Available online at www.sciencedirect.com

ScienceDirect

Journal of Approximation Theory 257 (2020) 105435

Journal of Approximation Theory

www.elsevier.com/locate/jat

Full Length Article

Approximative compactness of linear combinations of characteristic functions

Paul C. Kainen[a], Věra Kůrková[b,*], Andrew Vogt[a]

[a] Department of Mathematics and Statistics, Georgetown University, 37th and O Streets, N.W., Washington DC 20057, United States of America

[b] Institute of Computer Science of the Czech Academy of Sciences, Pod Vodárenskou věží 2, 18207 Prague, Czech Republic

Received 2 September 2019; received in revised form 18 March 2020; accepted 18 April 2020
Available online 30 April 2020

Communicated by H. Mhaskar

Abstract

Best approximation by the set of all n-fold linear combinations of a family of characteristic functions of measurable subsets is investigated. Such combinations generalize Heaviside-type neural networks. Existence of best approximation is studied in terms of approximative compactness, which requires convergence of distance-minimizing sequences. We show that for (Ω, μ) a measure space, in $L^p(\Omega, \mu)$ with $1 \leq p \leq \infty$ and for all $n \geq 1$, compact families of characteristic functions of sets (of finite measure for $p < \infty$) generate approximatively compact n-fold linear spans. Results are illustrated by examples of continuously parametrized sets.

MSC: 41A65; 41A45; 92B20; 52A37; 28A20; 68T05

Keywords: Approximative compactness; Compact sets of characteristic (indicator) functions; Symmetric difference metric; Hausdorff metric; Haar measure; Neural networks

1. Introduction

The theory of neural networks overlaps with approximation theory. Feedforward networks can be formally described as parametrized families of functions, and tools of approximation theory can be used to assess the capabilities of such networks (or, rather, of the sets of functions they encode) to efficiently approximate functions of interest (see, e.g., [17,20]). Properties like

* Corresponding author.
E-mail addresses: kainen@georgetown.edu (P.C. Kainen), vera@cs.cas.cz (V. Kůrková), vogta@georgetown.edu (A. Vogt).

https://doi.org/10.1016/j.jat.2020.105435
0021-9045/

existence, uniqueness, and accuracy of approximation operators (and, at given accuracy, the rate at which number of units increases with increasing number of network parameters) can help selection of proper computational models.

In traditional perceptron-type one-hidden-layer neural networks, the output signal y is a linear combination of parametrized functions acting on the input $x = (x_1, \dots, x_d) \in \mathbb{R}^d$, i.e.,

$$y = \sum_{j=1}^{n} c_j \sigma \left(\sum_{i=1}^{d} w_{ji} \cdot x_i + w_{j0} \right), \tag{1}$$

where c_j and w_{ji} are real *weights*, n is the number of computational units, and σ is a non-decreasing *activation* function. This activation function is often equal to, or an approximation to, the Heaviside function $\vartheta(t)$, which is 1 for $t \geq 0$ and 0 for $t < 0$. For neural networks with the Heaviside activation, dependence of y on x is expressed by a linear combination of characteristic (indicator) functions of half-spaces. These half-spaces may have finite measure for an appropriate measure or if inputs x are restricted to compact subsets. Our results show that characteristic functions of balls, ellipsoids, or polytopes can replace Heavisides as the activation functions, with similar approximation properties.

Training a neural network can be seen as optimizing the parameters (weights) in order to obtain a good approximation to target functions by the corresponding input–output functions. Typically, an L^2 norm is used to measure this closeness. To avoid over-fitting and to decrease vulnerability to outliers, other L^p norms may also be used.

A basic question in approximation theory is whether one can achieve *best approximation* to some element of the ambient function space by members of a given subset. Existence of a best approximation has been formalized by the concept of *proximinal* set (sometimes also called existence set) and existence of unique best approximation by the concept of *Chebyshev* set. A sufficient condition for proximinality is compactness. But compactness can be replaced with a weaker property, which requires only some sequences to have convergent subsequences — the ones which minimize distance to a set. This property, called *approximative compactness*, was introduced in [6] as a tool for exploring the geometry of Banach spaces. Approximative compactness implies proximinality, but the converse is not true. Indeed, in the unit sphere in an infinite-dimensional Hilbert space, all elements are at minimum distance from zero but there are sequences with no convergent subsequence. It was shown in [6] that for a Chebyshev subset of a smooth and uniformly convex Banach space, approximative compactness is equivalent to convexity. For example, the set of rational functions of given degrees of numerators and denominators is approximatively compact in $L^p([0, 1])$, $p \in (1, \infty)$, but is not boundedly compact and the (non-convex) families of fixed-degree rational functions cannot be Chebyshev sets [25, pp. 368–372], [5].

For neural networks based on the Heaviside function, we proved in [14] the *best-approximation property*: for every function $f \in L^p([0, 1]^d)$ with $p \in [1, \infty)$ and every n, there exist values of the weights for which the corresponding distance is the smallest (there may be several different combinations of weights providing the best approximation). However, while minimum distance is achievable, no such best (or even near best) approximation mapping is continuous [12,13].

In this paper we investigate best approximation by sets of n-fold linear combinations of characteristic functions of sets of finite measure. When the family of characteristic functions is compact in $L^p(\Omega, \mu)$ with $p \in [1, \infty]$, it is shown here that the set of all linear combinations of n elements is approximatively compact — and hence has the best approximation property.

Compact sets of characteristic functions in L^p can be obtained from continuously parametrized families restricted to a compact subset of parameter space. Another method to generate compact families of characteristic functions is to use continuous mappings from the set of all compact convex non-empty subsets of $\mathbb{R}^d$, with Hausdorff metric ρ_H, to the set of all finite-λ-measure subsets of $\mathbb{R}^d$ under the symmetric difference metric ρ_Δ. Its continuity follows from results in [7,23].

The relationship between ρ_H and ρ_Δ yields easy arguments for the compactness of various families of compact convex subsets of $\mathbb{R}^d$. For example, *the set of characteristic functions of polytopes, with at most k extreme points all in a prescribed compact subset of $\mathbb{R}^d$, is compact in L^p, $p \in [1, \infty)$.*

Our work evolved from Kůrková [16], where it is shown that families of n-fold linear combinations of characteristic functions of half-spaces are closed in $L^p([0, 1]^d, \lambda)$, $p \in (1, \infty)$. In this paper we allow $p \in [1, \infty]$ and generalize and extend Gurvits and Koiran [8] and earlier work of ours in [14]. Examples are given of compact G of the proper type for the case of a locally compact topological group with Haar measure, mostly when $(\Omega, \mu) = (\mathbb{R}^d, \lambda)$, where λ denotes Lebesgue measure.

The paper is organized as follows: Section 2 contains notation and background material and includes a "virtual convergence" lemma from [14]. Section 3 states our main theorem with some corollaries and remarks while Section 4 contains the proof. Section 5 has Theorem 2 and gives examples, while Section 6 contains a discussion and states open problems.

2. Background and notation

Let $\mathbb{N}_+$ denote the set of all positive integers, and for $n \in \mathbb{N}_+$, $[n] := \{1, \dots, n\}$. For any nonempty set S, let $\mathcal{P}(S)$ denote the set of all subsets of S. For two sets A and B the *symmetric difference* $A \,\Delta\, B$ is the set $(A \cup B) \setminus (A \cap B)$. For Ω a set and $S \subseteq \Omega$, the *characteristic* (or "indicator") function of S is the 0-1-function $\chi_S : \Omega \to \{0, 1\}$, with $\chi_S(x) = 1$ iff $x \in S$. Note that $\chi_\emptyset$ is the zero-function on Ω, where $\emptyset$ denotes the empty subset, and for all A, $B \subseteq \Omega$, we have the following equalities

$$\max\{\chi_A, \chi_B\} = \chi_{A\cup B}, \ \min\{\chi_A, \chi_B\} = \chi_{A\cap B} = \chi_A\, \chi_B \ \text{ and } \ |\chi_A - \chi_B| = \chi_{A\,\Delta\, B}. \tag{2}$$

For any non-empty subset G of a real vector space and $n \geq 1$ a fixed integer, the set of all n-fold linear combinations of elements from G is denoted

$$\operatorname{span}_n(G) := \left\{ \sum_{i=1}^{n} c_i g_i : c_i \in \mathbb{R}, g_i \in G \right\}.$$

In the following result, from [14], if $\{a_{jk}\}$ converges, then a_j is its limit.

