Chapter 10
Quasi-Newton Methods

In this chapter we take another approach toward the development of methods lying somewhere intermediate to steepest descent and Newton's method. Again working under the assumption that evaluation and use of the Hessian matrix is impractical or costly, the idea underlying quasi-Newton methods is to use an approximation to the inverse Hessian in place of the true inverse that is required in Newton's method. The form of the approximation varies among different methods—ranging from the simplest where it remains fixed throughout the iterative process, to the more advanced where improved approximations are built up on the basis of information gathered during the descent process.

The quasi-Newton methods that build up an approximation to the inverse Hessian are analytically the most sophisticated methods discussed in this book for solving unconstrained problems and represent the culmination of the development of algorithms through detailed analysis of the quadratic problem. As might be expected, the convergence properties of these methods are somewhat more difficult to discover than those of simpler methods. Nevertheless, we are able, by continuing with the same basic techniques as before, to illuminate their most important features.

In the course of our analysis we develop two important generalizations of the method of steepest descent and its corresponding convergence rate theorem. The first, discussed in Sect. 10.1, modifies steepest descent by taking as the direction vector a positive definite transformation of the negative gradient. The second, discussed in Sect. 10.8, is a combination of steepest descent and Newton's method. Both of these fundamental methods have convergence properties analogous to those of steepest descent.

D.G. Luenberger, Y. Ye, *Linear and Nonlinear Programming*, International Series in Operations Research & Management Science 228,
DOI 10.1007/978-3-319-18842-3_10

10.1 Modified Newton Method

A very basic iterative process for solving the problem

$$\text{minimize} \quad f(\mathbf{x})$$

which includes as special cases most of our earlier ones is

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k \mathbf{S}_k \nabla f(\mathbf{x}_k)^T \tag{10.1}$$

where $\mathbf{S}_k$ is a symmetric $n \times n$ matrix and where, as usual, α_k is chosen to minimize $f(\mathbf{x}_{k+1})$. If $\mathbf{S}_k$ is the inverse of the Hessian of f, we obtain Newton's method, while if $\mathbf{S}_k = \mathbf{I}$ we have steepest descent. It would seem to be a good idea, in general, to select $\mathbf{S}_k$ as an approximation to the inverse of the Hessian. We examine that philosophy in this section.

First, we note, as in Sect. 8.5, that in order that the process (10.1) be guaranteed to be a descent method for small values of α, it is necessary in general to require that $\mathbf{S}_k$ be positive definite. We shall therefore always impose this as a requirement.

Because of the similarity of the algorithm (10.1) with steepest descent[†] it should not be surprising that its convergence properties are similar in character to our earlier results. We derive the actual rate of convergence by considering, as usual, the standard quadratic problem with

$$f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{Q}\mathbf{x} - \mathbf{b}^T\mathbf{x}, \tag{10.2}$$

where $\mathbf{Q}$ is symmetric and positive definite. For this case we can find an explicit expression for α_k in (10.1). The algorithm becomes

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k \mathbf{S}_k \mathbf{g}_k, \tag{10.3a}$$

where

$$\mathbf{g}_k = \mathbf{Q}\mathbf{x}_k - \mathbf{b} \tag{10.3b}$$

$$\alpha_k = \frac{\mathbf{g}_k^T \mathbf{S}_k \mathbf{g}_k}{\mathbf{g}_k^T \mathbf{S}_k \mathbf{Q} \mathbf{S}_k \mathbf{g}_k}. \tag{10.3c}$$

We may then derive the convergence rate of this algorithm by slightly extending the analysis carried out for the method of steepest descent.

Modified Newton Method Theorem (Quadratic Case). *Let* $\mathbf{x}^*$ *be the uniqueminimum point of* f, *and define* $E(\mathbf{x}) = \frac{1}{2}(\mathbf{x} - \mathbf{x}^*)^T \mathbf{Q}(\mathbf{x} - \mathbf{x}^*)$.

[†] The algorithm (10.1) is sometimes referred to as the *method of deflected gradients*, since the direction vector can be thought of as being determined by deflecting the gradient through multiplication by $\mathbf{S}_k$.

Then for the algorithm (10.3) *there holds at every step k*

$$E(\mathbf{x}_{k+1}) \leqslant \left(\frac{B_k - b_k}{B_k + b_k}\right)^2 E(\mathbf{x}_k), \tag{10.4}$$

where b_k and B_k are, respectively, the smallest and largest eigenvalues of thematrix $\mathbf{S}_k\mathbf{Q}$.

Proof. We have by direct substitution

$$\frac{E(\mathbf{x}_k) - E(\mathbf{x}_{k+1})}{E(\mathbf{x}_k)} = \frac{(\mathbf{g}_k^T\mathbf{S}_k\mathbf{g}_k)^2}{(\mathbf{g}_k^T\mathbf{S}_k\mathbf{Q}\mathbf{S}_k\mathbf{g}_k)(\mathbf{g}_k^T\mathbf{Q}^{-1}\mathbf{g}_k)}.$$

Letting $\mathbf{T}_k = \mathbf{S}_k^{1/2}\mathbf{Q}\mathbf{S}_k^{1/2}$ and $\mathbf{p}_k = \mathbf{S}_k^{1/2}\mathbf{g}_k$ we obtain

$$\frac{E(\mathbf{x}_k) - E(\mathbf{x}_{k+1})}{E(\mathbf{x}_k)} = \frac{(\mathbf{p}_k^T\mathbf{P}_k)^2}{(\mathbf{p}_k^T\mathbf{T}_k\mathbf{p}_k)(\mathbf{p}_k^T\mathbf{T}_k^{-1}\mathbf{p}_k)}.$$

From the Kantorovich inequality we obtain easily

$$E(\mathbf{x}_{k+1}) \leqslant \left(\frac{B_k - b_k}{B_k + b_k}\right)^2 E(\mathbf{x}_k),$$

where b_k and B_k are the smallest and largest eigenvalues of $\mathbf{T}_k$. Since $\mathbf{S}_k^{1/2}\mathbf{T}_k\mathbf{S}_k^{-1/2} = \mathbf{S}_k\mathbf{Q}$, we see that $\mathbf{S}_k\mathbf{Q}$ is similar to $\mathbf{T}_k$ and therefore has the same eigenvalues. ∎

This theorem supports the intuitive notion that for the quadratic problem one should strive to make $\mathbf{S}_k$ close to $\mathbf{Q}^{-1}$ since then both b_k and B_k would be close to unity and convergence would be rapid. For a nonquadratic objective function f the analog to $\mathbf{Q}$ is the Hessian $\mathbf{F}(\mathbf{x})$, and hence one should try to make $\mathbf{S}_k$ close to $\mathbf{F}(\mathbf{x}_k)^{-1}$.

Two remarks may help to put the above result in proper perspective. The first remark is that both the algorithm (10.1) and the theorem stated above are only simple, minor, and natural extensions of the work presented in Chap. 8 on steepest descent. As such the result of this section can be regarded, correspondingly, not as a new idea but as an extension of the basic result on steepest descent. The second remark is that this one simple result when properly applied can quickly characterize the convergence properties of some fairly complex algorithms. Thus, rather than an isolated result concerned with a specific form of algorithm, the theorem above should be regarded as a general tool for convergence analysis. It provides significant insight into various quasi-Newton methods discussed in this chapter.

A Classical Method

We conclude this section by mentioning the *classical modified Newton's method*, a standard method for approximating Newton's method without evaluating $\mathbf{F}(\mathbf{x}_k)^{-1}$ for each k. We set

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k[\mathbf{F}(\mathbf{x}_0)]^{-1}\nabla f(\mathbf{x}_k)^T. \tag{10.5}$$

In this method the Hessian at the initial point $\mathbf{x}_0$ is used throughout the process. The effectiveness of this procedure is governed largely by how fast the Hessian is changing—in other words, by the magnitude of the third derivatives of f.

10.2 Construction of the Inverse

The fundamental idea behind most quasi-Newton methods is to try to construct the inverse Hessian, or an approximation of it, using information gathered as the descent process progresses. The current approximation $\mathbf{H}_k$ is then used at each stage to define the next descent direction by setting $\mathbf{S}_k = \mathbf{H}_k$ in the modified Newton method. Ideally, the approximations converge to the inverse of the Hessian at the solution point and the overall method behaves somewhat like Newton's method. In this section we show how the inverse Hessian can be built up from gradient information obtained at various points.

Let f be a function on E^n that has continuous second partial derivatives. If for two points $\mathbf{x}_{k+1}$, $\mathbf{x}_k$ we define $\mathbf{g}_{k+1} = \nabla f(\mathbf{x}_{k+1})^T$, $\mathbf{g}_k = \nabla f(\mathbf{x}_k)^T$ and $\mathbf{p}_k = \mathbf{x}_{k+1} - \mathbf{x}_k$, then

$$\mathbf{g}_{k+1} - \mathbf{g}_k \cong \mathbf{F}(\mathbf{x}_k)\mathbf{p}_k. \tag{10.6}$$

If the Hessian, $\mathbf{F}$, is constant, then we have

$$\mathbf{q}_k \equiv \mathbf{g}_{k+1} - \mathbf{g}_k = \mathbf{F}\mathbf{p}_k, \tag{10.7}$$

and we see that evaluation of the gradient at two points gives information about $\mathbf{F}$. If n linearly independent directions $\mathbf{p}_0, \mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_{n-1}$ and the corresponding $\mathbf{q}_k$'s are known, then $\mathbf{F}$ is uniquely determined. Indeed, letting $\mathbf{P}$ and $\mathbf{Q}$ be the $n \times n$ matrices with columns $\mathbf{p}_k$ and $\mathbf{q}_k$ respectively, we have $\mathbf{F} = \mathbf{Q}\mathbf{P}^{-1}$.

It is natural to attempt to construct successive approximations $\mathbf{H}_k$ to $\mathbf{F}^{-1}$ based on data obtained from the first k steps of a descent process in such a way that if $\mathbf{F}$ were constant the approximation would be consistent with (10.7) for these steps. Specifically, if $\mathbf{F}$ were constant $\mathbf{H}_{k+1}$ would satisfy

$$\mathbf{H}_{k+1}\mathbf{q}_i = \mathbf{p}_i, \ 0 \leqslant i \leqslant k. \tag{10.8}$$

After n linearly independent steps we would then have $\mathbf{H}_n = \mathbf{F}^{-1}$.

For any $k < n$ the problem of constructing a suitable $\mathbf{H}_k$, with in general serves as an approximation to the inverse Hessian and which in the case of constant $\mathbf{F}$ satisfies (10.8), admits an infinity of solutions, since there are more degrees of freedom than there are constraints. Thus a particular method can take into account additional considerations. We discuss below one of the simplest schemes that has been proposed.

Rank One Correction

Since $\mathbf{F}$ and $\mathbf{F}^{-1}$ are symmetric, it is natural to require that $\mathbf{H}_k$, the approximation to $\mathbf{F}^{-1}$, be symmetric. We investigate the possibility of defining a recursion of the form

$$\mathbf{H}_{k+1} = \mathbf{H}_k + a_k \mathbf{z}_k \mathbf{z}_k^T, \tag{10.9}$$

which preserves symmetry. The vector $\mathbf{z}_k$ and the constant a_k define a matrix of (at most) rank one, by which the approximation to the inverse is updated. We select them so that (10.8) is satisfied. Setting i equal to k in (10.8) and substituting (10.9) we obtain

$$\mathbf{p}_k = \mathbf{H}_{k+1}\mathbf{q}_k = \mathbf{H}_k\mathbf{q}_k + a_k \mathbf{z}_k \mathbf{z}_k^T \mathbf{q}_k. \tag{10.10}$$

Taking the inner product with $\mathbf{q}_k$ we have

$$\mathbf{q}_k^T \mathbf{p}_k - \mathbf{q}_k^T \mathbf{H}_k \mathbf{q}_k = a_k \left(\mathbf{z}_k^T \mathbf{q}_k\right)^2 \tag{10.11}$$

On the other hand, using (10.10) we may write (10.9) as

$$\mathbf{H}_{k+1} = \mathbf{H}_k + \frac{(\mathbf{p}_k - \mathbf{H}_k\mathbf{q}_k)(\mathbf{p}_k - \mathbf{H}_k\mathbf{q}_k)^T}{a_k \left(\mathbf{z}_k^T \mathbf{q}_k\right)^2},$$

which in view of (10.11) leads finally to

$$\mathbf{H}_{k+1} = \mathbf{H}_k + \frac{(\mathbf{p}_k - \mathbf{H}_k\mathbf{q}_k)(\mathbf{p}_k - \mathbf{H}_k\mathbf{q}_k)^T}{\mathbf{q}_k^T(\mathbf{p}_k - \mathbf{H}_k\mathbf{q}_k)}. \tag{10.12}$$

We have determined what a rank one correction must be if it is to satisfy (10.8) for $i = k$. It remains to be shown that, for the case where $\mathbf{F}$ is constant, (10.8) is also satisfied for $i < k$. This in turn will imply that the rank one recursion converges to $\mathbf{F}^{-1}$ after at most n steps.

