{if } z < 0 \\ z & \text{if } z \in [0,1] \\ 1 & \text{if } z > 1 \end{cases} \tag{5}$$

defining the set $\Lambda_{[0,1]}$ of *interval-restricted linear functions*. It can easily be shown that the respective span $\mathcal{S}(\Lambda_{[0,1]})$ is dense in $\mathcal{C}[0,1]$ [17].

The following Lemma, proven in [5], relates sigmoidal functions to discriminatory functions:

Lemma 2. *Any bounded, measurable sigmoidal function is discriminatory with respect to the real inner product $\langle\cdot,\cdot\rangle$ and, hence, any continuous sigmoidal function is discriminatory.*

It turns out that also the function $\mathrm{ReLU}(z)$ from (3) is discriminatory with respect to the inner product $\langle\cdot,\cdot\rangle$. In fact, we now prove the following lemma about the discriminatory property of the $ReLU$-activation with repsect to a real inner product:

Lemma 3. *The $\mathrm{ReLU}(z)$ from (3) is discriminatory with respect to the real inner product $\langle\cdot,\cdot\rangle$ for $z(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x}\rangle + b$.*

Proof: We follow [10] and start with the case $n = 1$ (1-discriminatory), i.e. $\langle w, x\rangle = w \cdot x$ and $z(x) = w \cdot x + b$ for given w and b. For $w = 0$ we can rewrite an arbitrary $\lambda(z) \in \mathcal{S}(\Lambda_{[0,1]})$ into

$$\lambda(b) = \begin{cases} \mathrm{ReLU}(\lambda(b)) & \text{if } \lambda(b) \geq 0 \\ -\mathrm{ReLU}(-\lambda(b)) & \text{if } \lambda(b) \leq 0 \end{cases}$$

whereas for $w \neq 0$ we decompose $\lambda(z(x))$ into

$$\lambda(x) = \mathrm{ReLU}\left(w \cdot x - \frac{b}{w}\right) - \mathrm{ReLU}\left(w \cdot x + \frac{1-b}{w}\right) \tag{6}$$

using the linearity of the (inner) product $w \cdot x$. Applying this decomposition we prove immediately the assertion: Because $\lambda(z(x))$ is discriminatory according to the previous lemma we have that for the integral $I[\lambda] = \int \lambda(w \cdot x - b)\, d\mu(x)$ the equality $I[\lambda] = 0$ holds, which further implies that $\mu \equiv 0$ has to be valid. Hence, we get for the decomposition (6)

$$\begin{aligned} I[\lambda] \quad &= \quad \int \mathrm{ReLU}\left(w \cdot x - \frac{b}{w}\right) d\mu(x) \\ &\quad - \int \mathrm{ReLU}\left(w \cdot x + \frac{1-b}{w}\right) d\mu(x) \\ &\overset{\mu \equiv 0}{=} \quad 0 - 0 \end{aligned}$$

which is the desired result.

For $n > 1$ we consider the span $\mathcal{S}(G)$ of the set $G = \{g(z(\mathbf{x})) \,|\text{nonlinear } g \in \mathcal{C}([0,1])\}$ of continuous functions depending on $\mathbf{x}$ with parameters $\mathbf{w}$ and b. Let $h(\mathbf{x}) \in \mathcal{S}(G)$, arbitrarily given. According to Kolmogorov's representation theorem [14], [2] and [9] exist affine functions $g_k(z_\varepsilon(\mathbf{x})) \in \mathcal{C}([0,1])$ with $z_\varepsilon(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x}\rangle + \frac{b}{N(\varepsilon)}$ such that

$$\left| h(\mathbf{x}) - g\left(\sum_{k=1}^{N(\varepsilon)} g_k(z_\varepsilon(\mathbf{x}))\right)\right| < \frac{\varepsilon}{2}$$

for arbitrarily chosen $\varepsilon > 0$ using the non-linearity of g. Because $z(\mathbf{x}) = \sum_{k,j} w_k x_j \langle \mathbf{e}_k, \mathbf{e}_j\rangle + b$ is an affine (linear) function in each variable x_j the introduced functions g_k are affine (linear) functions of x_j, i.e. we have $g_k(z_\varepsilon(\mathbf{x})) = \sum_{j=1}^n \hat{g}_k(z_\varepsilon(x_j))$ with $z_\varepsilon(x_j) = x_j \cdot w_j + b_j$. Each of the continuous functions $\hat{g}_k$ can be further approximated by

$$\left|\hat{g}_k(z_\varepsilon(x_j)) - \sum_{l=1}^{N_k(\varepsilon)} \lambda_{k,l}(z_\varepsilon(x_j))\right| < \frac{\varepsilon}{2 \cdot N(\varepsilon) \cdot n}$$

with $\lambda_{k,l} \in \mathcal{S}(\Lambda_{[0,1]})$ which can be taken as combinations of ReLU-functions according to (6).

In consequence, we are able approximate each $h(\mathbf{x}) \in \mathcal{S}(G)$ with arbitrary precision which implies the n-discriminatory property using the first part of the proof. This completes the proof of the lemma. ∎

Remark 4. We emphasize that for (6) the linearity of the inner product with respect to the first argument was used.

Definition 5. Let X be a vector space over $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$ and $\varphi : X \to \mathbb{K}$ be a functional. If both properties

- positive homogeneity: $\varphi(\lambda \mathbf{x}) = \lambda \varphi(\mathbf{x})$ for $\lambda \in \mathbb{R}_+$ and $\varphi(i\mathbf{x}) = i\varphi(\mathbf{x})$ is valid in the complex case
- subadditivity: $\varphi(\mathbf{x} + \mathbf{y}) \leq \varphi(\mathbf{x}) + \varphi(\mathbf{y})$

hold, φ is denoted as *sublinear*.

