Near G-optimal Tchakaloff designs *

Len Bos[1], Federico Piazzon[2] and Marco Vianello[2]

October 18, 2019

Abstract

We show that the notion of polynomial mesh (norming set), used to provide discretizations of a compact set nearly optimal for certain approximation theoretic purposes, can also be used to obtain finitely supported near G-optimal designs for polynomial regression. We approximate such designs by a standard multiplicative algorithm, followed by measure concentration via Caratheodory-Tchakaloff compression.

2010 AMS subject classification: 62K05, 65C60, 65D32.

Keywords: near G-optimal designs, polynomial regression, norming sets, polynomial meshes, Dubiner distance, D-optimal designs, multiplicative algorithms, Caratheodory-Tchakaloff measure compression.

1 Introduction

In this paper we explore a connection of the approximation theoretic notion of polynomial mesh (norming set) of a compact set K with the statistical theory of optimal polynomial regression designs on K. We begin by recalling some basic definitions and properties.

Let $\mathbb{P}_n^d(K)$ denote the space of polynomials of degree not exceeding n restricted to a compact set $K \subset \mathbb{R}^d$, and $\|f\|_Y$ the sup-norm of a bounded function on the compact set Y. We recall that a *polynomial mesh* on K (with constant $c > 0$) is a sequence of norming subsets $X_n \subset K$ such that

$$\|p\|_K \le c\,\|p\|_{X_n}\ ,\quad \forall p \in \mathbb{P}_n^d(K)\ , \tag{1}$$

where $\mathrm{card}(X_n)$ grows algebraically in

$$N = N_n(K) = \dim(\mathbb{P}_n^d(K))\ . \tag{2}$$

Notice that necessarily $\mathrm{card}(X_n) \ge N$, since X_n is determining for $\mathbb{P}_n^d(K)$ (i.e., polynomials vanishing there vanish everywhere on K). With a little abuse of

*Work partially supported by the DOR funds and the Project BIRD 181249 of the University of Padova, and by the GNCS-INdAM. This research has been accomplished within the RITA "Research ITalian network on Approximation".

[1]Department of Computer Science, University of Verona, Italy
[2]Department of Mathematics, University of Padova, Italy
corresponding author: marcov@math.unipd.it

notation, below we shall call "polynomial mesh" both the entire sequence $\{X_n\}$ and (more frequently) the single norming set X_n.

Observe also that $N = \mathcal{O}(n^\beta)$ with $\beta \leq d$, in particular $N = \binom{n+d}{d} \sim n^d/d!$ on polynomial determining compact sets (i.e., polynomials vanishing there vanish everywhere in $\mathbb{R}^d$), but we can have $\beta < d$ for example on compact algebraic varieties, like the sphere in $\mathbb{R}^d$ where $N = \binom{n+d}{d} - \binom{n-2+d}{d}$.

The notion of polynomial mesh, though present in the literature for specific instances, was introduced in a systematic way in the seminal paper [8], and since then has seen an increasing interest, also in the computational framework. We recall among their properties that polynomial meshes are invariant under affine transformations, can be extended by algebraic transformations, finite union and product, are stable under small perturbations. Concerning *finite union*, for example, which is a powerful constructive tool, it is easily checked that if $X_n^{(i)}$ are polynomial meshes for K_i, $1 \leq i \leq s$, then

$$\|p\|_{\cup K_i} \leq \max\{c_i\} \, \|p\|_{\cup X_n^{(i)}} \, , \quad \forall p \in \mathbb{P}_n^d(\cup K_i) \, . \tag{3}$$

Polynomial meshes give good discrete models of a compact set for polynomial approximation, for example it is easily seen that the uniform norm of the unweighted Least Squares operator on a polynomial mesh, say $L_n : C(K) \to \mathbb{P}_n^d(K)$, is bounded as

$$\|L_n\| = \sup_{f \neq 0} \frac{\|L_n f\|_K}{\|f\|_K} \leq c \sqrt{\mathrm{card}(X_n)} \, . \tag{4}$$

Moreover, polynomial meshes contain extremal subsets of Fekete and Leja type for polynomial interpolation, and have been applied in polynomial optimization and in pluripotential numerics; cf., e.g., [5, 19, 23].

The class of compact sets which admit (constructively) a polynomial mesh is very wide. For example, it has been proved in [8] that a polynomial mesh with cardinality $\mathcal{O}(N^r) = \mathcal{O}(n^{rd})$ can always be constructed simply by intersection with a sufficiently dense uniform covering grid, on compact sets satisfying a Markov polynomial inequality with exponent r (in particular, on compact bodies with Lipschitz boundary, in which case $r = 2$).

From the computational point of view, it is however important to deal with low cardinality polynomial meshes. Indeed, polynomial meshes with $\mathrm{card}(X_n) = \mathcal{O}(N)$, that are said to be *optimal* (in the sense of cardinality growth), have been constructed on compact sets with different geometric structure, such as polygons and polyhedra, convex, starlike and even more general bodies with smooth boundary, sections of a sphere, ball and torus; cf., e.g., [15, 18, 27]. By (4), on optimal meshes we have $\|L_n\| = \mathcal{O}(\sqrt{\mathrm{card}(X_n)}) = \mathcal{O}(\sqrt{N})$; we stress, however, that even with an optimal mesh typically $\mathrm{card}(X_n) \gg N$.

The problem of reducing the sampling cardinality while maintaining estimate (4) (Least Squares compression) has been considered for example in [20, 21], where a strategy is proposed, based on weighted Least Squares on N_{2n} Caratheodory-Tchakaloff points extracted from the mesh by Linear or Quadratic Programming. Nevertheless, also reducing the Least Squares uniform operator norm though much more costly is quite relevant, and this will be addressed in the next Section via the theory of optimal designs.

2 Near optimal designs by polynomial meshes

Let μ be a probability measure supported on a compact set $K \subset \mathbb{R}^d$. In statistics, μ is usually called a design and K the design space. The literature on optimal designs dates back at least one century, and is so vast and ramificated that we can not even attempt any kind of survey. We may for example quote, among many others, a classical and a quite recent textbook [1, 10], and the new paper [11] that bear witness to the vitality of the field. Below we simply recall some relevant notions and results, trying to follow an apparently unexplored connection to the theory of polynomial meshes in the framework of polynomial regression.

