The Christoffel function for supervised learning: theory and practice

Srećko Đurašinović

▶ **To cite this version:**

Srećko Đurašinović. The Christoffel function for supervised learning: theory and practice. Machine Learning [stat.ML]. 2022. hal-03768886

HAL Id: hal-03768886
https://laas.hal.science/hal-03768886

Submitted on 5 Sep 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Master 2 Applied Mathematics, Statistics

Mathematics and Economic Decision

The Christoffel function for supervised learning: theory and practice

Internship report

Student:

Srećko Đurašinović

Supervisors:

Victor Magron

Full-Time Researcher, LAAS-CNRS

Jean-Bernard Lasserre

Emeritus Research Director, LAAS-CNRS

Keywords: Christoffel-Darboux Kernel, Polynomial Optimization, Machine Learning, Data Analysis

July, 2022

Summary

Notations 4
Introduction 5

1 Christoffel function - Selected properties **7**
1.1 Preliminaries 7
1.2 Reproducing Kernel Hilbert Spaces 8
1.3 Christoffel-Darboux kernel 9
1.4 Christoffel function 11
1.4.1 Some properties 13

2 Application in machine learning **17**
2.0.1 Introductory remarks 17
2.1 Supervised learning 17
2.2 Christoffel-Darboux classifier 18
2.2.1 Consistency result 20
2.3 Empirical evaluations 22
2.3.1 *Iris* data 23
2.3.2 Toy data sets 24
2.4 Extensions 26
Conclusion 28
Bibliography 29

Notations

- $\mathcal{M}(K), \mathcal{M}_+(K), \mathcal{P}(K)$ - respectively, set of finite, signed, regular Borel measures, set of positive finite measures, set of probabilities on compact set K
- $\mathcal{C}(K, K'), \mathcal{C}_b(K, K'), \mathcal{C}_0(K, K')$ - respectively, set of continuous, continuous and bounded, continuous vanishing at infinity functions from K to K'. If the functions are real-valued, we denote them simply by $\mathcal{C}(K)$ (resp. $\mathcal{C}_b(K), \mathcal{C}_0(K)$)
- $\mathcal{C}(K)^*$ - set of real-valued and continuous on $\mathcal{C}(K)$
- $\mathbb{R}^{n \times p}$ - set of matrices with $n \in \mathbb{N} \setminus \{0\}$ lines, $p \in \mathbb{N} \setminus \{0\}$ columns, with real coefficients/entries
- $\boldsymbol{I}_n$ - identity matrix of size $n \times n,\ n \in \mathbb{N} \setminus \{0\}$
- X^Y - space of X-valued functions defined on Y
- $\mathrm{Lip}_L(X, Y)$ - set of L-Lipschitz maps from X to Y
- $A^{\mathbb{N}}$ - set of sequences taking values in the set A
- $B(x, r)$ - closed Euclidean ball of radius $r > 0$, centered at x
- $\mathcal{B}(X)$ - Borel sigma-algebra on X

- $\langle ., . \rangle_{X \times Y}$ - generic pairing (bi-linear form) between the spaces X and Y (the subscript is usually omitted if X and Y are easily inferred)
- $[\![k, k']\!]$ - interval containing all integers from k to k' (included)
- $A \times B$ - Cartesian product of the sets A and B
- $\|.\|$ - generic norm on some normed vector space
- If M is a matrix, then $||M||_\infty = \sup_{||x||_\infty = 1} ||Mx||_\infty$ is the matrix $\|\|_\infty$ norm
- M^T - transposition of a matrix M
- $|.|$ - absolute value
- d_X - distance on some metric space X

- δ_a - the Dirac mass concentrated at the point a
- δ_{ij} - Kronecker delta
- $\mathbf{1}_A$ - Indicator function (equal to 1 if A is realised, and to 0 otherwise)
- $\wedge, \vee$ - the min and max operators, i.e. $a \wedge b := \min\{a, b\}$ and $a \vee b := \max\{a, b\}$

- id - the identity map
- 1 - the constant map equal to 1
- 0 - the constant map equal to 0
- π_X - projection on the set X
- $\mathrm{diam}(X)$ - diameter of a set X, i.e. $\mathrm{diam}(X) : \sup_{x_1, x_2 \in X} d_X(x_1, x_2)$
- $\mathrm{val}(P)$ - optimal value of the optimization program (P)
- $\binom{n}{p} = \frac{n!}{p!(n-p)!}$ - Binomial coefficient

- $\mathbb{R}[\boldsymbol{x}]$ - the ring of real-valued multivariate polynomials in variables $\boldsymbol{x} = (x_1, \ldots, x_p)$. The finite dimensional vector space consisting of polynomials of degree at most $d \in \mathbb{N}$ is denoted by $\mathbb{R}_d[\boldsymbol{x}] \subset \mathbb{R}[\boldsymbol{x}]$.
- $\mathbb{N}^p_d := \{\boldsymbol{\alpha} \in \mathbb{N}^p \ \mid \ |\boldsymbol{\alpha}| := \sum_{i=1}^p \alpha_i \leq d\}$ - the set of multi-indices, with cardinal $s(d) = \binom{p+d}{p}$.

Introduction

Since its inception, machine learning has provided us with many different methods and approaches for handling classification problems. Interested readers can find in [10, 12] a non-exhaustive overview of these methods. In this report, we introduce an additional tool - Christoffel-Darboux kernel - and we exploit its properties, mainly in the context of supervised learning.

Christoffel-Darboux kernel is a very well known tool among the researchers working on the approximation theory. For a long time, its properties were only studied in this context, focusing on the relationship with orthogonal polynomials [9] that can be used for interpolation, .

However, recent works [1, 2, 4, 11, 13] have shed a very different light on this particular kernel by showing its appealing potential in the realm of data science. Indeed, some of the most important features of the Christoffel function can be easily connected to the classical problems in statistics and machine learning: support estimation, outlier detection, graph recovery, free probabilities and many others.

An additional task of equally great importance in data science - supervised learning - can be tackled in this context. We argue that the Christoffel function provides a very intuitive way for solving classification problems, while displaying quite good theoretical properties at the same time [4, 5, 6].

This report will be organised as follows: in the first part, we will introduce the Christoffel function and study its properties that can be useful for solving problems that arise in data science. Then, we will particularly focus on the problem of supervised learning and present some theoretical guarantees that could justify the use of the Chirstoffel function in this context. Finally, the last part will be devoted to some empirical evaluations and comparisons with some state-of-art methods. Additionally, directions for further research and generalizations will be suggested.

Chapter 1

Christoffel function - Selected properties

1.1 Preliminaries

Let us start by introducing the notations that will be used throughout this report.
We let $\boldsymbol{v}_d(\boldsymbol{x}) := (\boldsymbol{x}^{\boldsymbol{\alpha}})_{\boldsymbol{\alpha}\in\mathbb{N}^p_d}$ denote the monomial basis of the vector space of $\mathbb{R}[\mathbf{x}]_d$ (we suppose that the elements of this basis are arranged in the *graded lexicographic order*[1]).

Definition 1.1.1 (Moment matrix)**.** [7] Let $\boldsymbol{\Omega} \subset \mathbb{R}^p$ be compact, with non-empty interior. Let ϕ be a Borel measure supported on $\boldsymbol{\Omega}$. The moment matrix of order (degree) $d \in \mathbb{N}$ associated to ϕ, denoted by $\boldsymbol{M}_d(\phi)$, is a $s(d) \times s(d)$ dimensional real symmetric matrix, whose rows and columns are indexed by monomials in $\boldsymbol{v}_d(\boldsymbol{x})$, and whose entry $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ is given by

$$\boldsymbol{M}_d(\phi)(\boldsymbol{\alpha}, \boldsymbol{\beta}) := \int_{\boldsymbol{\Omega}} \boldsymbol{x}^{\boldsymbol{\alpha}+\boldsymbol{\beta}} d\phi, \qquad \text{for any } \boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{N}^p_d. \tag{1.1}$$

For the purpose of illustration, with $p = d = 2$, we would obtain the following matrix in $\mathbb{R}^{6\times 6}$:

$$\boldsymbol{M}_2(\phi) = \begin{bmatrix} \int_{\boldsymbol{\Omega}} 1 d\phi & \int_{\boldsymbol{\Omega}} x_1 d\phi & \int_{\boldsymbol{\Omega}} x_2 d\phi & \int_{\boldsymbol{\Omega}} x_1^2 d\phi & \int_{\boldsymbol{\Omega}} x_1 x_2 d\phi & \int_{\boldsymbol{\Omega}} x_2^2 d\phi \\ . & \int_{\boldsymbol{\Omega}} x_1^2 d\phi & . & . & . & . \\ . & . & \int_{\boldsymbol{\Omega}} x_2^2 d\phi & . & . & . \\ . & . & . & \int_{\boldsymbol{\Omega}} x_1^4 d\phi & . & . \\ . & . & . & . & \int_{\boldsymbol{\Omega}} x_1^2 x_2^2 d\phi & . \\ \int_{\boldsymbol{\Omega}} x_2^2 d\phi & \int_{\boldsymbol{\Omega}} x_1 x_2^2 d\phi & \int_{\boldsymbol{\Omega}} x_2^3 d\phi & \int_{\boldsymbol{\Omega}} x_1^2 x_2^2 d\phi & \int_{\boldsymbol{\Omega}} x_1 x_2^3 d\phi x_2 & \int_{\boldsymbol{\Omega}} x_2^4 d\phi \end{bmatrix}. \tag{1.2}$$

If we consider $\boldsymbol{v}_d(\boldsymbol{x})$ to be a column vector of $\mathbb{R}^{s(d)}$, then we can write the moment matrix in a more compact form as

$$\boldsymbol{M}_d(\phi) = \int_{\boldsymbol{\Omega}} \boldsymbol{v}_d(\boldsymbol{x}) \boldsymbol{v}_d(\boldsymbol{x})^T d\phi, \tag{1.3}$$

where the integral should be understood component-wise.
Finally, $\boldsymbol{v}_d(\boldsymbol{x})$ can denote a more general polynomial basis, not necessarily the monomial one. The meaning will always be made clear from the context.

[1] To order monomials, we first compares their total degree, i.e. the sum of all their exponents, and in case of a tie, we apply *lexicographic* order. This one, in turn, first compares the exponents of x_1 in the monomials, and if they are the same, then the exponents of x_2 are compared, and so forth.

1.2 Reproducing Kernel Hilbert Spaces

In order to better describe the ambient space and the objects that we will manipulate throughout this report, let us briefly recall the idea of Reproducing Kernel Hilbert Spaces (RKHS).

Definition 1.2.1 (Kernel)**.** Let X be a set. A kernel is any symmetric map $k : X \times X \to \mathbb{R}$, i.e. $k(\boldsymbol{x}, \boldsymbol{y}) = k(\boldsymbol{y}, \boldsymbol{x})$, for all $\boldsymbol{x}, \boldsymbol{y} \in X$. If, in addition, the matrix $(k(\boldsymbol{x_i}, \boldsymbol{x_j}))_{i,j=1}^n$ is positive semi-definite for all $\boldsymbol{x_1}, \ldots, \boldsymbol{x_n} \in X$, the kernel k is called positive semi-definite.

Notice that, by definition, a kernel k is positive semi-definite if $\sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j k(\boldsymbol{x_i}, \boldsymbol{x_j}) \geq 0$ for all $n \in \mathbb{N}, \boldsymbol{x_1}, \ldots, \boldsymbol{x_n} \in X$, and $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$.

Definition 1.2.2 (RKHS)**.** For $X \subset \mathbb{R}^p$, let $(H, \langle ., . \rangle_H)$ be the Hilbert space of real-valued functions $f : X \to \mathbb{R}$ defined on X. We say that H is a Reproducing Kernel Hilbert Space if the evaluation functional $L_{\boldsymbol{x}}$ is continuous on H, i.e., if for any $\boldsymbol{x} \in X$, the functional $H \ni f \mapsto L_{\boldsymbol{x}}(f) := f(\boldsymbol{x})$ is continuous.

We can now connect explicitly the notions from the last two definitions.

