Dimension lower bounds for linear approaches to function approximation

Daniel Hsu

Abstract

This short note presents a linear algebraic approach to proving dimension lower bounds for linear methods that solve L^2 function approximation problems. The basic argument has appeared in the literature before (e.g., Barron, 1993) for establishing lower bounds on Kolmogorov n-widths. The argument is applied to give sample size lower bounds for kernel methods.

1 Introduction

Function approximation is an important problem in many areas, and it is increasingly important for the methods of approximation to be computationally tractable when they are to be used in applications. Linearity is a property that has both enabled the development of efficient algorithms for approximation, as well as the tractable mathematical analyses of such "linear methods" (defined below). However, it has also been recognized that linear methods may be severely limited for solving certain approximation problems. The purpose of this note is to demonstrate such limitations through a simple dimension argument.

We are primarily concerned with L^2 approximation of functions. Let $\mathcal{X}$ be a domain (typically a subset of $\mathbb{R}^d$), P be a probability distribution on $\mathcal{X}$, and $L^2(P)$ be the space of real-valued functions on $\mathcal{X}$ that are square-integrable with respect to P. For any class of functions $\mathcal{F} \subseteq L^2(P)$, a *linear method* for approximating functions from $\mathcal{F}$ is one that commits to choosing the approximation from a subspace $W \subseteq L^2(P)$ before getting any information about the target function from $\mathcal{F}$. This is the setup behind the concept of Kolmogorov n-widths, and the argument given in this note is largely based on a lower bound by Barron (1993, Lemma 6) for the Kolmogorov n-width of a certain class of functions. We do not know the lineage of this argument, but it has recurred in the literature several times in related contexts (e.g., Blum et al., 1994; Kamath et al., 2020; Daniely and Malach, 2020; Hsu et al., 2021). We present a result from Hsu et al. (2021) in a slightly more general form to establish a lower bound on the dimension of the subspace used by any linear method that is able to achieve small approximation error with respect to $\mathcal{F}$. The bound is given in terms of the number of near-orthogonal functions contained in $\mathcal{F}$.

We use the dimension lower bound to give a sample size lower bound for kernel methods (Schölkopf and Smola, 2002). There are many lower bounds for kernel methods in the literature (e.g., Ben-David et al., 2002; Warmuth and Vishwanathan, 2005; Khardon and Servedio, 2005; Wei et al., 2019; Kamath et al., 2020; Allen-Zhu and Li, 2020). Our goal is to simply show how such a lower bound follows easily from the dimension lower bound, and also to point out an aspect of kernel methods as they relate to learning with non-adaptive membership queries.

2 The dimension lower bound

The following theorem is from Hsu et al. (2021, Theorem 29) in a slightly more specialized form. (Also see Kamath et al., 2020, Theorem 19 for a very similar theorem.)

Theorem 1. *Let H denote a Hilbert space with inner product denoted by $\langle \cdot, \cdot \rangle_H$ and norm denoted by $\|\cdot\|_H$. Fix any $\varphi_1, \ldots, \varphi_N \in H$ with $\|\varphi_i\|_H^2 = 1$ for all $i = 1, \ldots, N$. Let $\mathbf{W}$ be a finite-dimensional subspace of H (and $\mathbf{W}$ is allowed to be random) with $r := \mathbb{E}[\dim(\mathbf{W})] < +\infty$. Define*

$$\epsilon := \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}\left[\inf_{g \in \mathbf{W}} \|g - \varphi_i\|_H^2 \right].$$

Then

$$r \geq N \cdot \frac{1 - \epsilon}{1 + \sqrt{\sum_{i \neq j} \langle \varphi_i, \varphi_j \rangle_H^2}}.$$

Equality holds when the φ_i form an orthonormal basis for H.