Lemma 1. *Let $\mathcal{U}$ be any family of subsets of $[n]$ and $\{a_{jk}\}_{k=1}^{\infty}$, $1 \leq j \leq n$, be a sequence in $\mathbb{R}^n$. If for each S in $\mathcal{U}$, there exists a real number $c_S \in \mathbb{R}$ such that*

$$c_S = \lim_{k\to\infty} \left\{ \sum_{j\in S} a_{jk} \right\},$$

then there exist real numbers $a_1, \dots, a_n$ such that for all S in $\mathcal{U}$

$$c_S = \sum_{j\in S} a_j.$$

Proof. If $\mathcal{U} = \emptyset$, the result holds vacuously. Otherwise, the (finite-dimensional) linear operator

$$(x_1, \dots, x_n) \mapsto \left(\sum_{i \in S_1} x_i, \ \dots \ , \sum_{i \in S_m} x_i \right)$$

has closed range, where $\mathcal{U} = \{S_1, \dots, S_m\}$ for $m = |\mathcal{U}| \geq 1$. ■

2.1. Approximative compactness

Let (X, s) be a metric space. Then the sequence $\{x_k\} \subset X$ **converges subsequentially** to y in X if there is a subsequence of $\{x_k\}$ which converges to y. For $x \in X$ and $M \subset X$, let $s(x, M) := \inf_{m \in M} s(x, m)$. If the metric is determined by a norm $\| \cdot \|$, then $\|x - M\| := s(x, M)$. The subset M is called **proximinal** (an "existence set") if $\forall x \in X$, $\exists y \in M$ s.t. $s(x, y) = s(x, M)$. For $x \in X$, an x-**distance-minimizing sequence** in M is a sequence $\{m_k\}$ in M such that

$$\lim_{k \to \infty} s(x, m_k) = s(x, M);$$

M is **approximatively compact** if for every element x in X and x-distance-minimizing sequence $\{m_k\}$ in M, there exists a point m_∞ of M to which $\{m_k\}$ converges subsequentially. The point m_∞ is a **best approximant** to f. A subset of X is **boundedly compact** if its intersection with any bounded set has compact closure. With obvious abbreviations, for any subset of X,

$$\begin{aligned} &\text{compact} \implies \text{(b.c. and closed)} \implies \text{a.c.} \implies \text{prox.} \\ &\quad \implies \text{closed (w.r.t. metric} - \text{induced topology).} \end{aligned}$$

See Singer [25, p. 365, pp. 382–384].

It is well known that in metric spaces (Simmons [24, p. 124]) compactness is equivalent to **sequential compactness** (every infinite sequence contains a convergent subsequence). It follows that for all $x \in X$, the **projection set** of x in M

$$\Pi_M(x) := \{y \in M : s(x, y) = s(x, M)\}$$

(the *set of best approximants*) is compact when M is approximatively compact.

Since our approach involves the interplay between compactness and approximative compactness, it is natural to ask which operations preserve compactness and approximative compactness. For instance, the Minkowski sum $A + B$ and the Cartesian product $A \times B$ of compact sets A and B is compact. For a class of Banach spaces including those which are uniformly convex, Efimov and Stechkin [6] showed that weakly closed sets are approximatively compact and approximative compactness of M implies that of the closed ε-neighborhood $M_{(\varepsilon)}$. Pyatyshev [21] proved that the closure of the Minkowski sum of two a.c. sets may not be a.c. and that the Minkowski sum of an a.c. set with a compact set is again a.c. Kainen [11] showed that, in an F-space (and so in a normed linear space), if M is a.c. and C is compact, then both the Minkowski sum $M + C$ and the Cartesian product $M \times C$ are a.c., and the metric projection $\Pi_M(C)$ of C to M is compact.

2.2. Measure spaces and metric spaces

In any metric space, let $B(x,r)$ be the closed ball centered at x with radius $r \geq 0$. In a normed linear space, put $B_r := B(0,r)$. A **pseudometric** satisfies the axioms of a metric, except that two distinct points can have distance zero. By defining an equivalence relation (having zero distance), one passes to a quotient space which is a metric space. If s and s' are two different (pseudo)metrics for a set X, then $(X,s) \cong (X,s')$ means that s and s' are **equivalent**, i.e., they induce the same topology on X. The next lemma is well-known (e.g., Kelley [15, p. 131]).

Lemma 2. *Let (X,s) be a metric space with $F : [0,\infty) \to \mathbb{R}$ continuous, nondecreasing, and subadditive such that $F(x) = 0$ iff $x = 0$. Then $(X, F \circ s)$ is a metric space equivalent to (X,s).*

As in Halmos [9], a σ-**ring** in a set Ω is a family of subsets of Ω closed under set-difference and countable unions. A **measurable pair** is a pair $(\Omega, \mathbf{S})$ where $\Omega = \bigcup \mathbf{S}$ and $\mathbf{S}$ is a σ-ring. A **measure** on $(\Omega, \mathbf{S})$ is a function μ from $\mathbf{S}$ to $[0,\infty]$ which is countably additive and assigns measure zero to the empty set. We write (Ω,μ), or $(\Omega, \mathbf{S}, \mu)$ if the σ-ring $\mathbf{S}$ needs to be specified.

Given any class $\mathbf{E}$ of subsets of Ω, there is a unique smallest σ-ring $\mathbf{S(E)}$ containing $\mathbf{E}$ [9, p. 24]. Now suppose that Ω is a locally compact Hausdorff topological space. A **Borel set** is any member of $\mathbf{S(C)}$, where $\mathbf{C}$ is the family of all compact subsets of Ω. A **Borel measure** is a measure μ defined on the class of all Borel sets such that $\mu(C) < \infty$ for every $C \in \mathbf{C}$ [9, p. 223].

A Borel measure μ is **regular** if for each Borel set E, we have

$$\sup\{\mu(C) : E \supseteq C,\ C \text{ is compact}\} = \mu(E) = \inf\{\mu(U) : E \subseteq U,\ U \text{ open}\}.$$

For a measure space (Ω,μ), the **associated metric space** $\mathcal{S}(\Omega,\mu)$ [9, p. 168] is the set of all (equivalence classes of) subsets of Ω which are μ-measurable with finite measure, with the distance between two such sets E and F equal to the measure of their symmetric difference

$$\rho_\Delta(E,F) := \mu(E \Delta F).$$

We identify two subsets of Ω with symmetric difference of measure zero.

When Ω is a locally compact topological group (necessarily Hausdorff by [9, p. 6]), a nonzero Borel measure μ on Ω is called a **Haar measure** if it is *left-invariant*, i.e.,

$$\mu(xE) = \mu(E), \tag{3}$$

where $xE = \{xy : y \in E\}$, xy denotes the group operation for $x \in \Omega$, and E is a Borel subset of Ω.

2.3. L^p-spaces and characteristic functions

Let (Ω,μ) be a measure space and let $p \in [1,\infty)$. Then $L^p := L^p(\Omega,\mu)$ is the set of equivalence classes of measurable functions $f : \Omega \to \mathbb{R}$ whose pth power is absolutely integrable, so f is in L^p iff

$$\int_\Omega |f|^p := \int_\Omega |f|^p \, d\mu := \int_\Omega |f(x)|^p \, \mu(dx) < \infty$$

where functions are equivalent if they agree except for a set of measure zero. For $1 \le p < \infty$, the L^p-norm is the pth root of this integral $\|f\|_p = (\int |f|^p)^{1/p}$. The set of (equivalence classes of) functions with absolute values having finite essential sup is denoted $L^\infty := L^\infty(\Omega, \mu)$ and $\|f\|_\infty :=$ ess sup$(|f|)$. For $f \in L^p(\Omega, \mathbf{S}, \mu)$, $p \in [1, \infty)$, and $A \in \mathbf{S}$, one defines an integral over A by $\int_A f := \int_\Omega f \chi_A$.

