GEOMETRY OF INTERPOLATION SETS IN DERIVATIVE FREE OPTIMIZATION

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

Abstract. We consider derivative free methods based on sampling approaches for nonlinear optimization problems where derivatives of the objective function are not available and cannot be directly approximated.

We show how the bounds on the error between an interpolating polynomial and the true function can be used in the convergence theory of derivative free sampling methods. These bounds involve a constant that reflects the quality of the interpolation set. The main task of such a derivative free algorithm is to maintain an interpolation sampling set so that this constant remains small, and at least uniformly bounded. This constant is often described through the basis of Lagrange polynomials associated with the interpolation set. We provide an alternative, more intuitive, definition for this concept and show how this constant is related to the condition number of a certain matrix. This relation enables us to provide a range of algorithms whilst maintaining the interpolation set so that this condition number or the geometry constant remain uniformly bounded. We also derive bounds on the error between the model and the function and between their derivatives, directly in terms of this condition number and of this geometry constant.

Key words. Multivariate Polynomial Interpolation, Error Estimates, Poisedness, Derivative Free Optimization.

AMS subject classifications. 65D05, 65G99, 90C30, 90C56

1. Introduction. Derivative free optimization (DFO) is an area of nonlinear optimization that deals with problems where the derivatives of the objective function (and, possibly, constraints) are not available. Often, legacy code makes it undesirable to introduce the necessary modifications to obtain the derivatives. There are practical instances where the evaluation of the function is the result of an experiment, resulting in the absence of derivatives.

It is also assumed that the derivatives cannot be approximated directly, either because they are too expensive or because the presence of noise in the function evaluations makes such approximation inaccurate. Among the most popular such methods are direct and pattern search ([2], [12], [19], [21]). DFO sampling methods include interpolation based trust-region methods ([7], [8], [14], [15], [16], [18]) and implicit filtering ([3], [11]). A number of surveys have been written on DFO methods ([5] [6, Chapter 9], [8], [12], [16], [22]).

Like their derivative based counterparts most of the derivative free sampling methods for unconstrained optimization build a linear or a quadratic model of the objective function and apply a trust region or a line search step to find the next iterate. While derivative based methods typically use a Taylor based model, or an approximation thereof, as the model of the objective function, these derivative free methods use interpolation, regression, or other sample based models. In order to be able to use the

extensive convergence theory that exists for derivative based methods, these derivative free methods need to satisfy properties similar to those of the former methods. For instance, it is necessary to show that as the trust region or line search step becomes small, the error in the gradient approximation approaches zero, and, hence, a descent step will eventually be found, unless the gradient of the true function is zero at the current iterate. One way to show this is to prove that the linear or quadratic approximation models satisfy Taylor-like error bounds on the function value and the gradient. (Pattern search methods are directional methods and do not use models of the function, but the main convergence issues related to these methods are similar.)

The purpose of this paper is to provide the necessary theoretical foundation for various DFO sampling methods to guarantee such bounds. We will focus on polynomial interpolation models because of their simplicity and because polynomials appear naturally in Taylor based models. In an upcoming paper, we will address polynomial regression and underdetermined polynomial interpolation models. In the current paper, we describe some of the already existing results from the multivariate polynomial interpolation literature, relate them to the DFO context, provide additional intuition and show how these results can be used in a DFO framework. In addition, we will show how the results can be used consistently in an algorithmic framework.

Consider a set of interpolation points given by

$$Y \;=\; \left\{ y^0, y^1, \ldots, y^p \right\},$$

where $p_1 = p + 1 = |Y|$ is a positive integer defining the number of points in the interpolation set. Let $m(x)$ denote an interpolating polynomial of degree d satisfying the interpolation conditions

(1.1) $$m(y^i) \;=\; f(y^i), \; i = 0, \ldots, p.$$

Typically, $p_1 = p + 1$ is the dimension of the space of polynomials of degree less than or equal to d:

$$p_1 \;=\; p + 1 \;=\; \begin{pmatrix} n + d \\ n \end{pmatrix}.$$

In the multivariate polynomial interpolation literature there exist several Taylor type bounds on the error between $m(x)$ and $f(x)$ and their derivatives. One of the most commonly used ones can be found in [4, Theorem 1]. In particular, for the error in the gradient such a bound can be simplified to

$$\|\nabla m(x) - \nabla f(x)\| \;\leq\; \frac{1}{(d+1)!} G \, \Lambda_Y \sum_{i=0}^{p} \|y^i - x\|^{d+1},$$

for x in the convex hull of Y. Here G is a constant that depends only on the function f, and Λ_Y is a constant that depends on the interpolation set Y. To guarantee convergence of a DFO method, we need to ensure that the error in the gradient converges to zero when the distances between all y^i and x go to zero. Also, the rate of convergence of this error should be preserved to equal the rate of convergence of $\sum_{i=0}^{p} \|y^i - x\|^{d+1}$. For that purpose, Λ_Y has to remain uniformly bounded for all sets Y used by the algorithm.

The interpolation set Y gets updated whenever a new iterate is found (and sometimes more frequently than that). Each update has to be designed so that the constant

Λ_Y associated with the set Y remains uniformly bounded and, if possible in certain situations, is reduced. Thus, it is crucial to understand the nature of the dependence of the constant Λ_Y on Y.

The classical way to describe Λ_Y is as an upper bound on the absolute value of a certain set of polynomials, called Lagrange polynomials, that form a basis in the space of polynomials of degree $\le d$. Powell [17] uses this definition to suggest an approach of maintaining the set Y by choosing points to enter Y or to leave Y so that the bound on the maximum absolute value of the Lagrange polynomials is, hopefully, reduced. However, he does not provide any bounds on the resulting polynomials. In [7] and [8], a somewhat different definition of Λ_Y is used. It is derived from the bound on the value of the so called Newton polynomials, which also form a basis in the space of polynomials of degree $\le d$. A mechanism for maintaining a uniform bound on the Newton polynomials is described in [7]. This method requires maintaining the set of Newton polynomials in each iteration, just as Powell's method requires maintaining the set of Lagrange polynomials. The drawbacks of these approaches are discussed in Section 5.

In this paper, we develop a theoretical framework that guarantees a uniform bound on the constant Λ_Y, under various approaches to maintaining the interpolation set Y.

We will show how the constant Λ_Y is related to the condition number of a certain multivariate extension of a Vandermonde matrix. Hence, we will provide a connection between the bound on Lagrange polynomials and this condition number. This connection is important because it enables us to relate our results to the classical bounds on polynomial interpolation such as described in [4]. We will derive Taylor-like bounds for linear and quadratic interpolation expressed via both the condition number and the bound on the Lagrange polynomials. It is often simpler to design algorithms that operate with the condition number of a matrix rather than directly with a bound on the Lagrange polynomials.

We provide examples of such algorithms for linear and quadratic interpolation. These algorithms can verify if a given interpolation set Y satisfies the condition needed to keep Λ_Y uniformly bounded and if not, modify this set so that the condition is satisfied. The theory in this paper provides a guarantee of the Taylor-like bounds if such an algorithm (or a new algorithm of a similar nature) is used in any interpolation based DFO method. The Taylor-like bound can then be used to develop a global convergence theory for such a method.

The paper is organized as follows. In Section 2 we introduce the notion of a poised interpolation set, discuss the definition of Lagrange polynomials and their use in the existing approximation bounds. In Section 3 we present a geometric interpretation of Lagrange polynomials (introducing the notion of Λ–poisedness). We then establish a connection between the bound on Lagrange polynomials for a given interpolation set Y and the bound on the condition number of a (multivariate extension of a) Vandermonde matrix. This matrix arises from the natural basis of normalized monomials and from Y. In Section 4, we derive the Taylor-like bounds, both in terms of this condition number and in terms of an upper bound on the Lagrange polynomials. (In the Appendix of this paper, we briefly sketch how this derivation extends to higher order instances using the cubic case.) Finally, in Section 5, we present the algorithms for maintaining the set Y and the bound Λ_Y.

1.1. Basic facts and notation. Here we introduce some further 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 $B(\Delta) = \{x \in \mathrm{I\!R}^m : \|x\| \leq \Delta\}$ to denote the closed ball in $\mathrm{I\!R}^m$ of radius $\Delta > 0$ centered at the origin (where m is inferred from the particular context). By $B_Y(\Delta)$ we will denote the smallest ball enclosing a given set Y, not necessarily centered at the origin (we indicate Δ, the radius of this ball, although, it is uniquely defined by Y). We use several properties of norms. In particular, given an $m \times n$ 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 $\alpha(x)$ and $\beta(x)$ 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 $\mathrm{I\!R}^n$ we will mean the following basis of normalized monomial functions

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

2. Polynomial interpolation and poisedness. Let us consider $\mathcal{P}$, the space of polynomials of degree $\leq d$ in $\mathrm{I\!R}^n$. Let $p_1 = p + 1$ be the dimension of this space (e.g., for $d = 1$, $p_1 = n + 1$ and for $d = 2$, $p_1 = (n + 1)(n + 2)/2$) and let $\phi = \{\phi_0(x), \phi_1(x), \ldots, \phi_p(x)\}$ be a basis in $\mathcal{P}$. This means that ϕ is a set of p_1 polynomials of degree $\leq d$ that span $\mathcal{P}$. Assume we are given a set $Y = \{y^0, y^1, \ldots, y^p\} \subset \mathrm{I\!R}^n$ of interpolation points, and let $m(x)$ denote a polynomial of degree $\leq d$ that interpolates a given function $f(x)$ at the points in Y. Since ϕ is a basis in $\mathcal{P}$, then $m(x) = \sum_{j=0}^p \alpha_j \phi_j(x)$, where α_j's are some coefficients. By determining the coefficients $\alpha_0, \ldots, \alpha_p$ we determine the interpolation polynomial $m(x)$. The coefficients $\alpha_0, \ldots, \alpha_p$ can be determined from the interpolation conditions

$$m(y^i) = \sum_{j=0}^p \alpha_j \phi_j(y^i) = f(y^i), \ i = 0, \ldots, p.$$

This is a linear system in terms of the interpolating coefficients. For the above system to have a unique solution, the matrix of the system

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

has to be nonsingular. It is easy to see that if $M(\phi, Y)$ is nonsingular for some basis ϕ then it is nonsingular for any basis of $\mathcal{P}$. In that case the set Y is said to be *poised* (or d-*unisolvent*, see [4]).

