Near optimal polynomial regression on norming meshes

Len Bos *, Federico Piazzon †, Marco Vianello †
*Department of Computer Science, University of Verona (Italy)
Email: leonardpeter.bos@univr.it
†Department of Mathematics, University of Padova (Italy)
Email: fpiazzon@math.unipd.it, marcov@math.unipd.it

Abstract—**We connect the approximation theoretic notions of polynomial norming mesh and Tchakaloff-like quadrature to the statistical theory of optimal designs, obtaining near optimal polynomial regression at a near optimal number of sampling locations on domains with different shapes.**

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

Keywords: optimal designs, polynomial regression, norming set, polynomial norming mesh, Tchakaloff-like quadrature.

I. Introduction

In this paper we apply the approximation theoretic notion of norming set (in particular, of “polynomial norming mesh”) on a multidimensional compact set K within the statistical theory of optimal polynomial regression designs on K. Moreover, we propose an approach to reduce the sampling cardinality, based on a recent implementation of Tchakaloff-like quadrature.

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

$$\|p\|_K \le c\,\|p\|_{X_n} \;,\; \forall p \in \mathbb{P}_n^d(K) \;,\; \mathrm{card}(X_n) = \mathcal{O}(N^s) \;, \quad (1)$$

where $s \ge 1$ and $N = N_n(K) = \dim(\mathbb{P}_n^d(K))$. Observe that necessarily $\mathrm{card}(X_n) \ge N$, since X_n is $\mathbb{P}_n^d(K)$-determining. On the other hand, $N = \mathcal{O}(n^\beta)$ with $\beta \le d$, 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 $N = \binom{n+d}{d} - \binom{n-2+d}{d}$.

We recall that norming meshes can be computed on a wide family of compact sets, containing for example all compact sets satisfying a Markov polynomial inequality, in particular all compact sets with Lipschitz boundary. Norming meshes with $\mathcal{O}(N)$ cardinality, often called “optimal”, are known for several classes of compact sets, for example polytopes and smooth bodies, cf. [13], [15]. It is also worth recalling that norming meshes are preserved by affine transformations, and can be incrementally constructed by finite unions and products; cf. the seminal paper [8].

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

$$\|L_n\| = \sup_{f \ne 0} \frac{\|L_n f\|_K}{\|f\|_K} \le c\,\sqrt{\mathrm{card}(X_n)}\;. \quad (2)$$

In addition, norming meshes play a role in the computation of good interpolations sets of Fekete and Leja type, and have been applied in the fields of polynomial optimization and pluripotential numerics; cf., e.g., [5], [16], [18], [19].

The problem of reducing the sampling cardinality keeping invariant estimate (2) (Least Squares compression) has been solved in [17], via weighted Least Squares on N_{2n} Caratheodory-Tchakaloff points extracted from the norming mesh by Linear or Quadratic Programming. Nevertheless, also reducing the Least Squares uniform operator norm, though much more costly is important in applications, and this will be addressed in the next section via the theory of optimal designs.

II. Near optimal designs by norming meshes

In statistics a design is a probability measure μ supported on a (discrete or continuous) compact set $K \subset \mathbb{R}^d$. The search for designs that optimize some property of statistical estimators (optimal designs) began at least one century ago; the corresponding literature is so vast and still growing that we can not even attempt any kind of survey. We may for example quote the classical book [1] and the very recent paper [10]. Below we recall some relevant notions and results, in order to connect the theory of optimal designs with the theory of norming meshes.

In what follows we 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 $\mathbb{P}_n^d(K)$. The diagonal of the reproducing kernel for μ in $\mathbb{P}_n^d(K)$ (often called the *Christoffel polynomial*)

$$K_n^\mu(x,x) = \sum_{j=1}^{N} p_j^2(x)\;, \quad (3)$$

where $\{p_j\}$ is any μ-orthonormal basis of $\mathbb{P}_n^d(K)$, plays a key role in the theory of optimal designs (it can be shown that $K_n^\mu(x,x)$ is independent of the choice of the orthonormal basis). Indeed, a relevant property is that

$$\|p\|_K \leq \sqrt{\max_{x\in K} K_n^\mu(x,x)}\,\|p\|_{L^2_\mu(K)}\ ,\ \ \forall p \in \mathbb{P}_n^d(K)\ . \tag{4}$$