We remark that every norm on a vector space X is sublinear. A central role in this paper plays the *Hahn-Banach-Theorem* which states the following [15], [23]:

Theorem 6. ***(Hahn-Banach-Theorem)*** *Variant* ***a****): Let X be a vector space over $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$ and $Y \subseteq X$ a subspace. Let $\varphi : X \to \mathbb{R}$ be a sublinear functional and $f : Y \to \mathbb{K}$ be a linear functional with $\Re(f(\mathbf{y})) \leq \varphi(\mathbf{y})$ for all $\mathbf{y} \in Y$. Then there exists a linear functional $F : X \to \mathbb{K}$ with $F|_Y = f$ and $\Re(F(\mathbf{x})) \leq \varphi(\mathbf{x})$ is valid for all $\mathbf{x} \in X$.*

An alternative formulation is the variant [27], [25] ***b****): Let X be a normed space and Y is a subspace $Y \subset X$. Let be $f \in X^*$ with $f|_Y = 0$. The subspace Y is dense in X iff under these assumptions always follows $f(\mathbf{x}) = 0$ for all $\mathbf{x} \in X$.*

The following theorem is known as the *Theorem of Dominated Convergence from Lebesgue* [15], [23]:

Theorem 7. ***(Dominated-Convergence-Theorem)*** *Let X be a measure space, μ a Borel-measure on X and $g : X \longrightarrow \mathbb{R}$ absolute integrable, $g \in \mathcal{L}^1(X)$. Let further $\{f_k\}$ be a sequence of measurable functions $f_k : X \longrightarrow \mathbb{R}$ such that $|f_k(\mathbf{x})| \leq g(\mathbf{x})$ holds for all $\mathbf{x} \in X$, i.e. g dominates all f_k. If the sequence $\{f_n\}$ converges point-wise to a function f, i.e. $f_k(\mathbf{x}) \overset{pointwise}{\underset{k\to\infty}{\longrightarrow}} f(\mathbf{x})$ then f is absolute integrable, i.e. $f \in \mathcal{L}^1(X)$ with*

$$\lim_{k\to\infty} \int f_k(\mathbf{x})\, d\mu(\mathbf{x}) = \int f(\mathbf{x})\, d\mu(\mathbf{x}) \ .$$

B. Cybenko's Results for Standard MLP

The main statement regarding the universal approximation property of MLP's is given by the following theorem. For the sake of later considerations we also give the proof of the theorem as provided in [5]. We will later make use of that proof structure.

Theorem 8. *Let $I_n = [0,1]^n \subset \mathbb{R}^n$ be the closed hypercube equipped with the Euclidean metric., Let σ be a continuous discriminatory function with respect to the inner product $\langle \cdot, \cdot \rangle$. Further, let*

$$\Pi = \left\{ \pi(\mathbf{x}) \in \mathcal{C}(I_n) \,|\, \pi(\mathbf{x}) = \sum_{j=1}^{N} \alpha_j \cdot \sigma(\langle \mathbf{w}_j, \mathbf{x} \rangle + b_j) \right\} \tag{7}$$

be the set of continuous functions consisting of finite sums of perceptrons (1) with an activation function $f = \sigma$. Then the set $\mathcal{P} = \operatorname{span}(\Pi)$ of functions $\pi(\mathbf{x})$ is dense in the space $\mathcal{C}(I_n)$ of continuous functions over I_n.

Proof: The set $\mathcal{P}$ is dense in $\mathcal{C}(I_n)$ iff for any function $g(\mathbf{x}) \in \mathcal{C}(I_n)$ and $\varepsilon > 0$ exists a function $\pi(\mathbf{x}) \in \mathcal{P}$ with $|\pi(\mathbf{x}) - g(\mathbf{x})| < \varepsilon$ for all $\mathbf{x} \in I_n$. This statement is proven if we can show that for the closure $\overline{\mathcal{P}}$ of $\mathcal{P}$ the equality $\overline{\mathcal{P}} = \mathcal{C}(I_n)$ holds. We apply a proof by contradiction:

Obviously, $\mathcal{P}$ is a linear subspace of $\mathcal{C}(I_n)$. Thus, the closure $\overline{\mathcal{P}}$ is a closed subspace of $\mathcal{C}(I_n)$. We remark that I_n is equipped with the Euclidean norm such that it is a Banach-space or, more precisely, a Hilbert space. Now we suppose that $\overline{\mathcal{P}} \neq \mathcal{C}(I_n)$, i.e. $\mathcal{P}$ is not dense in $\mathcal{C}(I_n)$ and show that this assumption leads to a contradiction:

It follows from the assumed equality according to the Hahn-Banach-theorem that there is a bounded linear functional L on $\mathcal{C}(I_n)$ with $L(h) \neq 0$, i.e. it is not completely vanishing for $h \in \mathcal{C}(I_n)$ but $L(\mathcal{P}) = L\left(\overline{\mathcal{P}}\right) = 0$ is valid. We remark that L is continuous and we have $L \in \mathcal{C}^*(I_n)$ being the dual space of $\mathcal{C}(I_n)$.

According to the Hilbert-space property of I_n we can apply the Riesz-Representation-Theorem (RRT, [23]), which states that the functional L can be written in the form

$$L(h) = \int_{I_n} h(\mathbf{x})\, d\mu(\mathbf{x}) \tag{8}$$

for some measure $\mu \in \mathcal{M}(I_n)$ and a continuous function $h \in \mathcal{C}(I_n)$. Yet, so far μ is unspecified.