It is natural to ask if the condition number of $M(\phi, Y)$ is an indication of how *well-poised* the set Y is. The answer, in general, is "no", since the condition number of $M(\phi, Y)$ depends on the choice of ϕ, moreover, for any given poised interpolation set Y, one can choose the basis ϕ so that the condition number of $M(\phi, Y)$ can equal anything between 1 and $+\infty$. Hence, the condition number of $M(\phi, Y)$ is not considered to be a good characteristic of poisedness of a set of points. However, we will return to this issue later in the paper and show that for a specific choice of ϕ, namely for the natural basis, it is helpful to use the condition number of $M(\phi, \hat{Y})$ as a constant associated with the well-poisedness of Y, where $\hat{Y}$ is a scaled version of Y.

4

2.1. Lagrange Polynomials. The most commonly used measure of poisedness in multivariate polynomial interpolation literature is based on the basis of Lagrange polynomials.

DEFINITION 2.1. *Given a set of interpolation points* $Y = \{y^0, y^1, \ldots, y^p\}$*, 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*

$$\mathcal{L}_j(y^i) \;=\; \delta_{ij} \;=\; \left\{ \begin{array}{ll} 1 & if \;\; i = j, \\ 0 & if \;\; i \neq j. \end{array} \right.$$

For any poised set Y there exists a unique basis of Lagrange polynomials. The classical measure of poisedness of Y is the 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

$$(2.1) \qquad \|\mathcal{D}^m m(x) - \mathcal{D}^m f(x)\| \;\leq\; \frac{1}{(d+1)!} G \sum_{i=0}^{p} \|y^i - x\|^{d+1} |\mathcal{D}^m \mathcal{L}_i(x)|,$$

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 is a Taylor bound for multivariate polynomial interpolation. Let now

$$\Lambda_Y \;=\; \max_i \max_x |\mathcal{L}_i(x)|,$$

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

$$(2.2) \qquad\qquad |m(x) - f(x)| \;\leq\; \frac{1}{(d+1)!} p_1 G \Lambda_Y \Delta^{d+1},$$

where Δ is the diameter of the convex hull of Y. See also [17] for a simple derivation of this bound. The bounds derived in [4] for the approximation of the derivatives can also be expressed in terms of Λ_Y (as we will show later in Subsection 4.4).

In this paper we will (somewhat loosely) say that a set Y is *well-poised* if the associated Λ_Y is bounded by some reasonable predefined constant. As we described in the introduction, it is crucial for the convergence of a derivative free sampling method to maintain the interpolation set Y so that the poisedness constant Λ_Y remains uniformly bounded, but the actual size of this bound is not important for theoretical purposes. However, for the efficiency of such a derivative free method it is important to be able to relatively easily improve Λ_Y by replacing some point in Y, if necessary, and to be able to readily recognize this necessity. Powell addresses these issues in [16] and [17]. He suggests replacing the i-th point in Y by another point x in the convex hull of Y so that the absolute value of the associated $\mathcal{L}_i(x)$ is maximized. However, he does not provide a proof that this procedure will guarantee a uniform bound on Λ_Y.

It can be shown, as it is done in [7] for a similar approach using Newton polynomials [20], that one can construct interpolation sets $Y \in B_Y(\Delta)$ so that Λ_Y is uniformly bounded for all Δ. The proofs for construction of such sets (like the one in [7]) have to rely on the particular basis of polynomials that is used to produce the bound Λ_Y and the same polynomials have to be used in the maintenance of well-poised sets Y from iteration to iteration. We would like to provide a framework that will fit various

approaches to maintaining Y. Having to replace the interpolation set Y completely to ensure well-poisedness (which might happen in [7]) is very unattractive algorithmically. We would like to show that well-poised sets can be maintained throughout the algorithm by imposing some threshold rule for accepting new interpolation points. To do so we need a good intuition on what it means for the set Y to have the values of the associated set of Lagrange polynomials bounded by a given Λ_Y in $B_Y(\Delta)$.

Ciarlet and Raviart in [4, Theorem 2] provide some geometric intuition. They split the constant Λ_Y into a product of two constants. One is the "stretchedness" coefficient for the convex hull of Y and the other involves the size of the Lagrange polynomials for a chosen set Y', which is obtained from Y by an affine transformation. This split provides a nice interpretation for the well-poisedness of a set, but still involves the bound of Lagrange polynomials, and therefore it is still not clear how to use the second constant algorithmically. Moreover, the "stretchedness" coefficient can be a misleading indicator of poisedness since there exist nearly nonpoised sets with small "stretchedness" coefficient. For example, a set of six points in $\mathbb{R}^2$ lying on a circle is nonpoised for quadratic interpolation, hence small perturbations of such set are nearly nonpoised sets with small "stretchedness" coefficient.

In the next section we provide what we feel is a good intuition (with some geometric perspective) for Lagrange polynomials and the bound Λ_Y. Then we will show how Λ_Y relates to the condition number of $M(\phi, Y)$, where ϕ is the natural basis.

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 by $B(1)$. We will relax this assumption at the end of the section, considering a ball $B(\Delta)$ of radius $\Delta > 0$. By working with smallest enclosing balls rather than convex hulls, we adjust the presentation to our DFO motivation.

Given a polynomial basis ϕ, let $\phi(x) = [\phi_0(x), \phi_1(x), \ldots, \phi_p(x)]^\top$ be a vector in $\mathbb{R}^{p_1}$ whose entries are the values of the elements of the polynomial basis at x (one can view ϕ as a mapping from $\mathbb{R}^n$ to $\mathbb{R}^{p_1}$). Given a poised set $Y = \{y^0, y^1, \ldots, y^p\} \subset B(1) \subset \mathbb{R}^n$ and an $x \in B(1)$, we can express the vector $\phi(x)$ in terms of the vectors $\phi(y^i)$, $i = 0, \ldots, p$, as

$$(3.1) \qquad \sum_{i=0}^{p} \lambda_i(x)\phi(y^i) \; = \; \phi(x).$$

or

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

LEMMA 3.1. *Given a poised set Y, $\{\lambda_i(x), i = 0, \ldots, p\}$ defined by (3.1) is the set of Lagrange polynomials for Y.*

Proof. It is trivial to see that $\lambda_j(y^i) = \delta_{ij}$, $\forall i, j = 0, \ldots, p$. The fact that $\lambda_i(x)$ is a polynomial of degree $\leq d$ can be observed from the fact that $\lambda_i(x)$ is a linear function of $\phi(x)$. Hence, $\lambda_i(x)$ is the i-th Lagrange polynomial for the set Y. □

It is interesting to note as an immediate corollary that $\lambda(x) = [\lambda_0(x), \ldots, \lambda_p(x)]^\top$ does not depend on the choice of ϕ. However, the geometric interpretation of $\lambda(x)$

involves ϕ: $\lambda(x)$ is the vector of coefficients for which $\phi(x)$ can be expressed in terms of $\phi(y^i)$, $i = 0, \ldots, p$. The upper bound on $\lambda(x)$ is then related to the maximum possible coefficient that may be required to express $\phi(x)$ in terms of $\phi(y^i)$, for some $x \in B(1)$. If the set of vectors $\{\phi(y^i), i = 0, \ldots, p\}$ spans $\mathbb{R}^{p_1}$ well, then the bound on $\lambda(x)$ is relatively small. We will use the following definition to indicate the degree of poisedness of a set.

DEFINITION 3.2. *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 $Y = \{y^0, y^1, \ldots, y^p\}$ *is said to be* Λ–*poised in* $B(1)$ *if and only if for any* $x \in B(1)$ *there exists* $\lambda(x) \in \mathbb{R}^{p+1}$:

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

Assume that, for a given Λ, the set $Y \subset B(1)$ is not Λ–poised. Then there exists a $z \in B(1)$ such that

$$\sum_{i=0}^{p} \lambda_i(z)\phi(y^i) = \phi(z), \text{ and } \|\lambda(z)\|_\infty \ge \Lambda/p_1^{\frac{1}{2}},$$

hence, w.l.o.g., $\lambda_1(z) > \Lambda/p_1^{\frac{1}{2}}$. Then, dividing this expression by Λ we have

$$\sum_{i=0}^{p} \frac{\lambda_i(z)}{\Lambda}\phi(y^i) = \sum_{i=0}^{p} \alpha_i(z)\phi(y^i) = \frac{\phi(z)}{\Lambda}, \text{ and } \alpha_1(z) > 1/p_1^{\frac{1}{2}}.$$

Hence,

$$\|\sum_{i=0}^{p} \alpha_i(z)\phi(y^i)\|_\infty \le \frac{\max_{x \in B(1)} \|\phi(x)\|_\infty}{\Lambda}.$$

If, for example, ϕ is the natural basis, then $\max_{x \in B(1)} \|\phi(x)\|_\infty \le 1$ and

$$\|\sum_{i=0}^{p} \alpha_i(z)\phi(y^i)\|_\infty \le \frac{1}{\Lambda} \text{ with } \alpha_1(z) > 1/p_1^{\frac{1}{2}}.$$

It is easy to see now that $\frac{1}{\Lambda}$ bounds, in some sense, the distance to linear dependency of the vectors $\phi(y^i)$, $i = 0, \ldots, p$. As Λ grows the system represented by these vectors becomes increasingly linearly dependent. But the actual distance to singularity depends on the choice of ϕ. This brings us back to the discussion of the condition number of $M(\phi, Y)$.

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

$$(3.2) \quad M = \begin{bmatrix} 1 & y_1^0 & \cdots & y_n^0 & \frac{1}{2}(y_1^0)^2 & y_1^0 y_2^0 & \cdots & \frac{1}{(d-1)!}(y_{n-1}^0)^{d-1}y_n^0 & \frac{1}{d!}(y_n^0)^d \\ 1 & y_1^1 & \cdots & y_n^1 & \frac{1}{2}(y_1^1)^2 & y_1^1 y_2^1 & \cdots & \frac{1}{(d-1)!}(y_{n-1}^1)^{d-1}y_n^1 & \frac{1}{d!}(y_n^1)^d \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots \\ 1 & y_1^p & \cdots & y_n^p & \frac{1}{2}(y_1^p)^2 & y_1^p y_2^p & \cdots & \frac{1}{(d-1)!}(y_{n-1}^p)^{d-1}y_n^p & \frac{1}{d!}(y_n^p)^d \end{bmatrix}.$$

This matrix is exactly the matrix $M(\bar{\phi}, Y)$ for

(3.3) $\qquad \bar{\phi} = \{1, x_1, x_2, \ldots, x_n, x_1^2/2, x_1 x_2, \ldots, x_{n-1}^{d-1} x_n/(d-1)!, x_n^d/d!\},$

the natural basis of monomials. Moreover, if we use the natural basis in the definition of Λ–poisedness we can write it as

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

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

(3.5) $\qquad\qquad\qquad\qquad\qquad 1 \leq \|M\| \leq p_1^{\frac{3}{2}}.$

The condition number of M is denoted by $\kappa(M) = \|M\|\|M^{-1}\|$. To bound $\kappa(M)$ in terms of Λ it is, then, sufficient to bound $\|M^{-1}\|$, and conversely, to bound Λ in terms of $\kappa(M)$ it is sufficient to bound it in terms of $\|M^{-1}\|$.