Theorem. *Let $\boldsymbol{F}$ be a fixed symmetric matrix and suppose that $\mathbf{p}_0, \boldsymbol{p}_1, \mathbf{p}_2, \ldots, \boldsymbol{p}_k$ are given vectors. Define the vectors $\mathbf{q}_i = \boldsymbol{F}\boldsymbol{p}_i$, $i = 0, 1, 2, \ldots, k$.*

Starting with any initial symmetric matrix $\mathbf{H}_0$ let

$$\boldsymbol{H}_{i+1} = \boldsymbol{H}_i + \frac{(\boldsymbol{p}_i - \boldsymbol{H}_i\boldsymbol{q}_i)(\boldsymbol{p}_i - \boldsymbol{H}_i\boldsymbol{q}_i)^T}{\boldsymbol{q}_i^T(\boldsymbol{p}_i - \boldsymbol{H}_i\boldsymbol{q}_i)}. \tag{10.13}$$

Then

$$\boldsymbol{p}_i = \boldsymbol{H}_{k+1}\boldsymbol{q}_i \quad for \quad i \leqslant k. \tag{10.14}$$

Proof. The proof is by induction. Suppose it is true for $\mathbf{H}_k$, and $i \leqslant k-1$. The relation was shown above to be true for $\mathbf{H}_{k+1}$ and $i = k$. For $i < k$

$$\mathbf{H}_{k+1}\mathbf{q}_i = \mathbf{H}_k\mathbf{q}_i + \mathbf{y}_k(\mathbf{p}_k^T\mathbf{q}_i - \mathbf{q}_k^T\mathbf{H}_k\mathbf{q}_i), \tag{10.15}$$

where

$$\mathbf{y}_k = \frac{(\mathbf{p}_k - \mathbf{H}_k\mathbf{q}_k)}{\mathbf{q}_k^T(\mathbf{p}_k - \mathbf{H}_k\mathbf{q}_k)}.$$

By the induction hypothesis, (10.15) becomes

$$\mathbf{H}_{k+1}\mathbf{q}_i = \mathbf{p}_i + \mathbf{y}_k\left(\mathbf{p}_k^T\mathbf{q}_i - \mathbf{q}_k^T\mathbf{p}_i\right).$$

From the calculation

$$\mathbf{q}_k^T\mathbf{p}_i = \mathbf{p}_k^T\mathbf{F}\mathbf{p}_i = \mathbf{p}_k^T\mathbf{q}_i,$$

it follows that the second term vanishes. ∎

To incorporate the approximate inverse Hessian in a descent procedure while simultaneously improving it, we calculate the direction $\mathbf{d}_k$ From

$$\mathbf{d}_k = -\mathbf{H}_k\mathbf{g}_k$$

and then minimize $f(\mathbf{x}_k + \alpha\mathbf{d}_k)$ with respect to $\alpha \geqslant 0$. This determines $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k\mathbf{d}_k$, $\mathbf{p}_k = \alpha_k\mathbf{d}_k$, and $\mathbf{g}_{k+1}$. Then $\mathbf{H}_{k+1}$ can be calculated according to (10.12).

There are some difficulties with this simple rank one procedure. First, the updating formula (10.12) preserves positive definiteness only if $\mathbf{q}_k^T(\mathbf{p}_k - \mathbf{H}_k\mathbf{q}_k) > 0$, which cannot be guaranteed (see Exercise 6). Also, even if $\mathbf{q}_k^T(\mathbf{p}_k - \mathbf{H}_k\mathbf{q}_k)$ is positive, it may be small, which can lead to numerical difficulties. Thus, although an excellent simple example of how information gathered during the descent process can in principle be used to update an approximation to the inverse Hessian, the rank one method possesses some limitations.

10.3 Davidon-Fletcher-Powell Method

The earliest, and certainly one of the most clever schemes for constructing the inverse Hessian, was originally proposed by Davidon and later developed by Fletcher and Powell. It has the fascinating and desirable property that, for a quadratic objective, it simultaneously generates the directions of the conjugate gradient method while constructing the inverse Hessian. At each step the inverse Hessian is updated by the sum of two symmetric rank one matrices, and this scheme is therefore often referred to as a *rank two correction procedure*. The method is also often referred to as the *variable metric method*, the name originally suggested by Davidon.

The procedure is this: Starting with any symmetric positive definite matrix $\mathbf{H}_0$, any point $\mathbf{x}_0$, and with $k = 0$,

Step 1. Set $\mathbf{d}_k = -\mathbf{H}_k\mathbf{g}_k$.

Step 2. Minimize $f(\mathbf{x}_k + \alpha\mathbf{d}_k)$ with respect to $\alpha \geqslant 0$ to obtain $\mathbf{x}_{k+1}$, $\mathbf{p}_k = \alpha_k\mathbf{d}_k$, and $\mathbf{g}_{k+1}$.

Step 3. Set $\mathbf{q}_k = \mathbf{g}_{k+1} - \mathbf{g}_k$ and

$$\mathbf{H}_{k+1} = \mathbf{H}_k + \frac{\mathbf{p}_k \mathbf{p}_k^T}{\mathbf{p}_k^T \mathbf{q}_k} - \frac{\mathbf{H}_k \mathbf{q}_k \mathbf{q}_k^T \mathbf{H}_k}{\mathbf{q}_k^T \mathbf{H}_k \mathbf{q}_k}. \tag{10.16}$$

Update k and return to Step 1.

Positive Definiteness

We first demonstrate that if $\mathbf{H}_k$ is positive definite, then so is $\mathbf{H}_{k+1}$. For any $\mathbf{x} \in E^n$ we have

$$\mathbf{x}^T \mathbf{H}_{k+1} \mathbf{x} = \mathbf{x}^T \mathbf{H}_k \mathbf{x} + \frac{(\mathbf{x}^T \mathbf{p}_k)^2}{\mathbf{p}_k^T \mathbf{q}_k} - \frac{(\mathbf{x}^T \mathbf{H}_k \mathbf{q}_k)^2}{\mathbf{q}_k^T \mathbf{H}_k \mathbf{q}_k}. \tag{10.17}$$

Defining $\mathbf{a} = \mathbf{H}_k^{1/2} \mathbf{x}$, $\mathbf{b} = \mathbf{H}_k^{1/2} \mathbf{q}_k$ we may rewrite (10.17) as

$$\mathbf{x}^T \mathbf{H}_{k+1} \mathbf{x} = \frac{(\mathbf{a}^T \mathbf{a})(\mathbf{b}^T \mathbf{b}) - (\mathbf{a}^T \mathbf{b})^2}{(\mathbf{b}^T \mathbf{b})} + \frac{(\mathbf{x}^T \mathbf{p}_k)^2}{\mathbf{p}_k^T \mathbf{q}_k}.$$

We also have

$$\mathbf{p}_k^T \mathbf{q}_k = \mathbf{p}_k^T \mathbf{g}_{k+1} - \mathbf{p}_k^T \mathbf{g}_k = -\mathbf{p}_k^T \mathbf{g}_k, \tag{10.18}$$

since

$$\mathbf{p}_k^T \mathbf{g}_{k+1} = 0, \tag{10.19}$$

because $\mathbf{x}_{k+1}$ is the minimum point of f along $\mathbf{p}_k$. Thus by definition of $\mathbf{p}_k$

$$\mathbf{p}_k^T \mathbf{q}_k = \alpha_k \mathbf{g}_k^T \mathbf{H}_k \mathbf{g}_k, \tag{10.20}$$

and hence

$$\mathbf{x}^T \mathbf{H}_{k+1} \mathbf{x} = \frac{(\mathbf{a}^T \mathbf{a})(\mathbf{b}^T \mathbf{b}) - (\mathbf{a}^T \mathbf{b})^2}{(\mathbf{b}^T \mathbf{b})} + \frac{(\mathbf{x}^T \mathbf{p}_k)^2}{\alpha_k \mathbf{g}_k^T \mathbf{H}_k \mathbf{g}_k}. \tag{10.21}$$

Both terms on the right of (10.21) are nonnegative—the first by the Cauchy–Schwarz inequality. We must only show they do not both vanish simultaneously. The first term vanishes only if a and $\mathbf{b}$ are proportional. This in turn implies that $\mathbf{x}$ and $\mathbf{q}_k$ are proportional, say $\mathbf{x} = \beta \mathbf{q}_k$. In that case, however,

$$\mathbf{p}_k^T \mathbf{x} = \beta \mathbf{p}_k^T \mathbf{q}_k = \beta \alpha_k \mathbf{g}_k^T \mathbf{H}_k \mathbf{g}_k \neq 0$$

from (10.20). Thus $\mathbf{x}^T \mathbf{H}_{k+1} \mathbf{x} > 0$ for all nonzero $\mathbf{x}$.

It is of interest to note that in the proof above the fact that α_k is chosen as the minimum point of the line search was used in (10.19), which led to the important conclusion $\mathbf{p}_k^T \mathbf{q}_k > 0$. Actually any α_k, whether the minimum point or not, that gives $\mathbf{p}_k^T \mathbf{q}_k > 0$ can be used in the algorithm, and $\mathbf{H}_{k+1}$ will be positive definite (see Exercises 8 and 9).

Finite Step Convergence

We assume now that f is quadratic with (constant) Hessian $\mathbf{F}$. We show in this case that the Davidon-Fletcher-Powell method produces direction vectors $\mathbf{p}_k$ that are $\mathbf{F}$-orthogonal and that if the method is carried n steps then $\mathbf{H}_n = \mathbf{F}^{-1}$.

Theorem. *If f is quadratic with positive definite Hessian $\mathbf{F}$, then for the Davidon-Fletcher-Powell method*

$$\boldsymbol{p}_i^T \boldsymbol{F} \boldsymbol{p}_j = 0, \qquad 0 \leqslant i < j \leqslant k \tag{10.22}$$

$$\boldsymbol{H}_{k+1} \boldsymbol{F} \boldsymbol{p}_i = \boldsymbol{p}_i \quad \textit{for} \quad 0 \leqslant i \leqslant k. \tag{10.23}$$

Proof. We note that for the quadratic case

$$\mathbf{q}_k = \mathbf{g}_{k+1} - \mathbf{g}_k = \mathbf{F}\mathbf{x}_{k+1} - \mathbf{F}\mathbf{x}_k = \mathbf{F}\mathbf{p}_k. \tag{10.24}$$

Also

$$\mathbf{H}_{k+1}\mathbf{F}\mathbf{p}_k = \mathbf{H}_{k+1}\mathbf{q}_k = \mathbf{p}_k \tag{10.25}$$

from (10.16).

We now prove (10.22) and (10.23) by induction. From (10.25) we see that they are true for $k = 0$. Assuming they are true for $k-1$, we prove they are true for k. We have

$$\mathbf{g}_k = \mathbf{g}_{i+1} + \mathbf{F}(\mathbf{p}_{i+1} + \cdots + \mathbf{p}_{k-1}).$$

Therefore from (10.22) and (10.19)

$$\mathbf{p}_i^T \mathbf{g}_k = \mathbf{p}_i^T \mathbf{g}_{i+1} = 0 \quad \text{for} \quad 0 \leqslant i < k. \tag{10.26}$$

Hence from (10.23)

$$\mathbf{p}_i^T \mathbf{F} \mathbf{H}_k \mathbf{g}_k = 0. \tag{10.27}$$

Thus since $\mathbf{p}_k = -\alpha_k \mathbf{H}_k \mathbf{g}_k$ and since $\alpha_k \neq 0$, we obtain

$$\mathbf{p}_i^T \mathbf{F} \mathbf{p}_k = 0 \quad \text{for} \quad i < k, \tag{10.28}$$

which proves (10.22) for k.

Now since from (10.23) for $k-1$, (10.24) and (10.28)

$$\mathbf{q}_k^T \mathbf{H}_k \mathbf{F} \mathbf{p}_i = \mathbf{q}_k^T \mathbf{p}_i = \mathbf{p}_k^T \mathbf{F} \mathbf{p}_i = 0, \quad 0 \leqslant i < k$$

we have

$$\mathbf{H}_{k+1}\mathbf{F}\mathbf{p}_i = \mathbf{H}_k \mathbf{F} \mathbf{p}_i = \mathbf{p}_i, \quad 0 \leqslant i < k.$$

This together with (10.25) proves (10.23) for k. ∎

Since the $\mathbf{p}_k$'s are $\mathbf{F}$-orthogonal and since we minimize f successively in these directions, we see that the method is a conjugate direction method. Furthermore,

if the initial approximation $\mathbf{H}_0$ is taken equal to the identity matrix, the method becomes the conjugate gradient method. In any case the process obtains the overall minimum point within n steps.