Definition 1.2.3 (Reproducing kernel)**.** A kernel k is called *reproducing kernel* if it satisfies the *reproducing property*:

$$\forall \boldsymbol{x} \in X, \forall f \in H, L_{\boldsymbol{x}}(f) = \langle k(., \boldsymbol{x}), f \rangle_H = f(\boldsymbol{x}). \tag{1.4}$$

Theorem 1.2.1. [4, 8] For $X \subset \mathbb{R}^p$, let H be the Hilbert space of real-valued functions defined on X. Let $k : X \times X \to \mathbb{R}$ be a kernel.

1. If H is a RKHS, then it admits a positive semi-definite reproducing kernel.

2. If k is positive semi-definite, there exists (a unique) RKHS with k as its reproducing kernel.

Proof. Let X, H and k be defined as in the statement of the theorem.

1. Let $\boldsymbol{x}, \boldsymbol{y} \in X$. Since H is RKHS, then, by definition, the linear functional $L_{\boldsymbol{x}}$ is continuous. Hence, the classical Riesz representation theorem ensures the there exists some $k_{\boldsymbol{x}} \in H$ such that $L_{\boldsymbol{x}}(f) = \langle f, k_{\boldsymbol{x}} \rangle_H$ for all $f \in H$. Using the same arguments, there exists $k_{\boldsymbol{y}} \in H$ such that $L_{\boldsymbol{y}}(f) = \langle f, k_{\boldsymbol{y}} \rangle_H$ for all $f \in H$. Let us define on the product space $X \times X$ the mapping k by $k(\boldsymbol{x}, \boldsymbol{y}) = \langle k_{\boldsymbol{x}}, k_{\boldsymbol{y}} \rangle_H$. By construction, k satisfies the reproducing property. Moreover, for any $n \in \mathbb{N}, \boldsymbol{x_1}, \ldots, \boldsymbol{x_n} \in X$, and $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$, we have

$$\sum_{i,j=1}^n \alpha_i \alpha_j k(\boldsymbol{x_i}, \boldsymbol{x_j}) = \sum_{i,j=1}^n \alpha_i \alpha_j \langle k_{\boldsymbol{x_i}}, k_{\boldsymbol{x_j}} \rangle_H \tag{1.5}$$

$$= \left\langle \sum_{i=1}^n \alpha_i k_{\boldsymbol{x_i}}, \sum_{i=1}^n \alpha_i k_{\boldsymbol{x_i}} \right\rangle_H = \left\| \sum_{i=1}^n \alpha_i k_{\boldsymbol{x_i}} \right\|_H^2 \geq 0, \tag{1.6}$$

   proving that k is positive semi-definite.

2. Consider the mapping $\phi : X \to \mathbb{R}^X$ such that $\phi(\boldsymbol{x}) = k(., \boldsymbol{x})$ for all $\boldsymbol{x} \in X$. Then, we let $V := \text{vect}(\{\phi(\boldsymbol{x}) \mid \boldsymbol{x} \in X\}) = \{f(.) = \sum_{i=1}^n \alpha_i k(., \boldsymbol{x_i}) \mid n \in \mathbb{N}, \alpha_i \in \mathbb{R}, \boldsymbol{x_i} \in X\}$. Moreover,

we can define for two elements in V, say $f(.) = \sum_{i=1}^{n_f} \alpha_i k(., \boldsymbol{x_i})$ and $g(.) = \sum_{i=1}^{n_g} \beta_i k(., \boldsymbol{y_i})$ the following form:

$$\langle f, g \rangle = \sum_{i=1}^{n_f} \sum_{j=1}^{n_g} \alpha_i \beta_j k(\boldsymbol{x_i}, \boldsymbol{y_j}).$$

The assumptions about k ensure that $\langle , \rangle$ is a valid scalar product. The closure of V would then form a RKHS, since for any $f \in \text{closure}(V)$,

$$|f(\boldsymbol{x})| = |L_{\boldsymbol{x}}(f)| = |\langle f, k(., \boldsymbol{x}) \rangle| \leq ||k(., \boldsymbol{x})||_H ||f||_H, \tag{1.7}$$

where the last inequality follows from the Cauchy-Schwartz inequality.

□

Thus, we have managed to obtain a Hilbert space satisfying some additional reproducing properties. This additional space structure can also be obtained with Christoffel-Darboux kernels that we will introduce in a sequel. It turns out that these features can be exploited in the context of machine learning.

1.3 Christoffel-Darboux kernel

In the spirit of the previous section, one may see the space $\mathbb{R}_d[\mathbf{x}]$ (which is a subspace of $L^2(\phi)$) as the RKHS.
If, for instance, the measure ϕ is compactly supported and absolutely continuous with respect to the Lebesgue measure, then we can show that the following bi-linear form

$$\langle , \rangle_\phi \colon \mathbb{R}_d[\boldsymbol{x}] \times \mathbb{R}_d[\boldsymbol{x}] \longrightarrow \mathbb{R} \tag{1.8}$$

$$(P, Q) \longmapsto \langle P, Q \rangle_\phi := \int_{\boldsymbol{\Omega}} PQ d\phi \tag{1.9}$$

defines a valid scalar product, making $(\mathbb{R}_d[\boldsymbol{x}], \langle , \rangle_\phi)$ a finite-dimensional Hilbert space of polynomial functions from $\mathbb{R}^p$ to $\mathbb{R}$ [4]. Since every linear map defined on a finite-dimensional space is continuous, we deduce that $(\mathbb{R}_d[\boldsymbol{x}], \langle , \rangle_\phi)$ is indeed a RKHS. Its associated reproducing kernel is known as the Christoffel-Darboux kernel. In order to define it properly, let us recall the next definition:

Definition 1.3.1. Let $(P_{\boldsymbol{\alpha}})_{\alpha \in \mathbb{N}^p}$ be a family of polynomials in $\mathbb{R}[\boldsymbol{x}]$. We say that this family is *orthonormal with respect to* ϕ if

$$\int_{\boldsymbol{\Omega}} P_{\boldsymbol{\alpha}}(\boldsymbol{x}) P_{\boldsymbol{\beta}}(\boldsymbol{x}) d\phi(\boldsymbol{x}) = \mathbf{1}_{\boldsymbol{\alpha}=\boldsymbol{\beta}}, \qquad \forall \boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{N}^p. \tag{1.10}$$

Definition 1.3.2 (Christoffel-Darboux kernel)**.** The *Christoffel-Darboux kernel* is the reproducing kernel associated to $(\mathbb{R}_d[\boldsymbol{x}], \langle , \rangle_\phi)$. It can be defined via the following relationship

$$K_d^\phi(\boldsymbol{x}, \boldsymbol{y}) := \sum_{i=1}^{s(d)} P_i(\boldsymbol{x}) P_i(\boldsymbol{y}), \quad \text{for all} \quad \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^p, \tag{1.11}$$

where $(P_i)_{i=1}^{s(d)}$ is an orthonormal basis of $\mathbb{R}_d[\boldsymbol{x}]$.

Remark 1.3.1. The kernel K_d^ϕ satisfies the reproducing property. Indeed, if $P \in \mathbb{R}_d[\mathbf{x}]$, then one can write $P = \sum_{i=1}^{s(d)} p_i P_i$, for some real numbers $(p_i)_{i=1}^{s(d)}$, implying that

$$\forall \boldsymbol{x} \in X, L_{\boldsymbol{x}}(P) = \langle K_d^\phi(., \boldsymbol{x}), P\rangle_\phi \tag{1.12}$$

$$= \int_{\boldsymbol{\Omega}} P(\boldsymbol{y}) K_d^\phi(\boldsymbol{y}, \boldsymbol{x}) d\phi(\boldsymbol{y}) = \int_{\boldsymbol{\Omega}} P(\boldsymbol{y}) \sum_{i=1}^{s(d)} P_i(\boldsymbol{x}) P_i(\boldsymbol{y}) d\phi(\boldsymbol{y}) \tag{1.13}$$

$$= \sum_{i=1}^{s(d)} P_i(\boldsymbol{x}) \int_{\boldsymbol{\Omega}} P(\boldsymbol{y}) P_i(\boldsymbol{y}) d\phi(\boldsymbol{y}) = \sum_{i=1}^{s(d)} P_i(\boldsymbol{x}) \sum_{j=1}^{s(d)} p_j \underbrace{\int_{\boldsymbol{\Omega}} P_j(\boldsymbol{y}) P_i(\boldsymbol{y}) d\phi(\boldsymbol{y})}_{\mathbf{1}_{j=i}} \tag{1.14}$$

$$= \sum_{i=1}^{s(d)} p_i P_i(\boldsymbol{x}) = P(\boldsymbol{x}). \tag{1.15}$$

An interesting feature of this kernel is that it can be explicitly computed from the moments, i.e., from the moment matrix. Indeed, if we let $\boldsymbol{v}_d(\boldsymbol{x}) := (P_1(\boldsymbol{x}), \ldots, P_{s(d)}(\boldsymbol{x})) \in \mathbb{R}^{s(d)}$ for some basis $(P_i)_{i=1}^{s(d)}$ of $\mathbb{R}_d[\boldsymbol{x}]$ (not necessarily orthonormal), then for any polynomial $P \in \mathbb{R}_d[\boldsymbol{x}]$ there exists some $\boldsymbol{p} \in \mathbb{R}^{s(d)}$ such that $P(\boldsymbol{x}) = \boldsymbol{p}^T \boldsymbol{v}_d(\boldsymbol{x})$ for any $\boldsymbol{x} \in \mathbb{R}^p$.
In particular, since $\langle ., . \rangle_\phi$ is a valid scalar product, then for any $P \not\equiv 0$, we have

$$\langle P, P\rangle_\phi = \int_{\boldsymbol{\Omega}} \boldsymbol{p}^T \boldsymbol{v}_d(\boldsymbol{x}) (\boldsymbol{p}^T \boldsymbol{v}_d(\boldsymbol{x}))^T d\phi(\boldsymbol{x}) = \boldsymbol{p}^T \left(\int_{\boldsymbol{\Omega}} \boldsymbol{v}_d(\boldsymbol{x}) \boldsymbol{v}_d(\boldsymbol{x})^T d\phi(\boldsymbol{x}) \right) \boldsymbol{p} = \boldsymbol{p}^T \boldsymbol{M}_d(\phi) \boldsymbol{p} > 0, \tag{1.16}$$

implying that $\boldsymbol{M}_d(\phi) \succ 0$. This allows us to derive another expression for the Christoffel-Darboux kernel, based on the moments. Namely,

$$K_d^\phi(\boldsymbol{x}, \boldsymbol{y}) := \boldsymbol{v}_d(\boldsymbol{x})^T \boldsymbol{M}_d(\phi)^{-1} \boldsymbol{v}_d(\boldsymbol{y}), \qquad \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^p. \tag{1.17}$$

That this new formulation satisfies the reproducing property is guaranteed by the uniqueness of the reproducing kernel, but it can also been seen from the following computation: $\forall \boldsymbol{x} \in X, \forall \boldsymbol{p} \in \mathbb{R}^{s(d)}$,

$$\langle K_d^\phi(., \boldsymbol{x}), P\rangle_\phi = \int_{\boldsymbol{\Omega}} P(\boldsymbol{y}) K_d^\phi(\boldsymbol{y}, \boldsymbol{x}) d\phi(\boldsymbol{y}) = \int_{\boldsymbol{\Omega}} \boldsymbol{p}^T \boldsymbol{v}_d(\boldsymbol{y}) \boldsymbol{v}_d(\boldsymbol{y})^T \boldsymbol{M}_d(\phi)^{-1} \boldsymbol{v}_d(\boldsymbol{x}) d\phi(\boldsymbol{y}) \tag{1.18}$$

$$= \boldsymbol{p}^T \underbrace{\int_{\boldsymbol{\Omega}} \boldsymbol{v}_d(\boldsymbol{y}) \boldsymbol{v}_d(\boldsymbol{y})^T d\phi(\boldsymbol{y}) \boldsymbol{M}_d(\phi)^{-1}}_{\boldsymbol{I}_{s(d)}} \boldsymbol{v}_d(\boldsymbol{x}) \tag{1.19}$$

$$= \boldsymbol{p}^T \boldsymbol{v}_d(\boldsymbol{x}) = P(\boldsymbol{x}). \tag{1.20}$$

Finally, we can observe that $\boldsymbol{M}_d(\phi) = (\langle P_i, P_j\rangle_\phi)_{i,j=1}^{s(d)}$, so that, if the polynomial basis is orthonormal, we obtain $\boldsymbol{M}_d(\phi) = \boldsymbol{I}_{s(d)}$, which shows explicitly that expressions (1.11) and (1.17) coincide.

