Positive Bases in Numerical Optimization

I.D. Coope and C.J. Price

Department of Mathematics & Statistics,
University of Canterbury,
Private Bag 4800, Christchurch, New Zealand.

Report Number: UCDMS2000/12 August 2000

Keywords: Positive bases, numerical optimization, derivative estimation, convergence.

Positive Bases in Numerical Optimization

I.D. Coope and C.J. Price

Department of Mathematics & Statistics

University of Canterbury, Private Bag 4800

Christchurch, New Zealand.

August, 2000

Abstract. The theory of positive bases introduced by C. Davis in 1954 does not appear in most modern texts on linear algebra but has re-emerged in publications in optimization journals. In this paper some simple properties of this highly useful theory are highlighted and applied to both theoretical and practical aspects of the design and implementation of numerical algorithms for nonlinear optimization.

Keywords: Positive bases, numerical optimization.

1 Introduction

The theory of positive bases introduced by C. Davis [5] in 1954 does not appear in most modern texts on linear algebra and has only recently re-emerged in publications in optimization journals (see, for example [12], [7], [8]). In this paper some simple properties of this highly useful theory are highlighted and applied to both theoretical and practical aspects of the design and implementation of algorithms for nonlinear optimization.

The paper is organized as follows. In the next section some simple properties of positive bases that are useful to optimization applications are outlined. In Section 3, grids or meshes

are formally introduced and the properties of positive bases are used to provide examples of very simple convergence proofs for some grid-based numerical optimization algorithms.

Section 4 is concerned with the very practical problem of estimating derivative information using function values at grid points and the inclusion of some numerical examples illustrates the usefulness of positive bases in this context.

2 Positive Independence & Positive Bases

A *positive combination* of the set of vectors $\{v_j \in R^n : j=1,\ldots,r\}$ is a linear combination

$$\alpha_1 v_1 + \cdots + \alpha_r v_r$$

with $\alpha_j \geq 0$; if all $\alpha_j > 0$ then it is a *strictly* positive combination.

2.1 Positive Independence

A set of vectors $\{v_j \in R^n : \ j=1,\ldots,r\}$ is *positively dependent* if one of them is a positive combination of the others (in particular, if any v_j is zero). Otherwise the set is *positively independent.* Any subset of a positively independent set is positively independent.

2.2 Positive Basis

A *positive basis* for a subspace $C \subset \mathbf{R}^n$ is a set of positively independent vectors whose span is C. In particular, a positive basis for $\mathbf{R}^n$ is such that every vector in $\mathbf{R}^n$ can be written as a positive combination of the positive basis vectors but no member of the positive basis is expressible as a positive combination of the remaining members of the basis. It is shown in [5] that the cardinality of a positive basis $\mathcal{V}_+$ for $\mathbf{R}^n$ satisfies $n+1 \leq |\mathcal{V}_+| \leq 2n$. Such positive bases are easily constructed as the following examples show.

Let $V = [v_1, \ldots, v_n]$ be a matrix whose columns form a basis $\mathcal{V} = \{v_1, \ldots, v_n\}$ for $\mathbf{R}^n$ and let $e = [1, \ldots, 1]^T$. Then the columns of

$$[V, -Ve] \text{ and } [V, -V] \tag{1}$$

are, respectively, positive bases of minimal and maximal cardinality. These simple examples of positive bases (1) will be referred to as being obtained by extending the basis $\mathcal{V}$ to $\mathcal{V}_+$. Other examples (and properties) of positive bases can be found in [5].

The usefulness of positive bases in an optimization context stems from the following simple result.