THEOREM 3.3. *If M is nonsingular and $\|M^{-1}\| \leq \Lambda$, then the set Y is $\sqrt{p_1}\Lambda$–poised in the unit ball $B(1)$ centered at 0. Conversely, if the set Y is Λ–poised in the unit ball $B(1)$ centered at 0, then M is nonsingular and*

(3.6) $\qquad\qquad\qquad\qquad\qquad \|M^{-1}\| \leq \theta\Lambda,$

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

Proof. Since the ℓ_2 norm is invariant under transposition, we can use $M^\top$ in the proof. If M is nonsingular and $\|M^{-1}\| \leq \Lambda$ then

$$\|\lambda(x)\| \leq \|M^{-\top}\|\|\bar{\phi}(x)\| \leq p_1^{\frac{1}{2}}\|M^{-\top}\|\|\bar{\phi}(x)\|_\infty \leq p_1^{\frac{1}{2}}\Lambda,$$

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

Proving the reverse relation is more complicated. First let us show that the matrix M 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 $M^\top$. This means that there exists a vector $v \neq 0$ in the null space of 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 at a contradiction.

Now we want to show that there exists a constant $\theta > 0$, independent of Y and of Λ, such that $\|M^{-\top}\| \leq \theta\Lambda$. From the definition of the matrix norm,

$$\|M^{-\top}\| = \max_{\|v\|=1} \|M^{-\top}v\|,$$

we can consider a vector $\bar{v}$ such that

(3.7) $\qquad\qquad\qquad \|M^{-\top}\| = \|M^{-\top}\bar{v}\|, \quad \|\bar{v}\| = 1.$

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

$$\|M^{-\top}\bar{v}\| = \|M^{-\top}\bar{\phi}(x)\| \leq \Lambda,$$

8

and from (3.7) the statement of the theorem holds with $\theta = 1$.

Notice that $\bar{v}$ does not necessarily belong to the image of $\bar{\phi}$, which means that there might be no $x \in B(1)$ such that $\bar{\phi}(x) = \bar{v}$, and hence we have that $\|M^{-\top}\bar{v}\| \neq \|M^{-\top}\bar{\phi}(x)\|$. However, we will show that there exists a constant $\theta > 0$ such that for any $\bar{v}$ which satisfies (3.7) there exists an $y \in B(1)$, such that

$$(3.8) \qquad \frac{\|M^{-\top}\bar{v}\|}{\|M^{-\top}\bar{\phi}(y)\|} \leq \theta.$$

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

To show that (3.8) holds, we first show that there exists $\sigma > 0$ such that for any $\bar{v}$ with $\|\bar{v}\| = 1$, there exists an $y \in B(1)$ such that $|\bar{v}^\top \bar{\phi}(y)| \geq \sigma$. 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) $\sigma > 0$ such that $\psi(\bar{v}) \geq \sigma\|\bar{v}\| = \sigma$. Hence, there exists a $y \in B(1)$ such that $|\bar{v}^\top \bar{\phi}(y)| \geq \sigma$.

Let $\bar{v}^\perp$ be the orthogonal projection of $\bar{\phi}(y)$ onto the subspace orthogonal to $\bar{v}$. Now, notice that from the definition (3.7) of $\bar{v}$ it follows that $\bar{v}$ is the right singular vector corresponding to the largest singular value of $M^{-\top}$. Then $M^{-\top}\bar{v}$ and $M^{-\top}\bar{v}^\perp$ are orthogonal vectors (since $M^{-\top}\bar{v}$ is a scaled left singular vector corresponding to the largest singular value and $M^{-\top}\bar{v}^\perp$ is a vector spanned by the other left singular vectors). Since $\|\bar{v}\| = 1$, $\bar{\phi}(y) = \bar{v}^\perp + (\bar{v}^\top \bar{\phi}(y))\bar{v}$. And from the orthogonality of $M^{-\top}\bar{v}^\perp$ and $M^{-\top}\bar{v}$

$$\|M^{-\top}\bar{\phi}(y)\| = \|M^{-\top}\bar{v}^\perp\| + |\bar{v}^\top \bar{\phi}(y)|\|M^{-\top}\bar{v}\|.$$

Hence $\|M^{-\top}\bar{\phi}(y)\| \geq |\bar{v}^\top \bar{\phi}(y)|\|M^{-\top}\bar{v}\|$. It follows from $|\bar{v}^\top \bar{\phi}(y)| \geq \sigma$ that

$$\|M^{-\top}\bar{\phi}(y)\| \geq \sigma\|M^{-\top}\bar{v}\|,$$

Assigning $\theta = 1/\sigma$ shows (3.8), concluding the proof of the bound on the norm of $M^{-\top}$. ∎

The constant θ can be estimated for specific values of d. For $d = 1$ we need to find $\sigma > 0$ such that for any $\bar{v} \in \mathbb{R}^{n+1}$, with $\|\bar{v}\| = 1$, there exists an $x \in B(1) \subset \mathbb{R}^n$ such that $|\bar{v}^\top \bar{\phi}(x)| \geq \sigma$, where $\bar{\phi}(x) = [1, x_1, \ldots, x_n]^\top$. Let $w = [\bar{v}_2, \ldots, \bar{v}_{n+1}]^\top$. It is easy to see that the optimal solution of problem $\max_{x \in B(1)} |\bar{v}^\top \bar{\phi}(x)|$ is either given by $w/\|w\|$ (with optimal value $\bar{v}_1 + \|w\|$) or by $-w/\|w\|$ (with optimal value $-\bar{v}_1 + \|w\|$). Thus, the optimal value is $|\bar{v}_1| + \|w\|$. Thus, since $\|\bar{v}\| = 1$, we can guarantee that $\sigma \geq 1$. Hence $\theta \leq 1$ for $d = 1$.

For $d = 2$ we will need the following lemma.

LEMMA 3.4. *Let $\hat{v}^\top \bar{\phi}(x)$ be a quadratic polynomial with $\bar{\phi}(x)$ defined by (3.3) and $\|\hat{v}\|_\infty = 1$ and let $B(1)$ be a (closed) ball of radius 1 centered at the origin. Then*

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

Proof. Since $\|\hat{v}\|_\infty = 1$ then at least one of the elements of $\hat{v}$ is 1 or -1, and thus one of the coefficients of the polynomial $q(x) = \hat{v}^\top \bar{\phi}(x)$, is equal to 1, -1, $1/2$, or $-1/2$. Let us consider only the cases where one of the coefficients of $q(x)$ is 1 or $1/2$. The cases -1 or $-1/2$ would be analyzed similarly.

The largest coefficient in absolute value in $\hat{v}$ corresponds to a term which is either a constant term, a linear term x_i, or a quadratic term $x_i^2/2$ or $x_i x_j$. Let us restrict all variables that do not appear in this term to zero. And let us consider only the unrestricted variables. Clearly the maximum of the absolute value of $q(x)$ over the set of unrestricted variables is a lower bound on the maximum over $B(1)$. We can have three cases.

- $q(x) = 1$. This case is trivial.
- $q(x) = x_i^2/2 + \alpha x_i + \beta$. It is easy to see that

$$\max_{x_i \in [-1,1]} |q(x)| = \max\{|q(-1)|, |q(0)|, |q(1)|\} \geq \frac{1}{4}.$$

- $q(x) = \alpha x_i^2/2 + x_i + \beta$. In this case we have

$$\max_{x_i \in [-1,1]} |q(x)| \geq \max\{|q(-1)|, |q(1)|\} \geq 1.$$

- $q(x) = \alpha x_i^2/2 + \beta x_j^2/2 + x_i x_j + \gamma x_i + \delta x_j + \epsilon$. This time we are considering the quadratic function over a two dimensional ball. By considering four points, $p_1 = (1/\sqrt{2}, 1/\sqrt{2})$, $p_2 = (1/\sqrt{2}, -1/\sqrt{2})$, $p_3 = (-1/\sqrt{2}, 1/\sqrt{2})$, and $p_4 = (-1/\sqrt{2}, -1/\sqrt{2})$, on the boundary of the ball, and looking at all the possible signs of $\alpha + \beta$, $\gamma + \delta$, and $\gamma - \delta$, we get

$$\max\{|q(0)|, |q(p_i)|, \ i = 1,2,3,4\} \geq \frac{1}{4}.$$

□

We can replace the constant θ of Theorem 3.3 by an upper bound, which is easily derived for the quadratic case. Recall that $\theta = 1/\sigma$, where

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

Given any $\bar{v}$ such that $\|\bar{v}\| = 1$, we can scale $\bar{v}$ by at most $\sqrt{p_1}$ to $\hat{v} = \alpha \bar{v}$, $0 < \alpha \leq \sqrt{p_1}$, such that $\|\hat{v}\|_\infty = 1$. Then

$$\sigma = \min_{\|\bar{v}\|=1} \max_{x \in B(1)} |\bar{v}^\top \bar{\phi}(x)| \geq \frac{1}{\sqrt{p_1}} \min_{\|\hat{v}\|_\infty=1} \max_{x \in B(1)} |\hat{v}^\top \bar{\phi}(x)| \geq \frac{1}{4\sqrt{p_1}}.$$

The last inequality is due to Lemma 3.4 applied to the polynomials of the form $\hat{v}^\top \bar{\phi}(x)$. Hence we have

$$(3.9) \qquad \theta \leq 4 p_1^{\frac{1}{2}}.$$

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

REMARK 3.1. *It is important to note that θ depends on the choice of $\bar{\phi}(\cdot)$. For example, if we scale every element of $\bar{\phi}(\cdot)$ by 2 then the appropriate θ will decrease by*

2. Here we are interested in the condition number of a specific matrix M arising in our algorithms and, hence, in a specific choice of $\bar{\phi}(\cdot)$.

We will now relax the assumption that the radius of the enclosing ball is 1. The attractive property of Lagrange polynomials is that they remain invariant under the scaling of set Y. A simple proof can be derived from our interpretation of Lagrange polynomials given in the definition of Λ–poisedness.