Proof. Let $\mathbf{u}_1, \ldots, \mathbf{u}_{\mathbf{d}}$ be an orthonormal basis for $\mathbf{W}$, with $\mathbf{d} := \dim(\mathbf{W})$. Let $\Pi_{\mathbf{W}}$ denote the orthogonal projection operator for $\mathbf{W}$. Then

$$\begin{aligned}
\epsilon &= \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}\left[\inf_{g \in \mathbf{W}} \|g - \varphi_i\|_H^2 \right] && \text{(definition)} \\
&= \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}\left[1 - \|\Pi_{\mathbf{W}} \varphi_i\|_H^2 \right] && \text{(Hilbert projection theorem)} \\
&= 1 - \frac{1}{N} \mathbb{E}\left[\sum_{i=1}^{N} \sum_{k=1}^{\mathbf{d}} \langle \mathbf{u}_k, \varphi_i \rangle_H^2 \right] && \text{(linearity of expectation)} \\
&= 1 - \frac{1}{N} \mathbb{E}\left[\sum_{k=1}^{\mathbf{d}} \sum_{i=1}^{N} \langle \mathbf{u}_k, \varphi_i \rangle_H^2 \right] && \text{(switching order of summations)} \\
&\geq 1 - \frac{1}{N} \mathbb{E}\left[\sum_{k=1}^{\mathbf{d}} \left(1 + \sqrt{\sum_{i \neq j} \langle \varphi_i, \varphi_j \rangle_H^2} \right) \right] && \text{(Fact 1)} \\
&= 1 - \frac{r}{N} \left(1 + \sqrt{\sum_{i \neq j} \langle \varphi_i, \varphi_j \rangle_H^2} \right) && \text{(linearity of expectation).}
\end{aligned}$$

By Parseval's identity, the inequality holds with equality when the φ_i form an orthonormal basis for H. □

Fact 1 (Boas, 1941; Bellman, 1944)**.** *For any $g, \varphi_1, \ldots, \varphi_N$ in an inner product space,*

$$\sum_{i=1}^{N} \langle g, \varphi_i \rangle^2 \leq \langle g, g \rangle^2 \left(\max_{1 \leq i \leq N} \langle \varphi_i, \varphi_i \rangle^2 + \sqrt{\sum_{i \neq j} \langle \varphi_i, \varphi_j \rangle^2} \right).$$

3 Lower bounds for kernel methods

We can use Theorem 1 to give a sample size lower bound for kernel methods. A kernel method based on n training examples $(x_1, y_1), \ldots, (x_n, y_n) \in \mathcal{X} \times \mathcal{Y}$ returns a function of the form

$$x \mapsto \sum_{i=1}^{n} \alpha_i K(x, x_i)$$

for some $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$ (which may depend on the training examples). Here K is a positive definite kernel function on the input space $\mathcal{X}$, and $\mathcal{Y}$ is the output space (e.g., $\{-1, 1\}$). The subspace of such functions has dimension at most n. Let $H = L^2(P)$ where P is the probability distribution on $\mathcal{X}$ that we care about, and let $\varphi_1, \ldots, \varphi_N$ be orthonormal functions in H. If a kernel method can guarantee expected mean squared error at most ϵ for every φ_i, then by Theorem 1, the sample size n must be at least $(1 - \epsilon)N$.

Note that the argument above holds as long as the subspace does not depend on the target function to be approximated. A typical approach is to obtain $\mathbf{x}_1, \ldots, \mathbf{x}_n$ as an iid sample from P, in which case a kernel method chooses a function from a (random) subspace $\mathbf{W}$ defined to be the span of the n the functions $x \mapsto K(x, \mathbf{x}_i)$ for $i = 1, \ldots, n$. The choice of the function within $\mathbf{W}$ is typically guided by the labels $y_1, \ldots, y_n$, and the labels may depend on the target function to be approximated (e.g., $y_i = \varphi_j(\mathbf{x}_i)$ for all $i = 1, \ldots, n$ if φ_j is the target function). However, the above argument also applies even if the $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are selected deterministically or in any other way, with the corresponding labels $y_1, \ldots, y_n$ only being revealed after commiting to these $\mathbf{x}_i$. This model of learning is a form of learning with membership queries (Angluin, 1988) where the queries are restricted to be non-adaptive.

