ON THE CHRISTOFFEL FUNCTION AND CLASSIFICATION IN DATA ANALYSIS

Jean-Bernard Lasserre

INSTITUT DE FRANCE
Académie des sciences

Comptes Rendus
Mathématique

Jean B. Lasserre

On the Christoffel function and classification in data analysis

Volume 360 (2022), p. 919-928

https://doi.org/10.5802/crmath.358

Les Comptes Rendus. Mathématique *sont membres du Centre Mersenne pour l'édition scientifique ouverte*
www.centre-mersenne.org
e-ISSN : 1778-3569

Comptes Rendus
Mathématique
2022, Vol. 360, p. 919-928
https://doi.org/10.5802/crmath.358

Algorithmic and computer tools / *Algorithmes et outils informatiques*

On the Christoffel function and classification in data analysis

Jean B. Lasserre [a]

[a] LAAS-CNRS and Institute of Mathematics, BP 54200, 7 Avenue du Colonel Roche, 31031 Toulouse cédex 4, France
E-mail: lasserre@laas.fr

Abstract. We show that the empirical Christoffel function associated with a cloud of finitely many points sampled from a distribution, can provide a simple tool for supervised classification in data analysis, with good generalization properties.

Résumé. Nous montrons que la fonction de Christoffel empirique associée à un échantillon fini de points peut fournir un outil simple pour la classification supervisée en analyse de données, avec de bonnes propriétés de généralisation.

2020 Mathematics Subject Classification. 41A30, 42C05, 47B32, 68T09, 94A16.

Funding. J.B. Lasserre is supported by the AI Interdisciplinary Institute ANITI funding through the french program "Investing for the Future PI3A" under the grant agreement number ANR-19-PI3A-0004.

Manuscript received 16 March 2022, revised and accepted 21 March 2022.

1. Introduction

In this note we are mainly concerned with supervised classification with noiseless deterministic labels where the objects of interest $\mathbf{x} \in \mathbf{X}$ belong to m classes with supports $\mathbf{X}_j \subset \mathbf{X} \subset \mathbb{R}^n$, $j \in [m]$ (with $[m] = \{1, \ldots, m\} =: \mathbf{Y}$). The supports satisfy $\mathbf{X}_i \cap \mathbf{X}_j = \emptyset$ for all i, j with $i \neq j$. The data set consists of clouds of finitely many points $(\mathbf{x}(i)) \subset \mathbf{X}_j$ sampled from an underlying distribution ϕ_j on $\mathbf{X}_j$, $j \in [m]$. In this situation, an *exact* classifier $f : \mathbf{X} \to \mathbf{Y}$, selects $j =: f(\mathbf{x})$ whenever $\mathbf{x} \in \mathbf{X}_j$. When constructing a classifier from a sample of data points, as e.g. in machine learning, a sensitive issue is its generalization properties when applied on a test set different from the training set. For the reader interested in recent developments on various techniques and issues in supervised and unsupervised classification, we refer to e.g. the book [1] and the many references therein.

Contribution. We first introduce a simple and natural ideal classifier $f_t : \mathbf{X} \to \mathbf{Y}$, with nice asymptotic properties as t increases. It is based on the Christoffel function Λ_t^μ associated with the joint distribution $\mathrm{d}\mu(\mathbf{x}, y)$ on $\mathbf{X} \times \mathbf{Y}$. As μ is supported on the graph $\{(\mathbf{x}, f(\mathbf{x})) : \mathbf{x} \in \mathbf{X}\}$ of the exact classifier f, recent results of [5] transported in our context suggest that the classifier $f_t(\mathbf{x}) := \arg\max_{y \in \mathbf{Y}} \Lambda_t^\mu(\mathbf{x}, y)$ should approximate f nicely. This is the case and indeed, by a slight modification of the definition of Λ_t^μ, we show that f_t is simply expressed in terms of the Christoffel functions $\Lambda_t^{\phi_j}$ of the ϕ_j; namely, $f_t(\mathbf{x}) = \arg\max_k \Lambda_t^{\phi_k}(\mathbf{x})$. Notice that this simple form of f_t mathematically justifies for supervised classification, the intuitive argument that $\Lambda_t^{\phi_j}(\mathbf{x}) > \Lambda_t^{\phi_k}(\mathbf{x})$, $\forall\, k \neq j$, whenever t is sufficiently large and $\mathbf{x} \in \mathbf{X}_j$. Indeed as $\mathbf{x} \in \mathbf{X}_j$ is outside the support $\mathbf{X}_k$ of ϕ_k, for every $k \neq j$, the "score" $\Lambda_t^{\phi_k}(\mathbf{x})$ is close to zero for sufficiently large t, as it decreases exponentially fast to zero (while the decrease of $\Lambda_t^{\phi_j}(\mathbf{x})$ is at most polynomial in t).

We next consider the practical case where we only have access to a discrete sample of points in each class $\mathbf{X}_j$ (e.g., the training set in Machine Learning) so that $\Lambda_t^{\phi_j}$ is not available. We provide a *data-driven* analogue of the previous result which, as expected, is in terms of the Christoffel functions $\Lambda_t^{\phi_{j,N}}$ associated with the discrete empirical measures $\phi_{j,N}$. Namely the empirical discrete analogue f_t^N of the classifier f_t simply reads $f_t^N(\mathbf{x}) = \arg\max_k \Lambda_t^{\phi_{k,N}}(\mathbf{x})$, and has same properties as f_t, but of course in an almost-sure sense with respect to random samples. In particular it shows good generalization properties. Indeed with $\varepsilon > 0$ fixed and t sufficiently large, with probability 1 (with respect to random samples), $f_t^N(\mathbf{x}) = j$ for every $j \in [m]$ and all $\mathbf{x} \in \mathbf{X}_j$ at distance at least ε from the boundary $\partial\mathbf{X}_j$, for sufficiently large N.

Finally, we also briefly discuss more general joint distributions of pairs $(\mathbf{x}, y)$ (where possibly $\mathbf{X}_i \cap \mathbf{X}_j \neq \emptyset$ for some (i, j)) which covers practical cases where some misclassification may occur and/or some ambiguity is allowed.

1.1. *Notation, definitions and preliminary results*

Let $\mathbb{R}[\mathbf{x}]$ denote the ring of real polynomials in the variables $\mathbf{x} = (x_1, \ldots, x_n)$ and $\mathbb{R}[\mathbf{x}]_t \subset \mathbb{R}[\mathbf{x}]$ be its subset of polynomials of total degree at most t. Let $\mathbb{N}^n_t := \{\boldsymbol{\alpha} \in \mathbb{N}^n : |\boldsymbol{\alpha}| \leq t\}$ (where $|\boldsymbol{\alpha}| = \sum_i \alpha_i$) with cardinal $s(t) = \binom{n+t}{n}$. Let $\mathbf{v}_t(\mathbf{x}) = (\mathbf{x}^{\boldsymbol{\alpha}})_{\boldsymbol{\alpha} \in \mathbb{N}^n_t}$ be the vector of monomials up to degree t.