LEMMA 3.5. *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 $\hat{Y} = \{\Delta y^0, \Delta y^1, \ldots, \Delta y^p\}$ for any $\Delta > 0$.*

Proof. From Lemma 3.1 we know that $\lambda_i(x)$, $i = 0, \ldots, p$, satisfy

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

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 3.1, that $\lambda_i(\Delta x)$, $i = 0, \ldots, p$, is the set of Lagrange polynomials associated with the scaled set. $\square$

The norm of the inverse of M and therefore the condition number $\kappa(M)$, though, do depend on the scaling of the interpolation set. Multiplying points in the set Y by Δ translates to multiplying the columns in M by different scalars (1, Δ, Δ^2, etc.), this means that for the scaled matrix $\hat{M}$, we have that $\|\hat{M}^{-1}\|, \kappa(\hat{M}) \to \infty$ when $\Delta \to 0$.

Given a set $Y \subset B(\Delta)$ we will scale Y by $1/\Delta$ to obtain $\hat{Y} \subset B(1)$. The condition number of the corresponding matrix $\hat{M}$ will be used as a measure of poisedness and as a geometric constant in the Taylor-like bounds presented in the next section. The error bounds are derived first in terms of the norm of the scaled inverse $\hat{M}^{-1}$, from which we can then use either the condition number $\kappa(\hat{M})$ or the bound Λ on the norm of the Lagrange polynomials given in the Λ–poisedness definition.

4. Error bounds. In this section we present a derivation of Taylor-like bounds for linear and quadratic interpolation. Similar bounds, as we mentioned earlier, can be found, for example, in [4]. But the derivation of our bounds is done in terms of the inverse of the scaled matrix $\hat{M}$, described in the previous section, rather than in terms of the bound on Lagrange polynomials. Also the proof that we present is, in our opinion, very simple and intuitive and should be useful in furthering the understanding of the nature of such bounds. The bounds are then written in terms of the condition number of the scaled matrix $\hat{M}$ as well as in terms of the bound on Lagrange polynomials (what we call the Λ–poisedness constant).

For the remainder of the paper we will make an additional assumption 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 M can be written now as

$$(4.1) \quad M \;=\; \begin{bmatrix} 1 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 \\ 1 & y_1^1 & \cdots & y_n^1 & \frac{1}{2}(y_1^1)^2 & y_1^1 y_2^1 & \cdots & \frac{1}{(d-1)!}(y_{n-1}^1)^{d-1} y_n^1 & \frac{1}{d!}(y_n^1)^d \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots \\ 1 & y_1^p & \cdots & y_n^p & \frac{1}{2}(y_1^p)^2 & y_1^p y_2^p & \cdots & \frac{1}{(d-1)!}(y_{n-1}^p)^{d-1} y_n^p & \frac{1}{d!}(y_n^p)^d \end{bmatrix}.$$

The error bounds will derived in terms of the submatrix $M_{p \times p}$ formed by the last p rows and columns of M,

$$M \;=\; \begin{bmatrix} 1 & 0 \\ e & M_{p \times p} \end{bmatrix},$$

in particular, in terms of the inverse of its scaled version $\hat{M}_{p \times p}$.

4.1. Error estimates in the linear case. We now consider interpolation of a function $f(x)$ by a linear polynomial $m(x)$:

$$(4.2) \quad m(x) \;=\; c + g^\top x \;=\; c + \sum_{k=1}^{n} g_k x_k.$$

Assume that f is continuously differentiable in an open domain Ω containing $B(\Delta)$ and that ∇f is Lipschitz continuous in Ω with constant $\gamma_L > 0$.

The interpolation set satisfies $Y = \{y^0, y^1, \dots, y^n\} = \{0, y^1, \dots, y^n\} \subset B(\Delta)$, where $B(\Delta)$ is a ball of radius Δ centered at the origin. The unknown coefficients $c, g_1, \dots, g_n$ are defined by the linear system arising from the interpolating conditions (1.1).

We are interested in the error between the function and the polynomial and between their gradients at any given x in the ball $B(\Delta)$. We can write

$$(4.3) \quad m(x) \;=\; f(x) + e^f(x)$$

and

$$(4.4) \quad \nabla m(x) \;=\; g \;=\; \nabla f(x) + e^g(x).$$

Subtracting (4.3) from each of the $n+1$ equalities in (1.1) and using (4.2) yields

$$\begin{aligned} (-x)^\top g &= f(0) - f(x) - e^f(x), \\ (y^i - x)^\top g &= f(y^i) - f(x) - e^f(x), \quad i = 1, \dots, n. \end{aligned}$$

Expanding f by a Taylor's formula of order one around x for all the interpolation points, we obtain from (4.4)

$$(4.5) \quad \begin{aligned} & (-x)^\top e^g(x) \\ &= \int_0^1 \left(\nabla f(x - tx) - \nabla f(x) \right)^\top (-x) dt - e^f(x) \\ &= \mathcal{O}(\Delta^2) - e^f(x), \end{aligned}$$

$$(4.6) \quad \begin{aligned} & (y^i - x)^\top e^g(x) \\ &= \int_0^1 \left(\nabla f(x + t(y^i - x)) - \nabla f(x) \right)^\top (y^i - x) dt - e^f(x) \\ &= \mathcal{O}(\Delta^2) - e^f(x), \quad i = 1, \dots, n. \end{aligned}$$

Now we subtract the first equation from the rest, canceling $e^f(x)$ and obtaining

$$(y^i)^\top e^g(x) = \mathcal{O}(\Delta^2), \quad 1 \le i \le n,$$

or, using matrix notation,

(4.7) $$L_{n \times n} e^g(x) = \mathcal{O}(\Delta^2),$$

where

$$L_{n \times n} = \begin{bmatrix} (y^1)^\top \\ \vdots \\ (y^n)^\top \end{bmatrix} = \begin{bmatrix} y_1^1 & \cdots & y_n^1 \\ \vdots & \ddots & \vdots \\ y_1^n & \cdots & y_n^n \end{bmatrix}.$$

Notice that the matrix $L_{n \times n}$ is exactly the same as the matrix $M_{p \times p}$ described above, restricted to the linear case. It is important to remark that the coefficient matrix in this linear system does not depend on the point x.

Let us estimate an upper bound on the right hand side vector in (4.7). Each element of this vector is the difference of two quantities of the form $\int_0^1 (\nabla f(x + t(y^i - x)) - \nabla f(x))^\top (y^i - x) dt$ and $\int_0^1 (\nabla f(x - tx) - \nabla f(x))^\top (-x) dt$. Because $x \in B(\Delta)$ and $y^i \in B(\Delta)$, these quantities can be bounded above by $2\gamma_L \Delta^2$ and $\gamma_L \Delta^2/2$, respectively, where γ_L is the Lipschitz constant of ∇f in Ω (see [10, Lemma 4.1.12]). Hence the ℓ_∞ norm of the right hand side can be bounded by $(5/2)\gamma_L \Delta^2$, and the bound on the ℓ_2 norm is

(4.8) $$\|L_{n \times n} e^g(x)\| \le \frac{5}{2} n^{1/2} \gamma_L \Delta^2.$$

This suggests that the error bound between the function and the fully linear interpolating polynomial is of the order of Δ^2 in the case of function values and of the order of Δ in the case of gradient values, which we prove below.

The error bounds in the following theorem are stated in terms of the size of the inverse of the scaled matrix

$$\hat{L}_{n \times n} = L_{n \times n}/\Delta.$$

This scaled matrix is the same as the matrix $L_{n \times n}$ corresponding to the scaled set $\hat{Y} = Y/\Delta \subset B(1)$. The smaller $\|\hat{L}_{n \times n}^{-1}\|$ is the better the error estimates are.

THEOREM 4.1. *Let $Y = \{0, y^1, \ldots, y^n\}$ be a poised set of interpolation 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 fully linear interpolation model and the gradient of the function satisfies*

(4.9) $$\|e^g(x)\| \le (5n^{\frac{1}{2}} \gamma_L \|\hat{L}_{n \times n}^{-1}\|/2) \, \Delta,$$

- *the error between the fully linear interpolation model and the function satisfies*

$$|e^f(x)| \le (5n^{\frac{1}{2}} \gamma_L \|\hat{L}_{n \times n}^{-1}\|/2 + \gamma_L/2) \, \Delta^2.$$

13

Proof. From (4.8) and $L_{n \times n} = \Delta \hat{L}_{n \times n}$, we have that

$$\|e^g(x)\| \leq \|L_{n \times n}^{-1}\|(5/2)n^{\frac{1}{2}}\gamma_L\Delta^2 \leq \|\hat{L}_{n \times n}^{-1}\|(5/2)n^{\frac{1}{2}}\gamma_L\Delta,$$

From this and the detailed form (4.5), we obtain

$$|e^f(x)| \leq (5n^{\frac{1}{2}}\gamma_L\|\hat{L}_{n \times n}^{-1}\|/2 + \gamma_L/2)\Delta^2.$$

□

4.2. Error estimates in the quadratic case. In the fully quadratic case we assume that we have a poised set $Y = \{y^0, y^1, \ldots, y^p\} = \{0, y^1, \ldots, y^p\}$ of $p_1 = p + 1 = (n+1)(n+2)/2$ interpolation points 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 therefore possible to build the fully quadratic interpolation model

$$(4.10) \quad m(x) = c + g^\top x + \frac{1}{2}x^\top Hx = c + \sum_{1 \leq k \leq n} g_k x_k + \frac{1}{2}\sum_{1 \leq k,\ell \leq n} h_{k\ell}x_k x_\ell,$$

where H is a symmetric matrix of order n. The unknown coefficients c, $g_1, \ldots, g_n$, and $h_{k\ell}, 1 \leq \ell \leq k \leq n$, are defined by the interpolating conditions (1.1).

