Annals of Operations Research 101, 143–147, 2001

Continuity of Approximation by Neural Networks in L^p Spaces

PAUL C. KAINEN
Department of Mathematics, Georgetown University, Washington, DC 20057, USA

VĚRA KŮRKOVÁ
Institute of Computer Science, Academy of Sciences of the Czech Republic, P.O. Box 5, 182 07 Prague 8, Czech Republic

ANDREW VOGT
Department of Mathematics, Georgetown University, Washington, DC 20057, USA

Abstract. Devices such as neural networks typically approximate the elements of some function space X by elements of a nontrivial finite union M of finite-dimensional spaces. It is shown that if $X = L^p(\Omega)$ ($1 < p < \infty$ and $\Omega \subset R^d$), then for any positive constant Γ and any continuous function ϕ from X to M, $\|f - \phi(f)\| > \|f - M\| + \Gamma$ for some f in X. Thus, no continuous finite neural network approximation can be within any positive constant of a best approximation in the L^p-norm.

Keywords: Chebyshev set, strictly convex space, boundedly compact, continuous selection, near best approximation

AMS subject classification: primary 82C32; secondary 41A65, 41A46, 41A50, 41A52

1. Introduction

A neural network consists of d input nodes, h hidden nodes, and 1 or more output nodes, connected by an architecture of weights and activation functions. Without loss of generality we study the case where there is exactly one output node and the neural network is used for functional approximation. In this case the objective is to approximate a function $f : \Omega \to R$ where Ω is a subset of R^d. Given f, we seek a function of the form

$$x \longmapsto \sum_{j=1}^{h} w_j g(a_j, x), \quad x \in \Omega, \tag{1}$$

that best approximates f. The functions $g(a_1, \cdot), \ldots, g(a_h, \cdot)$ are chosen from a parametric family of functions $g(a, \cdot)$ with parameter a in some subset A of R^m, while the weights $w_1, \ldots, w_h$ are real numbers. A typical case is a family of Heaviside functions: $g(a, x) = H(c + b_1 x_1 + \cdots + b_d x_d)$, where $H(u) = 1$ for $u \geqslant 0$ and $H(u) = 0$ otherwise, $a = (c, b_1, \ldots, b_d)$, and $m = d + 1$. Other cases are based on the logistic function, radial Gaussians, etc. We suppose that f and each function $g(a, \cdot)$ are members of the space $L^p(\Omega)$, and that approximation is with respect to the L^p-norm for some fixed p

satisfying $1 < p < \infty$. The set Ω is assumed to be the closure of a nonempty open subset of R^d. A function of the form (1) can also be described as follows: it is a member of $\text{span}_h G$ where $G = \{g(a, \cdot)\colon a \in A\}$ and $\text{span}_h G$ denotes the union of all subspaces spanned by h elements of G. As in the examples mentioned above, the family G without loss of generality can be taken to be linearly independent.

The issues of interest can be expressed alternatively in the languages of optimization theory or approximation theory. With respect to optimization theory, we wish to find a function in $\text{span}_h G$ that minimizes the L^p-distance to f, determine whether this function is unique or not, and determine whether it varies continuously as f is varied. See [2] for a treatment of well-posedness of optimization problems. With respect to approximation theory, the issues are whether f has a best approximation, whether it is unique, and whether any best approximation operator is continuous. For example, when the function space is a uniformly convex Banach space, best approximation by closed convex subsets is unique and continuous. However, $\text{span}_h G$ is not convex, and this is the situation considered below. We also discuss the more general issue of whether a near best, rather than best, approximation function can be continuous.

2. Best approximation

Let X be a normed real vector space. If X is sequentially complete in the norm, X is called a *Banach* space. Let X^* be the space of all continuous linear functionals $h : X \to R$, with the norm $\|h\| = \sup\{|h(x)|\colon \|x\| \leqslant 1\}$. Let M be a subset of X. (Our prototype is $X = L^p(\Omega)$ with the L^p-norm, and $M = \text{span}_h G$.) For x in X let $\|x - M\| = \inf\{\|x - m\|\colon m \in M\}$. Then we denote by $P_M(x)$ the set $\{m \in M\colon \|x - m\| = \|x - M\|\}$. An element of $P_M(x)$ is called a *best approximation* to x. Existence, uniqueness, and continuity can then be rephrased in the following terms. If $P_M(x)$ is nonempty, a best approximation for x exists. If this is true for all x in X, then M is said to be *proximinal*. If $P_M(x)$ is a singleton set, the best approximation is unique. If this is true for all x in X, then M is said to be a *Chebyshev* set. Finally, a (continuous) best approximation of X by M exists when the set-valued map $x \mapsto P_M(x)$ has a (continuous) selection defined on all of X, i.e., a (continuous) map $\phi : X \to M$ such that $\phi(x)$ is in $P_M(x)$ for all x.