Because for the continuous function $\sigma(\langle \mathbf{w}, \mathbf{x} \rangle + b) \in \overline{\mathcal{P}}$ is valid for all $\mathbf{w}$ and b we must have that

$$L(\sigma) = \int_{I_n} \sigma(\langle \mathbf{w}, \mathbf{x} \rangle + b)\, d\mu(\mathbf{x}) = 0$$

holds for all choices $\mathbf{w}$ and b according to $L\left(\overline{\mathcal{P}}\right) = 0$. Since σ is assumed to be discriminatory, the zero integral implies that $\mu \equiv 0$ has to be valid, which further implies, however, that $L(h) \equiv 0$ for any $h \in \mathcal{C}(I_n)$. This contradicts the assumption $\overline{\mathcal{P}} \neq \mathcal{C}(I_n)$. Hence, $\mathcal{G}$ is dense in $\mathcal{C}(I_n)$ which completes the proof. ∎

According to this result and the Lemma 3 we can conclude that also the ReLU-activation ensures the universal approximation property.

Remark 9. In the proof of the Cybenko-theorem the Hilbert-space property of I_n was explicitly used which is guaranteed by the Euclidean metric/norm. Further, the Euclidean norm in I_n is consistent with the mathematical structure of the discriminatory functions $\sigma(\langle \mathbf{w}, \mathbf{x} \rangle + b)$ containing the Euclidean inner product in the argument.

Remark 10. We explicitly remark that the validity of the RRT provided by eq. (8) is essential to complete the proof. The RRT, however, originally requires the Hilbert-space property.

III. Generalizations of Cybenko's Results for MLPs with Generalized Inner Products

In this chapter we generalize the Cybenko-Theorem (8). First, we make the easy step to kernel-based inner products

replacing the inner product in perceptrons. Thereafter, we consider more general inner product variants, namely, semi-inner products and variants thereof.

A. Kernels for Hilbert-Spaces

Obviously, the proof of the Cybenko-theorem remains valid if we replace the Euclidean inner product $\langle \mathbf{w}, \mathbf{x} \rangle$ in the standard perceptron (1) by an arbitrary inner product and use the resulting norm as norm for the n-dimensional real space $\mathbb{R}^n$. We can continue this idea and, more generally, replace the inner product by a kernel κ, i.e. we consider

$$\kappa(\mathbf{w}, \mathbf{x}) = \langle \phi(\mathbf{w}), \phi(\mathbf{x}) \rangle$$

with $\phi(\mathbf{w}) \in \mathcal{H}$ where $\mathcal{H}$ is a reproducing kernel Hilbert space (RKHS) [26]. Then $\mathcal{I}_n = \phi(I_n)$ is compact in the Hilbert space $\mathcal{H}$ and the Cybenko's theorem is still applicable also for $\mathcal{I}_n$.

B. Semi-Inner Products

In the second, more challenging case we want to exchange in the perceptron (1) the inner product $\langle \mathbf{w}, \mathbf{x} \rangle$ by a semi-inner product (SIP) $[\mathbf{w}, \mathbf{x}]$ [20].

Definition 11. A mapping $[\cdot, \cdot] : \mathcal{B} \times \mathcal{B} \to \mathbb{C}$ is called a semi-inner product (SIP) if the following relations are fulfilled:

1) linearity: $[\lambda \mathbf{x} + \mathbf{z}, \mathbf{y}] = \lambda [\mathbf{x}, \mathbf{y}] + [\mathbf{z}, \mathbf{y}]$ for $\lambda \in \mathbb{C}$
2) positiveness: $[\mathbf{x}, \mathbf{x}] > 0$ for $\mathbf{x} \neq \mathbf{0}$
3) Cauchy-Schwarz-inequality: $|[\mathbf{x}, \mathbf{y}]|^2 \leq [\mathbf{x}, \mathbf{x}] [\mathbf{y}, \mathbf{y}]$

LUMER has shown that a SIP always generates a norm by $\|\mathbf{x}\| = \sqrt{[\mathbf{x}, \mathbf{x}]}$ as well as he has proofed that every Banach-space with norm $\|\mathbf{x}\|_{\mathcal{B}}$ is equipped with a SIP generating this norm [20]. Generally, there may exist several SIPs generating a given norm. Additional requirements are needed to ensure uniqueness. Further, given a norm, generally there is no constructive way to derive a respective SIP. Despite this impossibility, one can show that the homogeneity property $[\mathbf{x}, \lambda \mathbf{y}] = \overline{\lambda} [\mathbf{x}, \mathbf{y}]$ can be imposed without causing any significant restriction of the LUMER results [7].

Now we equip I_n with the norm $\|\mathbf{x}\| = \sqrt{[\mathbf{x}, \mathbf{x}]}$ denoted as $I_n^{\mathcal{B}} \subset \mathbb{R}_{\mathcal{B}}^n$. Thus $\mathbb{R}_{\mathcal{B}}^n$ becomes an n-dimensional real Banach-space. Considering now *Banach-like perceptrons (B-perceptron)* with output

$$O(\mathbf{w}, \mathbf{x}) = f([\mathbf{w}, \mathbf{x}] + b) \tag{9}$$

using real SIPs, we cannot simply apply the original Cybenko-theorem to show approximation completeness, because its proof requires the Hilbert-space property needed to apply the RRT. However, as mentioned before, $I_n^{\mathcal{B}}$ is not conained in a Hilbert space. Fortunately, there exist variants of the RRT which suppose weaker but special Banach-spaces instead of a Hilbert-space.