Theorem 1

If the set of vectors $\mathcal{V}_+$ is a positive basis, then

$$v^T g \geq 0 \quad \forall v \in \mathcal{V}_+ \quad \Rightarrow \quad g = 0.$$

PROOF: Let the members of $\mathcal{V}_+$ be v_i for $i = 1, \ldots, |\mathcal{V}_+|$. Then

$$-g = \sum_{i=1}^{|\mathcal{V}_+|} \eta_i v_i \quad \text{where} \quad \eta_i \geq 0 \quad i = 1, 2, \ldots, |\mathcal{V}_+|$$

and so

$$0 \geq -g^T g = \sum_{i=1}^{|\mathcal{V}_+|} \eta_i v_i^T g \geq 0$$

The only possibility is $g = 0$. □

3 Grids or Meshes in $\mathbf{R}^n$

A grid $\mathcal{G}_{\mathcal{V}}(h, x_o)$ is defined by a mesh size h, a point x_o on the grid, and a set of n linearly independent basis vectors $\mathcal{V}$, where

$$\mathcal{V} = \{v_j \in \mathbf{R}^n \ : \ j = 1, \ldots, n\}.$$

The points on the grid $\mathcal{G}$ are:

$$\mathcal{G}_{\mathcal{V}}(h) = \left\{ x \in R^n : x = x_o + h \sum_{i=1}^{n} \eta_i v_i \right\}.$$

with η_i integer. If the origin (or any other point) is known to lie on the grid then the dependence on x_o will usually be suppressed. The vectors $hv_1, \ldots, hv_n$ are the steps between adjacent grid points along each of the principal axes of the grid $\mathcal{G}$.

3.1 Grid Local Minima

Form a positive basis by extending the basis $\mathcal{V}$ to $\mathcal{V}_+$. A point $\hat{x} \in \mathcal{G}_{\mathcal{V}}(h, x_0)$ is a *grid local minimum* of f with respect to the positive basis $\mathcal{V}_+$ if and only if

$$f(\hat{x}) \leq f(\hat{x} + hv_i) \quad \forall v_i \in \mathcal{V}_+.$$

Normally each point $\hat{x} + hv_i$ is required to lie on $\mathcal{G}_{\mathcal{V}}(h, x_0)$. A necessary and sufficient condition for this is that each member of $\mathcal{V}_+$ is an integer combination of the members of $\mathcal{V}$.

Theorem 2

Let $\{\hat{x}_k \in \mathcal{G}_{\mathcal{V}}(h_k)\}$ be a sequence of grid local minima (with respect to $\mathcal{V}_+$) of f lying in a compact set $X \subset \mathbf{R}^n$ and let $f : \mathbf{R}^n \to \mathbf{R}$ be continuously differentiable on X. If $\lim_{k\to\infty} h_k = 0$ then all limit points of the sequence $\{\hat{x}_k\}$ are stationary points of f.

PROOF: Since $\hat{x}_k$ is a grid local minimizer on the grid $\mathcal{G}_{\mathcal{V}}(h_k)$

$$f\left(\hat{x}_k + h_k v_i\right) \geq f\left(\hat{x}_k\right), \quad i = 1, \ldots, |\mathcal{V}_+| \tag{2}$$

Letting $g(x), \hat{g}_k$ denote $\nabla f(x), \nabla f(\hat{x}_k)$ respectively, the definition of a derivative gives

$$\begin{aligned} f\left(\hat{x}_k + h_k v_i\right) - f\left(\hat{x}_k\right) &= \int_{s=0}^{h_k} v_i^T \left[g\left(\hat{x}_k + s v_i\right) - \hat{g}_k + \hat{g}_k\right] \, ds \\ &= h_k v_i^T \hat{g}_k + E \end{aligned} \tag{3}$$

where

$$E = \int_{s=0}^{h_k} v_i^T \left[g\left(\hat{x}_k + s v_i\right) - \hat{g}_k\right] \, ds$$

Using the Lipschitz condition, $||g(y) - g(x)|| \leq L||y - x||$, with Lipschitz constant L (which exists since X is compact) gives

$$|E| \leq \int_{s=0}^{h_k} L\|v_i\|^2 s \, ds \leq \tfrac{1}{2} L K^2 h_k^2, \tag{4}$$

where $\|v_i\| \leq K \quad \forall v_i \in \mathcal{V}_+$. Therefore, (2), (3) and (4) provide the bounds

$$v_i^T \hat{g}_k \geq -\tfrac{1}{2} L K^2 h_k, \quad \forall v_i \in \mathcal{V}_+. \tag{5}$$