The space X is *strictly convex* iff whenever x and y are distinct unit vectors in X all nontrivial convex combinations of the two have norm less than 1. All L^p-spaces are uniformly convex, and all uniformly convex spaces are strictly convex (see, e.g., [3, p. 232]).

Theorem 2.1. If M is a subset of a strictly convex subspace X and there exists a continuous best approximation of X by M, then M is a Chebyshev set.

For a proof of this theorem, see [5, theorem 3.2] or [6, theorem 2.3]. The idea of the proof is simply that if two points m_1 and m_2 lie in $P_M(x)$, then the best approximation map restricted to the union of the two intervals $[m_1, x] \cup [x, m_2]$ cannot be continuous. Continuity would demand that the intervals map to $\{m_1, m_2\}$, a disconnected set.

Theorem 2.2. Let $1 < p < \infty$ and let G be a linearly independent subset of $L^p(\Omega)$ with $|G| > h$. Then there is no continuous best approximation of $L^p(\Omega)$ by $\text{span}_h G$.

Proof. Let $X = L^p(\Omega)$ and $M = \text{span}_h G$. Then X is strictly convex, and so is its dual X^*. Hence, by a theorem of Vlasov (see [4]), if M is a Chebyshev set for which the best approximation of X by M is continuous, then M is convex. By theorem 2.1 if P_M has a continuous selection, M is a Chebyshev set and thus is convex. A convex union of subspaces is itself a subspace, but [5, lemma 4.1] shows that this is impossible when G is as above. □

An example indicates the importance of strict convexity. Let $X = R^2$ with the L^1-norm $\|(x, y)\| = |x| + |y|$, and let $M = \{(x, y): y = \pm x\}$. Then X is not a strictly convex space, M is not convex and not Chebyshev, M is a union of subspaces, and $P_M((x, y))$ is always an interval or a pair of intervals meeting at $(0, 0)$. Furthermore, there does exist a continuous best approximation $\phi: X \to M$, namely, the function defined by $\phi(x, y) = (\min\{x, y\}, \min\{x, y\})$ for (x, y) in the first quadrant, with a symmetric definition in the other quadrants.

3. Near best approximation

We consider a generalization of best approximation. For a nonnegative number ε, an *ε-near best approximation* of X by M is a function $\phi: X \to M$ such that $\|\phi(x) - x\| \leqslant \|x - M\| + \varepsilon$ for all x in X.

The set $P_M(x)$ of best approximants of x by elements in M is the intersection of M with the smallest closed ball centered at x which intersects M, the ball of radius $\|x - M\|$. Similarly, $P_{M,\varepsilon}(x)$ is the intersection of M with a closed ball centered at x of radius $\varepsilon + \|x - M\|$. A function from X to M is an ε-near best approximation of X by M if and only if it is a selection for the set-valued mapping $P_{M,\varepsilon}$.

A set M is *boundedly compact* iff the closure of its intersection with any ball of finite radius is compact (cf. [7, p. 365]). A closed, boundedly compact set is always proximinal. When M is boundedly compact and closed, then the sets $P_M(x)$ and $P_{M,\varepsilon}(x)$ are compact.

Theorem 3.1. If M is a closed, boundedly compact subset of a strictly convex space X and for every $\varepsilon > 0$ there is a continuous ε-near best approximation of X by M, then M is a Chebyshev set.

Here is a sketch of the proof of theorem 3.1; for details, see [6, theorem 2.3].

It suffices to show that for an arbitrary point x in X, $P_M(x)$ is a singleton. Plainly, this will not be affected by translation or scalar multiplication so without loss of generality, take $x = 0$ and assume that $\|x - M\| = 1$. For every positive integer n, let ϕ_n be a $1/n$-near best approximation of X by M. Since the sets of near best approximants are compact (by bounded compactness of M), it follows from the Schauder fixed

point theorem that there is a point x_n on the boundary of the unit ball B satisfying $x_n = -\pi \circ \phi_n(x_n)$, where π is the normalization map $\pi(y) = y/\|y\|$. Moreover, $\|x_n - M\| \geqslant 2 - 1/n$. Again by bounded compactness of M, the sequence x_n has a subsequence converging to a point x_∞. It follows from strict convexity of X that if m is any point in $P_M(x)$, then $m = -x_\infty$.

Theorem 3.2. Let $1 < p < \infty$, let G be a finite linearly independent subset of $L^p(\Omega)$ with $|G| > h$, and let $\phi : L^p(\Omega) \to \text{span}_h G$ be a continuous function. Then for every $\Gamma > 0$ there exists a function f in $L^p(\Omega)$ such that $\|f - \phi(f)\| > \|f - \text{span}_h G\| + \Gamma$.

