

# GEOMETRY OF SAMPLE SETS IN DERIVATIVE-FREE OPTIMIZATION: POLYNOMIAL REGRESSION AND UNDERDETERMINED INTERPOLATION

ANDREW R. CONN\*, KATYA SCHEINBERG†, AND LUÍS N. VICENTE‡

**Abstract.** In recent years, there has been a considerable amount of work on the development of numerical methods for derivative-free optimization problems. Some of this work relies on the management of the geometry of sets of sampling points for function evaluation and model building.

In this paper, we continue the work developed in [8] for complete or determined interpolation models (when the number of interpolation points equals the number of basis elements), considering now the cases where the number of points is higher (regression models) and lower (underdetermined models) than the number of basis components.

We show that regression and underdetermined models essentially have similar properties to the interpolation model in that the mechanisms and concepts which control the quality of the sample sets, and hence of the approximation error bounds, of the interpolation models can be extended to the over- and underdetermined cases. We also discuss the trade-offs between using a fully determined interpolation model and the over- or underdetermined ones.

**1. Introduction.** A class of nonlinear optimization methods called derivative-free methods has been extensively developed in the past decade. These methods do not rely on derivative information of the objective function or constraints, but rather sample the functions considered. Some of these methods use the information directly extracted from sampling the functions, without attempting to form models or to approximate derivatives. These methods are called *direct-search methods* and we do not consider them here. Another popular approach (see [8] and references therein) is to use polynomial interpolation to build a model of the objective function (or constraints). There have been a number of practical algorithms, as well as supporting theory, based on this idea.

It is natural to consider an extension of this approach: if the number of sample points is more than the number of degrees of freedom of an approximating polynomial of a given degree then the interpolation is overdetermined. It might be beneficial to use polynomial regression on a larger set of sample points instead of using interpolation on a subset of these points. Such a situation may arise when function evaluations are not too expensive (so having more sample points than the number of degrees of freedom is not prohibitive) but are computed with noise (hence regression models might have a desirable smoothing affect). The most recent implementation of implicit filtering methods [2, 3] makes use of regression simplex gradients and diagonal simplex Hessians. Simplex gradients based on regression are also used in [11, 12]. But despite the usefulness of the idea there has been little other work employing regression for derivative-free optimization. One of the reasons, in our opinion, is the lack of supporting theory. In this paper we extend the properties of polynomial interpolation described in [8] to the case of least-squares regression. These properties provide the foundation for using least-squares regression models in an algorithmic framework with global convergence properties (see [6, 7, 10]).

---

\*Department of Mathematical Sciences, IBM T.J. Watson Research Center, Route 134, P.O. Box 218, Yorktown Heights, New York 10598, USA ([arconn@us.ibm.com](mailto:arconn@us.ibm.com)).

†Department of Mathematical Sciences, IBM T.J. Watson Research Center, Route 134, P.O. Box 218, Yorktown Heights, New York 10598, USA ([katya@us.ibm.com](mailto:katya@us.ibm.com)).

‡CMUC, Department of Mathematics, University of Coimbra, 3001-454 Coimbra, Portugal ([lnv@mat.uc.pt](mailto:lnv@mat.uc.pt)). Support for this author was provided by FCT under grant POCI/59442/2004.

Another extension of the interpolation models considered in [8] is the case when the number of sample points is smaller than the number of degrees of freedom of the interpolation polynomial. The interpolation is underdetermined in this case. Instead of lowering the degree of the interpolant, it is often beneficial to use all the available sample points to try to extract as much information for the higher order interpolation as possible. Such situations often arise when the function evaluations are expensive and it is prohibitive to sample enough points to provide a *fully* quadratic interpolation model at each iteration, yet using only first-order information is inefficient. In fact, underdetermined interpolation models have been extensively used in practical implementations of derivative-free methods based on second-order approximation. The DFO code described in [1] uses minimum-norm underdetermined interpolation models in a trust-region like method, at the early iterations when not enough points are available for complete interpolation. The approach in [17] intentionally uses incomplete or underdetermined interpolation throughout the course of the optimization algorithm. There, the degrees of freedom in the underdetermined interpolation systems are used to construct models that minimize the Frobenius norm of the change of the second derivative of the quadratic models.

Although the underdetermined case appears to be very similar to the fully determined one, the theory in [8] does not immediately extend. As in the overdetermined case (or regression) we want to show that the underdetermined models (if used properly) can be used in a globally convergent algorithm.

Essentially, as in [8], in order to use polynomial models in a globally convergent algorithmic framework the models have to satisfy Taylor-like error bounds. That is, at each iteration of the algorithm, the linear (quadratic) model which is used has to approximate the true function in a given neighborhood around the current iterate at least as well (up to a uniformly bounded constant) as the first (second) order Taylor expansion at that iterate. In [10] it is shown that a derivative-free trust-region algorithm, based on models that satisfy the above property, is globally convergent. In [8] we have shown that polynomial interpolation models indeed satisfy the desired property. Moreover, algorithms for constructing such models are presented in [8]. In particular, in that paper, we introduced the concept of  $\Lambda$ -poisedness as a condition on the set of sample points, say  $Y$ , based on which the interpolation model is constructed. This condition is equivalent to imposing a bound on the maximum sum of the absolute values of the Lagrange polynomials associated with  $Y$ , also known as the Lebesgue constant of  $Y$ . There are Taylor-like error bounds based on the Lebesgue constant of  $Y$  in the approximation theory literature. Many of them are difficult to interpret in terms suitable for optimization and it is unclear how one can generate or maintain sample sets with bounded Lebesgue constants. In [8] we show how this can be done algorithmically by maintaining lower bounds on pivots during a factorization of a certain Vandermonde matrix. We also show that the size of the pivots is related to the size of the Lebesgue constant by using the concept of  $\Lambda$ -poisedness.

Lagrange polynomials and Lebesgue constants can be defined for overdetermined and underdetermined interpolation [4]. Approximation theory definitions in their original form are ill-suited for an optimization audience. We simplify these definitions and also derive a convenient interpretation via the  $\Lambda$ -poisedness condition, where  $\Lambda$  is the bound of the maximum  $\ell_2$ -norm of the values of the Lagrange polynomials. We use this interpretation to connect Lebesgue constants (or their modified version,  $\Lambda$ ) of  $Y$  for regression, underdetermined interpolation, and full interpolation. Specifically, given an overdetermined interpolation problem one can select a subset of sample

points that defines a unique interpolating polynomial of a given degree. For example, one can select the subset which has the best Lebesgue constant. We show how this Lebesgue constant relates to the Lebesgue constant of the whole sample set with respect to regression. We also discuss (and show with the help of some basic computational examples) that regression based on the whole sample set often provides a more accurate model than the interpolation model based on the subset with the smallest Lebesgue constant. To fully show the benefit of regression models, however, a separate computational study would be required.

In the case of underdetermined interpolation one might restrict the space of polynomials to a subspace of an appropriate dimension and construct a uniquely defined interpolating polynomial in that subspace. The subspace can be selected to minimize the appropriate Lebesgue constant. We compare such interpolation to the minimum-norm solution for the underdetermined problem. We show how the Lebesgue constant with respect to interpolation in the ‘best’ subspace relates to the Lebesgue constant with respect to minimum-norm interpolation. As in the case of regression, we argue, and support by simple computational evidence, that minimum-norm interpolation is often better than other alternatives. The minimum-norm interpolation case does not require additional computational evidence, however, since it has shown its good performance in, for example, [1] and [17].

Finally, as in [8] we connect  $\Lambda$ -poisedness, and thus, the Lebesgue constants, to the size of the pivots of the appropriate Vandermonde matrices to which the algorithms of [8] can be applied. We also provide the Taylor error bounds in a simple form for the case of overdetermined interpolation that can be readily used in the convergence theory of optimization algorithms (see, for example, [10]).

The paper is organized as follows. In Section 2, we present the building blocks for polynomial regression, introducing Lagrange polynomials, the Lebesgue constant and  $\Lambda$ -poisedness and showing all the corresponding algebraic and geometrical properties, analogously to the case of fully determined interpolation. The error bounds for regression are stated in Section 3. A few issues concerning the usefulness of regression models and their relation to interpolation models are addressed in Section 4. Section 5 covers the underdetermined case. We end the paper in Section 6 stating some concluding remarks and the perspectives that are opened by our results.

**1.1. Basic facts and notation.** Here we introduce some notation and also state some facts from linear algebra that will be used in the paper.

By  $\|\cdot\|_k$ , with  $k \geq 1$ , we denote the standard  $\ell_k$  vector norm or the corresponding matrix norm. By  $\|\cdot\|$  (without the subscript) we denote the  $\ell_2$ -norm. We use  $B(\Delta) = \{x \in \mathbb{R}^m : \|x\| \leq \Delta\}$  to denote the closed ball in  $\mathbb{R}^m$  of radius  $\Delta > 0$  centered at the origin (where  $m$  is inferred from the particular context). We use several properties of norms. In particular, given a  $m \times n$  matrix  $A$ , we use the facts

$$\|A\|_2 \leq m^{\frac{1}{2}} \|A\|_\infty, \quad \|A\|_F \leq n^{\frac{1}{2}} \|A\|_2, \quad \|A\|_2 = \|A^\top\|_2.$$

We will use the standard ‘big-O’ notation written as  $\mathcal{O}(\cdot)$  to say, for instance, that if for two scalar or vector functions  $\beta(x)$  and  $\alpha(x)$  one has  $\beta(x) = \mathcal{O}(\alpha(x))$  then there exists a constant  $C > 0$  such that  $\|\beta(x)\| \leq C\|\alpha(x)\|$  for all  $x$  in its domain.

By the *natural basis* of the space of polynomials of degree at most  $d$  in  $\mathbb{R}^n$ , we will mean the following basis of monomial functions

$$\{1, x_1, x_2, \dots, x_n, x_1^2/2, x_1x_2, \dots, x_{n-1}^{d-1}x_n/(d-1)!, x_n^d/d!\}.$$

Given a matrix  $M \in \mathbb{R}^{\ell \times k}$ , such that  $\ell > k$ , we will use  $M = U\Sigma V^\top$  to denote the reduced singular value decomposition, where  $\Sigma$  is a diagonal  $k \times k$  matrix formed by the singular values. The columns of the matrix  $U \in \mathbb{R}^{\ell \times k}$  are orthonormal and form the left singular vectors of  $M$ . The matrix  $V \in \mathbb{R}^{k \times k}$  is orthogonal and its columns are the right singular vectors of  $M$ . If  $M$  has full column rank then  $\Sigma$  is invertible. Analogously, if  $k > \ell$  then the reduced singular value decomposition  $M = U\Sigma V^\top$  is such that  $\Sigma$  is a diagonal  $\ell \times \ell$  matrix,  $U$  is an  $\ell \times \ell$  orthogonal matrix, and  $V$  is a  $k \times \ell$  matrix with orthonormal columns.

We present here a lemma that will be useful later in the paper.

LEMMA 1.1. *Consider a set  $Z = \{z^1, \dots, z^m\} \subset \mathbb{R}^n$ , with  $m > n$ . Let  $I \subset \{1, \dots, m\}$  be a subset of indices with  $|I| = n$ . It is possible to choose  $I$  so that for any  $x \in \mathbb{R}^n$  such that*

$$x = \sum_{i=1}^m \lambda_i z^i, \quad |\lambda_i| \leq \Lambda,$$

for some  $\Lambda > 0$ , we can write

$$x = \sum_{i \in I} \gamma_i z^i, \quad |\gamma_i| \leq (m - n + 1)\Lambda.$$

*Proof.* Consider an  $n \times n$  matrix  $A$  whose columns are the vectors  $z^i$ ,  $i \in I$ . Among all possible sets  $I$ , choose the one that corresponds to the matrix  $A$  with the largest absolute value of the determinant. We will show that this  $I$  satisfies the statement of the lemma.

Let  $\bar{I} = \{1, \dots, m\} \setminus I$  and let  $Z_{\bar{I}}$  be the subset of  $Z$  containing those points whose indices are in  $\bar{I}$ . First, we will show that for any  $z^j$ ,  $j \in \bar{I}$ ,

$$z^j = \sum_{i \in I} \alpha_i^j z^i, \quad |\alpha_i^j| \leq 1.$$

By Cramer's rule  $\alpha_i^j = \det(A_{z^j, i}) / \det(A)$ , where  $A_{z^j, i}$  corresponds to the matrix  $A$  with its  $i$ -th column replaced by the vector  $z^j$ . Since by the selection of  $I$ ,  $|\det(A)| \geq |\det(A_{z^j, i})|$  for any  $j \in \bar{I}$ ,  $|\alpha_i^j| \leq 1$ .

