Pré-Publicações do Departamento de Matemática
Universidade de Coimbra
Preprint Number 05–15

GEOMETRY OF SAMPLE SETS IN DERIVATIVE FREE OPTIMIZATION. PART II: POLYNOMIAL REGRESSION AND UNDERDETERMINED INTERPOLATION

A. R. CONN, K. SCHEINBERG AND L. N. VICENTE

ABSTRACT: In the recent years, there has been a considerable amount of work in 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 [7] 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 how the notion of Λ–poisedness introduced in [7] to quantify the quality of the sample sets can be extended to the nondetermined cases, by extending first the underlying notion of bases of Lagrange polynomials. We also show that Λ–poisedness is equivalent to a bound on the condition number of the matrix arising from the sampling conditions. We derive bounds for the errors between the function and the (regression and underdetermined) models and between their derivatives.

KEYWORDS: Multivariate polynomial interpolation, polynomial regression and underdetermined interpolation, error estimates, poisedness, derivative free optimization.
AMS SUBJECT CLASSIFICATION (2000): 65D05, 65G99, 90C30, 90C56.

1. Introduction

A class of nonlinear optimization methods called derivative free optimization methods has been extensively developed in the past decade. Some of these methods do not rely on derivative information of the objective function or constraints, but rather approximate these functions using sample models. Some of the popular approaches (see [7] for a more extensive list of references) use polynomial interpolation to build the surrogate model of the objective function (or constraints). In addition, there is both interest and advantages in considering the use of regression models for derivative free optimization.

Date: July 20, 2005.
Support for this work was provided by Centro de Matemática da Universidade de Coimbra and by FCT under grant POCI/59442/2004.

To be able to employ the well developed convergence theory of derivative based methods, the models used in derivative free methods that use trust region or line search frameworks have to satisfy Taylor-like error bounds. Essentially, this requirement is in order to guarantee that when the steps are forced to become small the approximate models are forced to become good.

In [7] we studied the error bounds for the case of polynomial interpolation. These bounds depend on the geometry of the interpolation set. It is important to characterize this geometry and to be able to measure it. Let Y denote the interpolation set. In [7] we discussed two constants that can be used to characterize the geometry, or the so-called *well-poisedness*, of the interpolation set Y. One of these constants is a condition number of a matrix $M(\phi, Y)$ whose entries are determined by evaluating elements of a polynomial basis ϕ at the points in Y. This constant is usually considered to be a bad measure of poisedness, since its value depends on the polynomial basis ϕ that is chosen and since it can be arbitrarily large if Y is scaled by a sufficiently small factor. On the other hand, this measure is convenient to use from an algorithmic perspective, since one can maintain a factorization of $M(\phi, Y)$ to control its condition number.

To fix the drawbacks of this measure, we did the following in [7]: we considered the condition number of the matrix associated with the natural polynomial basis $\bar{\phi}$ (the basis of monomials) and we scaled the set Y to fit exactly into a unit ball. The basis of monomials arises naturally in the algorithmic framework. By using the scaled set Y we ensure that the measure of poisedness is independent of the original scaling. Thus, we have an algorithmically desirable measure of poisedness of Y. To make it theoretically useful, we connected it to the measure of geometry, which appears in most existing Taylor-like bounds for polynomial interpolation.

This measure of geometry is essentially an upper bound on the absolute value of the Lagrange polynomials associated with Y (see [4]). We give the definition of Lagrange polynomials later in the paper. Such a constant is difficult to use in an algorithmic framework for which one wants to prove global convergence. Powell [9] uses the values of Lagrange polynomials to control the poisedness of the interpolation set. However, he does not prove that his method will provide sets with uniformly bounded absolute values of the Lagrange polynomials. In [6], the authors did manage to prove global convergence of their methods based on a similar measure of poisedness but some rather unnatural conditions had to be imposed on the algorithmic framework.

In [7], we also showed that the two measures of poisedness, the Lagrange polynomial bound and the condition number of $M(\bar{\phi}, Y)$, suitably scaled, are interchangeable in some sense. Firstly, we introduced a definition of "Λ–poisedness", where Λ is a well-poisedness constant for Y, which is geometrically intuitive, and it is "equivalent" to the bound on the Lagrange polynomials. Using this definition, we were able to connect this bound with the condition number of the matrix $M(\bar{\phi}, Y)$. Our result showed that these two measures of well-poisedness differ only by a constant factor. We then introduced two algorithms based on the condition number measure of poisedness that were guaranteed to generate a well-poised set. These algorithms are very important in the context of derivative free optimization methods, since to ensure global convergence one needs to maintain a uniformly bounded poisedness constant throughout the algorithm.

In this paper, we extend the results of [7] to the cases of polynomial least squares regression and incomplete or underdetermined polynomial interpolation. We will start by introducing the concept of Lagrange polynomials for least squares regression. As far as we are aware, this generalized definition of Lagrange polynomials does not appear elsewhere in the literature. We will show that the Lagrange polynomials for regression enjoy many properties of the Lagrange polynomials for interpolation. We will extend the geometric definition of Λ-poisedness, which in this case is also equivalent to the bound on Lagrange polynomials for regression. Following results in [7], we will then show the connection between Λ and the condition number of the matrix $M(\bar{\phi}, Y)$. We will conclude the first part of the paper by extending the Taylor-like bounds for linear and quadratic least squares regressions.

In the second part of the paper, we will address the case of incomplete or underdetermined polynomial interpolation. This case is more complicated than complete interpolation, since the choice of the interpolating polynomial is not unique. We will extend the results from interpolation to incomplete interpolation, whenever they are applicable. We will also show the bounds that can be guaranteed for underdetermined interpolation.

A number of the ideas explored in this paper have already been tried in the practical context of derivative free optimization, stressing the need for a more comprehensive theoretical study:

- The DFO code developed in [1] uses minimum norm underdetermined interpolation models in a trust region like method, at early iterations of the method when not enough points are available for complete interpolation. Colson and Toint [5] use a similar strategy in the context of exploiting partial separability in derivative free optimization.

- The approach in [10] intentionally uses incomplete or underdetermined interpolation throughout the course of the optimization algorithm. 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.
- The implicit filtering method [2, 3] is a line search quasi-Newton method based on the negative simplex gradient. Note that the simplex gradient coincides with the gradient of the corresponding interpolation model. Similarly, the Hessian of an interpolation model can be called a simplex Hessian. The most recent implementation of implicit filtering makes use of regression simplex gradients and diagonal simplex Hessians.
- Finally, several types of simplex derivatives (including regression and underdetermined) have been used in the context of direct/pattern search methods to enhance their numerical performance [8]. This requires the identification of Λ–poised sets of sampling points where the function under consideration has been previously evaluated.

The paper is organized as follows. In Section 2, we present the building blocks for regression interpolation, introducing Lagrange polynomials and Λ–poisedness and showing all the corresponding algebraic and geometrical properties, analogous to the case of complete interpolation. The error bounds for regression interpolation are stated in Section 3. A few issues about the usefulness of regression interpolation models and their relation to interpolation models are addressed in Section 4. Section 5 covers the underdetermined case.

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 ℓ_k vector norm or the corresponding matrix norm. By $\|\cdot\|$ (without the subscript) we denote the ℓ_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

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

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 normalized monomial functions

$$\{1, x_1, x_2, \ldots, x_n, x_1^2/2, x_1 x_2, \ldots, x_{n-1}^{d-1} x_n/(d-1), x_n^d/d\}, \tag{1}$$

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 Σ is a diagonal $k \times k$ matrix formed by the (nonnegative) 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 Σ is invertible. Analogously, if $k > \ell$ then the reduced singular value decomposition $M = U\Sigma V^\top$ is such that Σ is a diagonal $\ell \times \ell$ matrix, U is an $\ell \times \ell$ and 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 $Z = \{z^1, \ldots, z^m\} \subset \mathbb{R}^n$, with $m > n$. Let $I \subset \{1, \ldots, m\}$ 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 $\bar{I} = \{1, \ldots, m\} \backslash I$ and let $Z_{\bar{I}}$ be the subset of Z containing points whose indices are in $\bar{I}$. First, we will show that for any z^j, $j \in \bar{I}$,

$$z^j \;=\; \sum_{i \in I} \alpha_i^j z^i, \quad |\alpha_i^j| \;\leq\; 1. \tag{2}$$

By Kramer's rule, (2) holds with $\alpha_i^j = \det(A_{z^j,i}) / \det(A)$, where $A_{z^j,i}$ equals matrix A whose i-th column is replaced by vector z^j. Since by the selection of I, $|\det(A)| \geq |\det(A_{z^j,i})|$ for any $j \in \bar{I}$, then $|\alpha_i^j| \leq 1$.