The support of a Borel measure μ on $\mathbb{R}^n$ is the smallest closed set A such that $\mu(\mathbb{R}^n \setminus A) = 0$, and such a set A is unique.

Moment matrix. Let ϕ be a Borel measure whose support $\boldsymbol{\Omega} \subset \mathbb{R}^n$ is compact with nonempty interior. Its *moment matrix* of order (or degree) t is the real symmetric matrix $\mathbf{M}_t(\phi)$ with rows and columns indexed by $\mathbb{N}^n_t$, and with entries

$$\mathbf{M}_t(\phi)(\boldsymbol{\alpha}, \boldsymbol{\beta}) := \int_{\boldsymbol{\Omega}} \mathbf{x}^{\boldsymbol{\alpha}+\boldsymbol{\beta}}\, \mathrm{d}\phi = \phi_{\boldsymbol{\alpha}+\boldsymbol{\beta}}, \quad \boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{N}^n_t.$$

Then necessarily $\mathbf{M}_t$ is positive semidefinite for all t, denoted $\mathbf{M}_t(\phi) \succeq 0$.

Christoffel function. If $\boldsymbol{\Omega}$ has nonempty interior then $\mathbf{M}_t$ is positive definite for all t, denoted $\mathbf{M}_t(\phi) \succ 0$. Let $(P_{\boldsymbol{\alpha}})_{\boldsymbol{\alpha} \in \mathbb{N}^n} \subset \mathbb{R}[\mathbf{x}]$ be a family of polynomials, orthonormal with respect to ϕ, i.e.,

$$\int_{\boldsymbol{\Omega}} P_{\boldsymbol{\alpha}}\, P_{\boldsymbol{\beta}}\, \mathrm{d}\phi = \delta_{\boldsymbol{\alpha}=\boldsymbol{\beta}}, \quad \forall\, \boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{N}^n.$$

Then the Christoffel function (CF) $\Lambda_t^\phi : \mathbb{R}^n \to \mathbb{R}_+$ associated with ϕ, is defined by

$$\mathbf{x} \mapsto \Lambda_t^\phi(\mathbf{x}) := \left[\sum_{\boldsymbol{\alpha} \in \mathbb{N}^n_t} P_{\boldsymbol{\alpha}}(\mathbf{x})^2 \right]^{-1}, \quad \forall\, \mathbf{x} \in \mathbb{R}^n, \tag{1}$$

and recalling that $\mathbf{M}_t(\mu)$ is nonsingular, it turns out that

$$\Lambda_t^{\phi}(\mathbf{x}) = \left[\, \mathbf{v}_t(\mathbf{x})^T \mathbf{M}_t(\phi)^{-1} \mathbf{v}_t(\mathbf{x}) \,\right]^{-1}, \quad \forall\, \mathbf{x} \in \mathbb{R}^n. \tag{2}$$

An equivalent and variational definition is also

$$\Lambda_t^{\phi}(\mathbf{x}) = \inf_{p \in \mathbb{R}[\mathbf{x}]_t} \left\{ \int_{\mathbf{\Omega}} p^2 \, \mathrm{d}\phi \,:\, p(\mathbf{x}) = 1 \right\}, \quad \forall\, \mathbf{x} \in \mathbb{R}^n. \tag{3}$$

One interesting and distinguishing feature of the CF is that as t increases, $\Lambda_t^{\phi}(\mathbf{x}) \downarrow 0$ exponentially fast for every $\mathbf{x} \notin \text{support}(\phi)$. In other words Λ_t^{ϕ} identifies the support of ϕ when t is sufficiently large. In addition, at least in dimension $n = 2$ or $n = 3$, one may visualize this property even for small t, as the resulting superlevel sets $\mathbf{\Omega}_\gamma := \{\mathbf{x} : \Lambda_t^{\phi}(\mathbf{x}) \geq \gamma\}$, $\gamma \in \mathbb{R}$, capture the shape of $\mathbf{\Omega}$ quite well; see e.g. [2].

1.2. *Setting*

Let $\mathbf{Y} := \{1, 2, \ldots, m\}$ be the set of m classes, and for each (class) $j \in \mathbf{Y}$, let $\mathbf{X}_j \subset \mathbb{R}^n$ be the set of points in the class j, assumed to be open with compact closure $\overline{\mathbf{X}}_j$, Let $\mathbf{X} := \bigcup_{j=1}^m \mathbf{X}_j \subset \mathbb{R}^n$ be the open set (with compact closure $\overline{\mathbf{X}} = \bigcup_{j=1}^m \overline{\mathbf{X}}_j$) of all points to be classified and μ be the joint probability distribution of $(\mathbf{x}, y)$ on $\overline{\mathbf{X}} \times \mathbf{Y}$. Write

$$\mathrm{d}\mu(\mathbf{x}, y) = \varphi(\mathrm{d}y|\mathbf{x})\, \phi(\mathrm{d}\mathbf{x}), \tag{4}$$

where μ has been disintegrated into its marginal ϕ on $\overline{\mathbf{X}}$ and its conditional probability distribution $\varphi(dy|\mathbf{x})$ on $\mathbf{Y}$ given $\mathbf{x} \in \overline{\mathbf{X}}$. Next, each point $\mathbf{x} \in \mathbf{X}$ belongs to only one class and therefore $\mathbf{X}_i \cap \mathbf{X}_j = \emptyset$ for all pairs (i, j) with $i \neq j$, and so we may and will assume that $\phi(\overline{\mathbf{X}}_i \cap \overline{\mathbf{X}}_j) = 0$ for all pairs (i, j) with $i \neq j$. Therefore one may write $\mu = \sum_{j=1}^m \mu_j$, with

$$\mathrm{d}\mu_j(\mathbf{x}, y) = \delta_{\{j\}}(\mathrm{d}y)\, \phi_j(\mathrm{d}\mathbf{x}), \quad j \in \mathbf{Y},$$

for some marginals ϕ_j on $\overline{\mathbf{X}}_j$, $j \in \mathbf{Y}$. In particular:

$$\phi(A) = \mu(A \times \mathbf{Y}) = \sum_{j=1}^m \mu_j(A \times \mathbf{Y}) = \sum_{j=1}^m \phi_j(A), \quad \forall\, A \in \mathscr{B}(\mathbf{X}),$$

and therefore $\phi = \sum_{j=1}^m \phi_j$. Next, let $f : \overline{\mathbf{X}} \to \mathbf{Y}$ be the exact classifier

$$\mathbf{x} \mapsto f(\mathbf{x}) = \begin{cases} \sum_{j=1}^m j \cdot 1_{\mathbf{X}_j}(\mathbf{x}) & \text{if } \mathbf{x} \in \mathbf{X}, \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

So $f(\mathbf{x})$ identifies the class of $\mathbf{x} \in \mathbf{X}$ and returns $f(\mathbf{x}) = 0$ if $\mathbf{x}$ belong to some intersection $\overline{\mathbf{X}}_i \cap \overline{\mathbf{X}}_j$. Notice that one may write

$$\mathrm{d}\mu(\mathbf{x}, y) = \delta_{f(\mathbf{x})}(\mathrm{d}y)\, \phi(\mathrm{d}\mathbf{x}), \tag{6}$$

and so the joint distribution μ is supported on the graph $\mathbf{G} := \{(\mathbf{x}, f(\mathbf{x})) : \mathbf{x} \in \overline{\mathbf{X}}\}$ of the function f.

