

Indexing Spatio-Temporal Trajectories with Orthogonal Polynomials

Elena Braynova

Department of Computer Science, Worcester State College

Worcester, MA 01602

ebraynova@worcester@edu

Abstract—In this paper we consider d -dimensional spatio-temporal data ($d \geq 1$) and ways to approximate and index it. We focus on indexing such data for similarity matching using orthogonal polynomial approximations. There are many ways to choose an approximation scheme for d -dimensional spatio-temporal trajectories. Some of them have been studied before. In this paper we extend the approach proposed in [6] and show that not only Chebychev orthogonal polynomials but any orthogonal polynomial scheme satisfies Lower Bounding Lemma and, therefore, can be successfully used for approximating and indexing d -dimensional spatio-temporal trajectories. The basic result of the paper is Lower Bounding Lemma (a generalization of [6] result). That is, we prove that the Euclidean distance between two d -dimensional trajectories is lower bounded by Euclidean distance between the two vectors of orthogonal polynomial coefficients.

Keywords: *spatio-temporal data, approximating, indexing, orthogonal polynomials, data mining*

I. INTRODUCTION

Spatio-temporal objects appear in many applications. In general, a spatio-temporal object is an object described by d -attributes, where the attributes are functions of time. A d -dimensional spatio-temporal trajectory of an object is a sequence of the form $\{(t_1, v_1), (t_2, v_2), \dots, (t_n, v_n)\}$, where $t_i \in R$, $t_1 < t_2 < \dots < t_n$, v_i is a d -dimensional vector describing the object at time t_i , $i = 1, 2, \dots, n$, R is a set of real numbers. We can say that v_i represents the coordinates of the object at time t_i . The examples of spatio-temporal data are transportation, satellite and earth change data. This type data is usually modeled as multi-dimensional trajectories. 1-dimensional spatio-temporal trajectories are called time series and are used to represent stock prices, salaries histories and some other types of data. Many financial, medical, and scientific databases store time series data.

Working with spatio-temporal data is difficult. The corresponding databases are very large and grow rapidly. Multigigabyte databases are common. Typical examples are Weblog, Space Shuttle, and Macho databases. Macho database, for example, contains several terabytes of data and is updated with several gigabytes a day. Since most of the data lives on disk or tape, we need techniques to perform dimensionality reduction on the data, to index and represent it efficiently. There are different approaches to represent spatio-temporal data. One of the approaches is to approximate this data. Spatio-temporal data can be approximated using polynomials, rational

functions, Fourier transforms, splines, non-linear regressions, and etc. Some of these approaches have been studied before [1, 2, 3, 4, 5]. Most of proposed indexing techniques are based on piecewise approximations, where each piece is either constant or a linear function (polynomial of degree 1). We believe that most of the d -dimensional spatio-temporal data is generated by the processes that have non-linear nature and therefore should be better represented by polynomial or even rational functions. In this paper we focus on orthogonal polynomial approximations of d -dimensional trajectories. We believe that most the processes generated spatio-temporal data are smooth and continuous, so continuous type approximations should better represent the data, than piecewise discontinuous type approximations. We consider orthogonal polynomials as the ones that have small computational cost and minimize the maximum deviation from the true data values. Chebychev polynomial approximations are studied in [6]. The authors explored how to use Chebychev polynomials as a basis for approximating and indexing d -dimensional trajectories, and showed that this approximation scheme is almost identical to the *minimax* polynomials. In this paper we extend the ideas presented in [6]. We study orthogonal polynomial approximation scheme in general. We prove property similar to the Lower Bounding Lemma [6]. We show that any orthogonal polynomial schema is almost identical to the *minimax* polynomials and can be used for spatio-temporal data representation.

The paper is organized as follows. Section 2 recalls basic definitions and properties used in the paper. In section 3, we show how to approximate a time series using orthogonal polynomials in general. The closeness of orthogonal polynomials to the *minimax* polynomials is proved in Section 4. Section 5 generalizes the results for time series to d -dimensional spatio-temporal trajectories. Section 6 concludes the paper and discusses future work.