Now consider any x such that

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

We have

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

■

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 $\phi = \{\phi_0(x), \phi_1(x), \ldots, \phi_q(x)\}$ be a basis for $\mathcal{P}$. This means that ϕ is a set of q_1 polynomials of degree $\leq d$ that span $\mathcal{P}$. Assume we are given a set $Y = \{y^0, y^1, \ldots, y^p\} \subset \mathbb{R}^n$ 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 ϕ is a basis in $\mathcal{P}$, then $m(x) = \sum_{k=0}^{q} \alpha_k \phi_k(x)$, where α_k's are the unknown coefficients. By determining the coefficients $\alpha = [\alpha_0, \ldots, \alpha_q]^\top$ we determine the interpolation polynomial $m(x)$. The coefficients α can be determined from the least squares regression conditions

$$m(y^i) \;=\; \sum_{k=0}^{q} \alpha_k \phi_k(y^i) \overset{\ell.\text{s.}}{=} f(y^i), \quad i = 0, \ldots, p.$$

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

$$M(\phi, Y) = \begin{bmatrix} \phi_0(y^0) & \phi_1(y^0) & \cdots & \phi_q(y^0) \\ \phi_0(y^1) & \phi_1(y^1) & \cdots & \phi_q(y^1) \\ \vdots & \vdots & \vdots & \vdots \\ \phi_0(y^p) & \phi_1(y^p) & \cdots & \phi_q(y^p) \end{bmatrix} \tag{3}$$

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 in the case of interpolation, more generally if $M(\phi, Y)$ has full column rank for some choice of ϕ then this is so for any basis of $\mathcal{P}$. Hence, we will call a set Y poised with respect to the polynomial least squares regression if the appropriate $M(\phi, Y)$ has full column rank for some choice of the basis ϕ.

Let us show now that in the full column rank case the least squares regression polynomial does not depend on the choice of the basis ϕ. Consider a different basis $\psi(x)$ related to $\phi(x)$ by $\phi(x) = P\psi(x)$, where P is $q_1 \times q_1$ and nonsingular. Then $M(\phi, Y) = M(\psi, Y)P^\top$. Let α^ϕ (resp. α^ψ) be the vector of coefficients of the least squares regression polynomial for the basis $\phi(x)$ (resp. $\psi(x)$). To show that the least squares regression polynomial is the same for both bases, it is sufficient to show that $\alpha^\psi = P^\top \alpha^\phi$. Let f_Y denote a vector whose elements are $f(y^i)$, $i = 0, \ldots, p$. Since α^ϕ is the least squares solution to the system $M(\phi, Y)\alpha_i^\phi = f_Y$ then

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

The last equality follows from the fact that α^ψ is the least squares solution to the system $M(\psi, Y)\alpha_i^\psi = f_Y$. We have shown that if Y is poised, then the least squares regression polynomial is unique and independent of the choice of ϕ.

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 function. In [7], 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 (and to underdetermined interpolation).

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 ϕ and, moreover, for any given poised interpolation set Y, one can choose the basis ϕ so that the ratio of the largest over the smallest singular values of $M(\phi, Y)$ can equal anything between 1 and ∞.

The most commonly used measure of poisedness in the multivariate polynomial interpolation literature is based on the basis of Lagrange polynomials. We will briefly describe the concept and its use in polynomial interpolation. Then we will turn to the extension of this concept to the case of least squares regression.

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

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

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 a region of interest. In [4, Theorem 1], it is shown that for any x in the convex hull of Y

$$\|\mathcal{D}^m m(x) - \mathcal{D}^m f(x)\| \;\leq\; \frac{1}{(d+1)!} G \sum_{i=0}^{q} \|y^i - x\|^{d+1} |\mathcal{D}^m \mathcal{L}_i(x)|, \qquad (4)$$

where $\mathcal{D}^m$ denotes the m-th derivative of a function and G is an upper bound on $\mathcal{D}^{d+1} f(x)$. This inequality is a Taylor bound for multivariate polynomial interpolation. Let now

$$\Lambda_Y \;=\; \max_i \; \max_{x \in B_Y(\Delta)} |\mathcal{L}_i(x)|,$$

where i varies in $\{0, \ldots, p\}$ and x in the convex hull of Y. The Taylor bound for function value approximation can be simplified as $(m = 0)$:

$$|m(x) - f(x)| \;\leq\; \frac{1}{(d+1)!} q_1 G \Lambda_Y \Delta^{d+1}, \qquad (5)$$

where Δ is the diameter of the convex hull of Y. See [9] for a simple derivation of this bound. For further discussion see also [7].

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

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

where $f(y^i)$ are the values that are interpolated. We will see that the same is true in the regression case.

Our goal in this paper is to extend the concept of well-poisedness to polynomial least squares regression (and to polynomial underdetermined interpolation). Once we obtain a constant Λ_Y that characterizes well-poisedness of the regression sampling set, we will use it to derive a Taylor bound on the error of the polynomial least squares regression. We will start by extending the notion of Lagrange polynomials.

2.1. Lagrange polynomials for regression. Let $Y = \{y^0, y^1, \ldots, y^p\}$ be the interpolation set, and $\phi = \{\phi_0(x), \phi_1(x), \ldots, \phi_q(x)\}$ 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 $Y = \{y^0, y^1, \ldots, y^p\}$, with $p > q$, a set of $p_1 = p + 1$ polynomials $\mathcal{L}_j(x)$, $j = 0, \ldots, p$, of degree $\leq d$, is called a set of Lagrange regression polynomials if*

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

This set of polynomials is an extension of the traditional Lagrange polynomials to the case when $p > q$. 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

$$\mathcal{L}_j(x) = \sum_{i=0}^{q} (\alpha_j)_i \phi_i(x),$$

where $(\alpha_j)_i$ is the i-th element of the vector α_j.

Then the Lagrange interpolation conditions (in the least squares sense) can be written as

$$M(\phi, Y)\alpha_j \stackrel{\ell.\text{s.}}{=} e_{j+1}, \quad j = 0, \ldots, p,$$

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

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

where α is the matrix whose columns are α_j, $j = 0, \ldots, p$.

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, the polynomials $\mathcal{L}_j(x)$, $j = 0, \ldots, p$, do not depend on the choice of ϕ.

Let $M(\phi, Y) = U_\phi \Sigma_\phi {V_\phi}^\top$ be the reduced singular value decomposition (meaning that U_ϕ is a $p_1 \times q_1$ matrix with orthonormal columns, Σ_ϕ is a $q_1 \times q_1$ matrix of nonzero singular values, and V_ϕ is a $q_1 \times q_1$ orthonormal matrix). 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 $j = 0, \ldots, p$.