Now, by (3) we get immediatly $\int_K K_n^\mu(x,x)\,d\mu = N$, which implies that $\max_{x\in K} K_n^\mu(x,x) \geq N$. A probability measure $\mu^* = \mu^*(K)$ is then called a G-optimal design for polynomial regression of degree n on K if

$$\min_\mu \max_{x\in K} K_n^\mu(x,x) = \max_{x\in K} K_n^{\mu^*}(x,x) = N\ . \tag{5}$$

Observe that, since $\int_K K_n^\mu(x,x)\,d\mu = N$ for every μ, an optimal design has also the following property $K_n^{\mu^*}(x,x) = N$, $\mu^* - a.e.$ in K.

A cornerstone of optimal design theory, the well-known Kiefer-Wolfowitz General Equivalence Theorem [12], says that the difficult min-max problem (5) is equivalent to the much simpler maximization

$$\max_\mu det(G_n^\mu)\ ,\ \ G_n^\mu = \left(\int_K q_i(x)q_j(x)\,d\mu\right)_{1\leq i,j\leq N}\ , \tag{6}$$

where G_n^μ is the Gram matrix of μ in a fixed polynomial basis $\{q_i\}$ (also called the information matrix in statistics). This kind of optimality is called D-optimality, and ensures that an optimal measure always exists, since the set of Gram matrices of probability measures is compact (and convex); see e.g. [1], [2], [4] for a 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 atomic optimal measure always exists by Tchakaloff's Theorem on positive quadratures of degree $2n$ for K; cf. [20] for a general proof of Tchakaloff's Theorem. Moreover, it has been proved in [2], [3] that optimal designs converge weakly as $n \to \infty$ to the pluripotential theoretic equilibrium measure of the compact set.

G-optimality can be interpreted in a probabilistic as well as in a deterministic framework. 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, denoting by $L_n^{\mu^*} : C(K) \to \mathbb{P}_n^d(K)$ the corresponding weighted Least Squares projection operator, in view of (4) and (5) for every $f \in C(K)$ we have the chain of estimates

$$\frac{\|L_n^{\mu^*} f\|_K}{\sqrt{N}} \leq \|L_n^{\mu^*} f\|_{L^2_{\mu^*}(K)} \leq \|f\|_{L^2_{\mu^*}(K)} \leq \|f\|_K\ , \tag{7}$$

and thus $\|L_n^{\mu^*}\| \leq \sqrt{N}$, i.e. a G-optimal measure minimizes (the estimate of) the weighted Least Squares operator norm.

There is a vast literature on the computation of D-optimal designs, with many different approaches and methods. A classical approach is given by the discretization of K and then the D-optimization over the discrete set. In the discretization framework, the possible role of norming meshes seems apparently overlooked. A simple but meaningful result is given in the following proposition.

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

Then for every $n, m \in \mathbb{N}^+$, the probability measure

$$\nu = \nu(n,m) = \mu^*(X_{2mn}) \tag{8}$$

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

$$\max_{x\in K} K_n^\nu(x,x) \leq c_m\, N\ ,\ \ c_m = c^{1/m}\ . \tag{9}$$

Proposition 1 shows that norming meshes are good discretizations of a compact set for the purpose of computing a near G-optimal measure, and that G-optimality maximum condition (5) is approached at a rate proportional to $1/m$, since $c_m \sim 1 + \log(c)/m$. Recalling the statistical notion of G-efficiency on K we have

$$\mathrm{G_{eff}}(\nu) = \frac{N}{\max_{x\in K} K_n^\nu(x,x)} \geq c^{-1/m}\ , \tag{10}$$

whereas concerning the norm of the corresponding weighted Least Squares projection operator

$$\|L_n^\nu\| \leq \sqrt{c_m\, N}\ , \tag{11}$$

i.e. the discrete probability measure ν nearly minimizes (the estimate of) such a norm.