Before we will characterize those Banach-spaces, we have to extend the definition of a discriminatory functions:

Definition 12. The function σ is n-discriminatory with respect to the real-valued linear functional $l(\mathbf{w}, \mathbf{x})$ in $\mathbf{x}$, if for a measure $\mu \in \mathcal{M}(I_n)$ of the closed (compact) subset $I_n = [0,1]^n \subset \mathbb{R}^n$ with the property

$$\int_{I_n} \sigma(l(\mathbf{w}, \mathbf{x}) + b) \, d\mu(\mathbf{x}) = 0$$

for all $\mathbf{w} \in \mathbb{R}^n$ and $b \in \mathbb{R}$ the implication $\mu \equiv 0$ follows. The function σ is said to be discriminatory with respect to the real-valued linear functional $l(\mathbf{w}, \mathbf{x})$ in $\mathbf{x}$, if it is n-discriminatory with respect to the real-valued linear functional $l(\mathbf{w}, \mathbf{x})$ for all n.

Lemma 13. *Any bounded, measurable sigmoidal function is discriminatory with respect to the real-valued linear functional $l(\mathbf{w}, \mathbf{x})$ and, hence, any continuous sigmoidal function is discriminatory.*

Proof: The proof we give here follows the argumentation in [5]. Doing so, we suppose a sigmoid function σ and a real-valued linear functional $l(\mathbf{w}, \mathbf{x})$ with $\int_{I_n} \sigma(l(\mathbf{w}, \mathbf{x}) + b) \, d\mu(\mathbf{x}) = 0$ for given signed measure μ. We have to show that $\mu \equiv 0$ follows.

For this purpose we consider the function

$$\sigma_\lambda(\mathbf{x}) = \sigma(\lambda \cdot (l(\mathbf{w}, \mathbf{x}) + b) + \varphi)$$

which converges point-wise and boundedly to the function

$$\gamma(\mathbf{x}) = \begin{cases} 1 & \text{for} \quad l(\mathbf{w}, \mathbf{x}) + b > 0 \\ 0 & \text{for} \quad l(\mathbf{w}, \mathbf{x}) + b < 0 \\ \sigma(\varphi) & \text{for} \quad l(\mathbf{w}, \mathbf{x}) + b = 0 \end{cases}$$

in the limit $\lambda \longrightarrow +\infty$, i.e. $\sigma_\lambda(\mathbf{x}) \xrightarrow[\lambda \to +\infty]{\text{pointwise}} \gamma(\mathbf{x})$. Hence, $|\sigma_\lambda(\mathbf{x})| \leq \gamma(\mathbf{x})$ is valid. Applying the Dominant-Convergence-theorem 7 we have

$$\int_{I_n} \gamma(\mathbf{x}) \, d\mu(\mathbf{x}) = \lim_{\lambda \to +\infty} \int_{I_n} \sigma_\lambda(\mathbf{x}) \, d\mu(\mathbf{x})$$

with $\int_{I_n} \sigma_\lambda(\mathbf{x}) \, d\mu(\mathbf{x}) = 0$ according to the assumed discriminatory property of σ. Thus we can further calculate for an arbitrary choice of $\mathbf{w}$, b, and φ

$$\begin{aligned} \int_{I_n} \gamma(\mathbf{x}) \, d\mu(\mathbf{x}) &= \int_{X_{\mathbf{w},b}^+} 1 d\mu(\mathbf{x}) + \int_{X_{\mathbf{w},b}^-} 0 d\mu(\mathbf{x}) \\ &\quad + \int_{X_{\mathbf{w},b}^0} \sigma(\varphi) \, d\mu(\mathbf{x}) \qquad (10) \\ &= \mu\left(X_{\mathbf{w},b}^+\right) + \sigma(\varphi) \, \mu\left(X_{\mathbf{w},b}^0\right) \qquad (11) \\ &= 0 \end{aligned}$$

using the definition of $\gamma(\mathbf{x})$ in the first equation together with the half-planes $X_{\mathbf{w},b}^+ = \{\mathbf{x} \in I_n | l(\mathbf{w}, \mathbf{x}) + b > 0\}$ and $X_{\mathbf{w},b}^- = \{\mathbf{x} \in I_n | l(\mathbf{w}, \mathbf{x}) + b < 0\}$ whereas $X_{\mathbf{w},b}^0 = \{\mathbf{x} \in I_n | l(\mathbf{w}, \mathbf{x}) + b = 0\}$ is a hyperplane according to the linearity of $l(\mathbf{w}, \mathbf{x})$. For $\varphi \to +\infty$ we observe $\sigma \to 1$, because σ is sigmoid. Hence,

$$\mu\left(X_{\mathbf{w},b}^+\right) + \mu\left(X_{\mathbf{w},b}^0\right) = 0$$

must be valid in (11). Otherwise, if $\varphi \to -\infty$ we observe that $\sigma \to 0$ holds in (11) and, therefore, $\mu\left(X_{\mathbf{w},b}^+\right) = 0$ must be valid. Thus we have shown that the measures of all half-planes

are zero. It remains to show that from this property it follows that the measure μ has to be zero. This would be trivial for positive measures, but this is not assumed here.