Assume that $supp(\mu)$ is determining for $\mathbb{P}^d(K)$ (the space of d-variate real polynomials restricted to K); for a fixed degree n, we could even assume that $supp(\mu)$ is determining for $\mathbb{P}^d_n(K)$. We recall a function that plays a key role in the theory of optimal designs, the diagonal of the reproducing kernel for μ in $\mathbb{P}^d_n(K)$ (often called *Christoffel polynomial*), namely

$$K^\mu_n(x,x) = \sum_{j=1}^N p_j^2(x) \ , \tag{5}$$

where $\{p_j\}$ is any μ-orthonormal basis of $\mathbb{P}^d_n(K)$, for example that obtained from the standard monomial basis by applying the Gram-Schmidt orthonormalization process (it can be shown that $K^\mu_n(x,x)$ does not depend on the choice of the orthonormal basis, cf. (7) below). It has the important property that

$$\|p\|_K \le \sqrt{\max_{x\in K} K^\mu_n(x,x)}\, \|p\|_{L^2_\mu(K)} \ , \quad \forall p \in \mathbb{P}^d_n(K) \ , \tag{6}$$

and also the following relevant characterization

$$K^\mu_n(x,x) = \max_{p\in\mathbb{P}^d_n(K),\, p(x)=1} \frac{1}{\int_K p^2(x)\, d\mu} \ . \tag{7}$$

Now, by (5) $\int_K K^\mu_n(x,x)\, d\mu = N$, which entails that $\max_{x\in K} K^\mu_n(x,x) \ge N$. A probability measure $\mu^* = \mu^*(K)$ is then called a G-optimal design for polynomial regression of degree n on K if

$$\min_\mu \max_{x\in K} K^\mu_n(x,x) = \max_{x\in K} K^{\mu^*}_n(x,x) = N \ . \tag{8}$$

Observe that, since $\int_K K^\mu_n(x,x)\, d\mu = N$ for every μ, an optimal design has the following property

$$K^{\mu^*}_n(x,x) = N \ \ \mu^* - a.e. \text{ in } K \ . \tag{9}$$

As is well-known, by the celebrated Kiefer-Wolfowitz General Equivalence Theorem [14] the difficult min-max problem (8) is equivalent to the much simpler maximization

$$\max_\mu det(G^\mu_n) \ , \quad G^\mu_n = \left(\int_K q_i(x) q_j(x)\, d\mu \right)_{1\le i,j\le N} \ , \tag{10}$$

where G^μ_n is the Gram matrix of μ in a fixed polynomial basis $\{q_i\}$ (also called information matrix in statistics). Such an optimality is called D-optimality,

and entails that an optimal measure exists, since the set of Gram matrices of probability measures is compact (and convex); see e.g. [1, 2, 4] for a quite general proof of these results, valid for both continuous and discrete compact sets. An optimal measure is not unique and not necessarily discrete (unless K is discrete itself), but an equivalent discrete optimal measure always exists by the Tchakaloff Theorem on positive quadratures of degree $2n$ for K; cf. [24] for a general proof of the Tchakaloff Theorem. Moreover, the asymptotics of optimal designs as the degree n goes to ∞ can be described using multivariate pluripotential theory, see [2, 3].

G-optimality has two important interpretations in terms of probabilistic and deterministic polynomial regression. From a statistical point of view, it is the probability measure that minimizes the maximum prediction variance by n-th degree polynomial regression, cf. [1].

From the approximation theory point of view, calling $L_n^{\mu^*}$ the corresponding weighted Least Squares operator, by (6) we can write for every $f \in C(K)$

$$\begin{aligned} \|L_n^{\mu^*} f\|_K &\le \sqrt{\max_{x\in K} K_n^{\mu^*}(x,x)}\, \|L_n^{\mu^*} f\|_{L^2_{\mu^*}(K)} \le \sqrt{N}\, \|L_n^{\mu^*} f\|_{L^2_{\mu^*}(K)} \\ &\le \sqrt{N}\, \|f\|_{L^2_{\mu^*}(K)} \le \sqrt{N}\, \|f\|_K \ , \ \text{ i.e. } \ \|L_n^{\mu^*}\| \le \sqrt{N} \ , \end{aligned} \tag{11}$$

which shows that a G-optimal measure minimizes (the estimate of) the weighted Least Squares uniform operator norm.

The computational literature on D-optimal designs is huge, with a variety of approaches and methods. A classical approach is given by the discretization of K and then the D-optimization over the discrete set; see e.g. the references in [11] (where however a different approach is proposed, based on a moment-sum-of-squares hierarchy of semidefinite programming problems). In the discretization framework, the possible role of polynomial meshes seems apparently overlooked. We summarize the corresponding simple but meaningful near G-optimality result by the following Proposition.

Proposition 1 *Let $K \subset \mathbb{R}^d$ be a compact set, admitting a polynomial mesh $\{X_n\}$ with constant c.*

Then for every $n \in \mathbb{N}$ and $m \in \mathbb{N}$, $m \ge 1$, the probability measure

$$\nu = \nu(n,m) = \mu^*(X_{2mn}) \tag{12}$$

is a near G-optimal design on K, in the sense that

$$\max_{x\in K} K_n^{\nu}(x,x) \le c_m\, N \ , \ \ c_m = c^{1/m} \ . \tag{13}$$

Proof. First, observe that for every $p \in \mathbb{P}^d_{2n}(K)$

$$\|p^m\|_K = \|p\|_K^m \le c\, \|p^m\|_{X_{2mn}} = c\, \|p\|^m_{X_{2mn}} \ ,$$

and thus

$$\|p\|_K \le c^{1/m}\, \|p\|_{X_{2mn}} \ .$$

Now, X_{2mn} is clearly $\mathbb{P}^d_n(K)$-determining and hence denoting by $\nu = \mu^*(X_{2mn})$ an optimal measure for degree n on X_{2mn}, which exists by the General Equivalence Theorem with $supp(\nu) \subseteq X_{2mn}$, we get

$$\max_{x\in X_{2mn}} K_n^{\nu}(x,x) = N_n(X_{2mn}) = N_n(K) = N \ .$$

Since $K_n^\nu(x,x)$ is a polynomial of degree $2n$, we finally obtain

$$\max_{x\in K} K_n^\nu(x,x) \le c^{1/m} \max_{x\in X_{2mn}} K_n^\nu(x,x) \le c^{1/m}\, N\ . \quad \square$$