II. ORTHOGONAL POLYNOMIALS – BASIC DEFINITIONS, EXAMPLES, PROPERTIES

In this section we review basic definitions and properties of orthogonal polynomials used in the paper. We also recall a few well known types of orthogonal polynomials that, as it will be shown in the paper, can be used for approximating and indexing d -dimensional spatio-temporal data.

Definition 1: A system $\{f_i(t)\}$ of real-valued and almost everywhere defined on the interval $I = [x_0, x_1]$ functions is called *orthogonal system* on the interval I if the following condition holds true:

$$\int_I f_k(t) f_j(t) dt = X_k \delta_{k,j} \quad (1)$$

where $\delta_{k,j} = 1$ for $k = j$ and $\delta_{k,j} = 0$ otherwise, and $X_k \neq 0$ for all $k, k \in N$.

The system is called *orthogonal* and *normalized* if it satisfies (1) and $X_k=1$ for all $k, k \in N$. The two terms are usually reduced to the single term *orthonormal* or *orthonormalized*. A non-normalized system of orthogonal functions can be always normalized by dividing each $f_i(t)$ function by X_k . Systems of orthogonal functions are special cases of linear independent functions systems and can be used as a basis for the corresponding functional space. In fact, any system of m linear independent functions can be transformed into a system $\{f_i(t)\}$ of m orthogonal functions. So we can say that the number of orthogonal systems of a particular type is equal to the number of functional bases of the same type.

Definition 2: A system $\{p_i(t)\}$ is called *orthogonal polynomial system* on the interval I if

- i) $p_i(t)$ is a polynomial of degree i and
- ii) the system $\{p_i(t)\}$ satisfies (1)

If $\{p_i(t)\}$ is normalized, it is called *orthonormal polynomial system*. Let us recall a few well known and very often used for approximating orthogonal polynomials.

Definition 3: (Chebyshev polynomials) The Chebyshev polynomials are defined by the recurrence relation

$$P_m(t) = 2P_{m-1}(t) - P_{m-2}(t),$$

for all $m \geq 2$ with $P_0(t) = 1$ and $P_1(t) = t$.

From the above definition, the first few Chebyshev polynomials are:

$$\begin{aligned} P_0(t) &= 1 \\ P_1(t) &= t \\ P_2(t) &= 2t^2 - 1 \\ P_3(t) &= 4t^3 - 3t \\ P_4(t) &= 8t^4 - 8t^2 + 1 \end{aligned}$$

Definition 4: (Legendre polynomials) The Legendre polynomials are defined by the following formula:

$$L_n(t) = \frac{1}{2^n n!} \frac{d^n}{dt^n} (t^2 - 1)^n \quad (2)$$

Such defined Legendre polynomials are orthogonal, but not normalized. They can be easily normalized by $\sqrt{\frac{2n+1}{2}}$ factor. The Jacobi polynomials form a class of orthogonal polynomials which contains the Chebyshev as well as the Legendre polynomials. Their definition can be given by generalized formula (2).

Definition 5: (Jacobi polynomials) The Jacobi polynomials are defined by the following formula:

$$(1-t)^\alpha (1+t)^\beta P_n^{(\alpha, \beta)}(t) = \frac{(-1)^n}{2^n n!} \frac{d^n}{dt^n} [(1-t)^{n+\alpha} (1+t)^{n+\beta}]$$

where we suppose that $\alpha > -1, \beta > -1$.

Such defined Jacobi polynomials are orthogonal, but not normalized. They can be easily normalized by corresponding factors.

Definition 6: (Hermite polynomials) The Hermite polynomials are defined by the following formula:

$$H_n(t) = (-1)^n e^{t^2} \frac{d^n}{dt^n} e^{-t^2}$$

Any orthogonal system is a linear independent system and can be used as a base for approximating any function.