The Christoffel function is a powerful tool from the theory of approximation and orthogonal polynomials and one of its distinguishing features is its ability to identify the support of the underlying measure. So the CF Λ_t^{μ} associated with μ is an appropriate tool to approximate f since the graph of f is precisely the support of the measure μ in (6). In Marx et al. [5] the authors propose to approximate f (when $\|f\|_\infty < M$) by:

$$\widehat{f}_t(\mathbf{x}) := \arg\min_y \Lambda_t^{\mu + \varepsilon\mu_0}(\mathbf{x}, y)^{-1}, \quad \forall\, \mathbf{x} \in \mathbf{X}, \tag{7}$$

with a small $\varepsilon > 0$ and where μ_0 is a measure with a density w.r.t. Lebesgue measure, positive on $\mathbf{X} \times [-M, M]$. They prove nice theoretical convergence guarantees as t increases; see [5] for more details. Notice that in the present supervised classification framework, the function to

approximate is a *step function* so that the support of its graph is contained in a real algebraic variety, an even more specific case.

2. Main result

We first consider the ideal case of approximating the classifier f in (5) via the CF of the joint distribution μ in (6). Then we next consider the more practical setting (as in machine learning) where we only have access to a finite sample (the training set). In this case we use the *empirical* measures $\phi_{j,N}$ associated with the points of the sample in class "j". In Section 2.2 we invoke results from [4] that relate the degree t of the CF $\Lambda_t^{\phi_{j,N}}$ with the size N of the sample to ensure that important asymptotic properties of $\Lambda_t^{\phi_{j,N}}$ and $\Lambda_t^{\phi_j}$ as t and N increase, coincide.

2.1. *The CF on a real variety*

Let $q \in \mathbb{R}[\mathbf{x},y]$ be the polynomial $(\mathbf{x},y) \mapsto v(\mathbf{x},y) = \prod_{i=1}^m (y-i)$ and let μ be the probability measure on $\mathbf{\Omega} = \overline{\mathbf{X}} \times \mathbf{Y}$ defined in (6). Its support $\mathbf{\Omega}$ is contained in the real algebraic variety $V := \{(\mathbf{x},y) : v(\mathbf{x},y) = 0\} = \mathbb{R}^n \times \mathbf{Y}$ and the ideal $\mathscr{I} = \langle v \rangle \subset \mathbb{R}[\mathbf{x},y]$ generated by the polynomial v is the ideal of polynomials that vanish on V.

Observe that the moment matrix $\mathbf{M}_t(\mu)$ is singular since the vector $\mathbf{v}$ of coefficients of the polynomial v is in the kernel of $\mathbf{M}_t(\mu)$ as soon as $t \geq m$. Indeed $\mathbf{v}^T\mathbf{M}_t(\mu)\mathbf{v} = \int v^2 \,\mathrm{d}\mu = 0$ because the support of μ is contained in V. So the definition (2) of the CF is not valid any more. Denote by $L_t^2(\mu) \subset \mathbb{R}[\mathbf{x},y]$ the space of polynomials on V of total degree at most t (and degree at most $m-1$ in the variable y), equipped with the inner product and norm inherited from $L^2(\mathbf{\Omega},\mu)$. It turns out that $L_t^2(\mu)$ is a RKHS (Reproducing Kernel Hilbert Space). Then in this context, the variational definition (3) of the CF associated with μ reads

$$(\mathbf{x},y) \mapsto \Lambda_t^\mu(\mathbf{x},y) := \inf_{p \in L_t^2(\mu)} \left\{ \int_{\mathbf{\Omega}} p^2 \,\mathrm{d}\mu : p(\mathbf{x},y) = 1 \right\}, \quad \forall\, (\mathbf{x},y) \in V. \tag{8}$$

The set $\Gamma_t := \{\mathbf{x}^{\boldsymbol{\alpha}}\, y^k : k \leq m-1;\, |\boldsymbol{\alpha}| + k \leq t\} \subset \mathbb{R}[\mathbf{x},y]$ is a monomial basis of $L_t^2(\mu)$. Let $\mathbf{M}_t'(\mu)$ be the moment matrix associated with μ in (6) with rows and columns indexed by all monomials $(\mathbf{x}^{\boldsymbol{\alpha}}\, y^k)$ of Γ_t (and not all monomials $\mathbf{x}^{\boldsymbol{\alpha}}\, y^k$ of total degree at most t), e.g., listed according to the lexicographic ordering. Then $\mathbf{M}_t'(\mu)$ is non singular and

$$\Lambda_t^\mu(\mathbf{x},y)^{-1} = \mathbf{v}_t'(\mathbf{x},y)^T \mathbf{M}_t'(\mu)^{-1} \mathbf{v}_t'(\mathbf{x},y), \quad \forall\, (\mathbf{x},y) \in V,$$

where $\mathbf{v}_t'(\mathbf{x},y)$ is the vector of all monomials of Γ_t. Alternatively

$$\Lambda_t^\mu(\mathbf{x},y)^{-1} = \sum_{(\boldsymbol{\alpha},k) \in \Gamma_t} \frac{1}{\lambda_{\boldsymbol{\alpha},k}}\, Q_{\boldsymbol{\alpha},k}(\mathbf{x},y)^2,$$

where the $Q_{\boldsymbol{\alpha},k}$'s and the $\lambda_{\boldsymbol{\alpha},k}$'s are the eigenvectors and their respective eigenvalues associated with $\mathbf{M}_t'(\mu)$. Following [5], one may consider the perturbed measure $\mu + \varepsilon\mu_0$ where μ_0 is a probability uniformly distributed on $\overline{\mathbf{X}} \times [0,m]$ and $\varepsilon > 0$ is a small parameter. Then

$$\Lambda_t^{\mu+\varepsilon\mu_0}(\mathbf{x},y)^{-1} = \mathbf{v}_t(\mathbf{x},y)^T \mathbf{M}_t(\mu+\varepsilon\mu_0)^{-1} \mathbf{v}_t(\mathbf{x},y), \quad \forall\, (\mathbf{x},y) \in \mathbb{R}^n \times \mathbb{R}.$$

Recall that $\mathbf{v}_t(\mathbf{x},y)$ is the vector of *all* monomials $\mathbf{x}^{\boldsymbol{\alpha}}\, y^k$ of degree at most t, and $\mathbf{M}_t(\mu+\varepsilon\mu_0)$ is the moment matrix of $\mu+\varepsilon\mu_0$ of order t, which is non singular for all $\varepsilon > 0$. In addition $\Lambda_t^{\mu+\varepsilon\mu_0}$ is defined for *all* $(\mathbf{x},y) \in \mathbb{R}^{n+1}$ whereas Λ_t^μ in (8) is defined for all $(\mathbf{x},y) \in V$. Then as proved in [5] the classifier $\widehat{f}_t$ in (7) approximates f as t increases. Next we show that by a slightly change of the vector space $L_t^2(\mu)$ in (8), the resulting CF has nice additional properties that can be exploited to provide the resulting classifier (7) with a clear and more intuitive interpretation.

