Journal of Computational Physics 326 (2016) 931–946

Contents lists available at ScienceDirect

Journal of Computational Physics

www.elsevier.com/locate/jcp

On a near optimal sampling strategy for least squares polynomial regression ☆

Yeonjong Shin, Dongbin Xiu *

Department of Mathematics, The Ohio State University, Columbus, OH 43210, USA

ARTICLE INFO

Article history:
Received 16 March 2016
Received in revised form 31 August 2016
Accepted 12 September 2016
Available online 16 September 2016

Keywords:
Polynomial regression
Least squares
Christoffel function

ABSTRACT

We present a sampling strategy of least squares polynomial regression. The strategy combines two recently developed methods for least squares method: Christoffel least squares algorithm and quasi-optimal sampling. More specifically, our new strategy first choose samples from the pluripotential equilibrium measure and then re-order the samples by the quasi-optimal algorithm. A weighted least squares problem is solved on a (much) smaller sample set to obtain the regression result. It is then demonstrated that the new strategy results in a polynomial least squares method with high accuracy and robust stability at almost minimal number of samples.

1. Introduction

This paper is concerned with polynomial regression using least squares approach. As one of the most used methods in practice for approximating unknown functions, the performance of polynomial least squares regression depends on sample points, where data of the unknown function are collected. The choices of the samples typically follow two distinct approaches: random samples and deterministic samples. In random sampling, i.e., Monte Carlo (MC) method, one draws the samples from a probability measure, which is often defined by the underlying problem, whereas in deterministic sampling the samples follow certain fixed and deterministic rules to fill up the space systematically. For example, quasi Monte Carlo methods (QMC), lattice rules, orthogonal arrays, etc. Studies have been devoted to determine an "optimal" set, which aims to produce the best possible least squares solution with minimal number of points. Different design criteria have been proposed, resulting in a collection of optimality designs, e.g., A-optimality, D-optimality, V-optimality, etc. The study of these generally falls into the topic of design of experiments (DOE), see, for example, [1,2,7,11–13,19,21], and the references therein. In a recent work of [20], a design criterion is derived by directly minimizing the difference between the finite set least squares solution and the ideal solution from a dense set. The results in the so-called quasi-optimal set.

The polynomial least squares solution is obtained by solving an over-determined linear system of equations for the coefficients. Let the size of the system be $M \times p$, where M is the number of samples and p is the number of unknown coefficients in the polynomial. Ideally, one would prefer M to be as large as possible. However, since collecting sample data requires resources, small number of samples are preferred from a practical point of view. A well accepted rule-of-thumb is to use linear over-sampling by letting $M = \alpha p$, where $\alpha = 1.5 \sim 3$. On the other hand, some recent mathematical analysis revealed that such a linear over-sampling is asymptotically unstable for polynomial least squares using Monte Carlo sampling and quasi Monte Carlo, cf. [8,16,15,14]. In fact, for asymptotically stable polynomial regression, one should have at least $M \propto$

☆ This work is in part supported by AFOSR FA95501410022, DARPA N660011524053, and National Science Foundation DMS 1418771.

* Corresponding author.
E-mail addresses: shin.481@osu.edu (Y. Shin), xiu.16@osu.edu (D. Xiu).

http://dx.doi.org/10.1016/j.jcp.2016.09.032

$p \log p$ and in many cases $M \propto p^2$. We remark that these results are for bounded domains, i.e., d-dimensional hypercube. In a more recent work [17], a method termed Christoffel least squares (CLS) was proposed. Here one computes a weighted least squares problem, where the weights are derived from the Christoffel function of orthogonal polynomials. Instead of using the standard MC samples from the measure of the underlying problem, one samples instead from the pluripotential equilibrium measure. Analysis and extensive numerical tests in both bounded domain and unbounded domains were presented in [17] and demonstrated much improved stability property.