Proposition 1 shows that polynomial meshes are good discretizations of a compact set for the purpose of computing a near G-optimal measure, and that G-optimality maximum condition (8) is approached at a rate proportional to $1/m$, since $c_m \sim 1+\log(c)/m$. In terms of the statistical notion of G-efficiency on K we have

$$\mathrm{G_{eff}}(\nu) = \frac{N}{\max_{x\in K} K_n^\nu(x,x)} \ge c^{-1/m} \sim 1-\log(c)/m\ . \tag{14}$$

It is worth showing that a better rate proportional to $1/m^2$ can be obtained on certain compact sets, where an (optimal) polynomial mesh can be constructed via the approximation theoretic notion of Dubiner distance.

We recall that the *Dubiner distance* on a compact set K, introduced in the seminal paper [13]), is defined as

$$dub_K(x,y) = \sup_{deg(p)\ge 1,\, \|p\|_K\le 1} \left\{ \frac{1}{deg(p)}\ |\arccos(p(x)) - \arccos(p(y))| \right\}\ . \tag{15}$$

Among its basic properties, we recall that it is invariant under invertible affine transformations, i.e., if $\sigma(x) = Ax+b$, $det(A)\neq 0$, then

$$dub_K(x,y) = dub_{\sigma(K)}(\sigma(x),\sigma(y))\ . \tag{16}$$

The notion of Dubiner distance plays a deep role in multivariate polynomial approximation, cf. e.g. [6, 13]. Unfortunately, such a distance is explicitly known only in the univariate case on intervals (where it is the *arccos* distance by the Van der Corput-Schaake inequality), and on the cube, simplex, sphere and ball (in any dimension), cf. [6, 13]. On the other hand, it can be estimated on some classes of compact sets, for example on smooth convex bodies via a tangential Markov inequality on the boundary, cf. [23]. Its connection with the theory of polynomial meshes is given by the following elementary but powerful Lemma [23]; for the reader's convenience, we recall also the simple proof.

Lemma 1 *Let $Y_n = Y_n(\alpha)$, $n\ge 1$, be a sequence of finite sets of a compact set $K\subset\mathbb{R}^d$, whose covering radius with respect to the Dubiner distance does not exceed α/n, where $\alpha\in(0,\pi/2)$, i.e.*

$$r(Y_n) = \max_{x\in K} dub_K(x,Y_n) = \max_{x\in K}\min_{y\in Y_n} dub_K(x,y) \le \frac{\alpha}{n}\ . \tag{17}$$

Then, $\{Y_n\}$ is a polynomial mesh on K with constant $c = 1/\cos(\alpha)$.

Proof. First, possibly normalizing and/or multiplying p by -1, we can assume that $\|p\|_K = p(\hat{x}) = 1$ for a suitable $\hat{x}\in K$. Since (17) holds for Y_n, there exists $\hat{y}\in Y_n$ such that

$$|\arccos(p(\hat{x})) - \arccos(p(\hat{y}))| = |\arccos(p(\hat{y}))| \le \frac{\alpha\, deg(p)}{n} \le \alpha < \frac{\pi}{2}\ .$$

Now the arccos function is monotonically decreasing and nonnegative, thus we have that $p(\hat{y}) \geq \cos(\alpha) > 0$, and finally

$$\|p\|_K = 1 \leq \frac{p(\hat{y})}{\cos\alpha} \leq \frac{1}{\cos\alpha} \|p\|_{Y_n} . \quad \square$$

By Lemma 1 we can now prove the following proposition on near G-optimality by polynomial meshes constructed via the Dubiner distance.

Proposition 2 *Let $K \subset \mathbb{R}^d$ be a compact set and $\{Y_n(\alpha)\}$ be the polynomial mesh of Lemma 1.*

Then for every $n \in \mathbb{N}$ and $m > 1$, the probability measure

$$\nu = \nu(n,m) = \mu^*(Y_{2n}(\pi/(2m))) \tag{18}$$

is a near G-optimal design on K, in the sense that

$$\max_{x\in K} K_n^\nu(x,x) \leq c_m N , \;\; c_m = \frac{1}{\cos(\pi/(2m))} . \tag{19}$$

The proof follows essentially the lines of that of Proposition 1, with $Y_{2n}(\pi/(2m))$ replacing X_{2mn}, observing that by Lemma 1 for every $p \in \mathbb{P}^d_{2n}(K)$ we have $\|p\|_K \leq c_m \|p\|_{Y_{2n}(\pi/(2m))}$. We stress that in this case $c_m \sim 1 + \pi^2/(8m^2)$, i.e. G-optimality is approached at a rate proportional to $1/m^2$. In terms of G-efficiency we have in this case

$$\mathrm{G}_{\mathrm{eff}}(\nu) \geq \cos(\pi/(2m)) \sim 1 - \pi^2/(8m^2) . \tag{20}$$

We recall that optimal polynomial meshes like those in Proposition 2 have been recently constructed in the framework of polynomial optimization on some compact sets where the Dubiner distance is known or can be estimated, such as the cube, the sphere, convex bodies with smooth boundary; cf. [22, 23, 30].

Similar results can be obtained for compact sets of the general form

$$K = \sigma(I \times \Theta) , \;\; \sigma = (\sigma_\ell(t,\theta))_{1\leq\ell\leq d} ,$$

$$t \in I = I_1 \times \cdots \times I_{d_1} , \;\; \theta \in \Theta = \Theta_1 \times \cdots \times \Theta_{d_2+d_3} , \tag{21}$$

$$\sigma_\ell \in \bigotimes_{i=1}^{d_1} \mathbb{P}_1(I_i) \otimes \bigotimes_{j=1}^{d_2+d_3} \mathbb{T}_1(\Theta_j) \;\; , \;\; 1 \leq \ell \leq d , \tag{22}$$

where $d_1, d_2, d_3 \geq 0$, and $I_i = [a_i, b_i]$, $1 \leq i \leq d_1$ (algebraic variables), $\Theta_j = [u_j, v_j]$ with $v_j - u_j = 2\pi$, $1 \leq j \leq d_2$ (periodic trigonometric variables) and $v_j - u_j < 2\pi$, $d_2 + 1 \leq j \leq d_2 + d_3$ (subperiodic trigonometric variables). Here and below $\mathbb{T}_n = span(1, \cos(\theta), \sin(\theta) \ldots, \cos(n\theta), \sin(n\theta))$ denotes the space of univariate trigonometric polynomials of degree not exceeding n. Notice that the mapping σ can be non-injective.