Definition 7: Let $F(t)$ be a function defined on the interval I and $\{f_i(t)\}$ be an orthogonal system defined on the same interval I . The series $\sum_{k=0}^{\infty} a_k f_k(t)$ is called a series expansion of $F(t)$ by the orthogonal system $\{f_i(t)\}$. The coefficients a_k are computed by the formula:

$$a_k = \int_I F(t) f_k(t) dt \quad (3)$$

$F_n(t) = \sum_{k=0}^n a_k f_k(t)$ is called an approximation of function $F(t)$ over system $\{f_i(t)\}$ of degree n . Approximating function $F(t)$ by $F_n(t)$ we will write: $F(t) \approx F_n(t)$.

If $F(t)$ is a polynomial of degree m and $\{f_i(t)\}$ is a system of orthogonal polynomials, then all approximations $F_n(t)$ of degree $n, n \geq m$, of $F(t)$ over $\{f_i(t)\}$ are exact:

$$F(t) = F_n(t).$$

In general, expanding function F over an orthogonal system $\{f_i(t)\}$, it is natural to ask *How well $F(t)$ is represented by its series expansion? How close are the values of $F(t)$ to the corresponding values of $F_n(t)$ -approximations?* The corresponding property is referred as convergence property of a series expansion to the expanded function. In our paper we do not use any convergence results. Let us just mention that convergence properties for even well known orthogonal polynomials and basic types functions are not trivial. There are still many open problems in Approximation Theory exploring whether a certain function can be expanded in a convergent to it series of a particular orthogonal system. As it was shown [7] such simple properties of a function as continuity or boundedness is not enough to answer the question. In this paper we use the following property.

Theorem 1: (Bessel Inequality) Given function $F(t)$ and orthogonal system $\{f_i(t)\}$, both defined on the interval I . Let $F(t)$ be expanded in a series of the orthogonal system $\{f_i(t)\}$:

$F(t) = \sum_{k=0}^{\infty} a_k f_k(t)$, where a_k are computed by (3). Then the following inequality holds:

$$\sum_{k=0}^{m-1} a_k^2 \leq \sum_{k=0}^{\infty} a_k^2 \leq \int_I F(t)^2 dt \quad (4)$$

Nowadays orthogonal polynomials are used for approximation and interpolation of data various nature. Most popular among orthogonal schemes are Chebyshev, Legendre, Jacobi, and Hermite polynomials. Chebyshev approximation scheme is studied in [6]. We attempt to extend the idea of using Chebyshev polynomials for indexing d -dimensional spatio-temporal data to any orthogonal polynomial scheme and show

that any of the orthogonal polynomial approximations (if have a reasonable computational cost) can be successfully used for indexing d -dimensional spatio-temporal data.

III. INDEXING TIME SERIES

In this section, without lost of generality, we focus on time series (1-dimensional spatio-temporal trajectories). In section 5, we generalize to case $d, d > 1$.

Given a set of time series of size N and a system $\{f_i(t)\}$ of orthogonal polynomials. We will approximate the time series by $\{f_i(t)\}$ -polynomials of degree n , where $n \ll N$. In the paper we focus on indexing time series data for similarity matching using Euclidean distance. Similarity search is not only useful in its own, to explore a given data, but it is an important element of many data mining applications such as clustering, classification, and mining association rules. In this paper we make the following assumption for time series data:

- Every time series has the same length, N (same-length assumption).
- Every time series occurs at the same set of time points $\{t_1, t_2, \dots, t_N\}$ (same-set assumption).

The *same-set* assumption includes *same-length* assumption. The assumptions are made to facilitate proofs of basic results. If the time series are not of same length, padding techniques may be used. If *same-set* assumption is not met, interpolation techniques may be applied. The basic results of the paper remain true without the above assumptions.

