The Christoffel function, moments & sums-of-squares

Jean B. Lasserre

LAAS-CNRS, Institute of Mathematics & ANITI, Toulouse, France

Semidefinite Programming & Polynomial Optimization
CWI, 2022

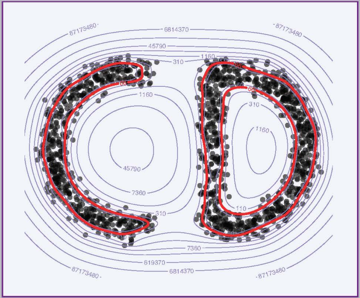
Research funded by ANITI institute, under grant ANR-19-PIA3-0004 and ANR-NUsCAP-20-CE48-0014

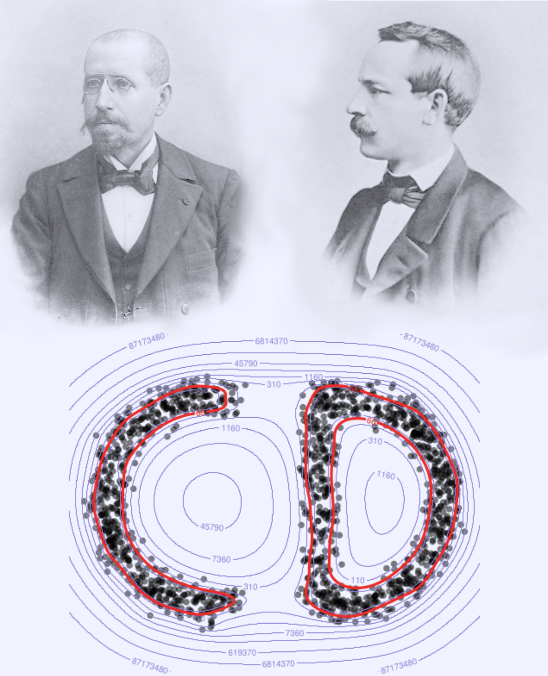
- The Christoffel function
- Some applications
- A link with sums-of-squares

Cambridge Monographs on Applied and Computational Mathematics

The Christoffel–Darboux Kernel for Data Analysis

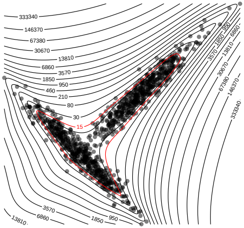
Jean Bernard Lasserre, Edouard Pauwels and Mihai Putinar




87173480
6814370
87173480
45790
310
1160
66
66
1160
310
1160
45790
7360
310
110
7360
619370
87173480
6814370
87173480

Motivation

Consider the following cloud of $2D$-points (data set) below



The red curve is the level set

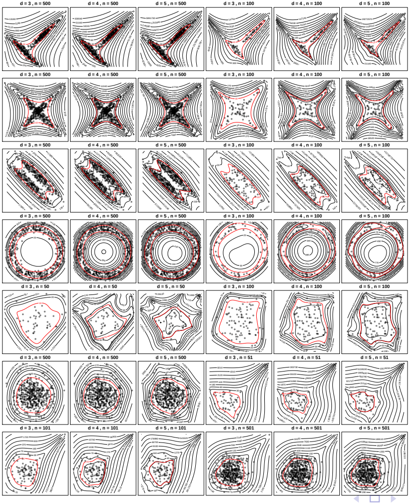
$$S_\gamma \; := \; \{\, \mathbf{x} : \; Q_d(\mathbf{x}) \leq \gamma \,\}, \qquad \gamma \in \mathbb{R}_+$$

of a certain polynomial $Q_d \in \mathbb{R}[x_1, x_2]$ of degree $2d$.

☞ Notice that S_γ captures quite well the shape of the cloud.

Not a coincidence!

☞ Surprisingly, low degree d for Q_d is often enough to get a pretty good idea of the shape of Ω (at least in dimension $p = 2, 3$)



Cook up your own convincing example

Perform the following simple operations on a preferred cloud of $2D$-points: So let $d = 2$, $p = 2$ and $s(d) = \binom{p+d}{p}$.

- Let $\mathbf{v}_d(\mathbf{x})^T = (1, x_1, x_2, x_1^2, x_1 x_2, \ldots, x_1 x_2^{d-1}, x_2^d)$. be the vector of all monomials $x_1^i x_2^j$ of total degree $i + j \leq d$
- Form the real symmetric matrix of size $s(d)$

$$\mathbf{M}_d := \frac{1}{N} \sum_{i=1}^{N} \mathbf{v}_d(\mathbf{x}(i))\, \mathbf{v}_d(\mathbf{x}(i))^T,$$

where the sum is over all points $(\mathbf{x}(i))_{i=1\ldots,N} \subset \mathbb{R}^2$ of the data set.

So the matrix $N \cdot \mathbf{M}_d$ reads:

$$\sum_i \begin{bmatrix} 1 & x_1(i) & x_2(i) & x_1(i)^2 & \dots & x_2(i)^d \\ x_1(i) & x_1(i)^2 & x_1(i)\,x_2(i) & x_1(i)^3 & \dots & x_1(i)\,x_2(i)^d \\ x_2(i) & x_1(i)x_2(i) & x_2(i)^2 & x_1(i)^2x_2(i) & \dots & x_2(i)^{d+1} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ x_2(i)^d & x_1(i)\,x_2(i)^d & x_2(i)^{d+1} & x_1(i)^2x_2(i)^d & \dots & x_2(i)^{2d} \end{bmatrix}$$

☞ Note that typically, $\mathbf{M}_d$ is what is called the MOMENT-matrix of the empirical measure

$$\mu^N := \frac{1}{N}\sum_{i=1}^{N} \delta_{\mathbf{x}(i)}$$

associated with a sample of size N, drawn according to an unknown measure μ.

☞ The (usual) notation $\delta_{\mathbf{x}(i)}$ stands for the DIRAC measure supported at the point $\mathbf{x}(i)$ of $\mathbb{R}^2$.

- Next, form the SOS polynomial:

$$\mathbf{x} \mapsto Q_d(\mathbf{x}) := \mathbf{v}_d(\mathbf{x})^T \mathbf{M}_d^{-1} \mathbf{v}_d(\mathbf{x}).$$

$$= (1, x_1, x_2, x_1^2, \ldots, x_2^d) \mathbf{M}_d^{-1} \begin{pmatrix} 1 \\ x_1 \\ x_2 \\ x_1^2 \\ \ldots \\ x_2^d \end{pmatrix}$$

- Plot some level sets

$$S_\gamma := \{ \mathbf{x} \in \mathbb{R}^2 : Q_d(\mathbf{x}) = \gamma \}$$

for some values of γ, the thick one representing the particular value $\gamma = \binom{2+d}{2}$.

The Christoffel function $\Lambda_d : \mathbb{R}^p \to \mathbb{R}_+$ is the reciprocal

$$\mathbf{x} \mapsto Q_d(\mathbf{x})^{-1}, \quad \forall \mathbf{x} \in \mathbb{R}^p$$

of the SOS polynomial Q_d.

☞ It has a rich history in Approximation theory and Orthogonal Polynomials.

☞ Among main contributors: Nevai, Totik, Króo, Lubinsky, Simon, . . .

☞ ... The CF seems to be not so well-known in data analysis

The Christoffel function $\Lambda_d : \mathbb{R}^p \to \mathbb{R}_+$ is the reciprocal

$$\mathbf{x} \mapsto Q_d(\mathbf{x})^{-1}, \quad \forall \mathbf{x} \in \mathbb{R}^p$$

of the SOS polynomial Q_d.

☞ It has a rich history in Approximation theory and Orthogonal Polynomials.

☞ Among main contributors: Nevai, Totik, Króo, Lubinsky, Simon, . . .

☞ ... The CF seems to be not so well-known in data analysis

Let μ be a (positive) measure supported on a compact set $\Omega \subset \mathbb{R}^p$ with nonempty interior.

> A family $(P_\alpha)_{\alpha\in\mathbb{N}^p} \subset \mathbb{R}[\mathbf{x}]$ is orthonormal w.r.t. μ if
>
> $$\int_\Omega P_\alpha(\mathbf{x})\, P_\beta(\mathbf{x})\, \mu(d\mathbf{x}) \,=\, \delta_{\alpha=\beta}\,, \quad \forall \alpha, \beta \in \mathbb{N}^p\,.$$

☞ Here $\delta_{\alpha=\beta}$ is the standard Kronecker symbol

How to construct a family $(P_\alpha)_{\alpha\in\mathbb{N}^p}$

Let $\mathbb{N}^p_t := \{\alpha \in \mathbb{N}^p : \sum_i \alpha_i \leq t\}$ and suppose that all moments

$$\mu_\alpha \;:=\; \int_\Omega \mathbf{x}^\alpha \, d\mu\,, \quad \forall \alpha \in \mathbb{N}^p_{2t}\,,$$

are available.