The class (21)-(22) contains many common domains in applications, which have in some sense a tensorial structure. For example in the 2-dimensional case *convex quadrangles* (with triangles as special degenerate cases) fall into this class, because they are bilinear transformations of a square (by the so-called

Duffy transform), with $d_1 = 2$, $d_2 = d_3 = 0$. Similarly the *disk* described in polar coordinates ($d_1 = d_2 = 1$, $d_3 = 0$), the *2-sphere* in spherical coordinates ($d_1 = 0$, $d_2 = d_3 = 1$), the *torus* in toroidal-poloidal coordinates ($d_1 = 0$, $d_2 = 2$, $d_3 = 0$), cf. [7, 27].

Moreover, many examples of sections of disk, sphere, ball, surface and solid torus can be written as (21)-(22). For example, a *circular sector* of the unit disk with angle 2ω, $\omega < \pi$, can be described by such a σ with $d_1 = d_3 = 1$, $d_2 = 0$, e.g.,

$$\sigma(t, \theta) = (t\cos(\theta), t\sin(\theta)) \ , \quad (t, \theta) \in [0, 1] \times [-\omega, \omega] \ , \tag{23}$$

(polar coordinates). Similarly, a *circular segment* with angle 2ω (one of the two portions of the disk cut out by a line) can be described by such a σ with $d_1 = d_3 = 1$, $d_2 = 0$, e.g.,

$$\sigma(t, \theta) = (\cos(\theta), t\sin(\theta)) \ , \quad (t, \theta) \in [-1, 1] \times [-\omega, \omega] \ . \tag{24}$$

On the other hand, a *toroidal rectangle* is described with $d_3 = 2$, $d_1 = d_2 = 0$, by the trasformation

$$\sigma(\theta) = ((R + r\cos(\theta_1))\cos(\theta_2), (R + r\cos(\theta_1))\sin(\theta_2), r\sin(\theta_1)) \ , \tag{25}$$

$\theta = (\theta_1, \theta_2) \in [\omega_1, \omega_2] \times [\omega_3, \omega_4]$, where R and r are the major and minor radii of the torus. In the degenerate case $R = 0$ we get a so-called *geographic rectangle* of a sphere of radius r, i.e. the region between two given latitudes and longitudes. For other planar, surface and solid examples we refer the reader to [26, 27].

By the geometric structure (21)-(22), we have that if $p \in \mathbb{P}_n^d(K)$ then

$$p \circ \sigma \in \bigotimes_{i=1}^{d_1} \mathbb{P}_n(I_i) \otimes \bigotimes_{j=1}^{d_2+d_3} \mathbb{T}_n(\Theta_j) \ , \tag{26}$$

and this allows us to construct product-like polynomial meshes on such domains. Indeed, in the univariate case Chebyshev-like optimal polynomial meshes are known for algebraic polynomials and for trigonometric polynomials (even on subintervals of the period). This result is stated in the following

Lemma 2 *Let $K \subset \mathbb{R}^d$ be a compact set of the form (21)-(22). Then, for every fixed $m > 1$, K possesses a polynomial mesh $\{Z_n(m)\}$ with constant $c = (1/\cos(\pi/(2m)))^{d_1+d_2+d_3}$ and cardinality not exceeding $(mn)^{d_1+d_2}(2mn)^{d_3}$.*

The proof is essentially that given in [27] by resorting to algebraic-trigonometric Chebyshev-like grids mapped by σ, with minor modifications to take into account the later results on subperiodic trigonometric Dubiner distance given in [31]. If Chebyshev-Lobatto-like grids are used, mn and $2mn$ have to be substituted by $mn + 1$ and $2mn + 1$, respectively. If m is not an integer, all these quantities should be substituted by their ceiling (the least integer not smaller than).

By Lemma 2 we get immediately the following proposition.

Proposition 3 *Let $K \subset \mathbb{R}^d$ be a compact set of the form (21)-(22) and $\{Z_n(m)\}$ the polynomial mesh of Lemma 2.*

Then for every $n \in \mathbb{N}$ and $m > 1$, the probability measure

$$\nu = \nu(n,m) = \mu^*(Z_{2n}(m)) \tag{27}$$

is a near G-optimal design on K, in the sense that

$$\max_{x\in K} K_n^\nu(x,x) \le c_m\, N\ ,\quad c_m = \left(\frac{1}{\cos(\pi/(2m))}\right)^{d_1+d_2+d_3}\ . \tag{28}$$

Concerning G-efficiency we have now

$$\mathrm{G}_{\mathrm{eff}}(\nu) \ge (\cos(\pi/(2m)))^{d_1+d_2+d_3} \sim 1-(d_1+d_2+d_3)\pi^2/(8m^2)\ . \tag{29}$$

Remark 1 Observe that by Propositions 1-3, reasoning as in (11) we get

$$\|L_n^\nu\| \le \sqrt{c_m\, N}\ , \tag{30}$$

i.e. the discrete probability measure ν nearly minimizes (the estimate of) the weighted Least Squares uniform operator norm.

3 Caratheodory-Tchakaloff design concentration

Propositions 1-3 and the General Equivalence Theorem suggest a standard way to compute near G-optimal designs. First, one constructs a polynomial mesh such as X_{2mn} or $Y_{2n}(\pi/(2m))$ or $Z_{2n}(m)$, then computes a D-optimal design for degree n on the mesh by one of the available algorithms. Observe that such designs will be in general approximate, that is we compute a discrete probability measure $\tilde\nu \approx \nu$ such that on the polynomial mesh

$$\max_{x\in mesh} K_n^{\tilde\nu}(x,x) \le \tilde N \approx N \tag{31}$$

(with $\tilde N$ not necessarily an integer), nevertheless estimates (13), (19) and (30) still hold with $\tilde\nu$ and $\tilde N$ replacing ν and N, respectively.