Thus, we now fix $\mathbf{w}$ and consider the linear functional

$$F(h) = \int_{I_n} h(l(\mathbf{w}, \mathbf{x}))\, d\mu(\mathbf{x})$$

for a bounded measurable function h. Hence, $F(h)$ is a bounded functional on $\mathcal{L}^\infty(\mathbb{R})$ because μ is a finite signed measure. We consider two choices for h: First we take the indicator function $\mathbf{1}_{[b,\infty)}$ obtaining

$$\begin{aligned} F\left(\mathbf{1}_{[b,\infty)}\right) &= \int_{I_n} \mathbf{1}_{[b,\infty)}(l(\mathbf{w}, \mathbf{x}))\, d\mu(\mathbf{x}) \\ &= \mu\left(X^+_{\mathbf{w},b}\right) + \mu\left(X^0_{\mathbf{w},b}\right) \\ &= 0 \end{aligned}$$

for the functional. Second, we have the indicator function $\mathbf{1}_{(b,\infty)}$ obtaining

$$\begin{aligned} F\left(\mathbf{1}_{(b,\infty)}\right) &= \int_{I_n} \mathbf{1}_{(b,\infty)}(l(\mathbf{w}, \mathbf{x}))\, d\mu(\mathbf{x}) \\ &= \mu\left(X^+_{\mathbf{w},b}\right) \\ &= 0 \end{aligned}$$

for the open interval (b, ∞). We can decompose indicator functions $h_{\mathbf{1}}$ of arbitrary sets into sums of indicator functions of the above types. Due to the linearity of the functional F (linearity of the integral operator) all these integrals vanish and, hence, $F(h_{\mathbf{1}})$ vanishes for indicator functions. Yet, indicator functions are dense in $\mathcal{L}^\infty(\mathbb{R})$ and, therefore, $F(h) = 0$ for all $h \in \mathcal{L}^\infty(\mathbb{R})$ has to be valid.

In the last step of the proof we consider the functions $h_s(z) = \sin(z)$ and $h_c(z) = \cos(z)$, which are both in $\mathcal{L}^\infty(\mathbb{R})$. We take $z(\mathbf{x}) = l(\mathbf{w}, \mathbf{x})$ and calculate

$$\begin{aligned} F(h_c + i \cdot h_s) &= \int_{I_n} h_c(z(\mathbf{x})) + i \cdot h_s(z(\mathbf{x}))\, d\mu(\mathbf{x}) \\ &= \int_{I_n} \exp(i \cdot z(\mathbf{x}))\, d\mu(\mathbf{x}) \end{aligned}$$

which is the Fourier-transform of the linear functional $l(\mathbf{w}, \mathbf{x})$ with an arbitrary chosen parameter $\mathbf{w}$. However, the Fourier-transform has to be zero in any case which is only possible for $\mu \equiv 0$, which completes the proof. ∎

Lemma 14. *The* $\mathrm{ReLU}(z)$ *from (3) is discriminatory* for $z(\mathbf{x}) = l(\mathbf{w}, \mathbf{x}) + b$, *where* $l(\mathbf{w}, \mathbf{x})$ *is a real-valued linear functional in* $\mathbf{x}$ *and* $\mathbf{w}$.

Proof: The proof is in complete analogy to the proof for Lemma 3: Because in this proof only the linearity of the inner product was used as the essential property of the inner product, the argumentation remains valid also for linear functionals. ∎

Now we start to characterize special Banach-spaces such that we can take them for a Cybenko-like theorem. In particular, we have to identify those Banach-spaces which preserve the possibility to apply an appropriate RRT as it was emphasized in Remark 10

Theorem 15. *Let* $\mathcal{B}$ *be an uniformly convex Banach space with continuous SIP* $[\cdot, \cdot]$. *Then a RRT analogously to (8) is valid.*

Proof: The proof can be found in [7, Theorem 6]. ∎

The theorem can be extended to:

Theorem 16. *Let* $\mathcal{B}$ *be a reflexive Banach space. Then a RRT analogously to (8) is valid.*

Proof: Let $\mathcal{B}$ be a reflexive Banach space and $h \in \mathcal{B}^* = \mathcal{C}(\mathcal{B})$. Then exists a SIP $[\cdot, \cdot]$ and an element $\beta \in \mathcal{B}$ such that $\varphi(\mathbf{x}) = [\mathbf{x}, \beta]$ is a continuous linear functional [6]. Hence, the respective SIP determines a RRT analogously to (8). ∎

Both theorems are related according to the following lemma:

Lemma 17. *Every smooth (continuous) uniformly convex Banach space is also reflexive and strictly convex. The reverse direction is not valid. Hence, Theorem 15 is a special case of Theorem 16.*

Proof: The proof can be found in [6]. ∎

Now we are able to formulate a theorem which states the universal approximation property for perceptron networks consisting of Banach-like perceptrons.

Theorem 18. ***(Cybenko theorem for Banach-like perceptron networks)*** *Let* σ *be a continuous general discriminatory function with respect to the SIP* $[\cdot, \cdot]$ *for* $I_n^{\mathcal{B}} \subset \mathbb{R}^n_{\mathcal{B}}$ *equipped with the norm* $\|\mathbf{x}\| = \sqrt{[\mathbf{x}, \mathbf{x}]}$ *such that* $\mathbb{R}^n_{\mathcal{B}}$ *is a reflexive* n*-dimensional real Banach-space. Additionally, let*

$$\Pi_{\mathcal{B}}(\mathbf{x}) = \sum_{j=1}^{N} \alpha_j \cdot \sigma([\mathbf{w}_j, \mathbf{x}] + b_j) \tag{12}$$

be the finite sum of Banach-like perceptrons (9) with activation function $f = \sigma$. *Then* $\Pi_{\mathcal{B}}(\mathbf{x})$ *is an universal approximator.*

Proof: The proof is in complete analogy to the proof of the Cybenko-theorem. The application of the Hahn-Banach-theorem is not affected by the weaker assumption regarding the Banach-space. The existence of a respective RRT is guaranteed by the previous lemmata. ∎

The most famous examples for (real) Banach-spaces are the spaces L^p and l^p. The latter one is equipped with the unique SIP

$$[\mathbf{w}, \mathbf{x}]_p = \frac{1}{\|\mathbf{x}\|_p^{p-2}} \sum_k w_k \cdot |x_k|^{p-1} \cdot \mathrm{sgn}(x_k) \tag{13}$$

with $1 \leq p < \infty$ [7]. Thus we can equip $I_n^{\mathcal{B}}$ with the SIP $[\mathbf{w}, \mathbf{x}]_p$. Further, the following lemma holds:

Lemma 19. *Both* L^p *and* l^p *are uniformly convex for* $1 < p < \infty$.