Analogous to the linear case, we consider a point x in the ball $B(\Delta)$, for which we will try to estimate the error in the function value

$$(4.11) \qquad\qquad m(x) = f(x) + e^f(x),$$

in the gradient

$$(4.12) \qquad\qquad \nabla m(x) = Hx + g = \nabla f(x) + e^g(x),$$

and, in this quadratic case, also in the Hessian

$$H = \nabla^2 f(x) + E^H(x).$$

The error in the gradient has n components $e_k^g(x), k = 1, \ldots, n$, just as in the linear case. Since the Hessians of f and m are symmetric, we only need to consider the error in the second-order derivatives in the diagonal elements and in the elements below the diagonal

$$h_{k\ell} = \nabla_{k\ell}^2 f(x) + E_{k\ell}^H(x), \quad 1 \leq \ell \leq k \leq n.$$

Using (4.10) and subtracting (4.11) from all the $p_1 = p + 1$ equalities in (1.1), we have that

$$-x^\top g + \frac{1}{2}x^\top Hx - x^\top Hx$$
$$= f(0) - f(x) - e^f(x),$$
$$(y^i - x)^\top g + \frac{1}{2}(y^i - x)^\top H(y^i - x) + (y^i - x)^\top Hx$$
$$= f(y^i) - f(x) - e^f(x), \quad i = 1, \ldots, p.$$

14

Now we expand f by a Taylor's formula of order two around x for all the $p_1 = p + 1$ interpolation points and use the notation for the error in the gradient given by (4.12), to get

$$-x^\top e^g(x) + \frac{1}{2} x^\top [H - \nabla^2 f(x)] x$$

(4.13)
$$= \mathcal{O}(\Delta^3) - e^f(x),$$

$$(y^i - x)^\top e^g(x) + \frac{1}{2}(y^i - x)^\top [H - \nabla^2 f(x)](y^i - x)$$

(4.14)
$$= \mathcal{O}(\Delta^3) - e^f(x), \quad i = 1, \ldots, p.$$

The next step, as in the linear case, is to subtract the first of these equations from the other equations, canceling $e^f(x)$ and obtaining

$$(y^i)^\top (e^g(x) - E^H(x)x) + \frac{1}{2}(y^i)^\top [H - \nabla^2 f(x)](y^i) = \mathcal{O}(\Delta^3), \quad 1 \le i \le p.$$

The linear system that we need to analyze in this quadratic case can be written as

$$\sum_{1 \le k \le n} y_k^i t_k(x) + \frac{1}{2} \sum_{1 \le k \le n} (y_k^i)^2 E_{kk}^H(x) + \sum_{1 \le \ell < k \le n} [y_k^i y_\ell^i] E_{k\ell}^H(x)$$

$$= \mathcal{O}(\Delta^3), \quad 1 \le i \le p,$$

or, in matrix form, as

(4.15)
$$Q_{p \times p} \begin{bmatrix} t(x) \\ e^H(x) \end{bmatrix} = \mathcal{O}(\Delta^3),$$

with

(4.16)
$$t(x) = e^g(x) - E^H(x)x = e^g(x) - [H - \nabla^2 f(x)]x.$$

Here $e^H(x)$ is a vector of dimension $n + n(n-1)/2$ storing the elements $E_{kk}^H(x)$, $k = 1, \ldots, n$ and $E_{k\ell}^H(x)$, $1 \le \ell < k \le n$.

Notice that the matrix $Q_{p \times p}$ is exactly the same as matrix $M_{p \times p}$, described in the beginning of this section, restricted to the quadratic case. Once again we remark that the matrix $Q_{p \times p}$ defining this linear system does not depend on the point x.

We will estimate an upper bound on the right hand side vector in (4.15). Each element of this vector is the difference of two terms that can be bounded by $\gamma_Q \|y^i - x\|^3/6$ and $\gamma_Q \|x\|^3/6$, respectively, where γ_Q is the Lipschitz constant of $\nabla^2 f$ in Ω (see [10, Lemma 4.1.14]). Since $\|y^i - x\| \le 2\Delta$ and $\|x\| \le \Delta$, the difference can be bounded by $3\Delta^3/2$. Hence the ℓ_∞ norm of the right hand side can be bounded by that amount, and a bound on the ℓ_2 norm is

(4.17)
$$\left\| Q_{p \times p} \begin{bmatrix} t(x) \\ e^H(x) \end{bmatrix} \right\| \le \frac{3}{2} p^{\frac{1}{2}} \gamma_Q \Delta^3.$$

We will derive the error bounds in terms of the norm of the inverse of the scaled matrix

(4.18)
$$\hat{Q}_{p \times p} = Q_{p \times p} \begin{bmatrix} D_\Delta^{-1} & 0 \\ 0 & D_{\Delta^2}^{-1} \end{bmatrix},$$

15

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 $p - n$ with Δ^2 in the diagonal entries. This scaled matrix is the same as the matrix $Q_{p \times p}$ corresponding to the scaled set $\hat{Y} = Y/\Delta \subset B(1)$.

The next theorem generalizes the error estimates obtained in Theorem 4.1 for the linear case to the quadratic case. 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. The smaller $\|\hat{Q}_{p \times p}^{-1}\|$ is the better the error estimates are.

THEOREM 4.2. *Let $Y = \{0, y^1, \ldots, y^p\}$, with $p_1 = p + 1 = (n+1)(n+2)/2$, be a 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 this ball 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 fully quadratic interpolation model and the Hessian of the function satisfies*

$$\|E^H(x)\| \leq (\alpha_Q^H p^{\frac{1}{2}} \gamma_Q \|\hat{Q}_{p \times p}^{-1}\|)\,\Delta,$$

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

$$\|e^g(x)\| \leq (\alpha_Q^g p^{\frac{1}{2}} \gamma_Q \|\hat{Q}_{p \times p}^{-1}\|)\,\Delta^2,$$

- *the error between the fully quadratic interpolation model and the function satisfies*

$$|e^f(x)| \leq (\alpha_Q^f p^{\frac{1}{2}} \gamma_Q \|\hat{Q}_{p \times p}^{-1}\| + \beta_Q^f \gamma_Q)\,\Delta^3,$$

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:

$$\alpha_Q^H = \frac{3\sqrt{2}}{2}, \quad \alpha_Q^g = \frac{3(1 + \sqrt{2})}{2}, \quad \alpha_Q^f = \frac{6 + 9\sqrt{2}}{4}, \quad \beta_Q^f = \frac{1}{6}.$$

Proof. Let us first write the left-hand-side of the system (4.15) in the form

$$Q_{p \times p} \begin{bmatrix} D_\Delta^{-1} & 0 \\ 0 & D_{\Delta^2}^{-1} \end{bmatrix} \begin{bmatrix} D_\Delta t(x) \\ D_{\Delta^2} e^H(x) \end{bmatrix}.$$

Then, using the bound (4.17) and the notation (4.18) we obtain

(4.19)
$$\left\| \begin{bmatrix} D_\Delta t(x) \\ D_{\Delta^2} e^H(x) \end{bmatrix} \right\| \leq \frac{3}{2} p^{\frac{1}{2}} \gamma_Q \|\hat{Q}_{p \times p}^{-1}\| \Delta^3,$$

from which we get

$$\|D_{\Delta^2} e^H(x)\| \leq \frac{3}{2} p^{\frac{1}{2}} \gamma_Q \|\hat{Q}_{p \times p}^{-1}\| \Delta^3,$$

yielding the bound $\|e^H(x)\| \leq (3/2) p^{\frac{1}{2}} \gamma_Q \|\hat{Q}_{p \times p}^{-1}\| \Delta$. The error in the Hessian is therefore given by

$$\|E^H(x)\| \leq \|E^H(x)\|_F \leq \sqrt{2}\|e^H(x)\| \leq \frac{3\sqrt{2}}{2} p^{\frac{1}{2}} \gamma_Q \|\hat{Q}_{p \times p}^{-1}\| \Delta.$$

16

Now, we would like to derive the bound on $\|e^g(x)\|$. From (4.19) we also have

$$\|D_\Delta t(x)\| \leq \frac{3}{2} p^{\frac{1}{2}} \gamma_Q \|\hat{Q}_{p\times p}^{-1}\| \Delta^3,$$

and

$$\|t(x)\| \leq \frac{3}{2} p^{\frac{1}{2}} \gamma_Q \|\hat{Q}_{p\times p}^{-1}\| \Delta^2,$$

and therefore, from (4.16),

$$
\begin{aligned}
\|e^g(x)\| &\leq \|t(x)\| + \|E^H(x)\| \|x\| \\
&\leq \frac{3}{2} p^{\frac{1}{2}} \gamma_Q \|\hat{Q}_{p\times p}^{-1}\| \Delta^2 + \left(\frac{3\sqrt{2}}{2} p^{\frac{1}{2}} \gamma_Q \|\hat{Q}_{p\times p}^{-1}\| \Delta \right) \Delta \\
&= \frac{3(1+\sqrt{2})}{2} p^{\frac{1}{2}} \gamma_Q \|\hat{Q}_{p\times p}^{-1}\| \Delta^2.
\end{aligned}
$$

Here we have used the fact that x is in the ball $B(\Delta)$ centered at the origin.

Finally, from the detailed version of (4.13) and the bounds on $\|e^g(x)\|$ and $\|E^H(x)\|$ we have

$$
\begin{aligned}
|e^f(x)| &\leq \|e^g(x)\| \Delta + \|E^H(x)\| \Delta^2/2 + \gamma_Q \Delta^3/6 \\
&\leq \frac{6+9\sqrt{2}}{4} p^{\frac{1}{2}} \gamma_Q \|\hat{Q}_{p\times p}^{-1}\| \Delta^3 + \frac{\gamma_Q}{6} \Delta^3.
\end{aligned}
$$

□

We would like to emphasize that the goal of this section is not so much the novelty of the bounds, in general, but their usefulness for the reader with primarily optimization motivations, not to mention the simplicity of the derivation of these bounds.

4.3. Interpreting the error bounds in terms of the condition number of the scaled matrix. Since at least one of the $\hat{y}^i$'s of the set $\hat{Y}$ has norm 1, we know similarly as in (3.5) that $\|\hat{L}_{n\times n}\| \geq 1$ in the linear case and $\|\hat{Q}_{p\times p}\| \geq 1$ in the quadratic case. Thus, in both instances we have that

$$\|\hat{L}_{n\times n}^{-1}\| \leq \kappa(\hat{L}_{n\times n}) \quad \text{and} \quad \|\hat{Q}_{p\times p}^{-1}\| \leq \kappa(\hat{Q}_{p\times p}),$$

and the error estimates of Theorems 4.1 and 4.2 can be stated by replacing the norms of the inverses of $\hat{M}_{p\times p}$ by their corresponding condition numbers $\kappa(\hat{M}_{p\times p})$.

In fact, we can even go one step further and bound $\kappa(\hat{M}_{p\times p})$ by $\kappa(\hat{M})$. In fact, it is easy to see that $\|\hat{M}_{p\times p}\| \leq \|\hat{M}\|$, also

$$\hat{M}^{-1} = \begin{bmatrix} 1 & 0 \\ \vdots & \hat{M}_{p\times p}^{-1} \end{bmatrix}.$$

Hence, $\|\hat{M}_{p\times p}^{-1}\| \leq \|\hat{M}^{-1}\|$, and $\kappa(\hat{M}_{p\times p}) \leq \kappa(\hat{M})$.