Again, we can not even attempt to survey the vast literature on computational methods for D-optimal designs; we may quote among others the class of exchange algorithms and the class of multiplicative algorithms, cf. e.g. [10, 17, 29] and the references therein.

Our computational strategy is in brief the following. We first approximate a D-optimal design for degree n on the polynomial mesh by a standard multiplicative algorithm, and then we concentrate the measure via Caratheodory-Tchakaloff compression of degree $2n$, keeping the Christoffel polynomial, and thus G-efficiency, invariant. Such a compression is based on a suitable implementation of a discrete version of the well-known Tchakaloff Theorem, which in general asserts that any (probability) measure has a representing atomic measure with the same polynomial moments up to a given degree, with cardinality not exceeding the dimension of the corresponding polynomial space; cf. e.g. [24] and [20, 25] and the references therein. In such a way we get near optimality with respect to both, G-efficiency and support cardinality, since the latter will not exceed $N_{2n} = \dim(\mathbb{P}^d_{2n}(K))$.

To simplify the notation, in what follows we shall denote by $X = \{x_i\}$ either the polynomial mesh $X = X_{2mn}$ or $X = Y_{2n}(\pi/(2m))$ or $X = Z_{2n}(m)$ (cf. Propositions 1-3), by M its cardinality, by $w = \{w_i\}$ the weights of a probability measure on X ($w_i \geq 0$, $\sum w_i = 1$), and by $K_n^w(x, x)$ the corresponding Christoffel polynomial.

The first step is the application of the standard Titterington's multiplicative algorithm [28] to compute a sequence $w(\ell)$ of weight arrays

$$w_i(\ell + 1) = w_i(\ell) \, \frac{K_n^{w(\ell)}(x_i, x_i)}{N} \; , \quad 1 \leq i \leq M \; , \quad \ell \geq 0 \; , \tag{32}$$

where we take $w(0) = (1/M, \ldots, 1/M)$. Observe that the weights $w_i(\ell + 1)$ determine a probability measure on X, since they are clearly nonnegative and $\sum_i w_i(\ell) \, K_n^{w(\ell)}(x_i, x_i) = N$. The sequence $w(\ell)$ is known to converge for any initial choice of probability weights to the weights of a D-optimal design (with a nondecreasing sequence of Gram determinants), cf. e.g. [12] and the references therein.

In order to implement (32), we need an efficient way to compute the right-hand side. Denote by $V_n = (\phi_j(x_i)) \in \mathbb{R}^{M \times N}$ the rectangular Vandermonde matrix at X in a fixed polynomial basis $(\phi_1, \ldots, \phi_N)$, and by $D(w)$ the diagonal matrix of a weight array w. In order to avoid severe ill-conditioning that may already occur for low degrees, we have discarded the monomial basis and used the product Chebyshev basis of the smallest box containing X, a choice that turns out to work effectively in multivariate instances; cf. e.g. [5, 19, 21].

By the QR factorization

$$D^{1/2}(w) \, V_n = QR \; ,$$

with $Q = (q_{ij})$ orthogonal (rectangular) and R square upper triangular, we have that $(p_1, \ldots, p_N) = (\phi_1, \ldots, \phi_N)R^{-1}$ is a w-orthonormal basis and

$$w_i \, K_n^w(x_i, x_i) = w_i \, \sum_{j=1}^{N} p_j^2(x_i) = \sum_{j=1}^{N} q_{ij}^2 \; , \quad 1 \leq i \leq M \; . \tag{33}$$

Thus we can update the weights at each step of (32) by a single QR factorization, using directly the squared 2-norms of the rows of the orthogonal matrix Q.

The convergence of (32) can be slow, but a few iterations usually suffice to obtain an already quite good design on X. Indeed, in all our numerical tests with bivariate polynomial meshes, after 10 or 20 iterations we already get 90% G-efficiency on X, and 95% after 20 or 30 iterations; cf. Figure 1-left and 2-left for typical convergence profiles. On the other hand, 99% G-efficiency would require hundreds, and $99,9\%$ thousands of iterations. When a G-efficiency very close to 1 is needed, one should choose one of the more sophisticated approximation algorithms available in the literature, cf. e.g. [11, 12, 17] and the references therein.

Though the designs given by (32) will concentrate in the limit on the support of an optimal design, which typically is of relatively low cardinality (with respect to M), this will be not numerically evident after only a small number of iterations. Hence, *in practice*, the support of the optimal measure is not readily identified, and a practioner may be presented with a measure with

full support (albeit with many small weights). However, using Tchakaloff compression (described below) the cardinality of the support can be immediately reduced providing the practioner with a typically much smaller, and hence more pratical, design set.

Let $V_{2n} \in \mathbb{R}^{M \times N_{2n}}$ be the rectangular Vandermonde matrix at X with respect to a fixed polynomial basis for $\mathbb{P}^d_{2n}(X) = \mathbb{P}^d_{2n}(K)$ (recall that the chosen polynomial mesh is determining on K for polynomials of degree up to $2n$), and w the weight array of a probability measure supported on X (in our instance, the weights produced by (32) after a suitable number of iterations, to get a prescribed G-efficiency on X). In this fully discrete framework the Tchakaloff Theorem is equivalent to the existence of a sparse nonnegative solution u to the underdetermined moment system

$$V_{2n}^t u = b = V_{2n}^t w \ , \quad u \geq 0 \ , \tag{34}$$

where b is the vector of discrete w-moments of the polynomial basis up to degree $2n$. The celebrated Caratheodory Theorem on conical finite-dimensional linear combinations [9], ensures that such a solution exists and has no more than N_{2n} nonzero components.