Finally, (10.23) shows that $\mathbf{p}_0, \mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_k$ are eigenvectors corresponding to unity eigenvalue for the matrix $\mathbf{H}_{k+1}\mathbf{F}$. These eigenvectors are linearly independent, since they are $\mathbf{F}$-orthogonal, and therefore $\mathbf{H}_n = \mathbf{F}^{-1}$.

10.4 The Broyden Family

The updating formulae for the inverse Hessian considered in the previous two sections are based on satisfying

$$\mathbf{H}_{k+1}\mathbf{q}_i = \mathbf{p}_i, \quad 0 \leqslant i \leqslant k, \tag{10.29}$$

which is derived from the relation

$$\mathbf{q}_i = \mathbf{F}\mathbf{p}_i, \qquad 0 \leqslant i \leqslant k, \tag{10.30}$$

which would hold in the purely quadratic case. It is also possible to update approximations to the Hessian $\mathbf{F}$ itself, rather than its inverse. Thus, denoting the kth approximation of $\mathbf{F}$ by $\mathbf{B}_k$, we would, analogously, seek to satisfy

$$\mathbf{q}_i = \mathbf{B}_{k+1}\mathbf{p}_i, \quad 0 \leqslant i \leqslant k. \tag{10.31}$$

Equation (10.31) has exactly the same form as (10.29) except that $\mathbf{q}_i$ and $\mathbf{p}_i$ are interchanged and $\mathbf{H}$ is replaced by B. It should be clear that this implies that any update formula for $\mathbf{H}$ derived to satisfy (10.29) can be transformed into a corresponding update formula for B. Specifically, given any update formula for $\mathbf{H}$, the *complementary* formula is found by interchanging the roles of $\mathbf{B}$ and $\mathbf{H}$ and of $\mathbf{q}$ and $\mathbf{p}$. Likewise, any updating formula for $\mathbf{B}$ that satisfies (10.31) can be converted by the same process to a complementary formula for updating H. It is easily seen that taking the complement of a complement restores the original formula.

To illustrate complementary formulae, consider the rank one update of Sect. 10.2, which is

$$\mathbf{H}_{k+1} = \mathbf{H}_k + \frac{(\mathbf{p}_k - \mathbf{H}_k\mathbf{q}_k)(\mathbf{p}_k - \mathbf{H}_k\mathbf{q}_k)^T}{\mathbf{q}_k^T(\mathbf{p}_k - \mathbf{H}_k\mathbf{q}_k)}. \tag{10.32}$$

The corresponding complementary formula is

$$\mathbf{B}_{k+1} = \mathbf{B}_k + \frac{(\mathbf{q}_k - \mathbf{B}_k\mathbf{p}_k)(\mathbf{q}_k - \mathbf{B}_k\mathbf{p}_k)^T}{\mathbf{p}_k^T(\mathbf{q}_k - \mathbf{B}_k\mathbf{p}_k)}. \tag{10.33}$$

Likewise, the Davidon-Fletcher-Powell (or simply DFP) formula is

$$\mathbf{H}_{k+1}^{\mathrm{DFP}} = \mathbf{H}_k + \frac{\mathbf{p}_k \mathbf{p}_k^T}{\mathbf{p}_k^T \mathbf{q}_k} - \frac{\mathbf{H}_k \mathbf{q}_k \mathbf{q}_k^T \mathbf{H}_k}{\mathbf{q}_k^T \mathbf{H}_k \mathbf{q}_k}, \tag{10.34}$$

and its complement is

$$\mathbf{B}_{k+1} = \mathbf{B}_k + \frac{\mathbf{q}_k \mathbf{q}_k^T}{\mathbf{q}_k^T \mathbf{p}_k} - \frac{\mathbf{B}_k \mathbf{p}_k \mathbf{p}_k^T \mathbf{B}_k}{\mathbf{p}_k^T \mathbf{B}_k \mathbf{p}_k}. \tag{10.35}$$

This last update is known as the Broyden-Fletcher-Goldfarb-Shanno update of $\mathbf{B}_k$, and it plays an important role in what follows.

Another way to convert an updating formula for $\mathbf{H}$ to one for $\mathbf{B}$ or vice versa is to take the inverse. Clearly, if

$$\mathbf{H}_{k+1} \mathbf{q}_i = \mathbf{p}_i, \quad 0 \leqslant i \leqslant k, \tag{10.36}$$

then

$$\mathbf{q}_i = \mathbf{H}_{k+1}^{-1} \mathbf{p}_i, \quad 0 \leqslant i \leqslant k, \tag{10.37}$$

which implies that $\mathbf{H}_{k+1}^{-1}$ satisfies (10.31), the criterion for an update of B. Also, most importantly, the inverse of a rank two formula is itself a rank two formula.

The new formula can be found explicitly by two applications of the general inversion identity (often referred to as the Sherman-Morrison formula)

$$\left[\mathbf{A} + \mathbf{a}\mathbf{b}^T\right]^{-1} = \mathbf{A}^{-1} - \frac{\mathbf{A}^{-1}\mathbf{a}\mathbf{b}^T\mathbf{A}^{-1}}{1 + \mathbf{b}^T\mathbf{A}^{-1}\mathbf{a}}, \tag{10.38}$$

where A is an $n \times n$ matrix, and a and $\mathbf{b}$ are n-vectors, which is valid provided the inverses exist. (This is easily verified by multiplying through by $\mathbf{A} + \mathbf{a}\mathbf{b}^T$.)

The Broyden-Fletcher-Goldfard-Shanno update for $\mathbf{B}$ produces, by taking the inverse, a corresponding update for $\mathbf{H}$ of the form

$$\mathbf{H}_{k+1}^{\mathrm{BFGS}} = \mathbf{H}_k + \left(\frac{1 + \mathbf{q}_k^T \mathbf{H}_k \mathbf{q}_k}{\mathbf{q}_k^T \mathbf{q}_k} \right) \frac{\mathbf{p}_k \mathbf{p}_k^T}{\mathbf{p}_k^T \mathbf{q}_k} - \frac{\mathbf{p}_k \mathbf{q}_k^T \mathbf{H}_k + \mathbf{H}_k \mathbf{q}_k \mathbf{q}_k^T}{\mathbf{q}_k^T \mathbf{p}_k}. \tag{10.39}$$

This is an important update formula that can be used exactly like the DFP formula. Numerical experiments have repeatedly indicated that its performance is superior to that of the DFP formula, and for this reason it is now generally preferred.

It can be noted that both the DFP and the BFGS updates have symmetric rank two corrections that are constructed from the vectors $\mathbf{p}_k$ and $\mathbf{H}_k \mathbf{q}_k$. Weighted combinations of these formulae will therefore also be of this same type (symmetric, rank two, and constructed from $\mathbf{p}_k$ and $\mathbf{H}_k \mathbf{q}_k$). This observation naturally leads to consideration of a whole collection of updates, known as the Broyden family, defined by

$$\mathbf{H}^{\phi} = (1 - \phi)\mathbf{H}^{\mathrm{DFP}} + \phi \mathbf{H}^{\mathrm{BFGS}}, \tag{10.40}$$

where ϕ is a parameter that may take any real value. Clearly $\phi = 0$ and $\phi = 1$ yield the DFP and BFGS updates, respectively. The Broyden family also includes the rank one update (see Exercise 12).

An explicit representation of the Broyden family can be found, after a fair amount of algebra, to be

$$\mathbf{H}_{k+1}^{\phi} = \mathbf{H}_k + \frac{\mathbf{p}_k \mathbf{p}_k^T}{\mathbf{p}_k^T \mathbf{q}_k} - \frac{\mathbf{H}_k \mathbf{q}_k \mathbf{q}_k^T \mathbf{H}_k}{\mathbf{q}_k^T \mathbf{H}_k \mathbf{q}_k} + \phi \mathbf{v}_k \mathbf{v}_k^T = \mathbf{H}_{k+1}^{\mathbf{DFP}} + \phi \mathbf{v}_k \mathbf{v}_k^T, \qquad (10.41)$$

where

$$\mathbf{v}_k = (\mathbf{q}_k^T \mathbf{H}_k \mathbf{q}_k)^{1/2} \left(\frac{\mathbf{p}_k}{\mathbf{p}_k^T \mathbf{q}_k} - \frac{\mathbf{H}_k \mathbf{q}_k}{\mathbf{q}_k^T \mathbf{H}_k \mathbf{q}_k} \right).$$

This form will be useful in some later developments.

A *Broyden method* is defined as a quasi-Newton method in which at each iteration a member of the Broyden family is used as the updating formula. The parameter ϕ is, in general, allowed to vary from one iteration to another, so a particular Broyden method is defined by a sequence $\phi_1, \phi_2, \ldots$, of parameter values. A *pure* Broyden method is one that uses a constant ϕ.

Since both $\mathbf{H}^{\mathrm{DFP}}$ and $\mathbf{H}^{\mathrm{BFGS}}$ satisfy the fundamental relation (10.29) for updates, this relation is also satisfied by all members of the Broyden family. Thus it can be expected that many properties that were found to hold for the DFP method will also hold for any Broyden method, and indeed this is so. The following is a direct extension of the theorem of Sect. 10.3.

Theorem. *If f is quadratic with positive definite Hessian $\mathbf{F}$, then for a Broydenmethod*

$$\begin{aligned} \boldsymbol{p}_i^T \boldsymbol{F} \boldsymbol{p}_j &= 0, \qquad 0 \leqslant i < j \leqslant k \\ \boldsymbol{H}_{k+1} \boldsymbol{F} \boldsymbol{p}_i &= \boldsymbol{p}_i \quad \textit{for } 0 \leqslant i \leqslant k. \end{aligned}$$

Proof. The proof parallels that of Sect. 10.3, since the results depend only on the basic relation (10.29) and the orthogonality (10.19) because of exact line search. ∎

The Broyden family does not necessarily preserve positive definiteness of $\mathbf{H}^{\phi}$ for all values of ϕ. However, we know that the DFP method does preserve positive definiteness. Hence from (10.41) it follows that positive definiteness is preserved for any $\phi \geqslant 0$, since the sum of a positive definite matrix and a positive semidefinite matrix is positive definite. For $\phi < 0$ there is the possibility that $\mathbf{H}^{\phi}$ may become singular, and thus special precautions should be introduced. In practice $\phi \geqslant 0$ is usually imposed to avoid difficulties.

There has been considerable experimentation with Broyden methods to determine superior strategies for selecting the sequence of parameters ϕ_k.

The above theorem shows that the choice is irrelevant in the case of a quadratic objective and accurate line search. More surprisingly, it has been shown that even for the case of *nonquadratic* functions and accurate line searches, the points generated

by all Broyden methods will coincide (provided singularities are avoided and multiple minima are resolved consistently). This means that differences in methods are important only with inaccurate line search.

For general nonquadratic functions of modest dimension, Broyden methods seem to offer a combination of advantages as attractive general procedures. First, they require only that first-order (that is, gradient) information be available. Second, the directions generated can always be guaranteed to be directions of descent by arranging for $\mathbf{H}_k$ to be positive definite throughout the process. Third, since for a quadratic problem the matrices $\mathbf{H}_k$ converge to the inverse Hessian in at most n steps, it might be argued that in the general case $\mathbf{H}_k$ will converge to the inverse Hessian at the solution, and hence convergence will be superlinear. Unfortunately, while the methods are certainly excellent, their convergence characteristics require more careful analysis, and this will lead us to an important additional modification.

Partial Quasi-Newton Methods

There is, of course, the option of restarting a Broyden method every $m + 1$ steps, where $m + 1 < n$. This would yield a *partial quasi-Newton method* that, for small values of m, would have modest storage requirements, since the approximate inverse Hessian could be stored implicitly by storing only the vectors $\mathbf{p}_i$ and $\mathbf{q}_i$, $i \leqslant m+1$. In the quadratic case this method exactly corresponds to the partial conjugate gradient method and hence it has similar convergence properties.

10.5 Convergence Properties

The various schemes for simultaneously generating and using an approximation to the inverse Hessian are difficult to analyze definitively. One must therefore, to some extent, resort to the use of analogy and approximate analyses to determine their effectiveness. Nevertheless, the machinery we developed earlier provides a basis for at least a preliminary analysis.

Global Convergence

In practice, quasi-Newton methods are usually executed in a continuing fashion, starting with an initial approximation and successively improving it throughout the iterative process. Under various and somewhat stringent conditions, it can be proved that this procedure is globally convergent. If, on the other hand, the quasi-Newton methods are restarted every n or $n + 1$ steps by resetting the approximate inverse Hessian to its initial value, then global convergence is guaranteed by the presence of the first descent step of each cycle (which acts as a spacer step).