If $\hat{x}_\infty$ is a limit point of the sequence $\{\hat{x}_k\}$ of grid local minima then there is a convergent subsequence $\{\hat{x}_{k_j}\}$, say, which has the unique limit $\hat{x}_\infty$. In the limit $k_j \to \infty$, (5) and the continuity of g give

$$v_i^T g(\hat{x}_\infty) \geq 0, \quad \forall v_i \in \mathcal{V}_+$$

and so $g(\hat{x}_\infty) = 0$ by Theorem 1. The choice of limit point of the sequence of grid local minima was arbitrary, so each limit point of the sequence of grid local minima is a stationary point of the objective function. □

A simple illustration of an application of Theorem 2 is to the algorithm of Hooke and Jeeves [6]. This algorithm consists of *exploratory moves* and *pattern moves* being made on a simple grid ($V = I$). An important feature of this method is that the mesh size h is only reduced (typically by a factor of 10) when a grid local minimizer is located. Therefore, it is only required to show that the method does generate a grid local minimizer on each grid in order to establish convergence via Theorem 2.

Theorem 3

Let x_o be the starting point for the method of Hooke and Jeeves [6]. Let f be continuously differentiable on the level set $S = \{x : f(x) \leq f(x_o)\}$ and let the set S be bounded. Then the method of Hooke and Jeeves generates a sequence of grid local minimizers whose limit points are stationary points of f.

PROOF: The condition that S is bounded implies that any grid $\mathcal{G}(h)$ for finite h has finitely many grid points satisfying $f(x) < f(x_o)$ for $x \in \mathcal{G}(h)$. If an exploratory move fails then a grid local minimizer has been found. If an exploratory move succeeds, it is followed by a sequence of pattern and exploratory moves which, if strictly lowering the function value, are continued. An infinite subsequence of pattern moves on any one grid is impossible since the function is evaluated only at grid points and there is a finite

number of these because of the conditions on S. Therefore, the search on the current grid terminates after a finite number of function evaluations with a failing exploratory move which characterizes a grid local minimizer. □

In the above proof it is assumed that the original algorithm of Hooke and Jeeves is followed. This is described unambiguously in the flow charts of the original paper [6]. The important point is that *success* of an exploratory move or a sequence of pattern/exploratory moves is defined by achieving a *strictly* lower function value. Some alternative interpretations allow weak inequality to define success (see, for example, [11]) but this is not considered advisable since convergence is not then assured.

The exploratory moves in Hooke and Jeeves method are made by probing along directions of a maximal positive basis. Clearly, any positive basis which generates points on a grid could be used without affecting the applicability of Theorem 3. For example, if $\mathcal{G}_{\mathcal{V}}(h, x_o)$ is the grid with $\mathcal{V} = \{v_j \in \mathbf{R}^n \ : \ j = 1, 2, \ldots, n\}$, then the exploratory phase need only probe along the directions defined by the minimal positive basis $\mathcal{V}_+ = \{v_j \in \mathbf{R}^n \ : \ j = 1, 2, \ldots, n+1\}$, where $v_{n+1} = -\sum_1^n v_j$ because, if $f(x_o) \leq f(x_o + hv_j)$, for all $j \in \mathcal{V}_+$ then x_o is a grid local minimizer on $\mathcal{G}_{\mathcal{V}}(h, x_o)$. This observation is also made in [8].

3.2 A numerical example

The original method of Hooke and Jeeves was applied to Rosenbrock's function from the standard start, $x_o = [-1.2, 1]^T$, (definitions and properties of the test functions used in this paper can be found in [9]). An initial mesh length $h_1 = .1$ (with $V = I$) was used with successive mesh lengths $h_{k+1} = h_k/10$. This algorithm required 229 function evaluations to determine a grid local minimizer with mesh length 10^{-5} at the solution. A modified algorithm using exploratory moves by probing the minimal positive basis (on the same grids) resulted in termination after 176 function evaluations. Note, however, that use of the minimal positive basis may not always be more efficient.