For $p \in [1, \infty]$ and letting functions denote their equivalence class modulo sets of μ-measure zero, we introduce notation used repeatedly below. With $[f]$ denoting the equivalence class of f,

$$L^p_\chi := L^p_\chi(\Omega, \mu) := \{[\chi_A] : A \text{ measurable}\} \cap L^p \tag{4}$$

is the metric space with the L^p metric d_p. By (2), for all $A, B \in \mathcal{S}(\Omega, \mu)$ and $p \in [1, \infty)$,

$$d_p(\chi_A, \chi_B) := \|\chi_A - \chi_B\|_p = \mu(A \,\Delta\, B)^{1/p}. \tag{5}$$

Hence, for $p < \infty$, since $F(t) := t^{1/p}$ satisfies Lemma 2, we have shown the equivalence of the L^1 and L^p metrics for characteristic functions of (finite-measure) sets. This proves the second half of the next lemma; the first half is by definition.

Lemma 3. *L^1_χ is isometric to $(\mathcal{S}(\Omega, \mu), \rho_\Delta)$ and for $p \in [1, \infty)$, $L^1_\chi \cong L^p_\chi$.*

The space L^∞_χ consists of the set of all equivalence classes of characteristic functions of measurable subsets of Ω, not just those of finite measure, under essential-sup norm; hence, it is a discrete space with the $0/1$ distance.

3. Main theorem and its consequences

We state our theorem on the approximative compactness of $\mathrm{span}_n(G)$. Before giving the proof, we provide some corollaries and instantiate the concepts.

Theorem 1. *Let $p \in [1, \infty]$. For a measure space (Ω, μ), if $G \subset L^p_\chi(\Omega, \mu)$ is compact, then for $n \in \mathbb{N}_+$, $\mathrm{span}_n(G)$ is approximatively compact.*

Observe that the condition of $0/1$ values is inessential for the members of G provided the norms are bounded away from ∞ and 0. For $p \in [1, \infty)$, we call $G \subset L^p$ **rescalable to 0/1** if for each $g \in G$, there exists a real number ν_g and a measurable subset $\Omega_g \subset \Omega$ of finite measure such that $g = \nu_g \chi_{\Omega_g}$, where $\inf_{g \in G}\{|\nu_g|\} > 0$. Let $G^\nu := \{g/\nu_g : g \in G\}$. Then G^ν is compact if G is.

Corollary 1. *For $p \in [1, \infty)$, let $G \subset L^p(\Omega, \mu)$ be compact and rescalable to $0/1$. Then for $n \in \mathbb{N}_+$, $\mathrm{span}_n(G)$ is approximatively compact in $L^p(\Omega, \mu)$.*

Proof. Theorem 1 implies $\mathrm{span}_n(G^\nu)$ is approximatively compact and $\mathrm{span}_n(G) = \mathrm{span}_n(G^\nu)$. ■

The conclusion of Theorem 1 can also hold when G is non-compact. Indeed, G^ν may be compact when G is not. Let Γ be the family

$$\Gamma := \left\{\chi_{[0,1+\frac{1}{n}]} : n \in \mathbb{N}_+\right\} = \left\{\chi_{[0,2]}, \chi_{[0,3/2]}, \chi_{[0,4/3]}, \ldots\right\}.$$

For c any real number, define $\Gamma_c := \Gamma \cup \{c\chi_{[0,1]}\}$. Then in $L^p(\mathbb{R}, \lambda)$, for $p \in [1, \infty)$, (i) Γ_c is compact iff $c = 1$, (ii) $(\Gamma_c)^\nu = \Gamma_1$ for $c \neq 0$, and (iii) $\mathrm{span}_n(\Gamma_c) = \mathrm{span}_n(\Gamma_1)$ is approximatively compact for $c \neq 0$.

For another example, put $G = \{\chi_{[n,n+1]} : n \in \mathbb{Z}\}$. This G is a non-compact subset of $L^p(\mathbb{R}, \lambda)$ for $1 \le p \le \infty$. While $\text{span}_n(G)$ is approximatively compact for $p < \infty$, it is *not* approximatively compact if $p = \infty$ as the set of nearest neighbors of any nonzero constant is not compact.

Proposition 1. *Let G be a compact subset of a normed linear space $(X, \|\cdot\|)$ and suppose that either $0 \notin G$ or 0 is an isolated point in G. Then $\text{span}_1(G)$ is approximatively compact.*

Proof. Let x be any element in X and let $t := \|\text{span}_1(G) - x\|$ denote the distance from x to $\text{span}_1(G)$. Suppose that $\|\lambda_n g_n - x\| \to t$. If $\{g_n\} \to 0$, then $\{g_n\}$ is equal to 0 for all sufficiently large n and the distance-minimizing sequence $\lambda_n g_n$ is eventually equal to zero. Otherwise, choose $\delta > 0$ such that $\delta \le \inf_{n\ge1} \|g_n\|$. For $C > 0$ an arbitrary constant, choose N so large that for all $n \ge N$, we have $\|\lambda_n g_n - x\| \le C + t$. It follows that for all $n \ge N$,

$$|\lambda_n|\delta \le |\lambda_n|\|g_n\| \le \|\lambda_n g_n - x\| + \|x\| \le C + t + \|x\|$$

so $|\lambda_n| \le (C + t + \|x\|)/\delta$. As the sequence $\{\lambda_n\}$ is bounded, it converges subsequentially to λ_0, while $\{g_n\} \to g_0$, and $\|\lambda_0 g_0 - x\| = \|\text{span}_1(G) - x\|$. ■

We exhibit an example of a compact set G for which $\text{span}_1(G)$ is *not* approximatively compact. Let $G := \{(x, y) : (x - 1)^2 + y^2 = 1\} \subset \mathbb{R}^2$. Then $\text{span}_1(G) = \{(x, y) : x \ne 0\} \cup \{(0, 0)\}$, which is not closed and hence not approximatively compact. Section 6 contains a generalization of this example.

4. Proof of the main theorem

Case 1: $p = \infty$. The essential sup of the difference between two non-equivalent characteristic functions is 1, so L^∞_χ has the 0/1-metric. Hence, if G is an L^∞-compact family of characteristic functions, then G is finite, so $\text{span}_n(G)$ is a finite union of finite dimensional spaces. A finite-dimensional Banach space is boundedly compact and closed, thus approximatively compact; approximative compactness is obviously preserved under finite unions.

Case 2: $p \in [1, \infty)$. Let $f \in L^p$ and let there exist $a_{jk} \in \mathbb{R},\ g_{jk} \in G$ such that

$$\lim_{k\to\infty} \left\| \sum_{j\in[n]} a_{jk}\, g_{jk} - f \right\|_p = \|\text{span}_n(G) - f\|_p, \tag{6}$$

We introduce notation for the combinatorics of the corresponding subset families. Suppose $\mathcal{F} := \{g_1, g_2, \ldots, g_n\}$, $n \ge 1$ is any finite family of characteristic functions $g_i = \chi_{\Omega_i}$ for measurable sets $\Omega_1, \Omega_2, \ldots, \Omega_n \subseteq \Omega$, with $\mathcal{F} \subset L^p(\Omega, \mu)$. If $S \subseteq [n]$, we define $T_{\mathcal{F}}(S) \subseteq \Omega$ by

$$T_{\mathcal{F}}(S) := \left(\bigcap_{i\in S} \Omega_i\right) \cap \left(\bigcap_{j\notin S} \Omega_j^c\right); \tag{7}$$

that is,

$$T_{\mathcal{F}}(S) = \left\{x \in \Omega : \forall j \in [n],\ g_j(x) = 1 \iff j \in S\right\}. \tag{8}$$

One can check that for all $\mathcal{F}$ and all positive integers n, the following properties hold: (i) $\bigcup\{T_{\mathcal{F}}(S) : S \subseteq [n]\} = \Omega$, (ii) $T_{\mathcal{F}}(\emptyset) = \bigcap_{j=1}^n \Omega_j^c$, and (iii) for all $S_1, S_2 \subseteq [n]$, if $T_{\mathcal{F}}(S_1) = T_{\mathcal{F}}(S_2)$, then $S_1 = S_2$ or else $S_1 \neq S_2$ but $T_{\mathcal{F}}(S_1) = \emptyset = T_{\mathcal{F}}(S_2)$.