Given a time series $S = \{(t_1, v_1), (t_2, v_2), \dots, (t_N, v_N)\}$, where $t_i \in R, a = t_1 < t_2 < \dots < t_N = b, v_i \in R, i = 1, 2, \dots, N, I = [a, b]$. We construct function $S(t)$ of the form:

$S(t) = v_j$, if $t \in I_j$, $j = 1, \dots, N$ and intervals I_j are defined by the formular:

$$I_j = \begin{cases} [a, \frac{t_1+t_2}{2}], & j = 1 \\ [\frac{t_{j-1}+t_j}{2}, \frac{t_j+t_{j+1}}{2}], & 2 \leq j \leq N-1 \\ [\frac{t_{N-1}+t_N}{2}, b], & j = N \end{cases}$$

Function $S(t)$ is an interval function created based on the original time series S (a discrete function). Now we can apply (3) and compute a_k -coefficients of $\{f_i(t)\}$ -polynomial approximation of degree n for function $S(t)$. Namely

$$a_k = \int_I S(t) f_k(t) dt, \quad k = 1, \dots, n, n \leq N.$$

Thus, for any time series S and a system of orthogonal polynomials $\{f_i(t)\}$, we can compute a vector $A = (a_1, a_2, \dots, a_n)$ of size n , where a_k are coefficients of function $S(t)$ over system $\{f_i(t)\}$. We use vector A to index time series S reducing dimensionality from N (N points of S) to n (size of A), $n \leq N$ (in fact $n \ll N$).

IV. LOWER BOUNDING LEMMA

In this section we prove basic result of the paper, the Lower Bounding Lemma (generalization of Lower Bounding Lemma in [6]). We use the following notation. Given $A = (a_1, a_2, \dots, a_n)$ and $B = (b_1, b_2, \dots, b_n)$, two vectors of

size n , then Euclidean distance $D_E(A, B)$ between them is computed by the formula: $D_E(A, B) = \sqrt{\sum_{k=1}^n (a_k - b_k)^2}$

Euclidean distance is simple, natural for many d -dimensional spatio-temporal applications, and is frequently used for similarity search. Now let us establish the Lower Bounding Lemma.

Theorem 2: (Lower Bounding Lemma). Given two time series S_1 and S_2 . C_1 and C_2 are corresponding vectors of $\{f_i(t)\}$ -orthogonal polynomial approximation of degree n for S_1 and S_2 respectively. Then

$$D_E(C_1, C_2) \leq D_E(S_1, S_2),$$

where $D_E(A, B)$ is Euclidean distance between vectors A and B .

Proof: Given $S_1 = \{(t_1, v_1), (t_2, v_2), \dots, (t_n, v_n)\}$ and $S_2 = \{(t_1, w_1), (t_2, w_2), \dots, (t_n, w_n)\}$. It is clear that the Euclidean distance between S_1 and S_2 satisfies the following equality:

$$D_E^2(S_1, S_2) = \sum_{k=1}^N (v_k - w_k)^2 \quad (5)$$

Let the interval functions corresponding to S_1 and S_2 be f_1 and f_2 . Let C_1 and C_2 have the form: $C_1 = (a_1, a_2, \dots, a_n)$ and $C_2 = (b_1, b_2, \dots, b_n)$.

Then $D_E^2(C_1, C_2) = \sum_{k=1}^n (a_k - b_k)^2 \quad (6)$

Comparing the distances from (5) and (6) we have $D_E^2(C_1, C_2) = \sum_{k=1}^n (a_k - b_k)^2 \leq \sum_{k=1}^{\infty} (a_k - b_k)^2 \leq \int_I (f_1 - f_2)^2(t) dt \quad (7)$

But, at the other hand $\int_I (f_1 - f_2)^2(t) dt =$

$$\sum_{k=1}^N \int_{I_k} \frac{(v_k - w_k)^2}{|I_k|} dt = \sum_{k=1}^N (v_k - w_k)^2 \frac{|I_k|}{|I_k|} = \sum_{k=1}^N (v_k - w_k)^2 = D_E^2(S_1, S_2) \quad (8)$$

From (7) and (8) we obtain $D_E(C_1, C_2) \leq D_E(S_1, S_2)$. ■

V. GENERALIZATION OF LOWER BOUNDING LEMMA FOR

MULTI-DIMENSIONAL TRAJECTORIES

The above result can be easily extended to d -dimensional spatio-temporal trajectories. Let $S = \{(t_1, v_1), (t_2, v_2), \dots, (t_N, v_N)\}$ be a d -dimensional spatio-temporal trajectory, where v_i are d -dimensional vectors. We can decompose S into d 1-dimensional series S_1, S_2, \dots, S_d . Each of these 1-dimensional series can be approximated by C_i vectors, where C_i is a vector of $\{f_i(t)\}$ -orthogonal polynomial approximation of degree n for S_i . We consider the following generalization of Euclidean distance [6] between two d -dimensional spatio-temporal trajectories.