Now consider any  $x$  such that

$$x = \sum_{i=1}^m \lambda_i z^i, \quad |\lambda_i| \leq \Lambda.$$

We have

$$x = \sum_{i \in I} \lambda_i z^i + \sum_{j \in \bar{I}} \lambda_j \left( \sum_{i \in I} \alpha_i^j z^i \right) = \sum_{i \in I} \gamma_i z^i, \quad |\gamma_i| \leq (m - n + 1)\Lambda, \quad i \in I.$$

□

**2. Polynomial least-squares regression and poisedness.** Let us consider  $\mathcal{P}$ , the space of polynomials of degree  $\leq d$  in  $\mathbb{R}^n$ . Let  $q_1 = q + 1$  be the dimension of this space (e.g., for  $d = 1$ ,  $q_1 = n + 1$  and for  $d = 2$ ,  $q_1 = (n + 1)(n + 2)/2$ ) and let  $\phi = \{\phi_0(x), \phi_1(x), \dots, \phi_q(x)\}$  be a basis for  $\mathcal{P}$ . This means that  $\phi$  is a set of  $q_1$  polynomials of degree  $\leq d$  that span  $\mathcal{P}$ . Given a polynomial basis  $\phi$ , let

$\phi(x) = [\phi_0(x), \phi_1(x), \dots, \phi_q(x)]^\top$  be a vector in  $\mathbb{R}^{q_1}$  whose entries are the values of the elements of the polynomial basis at  $x$  (one can view  $\phi(x)$  as a mapping from  $\mathbb{R}^n$  to  $\mathbb{R}^{q_1}$ ).

Assume we are given a set  $Y = \{y^0, y^1, \dots, y^p\} \subset \mathbb{R}^n$  of  $p_1 = p + 1$  sample points. Let  $m(x)$  denote the polynomial of degree  $\leq d$  that approximates a given function  $f(x)$  at the points in  $Y$  via least-squares regression. We assume that the number of points satisfies  $p_1 > q_1$  (in other words that  $p > q$ ). Since  $\phi$  is a basis in  $\mathcal{P}$ ,  $m(x) = \sum_{k=0}^q \alpha_k \phi_k(x)$ , where the  $\alpha_k$ 's are the unknown coefficients. By determining the coefficients  $\alpha = [\alpha_0, \dots, \alpha_q]^\top$  we determine the polynomial  $m(x)$ . The coefficients  $\alpha$  can be determined from the least-squares regression conditions

$$m(y^i) = \sum_{k=0}^q \alpha_k \phi_k(y^i) \stackrel{\text{l.s.}}{=} f(y^i), \quad i = 0, \dots, p.$$

This problem is a linear least-squares problem in terms of  $\alpha$ . The above system has a unique solution in the least-squares sense if the matrix of the system

$$(2.1) \quad M(\phi, Y) = \begin{bmatrix} \phi_0(y^0) & \phi_1(y^0) & \cdots & \phi_q(y^0) \\ \phi_0(y^1) & \phi_1(y^1) & \cdots & \phi_q(y^1) \\ \vdots & \vdots & \vdots & \vdots \\ \phi_0(y^p) & \phi_1(y^p) & \cdots & \phi_q(y^p) \end{bmatrix}$$

has full column rank.

It is easy to see that if  $M(\phi, Y)$  is square and nonsingular, then the above problem becomes an interpolation problem. In that case, the set  $Y$  is said to be *poised* (or *d-unisolvent* [4]). Just as for interpolation, if  $M(\phi, Y)$  has full column rank for some choice of  $\phi$  then this is the case for any basis of  $\mathcal{P}$ . Hence, we will call a set  $Y$  poised with respect to polynomial least-squares regression if the appropriate  $M(\phi, Y)$  has full column rank for some choice of the basis  $\phi$ .

We now show that in the full column rank case the least-squares regression polynomial does not depend on the choice of the basis  $\phi$ . Since the set is poised, it is obvious that the least-squares regression polynomial  $m(x)$  exists and is unique for a given choice of basis. Consider now two different bases  $\psi(x)$  and  $\phi(x)$  related by  $\psi(x) = P^\top \phi(x)$ , where  $P$  is  $q_1 \times q_1$  and nonsingular. Then,  $M(\psi, Y) = M(\phi, Y)P$ . Let  $\alpha_\phi$  (resp.  $\alpha_\psi$ ) be the vector of coefficients of the least-squares regression polynomial for the basis  $\phi(x)$  (resp.  $\psi(x)$ ). Since  $\alpha_\phi$  is the least-squares solution to the system  $M(\phi, Y)\alpha_\phi = f_Y$  then

$$\begin{aligned} \alpha_\phi &= [M(\phi, Y)^\top M(\phi, Y)]^{-1} M(\phi, Y)^\top f_Y \\ &= [P^{-\top} M(\psi, Y)^\top M(\psi, Y) P^{-1}]^{-1} P^{-\top} M(\psi, Y)^\top f_Y \\ &= P [M(\psi, Y)^\top M(\psi, Y)]^{-1} M(\psi, Y)^\top f_Y = P \alpha_\psi. \end{aligned}$$

The last equality follows from the fact that  $\alpha_\psi$  is the least-squares solution to the system  $M(\psi, Y)\alpha_\psi = f_Y$ . Then, for any  $x$ ,

$$\alpha_\psi^\top \psi(x) = \alpha_\psi^\top P^\top \phi(x) = \alpha_\phi^\top \phi(x).$$

We have shown that if  $Y$  is poised, then the least-squares regression polynomial is unique and independent of the choice of  $\phi$ .

The condition of poisedness and the existence of the regression polynomial is not sufficient in practical algorithms or in the derivation of error bounds. One needs a condition of ‘sufficient’ poisedness, which we will refer to as ‘well poisedness’, characterized by a constant. This constant should be an indicator of how well the regression polynomial approximates the true function. In [8], we considered such constants for the case of polynomial interpolation. In this paper, we will extend the concepts and the results to the case of least-squares regression.

Since the column linear independence of  $M(\phi, Y)$  reflects the poisedness of the set  $Y$ , it is natural to consider some condition number related to  $M(\phi, Y)$  as a constant characterizing the well poisedness of  $Y$ . However, the singular values of  $M(\phi, Y)$  depend on the choice of  $\phi$  and, moreover, for any given poised interpolation set  $Y$ , one can choose the basis  $\phi$  so that the ratio of the largest over the smallest singular values of  $M(\phi, Y)$  can equal anything between 1 and  $\infty$ .

The most commonly used measure of poisedness in the multivariate polynomial interpolation literature is the Lebesgue constant and is related to the basis of Lagrange polynomials. We begin by briefly describing the concept and its use in polynomial interpolation.

**DEFINITION 2.1.** *Given a set of interpolation points  $Y = \{0, y^1, \dots, y^p\}$ , with  $p = q$ , where  $q + 1$  is the dimension of the space of polynomials of degree  $\leq d$ , a basis of  $p_1 = p + 1$  polynomials  $\mathcal{L}_j(x)$ ,  $j = 0, \dots, p$ , of degree  $\leq d$ , is called a basis of Lagrange polynomials if*

$$\mathcal{L}_j(y^i) = \delta_{ij} = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j. \end{cases}$$

For any poised set  $Y$  there exists a unique basis of Lagrange polynomials. A measure of poisedness of  $Y$  is given by an upper bound on the absolute value of the Lagrange polynomials in the region of interest. In [4, Theorem 1], it is shown that for any  $x$  in the convex hull of  $Y$

$$(2.2) \quad \|\mathcal{D}^k m(x) - \mathcal{D}^k f(x)\| \leq \frac{1}{(d+1)!} G \sum_{i=0}^p \|y^i - x\|^{d+1} \|\mathcal{D}^k \mathcal{L}_i(x)\|,$$

where  $\mathcal{D}^k$  denotes the  $k$ -th derivative of a function and  $G$  is an upper bound on  $\mathcal{D}^{d+1} f(x)$ . This inequality is an equivalent of the Taylor bound for multivariate polynomial interpolation. The Lebesgue constant of  $Y$  is defined as

$$\Lambda_Y = \max_{x \in B_Y(\Delta)} \sum_{i=0}^p |\mathcal{L}_i(x)|,$$

where  $B_Y(\Delta)$  is, for instance, a ball of radius  $\Delta$  containing  $Y$ . The Taylor bound for function value approximation can be simplified as ( $k = 0$ ):

$$(2.3) \quad |m(x) - f(x)| \leq \frac{1}{(d+1)!} G \Lambda_Y \Delta^{d+1}.$$

See [15] for a simple derivation of this bound. For further discussion see also [8].

We will present the extension of the definition of Lagrange polynomials to the cases of polynomial least-squares regression (in this section) and to the case of mini-

mum-norm underdetermined interpolation (in Section 5). As we have already mentioned in the introduction, generalized definitions of Lagrange polynomials exist and are used in approximation theory literature (see [4] and references therein for the multivariate case and [13] for the univariate case). However, these definitions involve concepts and structures which are used primarily in approximation theory and it is nontrivial to adapt these definitions to the optimization context. Here we present definitions in simple terms and provide straightforward proofs that support the validity of these extended definitions.

One of the most notable properties of Lagrange polynomials in the case of interpolation is that the interpolating polynomial  $m(x)$  has a simple representation in terms of them, given by

$$m(x) = \sum_{i=0}^p f(y^i) \mathcal{L}_i(x),$$

where  $f(y^i)$ ,  $i = 0, \dots, p$ , are the values that are interpolated. We will see that the same is true in the regression case.

**2.1. Lagrange polynomials for regression.** Let  $Y = \{y^0, y^1, \dots, y^p\}$  be the interpolation set, and  $\phi = \{\phi_0(x), \phi_1(x), \dots, \phi_q(x)\}$  be the basis of polynomials of a given degree. We are considering the case where  $p > q$  (i.e., more points than basis polynomials).

**DEFINITION 2.2.** *Given a set of sample points  $Y = \{y^0, y^1, \dots, y^p\}$ , with  $p > q$ , where  $q + 1$  is the dimension of the space of polynomials of degree  $\leq d$ , a set of  $p_1 = p + 1$  polynomials  $\mathcal{L}_j(x)$ ,  $j = 0, \dots, p$ , of degree  $\leq d$ , is called a set of Lagrange regression polynomials if*

$$\mathcal{L}_j(y^i) \stackrel{\ell.s.}{=} \delta_{ij} = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j. \end{cases}$$

Clearly these polynomials are no longer linearly independent, since there are too many of them. However, as we show below, many other properties of Lagrange interpolation polynomials are preserved.

Assume that the set  $Y$  is poised. We can write the  $j$ -th Lagrange polynomial as

$$\mathcal{L}_j(x) = \phi(x)^\top \Theta_\phi^j,$$

where  $\Theta_\phi^j$  is a vector with  $q_1$  components. Consequently the Lagrange interpolation conditions (in the least-squares sense) can be written as

$$M(\phi, Y) \Theta_\phi^j \stackrel{\ell.s.}{=} e_{j+1}, \quad j = 0, \dots, p,$$

where  $e_{j+1}$  is the  $(j + 1)$ -th column of the identity matrix of order  $p_1$ . In matrix notation, we have that

$$M(\phi, Y) \Theta_\phi \stackrel{\ell.s.}{=} I,$$

where  $\Theta$  is the matrix whose columns are  $\Theta_\phi^j$ ,  $j = 0, \dots, p$ . This is essentially an immediate consequence of Definition 2.2.

The set of Lagrange regression polynomials exists and is unique if the matrix  $M(\phi, Y)$  has full column rank. As we have shown above for any least-squares regression polynomial when  $Y$  is poised, the polynomials  $\mathcal{L}_j(x)$ ,  $j = 0, \dots, p$ , do not depend on the choice of  $\phi$ .

Let  $M(\phi, Y) = U_\phi \Sigma_\phi V_\phi^\top$  be the reduced singular value decomposition, defined in Section 1.1. We omit the dependence on  $Y$ , since we keep  $Y$  fixed in the discussion below. Thus  $\alpha = V_\phi \Sigma_\phi^{-1} U_\phi^\top$  or  $\alpha_j = V_\phi \Sigma_\phi^{-1} U_\phi^\top e_{j+1}$ , for  $j = 0, \dots, p$ .

We will now show that the regression polynomial  $m(x)$  can also be written as a linear combination of the Lagrange polynomials.

LEMMA 2.3. *Let  $Y = \{y^0, y^1, \dots, y^p\}$  be a poised set of sample points for the function  $f(x)$  and let  $m(x)$  be a polynomial of degree  $\leq d$  that approximates  $f(x)$  via least-squares regression on the points in  $Y$ . Let  $\{\mathcal{L}_j(x), j = 0, \dots, p\}$  be the set of Lagrange regression polynomials of degree  $\leq d$  given by Definition 2.2. Then*

$$m(x) = \sum_{i=0}^p f(y^i) \mathcal{L}_i(x).$$

*Proof.* It is true that  $m(x)$  can always be expressed as

$$m(x) = \sum_{i=0}^p \gamma_i \mathcal{L}_i(x).$$

Since  $\mathcal{L}$  has more elements than a basis, the solution  $\gamma$  is not unique. But all we need to show is that  $\gamma_i = f(y^i)$ ,  $i = 0, \dots, p$ , is one such solution.