We now utilize this construction in two separate instances — an infinite sequence coming from the data and a corresponding limit sequence.

As G is compact, so is the n-fold Cartesian product G^n; hence, the sequence of n-vectors $\{g_{jk} : j \in [n]\}$ from (6), has a subsequence $\{g_{jk'} : j \in [n]\}$ converging to $\{g_j : j \in [n]\} \in G^n$. Passing to such a subsequence, one has for each $j \in [n]$,

$$\lim_{k\to\infty} \|g_{jk} - g_j\|_p = 0. \tag{9}$$

The functions g_{jk} and g_j are characteristic functions for Ω_{jk} and Ω_j, resp.,

$$g_{jk} = \chi_{\Omega_{jk}} \text{ and } g_j = \chi_{\Omega_j}.$$

For each $k \geq 1$, we write T_k to denote $T_{\{g_{1k},\ldots,g_{nk}\}}$, so that

$$T_k(S) := \Big\{x \in \Omega : \forall j \in [n],\ g_{jk}(x) = 1 \iff j \in S\Big\}. \tag{10}$$

We write T for $T_{\mathcal{F}}$ when $\mathcal{F} = \{g_1, \ldots, g_n\}$.

The following two equations are immediate and will be used freely below:

$$\forall S \subseteq [n],\ \forall x \in T_k(S),\ \sum_{j\in[n]} a_{jk} g_{jk}(x) = \sum_{j\in S} a_{jk}, \tag{11}$$

and

$$\forall S \subseteq [n],\ \forall x \in T(S),\ \sum_{j\in[n]} a_j g_j(x) = \sum_{j\in S} a_j, \tag{12}$$

where $a_j, a_{jk} \in \mathbb{R}$. By definition, a sum over the empty set is the zero function.

In addition to (9) one may pass to a subsequence of k's such that for every $S \subseteq [n]$: Either there exists $c_S \in (-\infty, \infty)$ s.t. $\lim_{k\to\infty} \sum_{j\in S} a_{jk} = c_S$ or else $\lim_{k\to\infty} |\sum_{j\in S} a_{jk}| = \infty$. Since there are only finitely many sets S, this is possible. Note that $c_\emptyset = 0$.

It is useful to partition $\mathcal{P}[n]$ correspondingly, $\mathcal{P}[n] = \mathcal{U}_1 \cup \mathcal{U}_2$, where

$$\mathcal{U}_1 := \Big\{S \subseteq [n] : \lim_{k\to\infty} \sum_{j\in S} a_{jk} = c_S\Big\}, \tag{13}$$

$$\mathcal{U}_2 := \Big\{S \subseteq [n] : \lim_{k\to\infty} \Big|\sum_{j\in S} a_{jk}\Big| = \infty\Big\}. \tag{14}$$

This partition will enable the application of Lemmas 1 to 5. Observe that $\emptyset \in \mathcal{U}_1$ and that by (6), there exists $C > 0$ such that for each k

$$\int_\Omega \Big|f(x) - \sum_{j\in[n]} a_{jk} g_{jk}(x)\Big|^p d\mu(x) \leq C. \tag{15}$$

Let $v_k := \sum_{j\in[n]} a_{jk} g_{jk}$. By (11), we get for each k and each $S \in \mathcal{U}_2$,

$$\Big(\|f\|_p + \|f - v_k\|_p\Big)^p \geq \|v_k\|_p^p \geq \|v_k \chi_{T_k(S)}\|_p^p = \Big|\sum_{j\in S} a_{jk}\Big|^p \cdot \mu(T_k(S)). \tag{16}$$

The left-hand side is bounded above while $\Big|\sum_{j\in S} a_{jk}\Big|$ blows up, and we obtain

$$\lim_{k\to\infty} \mu(T_k(S)) = 0 \quad \text{for all} \quad S \in \mathcal{U}_2. \tag{17}$$

Lemma 4. *For every $S \subseteq [n]$, $\lim_{k\to\infty} \mu\Big(T_k(S) \Delta T(S)\Big) = 0$.*

Proof. First we note that for $p \in [1, \infty)$, for each $j \in [n]$, and for all $k \geq 1$,

$$\Big\| g_{jk} - g_j \Big\|_p^p = \int_{\Omega_{jk} \cup \Omega_j} \Big| g_{jk} - g_j \Big|^p = \int_{\Omega_{jk} \Delta \Omega_j} \Big| g_{jk} - g_j \Big|^p + \int_{\Omega_{jk} \cap \Omega_j} \Big| g_{jk} - g_j \Big|^p,$$

But $|g_{jk}(x) - g_j(x)|^p = 1$ for $x \in \Omega_{jk} \Delta \Omega_j$ and $= 0$ for $x \in \Omega_{jk} \cap \Omega_j$, so

$$\int_{\Omega_{jk} \Delta \Omega_j} |g_{jk} - g_j|^p = \mu(\Omega_{jk} \Delta \Omega_j) \quad \text{and} \quad \int_{\Omega_{jk} \cap \Omega_j} |g_{jk} - g_j|^p = 0.$$

By (9), we have $\|g_{jk} - g_j\|_p^p \to 0$, so for all $j \in [n]$

$$\lim_{k\to\infty} \mu(\Omega_{jk} \Delta \Omega_j) = 0. \tag{18}$$

It is elementary to check that for arbitrary subsets A, B of some set E, one has $A \,\Delta\, B = A^c \,\Delta\, B^c$ and also that for all subsets A, B, C, D of E

$$(A \cap B) \,\Delta\, (C \cap D) \subseteq (A \,\Delta\, C) \cup (B \,\Delta\, D). \tag{19}$$

Now for $k \geq 1$ and $S \in \mathcal{P}([n])$, by definition, we have

$$T_k(S) \Delta T(S) = \Big[\Big(\bigcap_{j\in S} \Omega_{jk}\Big) \cap \Big(\bigcap_{j\notin S} \Omega_{jk}^c\Big)\Big] \Delta \Big[\Big(\bigcap_{j\in S} \Omega_j\Big) \cap \Big(\bigcap_{j\notin S} \Omega_j^c\Big)\Big]. \tag{20}$$

By (7), (18), and (19), $\mu(T_k(S) \Delta T(S)) \to 0$. ■

As a consequence of Lemma 4 and Eq. (2), we have

$$\lim_{k\to\infty} \mu(T_k(S)) = \mu(T(S)) \ \text{ for all } \ S \subseteq [n]. \tag{21}$$

By definition of symmetric difference, $0 \leq \mu(A) \leq \mu(A \,\Delta\, B) + \mu(B)$ for any two sets A and B. Taking $A = T(S)$ and $B = T_k(S)$, and using Eq. (17) as well as Lemma 4, one sees that

$$\mu(T(S)) = 0 \quad \text{for all} \quad S \in \mathcal{U}_2. \tag{22}$$

Lemma 5. *Let $h \in L^p(\Omega, \mu)$, $p \in [1, \infty)$. Then there exist real numbers $a_1, a_2, \ldots, a_n$ such that*

$$\lim_{k\to\infty} \Big(\int_{T_k(S)} \Big| h - \sum_{j\in S} a_{jk} \Big|^p d\mu \Big) = \int_{T(S)} \Big| h - \sum_{j\in S} a_j \Big|^p d\mu \tag{23}$$

for all $S \in \mathcal{U}_1$.

Proof. With c_S as in (13), choose $a_1, \ldots, a_n$ as in Lemma 1 such that $c_S = \sum_{j\in S} a_j$ for all $S \in \mathcal{U}_1$.