In order to compute a sparse solution, we can resort to Linear or Quadratic Programming. We recall here the second approach, that turned out to be the most efficient in all the tests on bivariate discrete measure compression for degrees in the order of tens that we carried out, cf. [21]. It consists of seeking a sparse solution $\hat{u}$ to the NonNegative Least Squares problem

$$\|V_{2n}^t \hat{u} - b\|_2^2 = \min_{u \geq 0} \|V_{2n}^t u - b\|_2^2 \tag{35}$$

using the Lawson-Hanson active set algorithm [16], that is implemented for example in the Matlab native function `lsqnonneg`. The nonzero components of $\hat{u}$ determine the resulting design, whose support, say $T = \{x_i : \hat{u}_i > 0\}$, has at most N_{2n} points.

Observe that by construction $K_n^{\hat{u}}(x, x) = K_n^w(x, x)$ on K, since the underlying probability measures have the same moments up to degree $2n$ and hence generate the same orthogonal polynomials. Now, since

$$\max_{x \in K} K_n^w(x, x) \leq c_m \max_{x \in X} K_n^w(x, x) = \frac{c_m N}{\theta} \ ,$$

where θ is the G-efficiency of w on X, in terms of G-efficiency on K we have the estimate

$$\mathrm{G}_{\mathrm{eff}}(\hat{u}) = \mathrm{G}_{\mathrm{eff}}(w) \geq \frac{\theta}{c_m} \ , \tag{36}$$

cf. Propositions 1-3, while in terms of the uniform norm of the weighted Least Squares operator we get the estimate

$$\|L_n^{\hat{u}}\| \leq \sqrt{\frac{c_m N}{\theta}} \ . \tag{37}$$

We present now several numerical tests. All the computations have been made in Matlab R2017b on a 2.7 GHz Intel Core i5 CPU with 16GB RAM. As a first example we consider polynomial regression on the square $K = [-1, 1]^2$.

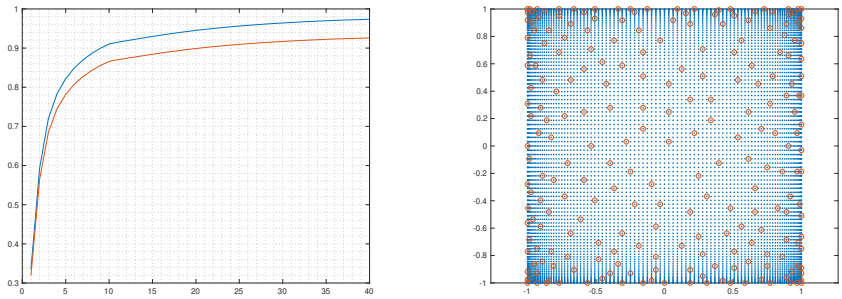

Figure 1: Left: G-efficiency of the approximate optimal designs computed by (32) on a 101×101 Chebyshev-Lobatto grid of the square (upper curve, $n = 10$, $m = 5$), and estimate (36) (lower curve); Right: Caratheodory-Tchakaloff compressed support (231 points) after $\ell = 22$ iterations ($\mathrm{G}_{\mathrm{eff}} \approx 0.95$).

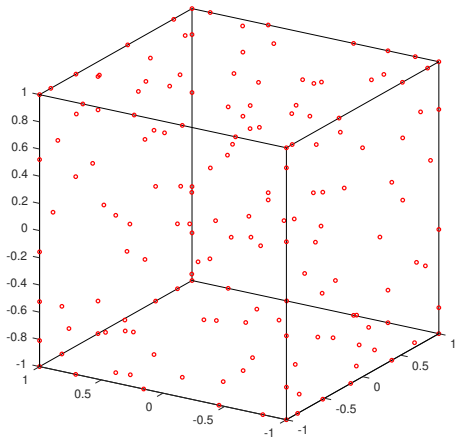

Figure 2: Caratheodory-Tchakaloff compressed support (165 points) on a $41 \times 41 \times 41$ Chebyshev-Lobatto grid of the cube for regression degree $n = 4$ (with $m = 5$), after $\ell = 35$ iterations ($\mathrm{G}_{\mathrm{eff}} \approx 0.95$).

Since $dub_{[-1,1]^2}(x, y) = \max\{\arccos|x_2 - x_1|, \arccos|y_2 - y_1|\}$, cf. [6], by Proposition 2 we can take as initial support $Y_{2n}(\pi/(2m))$ a $(2mn + 1) \times (2mn + 1)$ Chebyshev-Lobatto grid (here $c_m = 1/\cos(\pi/(2m))$, cf. [22]), apply the iteration (32) up to a given G-efficiency and then Caratheodory-Tchakaloff measure compression via (35).

The results corresponding to $n = 10$ and $m = 5$ are reported in Figure 1. Notice that (36) turns out to be an underestimate of the actual G-efficiency on K (the maximum has been computed at a much finer Chebyshev-Lobatto grid, say $Y_{2n}(\pi/(8m))$. All the information required for polynomial regression up to 95% G-efficiency is compressed into $231 = \dim(\mathbb{P}^2_{20})$ sampling nodes and weights, in about 1.7 seconds.

In Figure 2 we present a trivariate example, where $K = [-1, 1]^3$ and we

consider regression degree $n = 4$ and $m = 5$, with a corresponding $41 \times 41 \times 41$ Chebyshev-Lobatto grid. This polynomial mesh of about 68900 points is compressed into $165 = \dim(\mathbb{P}^3_8)$ sampling nodes and weights still ensuring 95% G-efficiency, in about 9 seconds.

In order to check the algorithm behavior on a more complicated domain, we take a 14-sided nonconvex polygon. An application of polygonal compact sets is the approximation of geographical regions; for example, the polygon of Figure 3 resembles a rough approximation of the shape of Belgium. The problem could be that of locating a near minimal number of sensors for near optimal polynomial regression, to sample continuous scalar or vector fields that have to be reconstructed or modelled on the whole region.

With polygons we can resort to triangulation and finite union as in (3), constructing on each triangle a polynomial mesh like $Z_{2n}(m)$ in Proposition 3 by the Duffy transform of a Chebyshev-grid of the square with approximately $(2mn)^2$ points; here $c_m = 1/\cos^2(\pi/(2m))$ for any triangle and hence for the whole polygon. The results corresponding to $n = 8$ and $m = 5$ are reported in Figure 3. The G-efficiency convergence profile is similar to that of Figure 1, and the whole polynomial mesh of about 84200 points is compressed into $153 = \dim(\mathbb{P}^2_{16})$ sampling nodes and weights still ensuring 95% G-efficiency, in about 8 seconds.