We also know that  $V_\phi \Sigma_\phi^{-1} U_\phi^\top f_Y$  is the vector of coefficients that expresses  $m(x)$  in terms of the basis  $\phi$ . Thus,

$$m(x) = \left( V_\phi \Sigma_\phi^{-1} U_\phi^\top f_Y \right)^\top \phi(x) = f_Y^\top \left( V_\phi \Sigma_\phi^{-1} U_\phi^\top \right)^\top \phi(x) = f_Y^\top \mathcal{L}(x)$$

and we have proved what we wanted.  $\square$

REMARK 2.1. *It is interesting to note that the extension of Lagrange polynomials does not apply to the case of  $\ell_1$ -norm and  $\ell_\infty$ -norm regressions. The reason why the properties of the Lagrange polynomials extend to the case of least-squares regression is because any least-squares regression polynomial is a linear function of the right hand side  $f_Y$ . This situation is no longer the case when  $\ell_1$ -norm and  $\ell_\infty$ -norm regressions are considered.*

In Section 3 an error bound similar to (2.2) is extended to the regression case. The Lebesgue constant  $\max_{x \in B_Y(\Delta)} \sum_{i=0}^p |\mathcal{L}_i(x)|$ , where  $\mathcal{L}_i(x)$ ,  $i = 0, \dots, p$ , are the Lagrange polynomials defined for regression, is the constant that controls the quality of the error bound. To be able to use regression models in a derivative-free optimization algorithm we need to be able to control the Lebesgue constant of the sample set  $Y$ . As in [8] we find the following interpretation of Lagrange regression polynomials convenient for understanding the behavior of the Lebesgue constant.

**2.2. Geometric interpretations of Lagrange regression polynomials.** Given a poised set  $Y = \{y^0, y^1, \dots, y^p\} \subset B(1) \subset \mathbb{R}^n$ , with  $p > q$  and  $x \in B(1)$ , we can express the vector  $\phi(x)$  in terms of the vectors  $\phi(y^i)$ ,  $i = 0, \dots, p$ , as

$$(2.4) \quad \sum_{i=0}^p \lambda_i(x) \phi(y^i) = \phi(x),$$

or, equivalently,

$$M(\phi, Y)^\top \lambda(x) = \phi(x), \quad \text{where } \lambda(x) = [\lambda_0(x), \dots, \lambda_p(x)]^\top.$$

This system is a simple extension of a similar system introduced in [8] for the case of polynomial interpolation. Unlike the system in [8], this new system is under-determined, hence it has multiple solutions. In order to establish uniqueness, we will consider the minimum  $\ell_2$ -norm solution.

LEMMA 2.4. *Given a poised set  $Y$ , the functions  $\lambda_i(x)$ ,  $i = 0, \dots, p$ , defined as the minimum  $\ell_2$ -norm solution of (2.4), form the set of Lagrange regression polynomials for  $Y$  given by Definition 2.2.*

*Proof.* We want to show that  $\mathcal{L}(x) = \lambda(x)$ , where  $\lambda(x)$  is the minimum  $\ell_2$ -norm solution to (2.4). We know that  $\lambda(x)$  satisfies

$$M(\phi, Y)^\top \lambda(x) = \phi(x),$$

where  $M(\phi, Y)$  is defined by (2.1), and, in particular, that  $\lambda(x)$  is the minimum  $\ell_2$ -norm solution to this system. Hence, given the reduced singular value decomposition of  $M(\phi, Y)^\top = V_\phi \Sigma_\phi U_\phi^\top$ , we have

$$\lambda(x) = U_\phi \Sigma_\phi^{-1} V_\phi^\top \phi(x) = \mathcal{L}(x).$$

Note that we have proved this result independently of the choice of  $\phi$ .  $\square$

A simple corollary of this result is that  $\lambda(x) = [\lambda_0(x), \dots, \lambda_p(x)]^\top$  does not depend on the choice of  $\phi$ . In [8], the well-poisedness condition for interpolation was introduced via a bound on  $\lambda(x)$  and was referred to as  $\Lambda$ -poisedness.

DEFINITION 2.5. *Let  $\Lambda > 0$  be given. Let  $\phi = \{\phi_0(x), \phi_1(x), \dots, \phi_p(x)\}$  be a basis in  $\mathcal{P}$ .*

*A set  $Y = \{y^0, y^1, \dots, y^p\}$ , with  $p = q$ , is said to be  $\Lambda$ -poised in  $B(1)$  (in an interpolation sense) if and only if for any  $x \in B(1)$  there exists a  $\lambda(x) \in \mathbb{R}^{p+1}$  such that*

$$\sum_{i=0}^p \lambda_i(x) \phi(y^i) = \phi(x) \quad \text{with} \quad \|\lambda(x)\| \leq \Lambda.$$

Clearly this definition is equivalent to having all Lagrange polynomials bounded by  $\Lambda$  in  $B(1)$  in the  $\ell_2$ -norm. We now introduce the analogous definition for a well-poised regression set.

DEFINITION 2.6. *Let  $\Lambda > 0$  be given. Let  $\phi = \{\phi_0(x), \phi_1(x), \dots, \phi_q(x)\}$  be a basis in  $\mathcal{P}$ .*

*A set  $Y = \{y^0, y^1, \dots, y^p\}$ , with  $p > q$ , is said to be  $\Lambda$ -poised in  $B(1)$  in a regression sense if and only if for any  $x \in B(1)$  there exists a  $\lambda(x) \in \mathbb{R}^{p+1}$  such that*

$$\sum_{i=0}^p \lambda_i(x) \phi(y^i) = \phi(x) \quad \text{with} \quad \|\lambda(x)\| \leq \Lambda.$$

Note that the difference between the two definitions is that, in the regression case,  $\lambda(x)$  may not be unique for every  $x$ . In fact, we are interested in the minimum-norm solution for  $\lambda(x)$  and the bound on its norm. It is sufficient to say that  $\|\lambda(x)\| \leq \Lambda$  for *some* solution  $\lambda(x)$ , because then, clearly, the same is true for the minimum-norm solution.

One can relate  $\Lambda$ -poisedness in the regression sense to  $\Lambda$ -poisedness in the interpolation sense, as it is shown in the next theorem.

**THEOREM 2.7.** *Given a set  $Y = \{y^0, y^1, \dots, y^p\}$ , with  $p > q$ , which is  $\Lambda$ -poised in the regression sense, there is a subset of  $q_1 = q+1$  points in  $Y$  which is  $(p-q+1)\sqrt{q_1}\Lambda$ -poised in the interpolation sense.*

*Conversely, if any subset of  $q_1$  points in  $Y$  is  $\Lambda$ -poised in the interpolation sense, then the set  $Y = \{y^0, y^1, \dots, y^p\}$ , is  $\Lambda$ -poised in the regression sense.*

*Proof.* The first implication follows from the Definitions 2.5 and 2.6 for  $\Lambda$ -poisedness in the interpolation and regression senses and from Lemma 1.1, with  $m = p_1$  and  $n = q_1$ . The second implication is immediate from the same definitions.  $\square$

Notice that, by definition,  $\Lambda$  is an *upper bound* on poisedness; that is, if  $Y$  is  $\bar{\Lambda}$ -poised, then it is also  $\Lambda$ -poised, for all  $\Lambda \geq \bar{\Lambda}$ , in other words,  $Y$  is *at least*  $\bar{\Lambda}$ -poised. Since  $\Lambda$  is an upper bound on the  $\ell_2$ -norm of the vector of values of Lagrange polynomials and the Lebesgue constant is the maximum  $\ell_1$ -norm of that vector, then clearly one can express the upper bound on the Lebesgue constant through  $\Lambda$  and vice versa. We will use the Lebesgue constant and  $\Lambda$  to mean the same thing, whenever possible, even though usually they do not have the same value.

Notice also that  $\phi(x)$  is in the definition of  $\Lambda$ -poisedness, yet, clearly,  $\Lambda$ -poisedness does not depend on the choice of  $\phi(x)$ , since the  $\lambda_i(x)$  are simply the Lagrange polynomials. One can view it as a definition defect, but we find it useful, since it enables us to substitute any choice of  $\phi(x)$  in (2.4) and use this expression together with the properties of the chosen  $\phi(x)$  and the bound on  $\lambda(x)$ . This is precisely what we do in the next subsection to show the relation between  $\Lambda$ -poisedness and the condition number of  $M(\phi, Y)$  when  $\phi(x)$  is the natural basis.

**2.3.  $\Lambda$ -poisedness and the condition number of  $M(\phi, Y)$ .** In [8] we provide algorithms for checking and improving the poisedness (or Lebesgue) constant of a sample set. This is done by factorizing the matrix  $M(\phi, Y)$  with  $\phi$  being the natural basis, and bounding the absolute value of the pivots away from zero during the process. The same algorithm can be applied in the regression case with the exception that the matrix that is factorized has more rows than columns. We will now show that the condition number of  $M(\phi, Y)$ , for the natural basis  $\phi$  is proportional to the Lebesgue constant of  $Y$ . Hence any algorithm that maintains and improves this condition number will also maintain and decrease  $\Lambda$ -poisedness of  $Y$ . In particular the algorithms in [8] can serve that purpose. However, other algorithms that maintain the condition number of  $M(\phi, Y)$  by looking at all of its rows might be better suited for the use of regression models.

For the remainder of the paper we will assume that the smallest enclosing ball containing  $Y$  is centered at the origin. This assumption can be made without loss of generality, since it can always be satisfied by a shift of coordinates. Furthermore, we will consider the sample set  $Y$  scaled in such a way that the smallest ball enclosing  $Y$  has radius 1. Recall that we denote such a ball by  $B(1)$ . Hence, we assume that one

of the elements of  $Y$  has norm 1. At the end of this subsection we will show how to incorporate the scaling into our results.

We will now show how  $\Lambda$ -poisedness in the regression sense relates to the condition number of the following matrix

$$(2.5) \bar{M} = \begin{bmatrix} 1 & y_1^0 & \cdots & y_n^0 & \frac{1}{2}(y_1^0)^2 & y_1^0 y_2^0 & \cdots & \frac{1}{(d-1)!}(y_{n-1}^0)^{d-1} y_n^0 & \frac{1}{d!}(y_n^0)^d \\ 1 & y_1^1 & \cdots & y_n^1 & \frac{1}{2}(y_1^1)^2 & y_1^1 y_2^1 & \cdots & \frac{1}{(d-1)!}(y_{n-1}^1)^{d-1} y_n^1 & \frac{1}{d!}(y_n^1)^d \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots \\ 1 & y_1^p & \cdots & y_n^p & \frac{1}{2}(y_1^p)^2 & y_1^p y_2^p & \cdots & \frac{1}{(d-1)!}(y_{n-1}^p)^{d-1} y_n^p & \frac{1}{d!}(y_n^p)^d \end{bmatrix}.$$

This is the same as  $\bar{M} = M(\bar{\phi}, Y)$ , where

$$(2.6) \quad \bar{\phi} = \{1, x_1, x_2, \dots, x_n, x_1^2/2, x_1 x_2, \dots, x_{n-1}^{d-1} x_n / (d-1)!, x_n^d / d!\}$$

is what we are referring to as the natural basis of monomials. Substituting  $\bar{\phi}$  in the definition of  $\Lambda$ -poisedness we can write

$$(2.7) \quad \bar{M}^\top \lambda(x) = \bar{\phi}(x) \quad \text{with} \quad \|\lambda(x)\| \leq \Lambda.$$

Also, since  $x \in B(1)$  and since at least one of the  $y^i$ 's has norm 1 (recall that  $B(1)$  is the smallest enclosing ball centered at the origin), then the norm of this matrix is always bounded by

$$(2.8) \quad 1 \leq \|\bar{M}\| \leq p_1^{\frac{1}{2}} q_1.$$

Let us consider the reduced SVD of  $\bar{M} = U\Sigma V^\top$ , and let  $\sigma_1$  (resp.  $\sigma_{q_1}$ ) denote the absolute value of the largest (resp. smallest) singular value of  $\bar{M}$ . We omit the dependence of  $U$ ,  $\Sigma$ , and  $V$  on  $Y$  for simplicity of the presentation. Then  $\|\bar{M}\| = \|\Sigma\| = \sigma_1$  and  $\|\Sigma^{-1}\| = 1/\sigma_{q_1}$ . The condition number of  $\bar{M}$  is denoted by  $\kappa(\bar{M}) = \sigma_1/\sigma_{q_1}$ . To bound  $\kappa(\bar{M})$  in terms of  $\Lambda$  it is, then, sufficient to bound  $\|\Sigma^{-1}\|$ . Conversely, to bound  $\Lambda$  in terms of  $\kappa(\bar{M})$  it is sufficient to bound it in terms of  $\|\Sigma^{-1}\|$ .