Local Convergence

The local convergence properties of quasi-Newton methods in the pure form discussed so far are not as good as might first be thought. Let us focus on the local convergence properties of these methods when executed with the restarting feature. Specifically, consider a Broyden method and for simplicity assume that at the beginning of each cycle the approximate inverse Hessian is reset to the identity matrix. Each cycle, if at least n steps in duration, will then contain one complete cycle of an approximation to the conjugate gradient method. Asymptotically, in the tail of the generated sequence, this approximation becomes arbitrarily accurate, and hence we may conclude, as for any method that asymptotically approaches the conjugate gradient method, that the method converges superlinearly (at least if viewed at the end of each cycle). Although superlinear convergence is attractive, the fact that in this case it hinges on repeated cycles of n steps in duration can seriously detract from its practical significance for problems with large n, since we might hope to terminate the procedure before completing even a single full cycle of n steps.

To obtain insight into the defects of the method, let us consider a special situation. Suppose that f is quadratic and that the eigenvalues of the Hessian, $\mathbf{F}$, of f are close together but all very large. If, starting with the identity matrix, an approximation to the inverse Hessian is updated m times, the matrix $\mathbf{H}_m\mathbf{F}$ will have m eigenvalues equal to unity and the rest will still be large. Thus, the ratio of smallest to largest eigenvalue of $\mathbf{H}_m\mathbf{F}$, the condition number, will be worse than for $\mathbf{F}$ itself. Therefore, if the updating were discontinued and $\mathbf{H}_m$ were used as the approximation to $\mathbf{F}^{-1}$ in future iterations according to the procedure of Sect. 10.1, we see that convergence would be poorer than it would be for ordinary steepest descent. In other words, the approximations to $\mathbf{F}^{-1}$ generated by the updating formulas, although accurate over the subspace traveled, do not necessarily improve and, indeed, are likely to worsen the eigenvalue structure of the iteration process.

In practice a poor eigenvalue structure arising in this manner will play a dominating role whenever there are factors that tend to weaken its approximation to the conjugate gradient method. Common factors of this type are round-off errors, inaccurate line searches, and nonquadratic terms in the objective function. Indeed, it has been frequently observed, empirically, that performance of the DFP method is highly sensitive to the accuracy of the line search algorithm—to the point where superior step-wise convergence properties can only be obtained through excessive time expenditure in the line search phase.

Example. To illustrate some of these conclusions we consider the six-dimensional problem defined by

$$f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{Q}\mathbf{x},$$

where

$$\mathbf{Q} = \begin{bmatrix} 40 & 0 & 0 & 0 & 0 & 0 \\ 0 & 38 & 0 & 0 & 0 & 0 \\ 0 & 0 & 36 & 0 & 0 & 0 \\ 0 & 0 & 0 & 34 & 0 & 0 \\ 0 & 0 & 0 & 0 & 32 & 0 \\ 0 & 0 & 0 & 0 & 0 & 30 \end{bmatrix}.$$

This function was minimized iteratively (the solution is obviously $\mathbf{x}^* = 0$) starting at $\mathbf{x}_0$ =(10, 10, 10, 10, 10, 10), with $f(\mathbf{x}_0) = 10{,}500$, by using, alternatively, the method of steepest descent, the DFP method, the DFP method restarted every six steps, and the self-scaling method described in the next section. For this quadratic problem the appropriate step size to take at any stage can be calculated by a simple formula. On different computer runs of a given method, different levels of error were deliberately introduced into the step size in order to observe the effect of line search accuracy. This error took the form of a fixed percentage increase over the optimal value. The results are presented below:

Case 1. No error in step size α

	Function value			
Iteration	Steepest descent	DFP	DFP (with restart)	Self-scaling
1	96.29630	96.29630	96.29630	96.29630
2	1.560669	6.900839×10^{-1}	6.900839×10^{-1}	6.900839×10^{-1}
3	2.932559×10^{-2}	3.988497×10^{-3}	3.988497×10^{-3}	3.988497×10^{-3}
4	5.787315×10^{-4}	1.683310×10^{-5}	1.683310×10^{-5}	1.683310×10^{-5}
5	1.164595×10^{-5}	3.878639×10^{-8}	3.878639×10^{-8}	3.878639×10^{-8}
6	2.359563×10^{-7}			

Case 2. 0.1 % error in step size α

	Function value			
Iteration	Steepest descent	DFP	DFP (with restart)	Self-scaling
1	96.30669	96.30669	96.30669	96.30669
2	1.564971	6.994023×10^{-1}	6.994023×10^{-1}	6.902072×10^{-1}
3	2.939804×10^{-2}	1.225501×10^{-2}	1.225501×10^{-2}	3.989507×10^{-3}
4	5.810123×10^{-4}	7.301088×10^{-3}	7.301088×10^{-3}	1.684263×10^{-5}
5	1.169205×10^{-5}	2.636716×10^{-3}	2.636716×10^{-3}	3.881674×10^{-8}
6	2.372385×10^{-7}	1.031086×10^{-5}	1.031086×10^{-5}	
7		3.633330×10^{-9}	2.399278×10^{-8}	

Case 3. 1 % error in step size α

	Function value			
Iteration	Steepest descent	DFP	DFP (with restart)	Self-scaling
1	97.33665	97.33665	97.33665	97.33665
2	1.586251	1.621908	1.621908	0.7024872
3	2.989875×10^{-2}	8.268893×10^{-1}	8.268893×10^{-1}	4.090350×10^{-3}
4	5.908101×10^{-4}	4.302943×10^{-1}	4.302943×10^{-1}	1.779424×10^{-5}
5	1.194144×10^{-5}	4.449852×10^{-3}	4.449852×10^{-3}	4.195668×10^{-8}
6	2.422985×10^{-7}	5.337835×10^{-5}	5.337835×10^{-5}	
7		3.767830×10^{-5}	4.493397×10^{-7}	
8		3.768097×10^{-9}		

Case 4. 10 % error in step size α

	Function value			
Iteration	Steepest descent	DFP	DFP (with restart)	Self-scaling
1	200.333	200.333	200.333	200.333
2	2.732789	93.65457	93.65457	2.811061
3	3.836899×10^{-2}	56.92999	56.92999	3.562769×10^{-2}
4	6.376461×10^{-4}	1.620688	1.620688	4.200600×10^{-4}
5	1.219515×10^{-5}	5.251115×10^{-1}	5.251115×10^{-1}	4.726918×10^{-6}
6	2.457944×10^{-7}	3.323745×10^{-1}	3.323745×10^{-1}	
7		6.150890×10^{-3}	8.102700×10^{-3}	
8		3.025393×10^{-3}	2.973021×10^{-3}	
9		3.025476×10^{-5}	1.950152×10^{-3}	
10		3.025476×10^{-7}	2.769299×10^{-5}	
11			1.760320×10^{-5}	
12			1.123844×10^{-6}	

We note first that the error introduced is reported as a percentage of the step size itself. In terms of the change in function value, the quantity that is most often monitored to determine when to terminate a line search, the fractional error is the square of that in the step size. Thus, a one percent error in step size is equivalent to a 0.01 % error in the change in function value.

Next we note that the method of steepest descent is not radically affected by an inaccurate line search while the DFP methods are. Thus for this example while DFP is superior to steepest descent in the case of perfect accuracy, it becomes inferior at an error of only 0.1 % in step size.

10.6 Scaling

There is a general viewpoint about what makes up a desirable descent method that underlies much of our earlier discussions and which we now summarize briefly in order to motivate the presentation of scaling. A method that converges to the exact solution after n steps when applied to a quadratic function on E^n has obvious appeal especially if, as is usually the case, it can be inferred that for nonquadratic problems repeated cycles of length n of the method will yield superlinear convergence. For problems having large n, however, a more sophisticated criterion of performance needs to be established, since for such problems one usually hopes to be able to terminate the descent process before completing even a single full cycle of length n. Thus, with these sorts of problems in mind, the finite-step convergence property serves at best only as a sign post indicating that the algorithm *might*, make rapid progress in its early stages. It is essential to insure that in fact it *will* make rapid progress at every stage. Furthermore, the rapid convergence at each step must not be tied to an assumption on conjugate directions, a property easily destroyed by inaccurate line search and nonquadratic objective functions. With this viewpoint it is natural to look for quasi-Newton methods that simultaneously possess favorable eigenvalue structure at each step (in the sense of Sect. 10.1) and reduce to the conjugate gradient method if the objective function happens to be quadratic. Such methods are developed in this section.

Improvement of Eigenvalue Ratio

Referring to the example presented in the last section where the Davidon-Fletcher-Powell method performed poorly, we can trace the difficulty to the simple observation that the eigenvalues of $\mathbf{H}_0\mathbf{Q}$ are all much larger than unity. The DFP algorithm, or any Broyden method, essentially moves these eigenvalues, one at a time, to unity thereby producing an unfavorable eigenvalue ratio in each $\mathbf{H}_k\mathbf{Q}$ for $1 \leqslant k < n$. This phenomenon can be attributed to the fact that the methods are sensitive to simple scale factors. In particular if $\mathbf{H}_0$ were multiplied by a constant, the whole process would be different. In the example of the last section, if $\mathbf{H}_0$ were scaled by, for instance, multiplying it by 1/35, the eigenvalues of $\mathbf{H}_0\mathbf{Q}$ would be spread above and below unity, and in that case one might suspect that the poor performance would not show up.

Motivated by the above considerations, we shall establish conditions under which the eigenvalue ratio of $\mathbf{H}_{k+1}\mathbf{F}$ is at least as favorable as that of $\mathbf{H}_k\mathbf{F}$ in a Broyden method. These conditions will then be used as a basis for introducing appropriate scale factors.

We use (but do not prove) the following matrix theoretic result due to Loewner.

Interlocking Eigenvalues Lemma. *Let the symmetric $n \times n$ matrix A have eigenvalues $\lambda_1 \leqslant \lambda_2 \leqslant \ldots \leqslant \lambda_n$. Let a be any vector in E^n and denote the eigenvalues of the matrix $\mathbf{A} + \mathbf{a}\mathbf{a}^T$ by $\mu_1 \leqslant \mu_2 \ldots \leqslant \mu_n$. Then $\lambda_1 \leqslant \mu_1 \leqslant \lambda_2 \leqslant \mu_2 \ldots \leqslant \lambda_n \leqslant \mu_n$.*

For convenience we introduce the following definitions:

$$\mathbf{R}_k = \mathbf{F}_k^{1/2}\mathbf{H}_k\mathbf{F}_k^{1/2}$$
$$\mathbf{r}_k = \mathbf{F}_k^{1/2}\mathbf{p}_k.$$

Then using $\mathbf{q}_k = \mathbf{F}_k^{1/2}\mathbf{r}_k$, it can be readily verified that (10.41) is equivalent to

$$\mathbf{R}_{k+1}^{\phi} = \mathbf{R}_k - \frac{\mathbf{R}_k\mathbf{r}_k\mathbf{r}_k^T\mathbf{R}_k}{\mathbf{r}_k^T\mathbf{R}_k\mathbf{r}_k} + \frac{\mathbf{r}_k\mathbf{r}_k^T}{\mathbf{r}_k^T\mathbf{r}_k} + \phi\mathbf{z}_k\mathbf{z}_k^T, \tag{10.42}$$

where

$$\mathbf{z}_k = \mathbf{F}^{1/2}\mathbf{V}_k = \sqrt{\mathbf{r}_k^T\mathbf{R}_k\mathbf{r}_k}\left(\frac{\mathbf{r}_k}{\mathbf{r}_k^T\mathbf{r}_k} - \frac{\mathbf{R}_k\mathbf{r}_k}{\mathbf{r}_k^T\mathbf{R}_k\mathbf{r}_k}\right).$$

Since $\mathbf{R}_k$ is similar to $\mathbf{H}_k\mathbf{F}$ (because $\mathbf{H}_k\mathbf{F} = \mathbf{F}^{1/2}\mathbf{R}_k\mathbf{F}^{1/2}$), both have the same eigenvalues. It is most convenient, however, in view of (10.42) to study $\mathbf{R}_k$, obtaining conclusions about $\mathbf{H}_k\mathbf{F}$ indirectly.