A slight variant of the CF. We now introduce a slight variant $\widehat{\Lambda}_t^\mu$ of the CF Λ_t^μ, defined by:

$$(\mathbf{x},y)\mapsto \widehat{\Lambda}_t^\mu(\mathbf{x},y) := \inf_{p\in\mathscr{L}_t^2(\mu)}\left\{\int_{\mathbf{\Omega}} p^2\,\mathrm{d}\mu : p(\mathbf{x},y)=1\right\},\quad \forall\,(\mathbf{x},y)\in V, \tag{9}$$

where $\mathscr{L}_t^2(\mu) =: \mathbb{R}[\mathbf{x},y]_{t,m-1}$ is the vector space of polynomials of degree at most t with respect to the variable $\mathbf{x}$ and at most $m-1$ with respect to the variable y, so that $L_t^2(\mu)\subset\mathscr{L}_t^2(\mu)\subset L_{t+m-1}^2(\mu)$.

Proposition 1. *Let Λ_t^μ and $\widehat{\Lambda}_t^\mu$ be as in* (8) *and* (9) *respectively. Then:*

$$\Lambda_{t+m-1}^\mu(\mathbf{x},y)\le\widehat{\Lambda}_t^\mu(\mathbf{x},y)\le\Lambda_t^\mu(\mathbf{x},y),\quad\forall\,(\mathbf{x},y)\in V. \tag{10}$$

Proof. Follows from $L_t^2(\mu)\subset\mathscr{L}_t^2(\mu)\subset L_{t+m-1}^2(\mu)$ and the definitions (8) and (9). □

Proposition 1 states that $\widehat{\Lambda}_t^\mu$ and Λ_t^μ are close but as we next see, $\widehat{\Lambda}_t^\mu$ has an interesting additional feature. Namely, it has a nice characterization in closed form which when exploited for classification leads to a classifier with a clear interpretation. Let $(\theta_j)_{j\in[m]}\subset\mathbb{R}[y]_{m-1}$ be the interpolation polynomials at the points $\{1,2,\ldots,m\}$ of $\mathbf{Y}$, i.e.,

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

which form an orthonormal family with respect to the uniform probability measure on $\mathbf{Y}$.

Theorem 2. *For each $j\in\mathbf{Y}$, let $(P_{\boldsymbol{\alpha}}^j)_{\boldsymbol{\alpha}\in\mathbb{N}}\subset\mathbb{R}[\mathbf{x}]$ be a family of polynomials that are orthonormal with respect to the marginal probability measure ϕ_j of μ_j, and let $\Lambda_t^{\phi_j}$ be the standard Christoffel function associated with ϕ_j on $\overline{\mathbf{X}}_j$. Then:*

(i) *The family $(\theta_j(y)\,P_{\boldsymbol{\alpha}}^j(\mathbf{x}))_{\boldsymbol{\alpha}\in\mathbb{N}_t^n}\subset\mathbb{R}[\mathbf{x},y]$ is an orthonormal basis of $\mathscr{L}_t^2(\mu)$.*

(ii) *The Christoffel function $\widehat{\Lambda}_t^\mu$ defined in* (9) *satisfies*

$$\widehat{\Lambda}_t^\mu(\mathbf{x},y)^{-1} = \sum_{j\in\mathbf{Y}}\theta_j(y)^2\sum_{\boldsymbol{\alpha}\in\mathbb{N}_t^n}P_{\boldsymbol{\alpha}}^j(\mathbf{x})^2,\quad\forall\,(\mathbf{x},y)\in\mathbb{R}^n\times\mathbf{Y} \tag{11}$$

$$= \sum_{j\in\mathbf{Y}}\theta_j(y)^2\,\Lambda_t^{\phi_j}(\mathbf{x})^{-1},\quad\forall\,(\mathbf{x},y)\in\mathbb{R}^n\times\mathbf{Y} \tag{12}$$

$$= \sum_{j\in\mathbf{Y}}\delta_{y=j}\,\Lambda_t^{\phi_j}(\mathbf{x})^{-1},\quad\forall\,(\mathbf{x},y)\in\mathbb{R}^n\times\mathbf{Y}. \tag{13}$$

Proof. (i). Every element of $\mathscr{L}_t^2(\mu)=\mathbb{R}[\mathbf{x},y]_{t,m-1}$ is of the form $\sum_{j=0}^{m-1}y^j\,q_j(\mathbf{x})$ with $q_j\in\mathbb{R}[\mathbf{x}]_t$. Hence consider a polynomial $u\in\mathscr{L}_t^2(\mu)$ in the form $u(\mathbf{x},y):=p(y)\,q(\mathbf{x})$ for some $q\in\mathbb{R}[\mathbf{x}]_t$ and some $p\in\mathbb{R}[y]_{m-1}$, arbitrary. Then as $(P_{\boldsymbol{\alpha}}^j)_{\boldsymbol{\alpha}\in\mathbb{N}_t^n}$ generates $\mathbb{R}[\mathbf{x}]_t$, observe that for every $j\in\mathbf{Y}$:

$$q(\mathbf{x}) = \sum_{\boldsymbol{\alpha}\in\mathbb{N}_t^n}q_{\boldsymbol{\alpha}}^j\,P_{\boldsymbol{\alpha}}^j(\mathbf{x}),\quad\forall\,\mathbf{x}\in\mathbb{R}^n,$$

for some coefficients $(q_{\boldsymbol{\alpha}}^j)_{\boldsymbol{\alpha}\in\mathbb{N}_t^n}$. Next, as the polynomials $(\theta_j)_{j\in\mathbf{Y}}$ generate $\mathbb{R}[y]_{m-1}$, write $p(y)=\sum_{j\in\mathbf{Y}}p_j\,\theta_j(y)$ for some coefficients $(p_j)_{j\in\mathbf{Y}}$, and therefore

$$u(\mathbf{x},y)=\sum_{j\in\mathbf{Y}}p_j\,(\theta_j(y)\,q(\mathbf{x}))=\sum_{\boldsymbol{\alpha}\in\mathbb{N}_t^n,\,j\in\mathbf{Y}}p_j\,q_{\boldsymbol{\alpha}}^j\left[\theta_j(y)\,P_{\boldsymbol{\alpha}}^j(\mathbf{x})\right].$$