☞ Then one may construct an orthonormal family $(P_\alpha)_{\alpha\in\mathbb{N}^p_t}$ from determinants of moment matrices associated with μ.

The moment matrix $\mathbf{M}_d(\mu)$ is the real symmetric matrix with rows and columns indexed by $(\mathbf{x}^\alpha)_{\alpha\in\mathbb{N}^p_d}$, and with entries

$$\mathbf{M}_d(\mu)(\alpha,\beta) \;:=\; \int_\Omega \mathbf{x}^{\alpha+\beta}\, d\mu \;=\; \mu_{\alpha+\beta}\,, \quad \forall \alpha,\beta\in\mathbb{N}^p_d\,.$$

☞ Illustrative example in dimension 2:

$$\mathbf{M}_1(\mu) \;:=\; \begin{pmatrix} & 1 & X_1 & X_2 \\ 1 & \mu_{00} & \mu_{10} & \mu_{01} \\ X_1 & \mu_{10} & \mu_{20} & \mu_{11} \\ X_2 & \mu_{01} & \mu_{11} & \mu_{02} \end{pmatrix}$$

is the *moment matrix of* μ *of "degree d=1"*.

One way to construct polynomials orthonormal w.r.t. μ

Fix an ordering of $\mathbb{N}^p$ (e.g. lexicographic ordering)

$$\underbrace{(0,0)}_{\text{degree0}},\underbrace{(1,0),(0,1)}_{\text{degree1}},\underbrace{(2,0),(1,1),(0,2)}_{\text{degree2}},(3,0),(2,1),\ldots$$

Then $P_{00}(\mathbf{x}) = 1$ for all $\mathbf{x} = (x_1, x_2) \in \mathbb{R}^2$.

$$Q_{10}(\mathbf{x}) := \det \begin{pmatrix} \mu_{00} & \mu_{10} \\ 1 & X_1 \end{pmatrix} = X_1 - \mu_{10}.$$

$$Q_{01}(\mathbf{x}) := \det \begin{pmatrix} \mu_{00} & \mu_{10} & \mu_{01} \\ \mu_{10} & \mu_{20} & \mu_{11} \\ 1 & X_1 & X_2 \end{pmatrix}$$

$$= \mu_{10}\mu_{11} - \mu_{01}\mu_{20} - X_1\,(\mu_{00}\mu_{11} - \mu_{10}\mu_{01}) + X_2\,(\mu_{00}\mu_{20} - \mu_{10}^2)$$

☞ Then normalize, i.e. $P_{10} = \theta Q_{10}$ with θ such that

$$\theta^2 \int_\Omega Q_{10}^2 \, d\mu = 1\,.$$

and similarly with $P_{01} = \theta\, Q_{01}$.

One way to construct polynomials orthonormal w.r.t. μ

Fix an ordering of $\mathbb{N}^p$ (e.g. lexicographic ordering)

$$\underbrace{(0,0)}_{\text{degree0}},\underbrace{(1,0),(0,1)}_{\text{degree1}},\underbrace{(2,0),(1,1),(0,2)}_{\text{degree2}},(3,0),(2,1),\ldots$$

Then $P_{00}(\mathbf{x}) = 1$ for all $\mathbf{x} = (x_1, x_2) \in \mathbb{R}^2$.

$$Q_{10}(\mathbf{x}) := \det \begin{pmatrix} \mu_{00} & \mu_{10} \\ 1 & X_1 \end{pmatrix} = X_1 - \mu_{10}.$$

$$Q_{01}(\mathbf{x}) := \det \begin{pmatrix} \mu_{00} & \mu_{10} & \mu_{01} \\ \mu_{10} & \mu_{20} & \mu_{11} \\ 1 & X_1 & X_2 \end{pmatrix}$$

$$= \mu_{10}\mu_{11} - \mu_{01}\mu_{20} - X_1\,(\mu_{00}\mu_{11} - \mu_{10}\mu_{01}) + X_2\,(\mu_{00}\mu_{20} - \mu_{10}^2)$$

☞ Then normalize, i.e. $P_{10} = \theta Q_{10}$ with θ such that

$$\theta^2 \int_{\Omega} Q_{10}^2 \, d\mu = 1\,.$$

and similarly with $P_{01} = \theta\, Q_{01}$.

Similarly,

$$Q_{20}(\mathbf{x}) := \det \begin{pmatrix} \mu_{00} & \mu_{10} & \mu_{01} & \mu_{20} \\ \mu_{10} & \mu_{20} & \mu_{11} & \mu_{30} \\ \mu_{01} & \mu_{11} & \mu_{02} & \mu_{21} \\ 1 & X_1 & X_2 & X_1^2 \end{pmatrix}$$

$$= X_1^2 \det \begin{pmatrix} \mu_{00} & \mu_{10} & \mu_{01} \\ \mu_{10} & \mu_{20} & \mu_{11} \\ \mu_{01} & \mu_{11} & \mu_{02} \end{pmatrix} - X_2 (\cdots) + X_1 (\cdots) - (\cdots).$$

and $P_{20} = \theta\, Q_{20}$ with θ such that

$$\theta^2 \int_{\Omega} Q_{20}^2 \, d\mu = 1\,.$$

The vector space $\mathbb{R}[\mathbf{x}]_d$ viewed as a subspace of $L^2(\mu)$ is a Reproducing Kernel Hilbert Space (RKHS) .
Its *reproducing kernel*

$$(\mathbf{x},\mathbf{y}) \mapsto K_d^{\mu}(\mathbf{x},\mathbf{y}) \;:=\; \sum_{|\alpha|\leq d} P_\alpha(\mathbf{x})\,P_\alpha(\mathbf{y})\,,\quad \forall\,\mathbf{x},\mathbf{y}\in\mathbb{R}^p\,,$$

is called the *Christoffel-Darboux kernel*.

The reproducing property

$$\mathbf{x} \mapsto q(\mathbf{x}) \;=\; \int_\Omega K_d^\mu(\mathbf{x}, y)\, q(y)\, d\mu(y)\,, \quad \forall q \in \mathbb{R}[\mathbf{x}]_d\,.$$

☞ useful to determine the best degree-d $L^2(\mu)$-polynomial approximation

$$\inf_{q\in\mathbb{R}[\mathbf{x}]_d} \|f - q\|_{L^2(\mu)}$$

of $f \in L^2(\mu)$. Indeed:

$$\begin{aligned}\mathbf{x} \mapsto \widehat{f_d}(\mathbf{x}) \;&:=\; \sum_{\alpha\in\mathbb{N}^p_d} \overbrace{\Big(\int_\Omega f(y)\, P_\alpha(y)\, d\mu\Big)}^{\widehat{f_{d,\alpha}}} P_\alpha(\mathbf{x}) \;\in\; \mathbb{R}[\mathbf{x}]_d \\ &=\; \arg\min_{q\in\mathbb{R}[\mathbf{x}]_d} \|f - q\|_{L^2(\mu)}\end{aligned}$$

and

$$\int_\Omega (f - \widehat{f_d})^2 \, d\mu \;\rightarrow\; 0 \quad \text{as } d \rightarrow \infty$$

or, equivalently:

$$\lim_{d\to\infty} \|f - \widehat{f_d}\|_{L^2(\mu)} \;=\; 0\,.$$

Recall that the support Ω of μ is compact with nonempty interior, and let $(P_\alpha)_{\alpha\in\mathbb{N}^p}$ be a family of orthonormal polynomials w.r.t. μ.