In [8] we showed that the well-poisedness constant  $\Lambda$  from Definition 2.5 and the condition number of  $\bar{M}$  (which is a square matrix in the case of interpolation) differ by a constant factor. The following theorem is an analog of Theorem 3.3 in [8].

**THEOREM 2.8.** *If  $\Sigma$  is nonsingular and  $\|\Sigma^{-1}\| \leq \Lambda$ , then the set  $Y$  is  $\sqrt{q_1}\Lambda$ -poised (according to Definition 2.6) in the unit ball  $B(1)$  centered at 0. Conversely, if the set  $Y$  is  $\Lambda$ -poised, according to Definition 2.6, in the unit ball  $B(1)$  centered at 0, then  $\Sigma$  is nonsingular and satisfies*

$$(2.9) \quad \|\Sigma^{-1}\| \leq \theta \Lambda,$$

where  $\theta > 0$  is dependent on  $n$  and  $d$  but independent of  $Y$  and  $\Lambda$ .

*Proof.* The proof is very similar to the proof in [8, Theorem 3.3], but there are a few extra steps. We include the proof for the sake of completeness.

If  $\Sigma$  is nonsingular and  $\|\Sigma^{-1}\| \leq \Lambda$  then the minimum-norm solution satisfies

$$\|\lambda(x)\| \leq \|U\Sigma^{-1}V^\top\| \|\phi(x)\| \leq q_1^{\frac{1}{2}} \|\Sigma^{-1}\| \|\phi(x)\|_\infty \leq q_1^{\frac{1}{2}} \Lambda$$

(we used the facts that  $\|\phi(x)\| \leq \sqrt{q_1} \|\phi(x)\|_\infty$  and  $\max_{x \in B(1)} \|\phi(x)\|_\infty \leq 1$ ).

Proving the other relation is more complicated. First let us show by contradiction that the matrix  $\Sigma$  is nonsingular. Let us assume it is singular. By the definition of  $\Lambda$ -poisedness, for any  $x \in B(1)$ ,  $\bar{\phi}(x)$  lies in the range space of  $\bar{M}^\top$ . This means that there exists a vector  $v \neq 0$  in the null space of  $\bar{M}$  such that for any  $x \in B(1)$  we get  $\bar{\phi}(x)^\top v = 0$ . Hence,  $\bar{\phi}(x)^\top v$  is a polynomial in  $x$  which is identically zero on a unit ball, which implies that all coefficients of this polynomial are zero, i.e.,  $v = 0$ . We have arrived at a contradiction.

Now we want to show that there exists a constant  $\theta > 0$ , independent of  $Y$  and of  $\Lambda$ , such that  $\|\Sigma^{-1}\| \leq \theta\Lambda$ . From the definition of the matrix norm, and from the fact that  $V$  has orthonormal columns

$$(2.10) \quad \|\Sigma^{-1}\| = \|\Sigma^{-1}V^\top\| = \max_{\|v\|=1} \|\Sigma^{-1}V^\top v\|,$$

and we can consider a vector  $\bar{v}$  at which the maximum is attained

$$(2.11) \quad \|\Sigma^{-1}V^\top\| = \|\Sigma^{-1}V^\top\bar{v}\|, \quad \|\bar{v}\| = 1.$$

Let us assume first that there exists an  $x \in B(1)$  such that  $\bar{\phi}(x) = \bar{v}$ . Then from the fact that  $Y$  is  $\Lambda$ -poised we have that

$$\|\Sigma^{-1}V^\top\bar{v}\| = \|U\Sigma^{-1}V^\top\bar{\phi}(x)\| \leq \Lambda,$$

and from (2.10) and (2.11) the statement of the theorem holds with  $\theta = 1$ .

Notice that  $\bar{v}$  does not necessarily belong to the image of  $\bar{\phi}(x)$ , which means that there might not be any  $x \in B(1)$  such that  $\bar{\phi}(x) = \bar{v}$ , and hence we have that  $\|\Sigma^{-1}V^\top\bar{v}\| \neq \|\Sigma^{-1}V^\top\bar{\phi}(x)\|$ . However, we will show that there exists a constant  $\theta > 0$  such that for any  $\bar{v}$  which satisfies (2.11) there exists an  $x \in B(1)$ , such that

$$(2.12) \quad \frac{\|\Sigma^{-1}V^\top\bar{v}\|}{\|\Sigma^{-1}V^\top\bar{\phi}(x)\|} \leq \theta.$$

Once we have shown that such constant  $\theta$  exists the result of the lemma follows from the definition of  $\bar{v}$ .

To show that (2.12) holds, we first show that there exists  $\gamma > 0$  such that for any  $\bar{v}$  with  $\|\bar{v}\| = 1$ , there exists an  $\bar{x} \in B(1)$  such that  $|\bar{v}^\top\bar{\phi}(\bar{x})| \geq \gamma$ . Consider

$$\psi(v) = \max_{x \in B(1)} |v^\top\bar{\phi}(x)|.$$

It is easy to show that  $\psi(v)$  is a norm in the space of vectors  $v$ . Since the ratio of any two norms in finite dimensional spaces can be uniformly bounded by a constant, there exists a (maximal)  $\gamma > 0$  such that  $\psi(\bar{v}) \geq \gamma\|\bar{v}\| = \gamma$ . Hence, there exists an  $\bar{x} \in B(1)$  such that  $|\bar{v}^\top\bar{\phi}(\bar{x})| \geq \gamma$ .

Let  $\bar{v}^\perp$  be the orthogonal projection of  $\bar{\phi}(\bar{x})$  onto the subspace orthogonal to  $\bar{v}$ . Now, notice that from the definition (2.11) of  $\bar{v}$ , it follows that  $\bar{v}$  is the right singular vector corresponding to the largest singular value of  $\Sigma^{-1}V^\top$ , i.e.,  $\bar{v}$  is equal to one of the columns of  $V$ . Then  $\Sigma^{-1}V^\top\bar{v}$  and  $\Sigma^{-1}V^\top\bar{v}^\perp$  are orthogonal vectors (since  $\Sigma^{-1}V^\top\bar{v}$  is a multiple of a column of an identity matrix and  $\Sigma^{-1}V^\top\bar{v}^\perp$  is a vector orthogonal to that column of the identity). Since  $\|\bar{v}\| = 1$ ,  $\bar{\phi}(\bar{x}) = \bar{v}^\perp + (\bar{v}^\top\bar{\phi}(\bar{x}))\bar{v}$ . Also, from the orthogonality of  $\Sigma^{-1}V^\top\bar{v}^\perp$  and  $\Sigma^{-1}V^\top\bar{v}$

$$\|\Sigma^{-1}V^\top\bar{\phi}(\bar{x})\| = \|\Sigma^{-1}V^\top\bar{v}^\perp\| + |\bar{v}^\top\bar{\phi}(\bar{x})|\|\Sigma^{-1}V^\top\bar{v}\|.$$

Hence  $\|\Sigma^{-1}V^\top \bar{\phi}(\bar{x})\| \geq |\bar{v}^\top \bar{\phi}(\bar{x})| \|\Sigma^{-1}V^\top \bar{v}\|$ . It follows from  $|\bar{v}^\top \bar{\phi}(\bar{x})| \geq \gamma$  that

$$\|\Sigma^{-1}V^\top \bar{\phi}(\bar{x})\| \geq \gamma \|\Sigma^{-1}V^\top \bar{v}\|,$$

Assigning  $\theta = 1/\gamma$  shows (2.12), concluding the proof of the bound on the norm of  $\Sigma^{-1}$ .  $\square$

The constant  $\theta$  can be estimated for specific values of  $d$ . As we have pointed out in [8], we have  $\theta \leq 1$  when  $d = 1$ . For  $d = 2$  the following lemma holds.

**LEMMA 2.9.** *Let  $\bar{v}^\top \bar{\phi}(x)$  be a quadratic polynomial with  $\bar{\phi}(x)$  defined by (2.6) and  $\|\bar{v}\|_\infty = 1$ , and let  $B(1)$  be a (closed) ball of radius 1 centered at the origin. Then*

$$\max_{x \in B(1)} |\bar{v}^\top \bar{\phi}(x)| \geq \frac{1}{4}.$$

For the proof of the lemma and further discussion see [8, Lemma 3.4].

We can replace the constant  $\theta$  of Theorem 2.8 by an upper bound, which is easily derived for the quadratic case. Recall that  $\theta = 1/\gamma$ , where

$$\gamma = \min_{\|\bar{v}\|=1} \max_{x \in B(1)} |\bar{v}^\top \bar{\phi}(x)|.$$

Given any  $\bar{v}$  such that  $\|\bar{v}\| = 1$ , we can scale  $\bar{v}$  by at most  $\sqrt{q_1}$  to  $\hat{v} = \alpha \bar{v}$ ,  $0 < \alpha \leq \sqrt{q_1}$ , such that  $\|\hat{v}\|_\infty = 1$ . Then

$$\gamma = \min_{\|\bar{v}\|=1} \max_{x \in B(1)} |\bar{v}^\top \bar{\phi}(x)| \geq \frac{1}{q_1^{\frac{1}{2}}} \min_{\|\hat{v}\|_\infty=1} \max_{x \in B(1)} |\hat{v}^\top \bar{\phi}(x)| \geq \frac{1}{4q_1^{\frac{1}{2}}}.$$

The last inequality is due to Lemma 2.9 applied to the polynomials of the form  $\hat{v}^\top \bar{\phi}(x)$ . Hence we have

$$(2.13) \quad \theta \leq 4q_1^{\frac{1}{2}}.$$

Specifying the bound on  $\theta$  for polynomials of degree higher than two is also possible. This bound will grow rapidly (most likely, exponentially) with the degree of the polynomial and its usefulness for higher degree interpolation and regression is unclear. But in our context we are motivated by optimization frameworks that rarely use polynomials of degree higher than 2.

We will now consider the set  $Y$  with arbitrary scaling. An attractive property of Lagrange polynomials is that they remain invariant under the scaling of the set  $Y$ . A simple proof can be derived from our interpretation of Lagrange polynomials given in the definition of  $\Lambda$ -poisedness.

**LEMMA 2.10.** *Let  $Y = \{y^0, y^1, \dots, y^p\}$  be an interpolation set and  $\{\lambda_i(x), i = 0, \dots, p\}$  be the set of Lagrange polynomials associated with  $Y$ . Then  $\{\lambda_i(\Delta x), i = 0, \dots, p\}$  is the set of Lagrange polynomials associated with  $\hat{Y}$ , where  $\hat{Y} = \{\Delta y^0, \Delta y^1, \dots, \Delta y^{p-1}\}$  for any  $\Delta > 0$ .*

*Proof.* From Lemma 2.4 we know that  $\lambda_i(x)$ ,  $i = 0, \dots, p$ , satisfy

$$\sum_{i=0}^p \bar{\phi}(y^i) \lambda_i(x) = \bar{\phi}(x),$$

where  $\bar{\phi}$  is the basis of monomials. If we scale each  $y^i$  and  $x$  by  $\Delta$ , this corresponds to scaling the above equations by different scalars ( $1, \Delta, \Delta^2$ , etc.). Clearly,  $\lambda(\Delta x)$  satisfies the scaled system of equations. That implies, again due to Lemma 2.4, that  $\lambda_i(\Delta x)$ ,  $i = 0, \dots, p$ , is the set of Lagrange polynomials associated with the scaled set.  $\square$

On the contrary, the norm of the inverse of  $\bar{M}$  and therefore the condition number  $\sigma(\bar{M})$  depends on the scaling of the interpolation set. When we multiply the set  $Y$  by  $\Delta$ , the columns of  $\bar{M}$  get multiplied by different scalars ( $1, \Delta, \Delta^2$ , etc.). So, the scaled matrix, say  $\hat{M}$ , is such that  $\|\hat{M}^{-1}\|, \kappa(\hat{M}) \rightarrow \infty$  when  $\Delta \rightarrow 0$ . To eliminate this effect we will scale a given set  $Y \subset B(\Delta)$  by  $1/\Delta$  to obtain  $\bar{Y} \subset B(1)$ . The condition number of the corresponding matrix  $\bar{M}$  is then suitable as a measure of well-poisedness of  $Y$ , since it is within a constant factor of the well-poisedness constant  $\Lambda$  (which is scaling independent), as we have shown in Theorem 2.8.