Proof. Put $X = L^p(\Omega)$ and $M = \text{span}_h G$ and assume the theorem is false, i.e., that for some $\Gamma > 0$ there is a Γ-near best approximation of X by M. Since M is positively homogeneous, it follows that for every $\varepsilon > 0$, there exists an ε-near best approximation of X by M; see [6, corollary 2.4]. Since G is linearly independent and $|G| > h$, M is not convex. Moreover, since G is finite, M is boundedly compact. Now apply theorem 3.1 to complete the proof. □

4. Multiple output nodes

In the introduction we observed that it was sufficient to consider the case of one output node. If the number of output nodes is k, the weights w_j in the introduction should be replaced by weights $w_{i,j}$ with $1 \leqslant i \leqslant k$ and $1 \leqslant j \leqslant h$. Then as before we seek an approximation of $f : \Omega \to R^k$ where f is in $L^p(\Omega, R^k)$, $\Omega \subseteq R^d$, and $\|f\|_p^p = \|f_1\|_p^p + \cdots + \|f_k\|_p^p$ with $f_1, \ldots, f_k : \Omega \longrightarrow R$ the component functions of f. The approximation is to be chosen from the set

$$M^{(k)} = \left\{F\colon\ \Omega \to R^k : \exists g_1, \ldots, g_h \in G \text{ with } F_j \in \text{span}(g_1, \ldots, g_h)\ \forall j\right\}.$$

Analogues of theorems 2.2 and 3.2 hold in this case under the same hypotheses on G. For the first theorem, it can be shown that $M^{(k)}$ is not convex. In the second theorem, $M^{(k)} \subseteq M \times \cdots \times M$ (k times), where $M = M^{(1)} = \text{span}_h G$, and $M^{(k)}$ is boundedly compact whenever M is.

5. Remarks

With a neural network, one approximates an unknown function by first choosing a finite set of basis functions that yield the outputs of the hidden units, and then choosing a suitable linear combination of the basis functions. There are two sets of parameters involved: weights which control the linear combination and parameters which select the given basis functions. Basis functions are continuously but nonlinearly parametrized. Since we only aim at approximation, it is not necessary to have a continuous family of basis functions available when the inputs are restricted to lie in a compact subset. Accordingly, the basis functions can be taken to lie in some finite (though large) set; that

is, the set of functions produced by their linear combinations is a finite union of finite dimensional subspaces.

Admittedly, the weights are also discrete in practice and this too prevents continuity. However, as technology improves, both parameters and weights can be chosen more precisely; theorem 3.2 indicates that continuous approximations will still be arbitrarily poor for some functions.

Theorem 3.2 says that no finite neural network approximation operator can be continuous unless its error, measured in an L^p-norm with $1 < p < \infty$, exceeds the minimum by more than any prescribed constant for at least one function in the domain. By continuity the same conclusion holds for an entire open set of functions.

What are the implications for nonlinear optimization? While there are known lower bounds in the case of continuous approximation schemes (DeVore et al. [1]) which force exponentially slow convergence, our results show that these lower bounds do not apply in the neural network case. If a neural network approximation technique can be developed which remains within some fixed additive constant of the best approximation error, then it might compensate for an occasional non-continuity by faster convergence.

Indeed, neural network approximation provides novel applications for nonlinear optimization. In particular, the sets of parametrized functions corresponding to neural networks determine nonconvex subsets of the ambient function space for which distance optimization is needed. These nonconvex subsets, however, have a regular structure as unions of convex sets. See [6] for some results on the topology and geometry of best and near best approximants in these sets.

Acknowledgement

V. Kůrková was partially supported by GA ČR grant 201/99/0092. Collaboration of V. Kůrková and A. Vogt was supported by an NRC COBASE grant.

References

[1] R. DeVore, R. Howard and C. Micchelli, Optimal nonlinear approximation, Manuscripta Mathematica 63 (1989) 469–478.
[2] A.L. Dontchev and T. Zolezzi, *Well-Posed Optimization Problems*, Lecture Notes in Mathematics 1543 (Springer, Berlin, 1993).
[3] E. Hewitt and K. Stromberg, *Real and Abstract Analysis* (Springer, New York, 1965).
[4] R. Huotari and W. Li, Continuities of metric projections and geometric consequences, J. Approx. Theory 90 (1997) 319–339.
[5] P.C. Kainen, V. Kůrková and A. Vogt, Approximation by neural networks is not continuous, Neurocomputing 29 (1999) 47–56.
[6] P.C. Kainen, V. Kůrková and A. Vogt, Geometry and topology of continuous best and near best approximations, J. Approx. Theory 105 (2000) 252–262.
[7] I. Singer, *Best Approximation in Normed Linear Spaces by Elements of Linear Subspaces* (Springer, New York, 1970).