Theorem

The Christoffel function $\Lambda_d^\mu : \mathbb{R}^p \to \mathbb{R}_+$ *is defined by:*

$$\xi \mapsto \Lambda_d^\mu(\xi)^{-1} = \sum_{|\alpha|\leq d} P_\alpha(\xi)^2 = K_d^\mu(\xi,\xi), \quad \forall \xi \in \mathbb{R}^p,$$

and it also satisfies the variational property:

$$\Lambda_d^\mu(\xi) = \min_{P\in\mathbb{R}[\mathbf{x}]_d} \left\{ \int_\Omega P^2\, d\mu : \; P(\xi) = 1 \right\}, \quad \forall \xi \in \mathbb{R}^p.$$

☞ Alternatively

$$\Lambda_d^\mu(\xi)^{-1} = \mathbf{v}_d(\xi)^T \mathbf{M}_d(\mu)^{-1}\, \mathbf{v}_d(\xi), \quad \forall \xi \in \mathbb{R}^p.$$

Recall that the support Ω of μ is compact with nonempty interior, and let $(P_\alpha)_{\alpha\in\mathbb{N}^p}$ be a family of orthonormal polynomials w.r.t. μ.

Theorem

The Christoffel function $\Lambda_d^\mu : \mathbb{R}^p \to \mathbb{R}_+$ is defined by:

$$\xi \mapsto \Lambda_d^\mu(\xi)^{-1} = \sum_{|\alpha|\leq d} P_\alpha(\xi)^2 = K_d^\mu(\xi,\xi), \quad \forall \xi \in \mathbb{R}^p,$$

and it also satisfies the variational property:

$$\Lambda_d^\mu(\xi) = \min_{P\in\mathbb{R}[\mathbf{x}]_d} \left\{ \int_\Omega P^2\, d\mu : \ P(\xi) = 1 \right\}, \quad \forall \xi \in \mathbb{R}^p.$$

☞ Alternatively

$$\Lambda_d^\mu(\xi)^{-1} = \mathbf{v}_d(\xi)^T \mathbf{M}_d(\mu)^{-1}\, \mathbf{v}_d(\xi), \quad \forall \xi \in \mathbb{R}^p.$$

Notice that we can also write:

$$\Lambda_d^{\mu}(\xi)^{-1} \,=\, K_d^{\mu}(\xi,\xi) \,=\, \sum_{|\alpha|\leq d} \frac{Q_\alpha(\xi)^2}{\lambda_\alpha}, \quad \forall \xi \in \mathbb{R}^p,$$

where the vector of coefficients of the polynomial Q_α is the normalized eigenvector of $\mathbf{M}_d(\mu)$ associated with the (positive) eigenvalue $\lambda_\alpha > 0$

☞ Importantly, and crucial for applications, the Christoffel function identifies the support Ω of the underlying measure μ.

Theorem

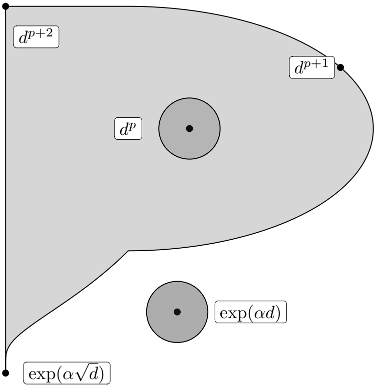
Let the support Ω of μ be compact with nonempty interior. Then:

- *For all* $\mathbf{x} \in \mathrm{int}(\Omega)$: $K_d^\mu(\mathbf{x}, \mathbf{x}) = O(d^p)$.
- *For all* $\mathbf{x} \in \mathrm{int}(\mathbb{R}^p \setminus \Omega)$: $K_d^\mu(\mathbf{x}, \mathbf{x}) = \Omega(\exp(\alpha d))$ *for some* $\alpha > 0$.

☞ In particular, as $d \to \infty$,

$$d^p \Lambda_d^\mu(\mathbf{x}) \to 0 \text{ very fast whenever } \mathbf{x} \notin \Omega.$$

Typical growth rates for $K_d^\mu(\mathbf{x},\mathbf{x}) = \Lambda_d^\mu(\mathbf{x})^{-1}$.



Some other properties

- Under some (restrictive) assumption on Ω and μ

$$\lim_{d\to\infty} s(d)\,\Lambda_d^{\mu}(\xi) \;=\; f_\mu(\xi)\,\omega(\xi)^{-1}$$

where ω is the density of an equilibrium measure intrinsically associated with Ω.
For instance with $p = 1$ and $\Omega = [-1, 1]$, $\omega(\xi) = \sqrt{1-\xi^2}$.

- If μ and ν have same support Ω and respective densities f_μ and f_ν w.r.t. Lebesgue measure on Ω, positive on Ω, then:

$$\lim_{d\to\infty} \frac{\Lambda_d^{\mu}(\xi)}{\Lambda_d^{\nu}(\xi)} = \frac{f_\mu(\xi)}{f_\nu(\xi)}, \quad \forall \xi \in \Omega\,.$$

☞ useful for density approximation

Some other properties

- Under some (restrictive) assumption on Ω and μ

$$\lim_{d\to\infty} s(d)\,\Lambda_d^\mu(\xi) \;=\; f_\mu(\xi)\,\omega(\xi)^{-1}$$

  where ω is the density of an equilibrium measure intrinsically associated with Ω.
  For instance with $p = 1$ and $\Omega = [-1, 1]$, $\omega(\xi) = \sqrt{1-\xi^2}$.
- If μ and ν have same support Ω and respective densities f_μ and f_ν w.r.t. Lebesgue measure on Ω, positive on Ω, then:

$$\lim_{d\to\infty} \frac{\Lambda_d^\mu(\xi)}{\Lambda_d^\nu(\xi)} \;=\; \frac{f_\mu(\xi)}{f_\nu(\xi)}, \quad \forall \xi \in \Omega\,.$$

☞ useful for density approximation

If Ω is not full-dimensional and is supported on a real variety $V \subset \mathbb{R}^p$, then for sufficiently large degree d:

$$d \mapsto \mathrm{rank}(\mathbf{M}_d) = q(d)$$

where $q \in \mathbb{R}[t]$ is the Hilbert polynomial associated with V and whose degree provides the dimension of V.

So one may use the rank of the moment matrix $\mathbf{M}_d$ to identify the dimension of the underlying variety.

☞ useful for manifold learning.

The Christoffel function can also be used in several important applications of Machine Learning (e.g. outlier detection, density approximation, manifold learning). In this case the measure μ is the empirical probability measure μ^N associated with a cloud of N points $\mathcal{C} \subset \mathbb{R}^p$ (the data of interest).

☞ Computing $\Lambda_d^{\mu^N}$ requires only one pass over the data & no optimization

The Christoffel function can also be used in several important applications of Machine Learning (e.g. outlier detection, density approximation, manifold learning). In this case the measure μ is the empirical probability measure μ^N associated with a cloud of N points $\mathcal{C} \subset \mathbb{R}^p$ (the data of interest).

☞ Computing $\Lambda_d^{\mu^N}$ requires only one pass over the data & no optimization

The Christoffel function can also be used in several important applications of Machine Learning (e.g. outlier detection, density approximation, manifold learning). In this case the measure μ is the empirical probability measure μ^N associated with a cloud of N points $\mathcal{C} \subset \mathbb{R}^p$ (the data of interest).

☞ Computing $\Lambda_d^{\mu^N}$ requires only one pass over the data & no optimization

☞ Rank-one update

Updating the Christoffel function when the cloud of N points one additional point ξ is added to the cloud of N points is easy.

$$(N+1)\,\mu^{N+1} \,=\, \sum_{i=1}^{N} \delta_{\mathbf{x}(i)} + \delta_{\xi} \,=\, N\,\mu^{N} + \delta_{\xi}$$

By Sherman-Morrison's rank-one update formula

$$\begin{aligned}((N+1)\,\mathbf{M}_d(\mu^{N+1}))^{-1} &= (N\,\mathbf{M}_d(\mu^N) + \mathbf{v}_d(\xi)\mathbf{v}_d(\xi)^T)^{-1}\\ &= (N\,\mathbf{M}_d(\mu^N))^{-1} - \\ &\quad \frac{1}{N^2}\,\frac{\mathbf{M}_d(\mu^N)^{-1}\mathbf{v}_d(\xi)\mathbf{v}_d(\xi)^T\mathbf{M}_d(\mu^N)^{-1}}{1+\mathbf{v}_d(\xi)\mathbf{M}_d(\mu^N)^{-1}\mathbf{v}_d(\xi)}\end{aligned}$$

and therefore

☞ one obtains the simple update formula:

$$\frac{1}{N+1}\,\Lambda_d^{\mu^{N+1}}(\mathbf{x}) \;=\; \frac{1}{N}\left[\Lambda_d^{\mu^N}(\mathbf{x}) - \frac{K_d^{\mu^N}(x,\xi)^2}{N(1+\Lambda_d^{\mu^N}(\mathbf{x}))}\right], \quad \forall \mathbf{x}$$

$$\frac{1}{N+1}\,\Lambda_d^{\mu^{N+1}}(\xi) \;=\; \frac{1}{N}\Lambda_d^{\mu^N}(\xi) - \frac{1}{N^2}\,\frac{\Lambda_d^{\mu^N}(\xi)^2}{1+\Lambda_d^{\mu^N}(\xi)}$$

☞ For instance one may decide to classify as outliers all points ξ such that $\Lambda_d^{\mu^N}(\xi) < \binom{p+d}{p}^{-1}$.