It is worth recalling that a better rate proportional to $1/m^2$ can be obtained on certain compact sets, such as triangles and quadrangles, cube, simplex, (sections of) sphere and ball, smooth convex bodies, where low cardinality norming meshes can be constructed via the approximation theoretic notion of Dubiner distance and suitable geometric transformations; cf. [6], [7], [19], [24].

III. Tchakaloff-like design concentration

Proposition 1 and the General Equivalence Theorem suggest a standard way to compute near G-optimal designs. First, one constructs a norming mesh such as X_{2mn}, then computes a D-optimal design for degree n on such a set 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 norming mesh

$$\max_{x\in mesh} K_n^{\tilde{\nu}}(x,x) \leq \tilde{N} \approx N \tag{12}$$

(with $\tilde{N}$ not necessarily an integer), nevertheless estimates (9)-(11) 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. [11], [14] and the references therein.

Our computational strategy is roughly the following. We first approximate a D-optimal design for degree n on the norming 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 the G-efficiency, invariant. Such a compression is based on a suitable implementation of a discrete version of the well-known Tchakaloff Theorem [20], [23], 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; for an implementation see e.g. [17], [21] 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 $X = X_{2mn}$, $M = card(X)$, $w = \{w_i\}$ are the weights of a probability measure on X ($w_i \geq 0$, $\sum w_i = 1$), and $K^w_n(x, x)$ is the corresponding Christoffel polynomial.

The first step is the application of the standard Titterington's multiplicative algorithm (cf. [14]) to compute a sequence $w(k)$ of weight arrays

$$w_i(k+1) = w_i(k)\,\frac{K_n^{w(k)}(x_i, x_i)}{N}\ ,\ \ 1 \leq i \leq M\ ,\ \ k \geq 0\ , \tag{13}$$

where we take $w(0) = (1/M, \ldots, 1/M)$. Observe that the weights $w_i(k+1)$ determine a probability measure on X, since they are clearly nonnegative and $\sum_i w_i(k)\, K_n^{w(k)}(x_i, x_i) = N$. The sequence $w(k)$ is known to converge as $k \to \infty$, 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. [11] and the references therein.

In order to implement (13), 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 relatively low degrees, instead of the standard monomial basis we have 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], [16], [17]. In view of the rectangular QR factorization $D^{1/2}(w)\, V_n = QR$ with $Q = (q_{ij})$ orthogonal (rectangular) and R square upper triangular, the polynomials $(p_1, \ldots, p_N) = (\phi_1, \ldots, \phi_N)R^{-1}$ form a w-orthonormal basis and we can write

$$w_i\, K^w_n(x_i, x_i) = w_i \sum_{j=1}^{N} p_j^2(x_i) = \sum_{j=1}^{N} q_{ij}^2\ ,\ \ 1 \leq i \leq M\ . \tag{14}$$

The latter equation shows that we can update the weights at each step of (13) by a single QR factorization, using directly the squared 2-norms of the rows of the orthogonal matrix Q.

The convergence of (13) can be slow, but a few iterations usually suffice to obtain a good design on X. Indeed, in all our numerical tests with bivariate norming meshs, after 10 or 20 iterations we already get 90% G-efficiency on X, and 95% after 20 or 30 iterations; cf. Figure 1-top for a typical convergence profile. 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 sought, one should adopt one of the more sophisticated approximation algorithms available in the literature, cf. e.g. [10], [11], [14] and the references therein.

Though the designs given by (13) will concentrate in the limit on the support of an optimal design, which typically is of relatively low cardinality (with respecy to M), the cardinality of the support can be reduced even after a small number of iterations by a suitable implementation of Tchakaloff's Theorem, that we describe below.