Finally, let us mention that changing the polynomial basis does not change the Christoffel-Darboux kernel, which is summarized in the next proposition.

Proposition 1.3.1. Let $\boldsymbol{v}_d(.)$ and $\boldsymbol{w}_d(.)$ be two different polynomial bases of $\mathbb{R}_d[\boldsymbol{x}]$. Let $\boldsymbol{M}_d(\phi)$ and $\boldsymbol{N}_d(\phi)$ be defined as in (1.3), with $\boldsymbol{v}_d(.)$ and $\boldsymbol{w}_d(.)$ respectively. Then, the associated Christoffel-Darboux kernels are equal.

Proof. Since $\boldsymbol{v}_d(.)$ and $\boldsymbol{w}_d(.)$ are two different basis of the vector space $\mathbb{R}_d[\boldsymbol{x}]$, there exists a positive definite matrix $\Pi \in \mathbb{R}^{s(d)\times s(d)}$ such that $\boldsymbol{w}_d(.) = \Pi \boldsymbol{v}_d(.)$. Let $\boldsymbol{x} \in \mathbb{R}^p$. Then,

$$\boldsymbol{w}_d(\boldsymbol{x})^T(\boldsymbol{N}_d(\phi))^{-1}\boldsymbol{w}_d(\boldsymbol{x}) = \boldsymbol{w}_d(\boldsymbol{x})^T \left(\int_{\boldsymbol{\Omega}} \boldsymbol{w}_d(\boldsymbol{x})\boldsymbol{w}_d(\boldsymbol{x})^T d\phi(\boldsymbol{x})\right)^{-1} \boldsymbol{w}_d(\boldsymbol{x}) \tag{1.21}$$

$$= \boldsymbol{v}_d(\boldsymbol{x})^T\Pi^T \left(\int_{\boldsymbol{\Omega}} \Pi\boldsymbol{v}_d(\boldsymbol{x})\boldsymbol{v}_d(\boldsymbol{x})^T\Pi^T d\phi(\boldsymbol{x})\right)^{-1} \Pi\boldsymbol{v}_d(\boldsymbol{x}) \tag{1.22}$$

$$= \boldsymbol{v}_d(\boldsymbol{x})^T\Pi^T \left(\Pi\int_{\boldsymbol{\Omega}} \boldsymbol{v}_d(\boldsymbol{x})\boldsymbol{v}_d(\boldsymbol{x})^T d\phi(\boldsymbol{x})\Pi^T\right)^{-1} \Pi\boldsymbol{v}_d(\boldsymbol{x}) \tag{1.23}$$

$$= \boldsymbol{v}_d(\boldsymbol{x})^T\Pi^T \left(\Pi\boldsymbol{M}_d(\phi)\Pi^T\right)^{-1} \Pi\boldsymbol{v}_d(\boldsymbol{x}) \tag{1.24}$$

$$= \boldsymbol{v}_d(\boldsymbol{x})^T\Pi^T(\Pi^T)^{-1}(\boldsymbol{M}_d(\phi))^{-1}\Pi^{-1}\Pi\boldsymbol{v}_d(\boldsymbol{x}) \tag{1.25}$$

$$= \boldsymbol{v}_d(\boldsymbol{x})^T(\boldsymbol{M}_d(\phi))^{-1}\boldsymbol{v}_d(\boldsymbol{x}). \tag{1.26}$$

□

What seems to distinguish the Christoffel-Darboux kernel from other kernels commonly used in practice is its profound connection with an underlying measure ϕ. Indeed, this measure is inducing an inner product on the ambient Reproducing Kernel Hilbert Space, which is suitable for many problems in statistics. Other commonly used kernels (polynomial kernel) are not induced by a positive measure.

1.4 Christoffel function

In the previous section, the Christoffel-Darboux kernel was introduced. Using the same notations, we can define the Christoffel function, the very central object of this report.

Definition 1.4.1 (Christoffel function)**.** The *Christoffel function* of order (degree) $d \in \mathbb{N}$, denoted by Λ_d^ϕ, is defined via the following relationship

$$\Lambda_d^\phi \colon \mathbb{R}^p \longrightarrow \mathbb{R}_+ \tag{1.27}$$

$$\boldsymbol{x} \longmapsto \Lambda_d^\phi(\boldsymbol{x}) := \min_{P\in\mathbb{R}_d[\boldsymbol{x}]} \left\{\int_{\boldsymbol{\Omega}} P^2(\boldsymbol{z})d\phi(\boldsymbol{z}),\ \ P(\boldsymbol{x}) = 1\right\}. \tag{1.28}$$

In other words, to any point $\boldsymbol{x} \in \mathbb{R}^p$, the Christoffel function associates the minimal squared $||.||_\phi$-norm among the polynomials whose value at $\boldsymbol{x}$ is equal to one. This constrained optimization problem is in fact equivalent to a particular quadratic programming problem. Indeed, recalling that any polynomial $P \in \mathbb{R}_d[\boldsymbol{x}]$ can be written as $\boldsymbol{p}^T\boldsymbol{v}_d(.)$, we deduce that the objective function becomes $\int_{\boldsymbol{\Omega}} \boldsymbol{p}^T\boldsymbol{v}_d(\boldsymbol{z})\boldsymbol{v}_d(\boldsymbol{z})^T\boldsymbol{p}d\phi(\boldsymbol{z}) = \boldsymbol{p}^T\boldsymbol{M}_d(\phi)\boldsymbol{p}$, so that

$$\Lambda_d^\phi(\boldsymbol{x}) := \min_{\boldsymbol{p}\in\mathbb{R}^{s(d)}} \left\{\boldsymbol{p}^T\boldsymbol{M}_d(\phi)\boldsymbol{p},\ \ \boldsymbol{p}^T\boldsymbol{v}_d(\boldsymbol{x}) = 1\right\}. \tag{1.29}$$

Hence, evaluating the Christoffel function at a particular point is the same as solving a specific convex quadratic programming. The interest of this formulation lies in the fact that this family of optimization problems can be handled very efficiently, even in large dimensions, by some numerical solvers, which is crucial in practical computations. However, the drawback is that for each point, one has to solve a completely new quadratic programming, i.e., no closed formula is available.

Let us consider an alternative way of defining this function:

Proposition 1.4.1. The Christoffel function of order (degree) $d \in \mathbb{N}$, defined in (1.28), satisfies the following relationship

$$\Lambda_d^\phi(\boldsymbol{x}) = \frac{1}{K_d^\phi(\boldsymbol{x},\boldsymbol{x})} = \frac{1}{\boldsymbol{v}_d(\boldsymbol{x})^T \boldsymbol{M}_d(\phi)^{-1} \boldsymbol{v}_d(\boldsymbol{x})}. \tag{1.30}$$

Proof. Let $\boldsymbol{x} \in \mathbb{R}^p$. First, we can see that the quotient in (1.30) is well defined thanks to the orthogonal polynomials definition of the Christoffel-Darboux kernel (1.11). Indeed, we have $K_d^\phi(\boldsymbol{x},\boldsymbol{x}) = \sum_{i=1}^{s(d)} P_i(\boldsymbol{x})^2 > 0$, for all $\boldsymbol{x} \in \mathbb{R}^p$.
Let us now consider the polynomial $P : \boldsymbol{z} \mapsto \frac{K_d^\phi(\boldsymbol{z},\boldsymbol{x})}{K_d^\phi(\boldsymbol{x},\boldsymbol{x})}$. By definition, $P \in \mathbb{R}_d[\boldsymbol{x}]$, and since $P(\boldsymbol{x}) = \frac{K_d^\phi(\boldsymbol{x},\boldsymbol{x})}{K_d^\phi(\boldsymbol{x},\boldsymbol{x})} = 1$, we deduce that P is admissible for the optimization problem in (1.28). This means that

$$\begin{aligned}
\Lambda_d^\phi(\boldsymbol{x}) \leq \int_{\boldsymbol{\Omega}} \frac{K_d^\phi(\boldsymbol{z},\boldsymbol{x})^2}{K_d^\phi(\boldsymbol{x},\boldsymbol{x})^2} d\phi(\boldsymbol{z}) &= \frac{1}{K_d^\phi(\boldsymbol{x},\boldsymbol{x})^2} \int_{\boldsymbol{\Omega}} K_d^\phi(\boldsymbol{z},\boldsymbol{x}) K_d^\phi(\boldsymbol{z},\boldsymbol{x}) d\phi(\boldsymbol{z}) && (1.31)\\
&= \frac{1}{K_d^\phi(\boldsymbol{x},\boldsymbol{x})^{\cancel{2}}} \cancel{K_d^\phi(\boldsymbol{x},\boldsymbol{x})} \quad \text{(reproducing property)} && (1.32)\\
&= \frac{1}{K_d^\phi(\boldsymbol{x},\boldsymbol{x})}. && (1.33)
\end{aligned}$$

On the other hand, if P is an admissible solution, then

$$\begin{aligned}
1 = P(\boldsymbol{x})^2 &= \left(\int_{\boldsymbol{\Omega}} K_d^\phi(\boldsymbol{z},\boldsymbol{x}) P(\boldsymbol{z}) d\phi(\boldsymbol{z}) \right)^2 \quad \text{(reproducing property)} && (1.34)\\
&\leq \int_{\boldsymbol{\Omega}} K_d^\phi(\boldsymbol{z},\boldsymbol{x})^2 d\phi(\boldsymbol{z}) \int_{\boldsymbol{\Omega}} P(\boldsymbol{z})^2 d\phi(\boldsymbol{z}) \quad \text{(Cauchy-Schwartz)} && (1.35)\\
&= K_d^\phi(\boldsymbol{x},\boldsymbol{x}) \int_{\boldsymbol{\Omega}} P(\boldsymbol{z})^2 d\phi(\boldsymbol{z}) \quad \text{(reproducing property)} && (1.36)
\end{aligned}$$

implying that $\frac{1}{K_d^\phi(\boldsymbol{x},\boldsymbol{x})} \leq \Lambda_d^\phi(\boldsymbol{x})$. Combining the two inequalities, we deduce that $\Lambda_d^\phi(\boldsymbol{x}) = \frac{1}{K_d^\phi(\boldsymbol{x},\boldsymbol{x})}$. Moreover, from computations leading to (1.33), we deduce that the polynomial P reaches the optimal value $\frac{1}{K_d^\phi(\boldsymbol{x},\boldsymbol{x})}$, so that it actually solves the problem in (1.28). □

Thus, evaluating the Christoffel function at some point $\boldsymbol{x} \in \mathbb{R}^p$ can be done by taking the reciprocal of the value of the Christoffel-Darboux kernel along the diagonal $\boldsymbol{y} = \boldsymbol{x}$. Indeed, recall that $\boldsymbol{v}_d(.)$ denotes a monomial basis of $\mathbb{R}_d[\boldsymbol{x}]$, so that $K_d^\phi(\boldsymbol{x},\boldsymbol{x}) = \boldsymbol{v}_d(\boldsymbol{x})^T \boldsymbol{M}_d(\phi)^{-1} \boldsymbol{v}_d(\boldsymbol{x})$ is nothing but a $2d$-degree polynomial evaluated at $\boldsymbol{x} \in \mathbb{R}^p$.
An advantage of this formulation is that it gives us an explicit closed-form formula for the Christoffel function, directly related to the Christoffel-Darboux kernel. However, this requires inverting an $s(d) \times s(d)$-dimensional matrix, which can be extremely costly in the high dimensional context, i.e., if p is high, or very unstable numerically, if d is high, for example.