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

Lemma 2.1. *Let $Y = \{y^0, y^1, \ldots, y^p\}$ be the 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 at the points in Y. Let $\{\mathcal{L}_j(x),\ j = 0, \ldots, p\}$ 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 γ is not unique. But all we need to show is that $\gamma_i = f(y^i)$, $i = 0, \ldots, p$, is one such solution. We know that $\alpha_j = V_\phi \Sigma_\phi^{-1} {U_\phi}^\top e_{j+1}$, where α_j is the vector of coefficients that expresses $\mathcal{L}_j(x)$ in terms of the basis ϕ. 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 ϕ. Thus,

$$\begin{array}{rcl} V_\phi \Sigma_\phi^{-1} U_\phi{}^\top f_Y & = & \sum_{j=0}^p f(y^j) V_\phi \Sigma_\phi^{-1} U_\phi{}^\top e_{j+1} \\ & = & \sum_{j=0}^p f(y^j)\alpha_j \\ & = & \alpha f_Y, \end{array}$$

and, hence, $m(x) = \sum_{j=0}^p f(y^j)\mathcal{L}_j(x)$. ■

Remark 2.1. *It is interesting to note that the extension of Lagrange polynomials to the case of regression seems to apply only to the case of the least squares regression. We will demonstrate what happens in the case of ℓ_1-norm regression. The case of ℓ_∞-norm regression is similar. 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 other regressions are considered.*

Let us try to extend the definition of Lagrange polynomials to the case of ℓ_1-norm regression.

Attempted Definition 2.1. *Given a set of sample points $Y = \{y^0, y^1, \ldots, y^p\}$, a set of p_1 polynomials $\mathcal{L}_j(x)$, $j = 0, \ldots, p$, of degree $\leq d$, is called a set of Lagrange regression polynomials if*

$$\mathcal{L}_j(y^i) \stackrel{\ell_1}{=} \delta_{ij} \ = \ \begin{cases} 1 & \textit{if} \ \ i = j, \\ 0 & \textit{if} \ \ i \neq j. \end{cases}$$

Under this definition all the attractive properties of Lagrange polynomials fail. For example, one can show that the set of $\mathcal{L}_i$ is not necessarily unique, and that the expression $m(x) = \sum_{i=0}^p f(y^i)\mathcal{L}_i(x)$ may fail to hold for any choice of $\mathcal{L}_i(x)$ $i = 0, \ldots, p$.

We demonstrate this by an example: let us consider the space of linear polynomials in $\mathbb{R}^2$ and the following sample set:

$$Y = \left\{ \begin{bmatrix} 1 \\ 1 \end{bmatrix} \begin{bmatrix} 1 \\ -1 \end{bmatrix} \begin{bmatrix} -1 \\ 1 \end{bmatrix} \begin{bmatrix} -1 \\ -1 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \end{bmatrix} \right\}.$$

The i-th Lagrange polynomial given by Definition 2.1 can be written as $a_i + b_i^1 x_1 + b_i^2 x_2$, where $[a_i, b_i^1, b_i^2]^\top$ minimizes the ℓ_1-norm of the residual of the *following system*

$$\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & -1 \\ 1 & -1 & 1 \\ 1 & -1 & -1 \\ 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} a_i \\ b_i^1 \\ b_i^2 \end{bmatrix} = e_{i+1}. \tag{6}$$

Such solutions are not unique, even though the matrix of the system has a full column rank. For instance, for $i = 0$, $[a_0, b_0^1, b_0^2]^\top = [0, 0.5, 0.5]^\top$ and $[a_0, b_0^1, b_0^2]^\top = [0, 0, 0]^\top$ are both solutions which achieve the minimum value of the ℓ_1-norm of the residual (clearly there is an infinite number of optimal solutions in this case).

It can be shown that any solution to the above system in the least ℓ_1-norm sense, for $i = 0, \ldots, 4$, has a component a_i equal to zero. So, all Lagrange ℓ_1 linear regression polynomials associated with the chosen set Y have a constant term equal to zero. If we now consider the polynomial $m(x) \equiv 1$, whose constant term is equal to 1, it is clear that it cannot be expressed as a linear combination of the polynomials given by Definition 2.1.

2.2. Geometric interpretations of Lagrange regression polynomials. Gi- ven a polynomial basis ϕ, let $\phi(x) = [\phi_0(x), \phi_1(x), \ldots, \phi_q(x)]^\top$ 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 $\mathbb{R}^{q_1}$). Given a poised set $Y = \{y^0, y^1, \ldots, y^p\} \subset B(1) \subset \mathbb{R}^n$, with $p > q$ and $x \in B(1)$, we can express the vector $\phi(x)$ in terms of the vectors $\phi(y^i)$, $i = 0, \ldots, p$, as

$$\sum_{i=0}^{p} \lambda_i(x)\phi(y^i) \;=\; \phi(x), \tag{7}$$

or, equivalently,

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

This system is a simple extension of a similar system introduced in [7] for the case of polynomial interpolation. Unlike the system in [7], this new system is underdetermined, hence it has multiple solutions. In order to establish uniqueness, we will consider the minimum ℓ_2-norm solution.

Lemma 2.2. *Given a poised set Y, the functions $\lambda_i(x)$, $i = 0, \ldots, p$, defined as the minimum ℓ_2-norm solution of (7), form the set of Lagrange regression polynomials for Y given by Definition 2.2.*

Proof: We want to show that $\mathcal{L}_i(x) = \lambda_i(x)$, where $\lambda(x)$ is the minimum ℓ_2-norm solution to (7). We know that $\lambda(x)$ satisfies

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

where $M(\phi, Y)$ is defined by (3), and, in particular, that $\lambda(x)$ is the minimum ℓ_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

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

Now it is clear that $\lambda_j(x)$ is a polynomial in x of the appropriate degree and that its coefficients, expressed through the basis ϕ, are given by the j-th row of $U_\phi \Sigma_\phi^{-1} V_\phi^\top$. Hence, these coefficients are given by the j-th column of $V_\phi \Sigma_\phi^{-1} U_\phi^\top$, as is the case of the j-th Lagrange regression polynomial. We have shown that $\lambda_j(x) = \mathcal{L}_j(x)$ independently of the choice of ϕ. $\blacksquare$

A simple corollary of this result is that $\lambda(x) = [\lambda_0(x), \ldots, \lambda_p(x)]^\top$ does not depend on the choice of ϕ. In [7], the well-poisedness condition for interpolation was introduced via a bound on $\lambda(x)$ and was referred to as Λ–poisedness.

Definition 2.3. *Let $\Lambda > 0$ be given. Let $\phi = \{\phi_0(x), \phi_1(x), \ldots, \phi_p(x)\}$ be a basis in $\mathcal{P}$.*

A set $Y = \{y^0, y^1, \ldots, y^p\}$, with $p = q$, is said to be Λ–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

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

Clearly this definition is equivalent to having all Lagrange polynomials bounded by Λ in $B(1)$ in the ℓ_2-norm. We now introduce the analogous definition for a well-poised regression set.

Definition 2.4. *Let $\Lambda > 0$ be given. Let $\phi = \{\phi_0(x), \phi_1(x), \ldots, \phi_q(x)\}$ be a basis in $\mathcal{P}$.*

A set $Y = \{y^0, y^1, \ldots, y^p\}$, with $p > q$, is said to be Λ–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

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

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 $\|\lambda(x)\| \le \Lambda$ for *some* solution $\lambda(x)$, because then, clearly, the same is true for the minimum norm solution.

Intuitively, Λ is a measure of (the inverse of) the distance to singularity of the set of vectors $\{\phi(y^0), \ldots, \phi(y^p)\}$. See [7] for more discussion.

Notice also that, by definition, Λ is an *upper bound* on poisedness; that is, if Y is $\bar{\Lambda}$–poised, then it is also Λ–poised, for all $\Lambda \ge \bar{\Lambda}$, in other words, Y is *at least* $\bar{\Lambda}$–poised.

2.3. Λ–poisedness and the condition number of $M(\phi, Y)$. 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, for most of this section we make an additional assumption that the radius of this smallest enclosing ball around the origin is one, and we will denote this ball $B(1)$. We will relax this assumption at the end of the section.