Example: learning parity functions. As a simple example, take $H = L^2(P)$ where P is the uniform distribution on the discrete hypercube $\{-1, 1\}^d$, and let $\varphi_1, \ldots, \varphi_N$ be the $N = 2^d$ parity functions (which take values in $\{-1, 1\}$). The parity functions form an orthonormal basis for H. Theorem 1 implies that every kernel method needs $n \geq (1 - \epsilon)2^d$ in order to guarantee expected mean squared error ϵ against every parity function. Or, let $\varphi_1, \ldots, \varphi_N$ be the $N = \binom{d}{k}$ parity functions that are k-sparse (i.e., only involve k variables). Then Theorem 1 implies a lower bound of $n \geq (1 - \epsilon)\binom{d}{k}$ for the same against k-sparse parity functions. As mentioned above, these lower bounds hold even if the kernel method is granted non-adaptive membership queries.

An interesting aspect of this lower bound for kernel methods was pointed out by Bubeck (2020), following the work of Allen-Zhu and Li (2020). Specifically, there are efficient algorithms (which are not kernel methods) for learning any parity function in the non-adaptive membership query model, that run in $\mathrm{poly}(d, 1/\epsilon)$ time and use sample size $n = \mathrm{poly}(d, 1/\epsilon)$. Moreover, the learning guarantee holds even if the labels y_i are corrupted by noise in the manner of the classification noise model of Angluin and Laird (1988). Such efficient algorithms are not known in the usual statistical model without membership queries (and are conjectured not to exist). Also, the class of k-sparse parity functions can be learned in $\mathrm{poly}(d, 1/\epsilon)$ time and with sample size $n = \mathrm{poly}(k, \log d, 1/\epsilon)$ (Feldman, 2007), again, in the non-adaptive membership query model. It is not known how to achieve this without using membership queries. So, the lower bounds we described above (based on Theorem 1) imply that kernel methods are not able to benefit from non-adaptive membership queries in the same way that general polynomial-time learning algorithms are.

References

Zeyuan Allen-Zhu and Yuanzhi Li. Backward feature correction: How deep learning performs deep learning. *arXiv preprint arXiv:2001.04413*, 2020.

Dana Angluin. Queries and concept learning. *Machine learning*, 2(4):319–342, 1988.

Dana Angluin and Philip Laird. Learning from noisy examples. *Machine Learning*, 2(4):343–370, 1988.

Andrew R. Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, 39(3):930–945, May 1993.

Richard Bellman. Almost orthogonal series. *Bulletin of the American Mathematical Society*, 50: 517–519, 1944.

Shai Ben-David, Nadav Eiron, and Hans Ulrich Simon. Limitations of learning via embeddings in euclidean half spaces. *Journal of Machine Learning Research*, 3(Nov):441–461, 2002.

Avrim Blum, Merrick Furst, Jeffrey Jackson, Michael Kearns, Yishay Mansour, and Steven Rudich. Weakly learning dnf and characterizing statistical query learning using fourier analysis. In *Twenty-Sixth Annual ACM Symposium on Theory of Computing*, 1994.

Ralph P. Boas, Jr. A general moment problem. *American Journal of Mathematics*, 63:361, 1941.

Sebastien Bubeck. Provable limitations of kernel methods, 2020. URL `https://www.youtube.com/watch?v=U-XsUB69mvc`.

Amit Daniely and Eran Malach. Learning parities with neural networks. In *Advances in Neural Information Processing Systems 33*, 2020.

Vitaly Feldman. Attribute-efficient and non-adaptive learning of parities and dnf expressions. *Journal of Machine Learning Research*, 8(7), 2007.

Daniel Hsu, Clayton Sanford, Rocco A Servedio, and Emmanouil-Vasileios Vlatakis-Gkaragkounis. On the approximation power of two-layer networks of random relus. *arXiv preprint arXiv:2102.02336*, 2021.

Pritish Kamath, Omar Montasser, and Nathan Srebro. Approximate is good enough: Probabilistic variants of dimensional and margin complexity. In *Thirty-Third Annual Conference on Learning Theory*, 2020.

Roni Khardon and Rocco A Servedio. Maximum margin algorithms with boolean kernels. *Journal of Machine Learning Research*, 6(Sep):1405–1429, 2005.

Bernhard Schölkopf and Alexander J Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

Manfred K Warmuth and SVN Vishwanathan. Leaving the span. In *Eighteenth Annual Conference on Computational Learning Theory*, 2005.

Colin Wei, Jason D Lee, Qiang Liu, and Tengyu Ma. Regularization matters: Generalization and optimization of neural nets vs their induced kernel. In *Advances in Neural Information Processing Systems 32*, 2019.