In [8] two examples of algorithms that guarantee  $\Lambda$ -poisedness of the interpolation sets are proposed. Each algorithm either verifies that the current interpolation set is  $\Lambda$ -poised for some given value of  $\Lambda$ , or if it is not, replaces the ‘bad’ points with new points to maintain a  $\Lambda$ -poised interpolation set. It is shown that as long as  $\Lambda$  is reasonably large, this procedure will always be successful. The same algorithms can be applied to the regression case, since as we will point out in the next section,  $\Lambda$ -poisedness of a set  $Y$  of  $q_1$  points implies  $\Lambda$ -poisedness, in the regression sense, of any larger superset of  $Y$ .

It would be interesting to investigate algorithms that target the maintenance of the regression set directly, rather than maintaining a good subset that is well poised in the interpolation sense. The properties of Lagrange polynomials for regression that we have described above provide a good foundation for such a schemes. A formal, theoretically supported method for interpolation sets based on Lagrange polynomials is described in [9] and can be extended to the regression case.

**3. Error bounds for least-squares regression.** In this section we present Taylor-like bounds for linear and quadratic least-squares regression in terms of the poisedness constant  $\Lambda$ . These bounds are extensions of the bounds on polynomial interpolation in [8]. We will present the bounds here without proofs, since they are straightforward adaptations of the proofs in [8].

As in [8], we will make an additional assumption, for the remainder of the section, that  $y^0 = 0$  — that is, one of the interpolation points is at the center of the region of interest, which, by an earlier assumption, is a ball of radius  $\Delta$  around the origin. This assumption is very natural in a derivative-free optimization setting, since the center of the region of interest is typically the current best iterate, which is usually an interpolation point. (Note that if this assumption is not satisfied, it can always be made so by shifting the coordinates so that  $y^0 = 0$ . Since all the points of  $Y$  are in  $B(\Delta)$ , then, after the shift, the points of the shifted interpolation set are all in  $B(2\Delta)$ .)

We will also assume that  $\Delta$  has the smallest possible value that satisfies  $Y \subset B(\Delta)$  and  $y^0 = 0$ . Under the assumption  $y^0 = 0$ , the matrix  $\bar{M}$  can be written now as

$$(3.1) \bar{M} = \begin{bmatrix} 1 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 \\ 1 & y_1^1 & \cdots & y_n^1 & \frac{1}{2}(y_1^1)^2 & y_1^1 y_2^1 & \cdots & \frac{1}{(d-1)!}(y_{n-1}^1)^{d-1} y_n^1 & \frac{1}{d!}(y_n^1)^d \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots \\ 1 & y_1^p & \cdots & y_n^p & \frac{1}{2}(y_1^p)^2 & y_1^p y_2^p & \cdots & \frac{1}{(d-1)!}(y_{n-1}^p)^{d-1} y_n^p & \frac{1}{d!}(y_n^p)^d \end{bmatrix}.$$

We first consider regression of a function  $f(x)$  by a linear polynomial  $m(x)$ :

$$(3.2) \quad m(x) = c + g^\top x = c + \sum_{k=1}^n g_k x_k.$$

The sample set satisfies  $Y = \{0, y^1, \dots, y^p\} \subset B(\Delta)$ , where  $B(\Delta)$  is a ball of radius  $\Delta$  centered at the origin.

**THEOREM 3.1.** *Let  $Y = \{0, y^1, \dots, y^p\}$  be a  $\Lambda$ -poised set of  $p + 1$  regression points ( $p > n$ ) contained in a (closed) ball  $B(\Delta)$  centered at 0. Assume that  $f$  is continuously differentiable in an open domain  $\Omega$  containing  $B(\Delta)$  and that  $\nabla f$  is Lipschitz continuous in  $\Omega$  with constant  $\gamma_L > 0$ .*

*Then, for all points  $x$  in  $B(\Delta)$ , we have that*

- *the error between the gradient of the linear regression model and the gradient of the function satisfies*

$$(3.3) \quad \|e^g(x)\| \leq (5p_1^{\frac{1}{2}}\gamma_L\Lambda/2)\Delta,$$

- *the error between the linear regression model and the function satisfies*

$$|e^f(x)| \leq (5p_1^{\frac{1}{2}}\gamma_L\Lambda/2 + \gamma_L/2)\Delta^2.$$

In the quadratic case we assume that we have a poised set  $Y = \{0, y^1, \dots, y^p\}$  of  $p_1 > (n + 1)(n + 2)/2$  sample points ( $p_1 = p + 1$ ) in a ball  $B(\Delta)$  of radius  $\Delta$  centered at the origin. In addition we will assume that  $f$  is twice continuously differentiable in an open domain  $\Omega$  containing this ball and that  $\nabla^2 f$  is Lipschitz continuous in  $\Omega$  with constant  $\gamma_Q > 0$ .

It is possible to build the quadratic regression model

$$(3.4) \quad m(x) = c + g^\top x + \frac{1}{2}x^\top Hx = c + \sum_{1 \leq k \leq n} g_k x_k + \frac{1}{2} \sum_{1 \leq k, \ell \leq n} h_{k\ell} x_k x_\ell,$$

where  $H$  is a symmetric matrix of order  $n$ .

As one might expect, the error estimates in the quadratic case are linear in  $\Delta$  for the second derivatives, quadratic in  $\Delta$  for the first derivatives, and cubic in  $\Delta$  for the function values, where  $\Delta$  is the radius of the smallest ball containing  $Y$ .

**THEOREM 3.2.** *Let  $Y = \{0, y^1, \dots, y^p\}$ , with  $p_1 > (n + 1)(n + 2)/2$  and  $p_1 = p + 1$ , be a  $\Lambda$ -poised set of interpolation points contained in a (closed) ball  $B(\Delta)$  centered at 0. Assume that  $f$  is twice continuously differentiable in an open domain  $\Omega$  containing  $B(\Delta)$  and that  $\nabla^2 f$  is Lipschitz continuous in  $\Omega$  with constant  $\gamma_Q > 0$ .*

*Then, for all points  $x$  in  $B(\Delta)$ , we have that*

- *the error between the Hessian of the quadratic regression model and the Hessian of the function satisfies*

$$\|E^H(x)\| \leq (\alpha_Q^H \sqrt{p_1 q_1} \gamma_Q \Lambda) \Delta,$$

- *the error between the gradient of the quadratic regression model and the gradient of the function satisfies*

$$\|e^g(x)\| \leq (\alpha_Q^g \sqrt{p_1 q_1} \gamma_Q \Lambda) \Delta^2,$$

- the error between the quadratic regression model and the function satisfies

$$|e^f(x)| \leq (\alpha_Q^f \gamma_Q \sqrt{p_1 q_1} \Lambda + \beta_Q^f \gamma_Q) \Delta^3,$$

where  $\alpha_Q^H$ ,  $\alpha_Q^g$ ,  $\alpha_Q^f$ , and  $\beta_Q^f$  are small positive constants dependent on  $d = 2$  and independent of  $n$  and  $Y$ :

$$\alpha_Q^H = 6\sqrt{2}, \quad \alpha_Q^g = 6(1 + \sqrt{2}), \quad \alpha_Q^f = 6 + 9\sqrt{2}, \quad \beta_Q^f = \frac{1}{6}.$$

The above error bounds can be extended to the case of polynomials of higher degrees. However, as we have mentioned, in the context of derivative-free methods, on which we are focusing, the linear and the quadratic cases are normally sufficient. These error bounds can be used to show global convergence of various optimization methods based on least-squares regression models as long as the sample sets for regression remain  $\Lambda$ -poised, with  $\Lambda$  uniformly bounded, throughout the progress of the methods.

**4. Regression versus interpolation.** Given  $p_1 > q_1$  sample points, where  $q_1$  is the number of points required for unique interpolation, one can select the ‘best’ subset of  $q_1$  sample points and use those points to interpolate the given function  $f$ . One natural question arises: will such interpolation provide consistently worse or better models than the models based on regression using all the  $p_1$  points? Given our theory, the answer to this question seems to be ‘neither’. On the one hand, it is clear that the poisedness constant  $\Lambda$ , as defined by Definition 2.6, reduces (or remains the same) as the number of sample points increases. On the other hand, the error bounds depend on  $p_1$ , hence when  $p_1$  increases, so does its contribution to the error bounds.

To better understand the quality of both models, we conducted the simple experiments described below for the following functions:

$$\begin{aligned} f_1(x, y) &= 107 \sin(4x)^5 + 101x^3 + y^2 + 5xy + x + y, \\ f_2(x, y) &= (10x^5)/(8.021 + y)^3 + y^4, \\ f_3(x, y) &= 107x^5 + 101x^3 + y^2 + x + y + 5xy. \end{aligned}$$

We only report results for the first function, since the results for the others seem to follow the same pattern as for the first.

We generated a set of  $p$  random points,  $(x_i, y_i) \in \mathbb{R}^2$ ,  $i = 1, \dots, p$ , in the *unit-radius* square centered at the origin ( $\{x \in \mathbb{R}^2 : \|x\|_\infty \leq 1\}$ ). Together with the origin  $(0, 0)$ , this gives us  $p_1 = p + 1$  points in  $\mathbb{R}^2$ .

In the linear case ( $q = 2$ ), we considered all possible pairs of points  $(x_i, y_i), (x_j, y_j)$ ,  $i, j = 1, \dots, p$ ,  $i \neq j$ , and selected the pair with the best condition number for the matrix  $\bar{M}$ . (In practice, we worked with the  $p \times q$  submatrix of  $\bar{M}$  obtained by removing its first row and column.) Then, we built a linear interpolation model of the true function for this sample set. We call this model the best subset model.

In the quadratic case ( $q = 5$ ), we considered all possible sets of 5 points and selected the set with the best condition number for the matrix  $\bar{M}$ . (We worked also with the  $p \times q$  submatrix of  $\bar{M}$  obtained by removing its first row and column.) We built a quadratic interpolation model of the true function for this sample set (called also the best subset model).

For each case (linear and quadratic), we compared the least-squares regression (LSR) model to the best subset (BS) model. The error between each model and the

#points	BS error	LSR error	BS worse	LSR worse
7	852770	794550	101	79
9	478160	392980	118	60
11	305470	236630	121	48
13	310080	209930	136	34
15	271000	193760	135	31
17	271890	189620	143	17
19	233790	174270	138	24
21	213590	160570	138	28
23	226930	158040	149	22
25	209190	156920	135	20
31	197770	149400	141	18

FIG. 4.1. Results comparing the errors between the function  $f_1$  and the corresponding best subset (BS) and least-squares regression (LSR) models, over 200 random runs.

true function was evaluated on a 0.1 step lattice of the unit radius square. We considered two types of error: (E1) maximum absolute error for all points in the lattice; (E2) error summed for all lattice points in the  $\ell_2$  sense. We repeated the experiment 200 times and counted the number of times when each model was significantly better ( $> 0.001$ ) than the other in *both* errors. When one of the errors was within 0.001 for both models no win was declared.

We report the results corresponding to the quadratic case in Figure 4.1 for  $p_1 = 7, 9, 11, 13, 15, 17, 19, 21, 23, 25, 31$  randomly generated points. We only report the error summed for all lattice points in the  $\ell_2$  sense. The error is reported in a cumulative way for all 200 runs.

For both models (LSR and BS), the error decreased (the approximation improved) as the number of points increased. The BS model becomes progressively worse compared with the LSR model — although this effect seemed to tail off once we had *enough* points. In any case, no model was consistently better than the other. For example, when using 21 points, of the 200 runs, the BS model was worse 138 times and LSR model was worse 28 times. (Note that the cumulative sum of the errors is as high as it is because the region is relatively large given the irregularities of the function. For example, again with 21 points, the error summed for all lattice points in the  $\ell_2$  sense, over the 200 runs, was 0.1011 when the *radius* of the square was scaled to 0.01.)

One possible advantage of using least-squares regression models is when there is noise in the evaluation of the true function  $f(x)$ . It is easy to show that if the noise is random and independently and identically distributed with mean zero, then the least-squares regression of the noisy function and the least-squares regression of the true function (based on the same sequence of sample sets) converge to each other (pointwise) as the number of sample points tends to infinity (see [9]).

Another possible advantage of using regression is when the function is not very smooth and occasional spikes make the interpolation unstable. In this case, when there are enough sample points available, it might be beneficial to use all of them to smooth out the effect of the spikes, although this statement has not been verified experimentally.

**5. Underdetermined interpolation.** We will now consider the case when  $p < q$ , that is the number of interpolation points in  $Y$  is smaller than the number of elements in the polynomial basis  $\phi$ . Then the matrix  $M(\phi, Y)$  has more columns than rows. The interpolation polynomials defined by

$$(5.1) \quad m(y^i) = \sum_{k=0}^q \alpha_k \phi_k(y^i) = f(y^i), \quad i = 0, \dots, p$$

are no longer unique.