We will now show how Λ–poisedness in the regression sense relates to the condition number of the following matrix

$$\bar{M} \;=\; \begin{bmatrix} 1 & y_1^0 & \cdots & y_n^0 & \frac{1}{2}(y_1^0)^2 & y_1^0 y_2^0 & \cdots & \frac{1}{d-1}(y_{n-1}^0)^{d-1} y_n^0 & \frac{1}{d}(y_n^0)^d \\ 1 & y_1^1 & \cdots & y_n^1 & \frac{1}{2}(y_1^1)^2 & y_1^1 y_2^1 & \cdots & \frac{1}{d-1}(y_{n-1}^1)^{d-1} y_n^1 & \frac{1}{d}(y_n^1)^d \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots \\ 1 & y_1^p & \cdots & y_n^p & \frac{1}{2}(y_1^p)^2 & y_1^p y_2^p & \cdots & \frac{1}{d-1}(y_{n-1}^p)^{d-1} y_n^p & \frac{1}{d}(y_n^p)^d \end{bmatrix}, \qquad (8)$$

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

$$\bar{\phi} \;=\; \left\{1, x_1, x_2, \ldots, x_n, x_1^2/2, x_1 x_2, \ldots, x_{n-1}^{d-1} x_n/(d-1), x_n^d/d\right\} \in \mathbb{R}^{q_1} \qquad (9)$$

is the natural basis of monomials. Substituting $\bar{\phi}$ in the definition of Λ–poisedness we can write

$$\bar{M}^\top \lambda(x) \;=\; \bar{\phi}(x) \quad \text{with} \quad \|\lambda(x)\| \le \Lambda. \qquad (10)$$

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

$$1 \leq \|\bar{M}\| \leq p_1^{\frac{1}{2}} q_1. \tag{11}$$

Let us consider the reduced SVD of $\bar{M} = U\Sigma V^\top$, and let σ_1 (resp. σ_{q_1}) denote the absolute value of the largest (resp. smallest) singular value of $\bar{M}$. We omit the dependence of U, Σ, 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 Λ it is, then, sufficient to bound $\|\Sigma^{-1}\|$. Conversely, to bound Λ is terms of $\kappa(\bar{M})$ it is sufficient to bound it in terms of $\|\Sigma^{-1}\|$.

In [7] we showed that the well-poisedness constant Λ from the Definition 2.3 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 [7].

Theorem 2.1. *If Σ is nonsingular and $\|\Sigma^{-1}\| \leq \Lambda$, then the set Y is $\sqrt{q_1}\Lambda$–poised (according to Definition 2.4) in the unit ball $B(1)$ centered at 0. Conversely, if the set Y is Λ–poised, according to Definition 2.4, in the unit ball $B(1)$ centered at 0, then Σ is nonsingular and satisfies*

$$\|\Sigma^{-1}\| \leq \theta\Lambda, \tag{12}$$

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

Proof: The proof is very similar to the proof in [7, Theorem 3.3], but there are a few extra steps. We include the proof for the sake of completeness.

If Σ 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 that the matrix Σ is nonsingular. Let us assume it is singular. By definition of Λ–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 arrived to a contradiction.

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

$$\|\Sigma^{-1}\| \;=\; \|\Sigma^{-1}V^\top\| \;=\; \max_{\|v\|=1} \|\Sigma^{-1}V^\top v\|, \tag{13}$$

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

$$\|\Sigma^{-1}V^\top\| \;=\; \|\Sigma^{-1}V^\top\bar{v}\|, \quad \|\bar{v}\| \;=\; 1. \tag{14}$$

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 Λ–poised we have that

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

and from (13 and 14) 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 be no $x \in B(1)$ 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 (14) there exists an $x \in B(1)$, such that

$$\frac{\|\Sigma^{-1}V^\top\bar{v}\|}{\|\Sigma^{-1}V^\top\bar{\phi}(x)\|} \;\leq\; \theta. \tag{15}$$

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

To show that (15) 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 (14) 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 $\Sigma^{-1}V^\top v^\perp$ 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 (15), concluding the proof of the bound on the norm of Σ^{-1}. ■

The constant θ can be estimated for specific values of d. For $d = 1$ we need to find $\gamma > 0$ such that for any $v \in \mathbb{R}^{n+1}$, with $\|v\| = 1$, there exists an $x \in B(1) \subset \mathbb{R}^n$ such that $\phi(x)^\top v \geq \gamma$, where $\phi(x) = [1, x_1, \ldots, x_n]^\top$. It is easy to see that the choice $x = [v_2, \ldots, v_{n+1}]^\top$ if $v_1 \geq 0$ and $x = [-v_2, \ldots, -v_{n+1}]^\top$ if $v_1 < 0$ guarantees that $\gamma = 1$. Hence $\theta = 1$ for $d = 1$. For $d = 2$ the following lemma holds.

Lemma 2.3. *Let $\bar{v}^\top \bar{\phi}(x)$ be a quadratic polynomial with $\bar{\phi}(x)$ defined by (9) and $\|\bar{v}\|_\infty = 1$, and let $B(1)$ be a (closed) ball of radius 1 centered at the origin. Then*

$$\max_{x \in B(1)} |\bar{v}^\top \bar{\phi}(x)| \; \geq \; \frac{1}{2}.$$

For the proof of the lemma and further discussion see [7, Lemma 3.4].

We can replace the constant θ of Theorem 2.1 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

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

The last inequality is due to Lemma 2.3 applied to the polynomials of the form $\hat{v}^\top \bar{\phi}(x)$. Hence we have

$$\theta \; \leq \; 2q_1^{\frac{1}{2}}. \tag{16}$$

Specifying the bound on θ for polynomials of higher degree is also possible, but is beyond the scope of this paper.

We will now relax the assumption that the enclosing ball has a unit radius. The 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 Λ–poisedness.

Lemma 2.4. *Let $Y = \{y^0, y^1, \ldots, y^p\}$ be an interpolation set and $\{\lambda_i(x)$, $i = 0, \ldots, p\}$ be the set of Lagrange polynomials associated with Y. Then $\{\lambda_i(\Delta\, x)$, $i = 0, \ldots, p\}$ is the set of Lagrange polynomials associated with $\hat{Y}$, where*
$\hat{Y} = \{\Delta\, y^0, \Delta\, y^1, \ldots, \Delta\, y^{p-1}\}$ for any $\Delta > 0$.

Proof: From Lemma 2.2 we know that $\lambda_i(x)$, $i = 0, \ldots, p$, 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 Δ, this corresponds to scaling the above equations by different scalars (1, Δ, Δ^2, etc.). Clearly, $\lambda(\Delta\, x)$ satisfies the scaled system of equations. That implies, again due to Lemma 2.2, that $\lambda_i(\Delta\, x)$, $i = 0, \ldots, p$, is the set of Lagrange polynomials associated with the scaled set. ■

On the contrary, the norm of the inverse of $\bar{M}$ and therefore the condition number $\sigma(\bar{M})$ depend on the scaling of the interpolation set. When we multiply the set Y by Δ, the columns of $\bar{M}$ get multiplied by different scalars (1, Δ, Δ^2, etc.). So, the scaled matrix, say $\hat{M}$, is such that $\|\hat{M}^{-1}\|, \kappa(\hat{M}) \to \infty$ when $\Delta \to 0$. To eliminate the scaling 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 Λ (which is scaling independent), as we have shown in Theorem 2.1.

In the next section, we present the Taylor-like error bounds on linear and quadra- tic least squares regressions. The derivation of these bounds can be found in [7] and is done first in terms of the singular values (Σ^{-1}) of the scaled $\bar{M}^{-1}$, from which one can then plug in either the condition number

$\kappa(\bar{M})$ or the bound Λ on the norm of the Lagrange polynomials given in the Λ–poisedness definition. In this paper, we choose to present these bounds expressed in terms of Λ.

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 Λ. These bounds are extensions of the bounds on polynomial interpolation in [7]. We will present the bounds here without proofs, since they are straightforward extensions of the proofs in [7].