☞ Such a strategy (even with relatively low degree d) is as efficient as more elaborated techniques, with only one parameter (the degree d), and with no optimization involved.

☞ Lass. & Pauwels (2016) Sorting out typicality via the inverse moment matrix SOS polynomial, NIPS 2016.
Lass. & Pauwels (2019) The empirical Christoffel function with applications in data analysis, Adv. Comp. Math. 45, pp. 1439–1468

☞ For instance one may decide to classify as outliers all points ξ such that $\Lambda_d^{\mu^N}(\xi) < \binom{p+d}{p}^{-1}$.

☞ Such a strategy (even with relatively low degree d) is as efficient as more elaborated techniques, with only one parameter (the degree d), and with no optimization involved.

☞ Lass. & Pauwels (2016) Sorting out typicality via the inverse moment matrix SOS polynomial, NIPS 2016.
Lass. & Pauwels (2019) The empirical Christoffel function with applications in data analysis, Adv. Comp. Math. 45, pp. 1439–1468

☞ For instance one may decide to classify as outliers all points ξ such that $\Lambda_d^{\mu^N}(\xi) < \binom{p+d}{p}^{-1}$.

☞ Such a strategy (even with relatively low degree d) is as efficient as more elaborated techniques, with only one parameter (the degree d), and with no optimization involved.

☞ Lass. & Pauwels (2016) Sorting out typicality via the inverse moment matrix SOS polynomial, NIPS 2016.
Lass. & Pauwels (2019) The empirical Christoffel function with applications in data analysis, Adv. Comp. Math. 45, pp. 1439–1468

Promising results in a recent collaboration with:

- L. Travé, K. Ducharlet (LAAS-CNRS) and the Carl Berger-Levrault company for detection of anomalies in data analysis of wireless sensors network used in several applications (e.g. units of air treatment, automatic bagage conveyor in airports (data in form of temporal series), ☞

☞ K. Ducharlet, L. Travé, J.B. Lasserre, M.V. Le Lann, Y. Miloudi. Leveraging the Christoffel Function for Outlier Detection in Data Streams, in preparation.

Manifold learning

☞ A measure μ on compact set Ω is completely determined by its moments and therefore it should not be a surprise that its moment matrix $\mathbf{M}_d(\mu)$ contains a lot of information.

We have already seen that its inverse $\mathbf{M}_d(\mu)^{-1}$ defines the Christoffel function.

☞ When μ is finitely supported (i.e., Ω is a finite set), then for sufficiently large d, the kernel of $\mathbf{M}_d(\mu)$[1] identifies the generators of a corresponding ideal of $\mathbb{R}[\mathbf{x}]$.

[1] Lass J.B., M. Laurent, P. Rostalski (2008) Semidefinite characterization and computation of zero-dimensional real radical ideals, Found. Comput. Math. 8, pp. 607–647.

☞ A measure μ on compact set Ω is completely determined by its moments and therefore it should not be a surprise that its moment matrix $\mathbf{M}_d(\mu)$ contains a lot of information.

We have already seen that its inverse $\mathbf{M}_d(\mu)^{-1}$ defines the Christoffel function.

☞ When μ is finitely supported (i.e., Ω is a finite set), then for sufficiently large d, the kernel of $\mathbf{M}_d(\mu)$[1] identifies the generators of a corresponding ideal of $\mathbb{R}[\mathbf{x}]$.

[1] Lass J.B., M. Laurent, P. Rostalski (2008) Semidefinite characterization and computation of zero-dimensional real radical ideals, Found. Comput. Math. 8, pp. 607–647.

☞ A measure μ on compact set Ω is completely determined by its moments and therefore it should not be a surprise that its moment matrix $\mathbf{M}_d(\mu)$ contains a lot of information.

We have already seen that its inverse $\mathbf{M}_d(\mu)^{-1}$ defines the Christoffel function.

☞ When μ is finitely supported (i.e., Ω is a finite set), then for sufficiently large d, the kernel of $\mathbf{M}_d(\mu)$[1] identifies the generators of a corresponding ideal of $\mathbb{R}[\mathbf{x}]$.

---

[1] Lass J.B., M. Laurent, P. Rostalski (2008) Semidefinite characterization and computation of zero-dimensional real radical ideals, Found. Comput. Math. 8, pp. 607–647.

More generally, if Ω is contained in a real algebraic variety V, then for sufficiently large d, the kernel of $\mathbf{M}_d(\mu)$ contains vectors of coefficients of polynomials that vanish on V.

In fact and remarkably,

$$\mathrm{rank}\,\mathbf{M}_d(\mu) \;=\; p(d)$$

for some univariate polynomial p (the Hilbert polynomial associated with the algebraic variety) which is of degree t if t is the dimension of the variety.

For instance $t = p - 1$ if the support is contained in the sphere $\mathbb{S}^{p-1}$ of $\mathbb{R}^p$.

For $\varepsilon > 0$ sufficiently small, the ε-perturbed Christoffel function

$$\mathbf{x} \mapsto \Lambda_d^{\mu}(\mathbf{x}) \;=\; \mathbf{v}_d(\mathbf{x})\,(\mathbf{M}_d(\mu) + \varepsilon\, I)^{-1}\,\mathbf{v}_d(\mathbf{x})$$

identifies correctly the support of Ω.

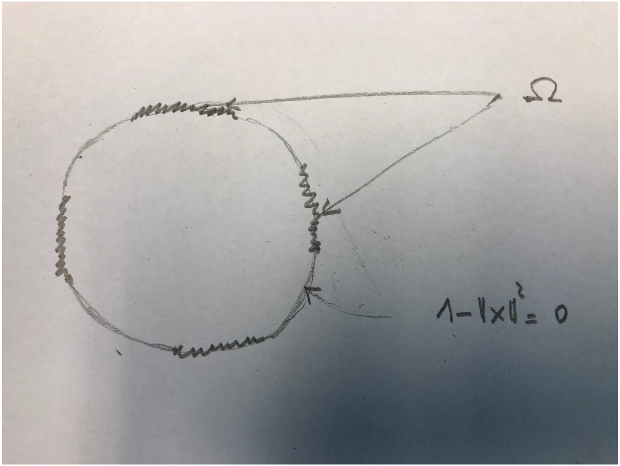
If the real algebraic variety $V \supset \Omega$ is irreducible, the empirical version of the CF (from a sample of data points on Ω)

$$\mathbf{x} \mapsto \Lambda_d^{\mu^N}(\mathbf{x}) \;=\; \mathbf{v}_d(\mathbf{x})\,(\mathbf{M}_d(\mu^N) + \varepsilon\, I)^{-1}\,\mathbf{v}_d(\mathbf{x})$$

also does it with probability 1.

☞ **Pauwels E., Putinar M., Lass. J.B. (2021).** Data analysis from empirical moments and the Christoffel function, Found. Comput. Math. 21, pp. 243–273.

For instance let $\Omega \subset \mathbb{S}^{p-1}$ (the Euclidean unit sphere of $\mathbb{R}^p$)



While the kernel identifies $\mathbb{S}^{p-1}$, the Christoffel function identifies $\Omega \subset \mathbb{S}^{p-1}$.

☞ Again this illustrates how quite sophisticated concepts of algebraic geometry are hidden and encapsulated in the moment matrix $\mathbf{M}_d(\mu)$.

☞ They can be exploited to extract various useful information on the data set.

☞ In addition, extraction of this information can be done via quite simple linear algebra techniques.

☞ Again this illustrates how quite sophisticated concepts of algebraic geometry are hidden and encapsulated in the moment matrix $\mathbf{M}_d(\mu)$.

☞ They can be exploited to extract various useful information on the data set.

☞ In addition, extraction of this information can be done via quite simple linear algebra techniques.

☞ Again this illustrates how quite sophisticated concepts of algebraic geometry are hidden and encapsulated in the moment matrix $\mathbf{M}_d(\mu)$.

☞ They can be exploited to extract various useful information on the data set.