**5.1. The choice of an underdetermined model.** The simplest approach to restrict the system so that it has a unique solution is to remove the last  $q - p$  columns of  $M(\phi, Y)$ . This causes the last  $q - p$  elements of the solution  $\alpha$  to be zero. Such an approach approximates some elements of  $\alpha$ , while it sets others to zero solely based on the order of the elements in the basis  $\phi$ . Clearly this approach is not very desirable, without any knowledge of, for instance, the sparsity structure of the gradient and the Hessian of the function  $f$ . There is also a more fundamental drawback: the first  $p_1$  columns of  $M(\phi, Y)$  may be linearly dependent. A natural conclusion would be that our sample points are not poised (in some sense) and we have to change them. However, if we had selected a different subset of  $p$  columns of  $M(\phi, Y)$ , it might have been well poised. From now on, we will use a notion of *sub-basis* of the basis  $\phi$  to mean a subset of  $p_1$  elements of the basis  $\phi$ . Selecting  $p_1$  columns of  $M(\phi, Y)$ , therefore, corresponds to selecting the appropriate sub-basis  $\tilde{\phi}$ . Let us consider the following example.

EXAMPLE 5.1.  $\phi = \{1, x_1, x_2, \frac{1}{2}x_1^2, x_1x_2, \frac{1}{2}x_2^2\}$ ,  $Y = \{y^0, y^1, y^2, y^3\}$ ,  $y^0 = [0, 0]^\top$ ,  $y^1 = [0, 1]^\top$ ,  $y^2 = [0, -1]^\top$ ,  $y^3 = [1, 0]^\top$ . The matrix  $M = M(\phi, Y)$  is given by

$$M = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0.5 \\ 1 & 0 & -1 & 0 & 0 & 0.5 \\ 1 & 1 & 0 & 0.5 & 0 & 0 \end{bmatrix}.$$

If we select the first four columns of  $M$  then the system is still not well defined, since the matrix is singular. Hence the set  $Y$  is not poised with respect to the sub-basis  $\tilde{\phi} = \{1, x_1, x_2, \frac{1}{2}x_1^2\}$ , and a new set of sample points is needed. Notice now that if another sub-basis was selected, for instance,  $\tilde{\phi} = \{1, x_1, x_2, \frac{1}{2}x_2^2\}$ , then the set  $Y$  is well poised and the matrix consisting of the first, the second, the third and the sixth columns of  $M$  is well-conditioned and a unique solution exists. If the Hessian of  $f$  happens to look like

$$\begin{bmatrix} 0 & 0 \\ 0 & \frac{\partial^2 f}{\partial x_2^2}(x) \end{bmatrix}$$

then the reduced system actually produces the full quadratic model of  $f$ .

If the sparsity structure of the derivatives of  $f$  is known in advance then this advantage can be exploited trivially by deleting appropriate columns from the system (5.1). A more sophisticated version of this idea is exploited in [5] for group partial separable functions. If no such structure is known, then there is no reason to select one set of columns over another except for *geometry considerations*. Hence it makes

sense to select those columns that produce the best geometry. The following definition of well poisedness is consistent with this approach.

DEFINITION 5.1. Let  $\Lambda > 0$  be given.

A set  $Y = \{y^0, y^1, \dots, y^p\} \subset B(1)$ , with  $p < q$ , where  $q + 1$  is the dimension of the space of polynomials of degree  $\leq d$ , is said to be  $\Lambda$ -poised in  $B(1)$  (in the sub-basis sense) if and only if there exists a sub-basis  $\tilde{\phi}(x)$  of  $p_1$  elements such that for any  $x \in B(1)$  the solution  $\lambda(x)$  of

$$(5.2) \quad \sum_{i=0}^p \lambda_i(x) \tilde{\phi}(y^i) = \tilde{\phi}(x)$$

satisfies  $\|\lambda(x)\| \leq \Lambda$ .

It is easy to show (as it is done for the complete interpolation case in [8]) that the functions  $\lambda_i(x)$  are, in fact, the Lagrange polynomials  $\tilde{\mathcal{L}}_i(x) = (\gamma^i)^\top \tilde{\phi}(x)$  for the sub-basis  $\tilde{\phi}(x)$ , satisfying

$$\tilde{\mathcal{L}}_i(y^j) = \delta_{ij}, \quad i, j = 0, \dots, p.$$

The approach to select a unique solution to (5.1) could then be the following. Given the sample set  $Y$ , select the sub-basis  $\tilde{\phi}(x)$  so that the poisedness constant  $\Lambda$  is minimized. Then consider the system with the appropriate columns of  $M(\phi, Y)$  and find the unique solution to the system. The following example shows the possible disadvantages of this approach.

EXAMPLE 5.2. Let us consider the purely linear case in  $\mathbb{R}^3$  for simplicity. An example for a quadratic case can be constructed in a similar manner. Consider  $\phi = \{1, x_1, x_2, x_3\}$  and  $Y = \{y^0, y^1, y^2\}$ , where, as always,  $y^0 = [0, 0, 0]^\top$ , and where  $y^1 = [1, 0, 0]^\top$  and  $y^2 = [0, 1, 1 - \epsilon]^\top$ . Assume  $f_Y = [0, b_1, b_2]^\top$ . The system (5.1) then becomes

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 - \epsilon \end{bmatrix} \alpha = \begin{bmatrix} 0 \\ b_1 \\ b_2 \end{bmatrix}.$$

The best sub-basis for  $Y$  is then  $\tilde{\phi} = \{1, x_1, x_2\}$ . If we select the appropriate columns of  $M(\phi, Y)$  and solve the reduced system, we obtain the following solution for the coefficients of  $m(x)$

$$\alpha = \begin{bmatrix} 0 \\ b_1 \\ b_2 \\ 0 \end{bmatrix}.$$

Now, if we consider  $y^2 = [0, 1 - \epsilon, 1]^\top$ , then the best sub-basis is  $\tilde{\phi} = \{1, x_1, x_3\}$  and the solution that we will find with this approach is

$$\alpha = \begin{bmatrix} 0 \\ b_1 \\ 0 \\ b_2 \end{bmatrix}.$$

Notice that the two possible solutions are very different from each other, yet as  $\epsilon$  goes to zero the two sets of points converge pointwise to each other. Hence, we see that the sub-basis approach suffers from a lack of robustness with respect to small perturbations in the sample set. We also notice that in the first (second) case the fourth (third) element of the coefficient vector is set to zero and the third (fourth) element is set to  $b_2$ . Hence, each solution is biased towards one of the basis components ( $x_2$  or  $x_3$ ) without using any actual information about the structure of  $f$ . A more suitable approach would be to treat all such components equally in some sense. This can be achieved by the minimum-norm solution of (5.1).

For this example, the minimum-norm solution in the first case is

$$\alpha^{mn} = M(\phi, Y)^\top (M(\phi, Y)M(\phi, Y)^\top)^{-1} f_Y = \begin{bmatrix} 0 \\ b_1 \\ \frac{b_2}{2-2\epsilon+\epsilon^2} \\ \frac{(1-\epsilon)b_2}{2-2\epsilon+\epsilon^2} \end{bmatrix}$$

and in the second case is

$$\alpha^{mn} = \begin{bmatrix} 0 \\ b_1 \\ \frac{(1-\epsilon)b_2}{2-2\epsilon+\epsilon^2} \\ \frac{b_2}{2-2\epsilon+\epsilon^2} \end{bmatrix}.$$

These two solutions converge to  $[0, b_1, b_2/2, b_2/2]^\top$  as  $\epsilon$  converges to zero. Hence, not only is the minimum-norm solution robust with respect to small perturbations of the data, but it also distributes evenly the elements of the gradient over the  $x_2$  and  $x_3$  basis components.

For the reasons described above it is beneficial to consider the minimum-norm solution of the system (5.1). The minimum-norm solution is expressed as

$$M(\phi, Y)^\top [M(\phi, Y)M(\phi, Y)^\top]^{-1} f_Y.$$

It is well known that a minimum-norm solution of an underdetermined system of linear equations is not invariant under linear transformations. In our case, this fact means that the minimum-norm solution depends on the choice of  $\phi$ . It is easy to show that the resulting interpolation polynomial also depends on the choice of  $\phi$  in the system (5.1).

This implies that depending on the choice of  $\phi$  we can obtain a better or a worse approximation to  $f$  by computing the minimum-norm interpolating polynomials. Ideally, we would like, for each set  $Y$ , to identify the ‘best’ basis  $\phi$ , which would generate the ‘best’ minimum-norm interpolating polynomial. However, it is a nontrivial task to define such a basis. First of all, one should define the best interpolating polynomial. The natural choice is the polynomial that has the smallest approximation error with respect to the function  $f$ . However, the definition of the best basis (and hence of the best polynomial) should only depend on  $Y$ .

In the next subsection, we will consider minimum-norm underdetermined interpolation for the specific choice of the natural basis  $\bar{\phi}$ . We will argue at the end of the next subsection that  $\bar{\phi}$  is a reasonable choice for the basis.

**5.2. Lagrange polynomials and  $\Lambda$ -poisedness for underdetermined interpolation.** We will consider the natural basis  $\bar{\phi}$  defined by (2.6) and the corresponding matrix  $\bar{M} = M(\bar{\phi}, Y)$  defined by (2.5). We omit the dependence on  $Y$ , since we keep  $Y$  fixed in the discussion below.

We will start by introducing the definition of the set of Lagrange polynomials for underdetermined interpolation.

DEFINITION 5.2. *Given a set of interpolation points  $Y = \{y^0, y^1, \dots, y^p\}$ , with  $p < q$ , where  $q + 1$  is the dimension of the space of polynomials of degree  $\leq d$ , a set of  $p_1 = p + 1$  polynomials  $\mathcal{L}_j(x) = \sum_{i=0}^q (\alpha_j)_i \bar{\phi}_i(x)$ ,  $j = 0, \dots, p$ , is called a set of Lagrange minimum-norm polynomials for the basis  $\bar{\phi}$  if it is a minimum-norm solution of*

$$\mathcal{L}_j(y^i) \stackrel{\text{m.n.}}{=} \delta_{ij} = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j. \end{cases}$$

The Lagrange minimum-norm polynomials are thus given by the minimum-norm solution of

$$\bar{M}\alpha_j \stackrel{\text{m.n.}}{=} e_{j+1}, \quad j = 0, \dots, p.$$

This set of polynomials is an extension of the traditional Lagrange polynomials to the case when  $p < q$ . Clearly these polynomials no longer compose a basis, since there are not enough of them. However, as in the regression case, many other properties of Lagrange interpolation polynomials are preserved.

The set of minimum-norm Lagrange polynomials exists and is unique if the matrix  $\bar{M}$  has full row rank. In this case, we will say that  $Y$  is poised. We again note that here the Lagrange polynomials generally depend on the choice of the basis  $\bar{\phi}$ , but it is easy to see that the poisedness of  $Y$  does not.

Just as in the case of standard Lagrange polynomials, the minimum-norm interpolating polynomial  $m(x)$  in the underdetermined case has a simple representation in terms of the minimum-norm Lagrange polynomials.

LEMMA 5.3. *Let  $Y = \{y^0, y^1, \dots, y^p\}$  be a set of poised sample points for the function  $f(x)$  and let  $m(x)$  be the minimum-norm interpolating polynomial (in terms of the basis  $\bar{\phi}$ ) of  $f(x)$  at the points in  $Y$ . Let  $\{\mathcal{L}_i(x), i = 0, \dots, p\}$  be the set of the minimum-norm Lagrange polynomials given in Definition 5.2. Then*

$$m(x) = \sum_{i=0}^p f(y^i) \mathcal{L}_i(x).$$

*Proof.* The proof is similar to the proof of Lemma 2.3.  $\square$

Let  $\bar{M} = U\Sigma V^\top$  be a reduced singular value decomposition of  $\bar{M}$ , defined in Section 1.1.

We will now show that, as in the case of polynomial interpolation [8] and regression, the geometric interpretations of the Lagrange polynomials can be easily derived. Thus, we will also have an analogous definition of  $\Lambda$ -poisedness.

Given a poised set  $Y = \{y^0, y^1, \dots, y^p\} \subset B(1) \subset \mathbb{R}^n$  and  $x \in B(1)$  we attempt to express the vector  $\bar{\phi}(x)$  in terms of the vectors  $\bar{\phi}(y^i)$ ,  $i = 0, \dots, p$ . Since the dimension

of the vector  $\bar{\phi}(x)$  is  $q_1 > p_1$ , it may no longer be possible to express it in terms of  $p_1$  vectors  $\bar{\phi}(y^i)$ ,  $i = 0, \dots, p$ . Hence, we will be looking for the least-squares solution to the following system

$$(5.3) \quad \sum_{i=0}^p \lambda_i(x) \bar{\phi}(y^i) \stackrel{\ell.s.}{=} \bar{\phi}(x).$$

This system is an extension of the similar systems introduced in [8] and of system (2.4) in Section 2.2. We have a kind of duality in that in the system (2.4) the minimum  $\ell_2$ -norm solution  $\lambda(x)$  to the system corresponded to the least-squares regression Lagrange polynomials, while in this case the least-squares solution  $\lambda(x)$  corresponds to the minimum  $\ell_2$ -norm Lagrange polynomials.