As in [7], 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 Δ around the origin. This assumption is very natural in a DFO 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 Δ 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

$$\bar{M} \;=\; \begin{bmatrix} 1 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 \\ 1 & y_1^1 & \cdots & y_n^1 & \frac{1}{2}(y_1^1)^2 & y_1^1 y_2^1 & \cdots & \frac{1}{d-1}(y_{n-1}^1)^{d-1} y_n^1 & \frac{1}{d}(y_n^1)^d \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots \\ 1 & y_1^p & \cdots & y_n^p & \frac{1}{2}(y_1^p)^2 & y_1^p y_2^p & \cdots & \frac{1}{d-1}(y_{n-1}^p)^{d-1} y_n^p & \frac{1}{d}(y_n^p)^d \end{bmatrix}. \tag{17}$$

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

$$m(x) \;=\; c + g^\top x \;=\; c + \sum_{k=1}^{n} g_k x_k. \tag{18}$$

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

Theorem 3.1. *Let $Y = \{0, y^1, \ldots, y^p\}$ be a Λ–poised set of $p > n$ regression points contained in a (closed) ball $B(\Delta)$ centered at 0. Assume that f is continuously differentiable in an open domain Ω containing $B(\Delta)$ and that ∇f is Lipschitz continuous in Ω 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*

$$\|e^g(x)\| \;\leq\; (5p_1^{\frac{1}{2}}\gamma_L\Lambda/2)\,\Delta, \tag{19}$$

- *the error between the fully linear interpolation 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 $Y = \{0, y^1, \ldots, y^p\}$ of $p_1 > (n+1)(n+2)/2$ sample points ($p_1 = p+1$) in a ball $B(\Delta)$ of radius Δ centered at the origin. In addition we will assume that f is twice continuously differentiable in an open domain Ω containing this ball and that $\nabla^2 f$ is Lipschitz continuous in Ω with constant $\gamma_Q > 0$.

It is possible to build the quadratic regression model

$$m(x) \;=\; c + g^\top x + \frac{1}{2}x^\top H x \;=\; c + \sum_{1\leq k\leq n} g_k x_k + \frac{1}{2}\sum_{1\leq k,\ell\leq n} h_{k\ell}x_k x_\ell, \tag{20}$$

where H is a symmetric matrix of order n.

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

Theorem 3.2. *Let $Y = \{0, y^1, \ldots, y^p\}$, with $p_1 > (n+1)(n+2)/2$ and $p_1 = p+1$, be a Λ–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 Ω containing $B(\Delta)$ and that $\nabla^2 f$ is Lipschitz continuous in Ω 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 α_Q^H, α_Q^g, α_Q^f, and β_Q^f are small positive constants dependent on $d = 2$ and independent of n and Y:

$$\alpha_Q^H = \frac{3\sqrt{2}}{2}, \quad \alpha_Q^g = \frac{3(1+\sqrt{2})}{2}, \quad \alpha_Q^f = \frac{6+9\sqrt{2}}{2}, \quad \beta_Q^f = \frac{1}{6}.$$

The above error bounds can be extended to the case of polynomials of higher degrees. However, in the context of derivative free methods, on which we are focusing, the linear and the quadratic cases are 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 Λ-poised, with Λ uniformly bounded, throughout the progress of the algorithm.

In [7] two examples of algorithms that guarantee Λ-poisedness interpolation sets are proposed. Each algorithm either verifies that the current interpolation set is Λ-poised for some given value of Λ, or if it is not, replaces the "bad" points with new points to maintain a Λ-poised interpolation set. It is shown that as long as Λ 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, Λ-poisedness of a set Y of q_1 points implies Λ-poisedness, in the regression sense, of any larger superset of Y.

It would be interesting to investigate any algorithms that target the maintenance of the regression set directly as it is done in [7], rather than maintaining a good subset that is well-poised in the interpolation sense. Using the algorithms in [7] we aim to select a square submatrix of matrix $\bar{M}$ with the best condition number. Instead it would be better to select the set of points that will increase the smallest singular value of the whole matrix $\bar{M}$. However, it is not clear at this point how to design such an algorithm in the nonlinear case.

Given the tools described above one can create a globally convergent algorithm by taking virtually any globally convergent method based on Taylor models. It suffices to replace the Taylor models by least squares polynomial regression models and to maintain well-poised regression sets by any method that guarantees Λ–poisedness.

4. Regression vs. interpolation

One can relate Λ–poisedness in the regression sense to Λ–poisedness in the interpolation sense, as it is shown in the next theorem.

Theorem 4.1. *Given a set $Y = \{y^0, y^1, \ldots, y^p\}$, which is Λ–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 Λ–poised in the interpolation sense, then the set $Y = \{y^0, y^1, \ldots, y^p\}$, is Λ–poised in the regression sense.

Proof: The first implication follows from the Definitions 2.3 and 2.4 for Λ–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. ∎

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? Based on our theory, the answer to this question seems to be "neither". On the one hand, it is clear that the poisedness constant Λ, as defined by Definition 2.4, 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 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, $(x_i, y_i) \in \mathbb{R}^2$, $i = 1, \ldots, p$, in the *unit radius* square centered at the origin ($\{x \in \mathbb{R}^2 : \|x\|_\infty \leq 1\}$). Together with the origin $(0,0)$, we have $p_1 = p + 1$ points in $\mathbb{R}^2$.

In the linear case ($q = 2$), we considered all possible pairs of points $(x_i, y_i), (x_j, y_j)$, $i, j = 1, \ldots, p$, $i \neq j$, and selected the pair with the best condition number of 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 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 of the matrix $\bar{M}$. (We again worked with the $p \times q$ submatrix of $\bar{M}$ obtained by removing its first row and column.) We, then, built a quadratic interpolation model of the function for this sample set (called also 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 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 ℓ_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 ≤ 0.001 no win was declared.

We report the results corresponding to the quadratic case in Figure 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 ℓ_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 ℓ_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

#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

FIGURE 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.

of sample sets) converge to each other (pointwise) as the number of sample points tends to infinity.

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 ϕ. Then the matrix

$$M(\phi, Y) = \begin{bmatrix} \phi_0(y^0) & \phi_1(y^0) & \cdots & \phi_q(y^0) \\ \phi_0(y^1) & \phi_1(y^1) & \cdots & \phi_q(y^1) \\ \vdots & \vdots & \vdots & \vdots \\ \phi_0(y^p) & \phi_1(y^p) & \cdots & \phi_q(y^p) \end{bmatrix} \tag{21}$$

has more columns than rows. The interpolation polynomials defined by

$$m(y^i) = \sum_{k=0}^{q} \alpha_k \phi_k(y^i) = f(y^i), \quad i = 0, \ldots, p. \tag{22}$$

are no longer unique, even if M has full row rank.

The simplest approach to restrict the system so that it has a unique solution is to remove the last $q-p$ columns of $M(\phi, Y)$ from the system. This causes the last $q-p$ elements of the solution α to be zero. Such an approach approximates some elements of α, while it sets others to zero solely based on the order of the elements in the basis ϕ. 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 ϕ to mean a subset of p_1 elements of the basis ϕ. 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 the advantage can be taken trivially by deleting appropriate columns from the system (22). If it is not 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 $Y = \{y^0, y^1, \ldots, y^p\} \subset B(1)$, with $p < q$, is said to be Λ–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

$$\sum_{i=0}^{p} \lambda_i(x) \tilde{\phi}(y^i) \;=\; \tilde{\phi}(x) \tag{23}$$

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

It is easy to show (as it is done for the complete interpolation case in [7]) 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

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

The approach to select a unique solution to (22) should then be the following. Given the sample set Y, select the sub-basis $\tilde{\phi}(x)$ so that the poisedness constant Λ 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 (22) 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

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

Notice that the two possible solutions are very different from each other, yet as ϵ goes to zero the two sets of points converge pointwise to each other. Hence, we see that the sub-basis approach suffers from lack of robustness with respect to small perturbation 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 ($b_2/(1+\epsilon)$). 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 (22).

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 ϵ converges to zero. Hence, not only the minimum norm solution is robust with respect to small perturbation of the data, but also it "evens out" the elements of the gradient over the x_2 and x_3 basis components.