Proof: The proof can be found in [12]. ∎

Corollary 20. *The compact set* $I_n^{\mathcal{B}}$ *with the SIP* $[\mathbf{w}, \mathbf{x}]_p$ *from (13) is contained in the uniformly reflexive Banach space* l_p *for* $1 < p < \infty$. *Hence, a RRT analogously to (8) is valid.*

Proof: Just applying Theorem 16 gives the desired result. ∎

The last corollary leads to the following statement:

Lemma 21. *A MLP using Banach-like perceptrons with output*

$$O_p(\mathbf{w}, \mathbf{x}) = f\left([\mathbf{w}, \mathbf{x}]_p + b\right) \tag{14}$$

according to (9) generated by the SIP $[\mathbf{w}, \mathbf{x}]_p$ *from (13) is an universal approximator in case of* $1 < p < \infty$.

Proof: The previous corollary about uniform convexity of the l_p-space together with Lemma 17 guarantees that Theorem 18 is applicable. ∎

The particular B-perceptron (14) is denoted as B_p-perceptron.

ZHANG & ZHANG considered generalized SIPs (gSIP) [31] extending a first attempt by NATH [21]. They considered SIPs $[\mathbf{w}, \mathbf{x}]_\xi$ for a function $\xi : \mathbb{R}_+ \to \mathbb{R}_+$ fulfilling the requirements 1) and 2) of Def. 11. The Cauchy-Schwarz-inequality is replaced by

$$\left|[\mathbf{w}, \mathbf{x}]_\xi\right| \leq \xi\left([\mathbf{w}, \mathbf{w}]_\xi\right) \cdot \psi\left([\mathbf{x}, \mathbf{x}]_\xi\right)$$

for a conjugate function $\psi : \mathbb{R}_+ \to \mathbb{R}_+$, i.e. $\xi(t) \cdot \psi(t) = t$ has to be valid. According to statement in [31] a RRT is also valid for generalized SIP-spaces: For a RRT regarding those gSIPs it is assumed that $\xi(t)$ is a so-called gauge function, i.e. $\xi(0) = 0$ and $\lim_{t\to\infty} \xi(t) = \infty$. If $\xi(t)$ is surjective onto $\mathbb{R}_+$ and $\zeta(t) = \frac{\xi^{-1}(t)}{t}$ is a gauge function on $\mathbb{R}_+$ then a RRT can be formulated, because the resulting Banach-space is reflexive and strictly convex [31].

C. Kernels for Banach-Spaces

In the last step we extend Cybenko's theorem to the case of kernels regarding reproducing kernel Banach spaces (RKBS). As stated in [30, Theorem 4], a RKBS is always reflexive. Thus, we suppose a kernel $\kappa_{\mathcal{B}}$ corresponding to the kernel feature map $\phi_{\mathcal{B}} : I_n \to \mathcal{I}_n \subset \mathcal{B}$ with $\mathcal{B}$ being a RKBS [18]. From Theorem 16 we can conclude that Cybenko's theorem is applicable, accordingly.

IV. NUMERICAL SIMULATIONS

In the simulation part we trained MLPs using B-perceptrons with SIP $[\mathbf{w}, \mathbf{x}]_p$ from (13) for the two well-known data sets MNIST and CIFAR10 [19], [16]. For the MNIST-problem, the gray-value images were vectorized and taken as an input for an MLP with only one hidden layer consisting of 32 B_p-perceptrons with sigmoid activation. For CIFAR10 we used a convolutional network with four convolutional layers and three max-pooling layers. The final dense layer was performed by 10 B_p-perceptrons with ReLU-activation. The convolutional layers were trained using the dense layer for $p = 2$. After this training, the convolutional layers were kept fix - only the dense layer was trained using different p-values.

Both networks were trained using cross-entropy loss for different p-values for the SIP $[\mathbf{w}, \mathbf{x}]_p$. The MNIST-results are depicted in Fig. 2 and Fig. 3.

The MLP is always capable to solve the classification problem appropriately. For large and small p-values, numerical instabilities and difficulties lead to a slightly decreased performance.

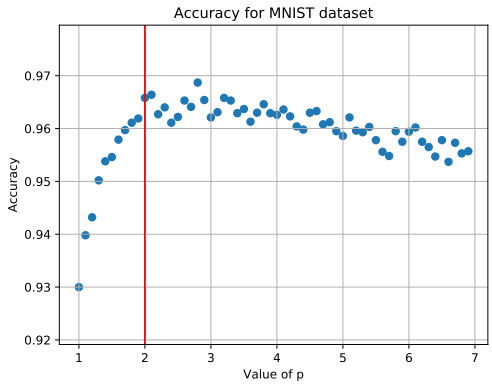

Figure 2. Obtained accuracies of an MLP with B_p-perceptrons for the MNIST data set depending on the p-value for the SIP $[\mathbf{w}, \mathbf{x}]_p$. We observe a broad range of p-values delivering the same good accuracy.

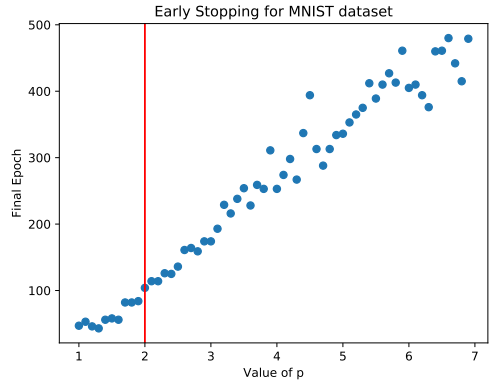

Figure 3. Investigation of the convergence behavior of MLPs with B_p-perceptrons for the MNIST data set depending on the p-value for the SIP $[\mathbf{w}, \mathbf{x}]_p$. A linear correlation between early stopping (number of learning epochs until convergence) and p-value is observable.