LEMMA 5.4. *Given a poised set  $Y$ , the functions  $\lambda_i(x)$ ,  $i = 0, \dots, p$ , defined by the least-squares solution of (5.3), form the set of minimum-norm Lagrange polynomials for  $Y$  given in Definition 5.2.*

*Proof.* The proof is similar to the proof of Lemma 2.4.  $\square$

The following definition of well poisedness is analogous to Definitions 2.5 and 2.6.

DEFINITION 5.5. *Let  $\Lambda > 0$  be given. Let  $\bar{\phi}$  be the natural basis of monomials.*

*A set  $Y = \{y^0, y^1, \dots, y^p\}$ , with  $p < q$ , where  $q + 1$  is the dimension of the space of polynomials of degree  $\leq d$ , is said to be  $\Lambda$ -poised in  $B(1)$  (in the minimum-norm sense) if and only if for any  $x \in B(1)$  there exists a unique  $\lambda(x) \in \mathbb{R}^{p+1}$  such that*

$$\sum_{i=0}^p \lambda_i(x) \bar{\phi}(y^i) \stackrel{\ell.s.}{=} \bar{\phi}(x), \quad \text{with} \quad \|\lambda(x)\| \leq \Lambda.$$

This definition is equivalent to having all the Lagrange polynomials bounded by  $\Lambda$  on  $B(1)$ .

The following theorem states that if a set is well poised in the minimum-norm sense then it is well poised in the sub-basis sense and vice versa.

THEOREM 5.6. *There exists a constant  $\theta$  independent of  $\Lambda$  and  $Y$  such that if a set  $Y$  is  $\Lambda$ -poised in the sub-basis sense, then it is  $\sqrt{p_1 q_1} \theta \Lambda$ -poised in the sense of Definition 5.5.*

*Conversely, if a set  $Y = \{y^0, y^1, \dots, y^p\} \subset B(1)$  is  $\Lambda$ -poised by Definition 5.5, then the set is  $(q - p + 1) \sqrt{p_1 q_1} \theta \Lambda$ -poised in the sub-basis sense.*

*Proof.* Assume that  $Y$  is  $\Lambda$ -poised in the sub-basis sense, and that  $\tilde{\phi}$  is the sub-basis. Let  $\tilde{\mathcal{L}}_i(x) = (\gamma^i)^\top \tilde{\phi}(x)$ ,  $i = 0, \dots, p$ , be the Lagrange polynomials for the sub-basis  $\tilde{\phi}$  (see Definition 5.1 and the paragraph following). Then

$$\max_{0 \leq i \leq p} \max_{x \in B(1)} \tilde{\mathcal{L}}_i(x) \leq \Lambda.$$

As it is shown in the proof of Theorem 2.8, there exists a constant  $\sigma$ , independent of  $Y$  and  $\Lambda$ , such that  $\max_{x \in B(1)} (\gamma^i)^\top \tilde{\phi}(x) \geq \sigma \|\gamma^i\|$  for each  $i$ . Hence,  $\theta \Lambda \geq \|\gamma^i\|$ , for each  $i$  and  $\theta = 1/\sigma$ .

Now let us consider the minimum-norm Lagrange polynomials for  $Y$ , given by  $\mathcal{L}_i(x) = (\alpha^i)^\top \bar{\phi}(x)$ ,  $i = 0, \dots, p$ . The vector  $\alpha^i$  is the minimum-norm solution of

$$\bar{M} \alpha^i = e_{i+1}.$$

Hence,  $\|\alpha^i\| \leq \|\gamma^i\|$ . Since  $\bar{\phi}_j(x) \leq 1$ , for all  $j = 0, \dots, q$  and all  $x \in B(1)$ , then

$$\max_{x \in B(1)} (\alpha^i)^\top \bar{\phi}(x) \leq \|\alpha^i\|_1 \leq \sqrt{q_1} \|\alpha^i\| \leq \sqrt{q_1} \|\gamma^i\| \leq \sqrt{q_1} \theta \Lambda.$$

We have shown that  $Y$  is  $\sqrt{p_1 q_1} \theta \Lambda$ -poised in the minimum-norm sense.

Now assume that  $Y$  is  $\Lambda$ -poised in the minimum-norm sense. We can apply Lemma 1.1, with  $m = q_1$  and  $n = p_1$ , to the columns of  $\bar{M}$  and conclude that we can select a subset of  $p_1$  columns such that the corresponding submatrix  $\bar{M}_{p_1}$  is nonsingular and

$$\bar{M}_{p_1} \gamma^i = e_{i+1}, \quad |\gamma_j^i| \leq (q-p+1) |\alpha_j^i|, \quad j = 0, \dots, p.$$

The selected columns determine a sub-basis  $\tilde{\phi}$  and the vector of coefficients  $\gamma^i$  determines the  $i$ -th Lagrange polynomial  $\tilde{\mathcal{L}}_i(x) = (\gamma^i)^\top \tilde{\phi}(x)$ . As before, we know that there exists a constant  $\sigma$ , independent of  $Y$  and  $\Lambda$ , such that  $\max_{x \in B(1)} (\alpha^i)^\top \tilde{\phi}(x) \geq \sigma \|\alpha^i\|$  for each  $i$ . Hence,  $\theta \Lambda \geq \|\alpha^i\|$ , for each  $i$  and  $\theta = 1/\sigma$ . On the other hand,

$$\max_{x \in B(1)} (\gamma^i)^\top \tilde{\phi}(x) \leq \|\gamma^i\|_1 \leq \sqrt{q_1} \|\gamma^i\| \leq (q-p+1) \sqrt{q_1} \|\alpha^i\| \leq (q-p+1) \sqrt{q_1} \theta \Lambda.$$

We have established, that  $Y$  is  $(q-p+1) \sqrt{p_1 q_1} \theta \Lambda$ -poised in the sub-basis sense. (The values of  $\theta$  for the specific cases of linear and quadratic interpolations are discussed after the proof of Theorem 2.8.)  $\square$

**REMARK 5.1.** *It is easy to see that the results of this subsection hold for any given basis  $\phi$ , as long as it remains fixed throughout the discussion. (Note that the constants in Theorem 5.6 vary with the choice of  $\phi$ .) Hence, the minimum-norm Lagrange polynomials can be defined for any basis. The definition of  $\Lambda$ -poisedness also can be introduced for any basis. However, for any given set  $Y$ , one can create different bases, which, when used in Definition 5.5, will result in different constants for  $\Lambda$ -poisedness of  $Y$ . Moreover, by varying the basis  $\phi$ , one can make the constant  $\Lambda$  as large or as close to 1 as desired. Clearly for the definition of  $\Lambda$ -poisedness to make sense, the  $\Lambda$  constant should be related to the quality of the geometry  $Y$  and the resulting interpolation. Hence, we consider only one basis (the basis  $\phi$ ).*

We choose  $\bar{\phi}$  as the basis because: (i) it appears naturally in Taylor bounds and their derivations; (ii) it is the obvious choice in algorithmic implementations; (iii) it is well scaled; (iv)  $\Lambda$ -poisedness of a set  $Y$  in terms of  $\bar{\phi}$  implies  $\mathcal{O}(\Lambda)$ -poisedness of  $Y$  in terms of any other basis  $\psi$ , such that  $\bar{\phi} = P\psi$  and  $\|P\| \|P^{-1}\| = \mathcal{O}(1)$  (the last statement is easy to show from the definition of  $\Lambda$ -poisedness and from Theorem 5.7 of the next section).

In the next section, we will use the properties of  $\bar{\phi}$  to show the relation between the poisedness constant  $\Lambda$  and the condition number of  $\bar{M}$ .

**REMARK 5.2.** *As we pointed out in the introduction of this paper, minimum-norm models for underdetermined interpolation have been developed in [17], by minimizing the Frobenius norm of the change of the second derivative of the models from one iteration of the optimization algorithm to the next.*

*Related to the need of updating such models, the author in [17] also proposed a definition of Lagrange polynomials for underdetermined interpolation. In the notation of our paper, the definition in [17] can be described as a modified Definition 5.2 where the norm being minimized is applied only to the components of the second-order terms of the Lagrange polynomials (see [9]).*

To ensure the existence and uniqueness of the Lagrange polynomials, the definition in [17] requires not only that  $\bar{M}$  has full row rank but also that  $Y$  contains a subset of  $n + 1$  points that are poised in the linear interpolation sense. We impose this requirement directly on  $Y$  when we derive the first-order error bound.

In the spirit of [15, 17], one can use the minimum-norm Lagrange polynomials to maintain ‘good’ sample sets for minimum-norm interpolation. The advantages and disadvantages of such a scheme are similar to those for the regression case which were discussed in the last paragraph of Section 2.3.

**5.3.  $\Lambda$ -poisedness and the condition number of  $M(\bar{\phi}, Y)$ .** We will again connect  $\Lambda$ -poisedness with the condition number of  $M(\bar{\phi}, Y)$ . As in the regression case we will scale  $Y$  in such a way that the smallest ball containing  $Y$  is  $B(1)$ .

Recall that under this scaling  $1 \leq \|\bar{M}\| \leq p_1^{1/2} q_1$ . Hence to bound the condition number of  $\bar{M}$  in terms of  $\Lambda$  (and vice versa) all we need to do is to bound  $\|\bar{M}^{-1}\|$  in terms of  $\Lambda$  (and vice versa). We will now present the analogue for the underdetermined case of Theorem 2.8 of this paper and Theorem 3.3 of [8]. Recall the reduced singular value decomposition of  $\bar{M} = U\Sigma V^\top$ .

**THEOREM 5.7.** *If  $\Sigma$  is nonsingular and  $\|\Sigma^{-1}\| \leq \Lambda$ , then the set  $Y$  is  $\sqrt{q_1}\Lambda$ -poised (according to Definition 5.5) in the unit ball  $B(1)$  centered at 0. Conversely, if the set  $Y$  is  $\Lambda$ -poised, according to Definition 5.5, in the unit ball  $B(1)$  centered at 0, then  $\Sigma$  is nonsingular and*

$$(5.4) \quad \|\Sigma^{-1}\| \leq \theta\Lambda,$$

where  $\theta > 0$  is dependent on  $n$  and  $d$  but independent of  $Y$  and  $\Lambda$ .

*Proof.* As in the proof of Theorem 2.8, it is trivial to show that if  $\Sigma$  is nonsingular and  $\|\Sigma^{-1}\| \leq \Lambda$  then the least-squares solution

$$\|\lambda(x)\| \leq \|U\Sigma^{-1}V^\top\| \|\bar{\phi}(x)\| \leq q_1^{\frac{1}{2}} \|\Sigma^{-1}\| \|\bar{\phi}(x)\|_\infty \leq q_1^{\frac{1}{2}} \Lambda,$$

since  $\max_{x \in B(1)} \|\bar{\phi}(x)\|_\infty \leq 1$ .

To prove the other relation we note, first, that the matrix  $\Sigma$  is nonsingular by the definition of  $\Lambda$ -poisedness. To prove that there exists a constant  $\theta > 0$ , independent of  $Y$  and of  $\Lambda$ , such that  $\|\Sigma^{-1}\| \leq \theta\Lambda$ , we would proceed exactly as in the proof of Theorem 2.8.  $\square$

To obtain the specific values of  $\theta$  in the linear and quadratic cases we can apply the results presented immediately after Theorem 2.8. To relax the assumption that the radius of the ball enclosing  $Y$  is 1, we can use the same arguments as at the end of Subsection 2.3.

**5.4. Error bounds for underdetermined interpolation.** We study the quality of the quadratic minimum-norm interpolation model in the general case. Recall that  $q_1 = q + 1 = (n + 1)(n + 2)/2$ . We will again assume that  $\Delta$  has the smallest possible value that satisfies  $Y \subset B(\Delta)$  and  $y^0 = 0$ . The derivation of the general error bound follows exactly the derivation in [8] for complete quadratic polynomial interpolation. We need to consider the submatrix  $\bar{M}_{p \times q}$  of  $\bar{M}$  obtained by removing its first row and its first column. Consider the reduced SVD of the scaled version of  $\bar{M}_{p \times q}$

$$\hat{M}_{p \times q} = \bar{M}_{p \times q} \begin{bmatrix} D_\Delta^{-1} & 0 \\ 0 & D_{\Delta^2}^{-1} \end{bmatrix} = U_{p \times p} \Lambda_{p \times p} V_{q \times p}^\top,$$

where  $D_\Delta$  is a diagonal matrix of dimension  $n$  with  $\Delta$  in the diagonal entries and  $D_{\Delta^2}$  is a diagonal matrix of dimension  $q - n$  with  $\Delta^2$  in the diagonal entries. The scaled matrix corresponds to using the scaled interpolation set  $\{0, y^1/\Delta, \dots, y^p/\Delta\}$  at the outset.