For the common integrand $\Big| h - \sum_{j\in S} a_{jk} \Big|^p$, we get for all $k \geq 1$

$$0 \leq \int_{T_k(S)} - \int_{T_k(S)\cap T(S)} = \int_{T_k(S)\setminus T(S)} \leq \int_{T_k(S)\Delta T(S)} \tag{24}$$

and, again for the integrand $\Big| h - \sum_{j\in S} a_{jk} \Big|^p$,

$$0 \leq \int_{T(S)} - \int_{T_k(S)\cap T(S)} = \int_{T(S)\setminus T_k(S)} \leq \int_{T_k(S)\Delta T(S)}. \tag{25}$$

We claim the right-hand term in (24) and (25) goes to zero as $k \to \infty$, i.e.,

$$\left\|(h - \sum_{j\in S} a_{jk})\, \chi_{T_k(S)\Delta T(S)}\right\|_p^p = \int_{T_k(S)\Delta T(S)} \left|h - \sum_{j\in S} a_{jk}\right|^p \to 0. \tag{26}$$

Let $c := \sup_k \left|\sum_{j\in S} a_{jk}\right|$; since $S \in \mathcal{U}_1$, c is *finite*. So, for all $k \in \mathbb{N}_+$, the following inequality holds

$$\left\|(h - \sum_{j\in S} a_{jk})\, \chi_{T_k(S)\Delta T(S)}\right\|_p \le \|h\chi_{T_k(S)\Delta T(S)}\|_p + \|c\chi_{T_k(S)\Delta T(S)}\|_p. \tag{27}$$

We claim that the right-hand side of (27) goes to zero. Indeed, the second term on the right is $c\,\mu(T_k(S)\Delta T(S))^{1/p}$ which goes to zero by Lemma 4, while the integrand in the first term on the right is dominated by the integrable function $|h|$, so by Lebesgue dominated convergence and Lemma 4, the first term goes to zero, and (26) follows.

Applying (26) to the limits of (24) and (25) we get (for $S \in \mathcal{U}_1$)

$$\lim_{k\to\infty}\left(\int_{T_k(S)} \left|h - \sum_{j\in S} a_{jk}\right|^p - \int_{T(S)} \left|h - \sum_{j\in S} a_{jk}\right|^p\right) = 0. \tag{28}$$

If S is non-empty, then $T(S)$ is a subset of Ω_j for every $j \in S$ and thus of finite measure; if $S = \emptyset$, both sums in (29) vanish. By definition (13) of $\mathcal{U}_1$ and Lemma 1, we get

$$\lim_{k\to\infty} \int_{T(S)} \left|h - \sum_{j\in S} a_{jk}\right|^p = \int_{T(S)} \left|h - \sum_{j\in S} a_j\right|^p \tag{29}$$

From (28) and (29), it follows that Eq. (23) holds. ■

Let $r := \|\mathrm{span}_n(G) - f\|_p$. Using (6), we have

$$r^p = \lim_{k\to\infty} \left\|\sum_{j\in[n]} a_{jk} g_{jk} - f\right\|_p^p$$

and the last term can be re-expressed as a sum over $\mathcal{U}_1$ plus a sum over $\mathcal{U}_2$,

$$\sum_{S\in\mathcal{U}_1} \lim_{k\to\infty} \left\|\Big(\sum_{j\in S} a_{jk} - f\Big)\chi_{T_k(S)}\right\|_p^p + \sum_{S\in\mathcal{U}_2} \lim_{k\to\infty} \left\|\Big(\sum_{j\in S} a_{jk} - f\Big)\chi_{T_k(S)}\right\|_p^p \tag{30}$$

$$\ge \sum_{S\in\mathcal{U}_1} \left\|\Big(\sum_{j\in S} a_j - f\Big)\chi_{T(S)}\right\|_p^p = \sum_{S\in\mathcal{U}_1} \left\|\Big(\sum_{j\in S} a_j g_j - f\Big)\chi_{T(S)}\right\|_p^p, \tag{31}$$

where (for the inequality) we ignore the second summand (over $\mathcal{U}_2$) and use Lemma 5; the last equality is just (12). By (22), $\mu(T(S)) = 0$ for $S \in \mathcal{U}_2$, so the last term in (31) is equal to $\left\|\sum_{j\in[n]} a_j g_j - f\right\|_p^p \ge \|f - \mathrm{span}_n(G)\|_p^p = r^p$. Therefore, all the inequalities are equalities and **$\mathrm{span}_n(G)$ is proximinal**.

This argument also establishes that

$$\lim_{k\to\infty} \sum_{S\in\mathcal{U}_2} \int_{T_k(S)} \left|f - \sum_{j\in S} a_{jk}\right|^p d\mu = 0. \tag{32}$$

We now show that $\mathrm{span}_n(G)$ is approximatively compact by proving that

$$\alpha_k := \left\|\sum_{j\in[n]} a_j g_j - \sum_{j\in[n]} a_{jk} g_{jk}\right\|_p^p \to 0 \quad \text{as} \quad k \to \infty.$$

Since $\mathcal{P}([n]) = \mathcal{U}_1 \cup \mathcal{U}_2$, for all $k \geq 1$, $\alpha_k = \beta_k + \gamma_k$, where

$$\beta_k := \sum_{S \in \mathcal{U}_1} \int_{T_k(S)} \Big| \sum_{j \in S} a_j g_j - \sum_{j \in S} a_{jk} \Big|^p d\mu$$

and

$$\gamma_k := \sum_{S \in \mathcal{U}_2} \int_{T_k(S)} \Big| \sum_{j \in S} a_j g_j - \sum_{j \in S} a_{jk} \Big|^p d\mu .$$

The sequence $\beta_k \to 0$ by Lemma 5, with $h = \sum_{j=1}^n a_j g_j$. To show that $\gamma_k \to 0$, we let $u := \sum_{j \in S} a_j g_j - f$ and $v_{k,S} := f - \sum_{j \in S} a_{jk}$. Then, for all k, one gets

$$\gamma_k = \sum_{S \in \mathcal{U}_2} \left\| (u + v_{k,S}) \chi_{T_k(S)} \right\|_p^p \leq \sum_{S \in \mathcal{U}_2} \left(\left\| u \chi_{T_k(S)} \right\|_p + \sum_{S \in \mathcal{U}_2} \left\| v_{k,S} \chi_{T_k(S)} \right\|_p \right)^p . \tag{33}$$

As $k \to \infty$, the second term in the parenthesis on the right-hand side converges to zero by (32). In addition, the first term in the parenthesis also goes to zero because of (17) and Lebesgue dominated convergence. As $\lim_{k \to \infty} \alpha_k = 0$, **$\mathrm{span}_n(G)$ is approximatively compact**. ■

5. Compact sets of 0-1 functions in $L^p(\mathbb{R}^d, \lambda)$

To use Theorem 1, one needs *compact families of characteristic functions in* $L^p(\Omega, \mu)$. These are provided in the next five examples.

Of course, a trivial way to achieve such compactness is to take a *finite* family. Corresponding families of subsets of $\mathbb{R}^d$ could be chosen at random or with the aim of well-covering the portion of $\mathbb{R}^d$ most important to the data. Correct statistical theory here might facilitate experimentation; e.g., Artstein & Vitale [1], Molchanov [18]. Or one might use a learning algorithm to find a finite collection of subsets of a universe (e.g., $\mathbb{R}^d$) which provides good fits to the class of data considered.

Another way to produce compact sets involves the continuous action of a topological group.

Example 1. Let E be any Borel set in $\mathbb{R}^d$ with finite Lebesgue-measure and let T be a compact subset of $\mathbb{R}^d$. The set of translates of E by elements t in T

$$\mathcal{F} := \{E + t : t \in T\}$$

is a compact subset of $\mathcal{S}(\mathbb{R}^d, \lambda)$ and the corresponding set of characteristic functions $\{\chi_A : A \in \mathcal{F}\}$ is compact in $L^p(\mathbb{R}^d, \lambda)$ for $p \in [1, \infty)$.

This example, in which $(\mathbb{R}^d, +)$ is the group and λ is the Haar measure, is a special case of the following theorem.