For the reasons described above we are interested in looking at the minimum norm solution of the system (22). The minimum norm solution is expressed as

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

It is well know that a minimum norm solution of an underdetermined system of linear equation is not invariant under linear transformation. In our case, this fact means that the minimum norm solution depends on the choice of ϕ. It is easy to show that the resulting interpolation polynomial also depends on the choice of ϕ in the system (22). Consider the following example:

Example 5.3. $Y = \{y^0, y^1, y^2\}$, $y^0 = [0, 0]^\top$, $y^1 = [1, 0]^\top$, $y^2 = [0, 1]^\top$, and $f = x_1^2 + x_2^2$. Let us consider two bases:

$$\phi \;=\; \left\{1, x_1, x_2, \frac{1}{2}x_1^2, x_1 x_2, \frac{1}{2}x_2^2\right\}$$

and

$$\psi \;=\; \left\{1, x_1 + x_2, x_1 - x_2, \frac{1}{2}(x_1 + x_2)^2, x_1 x_2, \frac{1}{2}(x_1 - x_2)^2\right\}.$$

Note that $f_Y = [0, 1, 1]^\top$. The systems under consideration are

$$M(\phi, Y)\alpha_\phi = f_Y$$

and

$$M(\psi, Y)\alpha_\psi \;=\; f_Y.$$

The minimum norm solution of the first system is $\alpha_\phi = [0, \frac{4}{5}, \frac{4}{5}, \frac{2}{5}, 0, \frac{2}{5}]^\top$ and the resulting polynomial

$$m_\phi(x) \;=\; \frac{4}{5}x_1 + \frac{4}{5}x_2 + \frac{1}{5}x_1^2 + \frac{1}{5}x_2^2.$$

The minimum norm solution of the second system is $\alpha_\psi = [0, \frac{2}{3}, 0, \frac{1}{3}, 0, \frac{1}{3}]^\top$ and the resulting polynomial

$$m_\psi(x) \;=\; \frac{2}{3}x_1 + \frac{2}{3}x_2 + \frac{1}{3}x_1^2 + \frac{1}{3}x_2^2.$$

The maximum of $|m_\phi(x)|$ over $B(1)$ is $\frac{1+4\sqrt{2}}{5} \simeq 1.33$ and the maximum of $|m_\psi(x)|$ over $B(1)$ is $\frac{1+2\sqrt{2}}{3} \simeq 1.28$. Furthermore,

$$f(x) - m_\phi(x) = -\frac{4}{5}x_1 - \frac{4}{5}x_2 + \frac{4}{5}x_1^2 + \frac{4}{5}x_2^2.$$

The maximum of $|f(x) - m_\phi(x)|$ over $B(1)$ is attained at the point $x = [-1/\sqrt{2}, -1/\sqrt{2}]^\top$ and equals $\frac{4}{5}(\sqrt{2}+1) \approx 1.93$. Also,

$$f(x) - m_\psi(x) = -\frac{2}{3}x_1 - \frac{2}{3}x_2 + \frac{2}{3}x_1^2 + \frac{2}{3}x_2^2.$$

The maximum of $|f(x) - m_\psi(x)|$ over $B(1)$ is attained at the point $x = [-1/\sqrt{2}, -1/\sqrt{2}]^\top$ and equals $\frac{2}{3}(\sqrt{2}+1) \simeq 1.61$.

This example implies that depending on the choice of ϕ we can obtain a better or a worse approximation of f by computing the minimum norm interpolating polynomials. Ideally, we would like, for each set Y, to identify the "best" basis ϕ, which would generate the "best" minimum norm interpolating polynomial. 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 w.r.t. 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 the 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 of the basis.

5.1. Lagrange polynomials and Λ-poisedness for underdetermined interpolation. We will consider the natural basis $\bar{\phi}$ defined by (9) and the corresponding matrix $\bar{M} = M(\bar{\phi}, Y)$ defined by (21). 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 $Y = \{y^0, y^1, \ldots, y^p\}$, with $p < q$, a set of $p_1 = p+1$ polynomials $\mathcal{L}_j(x) = \sum_{i=0}^q (\alpha_j)_i \bar{\phi}_i(x)$, $j = 0, \ldots, p$, is called a set of Lagrange minimum norm polynomials for the basis $\bar{\phi}$ if it is* *a minimum norm solution of*

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

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

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

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 in this case the Lagrange polynomials generally depend on the choice of the basis ϕ, 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 5.1. *Let $Y = \{y^0, y^1, \ldots, y^p\}$ be the set of 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 $\{\mathcal{L}_i(x), i = 0, \ldots, p\}$ 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: We know that $\alpha_j = \bar{M}^\top(\bar{M}\bar{M}^\top)^{-1}e_{j+1}$, where α_j is the vector of coefficients that express $\mathcal{L}_j(x)$ in terms of the basis $\bar{\phi}$. We also know that

$$\gamma = \bar{M}^\top(\bar{M}\bar{M}^\top)^{-1}f_Y$$

is the vector of coefficients that express $m(x)$ in terms of the basis $\bar{\phi}$. It is simple to see that $\gamma = \sum_{j=0}^{p} \alpha_j f(y^j)$, and hence, $m(x) = \sum_{j=0}^{p} \mathcal{L}_j(x) f(y^j)$. ∎

Let $\bar{M} = U\Sigma V^\top$ be a reduced singular value decomposition of $\bar{M}$ (meaning that U is a $p_1 \times p_1$ orthonormal matrix, Σ is a $p_1 \times p_1$ matrix of nonzero singular values, and V is a $q_1 \times p_1$ matrix with orthonormal columns).

We will now show that, as in the case of polynomial interpolation [7] and regression, the geometric interpretations of Lagrange polynomial can be easily derived. Thus, we will also have an analogous definition of Λ–poisedness.

Given a poised set $Y = \{y^0, y^1, \ldots, y^p\} \subset B(1) \subset \mathbb{R}^n$ and $x \in B(1)$ we attempt to express vector $\bar{\phi}(x)$ in terms of the vectors $\bar{\phi}(y^i)$, $i = 0, \ldots, p$. Since the dimension of vector $\bar{\phi}(x)$ is $q_1 > p_1$, it may no longer be possible to express it in terms of p_1 vectors $\bar{\phi}(y^i)$, $i = 0, \ldots, p$. Hence, we will be looking for the least squares solution to the following system

$$\sum_{i=0}^{p} \lambda_i(x)\bar{\phi}(y^i) \overset{\ell.\mathrm{s.}}{=} \bar{\phi}(x). \tag{24}$$

This system is an extension of the similar systems introduced in [7] and of system (7) in Section 2.2. Unlike system (7), where the minimum ℓ_2-norm solution $\lambda(x)$ to the system corresponded to the least squares regression Lagrange polynomials, in this case the least squares solution $\lambda(x)$ corresponds to the minimum ℓ_2-norm Lagrange polynomials.

Lemma 5.2. *Given a poised set Y, the functions $\lambda_i(x)$, $i = 0, \ldots, p$, defined by the least squares solution of (24), form the set of minimum norm Lagrange polynomials for Y given in Definition 5.2.*

Proof: The minimum norm solution α_j can be expressed as

$$\alpha_j \;=\; V\Sigma^{-1}U^\top e_{j+1}.$$

Hence, the vector of the coefficients of the j-th Lagrange polynomial, when expressed through the basis $\bar{\phi}(x)$, is just the $j+1$-th column of $V\Sigma^{-1}U^\top$.

Now consider the expression for $\lambda_j(x)$. We know that $\lambda(x)$ satisfies

$$\bar{M}^\top \lambda(x) \overset{\ell.\mathrm{s.}}{=} \bar{\phi}(x).$$

Hence, given the reduced singular value decomposition of $\bar{M}^\top = V\Sigma U^\top$,

$$\lambda(x) \;=\; U\Sigma^{-1}V^\top\bar{\phi}(x).$$

Now it is clear that $\lambda_j(x)$ is a polynomial in x of the appropriate degree and it is expressed through the basis $\bar{\phi}$ with a vector of coefficients equal to the

$j+1$-th row of $U\Sigma^{-1}V^\top$, which is the same as $j+1$-th column of $V\Sigma^{-1}U^\top$. We have shown that $\lambda_j(x) = \mathcal{L}_j(x)$, $j = 0, \ldots, p$. ■

The following definition of well-poisedness is analogous to Definitions 2.3 and 2.4.