In this paper, we propose a sampling strategy for polynomial least squares regression by combining the two aforementioned recent developments from [20] and [17]. We first generate the samples by using the Christoffel least squares (CLS) method [17]. This means that we re-formulate the problem into the weighted least squares polynomial regression and generate samples from the equilibrium measure (regardless of the measure defined by the problem). Prior to collecting the data at the samples, we then apply the subset selection criterion from [20] to determine the quasi-optimal subset, for any given number of samples specified by users. The data of the target function are then collected at this subset only and the corresponding weighted polynomial least squares problem is solved for the approximation of the target function. By doing so, the new strategy can produce, for any finite number of samples, the polynomial least squares regression result as close as possible to the CLS result obtained by an arbitrarily large number of samples, which has the near optimal stability property. In practice, this allows one to conduct the polynomial least squares regression using any given/affordable number of samples and obtain near the optimal approximation result. Moreover, the quasi-optimal subset selection from [20] can be implemented in a greedy manner, resulting in a ranking of the dense background samples from the CLS. This further allows one to add samples in a nested manner and obtain progressively better (and near optimal) approximation results. The combination of the quasi-optimal subset selection and the CLS method results in a near optimal sampling strategy for least squares regression using polynomials. Our numerical results demonstrate the advantage of this strategy. The results also suggest that the major contribution for the performance improvement stems from the quasi-optimal subset selection.

Upon presenting the basic setup and a brief review of the CLS method [17] and the quasi-optimal subset selection [20] in Section 2, we present the procedure for the new near optimal sampling strategy in Section 3, along with some of its theoretical properties. Extensive numerical results are then presented in Section 4 to demonstrate the performance of the method.

2. Preliminaries

Consider a function $f(x)$, where $x \in D \subseteq \mathbb{R}^d$, $d \geq 1$. Let $y = (y_1, \dots, y_M)$, where $y_i = f(x_i)$, $i = 1, \dots, M$, be the data vector. We are interested in approximating $f(x)$ using the data y. Here we consider polynomial approximation and let Π be a polynomial space with $\dim \Pi = p$. Let $\phi_i(x)$, $i = 1, \cdots, p$, be a set of basis functions for Π and

$$q(x; c) = \sum_{i=1}^{p} c_i \phi_i(x). \tag{2.1}$$

Requiring q to approximate f results in the following linear system of equations,

$$Ac = y,$$

where $c = (c_1, \dots, c_p)^T$ is the coefficient vector,

$$A = (a_{ij}), \qquad a_{ij} = \phi_j(x_i), \quad 1 \leq i \leq M, \quad 1 \leq j \leq p,$$

is the model matrix. Here we focus on the over-determined regression problem with $M > p$. The standard least squares method seeks to minimize $\|Ac - y\|^2$. Its solution is $c = (A^T A)^{-1} A^T y$. Throughout this paper all norms are the vector 2-norm, unless otherwise specified.

One can introduce weights in the least squares formulation. Let $W = \operatorname{diag}(w_1, \dots, w_M)$ be a diagonal matrix with positive entries, a weighted least squares problem can be written as

$$\min \|\sqrt{W} Ac - \sqrt{W} y\|^2 = \min \sum_{i=1}^{M} w_i (y_i - q_n(x_i; c))^2. \tag{2.2}$$

Its solution is

$$\hat{c} = (A^T W A)^{-1} A^T W y. \tag{2.3}$$

This is the formulation we shall consider in this paper.

The polynomial space Π, where the approximant is sought, can be any properly defined polynomial spaces. Typically, a “degree” is associated with the space. Throughout this paper, we shall use $n \geq 1$ to denote such a “degree”. The cardinality of the polynomial space, p, thus depends on n. For example, for the well known total degree polynomial space

$$\mathcal{P}_n^d = \operatorname{span}\{x^\alpha = x_1^{\alpha_1} \cdots x_d^{\alpha_d} : |\alpha| \leq n\}, \tag{2.4}$$