Theorem 2. *Let (Ω, μ) be a locally compact topological group with regular Haar measure μ, let $E \subseteq \Omega$ be a Borel set of finite measure. and let $T \subseteq \Omega$ be compact. Then $\{xE : x \in T\}$ is compact in $\mathcal{S}(\Omega, \mu)$, where $xE := \{xy : y \in E\}$ and xy denotes the product of x and y in Γ.*

Proof. The function $\Phi_E : x \mapsto xE$ is continuous from Ω to $\mathcal{S}(\Omega, \mu)$ when E is a Borel set of finite measure [9, pp. 266–268]. So for T compact in Ω, the family $\Phi_E(T) = \{xE : x \in T\}$ is compact in $\mathcal{S}(\Omega, \mu)$. ■

Let $\mathcal{H}$ denote the set of all closed half-space characteristic functions in $\mathbb{R}^d$ so $\mathcal{H} = \{\chi_{H_{e,b}} : (e, b) \in S^{d-1} \times \mathbb{R}\}$, where $H_{e,b} := \{y \in \mathbb{R}^d : e \cdot y \geq b\}$. Given a subset C of $\mathbb{R}^d$, let $\mathcal{H}|_C := \{H|_C : H \in \mathcal{H}\} \setminus \{\emptyset\}$, so $\mathcal{H}|_C$ is the set of characteristic functions of non-empty intersections of C with closed half-spaces. The next example generalizes [14].

Example 2. If C is a compact subset of $\mathbb{R}^d$, then $\mathcal{H}|_C$ is compact in $L^p(\mathbb{R}^d, \lambda)$, $p \in [1, \infty)$.

Proof. The restriction function defined by $A \mapsto E \cap A$ for E a fixed subset in $\mathcal{S}(\mathbb{R}^d, \lambda)$ is continuous [9, p. 168], so we may replace C by any ball B_R (centered at 0) of radius R sufficient to contain C. Non-empty subsets of B_R of the form $B_R \cap H_{e,b}$ will be called *solid caps*. Let $Y = S^{n-1} \times \mathbb{R}$ with metric $|(e, b) - (e', b')|$ equal to the sum of $|b - b'|$ and the geodesic distance in S^{d-1} between e and e'. Put $\Phi(e, b) := H_{e,b} \cap B_R$. If we can show that $\Phi : (Y, |\cdot|) \to (\mathcal{S}, \rho_\Delta)$ is continuous, then the set of solid caps is the image under Φ of a compact set in Y, i.e., $\mathcal{H}|_{B_R} = \Phi\Big(S^{d-1} \times [-R, R]\Big)$, and so is ρ_Δ-compact, hence $\mathcal{H}|_C$ is L^p-compact, $p \in [1, \infty)$.

To prove that Φ *is* continuous, let $\{(e_k, b_k)\} \subset Y$ and suppose for $(e, b) \in Y$,

$$\lim_{k\to\infty} |(e_k, b_k) - (e, b)| = 0.$$

The symmetric difference metric ρ_Δ satisfies the triangle inequality, so

$$\mu(\Phi(e_k, b_k))\Delta\Phi(e, b) \leq \mu(\Phi(e_k, b_k))\Delta\Phi(e_k, b) + \mu(\Phi(e_k, b))\Delta\Phi(e, b).$$

Since the first term on the right is independent of e_k, it suffices to consider the two extreme cases. (For $n \geq 1$, let $\kappa_n = \pi^{\frac{n}{2}}/\Gamma(\frac{n}{2} + 1)$ be the volume of the unit ball in $\mathbb{R}^n$.)

(i) All $e_k = e$. Then, as $b_k \to b$, the volume of the symmetric difference of these concentric caps is bounded by $|b - b_k|\, R^{d-1}\, \kappa_{d-1}$ and hence goes to zero.

(ii) All $b_k = b$. Then solid caps S and S_k corresponding to e and e_k, resp., are rotations of one another by $\vartheta = \arccos(e \cdot e_k)$, so $\lambda(S\Delta S_k) \to 0$ as $e_k \to e$ by continuity of the action of the orthogonal group O_d on $\mathbb{R}^d$. ■

One could also prove continuity using calculus, by expressing the volume of a solid cap when $|b| \leq R$ (required for a non-empty intersection) as

$$\lambda(B_R \cap H_{e,b}) = 2\kappa_{d-2} \int_{t=0}^{\sqrt{R^2-b^2}} \int_{s=b}^{\sqrt{R^2-t^2}} \Big(R^2 - t^2 - s^2\Big)^{\frac{d-1}{2}} \, ds\, dt.$$

The argument is straightforward but nontrivial. If the compact set C in Example 2 is convex, then the sets in the family $\mathcal{H}|_C$ are compact and convex.

For a metric space (X, s), the **Hausdorff pseudometric** [22, p. 47] between bounded non-empty subsets A and B is given by

$$\rho_H(A, B) := \inf\{\varepsilon > 0 : A_{(\varepsilon)} \supseteq B \ \text{ and } \ B_{(\varepsilon)} \supseteq A\},$$

where $A_{(\varepsilon)} := \{x \in X : s(x, A) \leq \varepsilon\}$. Both ρ_H and ρ_Δ are pseudometrics on the set of all compact non-empty subsets of $\mathbb{R}^d$ and so give metrics on the corresponding families of suitable equivalence classes. However, they differ radically in their treatment of the empty set $\emptyset$ which has infinite Hausdorff distance from all other sets but has finite ρ_Δ-distance to any finite measure set.

Theorem (*Groemer [7]*). *For compact convex non-empty subsets A and B of $\mathbb{R}^d$,*

$$\rho_\Delta(A, B) \le c_G\, \rho_H(A, B), \tag{34}$$

where

$$c_G := \frac{2\kappa_d}{-1 + 2^{1/d}} \Big(D/2\Big)^{d-1} \tag{35}$$

with $D := \max(\operatorname{diam}(A), \operatorname{diam}(B))$.

The above inequality permits one to check continuity w.r.t. $L^p(\mathbb{R}^d, \lambda)$ by showing that continuity holds w.r.t. Hausdorff distance. Using (34) and Lemma 3, we get the following.

Proposition 2. *Let $\mathcal{A}$ be a family of compact convex non-empty subsets of $\mathbb{R}^d$. Suppose $\mathcal{A}$ is compact in the topology induced by the Hausdorff metric. Then $\{\chi_A : A \in \mathcal{A}\}$ is a compact subset of $L^p(\mathbb{R}^d, \lambda)$ for $p \in [1, \infty)$.*

This provides an alternate route to Example 2 as the set of solid caps of a ball is compact w.r.t. Hausdorff distance.

The family $\mathcal{K}$, consisting of all **compact convex non-empty** subsets of $\mathbb{R}^d$, is boundedly compact w.r.t. the Hausdorff metric [19, pp. 7, 19, 30-31]. For $C \in \mathcal{K}$, let $\mathcal{K}_C$ be the family

$$\mathcal{K}_C := \{A \cap C : A \in \mathcal{K}\} \setminus \{\emptyset\} = \{C' \subseteq C : \emptyset \ne C' \text{ closed convex}\}.$$

Example 3. If $C \in \mathcal{K}$, then $\{\chi_A : A \in \mathcal{K}_C\}$ is compact in $L^p(\mathbb{R}^d, \lambda)$, $p \in [1, \infty)$.

Proof. The set $\mathcal{K}_C$ is bounded in the metric space $(\mathcal{K}, \rho_H)$ since the Hausdorff distance between any two (non-empty) closed subsets of C is bounded by the diameter of C. Further, $\mathcal{K}_C$ is closed. By bounded compactness of $(\mathcal{K}, \rho_H)$, $\mathcal{K}_C$ is compact and so, by Proposition 2, the corresponding family of characteristic functions is compact in L^p. ■

Example 4. For T compact $\subset \mathbb{R}^d$ and J compact $\subset [0, \infty)$, let

$$\mathcal{B} := \mathcal{B}(T, J) := \{\chi_{B(x,r)} : x \in T,\ r \in J\},$$

be the family of all characteristic functions of closed balls with center in T and radius r in J. Then $\mathcal{B}$ is a compact subset of $L^p(\mathbb{R}^d, \lambda)$, $p \in [1, \infty)$.