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 norming 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 (13) after a suitable number of iterations, to get a prescribed G-efficiency on X). In this fully discrete framework Tchakaloff's Theorem corresponds to the existence of a sparse solution u to the underdetermined moment system

$$V^t_{2n}u = b = V^t_{2n}w\ ,\ \ u \geq 0\ , \tag{15}$$

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. [17]. It consists of seeking a sparse solution $\hat{u}$ to the NonNegative Least Squares problem

$$\|V^t_{2n}\hat{u} - b\|_2^2 = \min_{u\geq 0} \|V^t_{2n}u - b\|_2^2 \tag{16}$$

using the Lawson-Hanson active set algorithm, 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 $K^{\hat{u}}_n(x, x) = K^w_n(x, x)$ on K, since the underlying probability measures have the same moments up to degree $2n$ and hence generate the same orthogonal polynomials. Now, since

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

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

$$\mathrm{G}_{\text{eff}}(\hat{u}) = \mathrm{G}_{\text{eff}}(w) \geq \frac{\theta}{c_m}\ , \tag{17}$$

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

$$\|L^{\hat{u}}_n\| \leq \sqrt{\frac{c_m N}{\theta}}\ . \tag{18}$$

We present now a bivariate example on a nonconvex polygon. An application of polygonal compact sets is the approximation of geographical regions; for example, the 27-sided polygon in Figure 1 resembling the shape of France. The problem could be that of locating a near minimal number of sampling stations (sensors) to reconstruct a scalar or vector field (such as rainfall, pollutants concentration, geomagnetic field, ...) by near optimal regression on the whole region.

With polygons we can resort to triangulation and finite union, constructing on each triangle a norming mesh 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 by finite union for the whole polygon, cf. [7], [8]. The results corresponding to $n = 8$ and $m = 5$ are reported in Figure 1; all the computations have been made in Matlab R2017b on a 2.7 GHz Intel Core i5 CPU with 16GB RAM. The whole norming mesh of about 168500 points is compressed into 153 sampling nodes and weights (a compression ratio of 3 orders of magnitude) still ensuring 95% G-efficiency, in about 22 seconds.

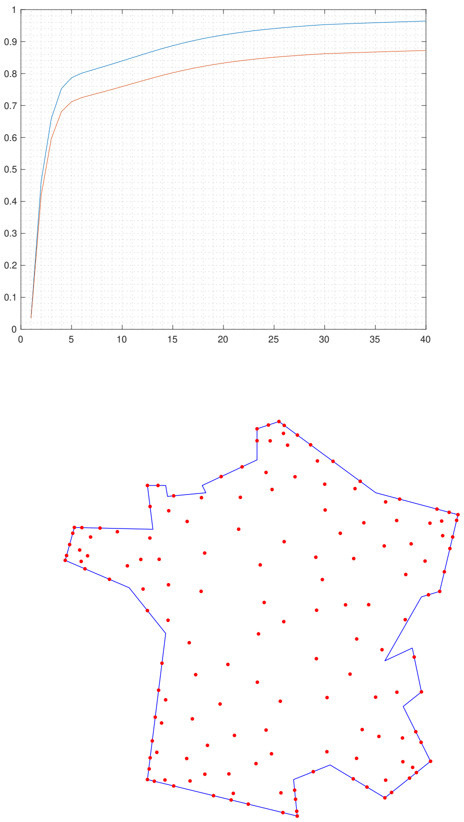

Fig. 1: Top: G-efficiency of the approximate optimal designs computed by (13) on a norming mesh with about 168500 points of a 27-sided nonconvex polygon (upper curve, $n = 8$, $m = 5$), and estimate (17) (lower curve); Bottom: Caratheodory-Tchakaloff compressed support (153 points) after $k = 35$ iterations ($G_{\text{eff}} = 0.95$).

Acknowledgments

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".

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, F. Piazzon and M. Vianello, Near G-Optimal Tchakaloff Designs, submitted, 2019.

[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] 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.

[11] 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.

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

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

[14] A. Mandal, W.K. Wong and Y. Yu, Algorithmic Searches for Optimal Designs, in: Handbook of Design and Analysis of Experiments, Chapman & Hall/CRC, New York, 2015.

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

[16] F. Piazzon, Pluripotential Numerics, Constr. Approx., published online 21 June 2018.

[17] F. Piazzon, A. Sommariva and M. Vianello, Caratheodory-Tchakaloff Least Squares, SampTA 2017, IEEE Xplore Digital Library.

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

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

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

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

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

[23] V. Tchakaloff, Formules de cubatures mécaniques à coefficients non négatifs. (French) Bull. Sci. Math. 81 (1957), 123–134.

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