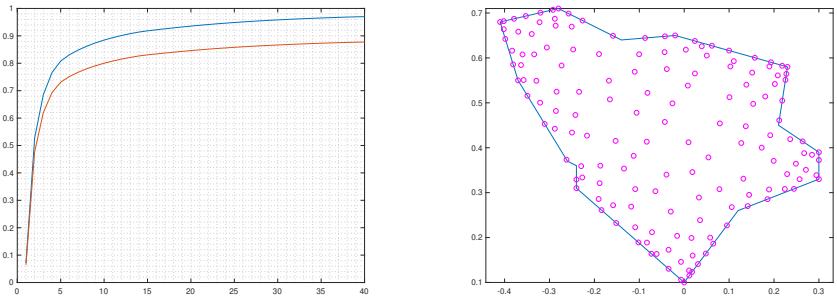

Figure 3: Left: G-efficiency of the approximate optimal designs computed by (32) on a polynomial mesh with about 84200 points of a 14-sided nonconvex polygon (upper curve, $n = 8$, $m = 5$), and estimate (36) (lower curve); Right: Caratheodory-Tchakaloff compressed support (153 points) after $\ell = 26$ iterations ($\mathrm{G}_{\mathrm{eff}} \approx 0.95$).

Remark 2 The practical implementation of *any* design requires an interpretation of the weights. As before, let $X := \{x_i\}_{i=1}^M \subset K$ be the support of a discrete measure with $w_i > 0$, $1 \le i \le M$. We will denote this measure by μ_X. Further, let $V_n := (\phi_j(x_i)) \in \mathbb{R}^{M \times N}$ be the Vandermonde evaluation matrix for

the basis $\{\phi_1, \cdots, \phi_N\}$. The Gram (information) matrix is then given by

$$\begin{aligned} G &= \left(\int_K \phi_i(x)\phi_j d\mu_X \right)_{1 \le i,j \le N} \\ &= \left(\sum_{k=1}^{M} w_k \phi_i(x_k)\phi_j(x_k) \right)_{1 \le i,j \le N} \\ &= V_n^t D(w) V_n \end{aligned}$$

where $D(w) \in \mathbb{R}^{M \times M}$ is the diagonal matrix of the weights w. Then the best least squares approximation from $\mathcal{P}_n^d(K)$, with respect to the measure μ_X, to observations y_i, $1 \le i \le M$, is given by

$$\sum_{j=1}^{N} c_j \phi_j(x), \qquad c = (V_n^t D(w) V_n)^{-1} V_n^t D(w) y. \tag{38}$$

For an optimal design, the determinant of the Gram matrix $V_n^t D(w) V_n$ is as large as possible and hence (38) could be used as a numerically stable (at least as much as possible) algorithm for computing an approximation to the given data. However, it is also useful to exploit the statistical meaning of the weights. Indeed, in comparison, underlying the statistical interpretation of least squares is the assumption that the observations are samples of a polynomial $p \in \mathcal{P}_n^d(K)$, at the design points x_i, each with an error ϵ_i assumed to be independent normal random variables $\epsilon_i \sim N(0, \sigma_i^2)$. One then minimizes the sum of the squares of the *normalized* variables,

$$\frac{\epsilon_i}{\sigma_i} = \frac{p(x_i) - y_i}{\sigma_i},$$

i.e., one minimizes

$$\sum_{i=1}^{M} \left(\frac{p(x_i) - y_i}{\sigma_i} \right)^2.$$

If we write $p(x) = \sum_{j=1}^{N} c_j \phi_j(x)$, then this may be expressed in matrix-vector form as

$$\| C^{-1/2}(V_n c - y) \|_2^2$$

where $C = D(\sigma_i^2) \in \mathbb{R}^{M \times M}$ is the (diagonal) covariance matrix. This is minimized by

$$c = (V_n^t C^{-1} V_n)^{-1} V_n C^{-1} y. \tag{39}$$

Comparing (39) with (38) we see that the weights w_i correspond the reciprocals of the variances, $w_i \sim 1/\sigma_i^2$, $1 \le i \le M$, but normalized so that $\sum_{i=1}^{M} w_i = 1$.

Now, if in principle any specific measurement has an error with a *fixed* variance σ^2 then the variance for an observation may be reduced by repeating the ith measurement m_i (say) times and then using the average $\overline{y}_i$ in place of y_i, with resulting error variance σ^2/m_i. Then $w_i \sim 1/\sigma_i^2 = m_i/\sigma^2$ which, after normalization, results in

$$w_i = \frac{m_i}{\sum_{j=1}^{M} m_j}, \quad 1 \le i \le M.$$

In other words, the weights indicate the percentage of the total measurement budget to use at the ith observation point.

However, the computed weights are rarely rational and thus to obtain a useable percentage some rounding must be done. It turns out that the effect of this rounding on the determinant of the Gram matrix can be readily estimated. Indeed we may calculate

$$\begin{aligned}\frac{\partial}{\partial w_i}\log(\det(G_n^\mu)) &= \frac{\partial}{\partial w_i}\mathrm{tr}(\log(G_n^\mu)) \\ &= \mathrm{tr}\left(\frac{\partial}{\partial w_i}\log(G_n^\mu)\right) \\ &= \mathrm{tr}\left((G_n^\mu)^{-1}\frac{\partial G_n^\mu}{\partial w_i}\right).\end{aligned}$$

But, as we may write

$$G_n^\mu = \sum_{k=1}^{M} w_k p(x_k)p^t(x_k)$$

where $p(x)$ is the vector $p(x) = [\phi_1(x), \phi_2(x), \cdots, \phi_N(x)]^t \in \mathbb{R}^M$, we have

$$\frac{\partial G_n^\mu}{\partial w_i} = p(x_i)p^t(x_i)$$

and hence

$$\begin{aligned}\frac{\partial}{\partial w_i}\log(\det(G_n^\mu)) &= \mathrm{tr}\left((G_n^\mu)^{-1}p(x_i)p^t(x_i)\right) \\ &= p^t(x_i)(G_n^\mu)^{-1}p(x_i) \\ &= K_n^\mu(x_i, x_i).\end{aligned}$$

For an optimal design $K_n^\mu(x_i, x_i) = N$, $1 \le i \le M$ and for our near optimal designs this is nearly so. Hence a perturbation in a weight results in a *relative* perturbation in the determinant amplified by around a factor of N. In adjusting the weights, some roundings will be up and others down and so these perturbations will tend to negate each other. In other words, the rounding strategy is entirely practical.