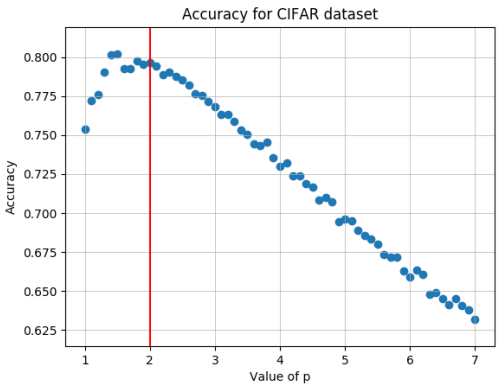

Figure 4. Obtained accuracies of CNN-networks with final dense layers consisting of B_p-perceptrons for the CIFAR10 data set depending on the p-value for the SIP $[\mathbf{w}, \mathbf{x}]_p$. We observe a broad range of p-values delivering the same good accuracy. Particulalrly, p-values lower than one provide good performance.

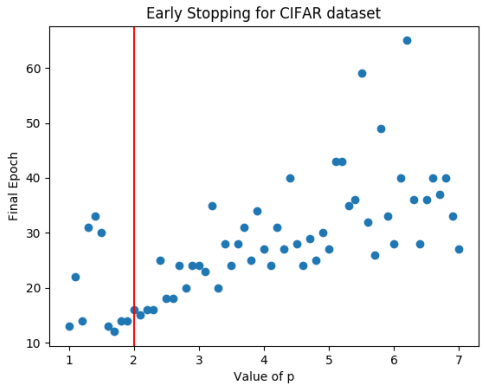

Figure 5. Investigation of the convergence behavior of CNN-networks with final dense layer using B_p-perceptrons for the CIFAR10 data set depending on the p-value for the SIP $[\mathbf{w}, \mathbf{x}]_p$. A rough linear correlation between early stopping (number of learning epochs until convergence) and p-value is observable.

For the CIFAR10 data set the results are depicted in Fig. 4 and Fig. 5.

Again, we can recognize a overall good performance for a wide range of p-values. The decrease of the performance for higher and very low p-values is again attributed to numerical difficulties. These can be observed also from the early-stopping analysis reflecting the somewhat instable convergence behavior.

In this paper we investigated the approximation completeness of multilayer perceptrons consisting of Banach-like perceptrons. These perceptrons use semi-inner products whereas usual perceptrons rely on the standard Euclidean inner product. Semi-inner products are related to Banach-spaces. We prove mathematically that for semi-inner products determining reflexive Banach-spaces the respective perceptron networks are approximation complete. The proof is valid for discriminatory activation functions which comprise both sigmoid and $ReLU$-functions. Numerical simulations accompany the theoretical considerations.

Future work will deal with indefinite inner products as well as will include the investigation of ResNets. Further, other more promising activation functions like *swish* (see [22], [29], [3]) should consideres as well as networks with bounded width [11]

Acknowledgment

Both A. Engelsberger and J. Ravichandran are supported by a PhD-grant of the European Social Fund (ESF).

References

[1] C. Bishop. *Pattern Recognition and Machine Learning*. Springer Science+Business Media, LLC, New York, NY, 2006.

[2] J. Braun and M. Griebel. On a constructive proof of Kolmogorov's superposition theorem. *Constructive Approximation*, 30:653–675, 2009.

[3] H. Chieng, N. Wahid, O. Pauline, and S. Perla. Flatten-T Swish: a thresholded ReLU-Swish-like activation function for deep learning. *International Journal of Advances in Intelligent Informatics*, 4(2):76–86, 2018.

[4] J. Clarkson. Uniformly convex spaces. *Transactions of the American Mathematical Society*, 40:396–414, 1936.

[5] G. Cybenko. Approximations by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2(4):303–314, 1989.

[6] G. D. Faulkner. Representation of linear functionals in a Banach space. *Rocky Mountain Journal of Mathematics*, 7(4):789–792, 1977.

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

[8] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016.

[9] A. Gorban. Approximation of continuous functions of several variables by an arbitrary nonlinear continuous function of one variable, linear functions, and their superpositions. *Applied Mathematical Letters*, 11(3):45–49, 1998.

[10] L. Guilhoto. An overview of artificial neural networks for mathematicians. http://math.uchicago.edu/ may/REU2018/REUPapers/Guilhoto.pdf, 2018.

[11] B. Hanin. Universal function approximation by deep neural networks with bounded width and ReLU activations. *Mathematics*, 7(992):1–9, 2019.

[12] O. Hanner. On the uniform convexity of L^p and l^p. *Arkiv för Matematik*, 3(19):239–244, 1956.

[13] J. A. Hertz, A. Krogh, and R. G. Palmer. *Introduction to the Theory of Neural Computation*, volume 1 of *Santa Fe Institute Studies in the Sciences of Complexity: Lecture Notes*. Addison-Wesley, Redwood City, CA, 1991.

[14] A. Kolmogorov. On the representation of continuous functions of several variables as superpositions of continuous functions of one variable and addition. *Doklady Academ Nauk SSSR*, 114(5):953–956, 1957.

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

[16] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25 (NIPS)*, pages 1097–1105. Curran Associates, Inc., 2012.