This function has been known for a very long time among the researchers working on the approximation theory and orthogonal polynomials. The next figure motivates the use of the Christoffel function in statistical applications:

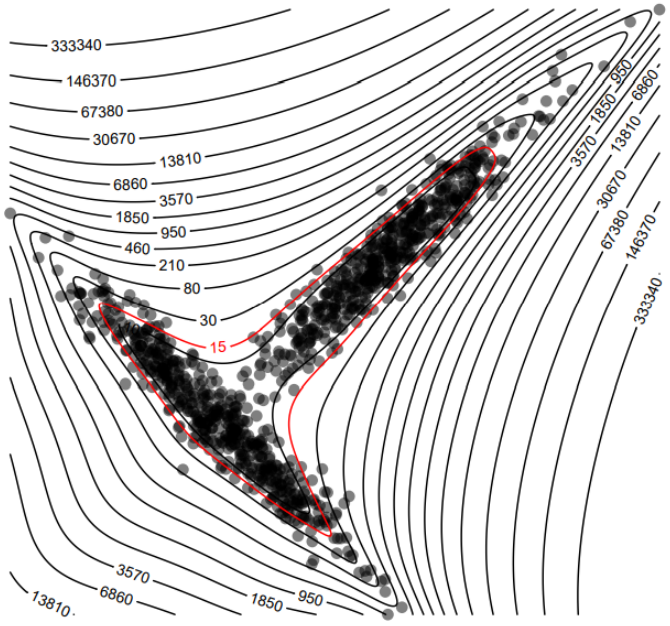


Figure 1.1: Level sets of a Christoffel function associated to the empirical measure [4, page 2].

The Figure 1.1. is obtained from an empirical measure $\phi_N := \sum_{i=1}^N \delta_{\boldsymbol{x}_i}$ associated to the cloud of data points $(\boldsymbol{x_i})_{i=1}^N$ in $\mathbb{R}^2$, with $N = 1000$. The red curve is just one particular level set associated to $\Lambda_4^{\phi_N}$. An immediate observation is that these level sets seem to capture very well the shape of the data points, i.e. the support of the underlying measure that the data was generated from. Although this property is true for more general measures and can be obtained from the true theoretical moments, the fact that it can be applied to the empirical measure, which is one of the central objects in statistics, increases the attractiveness of this function.

1.4.1 Some properties

Let us start by showing that the Christoffel function can identify pure points asymptotically.

Proposition 1.4.2. [3, 4] Let $\phi \in \mathcal{P}(\boldsymbol{\Omega})$, with $\boldsymbol{\Omega} \subset \mathbb{R}^p$ compact. Let $\boldsymbol{x} \in \mathbb{R}^p$. Then,

$$\lim_{d \to +\infty} \Lambda_d^\phi(\boldsymbol{x}) = \phi(\{\boldsymbol{x}\}). \tag{1.37}$$

Proof. Let $\boldsymbol{x} \in \mathbb{R}^p$. It is clear from the variational formulation in (1.28) that the limit in (1.37) exists, since the sequence $(\Lambda_d^\phi(\boldsymbol{x}))_{d\in\mathbb{N}}$ is bounded from below by zero and non-increasing.
Furthermore, using the properties of an integral with respect to a measure, we deduce that for any admissible $P \in \mathbb{R}_d[\boldsymbol{x}]$,

$$\int_{\boldsymbol{\Omega}} P(\boldsymbol{z})^2 d\phi(\boldsymbol{z}) \geq P(\boldsymbol{x})\phi(\{\boldsymbol{x}\}) = \phi(\{\boldsymbol{x}\}). \tag{1.38}$$

By taking the infimum and then the limit when $d \to +\infty$, we get that $\lim_{d\to+\infty} \Lambda_d^\phi(\boldsymbol{x}) \geq \phi(\{\boldsymbol{x}\})$. On the other hand, for any $d \in \mathbb{N}$, the polynomial $P : \boldsymbol{z} \mapsto (1 - ||\boldsymbol{z} - \boldsymbol{x}||^2)^d$ belongs to $\mathbb{R}_{2d}[\boldsymbol{x}]$ and

satisfies $P(\boldsymbol{x}) = 1$ so that

$$\Lambda^{\phi}_{2d+1}(\boldsymbol{x}) \leq \Lambda^{\phi}_{2d}(\boldsymbol{x}) \leq \int_{\boldsymbol{\Omega}} (1 - ||\boldsymbol{z} - \boldsymbol{x}||^2)^{2d} d\phi(\boldsymbol{z}) \tag{1.39}$$

$$\leq \int_{B(\boldsymbol{x}, d^{-\frac{1}{4}})} d\phi(\boldsymbol{z}) + \int_{\boldsymbol{\Omega} \setminus B(\boldsymbol{x}, d^{-\frac{1}{4}})} (1 - ||\boldsymbol{z} - \boldsymbol{x}||^2)^{2d} d\phi(\boldsymbol{z}) \tag{1.40}$$

$$\leq \int_{B(\boldsymbol{x}, d^{-\frac{1}{4}})} d\phi(\boldsymbol{z}) + \int_{\boldsymbol{\Omega} \setminus B(\boldsymbol{x}, d^{-\frac{1}{4}})} (1 - (d^{-\frac{1}{4}})^2)^{2d} d\phi(\boldsymbol{z}) \tag{1.41}$$

$$= \underbrace{\int_{B(\boldsymbol{x}, d^{-\frac{1}{4}})} d\phi(\boldsymbol{z})}_{\xrightarrow[d \to +\infty]{} \phi(\{\boldsymbol{x}\})} + \underbrace{\left(1 - \frac{1}{\sqrt{d}}\right)^{2d}}_{\xrightarrow[d \to +\infty]{} 0}. \tag{1.42}$$

Therefore, we conclude that $\lim_{d \to +\infty} \Lambda^{\phi}_d(\boldsymbol{x}) = \phi(\{\boldsymbol{x}\})$. □

The next result we present is related to the average value taken by the Christoffel function inside the support of its associated measure.

Proposition 1.4.3. [3, 4] Let $\boldsymbol{\Omega} \subset \mathbb{R}^p$ be compact and $\phi \in \mathcal{P}(\boldsymbol{\Omega})$ be an absolutely continuous probability measure. Let X be a random variable distributed according to ϕ. Let $d \in \mathbb{N}$. Then,

$$\mathbb{E}_{\phi}\left[\frac{1}{\Lambda^{\phi}_d(X)}\right] = s(d). \tag{1.43}$$

Proof. Let $P_1, \ldots, P_{s(d)}$ be an orthonormal basis of $\mathbb{R}[\boldsymbol{x}]$. By definition of an expectation,

$$\mathbb{E}_{\phi}\left[\frac{1}{\Lambda^{\phi}_d(X)}\right] = \mathbb{E}_{\phi}\left[K^{\phi}_d(X, X)\right] = \int_{\boldsymbol{\Omega}} K^{\phi}_d(\boldsymbol{x}, \boldsymbol{x}) d\phi(\boldsymbol{x}) \tag{1.44}$$

$$= \int_{\boldsymbol{\Omega}} \sum_{i=1}^{s(d)} P_i(\boldsymbol{x})^2 d\phi(\boldsymbol{x}) = \sum_{i=1}^{s(d)} \underbrace{\int_{\boldsymbol{\Omega}} P_i(\boldsymbol{x})^2 d\phi(\boldsymbol{x})}_{= ||P_i||_{\phi} = 1} \tag{1.45}$$

$$= s(d). \tag{1.46}$$

□

Thus, for an absolutely continuous measure on a compact support, the value of the Christoffel function decreases polynomially fast towards zero when we let $d \to +\infty$ (recall that $s(d) = \binom{p+d}{p} \sim d^p$).

Proposition 1.4.4. [4] Let $d \in \mathbb{N}$, and let $\phi \in \mathcal{P}(\boldsymbol{\Omega})$, with $\boldsymbol{\Omega} \subset \mathbb{R}^p$ compact. Then, for any $\boldsymbol{x_0} \in \mathbb{R}^p$ verifying $\text{dist}(\boldsymbol{x_0}, \boldsymbol{\Omega}) \geq \delta > 0$, one has

$$K^{\phi}_d(\boldsymbol{x_0}, \boldsymbol{x_0}) \geq 2^{\frac{d\delta}{\delta + \text{diam}(\boldsymbol{\Omega})} - 3}. \tag{1.47}$$

In order to prove the previous proposition, we use the following lemma:

Lemma 1.4.1. [9] Let $d \in \mathbb{N}^*$ and $\delta \in]0, 1[$. Then, there exists $Q \in \mathbb{R}_{2d}[\boldsymbol{x}]$ satisfying:

$$Q(\boldsymbol{0}) = 1, \quad ||\boldsymbol{x}|| \leq 1 \implies |Q(\boldsymbol{x})| \leq 1, \quad \text{and} \quad 0 < \delta \leq ||\boldsymbol{x}|| \leq 1 \implies |Q(\boldsymbol{x})| \leq 2^{1 - \delta d}. \tag{1.48}$$

Proof of the Proposition 1.4.4. Let $d \in \mathbb{N}^*$ and let $\delta \in]0,1[$. Set $\tilde{\delta} = \frac{\delta}{\delta + \text{diam}(\boldsymbol{\Omega})} \in]0,1[$. Consider $P \in \mathbb{R}_{2d}[\boldsymbol{x}]$ such that $P(\boldsymbol{x}) = Q\left(\frac{\boldsymbol{x}-\boldsymbol{x_0}}{\delta+\text{diam}(\boldsymbol{\Omega})}\right)$, where $Q \in \mathbb{R}_{2d}[\boldsymbol{x}]$ is the polynomial whose existence is assured by Lemma 1.4.1. Since $P(\boldsymbol{x_0}) = Q(\mathbf{0}) = 1$, we deduce that P is an admissible candidate in the optimization problem (1.28) defining $\Lambda^{\phi}_{2d}(\boldsymbol{x_0})$. Moreover, since $\text{dist}(\boldsymbol{x_0}, \boldsymbol{\Omega}) \geq \delta$, we deduce $\tilde{\delta} \leq \left|\left|\frac{\boldsymbol{x}-\boldsymbol{x_0}}{\delta+\text{diam}(\boldsymbol{\Omega})}\right|\right| \leq 1$ for all $\boldsymbol{x} \in \boldsymbol{\Omega}$. Hence, we obtain the following:

$$\Lambda^{\phi}_{2d}(\boldsymbol{x_0}) \leq \int P^2(\boldsymbol{x}) d\phi(\boldsymbol{x}) \leq \int \left(2^{1-\tilde{\delta}d}\right)^2 d\phi(\boldsymbol{x}) = 2^{2-2\tilde{\delta}d} \leq 2^{3-\tilde{\delta}2d}. \tag{1.49}$$

Additionally, since $\mathbb{R}_{2d}[\boldsymbol{x}] \subset \mathbb{R}_{2d+1}[\boldsymbol{x}]$, we deduce that $\Lambda^{\phi}_{2d+1}(\boldsymbol{x_0}) \leq \Lambda^{\phi}_{2d}(\boldsymbol{x_0})$. Combining the fact that $\tilde{\delta} \leq 1$ with the equation (1.49), we deduce that

$$\Lambda^{\phi}_{2d+1}(\boldsymbol{x_0}) \leq 2^{3-\tilde{\delta}(2d+1)}. \tag{1.50}$$

Since we have shown the validity of the formula for odd and even degrees separately, we deduce that for any $d \in \mathbb{N}^*, \Lambda^{\phi}_{d}(\boldsymbol{x_0}) \leq 2^{3-\tilde{\delta}d}$. Recalling the definition of $\tilde{\delta}$ and using the Proposition 1.4.1, we obtain

$$\Lambda^{\phi}_{d}(\boldsymbol{x_0}) \leq 2^{3-\tilde{\delta}d} \implies K^{\phi}_{d}(\boldsymbol{x_0}, \boldsymbol{x_0}) \geq 2^{\tilde{\delta}d-3} = 2^{\frac{d\delta}{\delta+\text{diam}(\boldsymbol{\Omega})}-3}. \tag{1.51}$$

□

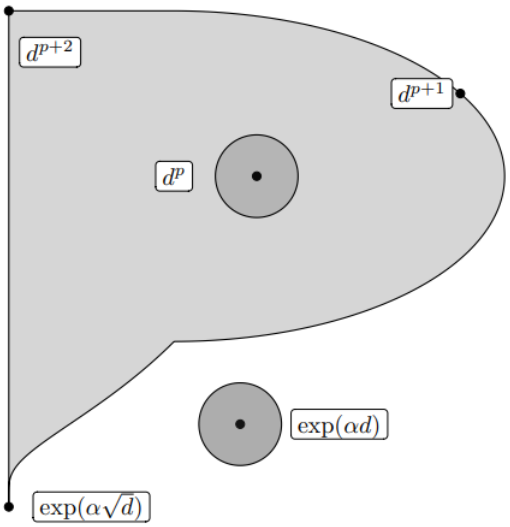


Figure 1.2: (Adapted (link) with permission) Different growth rates for $(\Lambda^{\phi}_{d}(\boldsymbol{x}))^{-1} = K^{\phi}_{d}(\boldsymbol{x}, \boldsymbol{x})$.