☞ In addition, extraction of this information can be done via quite simple linear algebra techniques.

☞ However

for non modest dimension of data, matrix inversion of $\mathbf{M}_d^{-1}$ does not scale well ...

☞ On the other hand

for evaluation $\Lambda_d^\mu(\xi)$ at a point $\xi \in \mathbb{R}^p$, the variational formulation

$$\Lambda_d^\mu(\xi) = \min_{P \in \mathbb{R}[\mathbf{x}]_d} \left\{ \int_\Omega P^2 \, d\mu : \; P(\xi) = 1 \right\}, \quad \forall \xi \in \mathbb{R}^p.$$

is the simple quadratic programming problem.

$$\min_{p \in \mathbb{R}^{s(d)}} \{ p^T \mathbf{M}_d p \; : \quad \mathbf{v}_d(\xi)^T p = 1 \},$$

which can be solved quite efficiently.

☞ However

for non modest dimension of data, matrix inversion of $\mathbf{M}_d^{-1}$ does not scale well ...

☞ On the other hand

for evaluation $\Lambda_d^\mu(\xi)$ at a point $\xi \in \mathbb{R}^p$, the variational formulation

$$\Lambda_d^\mu(\xi) \,=\, \min_{P\in\mathbb{R}[\mathbf{x}]_d} \left\{ \int_\Omega P^2\, d\mu :\; P(\xi) = 1 \right\}, \quad \forall \xi \in \mathbb{R}^p.$$

is the simple quadratic programming problem.

$$\min_{p\in\mathbb{R}^{s(d)}} \{ p^T \mathbf{M}_d p \;:\; \mathbf{v}_d(\xi)^T p \,=\, 1 \},$$

which can be solved quite efficiently.

Other non-polynomial kernels, some popular in ML (e.g. Gaussian kernels), can be very efficient, to provide a large class of functions on which efficient calculation in large dimension is possible. However they are not related (at least directly) to an underlying measure supported on the data points.

☞ Again, a distinguishing feature of the CD-kernel is its deep connexion with the underlying measure.

- It not only "encodes" the cloud of data points,
- but it also captures many essential features of the more complex measure supported on those data points.

☞ Should be seen as another item in the arsenal of kernel methods in ML.

Other non-polynomial kernels, some popular in ML (e.g. Gaussian kernels), can be very efficient, to provide a large class of functions on which efficient calculation in large dimension is possible. However they are not related (at least directly) to an underlying measure supported on the data points.

☞ Again, a distinguishing feature of the CD-kernel is its deep connexion with the underlying measure.

- It not only "encodes" the cloud of data points,
- but it also captures many essential features of the more complex measure supported on those data points.

☞ Should be seen as another item in the arsenal of kernel methods in ML.

Other non-polynomial kernels, some popular in ML (e.g. Gaussian kernels), can be very efficient, to provide a large class of functions on which efficient calculation in large dimension is possible. However they are not related (at least directly) to an underlying measure supported on the data points.

☞ Again, a distinguishing feature of the CD-kernel is its deep connexion with the underlying measure.

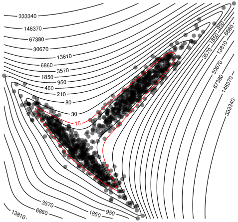
- It not only "encodes" the cloud of data points,
- but it also captures many essential features of the more complex measure supported on those data points.

☞ Should be seen as another item in the arsenal of kernel methods in ML.

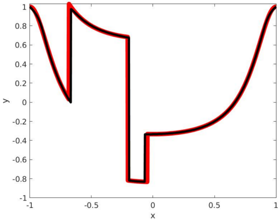
Next ...

We claim that the CF provides a simple and easy to use tool (no tuning nor optimization involved) which can help solve problems not only in data analysis, but also in approximation and interpolation of (possibly discontinuous) functions.

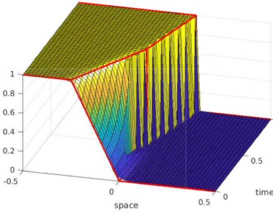
Outlier detection



Interpolation



Recovery



Some applications outside data analysis

A generic problem

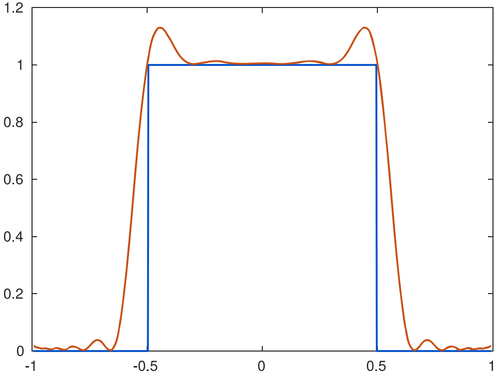
Recover an unknown function $f : \Omega \to \mathbb{R}$ from the sole knowledge of scalars

$$\mu_{\alpha,j} = \int_\Omega \mathbf{x}^\alpha \, f(\mathbf{x})^j \, d\phi(\mathbf{x}), \quad \alpha \in \mathbb{N}^n_d, \; |\alpha| + j \leq 2d$$

where $\Omega \subset \mathbb{R}^n$ is compact, d is fixed and ϕ is some given measure on Ω.

☞ f can be discontinuous and one would like to attenuate a classical Gibbs phenomenon as much as possible.

A typical Gibbs phenomenon occurs whenever one approximates a discontinuous function (in blue) by a polynomial (in red).



Applications:

- Recovery of functions (optimal solutions of optimal control problems, non-linear PDEs, etc.) from optimal solutions of SDP-relaxations of the Moment-SOS hierarchy
- density approximation
- interpolation

Ex: Application in optimal control

Consider the optimal control problem(OCP):

$$\min_{\mathbf{u}} \int_0^1 h(\mathbf{x}(t), \mathbf{u}(t))\, dt$$

$$\dot{\mathbf{x}}(t) \;=\; f(\mathbf{x}(t), \mathbf{u}(t)), \quad t \in [0,1], \quad + \text{ control/state constraints}$$

In the moment-SOS approach for optimal control one solves a hierarchy of semidefinite relaxations of increasing size.

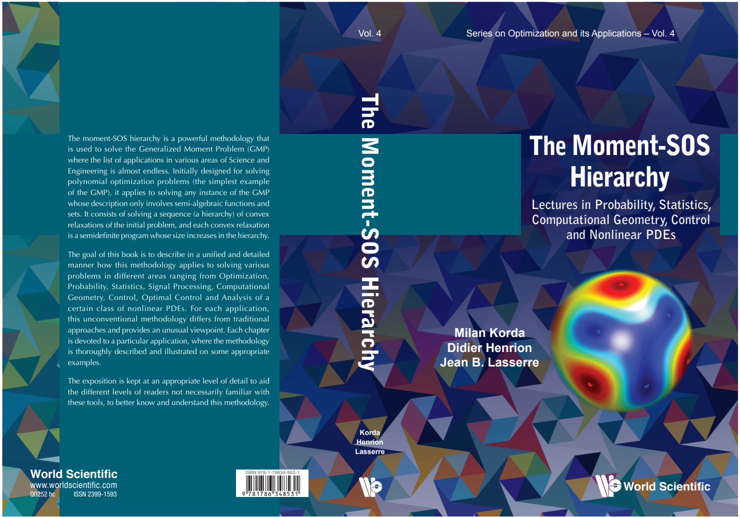
Ex: Application in optimal control

Consider the optimal control problem(OCP):

$$\min_{\mathbf{u}} \int_0^1 h(\mathbf{x}(t), \mathbf{u}(t))\, dt$$

$$\dot{\mathbf{x}}(t) \;=\; f(\mathbf{x}(t), \mathbf{u}(t)), \quad t \in [0,1], \quad + \text{ control/state constraints}$$

In the moment-SOS approach for optimal control one solves a hierarchy of semidefinite relaxations of increasing size.