Before proving the general theorem we shall consider the case $\phi = 0$ corresponding to the DFP formula. Suppose the eigenvalues of $\mathbf{R}_k$ are $\lambda_1, \lambda_2, \ldots, \lambda_n$ with $0 < \lambda_1 \leqslant \lambda_2 \leqslant \ldots \leqslant \lambda_n$. Suppose also that $1 \in [\lambda_1, \lambda_n]$. We will show that the eigenvalues of $\mathbf{R}_{k+1}$ are all contained in the interval $[\lambda_1, \lambda_n]$, which of course implies that $\mathbf{R}_{k+1}$ is no worse than $\mathbf{R}_k$ in terms of its condition number. Let us first consider the matrix

$$\mathbf{P} = \mathbf{R}_k - \frac{\mathbf{R}_k\mathbf{r}_k\mathbf{r}_k^T\mathbf{R}_k}{\mathbf{r}_k^T\mathbf{R}_k\mathbf{r}_k}.$$

We see that $\mathbf{P}\mathbf{r}_k = 0$ so one eigenvalue of $\mathbf{P}$ is zero. If we denote the eigenvalues of $\mathbf{P}$ by $\mu_1 \leqslant \mu_2 \leqslant \ldots \leqslant \mu_n$, we have from the above observation and the lemma on interlocking eigenvalues that

$$0 = \mu_1 \leqslant \lambda_1 \leqslant \mu_2 \leqslant \ldots \leqslant \mu_n \leqslant \lambda_n.$$

Next we consider

$$\mathbf{R}_{k+1} = \mathbf{R}_k - \frac{\mathbf{R}_k\mathbf{r}_k\mathbf{r}_k^T\mathbf{R}_k}{\mathbf{r}_k^T\mathbf{R}_k\mathbf{r}_k} + \frac{\mathbf{r}_k\mathbf{r}_k^T}{\mathbf{r}_k^T\mathbf{r}_k} = \mathbf{P} + \frac{\mathbf{r}_k\mathbf{r}_k^T}{\mathbf{r}_k^T\mathbf{r}_k}. \tag{10.43}$$

Since $\mathbf{r}_k$ is an eigenvector of $\mathbf{P}$ and since, by symmetry, all other eigenvectors of $\mathbf{P}$ are therefore orthogonal to $\mathbf{r}_k$, it follows that the only eigenvalue different in $\mathbf{R}_{k+1}$ from in $\mathbf{P}$ is the one corresponding to $\mathbf{r}_k$—it now being unity. Thus $\mathbf{R}_{k+1}$ has eigenvalues $\mu_2, \mu_3, \ldots, \mu_n$ and unity. These are all contained in the interval $[\lambda_1, \lambda_n]$. Thus updating does not worsen the eigenvalue ratio. It should be noted that this result in no way depends on α_k being selected to minimize f.

We now extend the above to the Broyden class with $0 \leqslant \phi \leqslant 1$.

Theorem. *Let the n eigenvalues of* $\mathbf{H}_k\boldsymbol{F}$ *be* $\lambda_1,\ \lambda_2,\ \ldots,\ \lambda_n$ *with* $0 < \lambda_1 \leqslant \lambda_2 \leqslant \ldots \leqslant \lambda_n$. *Suppose that* $1 \in [\lambda_1,\ \lambda_n]$. *Then for any* $\phi,\ 0 \leqslant \phi \leqslant 1$, *the eigenvalues of* $\mathbf{H}_{k+1}^{\phi}\boldsymbol{F}$, *where* $\mathbf{H}_{k+1}^{\phi}$ *is defined by (10.41), are all contained in* $[\lambda_1,\ \lambda_n]$.

Proof. The result shown above corresponds to $\phi = 0$. Let us now consider $\phi = 1$, corresponding to the BFGS formula. By our original definition of the BFGS update, $\mathbf{H}^{-1}$ is defined by the formula that is complementary to the DFP formula. Thus

$$\mathbf{H}_{k+1}^{-1} = \mathbf{H}_k^{-1} + \frac{\mathbf{q}_k\mathbf{q}_k^T}{\mathbf{q}_k^T\mathbf{p}_k} - \frac{\mathbf{H}_{k+1}^{-1}\mathbf{p}_k\mathbf{p}_k^T\mathbf{H}_k^{-1}}{\mathbf{p}_k^T\mathbf{H}_k^{-1}\mathbf{p}_k}.$$

This is equivalent to

$$\mathbf{R}_{k+1}^{-1} = \mathbf{R}_k^{-1} - \frac{\mathbf{R}_k^{-1}\mathbf{r}_k\mathbf{r}_k^T\mathbf{R}_k^{-1}}{\mathbf{r}_k^T\mathbf{R}_k^{-1}\mathbf{r}_k} + \frac{\mathbf{r}_k\mathbf{r}_k^T}{\mathbf{r}_k^T\mathbf{r}_k}, \tag{10.44}$$

which is identical to (10.43) except that $\mathbf{R}_k$ is replaced by $\mathbf{R}_k^{-1}$.

The eigenvalues of $\mathbf{R}_k^{-1}$ are $1/\lambda_n \leqslant 1/\lambda_{n-1} \leqslant \ldots \leqslant 1/\lambda_1$. Clearly, $1 \in [1/\lambda_n,\ 1/\lambda_1]$. Thus by the preliminary result, if the eigenvalues of $\mathbf{R}_{k+1}^{-1}$ are denoted $1/\mu_n < 1/\mu_{n-1} < \ldots < 1/\mu_1$, it follows that they are contained in the interval $[1/\lambda_n,\ 1/\lambda_1]$. Thus $1/\lambda_n < 1/\mu_n$ and $1/\lambda_1 > 1/\mu_1$. When inverted this yields $\mu_1 > \lambda_1$ and $\mu_n < \lambda_n$, which shows that the eigenvalues of $\mathbf{R}_{k+1}$ are contained in $[\lambda_1,\ \lambda_n]$. This establishes the result for $\phi = 1$.

For general ϕ the matrix $\mathbf{R}_{k+1}^{\phi}$ defined by (10.42) has eigenvalues that are all monotonically increasing with ϕ (as can be seen from the interlocking eigenvalues lemma). However, from above it is known that these eigenvalues are contained in $[\lambda_1,\ \lambda_n]$ for $\phi = 0$ and $\phi = 1$. Hence, they must be contained in $[\lambda_1,\ \lambda_n]$ for all $\phi,\ 0 \leqslant \phi \leqslant 1$. ∎

Scale Factors

In view of the result derived above, it is clearly advantageous to scale the matrix $\mathbf{H}_k$ so that the eigenvalues of $\mathbf{H}_k\mathbf{F}$ are spread both below and above unity. Of course in the ideal case of a quadratic problem with perfect line search this is strictly only necessary for $\mathbf{H}_0$, since unity is an eigenvalue of $\mathbf{H}_k\mathbf{F}$ for $k > 0$. But because of the inescapable deviations from the ideal, it is useful to consider the possibility of scaling every $\mathbf{H}_k$.

A scale factor can be incorporated directly into the updating formula. We first multiply $\mathbf{H}_k$ by the scale factor γ_k and then apply the usual updating formula. This is equivalent to replacing $\mathbf{H}_k$ by $\gamma_k\mathbf{H}_k$ in (10.42) and leads to

$$\mathbf{H}_{k+1} = \left(\mathbf{H}_k - \frac{\mathbf{H}_k\mathbf{q}_k\mathbf{q}_k^T\mathbf{H}_k}{\mathbf{q}_k^T\mathbf{H}_k\mathbf{q}_k} + \phi_k\mathbf{v}_k\mathbf{v}_k^T\right)\gamma_k + \frac{\mathbf{p}_k\mathbf{p}_k^T}{\mathbf{p}_k^T\mathbf{q}_k}. \tag{10.45}$$

This defines a two-parameter family of updates that reduces to the Broyden family for $\gamma_k = 1$.

Using $\gamma_0,\ \gamma_1,\ \ldots$ as arbitrary positive scale factors, we consider the algorithm: Start with any symmetric positive definite matrix $\mathbf{H}_0$ and any point $\mathbf{x}_0$, then starting with $k = 0$,

Step 1. Set $\mathbf{d}_k = -\mathbf{H}_k\mathbf{g}_k$.
Step 2. Minimize $f(\mathbf{x}_k + \alpha\mathbf{d}_k)$ with respect to $\alpha \geqslant 0$ to obtain $\mathbf{x}_{k+1},\ \mathbf{P}_k = \alpha_k\mathbf{d}_k$, and $\mathbf{g}_{k+1}$.
Step 3. Set $\mathbf{q}_k = \mathbf{g}_{k+1} - \mathbf{g}_k$ and

$$\mathbf{H}_{k+1} = \left(\mathbf{H}_k - \frac{\mathbf{H}_k\mathbf{q}_k\mathbf{q}_k^T\mathbf{H}_k}{\mathbf{q}_k^T\mathbf{H}_k\mathbf{q}_k} + \phi_k\mathbf{v}_k\mathbf{v}_k^T\right)\gamma_k + \frac{\mathbf{p}_k\mathbf{p}_k^T}{\mathbf{p}_k^T\mathbf{q}_k}$$

$$\mathbf{v}_k = (\mathbf{q}_k^T\mathbf{H}\mathbf{q}_k)^{1/2}\left(\frac{\mathbf{p}_k}{\mathbf{p}_k^T\mathbf{q}_k} - \frac{\mathbf{H}_k\mathbf{q}_k}{\mathbf{q}_k^T\mathbf{H}_k\mathbf{q}_k}\right). \tag{10.46}$$

The use of scale factors does destroy the property $\mathbf{H}_n = \mathbf{F}^{-1}$ in the quadratic case, but it does not destroy the conjugate direction property. The following properties of this method can be proved as simple extensions of the results given in Sect. 10.3.

1. If $\mathbf{H}_k$ is positive definite and $\mathbf{p}_k^T\mathbf{q}_k > 0$, (10.46) yields an $\mathbf{H}_{k+1}$ that is positive definite.
2. If f is quadratic with Hessian $\mathbf{F}$, then the vectors $\mathbf{P}_0,\ \mathbf{p}_1,\ \ldots,\ \mathbf{p}_{n-1}$ are mutually $\mathbf{F}$-orthogonal, and, for each k, the vectors $\mathbf{P}_0,\ \mathbf{p}_1,\ \ldots,\ \mathbf{p}_k$ are eigenvectors of $\mathbf{H}_{k+1}\mathbf{F}$.

We can conclude that scale factors do not destroy the underlying conjugate behavior of the algorithm. Hence we can use scaling to ensure good single-step convergence properties.

A Self-scaling Quasi-Newton Algorithm

The question that arises next is how to select appropriate scale factors. If $\lambda_1 \leqslant \lambda_2 \leqslant \ldots \leqslant \lambda_n$ are the eigenvalues of $\mathbf{H}_k\mathbf{F}$, we want to multiply $\mathbf{H}_k$ by γ_k where $\lambda_1 \leqslant 1/\gamma_k \leqslant \lambda_n$. This will ensure that the new eigenvalues contain unity in the interval they span.

Note that in terms of our earlier notation

$$\frac{\mathbf{q}_k^T\mathbf{H}_k\mathbf{q}_k}{\mathbf{p}_k^T\mathbf{q}_k} = \frac{\mathbf{r}_k^T\mathbf{R}_k\mathbf{r}_k}{\mathbf{r}_k^T\mathbf{r}_k}.$$

Recalling that $\mathbf{R}_k$ has the same eigenvalues as $\mathbf{H}_k\mathbf{F}$ and noting that for any $\mathbf{r}_k$

$$\lambda_1 \leqslant \frac{\mathbf{r}_k^T \mathbf{R}_k \mathbf{r}_k}{\mathbf{r}_k^T \mathbf{r}_k} \leqslant \lambda_n,$$

we see that

$$\gamma_k = \frac{\mathbf{p}_k^T \mathbf{q}_k}{\mathbf{q}_k^T \mathbf{H}_k \mathbf{q}_k} \tag{10.47}$$

serves as a suitable scale factor.

We now state a complete self-scaling, restarting, quasi-Newton method based on the ideas above. For simplicity we take $\phi = 0$ and thus obtain a modification of the DFP method. Start at any point $\mathbf{x}_0$, $k = 0$.

Step 1. Set $\mathbf{H}_k = \mathbf{I}$.
Step 2. Set $\mathbf{d}_k = -\mathbf{H}_k\mathbf{g}_k$.
Step 3. Minimize $f(\mathbf{x}_k + \alpha\mathbf{d}_k)$ with respect to $\alpha \geqslant 0$ to obtain α_k, $\mathbf{x}_{k+1}$, $\mathbf{p}_k = \alpha_k\mathbf{d}_k$, $\mathbf{g}_{k+1}$ and $\mathbf{q}_k = \mathbf{g}_{k+1} - \mathbf{g}_k$. (Select α_k accurately enough to ensure $\mathbf{p}_k^T\mathbf{q}_k > 0$.)
Step 4. If k is not an integer multiple of n, set

$$\mathbf{H}_{k+1} = \left(\mathbf{H}_k - \frac{\mathbf{H}_k\mathbf{q}_k\mathbf{q}_k^T\mathbf{H}_k}{\mathbf{q}_k^T\mathbf{H}_k\mathbf{q}_k}\right)\frac{\mathbf{p}_k^T\mathbf{q}_k}{\mathbf{q}_k^T\mathbf{H}_k\mathbf{q}_k} + \frac{\mathbf{p}_k\mathbf{p}_k^T}{\mathbf{p}_k^T\mathbf{q}_k}. \tag{10.48}$$

Add one to k and return to Step 2. If k is an integer multiple of n, return to Step 1.