The method of Hooke and Jeeves is not usually recommended for unconstrained opti-

mization because its rate of convergence is often too slow to be of practicable use. Like steepest descent for gradient based optimization algorithms it enjoys good theoretical global convergence properties but is disappointing in practice (although it usually outperforms steepest descent). Theorem 2 assumes that the grid basis vectors remain unchanged as the mesh size decreases. However, it is possible to prove analogous results for the cases where each grid $\mathcal{G}_{\mathcal{V}_k}(h_k, x_k)$, uses a different set of basis vectors, $\mathcal{V}_k$, and origin x_k. This is considered further in [2], [3], where it is shown that the extra flexibility allowed by translating and realigning the axes of the grid can result in considerable improvements in efficiency.

4 Estimating Directional Derivatives

The problem of estimating derivatives is important in many applications. Most of the successful algorithms for numerical optimization rely on gradient information and/or directional derivatives so it is important to be able to obtain reliable estimates by difference methods (or other methods) when analytical derivatives are not directly available. A good description of the difficulties involved are summarised, for example, in the recent text [10, pp 168–191] which includes a treatment of the relatively recent techniques of *automatic differentiation*. In this section, it is shown how positive bases can provide a useful contribution to the problem of determining reliable numerical values for gradients or directional derivatives.

It is convenient to introduce the notation $g_v(x) = v^T \nabla f(x)$ for the directional derivative of $f(x)$ in the direction v and $g_{_V}$ for the (column) vector of directional derivatives $[g_{v_1}(x), \ldots, g_{v_n}(x)]^T$ corresponding to directions taken from the columns of the matrix $V = [v_1, \ldots, v_n]$. The dependence on x will be suppressed if it is clear from the context (e.g. $g_{_V} = V^T g$). A forward difference approximation to $g_v(x)$ is obtained by ignoring the $O(h)$ term in the formula

$$g_v = \frac{f(x + hv) - f(x)}{h} \quad + O(h) \tag{6}$$

but in order to maintain good relative accuracy as g approaches zero it becomes necessary

to switch from forward to central differences using the approximation

$$g_v = \frac{f(x+hv) - f(x-hv)}{2h} \quad + O(h^2). \tag{7}$$

at the cost of one extra function evaluation (n for $g_{_V}$). Of course, it is not just a question of which formula to use but also what value for h gives the best accuracy. This paper is concerned with using function values at grid points so the value of h is set by the mesh size, therefore it is only the former problem that is considered further.

4.1 When to switch?

Let $\mathcal{V}_+ = [V, -Ve]$ be a minimal positive basis and write

$$\begin{aligned} f_0 &= f(x), \\ f_i &= f(x + hv_i), \quad i = 1, \ldots, n, \\ f_{n+1} &= f(x - hVe). \end{aligned}$$

Now solve for $g_{_V}$, in the least squares sense, the $(n+1)$ equations:

$$\begin{aligned} (f_i - f_0)/h &= g_{v_i}, \quad i = 1, \ldots, n \\ (f_{n+1} - f_0)/h &= -(g_{v_1} + \cdots + g_{v_n}). \end{aligned} \tag{8}$$

If this solution differs significantly from the forward difference estimate then switch to central differences. The least squares solution to the equations (8) is easy because the coefficient matrix is $[I : -e]^T$. The solution is

$$g_{v_i} = (f_i - \bar{f})/h \tag{9}$$

where $\bar{f}$ denotes the mean value of the $(n+1)$ function values $\{f_j\}_1^{n+1}$. Note that, ignoring the $O(h)$ terms, the values (6) and (9) will be the same (and exact) when f is an affine function. Thus these values will usually agree, within reasonable relative accuracy, at points remote from a stationary point. However, if x is sufficiently close to a strict local minimizer then the two values cannot make good agreement because $f_0 < f_j$ will be satisfied and inspection of the formulas (6) and (9) show that they agree if and only