Our bounds in terms of the condition number of $\hat{M}$ exhibit a constant multiplying $\hat{M}$ which is a linear function of $\sqrt{p_1}$. Curiously, the bound (2.1) for function values is not as sharp as ours if posed in terms of $\hat{M}$. If fact, for $m = 0$ and $d = 2$, we get from (2.1), (3.4), $\max_{x\in B(1)} \|\bar{\phi}(x)\|_\infty \leq 1$, and (3.5) that

$$
\begin{aligned}
|m(x) - f(x)| &\leq \frac{1}{6} G \|\lambda(x)\|_1 \Delta^3 \leq \frac{1}{6} p_1 G \|\lambda(x)\|_\infty \Delta^3 \\
&\leq \frac{1}{6} p_1 G \|\hat{M}^{-T}\| \max_{x\in B(1)} \|\bar{\phi}(x)\|_\infty \Delta^3 \leq \frac{1}{6} p_1 G \kappa(\hat{M}) \Delta^3,
\end{aligned}
$$

showing that the constant that multiplies Λ varies now linearly with $p_1 = p + 1$.

4.4. Interpreting the error bounds in terms of the bound on Lagrange polynomials. If we assume that the interpolation set Y is Λ–poised in $B(\Delta)$, which by the scaling invariance property of Lagrange polynomials implies that $\hat{Y} = Y/\Delta$ is Λ–poised in $B(1)$, then we can apply Theorem 3.3 and write

$$\|\hat{M}_{p \times p}^{-1}\| \ \leq \ \|\hat{M}^{-1}\| \ \leq \ \theta\Lambda,$$

where the first inequality is justified as in the previous section. So, for instance, the error in function values for the quadratic case becomes (see (3.9))

$$|m(x) - f(x)| \ \leq \ \left(4\alpha_Q^f \gamma_Q p_1 \Lambda + \beta_Q^f \gamma_Q\right) \Delta^3,$$

which is similar to the bound (2.2) for $d = 2$ (in the sense that what multiplies Λ varies linearly with $p_1 = p + 1$).

The error bound (2.1) for gradient and Hessian depends on a bound for the gradient and the Hessian (respectively) of Lagrange polynomials. In [4], the authors mention that such a bound would involve additional constants related to the geometry of Y, which may make the bounds unnecessarily loose. Actually, our approach of looking at Lagrange polynomials and considering Λ–poisedness provides an alternative way to derive a bound on the derivatives of the Lagrange polynomials that seems sharper and explicitly exhibits the same geometry constant Λ. For instance, for $m = 1$ and $d = 2$, we easily derive the following bound for the Jacobian matrix $\mathcal{D}\lambda(x)$:

$$\|\mathcal{D}\lambda(x)\|_\infty \ \leq \ \|\hat{M}^{-1}\| \ \max_{z \in B(1)} \|\mathcal{D}\bar{\phi}(z)\|_\infty \ \leq \ 2\theta\Lambda.$$

5. Ensuring well-poisedness. In a typical (interpolation based) trust-region or line search DFO sampling method, an interpolation set is maintained at each iteration. Either the new iterate needs to be included into the interpolation model, or the model needs to be improved. To guarantee the quality of the interpolation model, i.e., the appropriate error estimates in the function and in its derivatives, one needs to make sure that the poisedness of the interpolation set does not deteriorate arbitrarily from iteration to iteration. That can be guaranteed in at least two ways. One is to select a "good" interpolation set *a priori* and keep shifting and scaling at each iteration to place it inside the current region of interest. This idea is closely related to the use of a finite number of positive bases in pattern search methods (see [2], [12], [21]). An alternative approach (used in [7], [8], [13]) is to update the interpolation set by one or two interpolation points per iteration, while ensuring that it satisfies some sufficient well-poisedness condition. For instance, in a trust-region framework, this condition has to be satisfied whenever the new iterate is rejected and the trust-region radius is reduced. If such a condition is not satisfied then at least one "bad" point is replaced by a "good" point. In practice, it is more efficient to maintain well-poisedness throughout the algorithm, not just when it is necessary to pass the criterion needed for the convergence proof.

The methods of handling well-poisedness of the interpolation set that were proposed so far are not very satisfactory, in some sense. Powell, in [17], uses values of Lagrange polynomials as the criterion for update of the interpolation set. Each time a point in Y is replaced, it is done so that the absolute value of the corresponding Lagrange polynomial is maximized at the incoming point (depending on the situation this is achieved either by choosing the incoming point, or the outgoing point, and hence, the corresponding Lagrange polynomial). This strategy is reasonable, and

18

seems to work well in practice, however, there is no proof that such strategy will produce sets which are Λ-poised with a uniformly bounded Λ. In [7], a similar strategy based on Newton fundamental polynomials was used. The same problem as for Powell's method arises: there is no guarantee that the generated interpolation sets are Λ–poised with uniformly bounded Λ. To circumvent this difficulty and to guarantee convergence of the algorithm the authors of [7] introduced an extra step, where all interpolation points have to be replaced, in order to guarantee that a Λ–poised set is created. This step is very expensive, since it requires evaluation of f at all new sample points. In practice this step should be avoided as much as possible.

In contrast, the methods we propose below are both practical and theoretically sound. They guarantee that the set Y is Λ-poised for a specified (reasonably large) Λ and at the same time they can handle updates of the interpolation set by one element. These methods are also based on maintaining a set of polynomials and monitoring the values of these polynomials at interpolation points. The drawback of these methods, compared to that of Powell [17], is that the set of polynomials has to be (potentially) completely recomputed, at each update of Y. However, the set itself does not have to be recomputed. Only the points that "spoil" the Λ-poisedness have to be replaced. We will comment on performance of the methods at the end of the section.

The algorithms that we describe in this section build an interpolation set Y, or modify an already existing one, using some form of matrix factorization. We will first present (in Subsection 5.1) an algorithm based on Gaussian elimination, as it is the most intuitive and easy to explain, although, it might not be the most appealing from a numerical point of view. In Subsection 5.2 we will present a potentially more stable algorithm based on Gram-Schmidt orthogonalization. The proposed methods are not the only ones that can be used in derivative free methods. The main reason why we chose the two methods for this paper is because they are based on matrix factorization algorithms, and hence they are easy to analyze.

Gaussian elimination has already been proved to be a reasonable idea for matrices arising in polynomial interpolation. See, for instance, the paper by de Boor [9] and the references therein. However, in [9] it is assumed that the set of interpolation points is fixed data and the goal is to study the properties of the interpolant function spaces and related linear algebra algorithms for matrices arising from these spaces. Our optimization motivation is different, as the sets of sampling points for interpolation are no longer fixed and change considerably during the course of an optimization procedure. Consequently, our linear algebra algorithms focus instead on the quality of the interpolation set for a given fixed function interpolation space.

We will present both algorithms, the Gaussian elimination one and the Gram-Schmidt one, for the case of quadratic interpolation. The extension to higher degree interpolations is straightforward. We will use the standard Matlab notation to indicate submatrices. We describe the algorithms in the situation where the points lie in a (smallest enclosing) ball of radius 1. If the radius is $\Delta \neq 1$ then one should scale Y by $1/\Delta$, apply the chosen algorithm and then scale back the (possibly) new interpolation set. We will work under the assumption that $y^0 = 0$. The matrix that is factorized is M for LU and $M^\top$ for QR, where M is given by (4.1) (assuming $\Delta = 1$).

The algorithms will provide us with a bound of the type

$$(5.1) \qquad \|M^{-1}\| \;\leq\; \frac{c(p_1)\,\varepsilon_{growth}}{\xi},$$

where ε_{growth} is an estimate of the growth factor that occurs during the factorization, ξ is a lower bound on the absolute value of the pivots, and $c(p_1)$ is some power of

$p_1 = p + 1$. For instance, in the quadratic case and for the LU version, ξ can be any number in $(0, 1/4)$.

Since M corresponds to a set in $B(1)$, we get from (3.5) that the condition number of M can be bounded by

$$\kappa(M) \ \leq \ \frac{c(p_1)\, p_1^{\frac{3}{2}} \varepsilon_{growth}}{\xi}.$$

On the other hand, we already know from Theorem 3.3 that if $\|M^{-1}\|$ is bounded by Λ then Y is $\sqrt{p_1}\Lambda$–poised. These algorithms indicate us that, in this context, such a Λ can be any number satisfying

$$\Lambda \ \geq \ \frac{c(p_1)\, \varepsilon_{growth}}{\xi}.$$

We now have

$$\|\lambda(x)\|_1 \ \leq \ p_1 \|\lambda(x)\|_\infty \ \leq \ \frac{p_1 c(p_1)\, \varepsilon_{growth}}{\xi},$$

yielding a bound on the size of the Lagrange polynomials.

It is straightforward to adapt Algorithms 5.1 and 5.2 to the linear case, for which the threshold ξ for the absolute value of the pivots in the LU version is required to satisfy $\xi \in (0, 1)$. Moreover, in the linear case it is possible to identify the simplex geometry in Y that yields the smallest possible bound on the norm of M^{-1} (which is related to the concept of a uniform positive basis; see [1]).

5.1. Ensuring well-poisedness using Gaussian elimination. The outcome of the LU algorithm below is the interpolation set Y and the LU factors of the matrix M defined by (4.1) (assuming $\Delta = 1$ and $y^0 = 0$). The Gaussian elimination is performed by rows. Thus, since the points in Y appear by rows in M, the algorithm computes a new point y^i — or modifies the already existing one — only when the $i+1$-th row is being factorized (the first row of M, corresponding to $y^0 = 0$ is already factorized). The algorithm checks if the current set is well-poised, and if not, identifies "bad" points and replaces them by "good" points.

ALGORITHM 5.1 (Ensuring well-poisedness — LU).

Step 0 Let $\bar{\phi}$ denote the natural polynomial basis for $d \leq 2$ defined in (3.3). Choose some threshold ξ such that $0 < \xi < \frac{1}{4}$.

Step 1 Set $U_{1,1:p_1} = \phi(y^0)^\top = e_1^\top$, corresponding to $y^0 = 0$.

For $k = 2, \ldots, p_1$

Step k   • Assume that the first $k-1$ steps of Gaussian elimination have been completed, hence, we have the first $k-1$ rows of the upper triangular matrix U: $U_{1:k-1, 1:p_1}$.

      • Let $x \in Y$ be the sample point corresponding to the k-th row. Then the element $U_{k,k}$, i.e., the k-th pivot element in the Gaussian elimination process, can be expressed as

$$U_{k,k}(x) \ = \ \bar{\phi}_k(x) - \bar{\phi}_1(x)\frac{U_{1,k}}{U_{1,1}} - \cdots - \bar{\phi}_{k-1}(x)\frac{U_{k-1,k}}{U_{k-1,k-1}}.$$

20

Clearly $U_{k,k}(x)$ is a linear (when $k \leq n+1$) or quadratic (when $k > n+1$) polynomial in x, and can be written as $(v^k)^\top \bar{\phi}(x)$ with $v^k \in \mathbb{R}^{p_1}$ and $\|v^k\|_\infty \geq 1$.