We will make use of the following notation: given a symmetric matrix  $H$ ,  $\text{svec}(H)$  is a vector in  $\mathbb{R}^{n(n+1)/2}$  that stores the upper triangular part of  $H$  row by row, consecutively. The following theorem exhibits the error bound on the underdetermined quadratic interpolation model.

**THEOREM 5.8.** *Let  $Y = \{0, y^1, \dots, y^p\}$ , with  $p_1 < (n+1)(n+2)/2$  and  $p_1 = p+1$ , be a  $\Lambda$ -poised set of points (in the minimum-norm sense) contained in a (closed) ball  $B(\Delta)$  centered at 0. Assume that  $f$  is twice continuously differentiable in an open domain  $\Omega$  containing  $B(\Delta)$  and that  $\nabla^2 f$  is Lipschitz continuous in  $\Omega$  with constant  $\gamma_Q > 0$ .*

*Then, for all points  $x$  in  $B(\Delta)$ , we have that the error between the gradient of the quadratic minimum-norm model and the gradient of the function*

$$e^g(x) = \nabla m(x) - \nabla f(x)$$

*and the error between the Hessian of the quadratic minimum-norm model and the Hessian of the function*

$$E^H(x) = \nabla^2 m(x) - \nabla^2 f(x)$$

*satisfy*

$$\left\| V_{q \times p}^\top \begin{bmatrix} D_\Delta t(x) \\ D_{\Delta^2} e^H(x) \end{bmatrix} \right\| \leq (6(p_1 q_1)^{\frac{1}{2}} \gamma_Q \Lambda) \Delta^3,$$

where  $t(x) = e^g(x) - E^H(x)x$  and  $e^H(x) = \text{svec}(E^H(x))$ .

*Proof.* We apply the same algebraic manipulations and Taylor expansions to the interpolating conditions (5.1) as we did in [8]. We omit them here for the sake of brevity. As in [8], we obtain the following system

$$(5.5) \quad \bar{M}_{p \times q} \begin{bmatrix} D_\Delta^{-1} & 0 \\ 0 & D_{\Delta^2}^{-1} \end{bmatrix} \begin{bmatrix} D_\Delta t(x) \\ D_{\Delta^2} e^H(x) \end{bmatrix} = \mathcal{O}(\Delta^3).$$

The right-hand-side of this system is bounded in norm by  $3\sqrt{p}\gamma_Q\Delta^3/2$ . Since  $\|\Sigma_{p \times p}^{-1}\| \leq \|\Sigma^{-1}\| \leq \theta\Lambda \leq 4\sqrt{q_1}\Lambda$ , we obtain from the  $\Lambda$ -poisedness assumption that

$$\left\| V_{q \times p}^\top \begin{bmatrix} D_\Delta t(x) \\ D_{\Delta^2} e^H(x) \end{bmatrix} \right\| \leq \frac{3}{2} p^{\frac{1}{2}} \gamma_Q \|\Sigma_{p \times p}^{-1}\| \Delta^3 \leq 6(p_1 q_1)^{\frac{1}{2}} \gamma_Q \Lambda \Delta^3.$$

□

In the derivative-free optimization context, one is particularly interested in the error of the gradient and Hessian approximation at the center of the trust region. Hence, if we set  $x = 0$ , which means that we are evaluating the error at  $x = y^0 = 0$ , we obtain

$$\left\| V_{q \times p}^\top \begin{bmatrix} D_\Delta [g - \nabla f(x)] \\ D_{\Delta^2} [\text{svec}(H - \nabla^2 f(x))] \end{bmatrix} \right\| \leq 6(p_1 q_1)^{\frac{1}{2}} \gamma_Q \Lambda \Delta^3,$$

where  $m(x) = c + g^\top x + \frac{1}{2}x^\top Hx$ ,  $\nabla m(x) = Hx + g$ , and  $\nabla^2 m(x) = H$ .

The presence of  $V_{q \times p}^\top$  in the general error bound tells us that we can only measure the orthogonal projection of the error onto the  $p$ -dimensional linear subspace spanned by the rows of the matrix  $\bar{M}_{p \times q}$ , which seems reasonable since we do not have enough information to do better.

It is easy to see that if  $p = q$ , then  $V_{q \times p}$  is orthonormal and hence can be removed. In this case, one recovers the bound on full quadratic interpolation (see [8]).

It is also possible to show that if  $\nabla f$  and  $\nabla^2 f$  have a specific sparsity structure that corresponds to a certain sub-basis for which  $Y$  is  $\Lambda$ -poised, then a similar error bound can be established for the quadratic interpolation based on the sub-basis. We do not include the result and the proof here, because it is nothing more than repetition of the corresponding error bound result in [8].

In [9] we investigate the quality of a minimum-norm Frobenius interpolation model when the interpolation set  $Y$  contains a subset of  $n + 1$  points that are  $\Lambda$ -poised for linear interpolation.

**5.5. Numerical results for the underdetermined case.** As for the overdetermined case, we generated the same set of simple two-dimensional numerical examples. We report here the results for the function  $f_1(x, y) = 107 \sin(4x)^5 + 101x^3 + y^2 + 5xy + x + y$ .

We considered, for the quadratic case, the underdetermined situations where the number of points is  $p_1 = 5, 4, 3$  (including the origin). We built the minimum-norm (MN) model for each case.

Then we considered all possible sub-bases of cardinality  $p_1$  of the basis  $\bar{\phi}$  for the quadratic case, which has cardinality 6, and selected the resulting  $p_1 \times p_1$  sub-matrix of  $\bar{M}$  with the best condition number. We, then, built an interpolation model of the true function for this sub-basis. We call this model also the best basis (BB) model. (Here we again worked with the  $p \times q$  submatrix of  $\bar{M}$  obtained by removing its first row and column.)

We compared the minimum-norm (MN) model to the best basis (BB) model. The error between each model and the true function was evaluated on a 0.1 step lattice of the *unit radius* square. We considered two types of error (E1 and E2) as in the overdetermined numerical tests. We repeated the experiment 200 times and counted the number of times when each model was better than the other in both errors. When one of the errors was within 0.001 for both models no win was declared. The errors reported are cumulative for the 200 runs and computed using E2. The BB model was worse than the MN model, and moreover one should recall the lack of continuity in this solution, illustrated at the beginning of this section.

The results for the unit square in Figure 5.1 are followed by the results for a scaled square (scaled to have a *radius* of 0.01) in Figure 5.2. Again, as for the overdetermined case, the sum of the residuals is high in the unit square because the region is relatively large given the irregularities of the function. In the scaled square, the MN model behaved even better when compared to the BB model, since it was consistently better.

**6. Conclusions.** We have shown that many theoretical properties of fully determined polynomial interpolation models studied in [8] can be extended to the cases of polynomial least-squares regression and underdetermined interpolation. Underdetermined interpolation is extensively used in modern implementations of model-based derivative-free optimization methods, see for example [18]. Our very limited computational examples were included to give the reader an idea as to why using minimum-

#points	BB error	MN error	BB Worse	MN Worse
5	81950	653000	133	28
4	594320	323480	175	9
3	594750	368640	163	16

FIG. 5.1. Results comparing the errors between the function  $f_1$  and the corresponding best basis (BB) and minimum-norm (MN) models, over 200 random runs.

#points	BB error	MN error	BB worse	MN worse
5	19.7096	0.4839	58	0
4	324.4571	0.9466	193	0
3	88.7755	2.1193	118	0

FIG. 5.2. Results comparing the errors between the function  $f_1$  and the corresponding best basis (BB) and minimum-norm (MN) models, over 200 random runs, in the case where the square has a radius of 0.01.

norm incomplete interpolation can be beneficial. However, until recently no complete convergence theory had been developed. In the upcoming book [9], we develop a unifying theory for model-based methods and show, for example, that a modified version of the algorithm in [16] (which uses minimum-norm Frobenius interpolation) is globally convergent.

As we mention in the introduction, there has been only very limited use of regression models in derivative-free optimization. Perhaps, the main advantage of using the regression models and overdetermined sample sets is to reduce the effect of noise in the function values. The first attempt to adapt model-based derivative-free optimization to noisy problems was proposed in [14], where an interpolation-based trust-region algorithm used in [16] is applied to noisy functions by sampling the objective function at the sample points repeatedly. With the theory in [8] one can claim that the resulting interpolation models can be made sufficiently accurate since they are based on well-poised interpolation sets. However, with the theory provided in this paper, as well as in the book [9], it is now possible to have regression models based on sampling the objective function at, for instance, randomly selected sample points, as long as a well-poised sample set is generated. This approach may be advantageous, especially when sampling a noisy function at the same point repeatedly does not significantly reduce the level of noise.

In the book [9], algorithms for constructing  $\Lambda$ -poised sample sets for regression and minimum Frobenius norm interpolation are given. One can incorporate these algorithms into a model-based derivative-free method of choice and make use of the theory in this paper to provide theoretical convergence guarantees for the resulting optimization method.

**Acknowledgment.** The authors would like to thank C. T. Kelley and Ph. L. Toint for interesting discussions about earlier versions of this paper. We are also grateful to D. Coppersmith for his help with the proof of Lemma 1.1.

## REFERENCES

- [1] *COIN-OR Projects*: <http://www.coin-or.org/projects.html>.

- [2] D. M. BORTZ AND C. T. KELLEY, *The simplex gradient and noisy optimization problems*, in Computational Methods in Optimal Design and Control, Progress in Systems and Control Theory, edited by J. T. Borggaard, J. Burns, E. Cliff, and S. Schreck, vol. 24, Birkhäuser, Boston, 1998, pp. 77–90.
- [3] T. D. CHOI, O. J. ESLINGER, P. GILMORE, A. PATRICK, C. T. KELLEY, AND J. M. GABLONSKY, *IFFCO: Implicit filtering for constrained optimization, Version 2*. 1999.
- [4] P. G. CIARLET AND P. A. RAVIART, *General Lagrange and Hermite interpolation in  $R^n$  with applications to finite element methods*, Arch. Ration. Mech. Anal., 46 (1972), pp. 177–199.
- [5] B. COLSON AND PH. L. TOINT, *Optimizing partially separable functions without derivatives*, Optim. Methods Softw., 20 (2005), pp. 493–508.
- [6] A. R. CONN, N. I. M. GOULD, AND PH. L. TOINT, *Trust-Region Methods*, MPS-SIAM Series on Optimization, SIAM, Philadelphia, 2000.
- [7] A. R. CONN, K. SCHEINBERG, AND PH. L. TOINT, *On the convergence of derivative-free methods for unconstrained optimization*, in Approximation Theory and Optimization, Tributes to M. J. D. Powell, edited by M. D. Buhmann and A. Iserles, Cambridge University Press, Cambridge, 1997, pp. 83–108.
- [8] A. R. CONN, K. SCHEINBERG, AND L. N. VICENTE, *Geometry of interpolation sets in derivative free optimization*, Math. Program., 111 (2008), pp. 141–172.
- [9] ———, *Introduction to Derivative-Free Optimization*, MPS-SIAM Series on Optimization, SIAM, Philadelphia, to appear.
- [10] ———, *Convergence of general derivative-free trust-region algorithms to first and second order critical points*, SIAM J. Optim., (under final review).
- [11] A. L. CUSTÓDIO, J. E. DENNIS JR., AND L. N. VICENTE, *Using simplex gradients of nonsmooth functions in direct search methods*, IMA J. Numer. Anal., (to appear).
- [12] A. L. CUSTÓDIO AND L. N. VICENTE, *Using sampling and simplex derivatives in pattern search methods*, SIAM J. Optim., 18 (2007), pp. 537–555.
- [13] P. J. DAVIS, *Interpolation and Approximation*, Dover Publications, New York, 1975.
- [14] G. DENG AND M. C. FERRIS, *Adaptation of the UOBYQA algorithm for noisy functions*, in Proceedings of the 2006 Winter Simulation Conference, L. F. Perrone, F. P. Wieland, J. Liu, B. G. Lawson, D. M. Nicol, and R. M. Fujimoto, eds., 2006, pp. 312–319.
- [15] M. J. D. POWELL, *On the Lagrange functions of quadratic models that are defined by interpolation*, Optim. Methods Softw., 16 (2001), pp. 289–309.
- [16] ———, *UOBYQA: Unconstrained optimization by quadratic approximation*, Math. Program., 92 (2002), pp. 555–582.
- [17] ———, *Least Frobenius norm updating of quadratic models that satisfy interpolation conditions*, Math. Program., 100 (2004), pp. 183–215.
- [18] ———, *The NEWUOA software for unconstrained optimization without derivatives*, Tech. Report DAMTP 2004/NA08, Department of Applied Mathematics and Theoretical Physics, University of Cambridge, 2004.