Definition 5.3. *Let $\Lambda > 0$ be given. Let $\bar{\phi}$ be the natural basis of monomials.*

A set $Y = \{y^0, y^1, \ldots, y^p\}$, with $p < q$, is said to be Λ–poised in $B(1)$ (in the minimum norm sense) if and only if for any $x \in B(1)$ there exists an 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 Lagrange polynomials bounded by Λ in $B(1)$.

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 5.1. *There exists a constant θ independent of Λ and Y such that if a set Y is Λ–poised in the sub-basis sense, then it is $\sqrt{p_1 q_1}\theta\Lambda$–poised in the sense of Definition 5.3.*

Conversely, if a set $Y = \{y^0, y^1, \ldots, y^p\} \subset B(1)$ is Λ–poised by Definition 5.3, 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 Λ–poised in the sub-basis sense, and that $\tilde{\phi}$ is the sub-basis. Let $\tilde{\mathcal{L}}_i(x) = (\gamma^i)^\top \tilde{\phi}(x)$, $i = 0, \ldots, p$, be the Lagrange polynomials for the sub-basis $\tilde{\phi}$ (see Definition 5.1 and the paragraph following). Then $\max_i \max_{x \in B(1)} \tilde{\mathcal{L}}_i(x) \leq \Lambda$. As it is shown in the proof of Theorem 2.1, there exists a constant σ, independent of Y and Λ, 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 $\mathcal{L}_i(x) = (\alpha^i)^\top \bar{\phi}(x)$, $i = 0, \ldots, p$. The vector α^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 $j = 0, \ldots, q$ 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 Λ–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

$$\bar{M}_{p_1}\gamma^i \;=\; e_{i+1}, \quad |\gamma^i_j| \;\leq\; (q-p+1)|\alpha^i_j|, \quad j = 0,\ldots,p.$$

The selected columns determine a sub-basis $\tilde{\phi}$ and the vector of coefficients γ^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 σ, independent of Y and Λ, such that $\max_{x\in B(1)}(\alpha^i)^\top\bar{\phi}(x) \geq \sigma\|\alpha^i\|$ for each i. Hence, $\theta\Lambda \geq \|\alpha^i\|$, for each i and $\theta = 1/\sigma$. On the other hand,

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

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 θ for the specific cases of linear and quadratic interpolations are discussed after the proof of Theorem 2.1.)

■

Remark 5.1. *It is easy to see that the results of this subsection hold for any given basis ϕ, as long as it remains fixed throughout the discussion. (Note that the constants in Theorem 5.1 vary with the choice of ϕ.) Hence, the minimum norm Lagrange polynomials can be defined for any basis. The definition of Λ–poisedness also can be introduced for any basis. However, for any given set Y, one can create different bases, which, when used in Definition 5.3, will result in different constants for Λ–poisedness of Y. Moreover, by varying the basis ϕ, one can make the constant Λ as large or as close to 1 as desired. Clearly for the definition of Λ–poisedness to make sense, the Λ constant should be related to the quality of the geometry Y and the resulting interpolation. Hence, we consider only one basis (the basis $\bar{\phi}$).*

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) Λ-poisedness of a set Y in terms of $\bar{\phi}$ implies $\mathcal{O}(\Lambda)$–poisedness of Y in terms of any other basis ψ, such that $\bar{\phi} = P\psi$ and $\|P\|\|P^{-1}\| = \mathcal{O}(1)$ (the last statement is easy to show from the definition of Λ–poisedness and from Theorem 5.2 of the next section).

In the next section, we will use the properties of $\bar{\phi}$ to show the relation between the poisedness constant Λ 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* [10]*, 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 [10] *also proposed a definition of Lagrange polynomials for underdetermined interpolation. In the notation of our paper, the definition in* [10] *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.*

To ensure the existence and uniqueness of the Lagrange polynomials, the definition in [10] *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.*

5.2. Λ–poisedness and the condition number of $M(\bar{\phi}, M)$. We make again the assumption that the radius of the smallest ball around the origin enclosing Y is one, and we will denote this ball by $B(1)$. We will again relax this assumption at the end of the section.

Recall that under this assumption $1 \leq \|\bar{M}\| \leq p_1^{1/2} q_1$. Hence to bound the condition number of $\bar{M}$ in terms of Λ (and vice versa) all we need to do is to bound $\|\bar{M}^{-1}\|$ in terms of Λ (and vice versa). We will now present the analogue of Theorem 2.1 of this paper and Theorem 3.3 of [7] for the underdetermined case. Recall the reduced singular value decomposition of $\bar{M} = U \Sigma V^\top$.

Theorem 5.2. *If Σ is nonsingular and $\|\Sigma^{-1}\| \leq \Lambda$, then the set Y is $\sqrt{q_1}\Lambda$–poised (according to Definition 5.3) in the unit ball $B(1)$ centered at 0. Conversely, if the set Y is Λ–poised, according to Definition 5.3, in the unit ball $B(1)$ centered at 0, then Σ is nonsingular and*

$$\|\Sigma^{-1}\| \;\leq\; \theta\Lambda, \tag{25}$$

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

Proof: As in the proof of Theorem 2.1, it is trivial to show that if Σ 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 Σ is nonsingular by the definition of Λ–poisedness. To prove that there exists a constant $\theta > 0$, independent of Y and of Λ, such that $\|\Sigma^{-1}\| \leq \theta\Lambda$, we would proceed exactly as in the proof of Theorem 2.1. ∎

For obtaining the specific values of θ in linear and quadratic cases we can apply the results presented after Theorem 2.1. In particular, θ can be equal to one in linear interpolation and it satisfies (16) for quadratic interpolation. To relax the assumption that the radius of the ball enclosing Y is 1, we can use the same arguments as in the end of Subsection 2.3.

5.3. Error bounds for underdetermined interpolation. We start by studying 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 Δ has the smallest possible value that satisfies $Y \subset B(\Delta)$ and $y^0 = 0$. The derivation of the general error bound follows strictly the derivation in [7] 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 $\bar{M}_{p\times q}$

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

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

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

Theorem 5.3. *Let $Y = \{0, y^1, \ldots, y^p\}$, with $p_1 < (n + 1)(n + 2)/2$ and $p_1 = p+1$, be a Λ–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 Ω containing $B(\Delta)$ and that $\nabla^2 f$ is Lipschitz continuous in Ω 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

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

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

Proof: We apply the same algebraic manipulations and Taylor expansions to the interpolating conditions (22) as we did in [7]. We leave them out of this proof for the sake of brevity. As in [7], we obtain the following system

$$\bar{M}_{p\times q} \left[\begin{array}{cc} D_\Delta^{-1} & 0 \\ 0 & D_{\Delta^2}^{-1} \end{array} \right] \left[\begin{array}{c} D_\Delta t(x) \\ D_{\Delta^2} e^H(x) \end{array} \right] \;=\; \mathcal{O}(\Delta^3). \qquad (26)$$

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

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

■

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

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

where $m(x) = c + g^\top x + \frac{1}{2} x^\top H x$, $\nabla m(x) = Hx + g$, 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 $\bar{M}_{p\times q}$.

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 [7]).

It is also possible to show that if ∇f and $\nabla^2 f$ have specific sparsity structure that corresponds to a certain sub-basis for which Y is Λ–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 the repetition of the corresponding error bound result in [7].

Let us investigate the quality of a minimum norm interpolation model when the interpolation set Y is Λ–poised in the minimum norm sense and contains a subset of $n+1$ points that are Λ_L–poised for linear interpolation. We will show that the minimum norm model is at least as good, in order of accuracy, as the linear interpolation model based on these $n+1$ points.