Definition 8: Let S, R be d -dimensional spatio-temporal trajectories. Let their vectors of $\{f_i(t)\}$ -orthogonal polynomial approximations of degree n are $C = [C_1, C_2, \dots, C_d]$ and $D = [D_1, D_2, \dots, D_d]$ respectively. Define:

$$D_E(C, D) = \sqrt{\sum_{k=1}^d D_E^2(C_k, D_k)}$$

Then it is easy to show that the following generalization of Lower Bounding Lemma is true.

Corollary 1: Let S, R be d -dimensional spatio-temporal trajectories and C and D be the corresponding $\{f_i(t)\}$ -orthogonal polynomial approximations. Then

$$D_E(C, D) \leq D_E(S, R)$$

VI. CONCLUSION AND FUTURE WORK

In this paper, we explore how to use orthogonal polynomials for approximating and indexing d -dimensional spatio-temporal data. We prove Lower Bounding Lemma for time series and generalize it for d -dimensional case. We show that the property is true not only for Cheychev polynomials, but for any orthogonal polynomial scheme. The Lemma guarantees no false negatives in using any system of orthogonal polynomials for indexing as a filter. As tighter the lower bound as less is the number of false positives. The degree of orthogonal approximations is usually much less than size of d -dimensional trajectories ($n \ll N$) and cost to compute basic types orthogonal polynomials is low. In the paper we consider Euclidean distance. For future work we consider to investigate more advanced distance functions, such as time-warping [8] and longest common subsequence [9]. It is also interesting to study non -Euclidean distances and multi-variable orthogonal polynomials for approximating and indexing d -dimensional spatio-temporal data ($d \geq 1$). In section 5 we generalized our results using 1-dimensional projections of d -dimensional data. It is clear that projections are not always best approximations and multi-variable approximation schemes might be better in some cases. We also plan to conduct experiments and compare the performance of the basic orthogonal polynomial systems for approximating and indexing d -dimensional spatio-temporal data.

VII. REFERENCES

- [1] R. Agrawal, K. Lin, H. Sawhney and K. Shim, fast Similarity Search in the Presence of Noise, Scaling and Translation in Time-serie databases, In Proceeding of VLDB, 1995
- [2] C. Faloutsos, M. ranganathan and Y. Manolopoulos, Fast Subsequence Matching in Time-Series Databases, In Proceeding of SIGMOD, 1994.
- [3] D. Rafiei and A. Mendelzon, Efficient Retrieval of Similar Time Sequences using DFT, In Proceeding of FODO, 1998.
- [4] E. Keogh, K. Chakrabarti, M. Pazzani and S. Mehrotra, Locally Adaptive Dimensionality Reduction for Indexing Large Time Series Databases, In Proceeding of SIGMOD, 2001.
- [5] K. Chan and A. Fu, Efficient Time Series matching by Wavelets, Proceeding of ICDE, 1999.

- [6] Y.Cai, R.Ng, Indexing Spatio-Temporal Trajectories with Chebychev Polynomilas, In Proceeding of SIGMOD, 2004.
- [7] E.T.Whittaker and G.N.Watson, A course of modern analysis, Oxford U. Press, 1952
- [8] D.J. Berndt and J. Clifford, Using Dynamic time warping to find patterns in time series, Working Notes of the Knowledge Discovery in Databases Workshop, 1994.
- [9] M. Vlachos, G. Kollios and D. Gunopulos, Discovering similar multidimensional trajectories, In Proceeding of ICDE, 2002.