This algorithm was run, with various amounts of inaccuracy introduced in the line search, on the quadratic problem presented in Sect. 10.4. The results are presented in that section.

10.7 Memoryless Quasi-Newton Methods

The preceding development of quasi-Newton methods can be used as a basis for reconsideration of conjugate gradient methods. The result is an attractive class of new procedures.

Consider a simplification of the BFGS quasi-Newton method where $\mathbf{H}_{k+1}$ is defined by a BFGS update applied to $\mathbf{H} = \mathbf{I}$, rather than to $\mathbf{H}_k$. Thus $\mathbf{H}_{k+1}$ is determined without reference to the previous $\mathbf{H}_k$, and hence the update procedure is *memoryless*. This update procedure leads to the following algorithm: Start at any point $\mathbf{x}_0$, $k = 0$.

Step 1.

$$\text{Set } \mathbf{H}_k = \mathbf{I}. \tag{10.49}$$

Step 2.

$$\text{Set } \mathbf{d}_k = -\mathbf{H}_k\mathbf{g}_k. \tag{10.50}$$

Step 3. Minimize $f(\mathbf{x}_k + \alpha \mathbf{d}_k)$ with respect to $\alpha \geqslant 0$ to obtain α_k, $\mathbf{x}_{k+1}$, $\mathbf{p}_k = \alpha_k \mathbf{d}_k$, $\mathbf{g}_{k+1}$, and $\mathbf{q}_k = \mathbf{g}_{k+1} - \mathbf{g}_k$. (Select α_k accurately enough to ensure $\mathbf{p}_k^T \mathbf{q}_k > 0$.)
Step 4. If k is not an integer multiple of n, set

$$\mathbf{H}_{k+1} = \mathbf{I} - \frac{\mathbf{q}_k \mathbf{p}_k^T + \mathbf{p}_k \mathbf{q}_k^T}{\mathbf{p}_k^T \mathbf{q}_k} + \left(1 + \frac{\mathbf{q}_k^T \mathbf{q}_k}{\mathbf{p}_k^T \mathbf{q}_k}\right) \frac{\mathbf{p}_k \mathbf{p}_k^T}{\mathbf{p}_k^T \mathbf{q}_k}. \tag{10.51}$$

Add 1 to k and return to Step 2. If k is an integer multiple of n, return to Step 1.
Combining (10.50) and (10.51), it is easily seen that

$$\mathbf{d}_{k+1} = -\mathbf{g}_{k+1} + \frac{\mathbf{q}_k \mathbf{p}_k^T \mathbf{g}_{k+1} + \mathbf{p}_k \mathbf{q}_k^T \mathbf{g}_{k+1}}{\mathbf{p}_k^T \mathbf{q}_k} - \left(1 + \frac{\mathbf{q}_k^T \mathbf{q}_k}{\mathbf{p}_k^T \mathbf{q}_k}\right) \frac{\mathbf{p}_k \mathbf{p}_k^T \mathbf{g}_{k-1}}{\mathbf{p}_k^T \mathbf{q}_k}. \tag{10.52}$$

If the line search is exact, then $\mathbf{p}_k^T \mathbf{g}_{k+1} = 0$ and hence $\mathbf{p}_k^T \mathbf{q}_k = -\mathbf{p}_k^T \mathbf{g}_k$. In this case (10.52) is equivalent to

$$\mathbf{d}_{k+1} = -\mathbf{g}_{k+1} + \frac{\mathbf{q}_k^T \mathbf{g}_{k+1}}{\mathbf{p}_k^T \mathbf{q}_k} \mathbf{p}_k = -\mathbf{g}_{k+1} + \beta_k \mathbf{d}_k, \tag{10.53}$$

where

$$\beta_k = \frac{\mathbf{q}_k \mathbf{q}_{k+1}^T}{\mathbf{g}_k^T \mathbf{q}_k}.$$

This coincides exactly with the Polak-Ribiere form of the conjugate gradient method. Thus use of the BFGS update in this way yields an algorithm that is of the modified Newton type with positive definite coefficient matrix and which is equivalent to a standard implementation of the conjugate gradient method when the line search is exact.

The algorithm can be used without exact line search in a form that is similar to that of the conjugate gradient method by using (10.52). This requires storage of only the same vectors that are required of the conjugate gradient method. In light of the theory of quasi-Newton methods, however, the new form can be expected to be superior when inexact line searches are employed, and indeed experiments confirm this.

The above idea can be easily extended to produce a memoryless quasi-Newton method corresponding to any member of the Broyden family. The update formula (10.51) would simply use the general Broyden update (10.41) with $\mathbf{H}_k$ set equal to I. In the case of exact line search (with $\mathbf{p}_k^T \mathbf{g}_{k+1} = 0$), the resulting formula for $\mathbf{d}_{k+1}$ reduces to

$$\mathbf{d}_{k+1} = -\mathbf{g}_{k+1} + (1 - \phi) \frac{\mathbf{q}_k^T \mathbf{g}_{k+1}}{\mathbf{q}_k^T \mathbf{q}_k} \mathbf{q}_k + \phi \frac{\mathbf{q}_k^T \mathbf{g}_{k+1}}{\mathbf{p}_k^T \mathbf{q}_k} \mathbf{p}_k. \tag{10.54}$$

We note that (10.54) is equivalent to the conjugate gradient direction (10.53) only for $\phi = 1$, corresponding to the BFGS update. For this reason the choice $\phi = 1$ is generally preferred for this type of method.

Scaling and Preconditioning

Since the conjugate gradient method implemented as a memoryless quasi-Newton method is a modified Newton method, the fundamental convergence theory based on condition number emphasized throughout this part of the book is applicable, as are the procedures for improving convergence. It is clear that the function scaling procedures discussed in the previous section can be incorporated.

According to the general theory of modified Newton methods, it is the eigenvalues of $\mathbf{H}_k\mathbf{F}(\mathbf{x}_k)$ that influence the convergence properties of these algorithms. From the analysis of the last section, the memoryless BFGS update procedure will, in the pure quadratic case, yield a matrix $\mathbf{H}_k\mathbf{F}$ that has a more favorable eigenvalue ratio than $\mathbf{F}$ itself only if the function f is scaled so that unity is contained in the interval spanned by the eigenvalues of F. Experimental evidence verifies that at least an initial scaling of the function in this way can lead to significant improvement. Scaling can be introduced at every step as well, and complete self-scaling can be effective in some situations.

It is possible to extend the scaling procedure to a more general *preconditioning* procedure. In this procedure the matrix governing convergence is changed from $\mathbf{F}(\mathbf{x}_k)$ to $\mathbf{HF}(\mathbf{x}_k)$ for some $\mathbf{H}$. If $\mathbf{HF}(\mathbf{x}_k)$ has its eigenvalues all close to unity, then the memoryless quasi-Newton method can be expected to perform exceedingly well, since it possesses simultaneously the advantages of being a conjugate gradient method and being a well-conditioned modified Newton method.

Preconditioning can be conveniently expressed in the basic algorithm by simply replacing $\mathbf{H}_k$ in the BFGS update formula by $\mathbf{H}$ instead of I and replacing I by $\mathbf{H}$ in Step 1. Thus (10.51) becomes

$$\mathbf{H}_{k+1} = \mathbf{H} - \frac{\mathbf{H}\mathbf{q}_k\mathbf{p}_k^T + \mathbf{p}_k\mathbf{q}_k^T\mathbf{H}}{\mathbf{q}_k^T\mathbf{q}_k} + \left(1 + \frac{\mathbf{q}_k^T\mathbf{H}\mathbf{q}_k}{\mathbf{p}_k^T\mathbf{q}_k}\right)\frac{\mathbf{p}_k\mathbf{p}_k^T}{\mathbf{p}_k^T\mathbf{p}_k}, \tag{10.55}$$

and the explicit conjugate gradient version (10.52) is also modified accordingly.

Preconditioning can also be used in conjunction with an $(m + 1)$-cycle partial conjugate gradient version of the memoryless quasi-Newton method. This is highly effective if a simple $\mathbf{H}$ can be found (as it sometimes can in problems with structure) so that the eigenvalues of $\mathbf{HF}(\mathbf{x}_k)$ are such that either all but m are equal to unity or they are in m bunches. For large-scale problems, methods of this type seem to be quite promising.

*10.8 *Combination of Steepest Descent and Newton's Method

In this section we digress from the study of quasi-Newton methods, and again expand our collection of basic principles. We present a combination of steepest descent and Newton's method which includes them both as special cases. The resulting

combined method can be used to develop algorithms for problems having special structure, as illustrated in Chap. 13. This method and its analysis comprises a fundamental element of the modern theory of algorithms.

The method itself is quite simple. Suppose there is a subspace N of E^n on which the inverse Hessian of the objective function f is known (we shall make this statement more precise later). Then, in the quadratic case, the minimum of f over any linear variety parallel to N (that is, any translation of N) can be found in a single step. To minimize f over the whole space starting at any point $\mathbf{x}_k$, we could minimize f over the linear variety parallel to N and containing $\mathbf{x}_k$ to obtain $\mathbf{z}_k$; and then take a steepest descent step from there. This procedure is illustrated in Fig. 10.1. Since $\mathbf{z}_k$ is the minimum point of f over a linear variety parallel to N, the gradient at $\mathbf{z}_k$ will be orthogonal to N, and hence the gradient step is orthogonal to N. If f is not quadratic we can, knowing the Hessian of f on N, approximate the minimum point of f over a linear variety parallel to N by one step of Newton's method. To implement this scheme, that we described in a geometric sense, it is necessary to agree on a method for defining the subspace N and to determine what information about the inverse Hessian is required so as to implement a Newton step over N. We now turn to these questions.

Often, the most convenient way to describe a subspace, and the one we follow in this development, is in terms of a set of vectors that generate it. Thus, if $\mathbf{B}$ is an $n \times m$ matrix consisting of m column vectors that generate N, we may write N as the set of all vectors of the form $\mathbf{Bu}$ where $\mathbf{u} \in E^m$. For simplicity we always assume that the columns of $\mathbf{B}$ are linearly independent.

To see what information about the inverse Hessian is required, imagine that we are at a point $\mathbf{x}_k$ and wish to find the approximate minimum point $\mathbf{z}_k$ of f with respect to movement in N. Thus, we seek $\mathbf{u}_k$ so that

$$\mathbf{z}_k = \mathbf{x}_k + \mathbf{B}\mathbf{u}_k$$

approximately minimizes f. By "approximately minimizes" we mean that $\mathbf{z}_k$ should be the Newton approximation to the minimum over this subspace. We write

$$f(\mathbf{z}_k) \cong f(\mathbf{x}_k) + \nabla f(\mathbf{x}_k)\mathbf{B}\mathbf{u}_k + \frac{1}{2}\mathbf{u}_k^T\mathbf{B}^T\mathbf{F}(\mathbf{x}_k)\mathbf{B}\mathbf{u}_k$$

and solve for $\mathbf{u}_k$ to obtain the Newton approximation. We find

$$\begin{aligned}\mathbf{u}_k &= -(\mathbf{B}^T\mathbf{F}(\mathbf{x}_k)\mathbf{B})^{-1}\mathbf{B}^T\nabla f(\mathbf{x}_k)^T\\ \mathbf{z}_k &= \mathbf{x}_k - \mathbf{B}(\mathbf{B}^T\mathbf{F}(\mathbf{x}_k)\mathbf{B})^{-1}\mathbf{B}^T\nabla f(\mathbf{x}_k)^T.\end{aligned}$$

We see by analogy with the formula for Newton's method that the expression $\mathbf{B}(\mathbf{B}^T\mathbf{F}(\mathbf{x}_k)\mathbf{B})^{-1}\mathbf{B}^T$ can be interpreted as the inverse of $\mathbf{F}(\mathbf{x}_k)$ restricted to the subspace N.