Proof. If either T or J is empty, then $\mathcal{B}$ is also empty — hence, compact. Suppose $T \ne \emptyset \ne J$. It is easy to check that, for $|a - b| =$ Euclidean distance, $\rho_H(B(x, r), B(x', r')) \le |x - x'| + |r - r'|$ so $\Phi : \mathbb{R}^d \times [0, \infty) \to (\mathcal{K}, \rho_H)$ given by $\Phi(x, r) := B(x, r)$ is continuous.

Thus, the family $\{B(x, r) : x \in T, r \in J\} = \Phi(T \times J)$ is compact and non-empty. Hence, by Proposition 2, $\mathcal{B}$ is compact in L^p. ■

By continuity of intersection (w.r.t. ρ_Δ), restricting the members of $\mathcal{B}$ to T (that is, intersecting the balls with T), one gets a different compact family. Example 4 and the above remark can be extended to ellipsoids with bounded radii or eccentricity.

A *polytope* P in $\mathbb{R}^d$ is the convex hull of a finite set of points (e.g., [22, pp. 3, 94–95]),

$$P = \operatorname{conv}(\{x_1, \ldots, x_k\}),$$

where $\operatorname{conv}(\{x_1, \ldots, x_k\}) := \{\sum_{i=1}^k c_i x_i\}$ for nonnegative c_i summing to 1. There is a unique minimum subset $\{x_{i_1}, \ldots, x_{i_\ell}\}$ of $\{x_1, \ldots, x_k\}$ such that P is the convex hull of $\{x_{i_1}, \ldots, x_{i_\ell}\}$, constituting the *extreme points* of the polytope [22, p. 18].

Example 5. For T a compact subset of $\mathbb{R}^d$, let $\mathcal{P}(T,k)$ denote the set of all characteristic functions of polytopes with at most k extreme points all of which are in T. Then $\mathcal{P}(T,k)$ is a compact subset of $L^p(\mathbb{R}^n,\lambda)$, $p\in[1,\infty)$.

Proof. For $\mathbf{x}=(\mathbf{x}_1,\dots,\mathbf{x}_k)\in(\mathbb{R}^d)^k$, let $\Phi(\mathbf{x}):=\mathrm{conv}(\{\mathbf{x}_1,\dots,\mathbf{x}_k\})$, so $\Phi:(\mathbb{R}^d)^k\to\mathcal{K}$. For $\mathbf{y},\mathbf{y}'\in(\mathbb{R}^d)^k$, with $|a-b|=$ Euclidean distance, define

$$s(\mathbf{y},\mathbf{y}'):=s((\mathbf{y}_1,\dots,\mathbf{y}_k),(\mathbf{y}'_1,\dots,\mathbf{y}'_k)):=\max\{|\mathbf{y}_1-\mathbf{y}'_1|,\dots,|\mathbf{y}_k-\mathbf{y}'_k|\}.$$

For each i, $1\le i\le k$, $\mathbf{y}_i$ has distance at most $s(\mathbf{y},\mathbf{y}')$ from $\{\mathbf{y}'_1,\dots,\mathbf{y}'_k\}$, so

$$\rho_H\Big(\{\mathbf{y}_1,\dots,\mathbf{y}_k\},\{\mathbf{y}'_1,\dots,\mathbf{y}'_k\}\Big)\le s(\mathbf{y},\mathbf{y}').$$

But (e.g., [19, p. 30]) one has

$$\rho_H\Big(\mathrm{conv}(\{\mathbf{y}_1,\dots,\mathbf{y}_k\}),\mathrm{conv}(\{\mathbf{y}'_1,\dots,\mathbf{y}'_k\})\Big)\le\rho_H\Big(\{\mathbf{y}_1,\dots,\mathbf{y}_k\},\{\mathbf{y}'_1,\dots,\mathbf{y}'_k\}\Big).$$

Hence, $\rho_H(\Phi(\mathbf{y}),\Phi(\mathbf{y}'))\le s(\mathbf{y},\mathbf{y}')$ so $\Phi:((\mathbb{R}^d)^k,s)\to(\mathcal{K},\rho_H)$ is Lipschitz and $\mathcal{P}(T,k)=\Phi(T^k)$ is compact in L^p by Groemer's inequality. ■

6. Discussion

In this paper we proved that a compact subset G of $L^p_\chi=L^p(\Omega,\lambda)\cap\{[\chi_A]:A\subseteq\Omega\}$ determines an approximatively compact set $\mathrm{span}_n(G)$ for each $n\ge 1$. In a sense, characteristic functions of sets of finite measure stand midway between points in Ω and general functions in $L^p(\Omega,\mu)$ for $p\in[1,\infty]$; see, e.g., [10, p. 86]. The results proved here apply functional analysis in this useful case.

Observe that compactness of G is neither necessary nor sufficient to guarantee the approximative compactness of $\mathrm{span}_n(G)$. For example, if G is the standard orthonormal basis $\{e_j : j\in\mathbb{N}_+\}$ in $\ell_2(\mathbb{R})$, then G is not compact but $\mathrm{span}_n(G)=\mathrm{span}_n(\{0\}\cup\{e_j/j : j\in\mathbb{N}_+\})$ is approximatively compact by Theorem 1, as $\{0\}\cup\{e_j/j : j\in\mathbb{N}_+\}$ is compact.

Let $\{e_1,\dots,e_{n+1}\}$ be the standard orthonormal basis of $\mathbb{R}^{n+1}$. Let x,y be real numbers and let

$$G_0:=\{xe_1+ye_2:(x-1)^2+y^2=1\}\ \cup\ \{e_3,e_4,\dots,e_{n+1}\}\subset\mathbb{R}^{n+1},$$

which is a circle plus a finite number of points; hence G_0 is compact. Then

$$\mathrm{span}_n(G_0)=\left\{\sum_{j=1}^{n+1}a_j\,e_j : a_1\neq 0 \text{ or the product } a_2a_3\cdots a_{n+1} \text{ is zero}\right\},$$

and more generally for all k, $1\le k\le n$,

$$\begin{aligned}\mathrm{span}_k(G_0)=\Bigg\{&\sum_{j=1}^{n+1}a_j\,e_j : (i)\ a_1\neq 0 \text{ and exactly } n-k \text{ of the numbers}\\ &a_3,\dots,a_{n+1} \text{ equal zero, or}\\ &(ii)\ a_1=a_2=0 \text{ and at least } n-k-1 \text{ of the numbers } a_3,\dots,a_{n+1} \text{ equal zero, or}\\ &\quad (iii) \text{ at least } n-k+1 \text{ of the numbers } a_3,\dots,a_{n+1} \text{ equal zero}\Bigg\}.\end{aligned}$$

Let $h := \sum_{j=1}^{n+1} a_j\, e_j$, where $a_1 = 0$ but $a_2 \neq 0$ and exactly $n-k$ of the numbers $a_3, \ldots, a_{n+1}$ equal zero. Then h is not in $\operatorname{span}_k(G_0)$ but is in its closure. Hence, $\operatorname{span}_k(G_0)$ is not closed and so is not approximatively compact.

We ask whether or not there exists a generalization of Groemer's inequality (34) to $X = \mathbb{R}^d$ with a metric other than Euclidean metric and a Borel measure other than Lebesgue measure. If $\mathcal{K}(X)$ denotes the set of non-empty compact convex subsets of X, writing ρ_H and ρ_Δ, resp., for the corresponding Hausdorff and symmetric-difference metrics, is it true that the identity function is continuous from $(\mathcal{K}(X), \rho_H)$ to $(\mathcal{K}(X), \rho_\Delta)$?

As an important special case, one may consider a target function f, to be L^p-approximated, which is itself a characteristic function $f = \chi_E$, where E is a measurable subset of Ω and E is of finite measure when p is finite. Let G be a compact subset of $L^p_\chi(\Omega, \mu)$ $(1 \leq p \leq \infty)$. We can project f to the set $\Pi(f)$ consisting of the best approximations to f in $\operatorname{span}_n(G)$ w.r.t. the d_p-metric. But linear combinations of characteristic functions are not likely to be of that same kind.