The moment-SOS hierarchy is a powerful methodology that is used to solve the Generalized Moment Problem (GMP) where the list of applications in various areas of Science and Engineering is almost endless. Initially designed for solving polynomial optimization problems (the simplest example of the GMP), it applies to solving any instance of the GMP whose description only involves semi-algebraic functions and sets. It consists of solving a sequence (a hierarchy) of convex relaxations of the initial problem, and each convex relaxation is a semidefinite program whose size increases in the hierarchy.
The goal of this book is to describe in a unified and detailed manner how this methodology applies to solving various problems in different areas ranging from Optimization, Probability, Statistics, Signal Processing, Computational Geometry, Control, Optimal Control and Analysis of a certain class of nonlinear PDEs. For each application, this unconventional methodology differs from traditional approaches and provides an unusual viewpoint. Each chapter is devoted to a particular application, where the methodology is thoroughly described and illustrated on some appropriate examples.
The exposition is kept at an appropriate level of detail to aid the different levels of readers not necessarily familiar with these tools, to better know and understand this methodology.
World Scientific
www.worldscientific.com
00252 hc
ISSN 2399-1593
ISBN 978-1-78634-853-1
Vol. 4
The Moment-SOS Hierarchy
Korda
Henrion
Lasserre
Series on Optimization and its Applications – Vol. 4
The Moment-SOS Hierarchy
Lectures in Probability, Statistics, Computational Geometry, Control and Nonlinear PDEs
Milan Korda
Didier Henrion
Jean B. Lasserre
World Scientific

At an optimal solution of the step-d semidefinite relaxation one obtains an approximation

$$z_{\alpha,\beta,k} \approx \int \mathbf{x}^\alpha \, \mathbf{u}^\beta t^k \, d\mu(\mathbf{x}, \mathbf{u}, t) = \int_0^1 \mathbf{x}(t)^\alpha \, \mathbf{u}(t)^\beta t^k \, dt$$

of the moments up to order $2d$, of the measure μ supported on an optimal trajectory $\{(\mathbf{x}(t), \mathbf{u}(t)) : t \in [0, 1]\}$ of the OCP.

☞ Such a measure μ is called the occupation measure “up to time 1" associated with the trajectory $\{(\mathbf{x}(t), \mathbf{u}(t)) : t \in [0, 1]\}$ of the OCP.

Once $z_{\alpha,\beta,k}$ has been computed, it remains to recover the functions $t \mapsto \mathbf{x}_i(t), \mathbf{u}_j(t)$ from $z_{\alpha,\beta,k}$,

that is:

For each $i = 1,\ldots,n$, recover the function $f : [0,1] \to \mathbb{R}$:

$$t \mapsto f(t) \;:=\; \mathbf{x}_i(t)\,,\quad t \in [0,1]$$

from knowledge of the pseudo-moments $z_{\alpha,\beta,k}$.

☞ In fact for each i, one only needs to use $z_{\alpha,0,k}$ with $\alpha_j = 0$ for all $j \neq i$, that is, moments of the marginal of μ on (x_i, t) (which is a measure on $\mathbb{R}^2$).

☞ Same thing to recover $t \mapsto f(t) = \mathbf{u}_j(t)$ for each $j = 1,\ldots,m$.

Back to our recovery problem

Recall that we want to recover an unknown function $f : \Omega \to \mathbb{R}$ from the sole knowledge of moments

$$\begin{aligned} \mu_{\alpha,j} &= \int_{\Omega\times\mathbb{R}} \mathbf{x}^\alpha \, y^j \, d\mu(\mathbf{x}, y)\,, \quad \alpha \in \mathbb{N}^n_{2d}\,; \; |\alpha| + j \leq 2d \\ &= \int_{\Omega} \mathbf{x}^\alpha \, f(\mathbf{x})^j \, d\phi(\mathbf{x})\,, \quad \alpha \in \mathbb{N}^n_{2d}\,; \; |\alpha| + j \leq 2d \end{aligned}$$

of the measure

$$\mu = \delta_{\{f(\mathbf{x})\}}(dy)\, \phi(d\mathbf{x}) \quad \text{on } \Omega \times \mathbb{R}$$

with marginal ϕ on Ω, and supported on the graph

$$\boldsymbol{\Delta} := \{(\mathbf{x}, f(\mathbf{x})) : \mathbf{x} \in \Omega\} \quad \text{of } f.$$

Recall that the Christoffel function Λ_d^{μ} is a very appropriate tool to identify the (compact) support $\mathbf{\Delta}$ of a measure μ, from the sole knowledge of finitely many moments of μ (up to degree d).

Take home message

☞ Hence the Christoffel function Λ_d^{μ} is quite appropriate for approximating f from finitely many moments of μ.

Indeed the support $\mathbf{\Delta}$ of μ is precisely the graph $\{(\mathbf{x}, f(\mathbf{x})) : \mathbf{x} \in \Omega\}$ of f ...!

in a standard use of the CD kernel

to recover $f : [0,1] \to \mathbb{R}$, one considers the univariate measure on [0, 1]

$$d\mu(x) = f(x)\, 1_{[0,1]}(x)\, dx$$

and its moments

$$\mu_j = \int_0^1 x^j f(x)\, dx\,, \quad j = 0, 1, \ldots$$

☞ Notice that the support $\Omega = [0,1]$ of μ provides no information on the density f of μ.

whereas in a non-standard use of the CD kernel

to recover $f : [0,1] \to \mathbb{R}$, we consider the bivariate measure

$$d\mu(x,y) \,=\, \delta_{\{f(x)\}}(dy)\, 1_{[0,1]}(x)\, dx$$

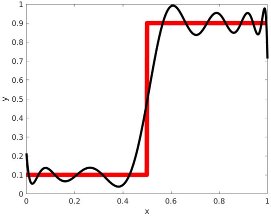
with support $\boldsymbol{\Delta} \subset [0,1] \times [0,M]$, and its moments

$$\mu_{i,j} \,=\, \int x^i y^j \, d\mu(x,y) \,=\, \int_0^1 x^i \, f(x)^j \, dx\,, \quad i,j = 0,1,\ldots$$

In the standard use of CD-kernel, one approximates $f : [0, 1] \to \mathbb{R}$ by its projection on $\mathbb{R}[\mathbf{x}]_n \subset L^2([0, 1])$:

$$x \mapsto \hat{f}_n(x) \;:=\; \sum_{j=0}^{n} \left(\int_0^1 f(y)\, L_j(y) dy \right) L_j(x)\,,$$

with an orthonormal basis $(L_j)_{j \in \mathbb{N}}$ of $L^2([0, 1])$.



Ex: Chebyshev interpolant

☞ Typical Gibbs phenomenon occurs.

The function $\hat{f}_n$ is a polynomial and also reads

$$\hat{f}_n(x) = \int_0^1 \mathbf{K}_n^\lambda(x,y)\, f(y)\, dy\,,$$

where λ is Lebesgue measure on $[0,1]$, and

$$\mathbf{K}_n^\lambda(x,y) = \sum_{i=0}^n L_i(x) L_i(y)$$

is the Christoffel-Darboux (CD) kernel associated with λ.

☞ So this standard "use" of the CD kernel does not avoid the Gibbs phenomenon.

Alternative Positive Kernels with better convergence properties have been proposed, still in the same framework:

Féjer, Jackson kernels, etc.

- Reproducing property of the CD kernel is LOST
- Preserve positivity (e.g when approximating a density)
- Better convergence properties than the CD kernel, in particular uniform convergence (for continuous functions) on arbitrary compact subsets

☞ F. Kirchner, E. de Klerk. Construction of multivariate polynomial approximation kernels via semidefinite programming arXiv:2203.05892

In a non standard use of CD-kernel

☞ we rather consider the graph $\mathbf{\Delta} \subset \mathbb{R}^2$ of f, i.e., the set

$$\mathbf{\Delta} := \{ (x, f(x)) : \quad x \in [0, 1] = \Omega, \} .$$

and the measure

$$d\mu(x, y) := \delta_{f(x)}(dy) \, 1_{[0,1]}(x) \, dx$$

whose (degenerate) support is $\mathbf{\Delta} \subset \mathbb{R}^2$.

Again, why should we do that as it implies going to $\mathbb{R}^2$ instead of staying in $\mathbb{R}$?

☞ ... because

- The support of μ is exactly the graph $\mathbf{\Delta}$ of f, and
- The CF $(x, y) \mapsto \Lambda_n^\mu(x, y)$ identifies the support of μ!
- Hence the CF $(x, y) \mapsto \Lambda_n^\mu(x, y)$ identifies f!