The Figure 1.2, together with the Proposition 1.4.3. and the Proposition 1.4.4. provides a rationale for the use of the Christoffel function in data analysis. Indeed, there exists some kind of dichotomy that is manifested through the asymptotically exponentially fast decay of the Christoffel function for the points not belonging to the support of its associated measure. On the other hand, as d increases, this decay is at most polynomial for the points inside the support. This is essentially the main mechanism through which the support of the underlying measure is identified. Moreover, we can also observe that different rates can be obtained for the points lying on the boundary, depending on its smoothness.

Finally, let us conclude this section by stating a theorem that relates the Christoffel functions associated to a certain measure and its empirical counterpart.

Theorem 1.4.1. [2, 4] Let $d \in \mathbb{N} \setminus \{0\}$, and let $\phi \in \mathcal{P}(\mathbf{\Omega})$, with $\mathbf{\Omega} \subset \mathbb{R}^p$ compact. Let $n \in \mathbb{N}$ and $(\boldsymbol{x}_i)_{i=1}^n$ be a sample of points in $\mathbb{R}^p$ drawn according to ϕ, and $\phi_n := \sum_{i=1}^n \delta_{\boldsymbol{x}_i}$ its associated empirical measure. Then,

$$||\Lambda_d^{\phi_n} - \Lambda_d^{\phi}||_\infty = \sup_{\boldsymbol{x} \in \mathbb{R}^p} \{|\Lambda_d^{\phi_n}(\boldsymbol{x}) - \Lambda_d^{\phi}(\boldsymbol{x})|\} \xrightarrow[n \to +\infty]{} 0 \quad \text{a.s.} \tag{1.52}$$

This theorem provides some convergence guarantees that justify the use of the empirical Christoffel function. In the context of data analysis, we are rarely aware of the probability measure that our data was generated from, which is why such guarantees are important.
Finally, a similar result as in Theorem 1.4.1. can be stated in terms of level sets. More precisely, we can certify that the level sets of the empirical and theoretical Christoffel function will asymptotically coincide, for any fixed degree. Once again, such a result puts us in a position to exploit the behavior of the Christoffel function when it comes to solving specific problems that arise in data analysis, which will be more detailed in the next part of the report.

Chapter 2

Application in machine learning

2.0.1 Introductory remarks

This section is going to be mainly focused on adapting the Christoffel function to the supervised learning context. Indeed, results from [1, 4, 11] have highlighted some remarkable capabilities of this function when it comes to recovering or approximating the graph of an unknown function.
Since the classification task can be viewed as a specific task of the graph approximation, it becomes then natural to connect these ideas with the Christoffel function [5]. However, such an approach requires a certain amount of slight modifications to be made to the Christoffel function, which will be detailed in a sequel. This will allow us to define a classifier based on the Christoffel function and to evaluate its performance in various classification tasks.

2.1 Supervised learning

Suppose that we want to solve a classification problem of assigning the points in $\mathbb{R}^p$ to one of the $m \geq 2$ classes in the set $\boldsymbol{Y} := \{1, \ldots, m\}$. Moreover, denote by $\boldsymbol{X}_j \subset \mathbb{R}^p$ the set of all points that fall into the given class $j \in \boldsymbol{Y}$. Let $\boldsymbol{X} = \bigcup_{j \in \boldsymbol{Y}} \boldsymbol{X}_j$ and suppose that $\boldsymbol{X}$ is open with compact closure.
Let μ be a joint probability distribution on the set $\boldsymbol{\Omega} = \boldsymbol{X} \times \boldsymbol{Y}$. One may always disintegrate μ into its marginal distribution ϕ on $\boldsymbol{X}$ and its conditional distribution $\nu_{\boldsymbol{x}}$ on $\boldsymbol{Y}$ given a point $\boldsymbol{x} \in \boldsymbol{X}$, so that $d\mu(\boldsymbol{x}, y) = \nu_{\boldsymbol{x}}(dy)\phi(d\boldsymbol{x})$ [6].
If the points to classify are perfectly separable, i.e. they can not belong to more than one class, then we can assume that the marginal distribution assigns no mass to the intersection of the supports, meaning that $\phi(\boldsymbol{X}_i \cap \boldsymbol{X}_j) = 0$, for all $(i, j) \in \boldsymbol{Y}^2$ with $i \neq j$. This allows us to write $\mu = \sum_{j \in \boldsymbol{Y}} \mu_j$, where for all $j \in \boldsymbol{Y}$, we have

$$d\mu_j(\boldsymbol{x}, y) = \delta_j(dy)\phi_j(d\boldsymbol{x}), \quad \text{with } \phi_j \in \mathcal{P}(\boldsymbol{X}_j). \tag{2.1}$$

This also implies that $\phi = \sum_{j \in \boldsymbol{Y}} \phi_j$, because $\phi(A) = \mu(A \times \boldsymbol{Y}) = \sum_{j \in \boldsymbol{Y}} \mu_j(A \times \boldsymbol{Y}) = \sum_{j \in \boldsymbol{Y}} \phi_j(A)$, for any Borel set $A \in \mathcal{B}(\boldsymbol{X})$.
Recall now that the goal of the supervised learning [5, 9, 12] is to recover a classifier f such that

$$f \colon \boldsymbol{X} \longrightarrow \boldsymbol{Y} \tag{2.2}$$

$$\boldsymbol{x} \longmapsto f(\boldsymbol{x}) := \begin{cases} \sum_{j=1}^{m} j \mathbf{1}_{\boldsymbol{X}_j}(\boldsymbol{x}), & \text{if } \boldsymbol{x} \in \boldsymbol{X} \\ 0, & \text{otherwise} \end{cases} \tag{2.3}$$

Thus, evaluating f at $\boldsymbol{x} \in \boldsymbol{X}$ gives either the class of the point $\boldsymbol{x}$ or zero (if the point belongs to different classes simultaneously).
As a byproduct of this construction, we obtain an alternative formulation of the underlying measure μ, namely

$$d\mu(\boldsymbol{x}, y) = \delta_{f(\boldsymbol{x})}(dy)\phi(d\boldsymbol{x}) \tag{2.4}$$

Hence, the graph $G := \{(\boldsymbol{x}, f(\boldsymbol{x})),\ \ \boldsymbol{x} \in \boldsymbol{X}\}$ of the classifier f is nothing else but the support of our underlying probability measure μ. This is why we can identify the graph recovery problem with the problem of learning a classifier, and connect these two using the Christoffel function.

2.2 Christoffel-Darboux classifier

This part is quite technical but its purpose is to explain how the Christoffel function needs to be modified in order to to encompass the problem of the supervised learning. Indeed, the usual moment matrix in the previously described classification framework will suffer from rank deficiency, requiring thus some additional refinements. Stated differently, the bi-linear form defined in (1.9) may not be positive definite, but we may still find an alternative way to circumvent the problem.

In this case, the support of the measure μ defined in (2.4) is contained in a real algebraic variety of the space $\mathbb{R}^p$, which is why (1.9) fails to define a valid scalar product on $L^2(\mu)$. One way to observe this is by considering the polynomial $Q \in \mathbb{R}[\boldsymbol{x}, y]$ such that $Q(\boldsymbol{x}, y) := \prod_{i=1}^{m}(y - i)$. Then, $V := \{(\boldsymbol{x}, y) \in \mathbb{R}^p \times \mathbb{R} \mid Q(\boldsymbol{x}, y) = 0\} = \mathbb{R}^p \times \boldsymbol{Y}$ is the real algebraic set containing the support of the measure μ. Thus, for any $0 \neq P \in \mathbb{R}[\boldsymbol{x}, y]$ that vanishes on V, and represented by the vector of coefficients $\boldsymbol{p}$, we would have

$$\boldsymbol{p}^T \boldsymbol{M}_d(\phi)\boldsymbol{p} = \int_{\boldsymbol{\Omega}} P^2 d\mu = 0, \tag{2.5}$$

justifying the non-invertibility of the moment matrix, so that equations (1.30) are not valid any more [4, 5]. In order to avoid these issues, one may define the Christoffel function as follows

$$\Lambda_d^{\mu} \colon V \longrightarrow \mathbb{R}_+ \tag{2.6}$$

$$(\boldsymbol{x}, y) \longmapsto \Lambda_d^{\mu}(\boldsymbol{x}, y) := \min_{P \in L_d^2(\mu)} \left\{ \int_{\boldsymbol{\Omega}} P^2(\boldsymbol{z}) d\phi(\boldsymbol{z}),\ \ P(\boldsymbol{x}, y) = 1 \right\}, \tag{2.7}$$

where $L_d^2(\mu) \subset \mathbb{R}[\boldsymbol{x}, y]$ is the space of polynomials on V of overall degree at most d (and at most $m-1$ in the variable y). A basis of such a set is thus given by $\{\boldsymbol{x}^{\boldsymbol{\alpha}} y^k \mid k \leq m-1, \sum_{i=1}^{p} \alpha_i + k \leq d\}$. Then, the associated moment matrix is non-singular and we can retrieve the standard formulations of the Christoffel function on the set V [5].

Alternatively, we could perturb the measure μ by considering $\mu + \epsilon\mu_0$, with $\epsilon > 0$ and μ_0 a uniform probability measure on $\boldsymbol{X} \times [0, m]$. Then,

$$\Lambda_d^{\mu+\epsilon\mu_0} \colon \mathbb{R}^p \times \mathbb{R} \longrightarrow \mathbb{R}_+ \tag{2.8}$$

$$(\boldsymbol{x}, y) \longmapsto \Lambda_d^{\mu+\epsilon\mu_0}(\boldsymbol{x}, y) := \boldsymbol{v}_d(\boldsymbol{x}, y)^T M_d(\mu + \epsilon\mu_0)^{-1} \boldsymbol{v}_d(\boldsymbol{x}, y) \tag{2.9}$$

is well-defined and allows to approximate f via the following relationship [11]:

$$\hat{f}(\boldsymbol{x}) := \arg\min_{y} \Lambda_d^{\mu+\epsilon\mu_0}(\boldsymbol{x}, y). \tag{2.10}$$

However, a more intuitive and a more classification-oriented interpretation can be obtained by implementing another minor modification to the Christoffel function. Let us consider

$$\hat{\Lambda}_d^\mu \colon V \longrightarrow \mathbb{R}_+ \tag{2.11}$$

$$(\boldsymbol{x}, y) \longmapsto \hat{\Lambda}_d^\mu(\boldsymbol{x}, y) := \min_{P \in \mathcal{L}_d^2(\mu)} \left\{ \int_{\boldsymbol{\Omega}} P^2(\boldsymbol{z}) d\phi(\boldsymbol{z}),\ \ P(\boldsymbol{x}, y) = 1 \right\}, \tag{2.12}$$

where $\mathcal{L}_d^2(\mu) := \mathbb{R}_{d,m-1}[\boldsymbol{x}, y]$ is the space of all polynomials of degree at most d with respect to $\boldsymbol{x}$ and at most $m-1$ with respect to the variable y.