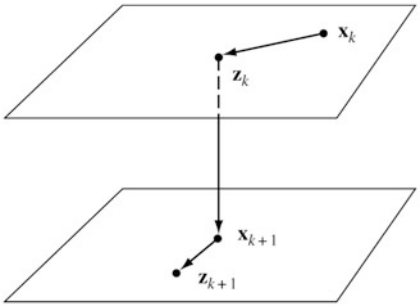

Fig. 10.1 Combined method

Example. Suppose

$$\mathbf{B} = \begin{bmatrix} \mathbf{I} \\ \mathbf{0} \end{bmatrix},$$

where $\mathbf{I}$ is an $m \times m$ identity matrix. This corresponds to the case where N is the subspace generated by the first m unit basis elements of E^n. Let us partition $\mathbf{F} = \nabla^2 f(\mathbf{x}_k)$ as

$$\mathbf{F} = \begin{bmatrix} \mathbf{F}_{11} & \mathbf{F}_{12} \\ \mathbf{F}_{21} & \mathbf{F}_{22} \end{bmatrix},$$

where $\mathbf{F}_{11}$ is $m \times m$. Then, in this case

$$(\mathbf{B}^T\mathbf{F}\mathbf{B})^{-1} = \mathbf{F}_{11}^{-1},$$

and

$$\mathbf{B}(\mathbf{B}^T\mathbf{F}\mathbf{B})^{-1}\mathbf{B}^T = \begin{bmatrix} \mathbf{F}_{11}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix},$$

which shows explicitly that it is the inverse of $\mathbf{F}$ on N that is required. The general case can be regarded as being obtained through partitioning in some skew coordinate system.

Now that the Newton approximation over N has been derived, it is possible to formalize the details of the algorithm suggested by Fig. 10.1. At a given point $\mathbf{x}_k$, the point $\mathbf{x}_{k+1}$ is determined through

$$\begin{aligned}
&\text{a)} \quad \text{Set} \mathbf{d}_k = -\mathbf{B}(\mathbf{B}^T\mathbf{F}(\mathbf{x}_k)\mathbf{B})^{-1}\mathbf{B}^T\nabla\mathbf{f}(\mathbf{x}_k)^T. \\
&\text{b)} \quad \mathbf{z}_k = \mathbf{x}_k + \beta_k\mathbf{d}_k, \text{ where } \beta_k \text{ minimizes } \mathbf{f}(\mathbf{x}_k + \beta\mathbf{d}_k). \\
&\text{c)} \quad \text{Set } \mathbf{p}_k = -\nabla\mathbf{f}(\mathbf{z}_k)^T. \\
&\text{d)} \quad \mathbf{x}_{k+1} = \mathbf{z}_k + \alpha_k\mathbf{p}_k, \text{ where } \alpha_k \text{ minimizes } \mathrm{f}(\mathbf{z}_\mathrm{k} + \alpha\mathbf{p}_\mathrm{k}).
\end{aligned} \tag{10.56}$$

The scalar search parameter β_k is introduced in the Newton part of the algorithm simply to assure that the descent conditions required for global convergence are met. Normally β_k will be approximately equal to unity. (See Sect. 8.5.)

Analysis of Quadratic Case

Since the method is not a full Newton method, we can conclude that it possesses only linear convergence and that the dominating aspects of convergence will be revealed by an analysis of the method as applied to a quadratic function. Furthermore, as might be intuitively anticipated, the associated rate of convergence is governed by the steepest descent part of algorithm (10.56), and that rate is governed by a Kantorovich-like ratio defined over the subspace orthogonal to N.

Theorem (Combined Method). *Let* $\mathbf{Q}$ *be an* $n \times n$ *symmetric positive definite matrix, and let* $\mathbf{x}^* \in E^n$. *Define the function*

$$E(\boldsymbol{x}) = \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}^*)^T \boldsymbol{Q}(\boldsymbol{x} - \boldsymbol{x}^*)$$

and let $\mathbf{b} = \boldsymbol{Q}\boldsymbol{x}^*$. *Let* $\mathbf{B}$ *be an* $n \times m$ *matrix of rank* m. *Starting at an arbitrary point* $\mathbf{x}_0$, *define the iterative process*

a) $\mathbf{u}_k = -(\boldsymbol{B}^T\boldsymbol{Q}\boldsymbol{B})^{-1}\boldsymbol{B}^T\boldsymbol{g}_k$, *where* $\mathbf{g}_k = \boldsymbol{Q}\boldsymbol{x}_k - \boldsymbol{b}$.

b) $\mathbf{z}_k = \boldsymbol{x}_k + \boldsymbol{B}\boldsymbol{u}_k$.

c) $\mathbf{p}_k = \boldsymbol{b} - \boldsymbol{Q}\boldsymbol{z}_k$.

d) $\mathbf{x}_{k+1} = \boldsymbol{z}_k + \alpha_k \boldsymbol{p}_k$, *where* $\alpha_k = \frac{\boldsymbol{p}_k^T \boldsymbol{p}_k}{\boldsymbol{p}_k^T \boldsymbol{Q} \boldsymbol{p}_k}$.

This process converges to $\mathbf{x}^*$, *and satisfies*

$$E(\boldsymbol{x}_{k+1}) \leqslant (1 - \delta)E(\boldsymbol{x}_k) \tag{10.57}$$

where δ, $0 \leqslant 8 \leqslant 1$, *is the minimum of*

$$\frac{(\boldsymbol{p}^T\boldsymbol{p})^2}{(\boldsymbol{p}^T\boldsymbol{Q}\boldsymbol{p})(\boldsymbol{p}^T\boldsymbol{Q}^{-1}\boldsymbol{p})}$$

over all vectors $\mathbf{p}$ *in the nullspace of* $\mathbf{B}^T$.

Proof. The algorithm given in the theorem statement is exactly the general combined algorithm specialized to the quadratic situation. Next we note that

$$\begin{aligned} \mathbf{B}^T\mathbf{p}_k &= \mathbf{B}^T\mathbf{Q}(\mathbf{x}^* - \mathbf{z}_k) = \mathbf{B}^T\mathbf{Q}(\mathbf{x}^* - \mathbf{x}_k) - \mathbf{B}^T\mathbf{Q}\mathbf{B}\mathbf{u}_k \\ &= -\mathbf{B}^T\mathbf{g}_k + \mathbf{B}\mathbf{Q}\mathbf{B}^T(\mathbf{B}^T\mathbf{Q}\mathbf{B})^{-1}\mathbf{B}^T\mathbf{g}_k = \mathbf{0}, \end{aligned} \tag{10.58}$$

which merely proves that the gradient at $\mathbf{z}_k$ is orthogonal to N. Next we calculate

$$\begin{aligned} 2\{E(\mathbf{x}_k) - E(\mathbf{z}_k)\} &= (\mathbf{x}_k - \mathbf{x}^*)^T\mathbf{Q}(\mathbf{x}_k - \mathbf{x}^*) - (\mathbf{z}_k - \mathbf{x}^*)^T\mathbf{Q}(\mathbf{z}_k - \mathbf{x}^*) \\ &= -2\mathbf{u}_k^T\mathbf{B}^T\mathbf{Q}(\mathbf{x}_k - \mathbf{x}^*) - \mathbf{u}_k^T\mathbf{B}^T\mathbf{Q}\mathbf{B}\mathbf{u}_k \\ &= -2\mathbf{u}_k^T\mathbf{B}^T\mathbf{g}_k + \mathbf{u}_k^T\mathbf{B}^T\mathbf{Q}\mathbf{B}(\mathbf{B}^T\mathbf{Q}\mathbf{B})^{-1}\mathbf{B}^T\mathbf{g}_k \\ &= -\mathbf{u}_k^T\mathbf{B}^T\mathbf{g}_k = \mathbf{g}_k^T\mathbf{B}(\mathbf{B}^T\mathbf{Q}\mathbf{B})^{-1}\mathbf{B}^T\mathbf{g}_k. \end{aligned} \tag{10.59}$$

Then we compute

$$\begin{aligned} 2\{E(\mathbf{z}_k) - E(\mathbf{x}_{k+1})\} &= (\mathbf{z}_k - \mathbf{x}^*)^T\mathbf{Q}(\mathbf{z}_k - \mathbf{x}^*) - (\mathbf{x}_{k+1} - \mathbf{x}^*)^T\mathbf{Q}(\mathbf{x}_{k+1} - \mathbf{x}^*) \\ &= -2\alpha_k\mathbf{p}_k^T\mathbf{Q}(\mathbf{z}_k - \mathbf{x}^*) - \alpha_k^2\mathbf{p}_k^T\mathbf{Q}\mathbf{p}_k \\ &= 2\alpha_k\mathbf{p}_k^T\mathbf{p}_k - \alpha_k^2\mathbf{p}_k^T\mathbf{Q}\mathbf{p}_k \\ &= \alpha_k\mathbf{p}_k^T\mathbf{p}_k = \frac{(\mathbf{p}_k^T\mathbf{p}_k)^2}{\mathbf{p}_k^T\mathbf{Q}\mathbf{p}_k}. \end{aligned} \tag{10.60}$$

Now using (10.58) and $\mathbf{p}_k = -\mathbf{g}_k - \mathbf{QBu}_k$ we have

$$\begin{aligned} 2E(\mathbf{x}_k) &= (\mathbf{x}_k - \mathbf{x}^*)^T\mathbf{Q}(\mathbf{x}_k - \mathbf{x}^*) = \mathbf{g}_k^T\mathbf{Q}^{-1}\mathbf{g}_k \\ &= (\mathbf{p}_k^T + \mathbf{u}_k^T\mathbf{B}^T\mathbf{Q})\mathbf{Q}^{-1}(\mathbf{p}_k + \mathbf{QBu}_k) \\ &= \mathbf{p}_k^T\mathbf{Q}^{-1}\mathbf{p}_k + \mathbf{u}_k^T\mathbf{B}^T\mathbf{QBu}_k \\ &= \mathbf{p}_k^T\mathbf{Q}^{-1}\mathbf{p}_k + \mathbf{g}_k^T\mathbf{B}(\mathbf{B}^T\mathbf{QB})^{-1}\mathbf{B}^T\mathbf{g}_k. \end{aligned} \tag{10.61}$$

Adding (10.59) and (10.60) and dividing by (10.61) there results

$$\begin{aligned} \frac{E(\mathbf{x}_k) - E(\mathbf{x}_{k+1})}{E(\mathbf{x}_k)} &= \frac{\mathbf{g}_k^T\mathbf{B}(\mathbf{B}^T\mathbf{QB})^{-1}\mathbf{B}^T\mathbf{g}_k + (\mathbf{p}_k^T\mathbf{p}_k)^2/\mathbf{p}_k^T\mathbf{Q}\mathbf{p}_k}{\mathbf{p}_k^T\mathbf{Q}^{-1}\mathbf{p}_k + \mathbf{g}_k^T\mathbf{B}(\mathbf{B}^T\mathbf{QB})^{-1}\mathbf{B}^T\mathbf{g}_k} \\ &= \frac{q + (\mathbf{p}_k^T\mathbf{p}_k)/(\mathbf{p}_k^T\mathbf{Q}\mathbf{p}_k)}{q + (\mathbf{p}_k^T\mathbf{Q}^{-1}\mathbf{p}_k)/(\mathbf{p}_k^T\mathbf{p}_k)}, \end{aligned}$$

where $q \geqslant 0$. This has the form $(q + a)/(q + b)$ with

$$a = \frac{\mathbf{p}_k^T\mathbf{p}_k}{\mathbf{p}_k^T\mathbf{Q}\mathbf{p}_k}, \quad b = \frac{\mathbf{p}_k^T\mathbf{Q}^{-1}\mathbf{p}_k}{\mathbf{p}_k^T\mathbf{p}_k}.$$

But for any $\mathbf{p}_k$, it follows that $a \leqslant b$. Hence

$$\frac{q + a}{q + b} \geqslant \frac{a}{b},$$

and thus

$$\frac{E(\mathbf{x}_k) - E(\mathbf{x}_{k+1})}{E(\mathbf{x}_k)} \geqslant \frac{(\mathbf{p}_k^T\mathbf{p}_k)^2}{(\mathbf{p}_k^T\mathbf{Q}\mathbf{p}_k)(\mathbf{p}_k^T\mathbf{Q}^{-1}\mathbf{p}_k)}.$$