In fact, we can approximate such a function $f \in L^p_\chi$ by other characteristic functions which are derived from $\Pi(f)$. For any element $h = \sum_{i=1}^n a_i \chi_{\Omega_i}$ and for any $t \in \mathbb{R}$, define the "level set"

$$A(h,t) := \left\{x \in \Omega : \sum_{i=1}^n a_i \chi_{\Omega_i}(x) \geq t\right\} = \{x \in \Omega : h(x) \geq t\}; \tag{36}$$

Cheang and Barron [2] used this construction to approximate the unit ball w.r.t. ρ_H or ρ_Δ. Observe that $A(h,t) = \Omega$ if $t \leq \min_{\emptyset \subseteq J \subseteq [n]} \sum_{j\in J} a_j$. Similarly, $A(h,t) = \emptyset$ if $t > \max_{\emptyset \subseteq J \subseteq [n]} \sum_{j\in J} a_j$. The parametrized family $\mathcal{F}(h) := \{A(h,t), t \in \mathbb{R}\}$ has at most 2^n elements. To approximate f, we propose using

$$Q(f) := \bigcup_{h\in\Pi(f)} \mathcal{F}(h). \tag{37}$$

If $Q(f)$ is compact, it constitutes a new family G' and allows an iteration of the previous process. Analogous constructions apply if f is replaced by a compact subset $C \subseteq L^p_\chi$ as in [11].

Characteristic functions of sets offer the possibility of *localized approximation*. Chui, Li, and Mhaskar [3] introduced this notion to describe a situation in which perturbation of a target function in some small region only requires the readjustment of a small subset of the weights. They argue that a one-hidden-layer network, using Heaviside activation, cannot provide localized approximation for input dimension $d > 1$ (though a two-hidden-layer network, which they describe, does give localized approximation). They further argue in [4] that a neural network (with sigmoidal function) having a single hidden layer cannot represent the characteristic function of a cube. But the networks we consider in Theorem 1 do not suffer from these limitations.

Indeed, our construction replaces closed half-spaces by subsets of $\mathbb{R}^d$, taken from a compact family, as the "parameter" for an indicator function (i.e., 0/1-neuron). If the target function is also a characteristic function, then localized approximation is given by a partition (or an approximate partition) of the domain into sets whose characteristic functions belong to a compact family G.

When the sets corresponding to G are polytopes, achieving each of the corresponding indicator functions can be obtained with a small two-layer Heaviside network. For example, the characteristic function of the d-cube can be obtained with $2d$ units in the first layer and one in the second layer.

To apply Theorem 1, one asks which families G give the best fit by $\mathrm{span}_n(G)$ to a particular class of natural data, e.g., street scenes of human activity? A variety of features or image-attributes such as color, texture, translucency, sharpness of edge, fundamental geometry and topology, extent, and motion-blur, among others, may be used to define regions-of-interest in the data, where various recognition engines could be applied. Results here may connect machine learning with the theory of random closed or compact sets [1] using functional approximation by characteristic functions.

CRediT authorship contribution statement

Paul C. Kainen: Conceptualization, Formal analysis, Writing original draft. **Věra Kůrková:** Conceptualization, Formal analysis, Writing original draft. **Andrew Vogt:** Conceptualization, Formal analysis, Writing original draft.

Acknowledgments

We thank Alexander Levin for the reference to Pyatyshev's work. We also thank the referees for suggesting improvements to the exposition and for pointing out a defect in our original statement of Lemma 1. V.K. was partially supported by the Czech Science Foundation, Czech Republic grant GA18-23827S and institutional support of the Institute of Computer Science, Czech Republic RVO 67985807.

References

[1] Z. Artstein, R.A. Vitale, A strong law of large numbers for random compact sets, Ann. Probab. 3 (5) (1975) 879–882, https://projecteuclid.org/euclid.aop/1176996275.

[2] G.H.L. Cheang, A.R. Barron, A better approximation for balls, J. Approx. Theory 104 (2000) 183–203, http://dx.doi.org/10.1006/jath.1999.3441.

[3] C.K. Chui, X. Li, H.N. Mhaskar, Neural networks for localized approximation, Math. Comput. 63 (208) (1994) 607–623.

[4] C.K. Chui, X. Li, H.N. Mhaskar, Limitations of the approximation capabilities of neural networks with one hidden layer, Adv. Comput. Math. 5 (1) (1996) 233–243.

[5] F. Deutsch, Existence of best approximations, J. Approx. Theory 28 (1980) 132–154, http://dx.doi.org/10.1016/0021-9045(80)90085-4.

[6] N.V. Efimov, S.B. Stechkin, Approximative compactness and Chebyshev sets, Sov. Math. Dokl. 2 (1961) 1226–1228; Dokl. Akad. Nauk. SSSR 140 (1961) 522–524, Orig. in Russian.

[7] H. Groemer, On the symmetric difference metric for convex bodies, Beitr. Algebra Geom. 41 (1) (2000) 107–114.

[8] L. Gurvits, P. Koiran, Approximation and learning of convex superpositions, J. Comput. System Sci. 55 (1997) 161–170, http://dx.doi.org/10.1006/jcss.1997.1506.

[9] P.R. Halmos, Measure Theory, D. Van Nostrand, Princeton, NJ, 1950, http://dx.doi.org/10.1007/978-1-4684-9440-2.

[10] E. Hille, R.S. Phillips, Functional Analysis and the Theory of Semigroups, American Mathematical Soc., Providence, RI, 1957.

[11] P.C. Kainen, Replacing points by compacta in neural network approximation, J. Franklin Inst. 341 (2004) 391–399, http://dx.doi.org/10.1016/j.jfranklin.2004.03.001.

[12] P.C. Kainen, V. Kůrková, A. Vogt, Approximation by neural networks is not continuous, Neurocomputing 29 (1999) 47–56.

[13] P.C. Kainen, V. Kůrková, A. Vogt, Geometry and topology of best and near-best approximations, J. Approx. Theory 105 (2000) 252–262, http://dx.doi.org/10.1006/jath.2000.3467.

[14] P.C. Kainen, V. Kůrková, A. Vogt, Best approximation by linear combinations of characteristic functions of half-spaces, J. Approx. Theory 122 (2003) 151–159, http://dx.doi.org/10.1016/S0021-9045(03)00072-8.

[15] J.L. Kelley, General Topology, D. Van Nostrand, Princeton, NJ, 1955.

[16] V. Kůrková, Approximation of functions by perceptron networks with bounded number of hidden units, Neural Netw. 8 (1995) 745–750, http://dx.doi.org/10.1016/0893-6080(95)00027.

[17] V. Kůrková, M. Sanguineti, Bounds on rates of variable-basis and neural-network approximation, IEEE Trans. Inform. Theory 47 (6) (2001) 2659–2665, http://dx.doi.org/10.1109/18.945285.

[18] I. Molchanov, Theory of Random Sets, Springer, London, 2005, http://dx.doi.org/10.1007/1-84628-150-4.

[19] M. Moszyńska, Selected Topics in Convex Geometry, Birkhäuser, Boston, 2006, http://dx.doi.org/10.1007/0-8176-4451-2.

[20] A. Pinkus, Approximation theory of the MLP model in neural networks, Acta Numer. 8 (1999) 143–195, http://dx.doi.org/10.1017/S0962492900002919.

[21] L.A. Pyatyshev, Operations on approximatively compact sets, Mathematical Notes 82 (2007) 653–659; Mat. Zametki 82 (2007) 729–735, http://dx.doi.org/10.1134/S0001434607110089, original in Russian.

[22] R. Schneider, Convex Bodies: The Brunn–Minkowski Theory, in: Encyclopedia of Mathematics and its Applications, Cambridge Univ. Press, Cambridge, 1993, http://dx.doi.org/10.1017/CBO9780511526282.

[23] G.C. Shephard, R.J. Webster, Metrics for sets of convex bodies, Mathematika 12 (1965) 73–88, http://dx.doi.org/10.1112/S0025579300005179.

[24] G.F. Simmons, Introduction to Topology and Modern Analysis, McGraw-Hill, NY, 1963.

[25] I. Singer, Best Approximation in Normed Linear Spaces, Springer, NY, 1970, http://dx.doi.org/10.1007/978-3-662-41583-2.