By construction, $L_d^2(\mu) \subset \mathcal{L}_d^2(\mu) \subset L_{d+m-1}^2$, so that for all $(\boldsymbol{x}, y) \in V$,

$$\Lambda_{d+m-1}^\mu(\boldsymbol{x}, y) \leq \hat{\Lambda}_d^\mu(\boldsymbol{x}, y) \leq \Lambda_d^\mu(\boldsymbol{x}, y), \tag{2.13}$$

which is how these two constructions connect to each other. However, the advantage of the formulation in (2.12) lies in the fact that it possesses a closed form expression connected to the interpolation polynomials at the points $\{1, \ldots, m\}$, which are defined through

$$y \mapsto \pi_j(y) := \frac{\prod_{i \neq j}(y - i)}{\prod_{i \neq j}(j - i)}, \quad j = 1, \ldots, m \tag{2.14}$$

and satisfy $\pi_j(i) = \delta_{ji}$, for all $i, j \in \boldsymbol{Y}$.
This connection can be stated more precisely in the form of the following theorem

Theorem 2.2.1. [5] Let $j \in \boldsymbol{Y}$ and let $(P_{\boldsymbol{\alpha}}^j)_{\boldsymbol{\alpha} \in \mathbb{N}_d^p}$ be a family of polynomials in $\mathbb{R}[\boldsymbol{x}]$ orthonormal with respect to $\phi_j \in \mathcal{P}(\boldsymbol{X}_j)$. Let $\Lambda_d^{\phi_j}$ be the associated Christoffel function defined as in (1.28). Then,

1. The family $(\pi_j P_{\boldsymbol{\alpha}}^j)_{\boldsymbol{\alpha} \in \mathbb{N}_d^p, j \in \boldsymbol{Y}}$ is an orthonormal basis of $\mathcal{L}_d^2(\mu)$;

2. $\hat{\Lambda}_d^\mu$ defined in (2.12) satisfies

$$(\hat{\Lambda}_d^\mu(\boldsymbol{x}, y))^{-1} = \sum_{j \in \boldsymbol{Y}} \pi_j(y)^2 \sum_{\boldsymbol{\alpha} \in \mathbb{N}_d^p} P_{\boldsymbol{\alpha}}^j(\boldsymbol{x})^2 = \sum_{j \in \boldsymbol{Y}} \delta_{y=j} (\Lambda_d^{\phi_j}(\boldsymbol{x}, y))^{-1}. \tag{2.15}$$

Proof.
1. The cardinality of the basis $(\pi_j P_{\boldsymbol{\alpha}}^j)_{\boldsymbol{\alpha} \in \mathbb{N}_d^p, j \in \boldsymbol{Y}}$ is $m \cdot s(d)$, which corresponds to the dimension of the space $\mathcal{L}_d^2(\mu) = \mathbb{R}_{d,m-1}[\boldsymbol{x}, y]$. Using the fact that $(\pi_j)_{j \in \boldsymbol{Y}}$ generates $\mathbb{R}_{m-1}[y]$ and that $(P_{\boldsymbol{\alpha}}^j)_{\boldsymbol{\alpha} \in \mathbb{N}_d^p}$ generates $\mathbb{R}_d[\boldsymbol{x}]$, we can easily deduce that every polynomial $R \in \mathbb{R}_{d,m-1}[\boldsymbol{x}, y]$ can be written in as $R(\boldsymbol{x}, y) = \sum_{\boldsymbol{\alpha} \in \mathbb{N}_d^p, j \in \boldsymbol{Y}} p_j r_{\boldsymbol{\alpha}}^j (\pi_j(y) P_{\boldsymbol{\alpha}}^j(\boldsymbol{x}))$. Finally, as long as $i \neq j$, we deduce from (2.14) that $\pi_i(y)\pi_j(y) = 0$ inside the support of μ, so that $\int \pi_i(y) P_{\boldsymbol{\alpha}}^i(\boldsymbol{x}) \pi_j(y) P_{\boldsymbol{\beta}}^j(\boldsymbol{x}) d\mu(\boldsymbol{x}, y) = 0$ as well. On the other hand, if $i = j$, then

$$\int \pi_i(y)^2 P_{\boldsymbol{\alpha}}^i(\boldsymbol{x}) P_{\boldsymbol{\beta}}^i(\boldsymbol{x}) d\mu(\boldsymbol{x}, y) = \sum_{j \in \boldsymbol{Y}} \int \pi_i(y)^2 P_{\boldsymbol{\alpha}}^i(\boldsymbol{x}) P_{\boldsymbol{\beta}}^i(\boldsymbol{x}) d\mu_j(\boldsymbol{x}, y) \quad \text{(from (2.1))} \tag{2.16}$$

$$= \int \pi_i(y)^2 P_{\boldsymbol{\alpha}}^i(\boldsymbol{x}) P_{\boldsymbol{\beta}}^i(\boldsymbol{x}) d\mu_i(\boldsymbol{x}, y) \quad \text{(by assumption)} \tag{2.17}$$

$$= \int P_{\boldsymbol{\alpha}}^i(\boldsymbol{x}) P_{\boldsymbol{\beta}}^i(\boldsymbol{x}) d\phi_i(\boldsymbol{x}, y) \quad \text{(from (2.1))} \tag{2.18}$$

$$= \mathbf{1}_{\boldsymbol{\alpha} = \boldsymbol{\beta}} \quad \text{(by definition)} \tag{2.19}$$

which proves orthogonality.

2. Consider the monomial basis of $\mathcal{L}^2_d(\mu)$, $(\boldsymbol{x}^{\boldsymbol{\alpha}} y^j)_{\boldsymbol{\alpha}\in\mathbb{N}^p_d, 0\le j\le m-1}$, and denote by $\hat{\boldsymbol{v}}(\boldsymbol{x}, y)$ the associated vector of monomials. Let $\hat{M}_d(\mu)$ be the corresponding moment matrix defined as in (1.3). Following (1.29), we deduce that

$$\hat{\Lambda}_d(\boldsymbol{x}, y) = \min_{\boldsymbol{p}\in\mathbb{R}^{m\cdot s(d)}} \left\{ \boldsymbol{p}^T \hat{M}_d(\mu)\boldsymbol{p},\ \ \boldsymbol{p}^T \hat{\boldsymbol{v}}_d(\boldsymbol{x}, y) = 1 \right\}. \tag{2.20}$$

We can then define the Lagrangian $(\boldsymbol{p}, \lambda) \mapsto L(\boldsymbol{p}, \lambda) := \boldsymbol{p}^T \hat{M}_d(\mu)\boldsymbol{p} - \lambda(1 - \boldsymbol{p}^T \hat{\boldsymbol{v}}_d(\boldsymbol{x}, y))$. The objective function being strictly convex, and the constraint being linear, we deduce that the first order condition for this optimization problem is both necessary and sufficient, and allows to characterize the optimal admissible solution $(\boldsymbol{p}^*, \lambda^*)$ as follows

$$2\hat{M}_d(\mu)\boldsymbol{p}^* = \lambda^* \hat{\boldsymbol{v}}_d(\boldsymbol{x}, y), \tag{2.21}$$

which after multiplying by $\boldsymbol{p}^*$ yields

$$\lambda^* = 2\hat{\Lambda}_d(\boldsymbol{x}, y) \quad \text{and} \quad \boldsymbol{p}^* = \hat{\Lambda}_d(\boldsymbol{x}, y)(\hat{M}_d(\mu))^{-1}\hat{\boldsymbol{v}}(\boldsymbol{x}, y). \tag{2.22}$$

This implies that the optimal polynomial in the definition of $\hat{\Lambda}_d(\boldsymbol{x}, y)$ writes, for all $(\boldsymbol{u}, z) \in \mathbb{R}^p \times \boldsymbol{Y}$,

$$\begin{aligned} p^*(\boldsymbol{u}, z) &= \hat{\boldsymbol{v}}(\boldsymbol{x}, y)^T \hat{\Lambda}_d(\boldsymbol{x}, y)(\hat{M}_d(\mu))^{-1}\hat{\boldsymbol{v}}(\boldsymbol{x}, y) \\ &= \hat{\Lambda}_d(\boldsymbol{x}, y) \sum_{\boldsymbol{\alpha}\in\mathbb{N}^p_d, j\in\boldsymbol{Y}} \pi_j(y)\pi_j(z) P^j_{\boldsymbol{\alpha}}(\boldsymbol{x}) P^j_{\boldsymbol{\alpha}}(\boldsymbol{u}) \quad \text{(from (1.11) and Proposition 1.4.1)} \end{aligned} \tag{2.23}$$

By evaluating p^* at $(\boldsymbol{x}, y)$, we deduce that

$$p^*(\boldsymbol{x}, y) = 1 = \hat{\Lambda}_d(\boldsymbol{x}, y) \sum_{\boldsymbol{\alpha}\in\mathbb{N}^p_d, j\in\boldsymbol{Y}} \pi_j(y)^2 P^j_{\boldsymbol{\alpha}}(\boldsymbol{x})^2 \tag{2.24}$$

$$= \hat{\Lambda}_d(\boldsymbol{x}, y) \sum_{j\in\boldsymbol{Y}} \pi_j(y)^2 \sum_{\boldsymbol{\alpha}\in\mathbb{N}^p_d} P^j_{\boldsymbol{\alpha}}(\boldsymbol{x})^2 \tag{2.25}$$

$$= \hat{\Lambda}_d(\boldsymbol{x}, y) \sum_{j\in\boldsymbol{Y}} \pi_j(y)^2 (\Lambda^{\phi_j}_d(\boldsymbol{x}))^{-1} \quad \text{(use definition)} \tag{2.26}$$

$$= \hat{\Lambda}_d(\boldsymbol{x}, y) \sum_{j\in\boldsymbol{Y}} \delta_{y=j} (\Lambda^{\phi_j}_d(\boldsymbol{x}))^{-1} \quad \text{(from (2.14))} \tag{2.27}$$

which are exactly those equalities stated in (2.15). □

Thus, if $y \in \boldsymbol{Y}$, we can easily and simply express the modified Christoffel function in terms of the class-specific standard Christoffel functions $\Lambda^{\phi_j}_d$, for $j = 1, \ldots, m$.

2.2.1 Consistency result

The expression that was derived in Theorem 2.2.1. is what motivates the following definition.

Definition 2.2.1. The *Christoffel-Darboux classifier* (CDC) of order $d \ge 1$ is a function $\hat{f}_d$ such that

$$\forall \boldsymbol{x} \in \boldsymbol{X}, \quad \hat{f}_d(\boldsymbol{x}) := \arg\max_{j\in[\![1,m]\!]} \Lambda^{\phi_j}_d(\boldsymbol{x}). \tag{2.28}$$

That this definition is natural follows from the following observation. Fix $k \in \boldsymbol{Y}$ and $\boldsymbol{x} \in \boldsymbol{X}_k$. From (1.37) and (1.47), we know that, if we increase the order d, then $\Lambda_d^{\phi_j}(\boldsymbol{x})$ should decrease exponentially fast towards zero for all $j \neq k$. On the other hand, $\Lambda_d^{\phi_k}(\boldsymbol{x})$ should manifest a decay which is not faster that polynomial.
So, if the degree is sufficiently large (say higher than a threshold $d_{\boldsymbol{x}} \in \mathbb{N}$), then after some time, we should necessarily have $\Lambda_d^{\phi_j}(\boldsymbol{x}) < \Lambda_d^{\phi_k}(\boldsymbol{x})$, for all $d \geq d_{\boldsymbol{x}}$ and all $j \neq k$. Consequently, those class-specific Christoffel functions can be interpreted as score-determining, so that every new point is classified to the class that yields the largest class-specific Christoffel function.

That this procedure represents a consistent way of defining a classifier is stated in the following theorem:

Theorem 2.2.2. [5] Let $j \in [\![1, m]\!], \epsilon > 0$ and $\boldsymbol{X}_j^\epsilon := \{\boldsymbol{x} \in \boldsymbol{X}_j \mid d(\boldsymbol{x}, \partial \boldsymbol{X}_j) > \epsilon\}$. Assume that for all $j \in [\![1, m]\!]$, ϕ_j is absolutely continuous with respect to the restriction of the Lebesgue measure on $\boldsymbol{X}_j$. Then, there exists $d_{\boldsymbol{x},\epsilon} \in \mathbb{N}$ such that for all $d \geq d_{\boldsymbol{x},\epsilon}$

$$\hat{f}_d(\boldsymbol{x}) = j, \quad \forall \boldsymbol{x} \in \boldsymbol{X}_j^\epsilon. \tag{2.29}$$

Due to the high number of technicalities, we choose to omit the proof of the Theorem 2.2.2. Nonetheless, the main message that should be taken from this theorem is that, by adjusting the order of the classifier $\hat{f}_d$, one can make sure that all the points, which are sufficiently far away from the class boundary, get correctly classified, which is quite a desirable feature.