Let us assume that Y is Λ–poised in minimum norm sense. We can write the scaled version of the first n rows of $\bar{M}_{p\times q}$ as

$$\hat{M}_{n\times q} \;=\; \left[\; \hat{M}^{lin}_{n\times n} \quad \hat{M}^{qua}_{n\times r} \;\right],$$

where $\hat{M}^{lin}_{n\times n}$ is an $n\times n$ matrix corresponding to the linear terms in $\bar{\phi}$, and $\hat{M}^{qua}_{n\times r}$ is the matrix consisting of the other $r = q-n$ columns (both referring to the first $n+1$ points). Let us assume (w.l.o.g.) that the subset $\{0, y^1, \ldots, y^n\}$ is Λ_L–poised for linear interpolation. From Theorem 2.1 (restricted to the case $p = q = n$), this fact implies that $\|(\hat{M}^{lin}_{n\times n})^{-1}\| \le \Lambda_L$ ($\theta = 1$ in the linear case). First, we show the following lemma.

Lemma 5.3. *Let $Y = \{0, y^1, \ldots, y^p\}$, with $p_1 < (n+1)(n+2)/2$ and $p_1 = p+1$, be a Λ–poised set of points (in the minimum norm sense) contained in a (closed) ball $B(\Delta)$ centered at 0. Let us assume also that $\{0, y^1, \ldots, y^n\}$ is Λ_L–poised for linear interpolation.*

Assume that f is twice continuously differentiable in an open domain Ω containing $B(\Delta)$ and that $\nabla^2 f$ is Lipschitz continuous in Ω with constant $\gamma_Q > 0$. Let β_f and β_H be upper bounds on the norm of f_Y and on the norm of the Hessian of f in $B(\Delta)$.

Then, the error between the Hessian of the minimum norm model and the Hessian of the function is bounded by

$$\|e^H(x)\| \;\le\; 2q_1^{\frac{1}{2}}\beta_f\Lambda + \beta_H.$$

Proof: We know that

$$\|e^H(x)\| \;\leq\; \|\mathrm{svec}(H)\| + \|\mathrm{svec}(\nabla^2 f(x))\| \;\leq\; \left\| \begin{bmatrix} Hx+g \\ \mathrm{svec}(H) \end{bmatrix} \right\| + \|\nabla^2 f(x)\|.$$

Since the $Hx+g$ and $\mathrm{svec}(H)$ are the components of the minimum norm solution, we have that

$$\|e^H(x)\| \;\leq\; \left\| V_{q\times p} \Sigma_{p\times p}^{-1} U_{p\times p}^\top f_Y \right\| + \|\nabla^2 f(x)\|.$$

From Theorem 5.2 and (16), we obtain

$$\|e^H(x)\| \;\leq\; \|\Sigma_{p\times p}^{-1}\| \|f_Y\| + \|\nabla^2 f(x)\| \;\leq\; 2 q_1^{\frac{1}{2}} \Lambda \beta_f + \beta_H.$$

■

Finally, we show the following result.

Theorem 5.4. *Let* $Y = \{0, y^1, \ldots, y^p\}$, *with* $p_1 < (n+1)(n+2)/2$ *and* $p_1 = p+1$, *be a* Λ*–poised set of points (in the minimum norm sense) contained in a (closed) ball* $B(\Delta)$ *centered at* 0. *Let us assume also that* $\{0, y^1, \ldots, y^n\}$ *is* Λ_L*–poised for linear interpolation.*

Assume that f *is twice continuously differentiable in an open domain* Ω *containing* $B(\Delta)$ *and that* $\nabla^2 f$ *is Lipschitz continuous in* Ω *with constant* $\gamma_Q > 0$. *Let* β_f *and* β_H *be upper bounds on the norm of* f_Y *and on the norm of the Hessian of* f *in* $B(\Delta)$.

Then, the error between the gradient of the minimum norm model and the gradient of the function satisfies

$$\|e^g(x)\| \;\leq\; n^{\frac{1}{2}} r \Lambda_L \|e^H(x)\|\, \Delta + \frac{3}{2} n^{\frac{1}{2}} \gamma_Q \Lambda_L\, \Delta^2 + \sqrt{2}\|e^H(x)\|\Delta,$$

where

$$\|e^H(x)\| \;\leq\; 2 q_1^{\frac{1}{2}} \beta_f \Lambda + \beta_H.$$

Proof: We can write the first n rows of (26) as

$$\hat{M}_{n\times n}^{lin} D_\Delta t(x) + \hat{M}_{n\times r}^{qua} D_{\Delta^2} e^H(x) \;=\; \mathcal{O}(\Delta^3),$$

where the right-hand-side is bounded in norm by $3\sqrt{n}\gamma_Q \Delta^3/2$. With the purpose of bounding $t(x)$, we write

$$D_\Delta t(x) \;=\; -\left(\hat{M}_{n\times n}^{lin}\right)^{-1} \hat{M}_{n\times r}^{qua} D_{\Delta^2} e^H(x) + \left(\hat{M}_{n\times n}^{lin}\right)^{-1} \mathcal{O}(\Delta^3)$$

and point out that

$$\left\| \left(\hat{M}^{lin}_{n\times n} \right)^{-1} \hat{M}^{qua}_{n\times r} \right\| \;\leq\; \Lambda_L n^{\frac{1}{2}} r.$$

with $r = q - n$. Thus,

$$\|D_\Delta t(x)\| \;\leq\; n^{\frac{1}{2}} r \Lambda_L \|e^H(x)\|\, \Delta^2 + \frac{3}{2} n^{\frac{1}{2}} \gamma_Q \Lambda_L \, \Delta^3$$

or, equivalently,

$$\|t(x)\| \;\leq\; n^{\frac{1}{2}} r \Lambda_L \|e^H(x)\|\, \Delta + \frac{3}{2} n^{\frac{1}{2}} \gamma_Q \Lambda_L \, \Delta^2.$$

We can now use the bound on $\|e^H(x)\|$ in Lemma 5.3 to conclude the proof of the theorem. ■

We remark that the poisedness constant in this bound appears "squared" ($\Lambda\Lambda_L$) and, as a result, the bound for the gradient of the minimum norm model is worse than the bound for linear interpolation. However, as we will argue in a moment, in practice the situation is usually the opposite.

5.4. 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 function for this sub-basis. We call this model the best basis (BB) model. (Here we also 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 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 significantly better (> 0.001) than the other in *both* errors. When one of the

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

FIGURE 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.

#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

FIGURE 3. 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.

errors was ≤ 0.001 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 2 are followed by the results for a scaled square (scaled to have a *radius* of 0.01) in Figure 3. 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.

Acknowledgement

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 greatful 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., (2005, to appear).

[6] 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.

[7] A. R. Conn, K. Scheinberg, and L. N. Vicente, *Geometry of sample sets in derivative free optimization. Part I: polynomial interpolation*, Tech. Report 03-09, Departamento de Matemática, Universidade de Coimbra, Portugal, 2003. Revised September 2004.

[8] A. L. Custódio and L. N. Vicente, *Using sampling and simplex derivatives in pattern search methods*, Tech. Report 04-35, Departamento de Matemática, Universidade de Coimbra, Portugal, 2004.

[9] M. J. D. Powell, *On the Lagrange functions of quadratic models that are defined by interpolation*, Optim. Methods Softw., 16 (2001), pp. 289–309.

[10] ——, *Least Frobenius norm updating of quadratic models that satisfy interpolation conditions*, Math. Program., 100 (2005), pp. 183–215.

A. R. Conn
Department of Mathematical Sciences, IBM T.J. Watson Research Center, Route 134, P.O. Box 218, Yorktown Heights, New York 10598, USA (arconn@watson.ibm.com).

K. Scheinberg
Department of Mathematical Sciences, IBM T.J. Watson Research Center, Route 134, P.O. Box 218, Yorktown Heights, New York 10598, USA (arconn@watson.ibm.com).

L. N. Vicente
Departamento de Matemática, Universidade de Coimbra, 3001-454 Coimbra, Portugal (lnv@mat.uc.pt).