Orthogonality. If $i\neq j$ then $\theta_i(y)\theta_j(y)=0$ everywhere on the support of μ and therefore

$$\int_{\mathbf{\Omega}}\theta_i(y)\,P_{\boldsymbol{\alpha}}^i(\mathbf{x})\,\theta_j(y)\,P_{\boldsymbol{\beta}}^j(\mathbf{x})\,\mathrm{d}\mu(\mathbf{x},y)=0,$$

whereas if $i = j$ then

$$\int_{\mathbf{\Omega}} \theta_i(y)^2\, P^i_{\boldsymbol{\alpha}}(\mathbf{x})\, P^i_{\boldsymbol{\beta}}(\mathbf{x})\, \mathrm{d}\mu(\mathbf{x},y) = \sum_{j=1}^m \int_{\mathbf{\Omega}} \theta_i(y)^2\, P^i_{\boldsymbol{\alpha}}(\mathbf{x})\, P^i_{\boldsymbol{\beta}}(\mathbf{x})\, \mathrm{d}\mu_j(\mathbf{x},y)$$
$$= \int_{\mathbf{\Omega}} \theta_i(y)^2\, P^i_{\boldsymbol{\alpha}}(\mathbf{x})\, P^i_{\boldsymbol{\beta}}(\mathbf{x})\, \mathrm{d}\mu_i(\mathbf{x},y) = \int_{\mathbf{\Omega}} P^i_{\boldsymbol{\alpha}}(\mathbf{x})\, P^i_{\boldsymbol{\beta}}(\mathbf{x})\, \mathrm{d}\phi_i(\mathbf{x}) = \delta_{\boldsymbol{\alpha}=\boldsymbol{\beta}}\,.$$

Next, as $\mathscr{L}^2_t(\mu) \subset \mathbb{R}[\mathbf{x},y]_{t,m-1}$, its cardinality is $r(t) := m\cdot\binom{n+t}{t}$ which is also the number of terms in the family $(\theta_j(y)\, P^j_{\boldsymbol{\alpha}}(\mathbf{x}))_{\boldsymbol{\alpha}\in\mathbb{N}^n_t,\, j\in\mathbf{Y}}$ which also generates $\mathscr{L}^2_t(\mu)$. Hence $(\theta_j(y)\, P^j_{\boldsymbol{\alpha}}(\mathbf{x}))_{\boldsymbol{\alpha}\in\mathbb{N}^n_t,\, j\in\mathbf{Y}}$ is an orthonormal basis of $\mathscr{L}^2_t(\mu)$.

(ii). Let $\widehat{\mathbf{M}}_t(\mu)$ be the moment matrix of degree t with rows and columns indexed by monomials $(\mathbf{x}^{\boldsymbol{\alpha}}\, y^j)_{\boldsymbol{\alpha}\in\mathbb{N}^n_t,\, 0\le j\le m-1}$ (e.g. with lexicographic ordering), and let $\widehat{\mathbf{v}}_t(\mathbf{x},y)$ be the vector of monomials $\mathbf{x}^{\boldsymbol{\alpha}}\, y^k$ listed with the same ordering. Observe that (9) reads

$$\widehat{\Lambda}^\mu_t(\mathbf{x},y) = \min_{\mathbf{p}} \{\, \mathbf{p}^T \widehat{\mathbf{M}}_t(\mu)\, \mathbf{p} :\, \langle \mathbf{p}, \widehat{\mathbf{v}}_t(\mathbf{x},y)\rangle = 1\,\}, \tag{14}$$

where $\mathbf{p} \in \mathbb{R}^{r(t)}$ is the vector of coefficients of $p \in \mathscr{L}^2_t(\mu)$ in that basis. Then (14) is a convex optimization problem whose optimal solution $\mathbf{p}^*$ satisfies $2\,\widehat{\mathbf{M}}_t(\mu)\,\mathbf{p}^* = \lambda^*\,\widehat{\mathbf{v}}_t(\mathbf{x},y)$ for some scalar λ^*. Hence $\lambda^* = 2\widehat{\Lambda}^\mu_t(\mathbf{x},y)$ and

$$\mathbf{p}^* = \widehat{\Lambda}^\mu_t(\mathbf{x},y)\, \widehat{\mathbf{M}}_t(\mu)^{-1}\, \widehat{\mathbf{v}}_t(\mathbf{x},y)$$

so that the corresponding polynomial $p^* \in \mathscr{L}^2_t(\mu)$ reads

$$p^*(\mathbf{u},z) = \widehat{\Lambda}^\mu_t(\mathbf{x},y)\, \widehat{\mathbf{v}}_t(\mathbf{u},z)^T\, \widehat{\mathbf{M}}_t(\mu)^{-1}\, \widehat{\mathbf{v}}_t(\mathbf{x},y)$$
$$= \widehat{\Lambda}^\mu_t(\mathbf{x},y) \sum_{\boldsymbol{\alpha}\in\mathbb{N}^n_t,\, j\in\mathbf{Y}} \theta_j(y)\,\theta_j(z)\, P^j_{\boldsymbol{\alpha}}(\mathbf{u})\, P^j_{\boldsymbol{\alpha}}(\mathbf{x}), \quad (\mathbf{u},z)\in\mathbb{R}^n\times\mathbf{Y},$$

and therefore

$$1 = p^*(\mathbf{x},y) = \widehat{\Lambda}^\mu_t(\mathbf{x},y) \sum_{\boldsymbol{\alpha}\in\mathbb{N}^n_t,\, j\in\mathbf{Y}} \theta_j(y)^2\, P^j_{\boldsymbol{\alpha}}(\mathbf{x})^2, \quad \forall\, (\mathbf{x},y)\in\mathbb{R}^n\times\mathbf{Y},$$

which is (11). In particular we also retrieve that

$$\widehat{\Lambda}^\mu_t(\mathbf{x},y)^{-1} = \widehat{\mathbf{v}}_t(\mathbf{x},y)^T\, \widehat{\mathbf{M}}_t(\mu)^{-1}\, \widehat{\mathbf{v}}_t(\mathbf{x},y), \quad \forall\, (\mathbf{x},y)\in V\,. \tag{15}$$

Next (12) follows from the definition of the Christoffel function associated with ϕ_j for each $j \in \mathbf{Y}$, and (13) follows from the properties of interpolation polynomials $(\theta_j)_{j\in\mathbf{Y}}$. □

So whenever $y \in \mathbf{Y}$, the Christoffel function $\widehat{\Lambda}^\mu_t(\mathbf{x},y)$ has a very simple expression (12), stated directly in terms of the Christoffel functions $(\Lambda^{\phi_j}_t(\mathbf{x}))_{j=1,\ldots,m}$ associated with the classes $j = 1,\ldots,m$. This is quite natural but is proper to the CF $\widehat{\Lambda}^\mu_t$ and not to the standard CF Λ^μ_t.