[17] V. Kůrková. Kolmogorov's theorem and multilayer neural networks. *Neural Networks*, 5:501–506, 1992.

[18] M. Lange, M. Biehl, and T. Villmann. Non-Euclidean principal component analysis by Hebbian learning. *Neurocomputing*, 147:107–119, 2015.

[19] Y. LeCun, C. Cortes, and C. Burges. The MNIST database, 1998.

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

[21] B. Nath. Topologies on generalized semi-inner product spaces. *Composito Mathematica*, 23(3):309–316, 1971.

[22] P. Ramachandran, B. Zoph, and Q. Le. Searching for activation functions. Technical Report arXiv:1710.05941v1, Google Brain, 2018.

[23] F. Riesz and B. Sz.-Nagy. *Vorlesungen über Functionalanalysis*. Verlag Harri Deutsch, Frankfurt/M., 4th edition, 1982.

[24] F. Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psych. Rev.*, 65:386–408, 1958.

[25] W. Rudin. *Functional Analysis*. MacGraw-Hill, Inc., New York, 2nd edition, 1991.

[26] I. Steinwart and A. Christmann. *Support Vector Machines*. Information Science and Statistics. Springer Verlag, Berlin-Heidelberg, 2008.

[27] H. Triebel. *Analysis und mathematische Physik*. BSB B.G. Teubner Verlagsgesellschaft, Leipzig, 3rd, revised edition, 1989.

[28] T. Villmann, S. Haase, and M. Kaden. Kernelized vector quantization in gradient-descent learning. *Neurocomputing*, 147:83–95, 2015.

[29] T. Villmann, J. Ravichandran, A. Villmann, D. Nebel, and M. Kaden. Investigation of activation functions for Generalized Learning Vector Quantization. In A. Vellido, K. Gibert, C. Angulo, and J. Guerrero, editors, *Advances in Self-Organizing Maps, Learning Vector Quantization, Clustering and Data Visualization – Proceedings of the 13th International Workshop on Self-Organizing Maps and Learning Vector Quantization, Clustering and Data Visualization, WSOM+2019, Barcelona*, volume 976 of *Advances in Intelligent Systems and Computing*, pages 179–188. Springer Berlin-Heidelberg, 2019.

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

[31] H. Zhang and J. Zhang. Generalized semi-inner products with applications to regularized learning. *Journal of Mathematical Analysis and Application*, 372:181–196, 2010.

APPENDIX

In this appendix we give some useful definitions regarding SIPs and Banach spaces, which are used in the text as well as some basic statements and remarks.

Definition 22. A Banach space $\mathcal{B}$ is denoted as *strictly convex* iff for $\mathbf{x}, \mathbf{y} \neq 0$ with $\|\mathbf{x}\| + \|\mathbf{y}\| = \|\mathbf{x} + \mathbf{y}\|$ we can always conclude that $\mathbf{x} = \lambda \mathbf{y}$ for some $\lambda > 0$.

Lemma 23. *A Banach space $\mathcal{B}$ with SIP $[\cdot, \cdot]$ is strictly convex iff for $\mathbf{x}, \mathbf{y} \neq 0$ with $[\mathbf{x}, \mathbf{y}] = \|\mathbf{x}\| \cdot \|\mathbf{y}\|$ we can always conclude that $\mathbf{x} = \lambda \mathbf{y}$ for some $\lambda > 0$.*

Proof: The proof can be found in [7]. ■

The following definition for the uniform convexity was introduced in [4]:

Definition 24. A Banach space $\mathcal{B}$ is denoted as *uniformly convex* iff for each $\varepsilon > 0$ exists a $\delta(\varepsilon) > 0$ such that if $\|\mathbf{x}\| = \|\mathbf{y}\| = 1$ with $\|\mathbf{x} - \mathbf{y}\| > \varepsilon$ then $\frac{\|(\mathbf{x}+\mathbf{y})\|}{2} < 1 - \delta(\varepsilon)$ is valid.

Definition 25. A Banach space $\mathcal{B}$ with SIP $[\cdot, \cdot]$ is denoted as *continuous* iff

$$\Re\{[\mathbf{x}, \mathbf{y} + \lambda \mathbf{x}]\} \underset{\lambda \to 0}{\longrightarrow} \Re\{[\mathbf{x}, \mathbf{y}]\}$$

is valid for $\lambda \in \mathbb{R}$. The space is *uniformly continuous* iff this limit is approached uniformly.

Definition 26. A Banach space $\mathcal{B}$ is denoted as *reflexive* iff the mapping $J : \mathcal{B} \to \mathcal{B}^{**} = (\mathcal{B}^*)^*$ is surjective, where the star indicates the dual space.

Theorem 27. *Let $\mathcal{B}$ be a Banach space. Then a necessary and sufficient condition for $\mathcal{B}$ to be reflexive is that for every $f \in \mathcal{B}^*$ exists an SIP $[\cdot, \cdot]$ and an element $\mathbf{y} \in \mathcal{B}$ with $f(\mathbf{x}) = [\mathbf{x}, \mathbf{y}]$ for all $\mathbf{x} \in \mathcal{B}$. If $\mathcal{B}$ is strictly convex then $\mathbf{y}$ is unique.*

Proof: The proof can be found in [6, Theorem 2]. ■

Definition 28. A Banach space $\mathcal{B}$ is denoted as *smooth* iff for each $\mathbf{x} \in \mathcal{B}$ with $\|\mathbf{x}\| = 1$ there exists a linear functional $f_{\mathbf{x}} \in \mathcal{B}^*$ with $f_{\mathbf{x}}(\mathbf{x}) = \|f_{\mathbf{x}}\|$. The existence of $f_{\mathbf{x}}$ is guaranteed by the Hahn-Banach-Theorem.