Available online at www.sciencedirect.com

SciVerse ScienceDirect

Journal of Approximation Theory 173 (2013) 33–55

JOURNAL OF Approximation Theory

www.elsevier.com/locate/jat

Full length article

Approximation by multivariate Bernstein–Durrmeyer operators and learning rates of least-squares regularized regression with multivariate polynomial kernels

Bing-Zheng Li

Department of Mathematics, Zhejiang University, Hangzhou 310027, PR China

Received 20 May 2012; received in revised form 14 April 2013; accepted 19 April 2013
Available online 9 May 2013

Communicated by Ding-Xuan Zhou

Abstract

In this paper, we establish error bounds for approximation by multivariate Bernstein–Durrmeyer operators in $L^p_{\rho_X}$ $(1 \le p < \infty)$ with respect to a general Borel probability measure ρ_X on a simplex $X \subset \mathbb{R}^n$. By the error bounds, we provide convergence rates of type $O(m^{-\gamma})$ with some $\gamma > 0$ for the least-squares regularized regression algorithm associated with a multivariate polynomial kernel (where m is the sample size). The learning rates depend on the space dimension n and the capacity of the reproducing kernel Hilbert space generated by the polynomial kernel.

Keywords: Approximation; Learning theory; Reproducing kernel Hilbert space; Polynomial kernel; Regularization error; Bernstein–Durrmeyer operators; Covering number; Regularization scheme

1. Introduction

In this paper, we study the approximation by multivariate Bernstein–Durrmeyer operators in $L^p_{\rho_X}$ with respect to an arbitrary Borel probability measure ρ_X on a simplex $X \subset \mathbb{R}^n$. The established error bounds are applied to the least-squares regression associated with multivariate polynomial kernels. Our purpose is to give learning rates for the regression algorithm.

Let X be a compact metric space, $Y = \mathbb{R}$ and ρ be a Borel probability measure on $Z := X \times Y$. The least-squares *error* (or the generalization error) [8,9,20] for a function $f : X \to Y$ is defined

E-mail address: libingzheng@zju.edu.cn.

0021-9045/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jat.2013.04.007

as

$$\mathcal{E}(f) = \int_Z (f(\mathbf{x}) - y)^2 d\rho. \tag{1.1}$$

The function that minimizes the least squares error is called the *regression function*, which is given by

$$f_\rho(\mathbf{x}) = \int_Y y d\rho(y|\mathbf{x}), \quad \mathbf{x} \in X. \tag{1.2}$$

Here $\rho(\cdot|\mathbf{x})$ is the conditional probability measure at $\mathbf{x} \in X$ induced by ρ.

The target of the regression problem is to learn the regression function or to find good approximations from random samples. The least-squares regularized algorithm for regression considered here is generated by a Mercer kernel [2,18] $K : X \times X \to \mathbb{R}$ which is a continuous, symmetric and positive semidefinite function in the sense that for any finite set of points $\{\mathbf{x}_1, \ldots, \mathbf{x}_l\} \subset X$, the matrix $(K(\mathbf{x}_i, \mathbf{x}_j))_{i,j=1}^l$ is positive semidefinite. The *reproducing kernel Hilbert space* (RKHS) $\mathcal{H}_K$ associated with the kernel K is defined [2] to be the completion of the linear span of the set of functions $\{K_\mathbf{x} := K(\mathbf{x}, \cdot) : \mathbf{x} \in X\}$ with the inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}_K} = \langle \cdot, \cdot \rangle_K$ given by $\langle K_\mathbf{x}, K_\mathbf{t} \rangle = K(\mathbf{x}, \mathbf{t})$. The reproducing property takes the form

$$\langle K_\mathbf{x}, g \rangle_K = g(\mathbf{x}), \quad \forall \mathbf{x} \in X,\ g \in \mathcal{H}_K.$$

Denote $C(X)$ as the space of continuous functions on X with the norm $\|\cdot\|$. Let $\kappa := \sup_{\mathbf{x} \in X} \sqrt{K(\mathbf{x}, \mathbf{x})}$. Then the above reproducing property tells us that

$$\|g\| \le \kappa \|g\|_K, \quad \forall g \in \mathcal{H}_K. \tag{1.3}$$

Denote $\mathbf{z} = \{z_i\}_{i=1}^m = \{(\mathbf{x}_i, y_i)\}_{i=1}^m \in Z^m$ a set of random samples independently drawn according to ρ. We call m the sample size. Define the empirical error $\mathcal{E}_\mathbf{z}(f)$ of $f : X \to \mathbb{R}$ as

$$\mathcal{E}_\mathbf{z}(f) = \frac{1}{m} \sum_{i=1}^m (f(\mathbf{x}_i) - y_i)^2.$$

It is a discretization of the least-squares error $\mathcal{E}(f)$. The least-squares regularized algorithm for regression associated with the Mercer kernel K is defined by the following least-squares optimization problem involving the set $\mathbf{z}$ of random samples

$$f_\mathbf{z} = f_{\mathbf{z},\lambda} = \arg \min_{f \in \mathcal{H}_K} \left\{ \mathcal{E}_\mathbf{z}(f) + \lambda \|f\|_K^2 \right\}. \tag{1.4}$$

Here $\lambda > 0$ is a constant, called the regularization parameter. Usually it is chosen to depend on $m : \lambda = \lambda(m)$, and $\lim_{m \to \infty} \lambda(m) = 0$.

The efficiency of the learning algorithm (1.4) is measured by the excess generalization error $\mathcal{E}(f_\mathbf{z}) - \mathcal{E}(f_\rho)$, which can be decomposed into a sample error and a regularization error as described in Proposition 1.1 below.

Throughout this paper, we assume that for some $M \ge 0$, $\rho(\cdot|\mathbf{x})$ is supported on $[-M, M]$ for almost every $\mathbf{x}$, that is, $|y| \le M$ almost surely. It follows from the definition (1.2) of f_ρ that almost everywhere

$$|f_\rho(\mathbf{x})| \le M. \tag{1.5}$$

We know that the efficiency of the algorithm (1.4) should be measured by the difference between $f_\mathbf{z}$ and the regression function f_ρ. Because of the least-squares nature, the measurement