It is important to notice that the classifier we defined in (2.28) remains purely theoretical, since we rarely know those population probability distributions $(\phi_k)_{k \in \boldsymbol{Y}}$. What we have instead at our disposal is the finite simple of labeled data $\{(\boldsymbol{x}_i, y_i) \in \boldsymbol{X} \times \boldsymbol{Y}\}_{i=1}^N$ often referred to as the *training data set*. These data points can be then used to construct empirical counterparts of the quantities defined in (2.28).

Indeed, we may consider the sequence of empirical measures $(\phi_{k,N})_{k \in \boldsymbol{Y}}$ and their associated empirical Christoffel functions $(\Lambda_d^{\phi_{k,N}})_{k \in \boldsymbol{Y}}$. The natural question that this construction raises is whether these discretizations still provide good approximations of the support of the underlying probability measures.
To answer this question, suppose that the training sets consists of N data points from each of the m classes. Compute for all $k \in \boldsymbol{Y}$, $\phi_{k,N} := \frac{1}{N} \sum_{i=1}^N \boldsymbol{x}_i^k$, and let $(\Lambda_d^{\phi_{k,N}}(\boldsymbol{x}))^{-1} = \boldsymbol{v}_d(\boldsymbol{x})^T (M_d(\phi_{k,N}))^{-1} \boldsymbol{v}_d(\boldsymbol{x})$. These functions are indeed well-defined as soon as the sample size is large enough and d is chosen carefully so that the empirical moment matrix is indeed invertible [2].

Then, the empirical analogue of the CDC is given by

$$\forall \boldsymbol{x} \in \boldsymbol{X}, \quad \hat{f}_d^N(\boldsymbol{x}) := \underset{j \in [\![1,m]\!]}{\arg\max}\, \Lambda_d^{\phi_{j,N}}(\boldsymbol{x}). \tag{2.30}$$

We have the following result:

Theorem 2.2.3. [5] Let $j \in [\![1, m]\!], \epsilon > 0$ and $\boldsymbol{X}_j^\epsilon := \{\boldsymbol{x} \in \boldsymbol{X}_j \mid d(\boldsymbol{x}, \partial \boldsymbol{X}_j) > \epsilon\}$. Let $(\boldsymbol{x}_i^j)_{i=1}^N \subset \boldsymbol{X}_j$ be an i.i.d. sample drawn from an absolutely continuous distribution $\phi_j \in \mathcal{P}(\boldsymbol{X}_j)$. Then, there exists $d_{\boldsymbol{x},\epsilon} \in \mathbb{N}$ such that for all $d \geq d_{\boldsymbol{x},\epsilon}$, and for sufficiently large sample size N,

$$\hat{f}_d^N(\boldsymbol{x}) = j, \quad \forall \boldsymbol{x} \in \boldsymbol{X}_j^\epsilon, \tag{2.31}$$

almost surely with respect to the random samples.

Proof. Let $j \in [\![1, m]\!]$ and $\epsilon > 0$. Let $\boldsymbol{x} \in \boldsymbol{X}_j^{\epsilon}$. From the Theorem 2.2.2, we deduce that it is possible to find a real constant $c_j > 0$ and $d_{\boldsymbol{x},\epsilon} \in \mathbb{N}$ such that $\Lambda_d^{\phi_j}(\boldsymbol{x}) - \Lambda_d^{\phi_k}(\boldsymbol{x}) > c_j$, for all $k \neq j$ and all $d \geq d_{\boldsymbol{x},\epsilon}$.
Furthermore, using the Theorem 1.4.1, we know that, for all $k \in [\![1, m]\!]$ and all $d \in \mathbb{N}^*$,

$$\sup_{\boldsymbol{x} \in \mathbb{R}^p} \{|\Lambda_d^{\phi_{k,N}}(\boldsymbol{x}) - \Lambda_d^{\phi_k}(\boldsymbol{x})|\} \xrightarrow[N \to +\infty]{} 0 \tag{2.32}$$

almost surely with respect to the random samples $(\boldsymbol{x}_i^k)_{i=1}^N \subset \boldsymbol{X}_k$. If we combine these two remarks, we deduce that for all $d > d_{\boldsymbol{x},\epsilon}$ and all $k \neq j$,

$$\Lambda_d^{\phi_{j,N}}(\boldsymbol{x}) > \Lambda_d^{\phi_{k,N}}(\boldsymbol{x}) + c_j, \tag{2.33}$$

provided that N is sufficiently large, which is the same as saying that $\hat{f}_d^N(\boldsymbol{x}) = j$. □

In other words, Theorem 2.2.3 provides some theoretical guarantees that could be used to justify the use of the CDC in the standard supervised learning framework. Implementing and evaluating this classifier is what the next sections are going to be mostly based on.

2.3 Empirical evaluations

At this stage, we are interested in implementing numerically the CDC classifier and measuring its performance on various data sets. More precisely, the results that will be presented below are obtained thanks to our preliminary code written in Python.

A significant amount of preliminary computations was necessary in order to successfully implement the CDC classifier. For example, the following functions were created:

- *gen_basis*(p, d) − generates the $p-$variate monomial basis of order d w.r.t. the graded lexicographic order;
- *MoEM*(*data,d*) − computes the empirical moment matrix (indexed by monomials) of order d associated to the points in *data*;
- *Christoffel*(*point,d,data*) − evaluates at *point* the empirical Christoffel function of order d associated to the cloud of points in *data*.

Next, inspired by the already available classifiers from the*scikit−learn* library, we create a a new **class CDKC()** of objects which will be responsible for implementing the formula in (2.30).
Some of the **main attributes** associated to this class of objects are:

- *emommatrix* - computes the empirical moment matrix from the input data;
- *predict(tobeclassified)* - assigns labels to the points in *tobeclasified*;
- *score(tobeclassified,true_label)* - percentage of correctly classified points;
- *confusion matrix(true_label,prediction)* - confusion matrix computation.

It is important to mention that the code has not been optimized, but rather used in order to obtain some benchmark comparisons. Furthermore, we emphasize the fact that the CDC implementation is dependant on only one tuning parameter, namely the degree d of the involved Christoffel functions.

2.3.1 *Iris* data

We first evaluate the performance of the CDC clasifier on the very famous *Iris data* set. This data set contains four features (length and width of sepals and petals) of 50 instances of three species of Iris flower, i.e. $\boldsymbol{Y} = \{1, 2, 3\}$ and $N = 50$, so that we have 150 labeled data points that were split into training and test categories via standard procedures.

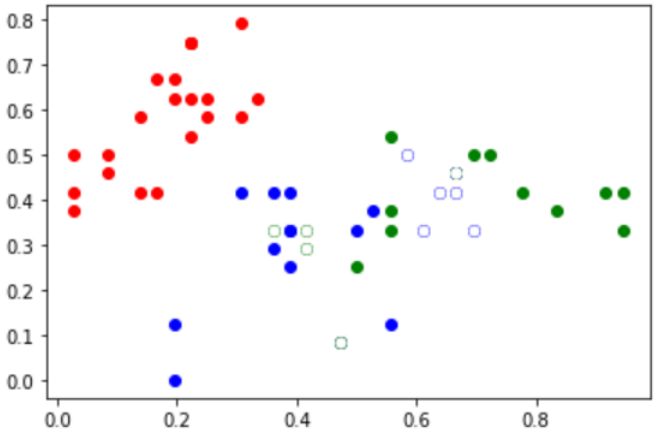


Figure 2.1: Scaled *Iris* data - CDC classification example with $d = 4$; points whose interior is not coloured are those that were miss-classified.

What we can observe from Figure 2.1. is that the red points were all correctly classified. Indeed, miss-classification occurred among blue and green points only since their supports are not disjoint, which violates the assumptions that we used in our theoretical framework. These intersecting supports make the exact classification impossible. Notice also that only two out of four features were used to solve the classification task.
Another interesting observation, which is rather intuitive, is that by considering all the features, one may increase the percentage of the correctly classified data. Indeed, in the 4-dimensional space, this data becomes much more easily separable.

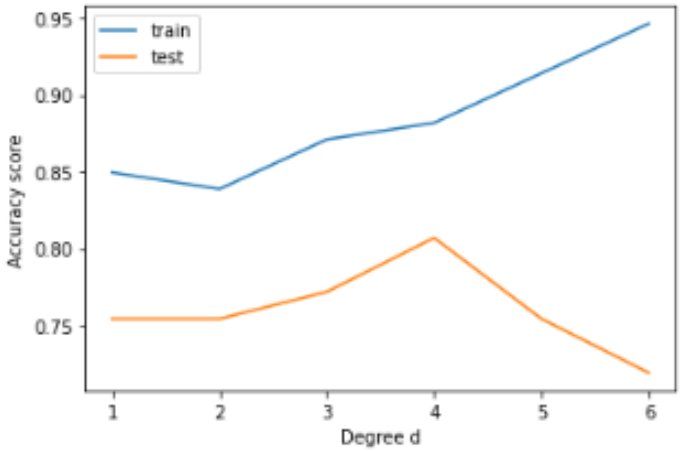


Figure 2.2: Score of the CDC classifier on the test and training sets as a function of the tuning parameter d.

As expected, if one increases the degree of the CDC classifier, its performance also increases and stabilizes around a certain threshold, if the data is not perfectly separable. However, by doing so, it seems that the levels sets of the Christoffel functions involved in the CDC expression become too specific to the data from the training set, and thus fail to generalize to the data in the test set. This over-fitting tendency seems to be captured by the Figure 2.2. However, an alternative explanation would be that the rather small sample size and degree are not sufficient to make the moment matrix invertible, so that this behavior is a numerical artefact resulting from the inversion of ill-conditioned matrices (the smallest eigenvalues of label-specific moment matrices are of order 10^{-10}).

Let us now compare the CDC classifier with other commonly used classifiers:

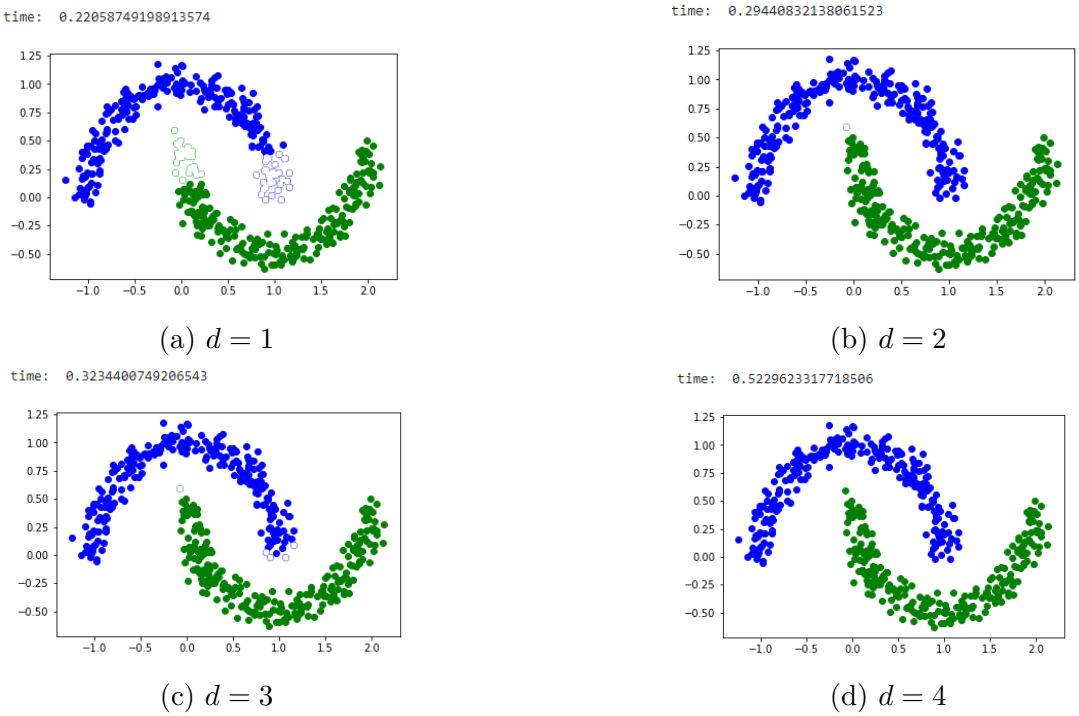
Comparison for $p = 2$ features		
Classifier	Precision (%)	Time (ms)
CDC ($d = 4$)	81	21.8
kNN	79	2.31
LDA	75	2.54
SVM	77	3.19
Tree	74	2.67

Table 2.1: Mean execution time and accuracy comparison

First, we notice that, even for small values of the tuning parameter, the CDC classifier is able to achieve great performance compared to the other commonly used classifiers. Furthermore, we observe that this method is much more costly to implement numerically, since the mean execution time is higher that what many other classifiers require. Moreover, by considering all the four features, the CDC classifier is able to correctly separate all the points by using only the moments up to order four, i.e., $d = 2$. However, the method's computational speed deteriorates significantly (200 ms on average), while the computing time of other classifiers remains quite stable. This indicates that the current implementation of the CDC classifier suffers from important dimensionality issues, so that the classifier is suitable for small-size problems only.