Interpolation

So suppose that you are given point evaluations $\{f(x_i)\}_{i\leq N}$ of an unknown function f on $[0,1]$, and again let

$$\mathbf{v}_d(x,y) := (1, x, y, x^2, x\,y, y^2, \ldots, x\,y^{d-1}, y^d)\,.$$

☞ Compute the degree-d empirical moment matrix:

$$\mathbf{M}_d(\mu) \;:=\; \sum_{i=1}^{N} \mathbf{v}_d((x_i, f(x_i))\,\mathbf{v}_d(x_i, f(x_i))^T\,,$$

of the empirical measure $d\mu(x,y) := \frac{1}{N}\sum_{i=1}^{N} \delta_{x(i),f(x(i))}$ on $\mathbb{R}^2$, by one pass over the data

Assume that is given the moment matrix $\mathbf{M}_d(\mu)$ of the measure

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

on the graph $\boldsymbol{\Delta} = \{(\mathbf{x}, f(\mathbf{x})) :\, \mathbf{x} \in \Omega\}$ of the function $f : \Omega \to \mathbb{R}$, and $\sup_{\mathbf{x}\in\Omega} |f(\mathbf{x})| < M$.

Let $\varepsilon > 0$ and λ be Lebesgue measure on $\Omega \times [-M, M]$

- ☞ Compute the Christoffel function

$$x \mapsto \Lambda_d^{\mu,\varepsilon}(\mathbf{x},y)^{-1} \;:=\; \mathbf{v}_d(\mathbf{x},y)^T \mathbf{M}_d(\mu+\varepsilon\,\lambda)^{-1}\,\mathbf{v}_d(\mathbf{x},y)\,.$$

Alternatively

$$x \mapsto \Lambda_d^{\mu,\varepsilon}(\mathbf{x},y)^{-1} \;:=\; \mathbf{v}_d(\mathbf{x},y)^T (\mathbf{M}_d(\mu)+\varepsilon\mathbf{I})^{-1}\,\mathbf{v}_d(\mathbf{x},y)\,.$$

- Approximate f by

$$\mathbf{x} \mapsto \hat{f}_{d,\varepsilon}(\mathbf{x}) \;:=\; \arg\min_{y} \Lambda_d^{\mu,\varepsilon}(\mathbf{x},y)^{-1}\,.$$

☞ minimize a univariate polynomial! (easy)

- ☞ Compute the Christoffel function

$$x \mapsto \Lambda_d^{\mu,\varepsilon}(\mathbf{x},y)^{-1} \;:=\; \mathbf{v}_d(\mathbf{x},y)^T \mathbf{M}_d(\mu+\varepsilon\,\lambda)^{-1}\,\mathbf{v}_d(\mathbf{x},y)\,.$$

Alternatively

$$x \mapsto \Lambda_d^{\mu,\varepsilon}(\mathbf{x},y)^{-1} \;:=\; \mathbf{v}_d(\mathbf{x},y)^T (\mathbf{M}_d(\mu)+\varepsilon\mathbf{I})^{-1}\,\mathbf{v}_d(\mathbf{x},y)\,.$$

- Approximate f by

$$\mathbf{x} \mapsto \hat{f}_{d,\varepsilon}(\mathbf{x}) \;:=\; \arg\min_{y} \Lambda_d^{\mu,\varepsilon}(\mathbf{x},y)^{-1}\,.$$

☞ minimize a univariate polynomial! (easy)

- ☞ Compute the Christoffel function

$$x \mapsto \Lambda_d^{\mu,\varepsilon}(\mathbf{x},y)^{-1} \;:=\; \mathbf{v}_d(\mathbf{x},y)^T \mathbf{M}_d(\mu+\varepsilon\,\lambda)^{-1}\,\mathbf{v}_d(\mathbf{x},y)\,.$$

Alternatively

$$x \mapsto \Lambda_d^{\mu,\varepsilon}(\mathbf{x},y)^{-1} \;:=\; \mathbf{v}_d(\mathbf{x},y)^T (\mathbf{M}_d(\mu)+\varepsilon\mathbf{I})^{-1}\,\mathbf{v}_d(\mathbf{x},y)\,.$$

- Approximate f by

$$\mathbf{x} \mapsto \hat{f}_{d,\varepsilon}(\mathbf{x}) \;:=\; \arg\min_{y} \Lambda_d^{\mu,\varepsilon}(\mathbf{x},y)^{-1}\,.$$

☞ minimize a univariate polynomial! (easy)

Choosing

$$\varepsilon := 2^{3-\sqrt{d}}$$

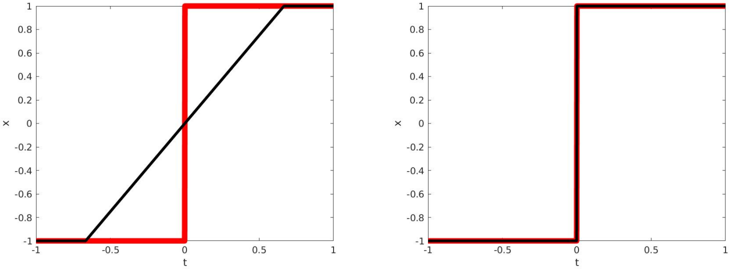
ensures convergence properties for bounded measurable functions, e.g. pointwise on open sets with no point of discontinuity.

Convergence properties as $d \uparrow$

- ☞ L^1-convergence
- ☞ pointwise convergence on open sets with no point of discontinuity, and so almost uniform convergence.
- ☞ L^1-convergence at a rate $O(d^{-1/2})$ for Lipschitz continuous f.

☞ Again note the central role played by the Moment Matrix!

S. Marx, E. Pauwels, T. Weisser, D. Henrion, J.B. Lass.
Semi-algebraic approximation using Christoffel-Darboux kernel,
Constructive Approximation, 2021



In black (left) the approximation with moments of order 2 and in black (right) the approximation with moments of order 4 (and ε is quite small)

☞ Observe the absence of any Gibbs phenomenon ...

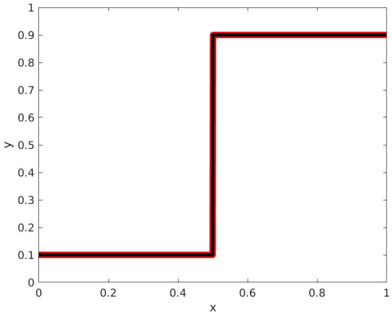
It is important to observe that the function

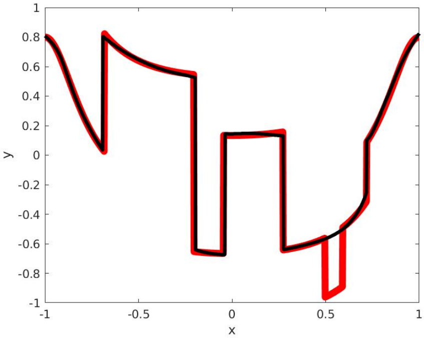
$$t \mapsto \hat{f}_{d,\varepsilon}(\mathbf{x}) := \arg\min_{y} \Lambda_d^{\mu,\varepsilon}(\mathbf{x}, y)^{-1}$$

is *not* a polynomial, but rather a semi-algebraic function, and therefore may capture discontinuities ...!

☞ Actually, the above step function is the arg min of an SOS of degree 6.

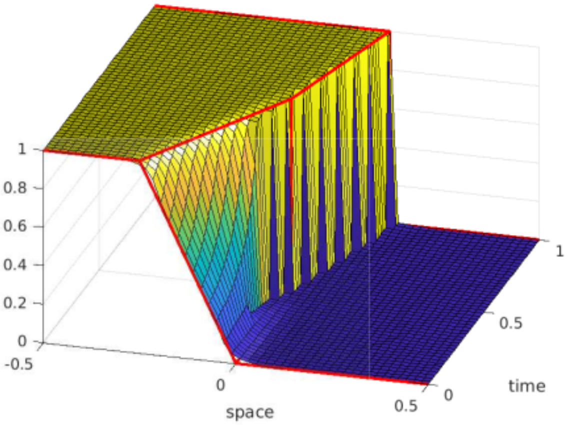
Ex: Interpolation





Ex: Recovery

Below : Recovery of a (discontinuous) solution of the Burgers Equation from knowledge of approximate moments of the occupation measure supported on the solution.



What if not all moments are available?