if $f_0 = \bar{f}$ which is impossible in this case. As an illustration, consider the problem of estimating $g_{_V}(x)$ for the well-known function of Rosenbrock, both at the standard start $x_1 = [-1.2, 1]^T$ and at the global minimizer $x_2 = [1, 1]^T$. Using $h = 10^{-6}$ and $V = I$ (so that $g_{_V}(x) = \nabla f(x)$), and applying the two formulas (6) and (9) in turn gives the approximations $[-215.5993, -87.9999]$ and $[-215.6000, -88.0006]$ respectively for $\nabla f(x_1)$. These both agree with each other to 5 significant figures and represent good approximations to the exact value $\nabla f(x_1) = [-215.6, -88]^T$. However, at x_2 applying the two formulas (6) and (9) in turn gives the approximations $10^{-4}[.40100, .99999]$ and $10^{-4}[.20033, -.00001]$ respectively for $\nabla f(x_2)$. Now there is very poor relative agreement between the two approximations which do not even agree in sign. The central difference formula at x_2 gives the much more accurate estimate $10^{-9}[.39997, -.00001]$. It could be argued that whenever the formula (6) gives a *small* value for the derivative then the extra function values in using (9) are well spent but it can be difficult to know in advance what is meant by *small* since it is not only problem dependent but also scale dependent. Therefore, an automatic approach is preferred. A strategy that has been successfully applied is to accept either of the estimates (6) or (9) as a suitable approximation to g_v if they agree to a relative tolerance of 10%. Otherwise, use formula (7) to estimate each component of $g_{_V}$. This test requires one extra function evaluation but it saves n when the test shows that central differences are not required and has the advantage of being completely automatic. This strategy was applied to a quasi-Newton algorithm based on the BFGS update using a line search based on the Goldstein conditions (see, for example, [10]) on the test functions described in [9]. For many problems the typical behaviour was to use forward differences for all but the last two or three iterations. However, for Meyer's function, only the first five of 332 iterations used forward differences. This is a very badly scaled problem which provides a severe test for any optimization algorithm. The strategy used above had no difficulty in finding the required solution.

5 Concluding remarks

It has been shown that positive bases can have useful applications in numerical optimization. The simple ideas outlined in this paper have also been applied and extended by the authors to provide significant improvements to the convergence properties of existing algorithms by minor modifications and to develop new convergent algorithms for numerical optimization that do not require the evaluation of analytical derivatives. Further examples can be found in the reports [2, 3, 4] and the Master's thesis [1].

References

[1] D. Byatt. Convergent variants of the Nelder-Mead algorithm. Master's thesis, University of Canterbury, Department of Mathematics & Statistics, Christchurch, New Zealand, June 2000.

[2] I. D. Coope and C. J. Price. A direct search conjugate directions algorithm for unconstrained minimization. Technical Report 188, Department of Mathematics & Statistics, University of Canterbury, Christchurch, New Zealand, 1999.

[3] I. D. Coope and C. J. Price. On the convergence of grid-based methods for unconstrained minimization. Technical Report 180, Department of Mathematics & Statistics, University of Canterbury, Christchurch, New Zealand, 1999.

[4] I. D. Coope and C. J. Price. Frame-based methods for numerical optimization. Technical report, Department of Mathematics & Statistics, University of Canterbury, Christchurch, New Zealand, in preparation.

[5] C. Davis. Theory of positive linear dependence. *American Journal of mathematics*, pages 733–746, 1954.

[6] R. Hooke and T. A. Jeeves. Direct search solution of numerical and statistical problems. *Journal of the Association for Computing Machinery (ACM)*, 8:212–219, 1961.

[7] R. M. Lewis and V. Torczon. Pattern search algorithms for bound constrained optimization. *SIAM Journal on Optimization*, 9(4):1082–1099, 1999.

[8] R. M. Lewis and V. Torczon. Rank ordering and positive bases in pattern search algorithms. *SIAM Journal on Optimization*, to appear.

[9] J. J. Moré, B. S. Garbow, and K. E. Hillstrom. Testing unconstrained optimization software. *ACM Trans. Math. Software*, 7(1):17–41, 1981.

[10] J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer series in operations research, Springer-Verlag, New York, N.Y., 1999.

[11] W. H. Swann. Direct search methods. In W. Murray, editor, *Numerical methods for unconstrained optimization*, pages 13–28. Academic Press, 1972.

[12] Wen-Ci Yu. Positive basis and a class of direct search techniques. *Scientia Sinica (Zhongguo Kexue), Special Issue 1 on Mathematics*, pages 53–68, 1979.