- Find $i^* = \operatorname{argmax}\{|(v^k)^\top \bar{\phi}(y^i)| : i = k-1, \ldots, p\}$.
  If $|(v^k)^\top \bar{\phi}(y^{i^*})| \geq \xi$ then set $x = y^{i^*}$ and swap the points y^{k-1} and y^{i^*} in Y.
  If $|(v^k)^\top \bar{\phi}(y^{i^*})| < \xi$ then find

$$x = y^{k-1} = \operatorname{argmax}_{x \in B(1)} |(v^k)^\top \bar{\phi}(x)|.$$

- Update the factorization

$$U_{k,i} \;=\; \bar{\phi}_i(x) - \bar{\phi}_1(x)\frac{U_{1,i}}{U_{1,1}} - \cdots - \bar{\phi}_{k-1}(x)\frac{U_{k-1,i}}{U_{k-1,k-1}}, \quad k < i \leq p_1.$$

Since $\|v^k\|_\infty \geq 1$ and $\xi < 1/4$ we know from Lemma 3.4 that

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

PROPOSITION 5.1. *Let a pivot threshold $\xi \in (0, 1/4)$ be given. Algorithm 5.1 computes a set Y of $p_1 = p+1 = (n+1)(n+2)/2$ points in the unit ball $B(1)$ centered at $y^0 = 0$ for which the pivots of the Gaussian elimination of M satisfy*

$$|D_{ii}| \;\geq\; \xi, \quad i = 1, \ldots, p_1.$$

The effort required by the algorithm for the Gaussian elimination as presented is of the order of $\mathcal{O}(n^6)$ floating point operations. The algorithm requires, moreover, in the worst case, the maximization of n linear functions and $p - n$ quadratic functions and the maximization of their symmetric counterparts, in a ball of radius 1. Strictly speaking we only need to guarantee the computation of a point with objective function value greater than or equal to $1/4$. This can be done by using the same arguments used in Lemma 3.4 to prove that this bound of $1/4$ is achievable, which has the advantage of reducing each pair of these optimization problems to a trivial enumeration.

The outcome of Algorithm 5.1 can be written in the form $M = LDU$ where $\|D^{-1}\| \leq \sqrt{p_1}/\xi$ and L and U are lower and upper triangular matrices, respectively, with ones in the diagonals. Thus,

$$\|M^{-1}\| \;\leq\; \frac{p_1^{\frac{1}{2}} \|L^{-1}\| \|U^{-1}\|}{\xi}.$$

The sizes of $\|L^{-1}\|$ and $\|U^{-1}\|$ are related to the growth factor of the factorization, and are expected to be of reasonable size for most practical instances. We obtain in this way the bound (5.1) with $\varepsilon_{growth} = \|L^{-1}\| \|U^{-1}\|$ and $c(p_1) = \sqrt{p_1}$.

5.2. Ensuring well-poisedness using Gram-Schmidt orthogonalization. We now consider an alternative way of maintaining the interpolation set Y, via Gram-Schmidt orthogonalization of the columns of the matrix $M^\top$ defined by (4.1), where the points in Y appear now columnwise. The outcome of this algorithm is the interpolation set Y and the QR factors of the matrix $M^\top$ (assuming $\Delta = 1$ and $y^0 = 0$).

Similarly to the LU algorithm, the QR algorithm computes a new point y^i — or modifies the already existing one — only when the $i+1$-th column is being factorized (the first column of $M^\top$, corresponding to $y^0 = 0$ is already factorized). This algorithm also checks if the current set is well-poised, and if not, identifies "bad" points and replaces them by "good" points.

Let us suppose that we have used the Gram-Schmidt orthogonalization scheme to generate an orthonormal basis $\{q^1, \ldots, q^{k-1}\}$ for the subspace generated by the vectors $\bar{\phi}(y^0), \bar{\phi}(y^1), \ldots, \bar{\phi}(y^{k-2})$, with $y^0 = 0$. At this stage of the orthogonalization process, we have computed a QR factorization of the first $k-1$ columns of the matrix $M^\top$:

$$M^\top_{1:p_1,1:k-1} = Q_{1:p_1,1:k-1} R_{1:k-1,1:k-1},$$

where $Q_{1:p_1,1:k-1} = \begin{bmatrix} q^1 & \cdots & q^{k-1} \end{bmatrix}$ and $R_{1:k-1,1:k-1}$ is an upper triangular matrix of order $k-1$. Given $\bar{\phi}(x)$, we know that the vector $q^k(x)/\|q^k(x)\|$, where

$$q^k(x) = \bar{\phi}(x) - \left(\bar{\phi}(x)^\top q^1\right) q^1 - \cdots - \left(\bar{\phi}(x)^\top q^{k-1}\right) q^{k-1},$$

forms together with $q^1, \ldots, q^{k-1}$ an orthonormal basis for the subspace generated by $\bar{\phi}(y^0), \bar{\phi}(y^1), \ldots, \bar{\phi}(y^{k-2}), \bar{\phi}(x)$.

We are interested in selecting or modifying the next point $y^{k-1} = x$ so that $q^k(x)$ has the largest possible norm. By maximizing $\|q^k(x)\|$ we are asking $\bar{\phi}(x)$ to be as close to being orthogonal to the subspace generated by $\bar{\phi}(y^0), \bar{\phi}(y^1), \ldots, \bar{\phi}(y^{k-2})$ as possible. Note that $\|q^k(y^{k-1})\|$ will be the k-th diagonal element of $R_{1:k,1:k}$. For the purpose of maximizing $\|q^k(x)\|$, let $\bar{q}^k, \ldots, \bar{q}^{p_1}$ be an orthonormal basis for the null space of $M_{1:p_1,1:k-1}$. We have that

$$q^k(x)^\top q^k(x) = \bar{\phi}(x)^\top \bar{\phi}(x) - \sum_{i=1}^{k-1} (\bar{\phi}(x)^\top q^i)^2 = \sum_{i=k}^{p_1} (\bar{\phi}(x)^\top \bar{q}^i)^2.$$

Let $\bar{v}^k$ be one of the vectors among $\bar{q}^k, \ldots, \bar{q}^{p_1}$. We know that $\|\bar{v}^k\| = 1$ and $\|\bar{v}^k\|_\infty \geq 1/\sqrt{p_1}$. From Lemma 3.4 we also know that

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

for $v^k = \bar{v}^k/\|\bar{v}^k\|_\infty$. Thus, for any ξ in $(0, 1/4)$ (or for any ξ in $(0,1)$ if $\bar{\phi}(x)$ is only formed by linear terms), there exists a point y^{k-1} such that

$$\|q_k\| = \|q_k(y^{k-1})\| > \frac{\xi}{p_1^{\frac{1}{2}}}.$$

The algorithm for ensuring well-poisedness via QR decomposition of the matrix $M^\top$ is presented below. At each iteration k, a new column $\bar{\phi}(y^{k-1})$ is orthogonalized after the point y^{k-1} has been appropriately selected or modified to increase the size of the next diagonal element of R in the QR factorization.

ALGORITHM 5.2 (Ensuring well-poisedness — QR).

Step 0 Let $\bar{\phi}$ denote the natural polynomial basis for $d \leq 2$ defined in (3.3). Choose some threshold ξ such that $0 < \xi < \frac{1}{4}$.

Step 1 Set $Q_{1:p_1,1} = q^1 = \bar{\phi}(y^0) = e_1$ and $R_{1,1} = \|\bar{\phi}(y^0)\| = 1$, corresponding to $y^0 = 0$.

For $k = 2, \ldots, p_1$

Step k • Assume that the first $k-1$ steps of the QR factorization have been completed, hence, we have $M_{1:p_1,1:k-1}^\top = Q_{1:p_1,1:k-1} R_{1:k-1,1:k-1}$, where the columns in $Q_{1:p_1,1:k-1}$ are orthonormal and $R_{1:k-1,1:k-1}$ is an upper triangular matrix.

       • Compute a vector v^k in the null space of $M_{1:p_1,1:k-1}$ with $\|v^k\|_\infty = 1$.

       • Find $i^* = \mathrm{argmax}\{|(v^k)^\top \bar{\phi}(y^i)| : i = k-1, \ldots, p\}$.
If $|(v^k)^\top \bar{\phi}(y^{i^*})| \geq \xi$ then set $x = y^{i^*}$ and swap the points y^{k-1} and y^{i^*} in Y.
If $|(v^k)^\top \bar{\phi}(y^{i^*})| < \xi$ then find

$$x = y^{k-1} = \mathrm{argmax}_{x \in B(1)} |(v^k)^\top \bar{\phi}(x)|.$$

       • Using Gram-Schmidt or modified Gram-Schmidt orthogonalize the k-th column $\bar{\phi}(y^{k-1})$ of $M^\top$, computing the k-th column q^k of Q, and the k-th column of R (with elements $R_{1,k}, \ldots, R_{k,k}$).

PROPOSITION 5.2. *Let a pivot threshhold $\xi \in (0, 1/4)$ be given. Algorithm 5.2 computes a set Y of $p_1 = p+1 = (n+1)(n+2)/2$ points in the unit ball $B(1)$ centered at $y^0 = 0$ for which the diagonal elements of R in the QR factorization of $M^\top$ satisfy*

$$|R_{ii}| \geq \frac{\xi}{p_1^{\frac{1}{2}}}, \quad i = 1, \ldots, p_1.$$

The algorithm could be improved in several ways. For instance, instead of computing only one vector v^k in the null space of $M_{1:k-1,1:p_1}$, we could compute an orthonormal basis for this subspace (normalized with respect to the ℓ_∞ norm), and then choose y^k that gives the maximum value of $\max_{x \in B(1)} |v^\top \bar{\phi}(x)|$ for all $p-k+2$ vectors v in this basis.

Algorithm 5.2 computes a QR factorization of $M^\top$ where, as we have seen before, the diagonal elements of R satisfy $|R_{ii}| > \xi/\sqrt{p_1}$, $i = 1, \ldots, p_1$. Now we write $M^\top = QR = QD\bar{R}$, where Q is orthogonal and $\bar{R}$ is a p_1-by-p_1 upper triangular matrix with ones in the diagonal. Thus,

$$\|M^{-1}\| \leq \frac{p_1 \|\bar{R}^{-1}\|}{\xi}.$$

The size of $\|\bar{R}^{-1}\|$ is related to the growth factor of the factorization, and is expected to be small in practice. We obtain in this way the bound (5.1) with $\varepsilon_{growth} = \|\bar{R}^{-1}\|$ and $c(p_1) = p_1$. Since the QR factorization of a matrix often has better numerical properties than the LU factorization, it is reasonable to expect the same from the comparison of Algorithms 5.1 and 5.2.