As is the case in density approximation, suppose that the function $f : [0,1] \to \mathbb{R}$ to approximate is only known via its Fourier-Legendre coefficients

$$\mu_{i,1} = \int_0^1 x^i f(x)\, dx\,, \quad i = 0,1,\ldots$$

and we do not have access to other moments

$$\mu_{i,j} = \int_0^1 x^i f(x)^j\, dx\,, \quad j > 1;\ i = 0,1,\ldots$$

of the measure $\mu(d(x,y)) = \delta_{f(x)}(dy)\, \underbrace{1_{[0,1]}(x)\, dx}_{\phi(dx)}$

By moment-matrix completion

Recall that $\boldsymbol{\phi} = (\phi_i)_{i\in\mathbb{N}}$ is the moment-sequence of Lebesgue measure ϕ on $[0,1]$, and consider the semidefinite programs indexed by $n \in \mathbb{N}$:

$$\mathbf{P}_n : \quad \inf_{\boldsymbol{\psi}} \quad \{\Theta_n(\boldsymbol{\psi}) : \mathbf{M}_n(\boldsymbol{\psi}) \succeq 0 \\ \psi_{i,0} = \phi_i \, (= \mu_{i,0}), \quad i \in \mathbb{N} \\ \psi_{i,1} = \mu_{i,1} \quad i \in \mathbb{N} \}\,,$$

where the inf is over all pseudo-moments $\boldsymbol{\psi} = (\psi_{i,j})_{i,j\in\mathbb{N}^2_{2n}}$, and Θ_n is a certain sparsity inducing linear functional.

(☞ An n-truncation of the diagonal of the infinite moment matrix $\mathbf{M}(\boldsymbol{\psi})$)

For every measure ν on $[0,1]\times\mathbb{R}$, let $\boldsymbol{\nu}=(\nu_{i,j})_{i,j\in\mathbb{N}}$ be the sequence of its moments.

Theorem

(i) For every $n\in\mathbb{N}$,

$$\Theta_n(\boldsymbol{\mu})\,\leq\,\Theta_n(\boldsymbol{\nu})\,,$$

for all measures ν on $[0,1]\times\mathbb{R}$ whose moment-sequence $\boldsymbol{\nu}$ is a feasible solution of $\mathbf{P}_n$.
(ii) Let ψ^n be an optimal solution of $\mathbf{P}_n$. Then

$$\lim_{n\to\infty}\psi^n_{i,j}\,=\,\mu_{i,j}\,=\,\int_{[0,1]}x^i\,f(x)^j\,dx\,,\quad\forall i,j=0,1,\ldots$$

☞ Hence one may approximate f accurately from finitely moments $\mu_{i,j}$ as described earlier.

☞ D. Henrion & J.B. Lass. Graph recovery from incomplete moment information (2021), Constructive Approximation.

Christoffel function and Positive polynomials

Let $\Omega \subset \mathbb{R}^n$ be the basic semi-algebraic set (with nonempty interior)

$$\Omega := \{\mathbf{x} \in \mathbb{R}^n : \; g_j(\mathbf{x}) \geq 0, \quad j = 1, \ldots, m\}$$

with $g_j \in \mathbb{R}[\mathbf{x}]_{d_j}$ and let $s_j = \lceil \deg(g_j)/2 \rceil$. Let $g_0 = 1$ with $s_0 = 0$.

With t fixed, its associated quadratic module

$$Q_t(\Omega) := \{ \sum_{j=0}^{m} \sigma_j \, g_j : \quad \sigma_j \in \Sigma[\mathbf{x}]_{t-s_j} \} \subset \mathbb{R}[\mathbf{x}]$$

is a convex cone with nonempty interior,

and with dual cone

$$Q_t(\Omega)^* := \{ y \in \mathbb{R}^{s(t)} : \mathbf{M}_{t-s_j}(g_j\, y) \succeq 0, \quad j = 0, \ldots, m \},$$

where $s(t) = \binom{n+t}{n}$.

Notice that if $\mathbf{M}_t(y)^{-1} \succ 0$ for all t

one may define a family of polynomials $(P_\alpha)_{\alpha\in\mathbb{N}^n} \subset \mathbb{R}[\mathbf{x}]$ orthonormal w.r.t. y, meaning that

$$L_y(P_\alpha \cdot P_\beta) = \delta_{\alpha=\beta}, \quad \alpha, \beta \in \mathbb{N}^n,$$

and exactly as for measures, the Christoffel function Λ_t^y

$$\mathbf{x} \mapsto \Lambda_t^y(\mathbf{x})^{-1} := \sum_{|\alpha|\le t} P_\alpha(\mathbf{x})^2 .$$

Theorem

For every $p \in \operatorname{int}(Q_t(\Omega))$ *there exists* $y \in \operatorname{int}(Q_t(\Omega)^*)$ *such that*

$$p(\mathbf{x}) \;=\; \sum_{j=0}^{m} \left(\mathbf{v}_{t-s_j}(\mathbf{x})^T \mathbf{M}_t(g_j\, y)^{-1}\, \mathbf{v}_{t-s_j}(\mathbf{x})\right)\, g_j(\mathbf{x})$$

$$\;=\; \sum_{j=0}^{m} \Lambda_{t-s_j}^{g_j \cdot y}(\mathbf{x})^{-1}\, g_j(\mathbf{x})$$

where $(g \cdot y)$ *is the sequence of pseudo-moments*

$$(g \cdot y)_\alpha \;:=\; \sum_{\gamma} g_\gamma\, y_{\alpha+\gamma}\,, \quad \alpha \in \mathbb{N}^n \quad (\textit{if } g(\mathbf{x}) = \textstyle\sum_\gamma g_\gamma\, \mathbf{x}^\gamma).$$

In addition $L_y(p) \;=\; \sum_{j=0}^{m} \binom{n+t-s_j}{n}$.

The proof combines

- ☞ a result by Nesterov on a one-to-one correspondence between $\mathrm{int}(Q_t(\Omega))$ and $\mathrm{int}(Q_t(\Omega)^*)$, and

- ☞ the fact that

$$\mathbf{v}_{t-s_j}(\mathbf{x})^T \mathbf{M}_t(g_j\, y)^{-1}\, \mathbf{v}_{t-s_j}(\mathbf{x}) \;=\; \Lambda^{g_j \cdot y}_{t-s_j}(\mathbf{x})^{-1}\,.$$

In other words:

In Putinar certificate

$$p = \sum_{j=0}^{m} \sigma_j \, g_j \,, \quad \sigma_j \in \mathbb{R}[\mathbf{x}]_{t-s_j} \,,$$

of strict positivity on Ω,

☞ one may always choose the SOS weights σ_j in the form

$$\sigma_j(\mathbf{x}) \;:=\; \Lambda^{g_j \cdot y}_{t-s_j}(\mathbf{x})^{-1} \,, \quad j = 0, \ldots, m \,,$$

for some sequence of pseudo-moments $y \in \mathrm{int}(Q_t(\Omega)^*)$.

☞ Question: Given p, what is the related linear functional y?

In other words:

In Putinar certificate

$$p = \sum_{j=0}^{m} \sigma_j \, g_j \,, \quad \sigma_j \in \mathbb{R}[\mathbf{x}]_{t-s_j} \,,$$

of strict positivity on Ω,

☞ one may always choose the SOS weights σ_j in the form

$$\sigma_j(\mathbf{x}) \;:=\; \Lambda_{t-s_j}^{g_j \cdot y}(\mathbf{x})^{-1} \,, \quad j = 0, \ldots, m \,,$$

for some sequence of pseudo-moments $y \in \mathrm{int}(Q_t(\Omega)^*)$.

☞ Question: Given p, what is the related linear functional y?

Disintegration

Recall that if μ is a measure on a Borel set $\Omega := X \times Y$, then it disintegrates as

$$d\mu(x,y) = \underbrace{\hat{\mu}(dy \,|\, x)}_{conditional} \; \underbrace{\phi(dx)}_{marginal}$$

with marginal ϕ on X and conditional $\hat{\mu}(dy|x)$ on Y given $x \in X$.

Let $d\mu(x,y)$ be a measure on $\Omega \subset \mathbb{R}^n \times \mathbb{R}$, with marginal ϕ on $\mathbb{R}^n$, and such that $\mathbf{M}_d(\mu) \succ 0$.

Theorem (Lass (2022))

The Christoffel function $(x,y) \mapsto \Lambda_d^{\mu}(x,y)$ *disintegrates into*

$$\Lambda_d^{\mu}(x,y) \,=\, {\Lambda_d}^{\phi}(x) \cdot \Lambda_d^{\nu_{x,d}}(y)\,, \quad \forall (x,y)\,,$$

for some measure $\nu_{x,d}$ *on* $\mathbb{R}$.

☞ valid even for linear functionals $\mu \in \mathbb{R}[x,y]^*$ with no representing measure. More details in:

J.B. Lass (2022) On the Christoffel function and classification in data analysis, *Comptes Rendus Mathematiques*

J.B. Lass (2022) A disintegration of the Christoffel function, *Comptes Rendus Mathematiques*

THANK YOU!