4 Summary

In this paper we have shown that polynomial meshes (norming sets) can be used as useful discretizations of compact sets $K \subset \mathbb{R}^d$ also for the purposes of (near) optimal statistical designs. We have further shown how the idea of Tchakaloff compression of a discrete measure can be efficiently used to concentrate the design measure onto a relatively small subset of its support, thus making any least squares calculation rather more pratical.

References

[1] A.K. Atkinson and A.N. Donev, Optimum Experimental Designs, Clarendon Press, Oxford, 1992.

[2] T. Bloom, L. Bos, N. Levenberg and S. Waldron, On the convergence of optimal measures, Constr. Approx. 32 (2010), 159–179.

[3] T. Bloom, L. Bos and N. Levenberg, The Asymptotics of Optimal Designs for Polynomial Regression, arXiv preprint: 1112.3735.

[4] L. Bos, Some remarks on the Fejér problem for Lagrange interpolation in several variables, J. Approx. Theory 60 (1990), 133–140.

[5] L. Bos, J.P. Calvi, N. Levenberg, A. Sommariva and M. Vianello, Geometric Weakly Admissible Meshes, Discrete Least Squares Approximations and Approximate Fekete Points, Math. Comp. 80 (2011), 1601–1621.

[6] L. Bos, N. Levenberg and S. Waldron, Pseudometrics, distances and multivariate polynomial inequalities, J. Approx. Theory 153 (2008), 80–96.

[7] L. Bos and M. Vianello, Low cardinality admissible meshes on quadrangles, triangles and disks, Math. Inequal. Appl. 15 (2012), 229–235.

[8] J.P. Calvi and N. Levenberg, Uniform approximation by discrete least squares polynomials, J. Approx. Theory 152 (2008), 82–100.

[9] C. Caratheodory, Über den Variabilittsbereich der Fourierschen Konstanten von positiven harmonischen Funktionen, Rend. Circ. Mat. Palermo 32 (1911), 193–217.

[10] G. Celant and M. Broniatowski, Interpolation and Extrapolation Optimal Designs 2 - Finite Dimensional General Models, Wiley, 2017.

[11] Y. De Castro, F. Gamboa, D. Henrion, R. Hess, J.-B. Lasserre, Approximate Optimal Designs for Multivariate Polynomial Regression, Ann. Statist. 47 (2019), 127–155.

[12] H. Dette, A. Pepelyshev and A. Zhigljavsky, Improving updating rules in multiplicative algorithms for computing D-optimal designs, Comput. Stat. Data Anal. 53 (2008), 312–320.

[13] M. Dubiner, The theory of multidimensional polynomial approximation, J. Anal. Math. 67 (1995), 39–116.

[14] J. Kiefer and J. Wolfowitz, The equivalence of two extremum problems, Canad. J. Math. 12 (1960), 363-366.

[15] A. Kroó, On optimal polynomial meshes, J. Approx. Theory 163 (2011), 1107–1124.

[16] C.L. Lawson and R.J. Hanson, Solving Least Squares Problems, Classics in Applied Mathematics 15, SIAM, Philadelphia, 1995.

[17] A. Mandal, W.K. Wong and Y. Yu, Algorithmic Searches for Optimal Designs, in: Handbook of Design and Analysis of Experiments (A. Dean, M. Morris, J. Stufken, D. Bingham Eds.), Chapman & Hall/CRC, New York, 2015.

[18] F. Piazzon, Optimal Polynomial Admissible Meshes on Some Classes of Compact Subsets of $\mathbb{R}^d$, J. Approx. Theory 207 (2016), 241–264.

[19] F. Piazzon, Pluripotential Numerics, Constr. Approx. 49 (2019), 227–263.

[20] F. Piazzon, A. Sommariva and M. Vianello, Caratheodory-Tchakaloff Subsampling, Dolomites Res. Notes Approx. DRNA 10 (2017), 5–14.

[21] F. Piazzon, A. Sommariva and M. Vianello, Caratheodory-Tchakaloff Least Squares, Sampling Theory and Applications 2017, IEEE Xplore Digital Library, DOI: 10.1109/SAMPTA.2017.8024337.

[22] F. Piazzon and M. Vianello, A note on total degree polynomial optimization by Chebyshev grids, Optim. Lett. 12 (2018), 63–71.

[23] F. Piazzon and M. Vianello, Markov inequalities, Dubiner distance, norming meshes and polynomial optimization on convex bodies, Optim. Lett., published online 01 January 2019.

[24] M. Putinar, A note on Tchakaloff's theorem, Proc. Amer. Math. Soc. 125 (1997), 2409–2414.

[25] A. Sommariva and M. Vianello, Compression of multivariate discrete measures and applications, Numer. Funct. Anal. Optim. 36 (2015), 1198–1223.

[26] A. Sommariva and M. Vianello, Polynomial fitting and interpolation on circular sections, Appl. Math. Comput. 258 (2015), 410–424.

[27] A. Sommariva and M. Vianello, Discrete norming inequalities on sections of sphere, ball and torus, J. Inequal. Spec. Funct. 9 (2018), 113–121.

[28] D.M. Titterington, Algorithms for computing d-optimal designs on a finite design space, in: Proc. 1976 Conference on Information Sciences and Systems, Baltimora, 1976.

[29] B. Torsney and R. Martin-Martin, Multiplicative algorithms for computing optimum designs, J. Stat. Plan. Infer. 139 (2009), 3947–3961.

[30] M. Vianello, Global polynomial optimization by norming sets on sphere and torus, Dolomites Res. Notes Approx. DRNA 11 (2018), 10–14.

[31] M. Vianello, Subperiodic Dubiner distance, norming meshes and trigonometric polynomial optimization, Optim. Lett. 12 (2018), 1659–1667.