An ideal classifier. Given the Christoffel function $\widehat{\Lambda}^\mu_t$ defined in (12) and inspired by (7), a natural candidate classifier is the function

$$\mathbf{x} \mapsto \widehat{f}_t(\mathbf{x}) := \arg\min_{y\in\mathbf{Y}} \widehat{\Lambda}^\mu_t(\mathbf{x},y)^{-1} = \arg\max_{y\in\mathbf{Y}} \widehat{\Lambda}^\mu_t(\mathbf{x},y), \quad \forall\, \mathbf{x}\in\mathbf{X}, \tag{16}$$

which in view of (13) reads:

$$\mathbf{x} \mapsto \widehat{f}_t(\mathbf{x}) := \arg\max_{k\in[m]} \Lambda^{\phi_k}_t(\mathbf{x}), \quad \forall\, \mathbf{x}\in\mathbf{X}. \tag{17}$$

Observe that the "max" in (17) is over $y \in [m]$ and not over the interval $[0,m]$. This is because in supervised classification, we know that $f(\mathbf{x}) \in [m]$ for all $\mathbf{x} \in \mathbf{X}$. The rationale is the following: Let $\mathbf{x} \in \mathbf{X}_j$ be fixed arbitrary, so that $\mathbf{x} \notin \overline{\mathbf{X}}_k$ for every $k \neq j$. As t increases, $\Lambda^{\phi_k}_t(\mathbf{x})$ decreases to

zero exponentially while $\Lambda_t^{\phi_j}(\mathbf{x})$ decreases not faster than t^n. Therefore for t sufficiently large, necessarily $\Lambda_t^{\phi_k}(\mathbf{x}) < \Lambda_t^{\phi_j}(\mathbf{x})$ for all $k \neq j$, and so by (13),

$$\mathbf{x} \in \mathbf{X}_j \Rightarrow \exists\, t_0 \text{ s.t. } \widehat{f}_t(\mathbf{x}) = \arg\min_{y \in \mathbf{Y}} \widehat{\Lambda}_t^{\mu}(\mathbf{x}, y)^{-1} = \arg\max_k \Lambda_t^{\phi_k}(\mathbf{x}) = j\,, \quad \forall\, t \geq t_0\,.$$

An even stronger *almost-uniform* result holds. Let $\partial\mathbf{X}_j$ denote the boundary of the set $\mathbf{X}_j \subset \mathbb{R}^n$.

Theorem 3. *Let $\widehat{f}$ be as in* (17) *and let $\mathbf{X}_j^{\varepsilon} := \{\mathbf{x} \in \mathbf{X}_j : d(\mathbf{x}, \partial\mathbf{X}_j) > \varepsilon\}$ where $\varepsilon > 0$ is fixed. Assume that for every $j \in [m]$, ϕ_j has a density w.r.t. Lebesgue measure λ restricted to $\mathbf{X}_j$, bounded from below by $c > 0$. Then there exists t_ε such that $\widehat{f}_t(\mathbf{x}) = j$ for all $\mathbf{x} \in \mathbf{X}_j^{\varepsilon}$ and all $t \geq t_\varepsilon$.*

Proof. Let $s(t) := \binom{n+t}{t}$ and denote by $\mathrm{diam}(S)$ the diameter of a bounded set $S \subset \mathbb{R}^n$. Let $\gamma_k := \phi_k(\mathbf{X}_k)$, $k \in [m]$. If $\mathbf{x} \in \mathbf{X}_j^{\varepsilon}$ then $d(\mathbf{x}, \overline{\mathbf{X}}_k) > \varepsilon$ for all $k \neq j$. Hence by [3, Lemma 6.6] (and using that $\Lambda_t^{\phi_k/\gamma_k} = \Lambda_t^{\phi_k}/\gamma_k$),

$$\frac{\gamma_k}{s(t)}\,\Lambda_t^{\phi_k}(\mathbf{x})^{-1} \geq 2^{\frac{t\varepsilon}{\varepsilon + \mathrm{diam}(\mathbf{X}_k)}-3}\, t^{-n}\,(\frac{n}{e})^n \exp(-n^2/t)\,, \quad \forall\, \mathbf{x} \in \mathbf{X}_j^{\varepsilon}\,.$$

On the other hand, as $d(\mathbf{x}, \partial\mathbf{X}_j) > \varepsilon$, we can invoke [3, Lemma 6.2] extended to measures with density w.r.t. Lebesgue bounded from below by $c > 0$ (see [3, Assumption 3.11]) and use $(t+1)(t+2)(t+3)/((n+t+1)(n+t+2)(n+2t+6)) \geq 1/2(n+1)^3$, to obtain

$$\frac{\gamma_j}{s(t)}\,\Lambda_t^{\phi_j}(\mathbf{x})^{-1} \leq \frac{2\,\lambda(\mathbf{X}_j)}{c\,\varepsilon^n \omega_n}\,(1+n)^3\,, \quad \forall\, \mathbf{x} \in \mathbf{X}_j^{\varepsilon}\,,$$

where ω_n is the n-dimensional area of $\mathbb{S}^{n+1}$. Hence clearly there exists t_ε such that $\Lambda_t^{\phi_k}(\mathbf{x})^{-1} > \Lambda_t^{\phi_j}(\mathbf{x})^{-1}$ for all $k \neq j$ and all $\mathbf{x} \in \mathbf{X}_j^{\varepsilon}$, whenever $t \geq t_\varepsilon$. In particular, as the functions $\Lambda_t^{\phi_k}$ are strictly positive, continuous and $\overline{\mathbf{X}}_j$ is compact, there exists $a_j > 0$ such that

$$\forall\, \mathbf{x} \in \mathbf{X}_j^{\varepsilon}: \quad \Lambda_t^{\phi_k}(\mathbf{x}) < \Lambda_t^{\phi_j}(\mathbf{x}) - a_j\,, \quad \forall\, k \neq j\,. \tag{18}$$

Therefore the result follows from (13) and (16). □

2.2. *Application to supervised classification with noiseless deterministic labels*

In supervised classification we do not have access to the CF Λ_t^{μ} or $\widehat{\Lambda}_t^{\mu}$. We only have access to a sample of N points $\mathrm{Tr}_N = \{(\mathbf{x}(i), y(i)) : i = 1, \ldots, N\} \subset \mathbf{X}$ (the training data set) and a sample of test points (the test data set). For instance, in a typical Machine Learning (ML) approach one tries to *learn* a classifier function f in (5) from the supervised data Tr_N by computing parameters of a deep neural network that minimize some loss function. Usually, the number of parameters is very large (compared to N) making the resulting solution sensitive to a classical *overfitting* phenomenon. One way to attenuate this overfitting phenomenon is to add an appropriate regularization term to the loss function in the criterion to minimize.

One reason behind this overfitting phenomenon is that in minimizing the loss function, each data point $(\mathbf{x}(i), y(i))$ is treated *separately*. Ideally one should somehow consider the entire *training* set Tr_N itself and not its members separately. This is precisely what the CF function approach does. Indeed the training set Tr_N is used to construct the empirical (discrete) analogues $\phi_{j,N}$ of the measures ϕ_j, and their associated empirical Christoffel function $\Lambda_t^{\phi_{j,N}}$, now obtained from empirical moments. Remarkably, even though the geometry of the support of $\phi_{j,N}$ is quite trivial, the CF $\Lambda_t^{\phi_{j,N}}$ is still close to $\Lambda_t^{\phi_j}$ in a certain sense and the training set Tr_N can still be used to infer properties of the underlying measures ϕ_j. Hence, and importantly, even though the mathematical object $\Lambda_t^{\phi_{j,N}}$ is built from *individual* items, it is in fact concerned with the *cloud* of points of Tr_N in class $\{j\}$, rather than the points $\mathbf{x}(i)$ in that class taken separately. However

of course, for $\Lambda_t^{\phi_{j,N}}$ to recover asymptotic properties of $\Lambda_t^{\phi_j}$, the sample size N and the degree t cannot be chosen independently; see [3, 4] for more details.