2.3.2 Toy data sets

We study the behaviour of the CDC classifier on various artificial data sets.
We start by considering the famous generic *Two moons* example. We control the level of the noise so that the two moon-like shapes remain disjoint (separable). Some of the obtained results are summarized by the means of the following figure:



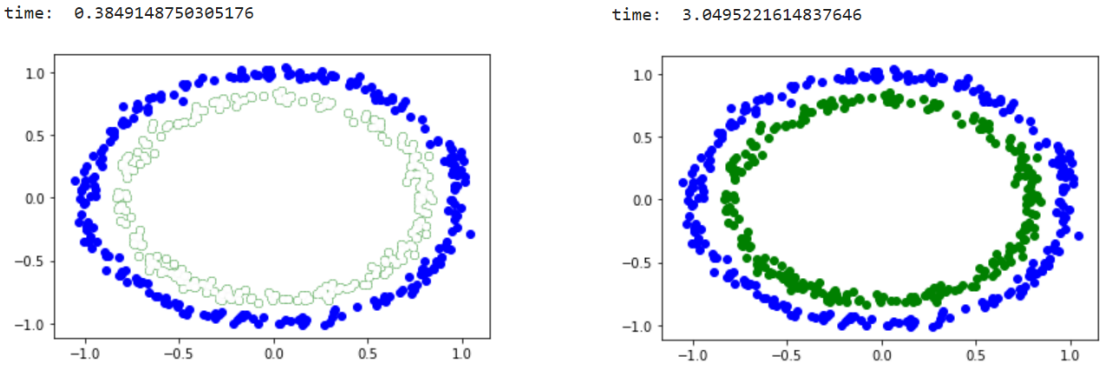
(a) $d = 1$ (b) $d = 2$

(c) $d = 3$ (d) $d = 4$

Figure 2.3: Performance of the CDC classifier on the *Two moons* data set, with $p = 2$ and $N = 2000$; Computation time is expressed in seconds.

Figure 2.3. displays some convergence properties of the CDC classifier that are in line with Theorem 2.2.3. Indeed, it suffices to consider the Christoffel functions of order not higher than four (i.e. empirical moments of order eight) in order to extract the level sets that perfectly separate the points in the depicted data cloud.
On the same data set, the LDA classifier was much faster but achieved only 88% accuracy, demonstrating some advantages of the CDC classifier in the context of non-linearly separable data sets. SVM classifier, on the other hand, manages to achieve the absolute efficiency by consuming somehow comparable amount of time ($123ms$) as $\hat{f}_d^N$.

Similar results are obtained for other data shapes.

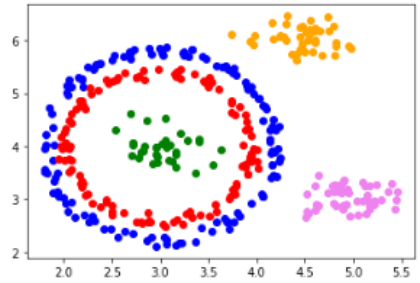


(a) $d = 1$ (b) $d = 4$

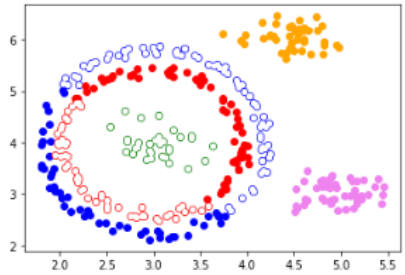
Figure 2.4: Embedded disks and CDC classifier, $p = 2$ and $N = 2000$.

What Figure 2.4. displays is a fifty percent accuracy gain obtained by considering polynomials of degree 8 instead of the much simpler degree two polynomials.
Another observation that can be made at this stage of analysis is related to the dependency of the CDC classifier on the sample size. It turns out that larger N is much less harmful in terms of computational complexity than, for example, larger p.

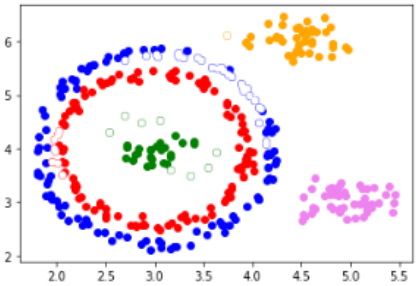
Let us end this section by including some multi-class performance evaluations.



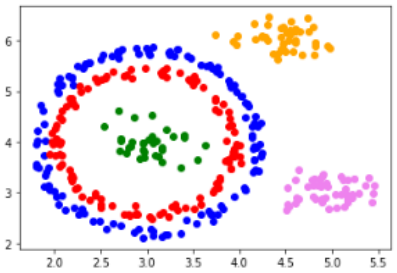
(a) CDC classifier ($d = 4$)



(b) LDA classifier



(c) QDA classifier



(d) SVM classifier

Figure 2.5: Comparison of various classifiers, $\boldsymbol{Y} = [\![1, 5]\!], p = 2$.

Even in the case of non-binary classification, the Christoffel-Darboux classifier seems to perform pretty well in terms of accuracy.

With only one tuning parameter, which has a very straightforward interpretation, the CDC classifier provides us with a classification tool that is deeply connected to the data generating process. This connection is established through the support identification, which takes place when the Christoffel-Darboux classifier is computed. This nice interpretability and quite natural parameter dependency are the main advantages of this classifier. On the other hand, being extremely computationally expensive in higher dimensions is the main drawback of this classification method.

2.4 Extensions

As emphasized previously, there are still many improvements to be made to the proposed classification method. Circumventing dimensionality issues and alleviating computational difficulties are among the main directions for the future work.

For example, it would be interesting to implement the CDC classifier using the relationship established in (1.29). This alternative way of computing the Christoffel function would make it

unnecessary to invert moment matrices of high dimensions, manifesting probably better numerical behaviour. However, this new method requires solving a completely new convex quadratic programming problem (which could be done efficiently using many solvers) whenever one needs to evaluate the Christoffel function at a different point. What is the trade-off between these two separate techniques remains to be investigated.

Alternatively, one may try to combine (1.29) or the variational formulation in (1.28) with neural networks and learn the Christoffel function. Indeed, in (1.28) one solves an optimization problem over the set of polynomials of a given degree. As such, that formulation is not tractable. But, one could consider optimizing the same objective function over the set of neural networks. Such an approach would result in an approximate CDC classifier which would not suffer from the curse of dimensionality. Since polynomials are continuous functions, and neural networks possess very interesting expressivity properties, one may hope to recover an optimal solution with desired error margin.

Additional research directions would include generalizing this approach to other tasks in the realm of data analysis. For example, it would be natural to examine how the Christoffel-function-based-approach behaves in the unsupervised learning framework (clustering). Here, one does not know *a priori* how many clusters actually exist, i.e. labels are not available. An efficient clustering algorithm should thus be able to deduce from the entire cloud of points (delineated by the level sets of one polynomial) how many natural sub-clouds exist, i.e., to decompose one *big* Christoffel function into multiple *small* Christoffel functions [6, 14]. Measuring the quality of such decompositions, while keeping the number of tuning parameters as small as possible is one of the main challenges.

Conclusion

In this report, we have provided some results related to the use of the Christoffel function in machine learning. More precisely, we have investigated how this tool from approximation theory domain can be adapted to handle supervised learning, i.e. classification tasks. Some of those results were purely theoretical, like in the first part, but found their use in practical implementations described in the second part.

As already mentioned, the scientific community has been aware for a very long time of many remarkable properties of the Christoffel-Darboux kernel, that were afterwards exploited mainly in the study of orthogonal polynomials. Recently, it was shown that this tool can be used in statistics as well, since the level sets of the Christoffel functions associated to empirical measures identify quite well the supports of those measures. This rather simple statement has motivated the content of the first part of the report. Indeed, getting familiar with the properties of the Christoffel function is a prerequisite for properly manipulating those properties when solving data science problems.

Furthermore, we have focused on solving the classification problem with a novel classifier based on the Christoffel function.

It turns out that a slightly modified Christoffel function represents the cornerstone for developing a classifier with some interesting properties. Indeed, we have demonstrated that such a classifier is quite easy to interpret, since it is profoundly connected to the data generating process. Moreover, it depends on only one tuning parameter, which is also a very much desired feature. In addition, some consistency-like results were stated that justify this approach from the theoretical point of view.

Finally, we have discussed numerical implementations and measured performance qualities of the proposed estimator. We have seen that, in terms of accuracy, the CDC classifier competes very well with many other state-of-the-art classifiers, and on different data sets. On the other hand, this classifier involves some pretty heavy computations that fail to make it suitable for the real-life classification problems, which are more than often high-dimensional. In any case, this approach represents a very promising direction for future works that could be useful in many machine learning applications.

Bibliography

[**1**] D. Henrion, J. B. Lasserre (2022) *Graph Recovery from Incomplete Moment Information*. Constructive Approximations https://doi.org/10.1007/s00365-022-09563-8

[**2**] E. Pauwels, M. Putinar, J. B. Lasserre (2018). *Data analysis from empirical moments and the Christoffel function.* arXiv:1810.08480 [stat.ML]

[**3**] E. Pauwels, J. B. Lasserre (2016). *Sorting out typicality with the inverse moment matrix SOS polynomial*. Advances in Neural Information Processing Systems 29 (NIPS 2016)

[**4**] J. B. Lasserre, E. Pauwels, M. Putinar (2022). *The Christoffel–Darboux Kernel for Data Analysis.* Cambridge University Press

[**5**] J. B. Lasserre (2022). *On the Christoffel function and classification in data analysis* . arXiv:2203.14571 [math.OC]

[**6**] J. B. Lasserre (2022). *A disintegration of the Christoffel function*. hal03624003

[**7**] J. B. Lasserre (2001). *Global optimization with polynomials and the problem of moments.* SIAM Journal on optimization, 11(3), 796-817.

[**8**] N. Aronszajn (1950). *Theory of Reproducing Kernels*. Transactions of the American Mathematical Society, Vol. 68, No. 3 (May, 1950), pp. 337-404

[**9**] P. Neval (1986). *Géza Freud, orthogonal polynomials and Christoffel functions. A case study.* Journal of Approximation Theory

[**10**] S. B. Kotsiantis, I. Zaharakis, and P. Pintelas (2007). *Supervised machine learning: A review of classification techniques*. Emerging artificial intelligence applications in computer engineering 160, no. 1: 3-24.

[**11**] S. Marx, E. Pauwels, T. Weisser, D. Henrion and J. B. Lasserre (2021). *Semi-algebraic Approximation Using Christoffel–Darboux Kernel*. https://doi.org/10.1007/s00365-021-09535-4

[**12**] S. L. Brunton and J. N. Kutz (2019). *Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control*. Cambridge University Press, Cambridge, UK

[**13**] S. T. Belinschi, V. Magron and V. Vinnikov (2021) . *Non-commutative Christoffel-Darboux Kernels*. arXiv preprint, arXiv:2106.06212.

[**14**] J. W. Helton, J. B. Lasserre and M. Putinar (2008). *Measures with zeros in the inverse of their moment matrix*. The Annals of Probability, 36(4), 1453-1471.