5.3. Practical considerations. A natural question that arises is if the proposed algorithms will be useful in practice. The answer will require the implementation of the algorithms within a good DFO package. It is possible that the proposed algorithms will not perform better than Powell's algorithm. However, we expect them to be reasonably practical. As we pointed out, these algorithms have the necessary

theoretical foundation to guarantee global convergence of DFO methods. It is also possible that, by using the theory of this paper, we may be able to show that algorithms like Powell's can also provide guarantees of uniform well-poisedness. Such proofs will be a subject of a separate study.

For this paper we have conducted a series of simple experiments in Matlab. We have generated 100 points, randomly in a unit cube centered at zero. From these points we selected sets of interpolation points, by repeatedly applying Powell's algorithm of maximizing the absolute value of Lagrange polynomials, our LU factorization algorithm, and our QR factorization algorithm. We then compared the resulting values of $\|M^{-1}\|$. All three algorithms produced very close results, with the value of $\|M^{-1}\|$ ranging most of the times between 10 and 100, and never exceeding 1000. This numerical behavior may be the consequence of the random nature of the chosen interpolation points. Note that without a framework of a real derivative free package, it is difficult to reproduce a more practical example. This simple experiment confirms our expectation that in "normal" circumstances our algorithms perform similarly to the existing algorithms used in practice.

Appendix: Extension to higher degree. The error estimates derived in this paper for linear and quadratic interpolation extend naturally to interpolation polynomials of higher degree. We will briefly sketch here the cubic case.

The procedure to derive the error estimates in the cubic case goes one step further than the quadratic case but the arguments used are the same. After subtracting the equation $m(x) - f(x) = e^f(x)$ on the error in the function from all the $p_1 = p + 1$ interpolating conditions (1.1) and expanding f by a Taylor's formula of order three around $x \in B(\Delta)$, we get the following analog of (4.13)–(4.14):

$$\sum_{1 \le k \le n} -e_k^g(x)x_k + \frac{1}{2} \sum_{1 \le k,\ell \le n} E_{k\ell}^H(x)x_k x_\ell$$
$$- \frac{1}{6} \sum_{1 \le k,\ell,m \le n} E_{k\ell m}^c(x)x_k x_\ell x_m \;=\; \mathcal{O}(\Delta^4) - e^f(x),$$

$$\sum_{1 \le k \le n} e_k^g(x)(y_k^i - x_k) + \frac{1}{2} \sum_{1 \le k,\ell \le n} E_{k\ell}^H(x)(y_k^i - x_k)(y_\ell^i - x_\ell)$$
$$+ \frac{1}{6} \sum_{1 \le k,\ell,m \le n} E_{k\ell m}^c(x)(y_k^i - x_k)(y_\ell^i - x_\ell)(y_m^i - x_m) \;=\; \mathcal{O}(\Delta^4) - e^f(x),$$

for $i = 1, \ldots, p$, where $p_1 = p + 1$ is the number of points in the interpolation set $Y = \{0, y^1, \ldots, y^p\} \subset B(\Delta)$, to be defined later. Here we have $E_{k\ell m}^c(x) = C_{k\ell m} - \frac{\delta^3 f}{\delta x_k \delta x_\ell \delta x_m}(x)$, where $C_{k\ell m}$ is the corresponding coefficient of the cubic interpolating polynomial. Subtracting the first of these equations from the others, yields

$$\sum_{1 \le k \le n} y_k^i \left(e_k^g(x) - \sum_{1 \le \ell \le n} E_{k\ell}^H(x)x_\ell - \frac{1}{2} \sum_{1 \le \ell,m \le n} E_{k\ell m}^c(x)x_\ell x_m \right)$$

$$+ \frac{1}{2} \sum_{1 \le k,\ell \le n} y_k^i y_\ell^i \left(E_{k\ell}^H(x) - \sum_{1 \le m \le n} E_{k\ell m}^c(x)x_m \right)$$

$$+ \frac{1}{6} \sum_{1 \le k,\ell,m \le n} y_k^i y_\ell^i y_m^i E_{k\ell m}^c(x) \;=\; \mathcal{O}(\Delta^4), \quad i = 1, \ldots, p.$$

We denote the matrix of this linear system by $C_{p \times p}$, which coincides with $M_{p \times p}$. The error bounds are given in terms of the norm of the inverse of $\hat{C}_{p \times p}$ corresponding to the scaled set $\hat{Y} = Y/\Delta$. The number of elements in the basis is given by

$$(5.2) \qquad p_1 = p + 1 = \begin{pmatrix} n + 3 \\ n \end{pmatrix}.$$

In the cubic case the error estimates include the error in the third derivatives.

THEOREM 5.3. *Let $Y = \{0, y^1, \ldots, y^p\}$, with p given by (5.2), be a poised set of interpolation points contained in a (closed) ball $B(\Delta)$ centered at 0. Assume that f is thrice continuously differentiable in an open domain Ω containing $B(\Delta)$ and that the vector of the third-order derivatives is Lipschitz continuous in Ω with constant $\gamma_C > 0$.*

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

- *the error between the vector of the third-order derivatives of the fully cubic interpolation model and the vector of the third-order derivatives of the function satisfies*

$$\|e^c(x)\| \leq (\alpha_C^c p^{\frac{1}{2}} \gamma_C \|\hat{C}_{p \times p}^{-1}\|) \, \Delta,$$

- *the error between the Hessian of the fully cubic interpolation model and the Hessian of the function satisfies*

$$\|E^H(x)\| \leq (\alpha_C^H p^{\frac{1}{2}} \gamma_C \|\hat{C}_{p \times p}^{-1}\|) \, \Delta^2,$$

- *the error between the gradient of the fully cubic interpolation model and the gradient of the function satisfies*

$$\|e^g(x)\| \leq (\alpha_C^g p^{\frac{1}{2}} \gamma_C \|\hat{C}_{p \times p}^{-1}\|) \, \Delta^3,$$

- *the error between the fully cubic interpolation model and the function satisfies*

$$|e^f(x)| \leq (\alpha_C^f p^{\frac{1}{2}} \gamma_C \|\hat{C}_{p \times p}^{-1}\| + \beta_C^f \gamma_C) \, \Delta^4,$$

where α_C^c, α_C^H, α_C^g, α_C^f, and β_C^f are small positive constants dependent on $d = 3$ and independent of n and Y.

From these bounds in terms of $\|\hat{C}_{p \times p}^{-1}\|$, we could derive bounds in terms of $\kappa(\hat{C}) = \kappa(\hat{M})$ and Λ, as explained in Subsections 4.3 and 4.4. Also, the extension to polynomial interpolation of degree higher than cubic would follow in a similar fashion.

REFERENCES

[1] P. ALBERTO, F. NOGUEIRA, H. ROCHA, AND L. N. VICENTE, *Pattern search methods for user-provided points: Application to molecular geometry problems*, SIAM J. Optim., 14 (2004), pp. 1216–1236.

[2] C. AUDET AND J. E. DENNIS, *Analysis of generalized pattern searches*, SIAM J. Optim., 13 (2003), pp. 889–903.

[3] D. M. BORTZ AND C. T. KELLEY, *The simplex gradient and noisy optimization problems*, vol. 24, Birkhäuser, Boston, 1998, pp. 77–90.

[4] P. G. CIARLET AND P. A. RAVIART, *General Lagrange and Hermite interpolation in R^n with applications to finite element methods*, Arch. Rational Mech. Anal., 46 (1972), pp. 177–199.

[5] B. Colson, *Trust-Region Algorithms for Derivative-Free Optimization and Nonlinear Bilevel Programming*, PhD thesis, Département de Mathématique, FUNDP, Namur, 2003.

[6] A. R. Conn, N. I. M. Gould, and Ph. L. Toint, *Trust-Region Methods*, MPS-SIAM Series on Optimization, SIAM, Philadelphia, 2000.

[7] A. R. Conn, K. Scheinberg, and Ph. L. Toint, *On the convergence of derivative-free methods for unconstrained optimization*, in Approximation Theory and Optimization, Tributes to M. J. D. Powell, edited by M. D. Buhmann and A. Iserles, Cambridge University Press, Cambridge, 1997, pp. 83–108.

[8] ———, *Recent progress in unconstrained nonlinear optimization without derivatives*, Math. Programming, 79 (1997), pp. 397–414.

[9] C. de Boor, *Gaussian elimination by segments and multivariate polynomial interpolation*, in Approximation and Computation: A Festschrift in Honor of Walter Gautschi, edited by R.V.M Zahar, Birkhäuser Verlag, 1994, pp. 87–96.

[10] J. E. Dennis and R. B. Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Prentice–Hall, Englewood Cliffs, (republished by SIAM, Philadelphia, in 1996, as Classics in Applied Mathematics, 16), 1983.

[11] C. T. Kelley, *Iterative Methods for Optimization*, SIAM, Philadelphia, 1999.

[12] T. G. Kolda, R. M. Lewis, and V. Torczon, *Optimization by direct search: New perspectives on some classical and modern methods*, SIAM Rev., 45 (2003), pp. 385–482.

[13] M. Marazzi and J. Nocedal, *Wedge trust region methods for derivative free optimization*, Math. Programming, 91 (2002), pp. 289–300.

[14] R. Oeuvray, *Trust-Region Methods Based on Radial Basis Functions with Application to Biomedical Imaging*, PhD thesis, EPFL, Lausanne, 2005.

[15] R. Oeuvray and M. Bierlaire, *BOOSTERS: A derivative-free algorithm based on radial basis functions*. Submitted for publication, 2005.

[16] M. J. D. Powell, *Direct search algorithms for optimization calculations*, Acta Numerica, 7 (1998), pp. 287–336.

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

[18] ———, *On trust region methods for unconstrained minimization without derivatives*, Math. Programming, 97 (2003), pp. 605–623.

[19] C. J. Price and I. D. Coope, *Frames and grids in unconstrained and linearly constrained optimization: A non-smooth approach*, SIAM J. Optim., 14 (2003), pp. 415–438.

[20] Th. Sauer, *On multivariate Lagrange interpolation*, Math. Comp., 64 (1995), pp. 1147–1170.

[21] V. Torczon, *On the convergence of pattern search algorithms*, SIAM J. Optim., 7 (1997), pp. 1–25.

[22] M. H. Wright, *Direct search methods: Once scorned, now respectable*, in Numerical Analysis 1995, Proceedings of the 1995 Dundee Biennal Conference in Numerical Analysis, D. F. Griffiths and G. A. Watson, eds., Pitman Res. Notes Math. Ser. 344, CRC Press, Boca Raton, Florida, 1996, pp. 191–208.