Setting. For every $j \in [m]$, let $\mathrm{Tr}_N^j = (\mathbf{x}(i))_{i\le N} \subset \mathbf{X}_j$ be a training set for class $\{j\}$, where $\mathbf{x}(i)$ are i.i.d. random vectors with common distribution ϕ_j whose support is $\mathbf{X}_j$. So the whole training set Tr_N has a total of mN points where the N points in each class j are sampled from ϕ_j. For every fixed t, a natural approach suggested by (17), consists of:

- Computing the Christoffel function $\Lambda_t^{\phi_{j,N}}$ associated with the empirical prob. measure

$$\phi_{j,N} := \frac{1}{N}\sum_{\mathbf{x}(i)\in\mathbf{X}_j} \delta_{\{\mathbf{x}(i)\}}, \quad \forall\, j \in [m]. \tag{19}$$

  Following (2), $\Lambda_t^{\phi_{j,N}}(\mathbf{x})^{-1} = \mathbf{v}_t(\mathbf{x})^T \mathbf{M}_t(\phi_{j,N})^{-1}\mathbf{v}_t(\mathbf{x})$, for all $\mathbf{x}\in\mathbb{R}^n$ and all $j\in[m]$. The moments of $\phi_{j,N}$ are easily obtained by

$$\phi_{j,N}(\boldsymbol{\alpha}) := \left\{\frac{1}{N}\sum_i \mathbf{x}(i)^{\boldsymbol{\alpha}} : \mathbf{x}(i)\in \mathrm{Tr}_N^j\right\}, \quad \forall\, \boldsymbol{\alpha}\in\mathbb{N}^n,$$

  and the moment matrix $\mathbf{M}_t(\phi_{j,N})$ is nonsingular for sufficiently large t.
- Following (17), introduce the empirical classifier

$$\mathbf{x} \mapsto \widehat{f}_t^N(\mathbf{x}) := \arg\max_k \Lambda_t^{\phi_{k,N}}(\mathbf{x}), \quad \forall\, \mathbf{x}\in\mathbf{X}. \tag{20}$$

Then the empirical version of Theorem 3 reads as follows:

Theorem 4. *For every $j\in[m]$, let $(\mathbf{x}(i))_{i\le N}\subset\mathbf{X}_j$ be i.i.d. random vectors according to a distribution ϕ_j whose support is $\overline{\mathbf{X}}_j$ and which satisfies the assumption in Theorem 3. Let $\phi_{j,N}$ be as in* (19)*, and $\widehat{f}_t^N$ be as in* (20)*. Given $\varepsilon>0$ fixed, let $\mathbf{X}_j^\varepsilon := \{\mathbf{x}\in\mathbf{X}_j : d(\mathbf{x},\partial\mathbf{X}_j)>\varepsilon\}$, $j\in[m]$. Then there exists t_ε such that for all $t>t_\varepsilon$ fixed, with probability* 1 *(with respect to the random samples $\mathrm{Tr}_N^k\subset\mathbf{X}_k$, $k\in[m]$), $\widehat{f}_t^N(\mathbf{x}) = j$ for all $\mathbf{x}\in\mathbf{X}_j^\varepsilon$, for sufficiently large N.*

Proof. By Theorem 3 there exist t_ε and $a_j>0$, such that $\Lambda_t^{\phi_j}(\mathbf{x}) - \Lambda_t^{\phi_k}(\mathbf{x}) > a_j$ for all $k\neq j$, all $t\ge t_\varepsilon$ and all $\mathbf{x}\in\mathbf{X}_j^\varepsilon$; see (18). On the other hand, with t fixed, by [3, Theorem 3.13],

$$\text{For every } j\in[m]\text{:}\quad \sup_{\mathbf{x}\in\mathbb{R}^n}\left|\Lambda_t^{\phi_{j,N}}(\mathbf{x}) - \Lambda_t^{\phi_j}(\mathbf{x})\right| \overset{a.s.}{\to} 0, \quad \text{as } N\to\infty,$$

where the *"a.s."* is with respect to the random sample $\mathrm{Tr}_N^j\subset\mathbf{X}_j$. Hence for every $t>t_\varepsilon$ fixed, with probability 1 (with respect to the m random samples $\mathrm{Tr}_N^k\subset\mathbf{X}_k$, $k\in[m]$)

$$\forall\, \mathbf{x}\in\mathbf{X}_j^\varepsilon,\ \forall\, k\neq j: \quad \Lambda_t^{\phi_{k,N}}(\mathbf{x}) < \Lambda_t^{\phi_{j,N}}(\mathbf{x}) - a_j/2,$$

for sufficiently large N, and therefore $\widehat{f}_t^N(\mathbf{x}) = j$ for all $\mathbf{x}\in\mathbf{X}_j^\varepsilon$, $j\in[m]$. □

2.3. *The general case where the supports $\mathbf{X}_j$ are not disjoint*

We next briefly consider the case where the assumption $\mathbf{X}_i\cap\mathbf{X}_j=\emptyset$ for all $i\neq j$, is *not* satisfied. That is, misclassifications occur or some points $\mathbf{x}\in\mathbf{X}$ may indeed belong to several classes, as can be the case in some practical situations.

So let μ on $\overline{\mathbf{X}}\times\mathbf{Y}$, $\varphi(\mathrm{d}y|\cdot)$ on $\mathbf{Y}$ and ϕ on $\overline{\mathbf{X}}$ be as in (4), but now with possibly $\mathbf{X}_i\cap\mathbf{X}_j\neq\emptyset$ for some $i\neq j$. For each $j\in[m]$, introduce the measures ϕ_j on $\mathbf{X}$, $j\in[m]$, defined by:

$$\phi_j(\mathrm{d}\mathbf{x}) := \varphi(j\,|\,\mathbf{x})\,\phi(\mathrm{d}\mathbf{x}), \quad j\in\mathbf{Y}. \tag{21}$$

So in contrast to Section 2.1, $\phi_j(\mathbf{X}_i)\neq 0$ is allowed for $i\neq j$. Then an analogue of Theorem 2 reads:

Theorem 5. *For each $j \in \mathbf{Y}$, let $(P^j_{\boldsymbol{\alpha}})_{\boldsymbol{\alpha}\in\mathbb{N}} \subset \mathbb{R}[\mathbf{x}]$ be a family of polynomials that are orthonormal with respect to the measure ϕ_j on $\overline{\mathbf{X}}$ defined in* (21)*, and let $\Lambda_t^{\phi_j}$ be the standard Christoffel function associated with ϕ_j. Then:*

(i) *The family $(\theta_j(y)\, P^j_{\boldsymbol{\alpha}}(\mathbf{x}))_{\boldsymbol{\alpha}\in\mathbb{N}^n_t} \subset \mathbb{R}[\mathbf{x}, y]$ is an orthonormal basis of $\mathscr{L}^2_t(\mu)$.*

(ii) *The Christoffel function $\widehat{\Lambda}^{\mu}_t$ defined in* (9) *satisfies*

$$\widehat{\Lambda}^{\mu}_t(\mathbf{x}, y)^{-1} = \sum_{j\in\mathbf{Y}} \theta_j(y)^2 \sum_{\boldsymbol{\alpha}\in\mathbb{N}^n_t} P^j_{\boldsymbol{\alpha}}(\mathbf{x})^2, \quad \forall\, (\mathbf{x}, y) \in \mathbb{R}^n \times \mathbf{Y} \tag{22}$$

$$= \sum_{j\in\mathbf{Y}} \theta_j(y)^2\, \Lambda^{\phi_j}_t(\mathbf{x})^{-1} = \sum_{j\in[m]} \delta_{y=j}\, \Lambda^{\phi_j}_t(\mathbf{x})^{-1}, \quad \forall\, (\mathbf{x}, y) \in \mathbb{R}^n \times \mathbf{Y}. \tag{23}$$

Proof. **(i).** If $i \neq j$ then $\theta_i(y)\theta_j(y) = 0$ everywhere on the support of μ and therefore

$$\int_{\boldsymbol{\Omega}} \theta_i(y)\, P^i_{\boldsymbol{\alpha}}(\mathbf{x})\, \theta_j(y)\, P^j_{\boldsymbol{\beta}}(\mathbf{x})\, \mathrm{d}\mu(\mathbf{x}, y) = 0,$$

whereas if $i = j$ then

$$\begin{aligned}\int_{\boldsymbol{\Omega}} \theta_i(y)^2\, P^i_{\boldsymbol{\alpha}}(\mathbf{x})\, P^i_{\boldsymbol{\beta}}(\mathbf{x})\, \mathrm{d}\mu(\mathbf{x}, y) &= \int_{\mathbf{X}} P^i_{\boldsymbol{\alpha}}(\mathbf{x})\, P^i_{\boldsymbol{\beta}}(\mathbf{x}) \left(\int_{\mathbf{Y}} \theta_i(y)^2 \varphi(\mathrm{d}y|\mathbf{x})\right) \phi(\mathrm{d}\mathbf{x}) \\ &= \int_{\mathbf{X}} P^i_{\boldsymbol{\alpha}}(\mathbf{x})\, P^i_{\boldsymbol{\beta}}(\mathbf{x}) \left(\sum_{j\in[m]} \theta_i(j)^2 \varphi(j\,|\mathbf{x})\right) \phi(\mathrm{d}\mathbf{x}) \\ &= \int_{\mathbf{X}} P^i_{\boldsymbol{\alpha}}(\mathbf{x})\, P^i_{\boldsymbol{\beta}}(\mathbf{x})\, \phi_i(\mathrm{d}\mathbf{x}) = \delta_{\boldsymbol{\alpha}=\boldsymbol{\beta}}.\end{aligned}$$

(ii). The rest of the proof is similar to that of Theorem 2. □

By (23), $\Lambda^{\phi_j}_t$ is easily obtained as $\widehat{\Lambda}^{\mu}_t(\cdot, j)$, $j \in [m]$, and $\widehat{\Lambda}^{\mu}_t(\mathbf{x}, y)$ is in turn obtained for instance via (15) from the moment matrix $\widehat{\mathbf{M}}_t(\mu)$ of the joint distribution μ. As the CF is an appropriate tool for support inference, notice that intersections of super level sets $\mathbf{G}_{i,\gamma} \cap \mathbf{G}_{j,\gamma}$, with $\mathbf{G}_{i,\gamma} := \{\mathbf{x} : \Lambda^{\phi_i}_t(\mathbf{x}) \geq \gamma\}$, should provide indications on whether $\mathbf{X}_i \cap \mathbf{X}_j = \emptyset$ if $i \neq j$.

Once again the CF $\widehat{\Lambda}^{\mu}_t$ has the simple and nice expression (23) only in terms of the CF's $\Lambda^{\phi_j}_t$, which suggests to define a classifier $\widehat{f}_t(\mathbf{x})$ exactly as in (16). The only difference is the meaning of ϕ_j and its implications. For instance if $\mathbf{X}_i \cap \mathbf{X}_j \neq \emptyset$ then a point $\mathbf{x} \in \mathbf{X}$ can belong to $\mathrm{supp}(\phi_i) \cap \mathrm{supp}(\phi_j)$ with $i \neq j$, and therefore the two scores $\Lambda^{\phi_i}_t(\mathbf{x})$ and $\Lambda^{\phi_j}_t(\mathbf{x})$ can be comparable even for large t, whereas before for sufficiently large t, the score $\Lambda^{\phi_j}_t(\mathbf{x})$ (with $j = \mathrm{class}(\mathbf{x})$) clearly dominates all other scores. In particular there is no analogue of Theorem 3. Evaluating how efficient the resulting classifier $\widehat{f}_t$ can be in a practical empirical context of sampled data as in Section 2.2, is beyond the scope of the present note.

3. Conclusion

The Christoffel function can provide a simple tool in supervised classification, with some theoretical guarantees. However to obtain $\Lambda^{\phi_{j,N}}_t$ *explicitly* one must handle matrices of size $\binom{n+t}{n}$ (inversion or eigenvectors) with a computational cost $O(t^n)$ that grows rapidly with t. Therefore so far, in this form this tool is limited to small dimension problems. On the other hand, evaluation of $\Lambda^{\phi_{j,N}}_t(\boldsymbol{\xi})$ at a point $\boldsymbol{\xi} \in \mathbb{R}^n$ via (8) only requires to solve a simple convex quadratic optimization problem, which can be done efficiently even for large n. Finally a detailed analysis of possible learning rates that can be obtained with this method remains to be done.

References

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

[2] J. B. Lasserre, E. Pauwels, "Sorting out typicality via the inverse moment matrix SOS polynomial", in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2016, p. 190-198.

[3] ———, "The empirical Christoffel function with applications in data analysis", *Adv. Comput. Math.* **45** (2019), no. 3, p. 1439-1468.

[4] J. B. Lasserre, E. Pauwels, M. Putinar, *The Christoffel–Darboux Kernel for Data Analysis*, Cambridge Monographs on Applied and Computational Mathematics, vol. 38, Cambridge University Press, 2022.

[5] S. Marx, E. Pauwels, T. Weisser, D. Henrion, J. B. Lasserre, "Semi-algebraic approximation using Christoffel–Darboux kernel", *Constr. Approx.* **54** (2021), no. 3, p. 391-429.