Finally,

$$E(\mathbf{x}_{k+1}) \leqslant E(\mathbf{x}_k)\left[1 - \frac{(\mathbf{p}_k^T\mathbf{p}_k)^2}{(\mathbf{p}_k^T\mathbf{Q}\mathbf{p}_k)(\mathbf{p}_k^T\mathbf{Q}^{-1}\mathbf{p}_k)}\right] \leqslant (1 - \delta)E(\mathbf{x}_k),$$

since $\mathbf{B}^T\mathbf{p}_k = \mathbf{0}$. ∎

The value δ associated with the above theorem is related to the eigenvalue structure of $\mathbf{Q}$. If $\mathbf{p}$ were allowed to vary over the whole space, then the Kantorovich inequality

$$\frac{(\mathbf{p}^T\mathbf{p})^2}{(\mathbf{p}^T\mathbf{Q}\mathbf{p})(\mathbf{p}^T\mathbf{Q}^{-1}\mathbf{p})} \geqslant \frac{4aA}{(a+A)^2}, \tag{10.62}$$

where a and A are, respectively, the smallest and largest eigenvalues of $\mathbf{Q}$, gives explicitly

$$\delta = \frac{4aA}{(a+A)^2}.$$

When $\mathbf{p}$ is restricted to the nullspace of $\mathbf{B}^T$, the corresponding value of δ is larger. In some special cases it is possible to obtain a fairly explicit estimate of δ. Suppose, for example, that the subspace N were the subspace spanned by m eigenvectors of $\mathbf{Q}$. Then the subspace in which $\mathbf{p}$ is allowed to vary is the space orthogonal to N and is thus, in this case, the space generated by the other $n - m$ eigenvectors of $\mathbf{Q}$. In this case since for $\mathbf{p}$ in $N^\perp$ (the space orthogonal to N), both $\mathbf{Q}\mathbf{p}$ and $\mathbf{Q}^{-1}\mathbf{p}$ are also in $N^\perp$, the ratio δ satisfies

$$\delta = \frac{(\mathbf{p}^T\mathbf{p})^2}{(\mathbf{p}^T\mathbf{Q}\mathbf{p})(\mathbf{p}^T\mathbf{Q}^{-1}\mathbf{p})} \geqslant \frac{4aA}{(a+A)^2},$$

where now a and A are, respectively, the smallest and largest of the $n - m$ eigenvalues of $\mathbf{Q}$ corresponding to $N^\perp$. Thus the convergence ratio (10.57) reduces to the familiar form

$$E(\mathbf{x}_{k+1}) \leqslant \left(\frac{A-a}{A+a}\right)^2 E(\mathbf{x}_k),$$

where a and A are these special eigenvalues. Thus, if $\mathbf{B}$, or equivalently N, is chosen to include the eigenvectors corresponding to the most undesirable eigenvalues of $\mathbf{Q}$, the convergence rate of the combined method will be quite attractive.

Applications

The combination of steepest descent and Newton's method can be applied usefully in a number of important situations. Suppose, for example, we are faced with a problem of the form

$$\text{minimize} \quad f(\mathbf{x}, \mathbf{y}),$$

where $\mathbf{x} \in E^n$, $\mathbf{y} \in E^m$, and where the second partial derivatives with respect to $\mathbf{x}$ are easily computable but those with respect to $\mathbf{y}$ are not. We may then employ Newton steps with respect to $\mathbf{x}$ and steepest descent with respect to $\mathbf{y}$.

Another instance where this idea can be greatly effective is when there are a few vital variables in a problem which, being assigned high costs, tend to dominate the value of the objective function; in other words, the partial second derivatives with respect to these variables are large. The poor conditioning induced by these variables

can to some extent be reduced by proper scaling of variables, but more effectively, by carrying out Newton's method with respect to them and steepest descent with respect to the others.

10.9 Summary

The basic motivation behind quasi-Newton methods is to try to obtain, at least on the average, the rapid convergence associated with Newton's method without explicitly evaluating the Hessian at every step. This can be accomplished by constructing approximations to the inverse Hessian based on information gathered during the descent process, and results in methods which viewed in blocks of n steps (where n is the dimension of the problem) generally possess superlinear convergence.

Good, or even superlinear, convergence measured in terms of large blocks, however, is not always indicative of rapid convergence measured in terms of individual steps. It is important, therefore, to design quasi-Newton methods so that their single step convergence is rapid and relatively insensitive to line search inaccuracies. We discussed two general principles for examining these aspects of descent algorithms. The first of these is the modified Newton method in which the direction of descent is taken as the result of multiplication of the negative gradient by a positive definite matrix $\mathbf{S}$. The single step convergence ratio of this method is determined by the usual steepest descent formula, but with the condition number of $\mathbf{SF}$ rather than just $\mathbf{F}$ used. This result was used to analyze some popular quasi-Newton methods, to develop the self-scaling method having good single step convergence properties, and to reexamine conjugate gradient methods.

The second principle method is the combined method in which Newton's method is executed over a subspace where the Hessian is known and steepest descent is executed elsewhere. This method converges at least as fast as steepest descent, and by incorporating the information gathered as the method progresses, the Newton portion can be executed over larger and larger subspaces.

At this point, it is perhaps valuable to summarize some of the main themes that have been developed throughout the four chapters comprising Part II. These chapters contain several important and popular algorithms that illustrate the range of possibilities available for minimizing a general nonlinear function. From a broad perspective, however, these individual algorithms can be considered simply as specific patterns on the analytical fabric that is woven through the chapters—the fabric that will support new algorithms and future developments.

One unifying element, that has reproved its value several times, is the Global Convergence Theorem. This result helped mold the final form of every algorithm presented in Part II and has effectively resolved the major questions concerning global convergence.

Another unifying element is the speed of convergence of an algorithm, which we have defined in terms of the asymptotic properties of the sequences an algorithm generates. Initially, it might have been argued that such measures, based on

properties of the tail of the sequence, are perhaps not truly indicative of the actual time required to solve a problem—after all, a sequence generated in practice is a truncated version of the potentially infinite sequence, and asymptotic properties may not be representative of the finite version—a more complex measure of the speed of convergence may be required. It is fair to demand that the validity of the asymptotic measures we have proposed be judged in terms of how well they predict the performance of algorithms applied to specific examples. On this basis, as illustrated by the numerical examples presented in these chapters, and on others, the asymptotic rates are extremely reliable predictors of performance—provided that one carefully tempers one's analysis with common sense (by, for example, not concluding that superlinear convergence is necessarily superior to linear convergence when the superlinear convergence is based on repeated cycles of length n). A major conclusion, therefore, of the previous chapters is the essential validity of the asymptotic approach to convergence analysis. This conclusion is a major strand in the analytical fabric of nonlinear programming.

10.10 Exercises

1. Prove (10.4) directly for the modified Newton method by showing that each step of the modified Newton method is simply the ordinary method of steepest descent applied to a scaled version of the original problem.
2. Find the rate of convergence of the version of Newton's method defined by (10.50), (10.51) of Chap. 8. Show that convergence is only linear if δ is larger than the smallest eigenvalue of $\mathbf{F}(\mathbf{x}^*)$.
3. Consider the problem of minimizing a quadratic function

$$f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{Q}\mathbf{x} - \mathbf{x}^T\mathbf{b},$$

where $\mathbf{Q}$ is symmetric and sparse (that is, there are relatively few nonzero entries in Q). The matrix $\mathbf{Q}$ has the form

$$\mathbf{Q} = \mathbf{I} + \mathbf{V},$$

where $\mathbf{I}$ is the identity and $\mathbf{V}$ is a matrix with eigenvalues bounded by $e < 1$ in magnitude.

(a) With the given information, what is the best bound you can give for the rate of convergence of steepest descent applied to this problem?

(b) In general it is difficult to invert $\mathbf{Q}$ but the inverse can be approximated by $\mathbf{I}$–$\mathbf{V}$, which is easy to calculate. (The approximation is very good for small e.) We are thus led to consider the iterative process

$$\mathbf{x}_{k-l} = \mathbf{x}_k - \alpha_k[\mathbf{I} - \mathbf{V}]\mathbf{g}_k,$$

where $\mathbf{g}_k = \mathbf{Q}\mathbf{x}_k - \mathbf{b}$ and α_k is chosen to minimize f in the usual way. With the information given, what is the best bound on the rate of convergence of this method?

(c) Show that for $e < (\sqrt{5} - 1)/2$ the method in part (b) is always superior to steepest descent.

4. This problem shows that the modified Newton's method is globally convergent under very weak assumptions.
   Let $a > 0$ and $b \geqslant a$ be given constants. Consider the collection P of all $n \times n$ symmetric positive definite matrices $\mathbf{P}$ having all eigenvalues greater than or equal to a and all elements bounded in absolute value by b. Define the point-to-set mapping $\mathbf{B} : E^n \to E^{n+n^2}$ by $\mathbf{B}(\mathbf{x}) = \{(\mathbf{x}, \mathbf{P}) : \mathbf{P} \in P\}$. Show that $\mathbf{B}$ is a closed mapping.
   Now given an objective function $f \in C^1$, consider the iterative algorithm

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k \mathbf{P}_k \mathbf{g}_k,$$

   where $\mathbf{g}_k = \mathbf{g}(\mathbf{x}_k)$ is the gradient of f at $\mathbf{x}_k$, $\mathbf{P}_k$ is any matrix from P and α_k is chosen to minimize $f(\mathbf{x}_{k+1})$. This algorithm can be represented by A which can be decomposed as $\mathbf{A} = \mathbf{SCB}$ where $\mathbf{B}$ is defined above, $\mathbf{C}$ is defined by $\mathbf{C}(\mathbf{x}, \mathbf{P}) = (\mathbf{x}, -\mathbf{P}\mathbf{g}(\mathbf{x}))$, and $\mathbf{S}$ is the standard line search mapping. Show that if restricted to a compact set in $\mathbf{E}^n$, the mapping A is closed.
   Assuming that a sequence $\{\mathbf{x}_k\}$ generated by this algorithm is bounded, show that the limit $\mathbf{x}^*$ of any convergent subsequence satisfies $\mathbf{g}(\mathbf{x}^*) = 0$.

5. The following algorithm has been proposed for minimizing unconstrained functions $f(\mathbf{x})$, $\mathbf{x} \in E^n$, without using gradients: Starting with some arbitrary point $\mathbf{x}_0$, obtain a direction of search $\mathbf{d}_k$ such that for each component of $\mathbf{d}_k$

$$f(\mathbf{x}_k = (\mathbf{d}_k)_i \mathbf{e}_i) = \min_{di} f(\mathbf{x}_k + d_i \mathbf{e}_i),$$

   where $\mathbf{e}_j$ denotes the ith column of the identity matrix. In other words, the ith component of $\mathbf{d}_k$ is determined through a line search minimizing $f(\mathbf{x})$ along the i**th** component.
   The next point $\mathbf{x}_{k+1}$ is then determined in the usual way through a line search along $\mathbf{d}_k$; that is,

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{d}_k,$$

   where $\mathbf{d}_k$ minimizes $f(\mathbf{x}_{k+1})$.

(a) Obtain an explicit representation for the algorithm for the quadratic case where

$$f(\mathbf{x}) = \frac{1}{2}(\mathbf{x} - \mathbf{x}^*)^T \mathbf{Q}(\mathbf{x} - \mathbf{x}^*) + f(\mathbf{x}^*).$$

(b) What condition on $f(\mathbf{x})$ or its derivatives will guarantee descent of this algorithm for general $f(\mathbf{x})$?

(c) Derive the convergence rate of this algorithm (assuming a quadratic objective). Express your answer in terms of the condition number of some matrix.

6. Suppose that the rank one correction method of Sect. 10.2 is applied to the quadratic problem (10.2) and suppose that the matrix $\mathbf{R}_0 = \mathbf{F}^{1/2}\mathbf{H}_0\mathbf{F}^{1/2}$ has $m < n$ eigenvalues less than unity and $n - m$ eigenvalues greater than unity. Show that the condition $\mathbf{q}_k^T(\mathbf{p}_k - \mathbf{H}_k\mathbf{q}_k) > 0$ will be satisfied at most m times during the course of the method and hence, if updating is performed only when this condition holds, the sequence $\{\mathbf{H}_k\}$ will not converge to $\mathbf{F}^{-1}$. Infer from this that, in using the rank one correction method, $\mathbf{H}_0$ should be taken very small; but that, despite such a precaution, on nonquadratic problems the method is subject to difficulty.
7. Show that if $\mathbf{H}_0 = \mathbf{I}$ the Davidon-Fletcher-Powell method is the conjugate gradient method. What similar statement can be made when $\mathbf{H}_0$ is an arbitrary symmetric positive definite matrix?
8. In the text it is shown that for the Davidon-Fletcher-Powell method $\mathbf{H}_{k+1}$ is positive definite if $\mathbf{H}_k$ is. The proof assumed that α_k is chosen to exactly minimize $f(\mathbf{x}_k + \alpha\mathbf{d}_k)$. Show that any $\alpha_k > 0$ which leads to $\mathbf{p}_k^T\mathbf{q}_k > 0$ will guarantee the positive definiteness of $\mathbf{H}_{k+1}$. Show that for a quadratic problem any $\alpha_k \neq 0$ leads to a positive definite $\mathbf{H}_{k+1}$.
9. Suppose along the line $\mathbf{x}_k + \alpha\mathbf{d}_k$, $\alpha > 0$, the function $f(\mathbf{x}_k + \alpha\mathbf{d}_k)$ is unimodal and differentiable. Let $\overline{\alpha}_k$ be the minimizing value of α. Show that if any $\alpha_k > \overline{\alpha}_k$ is selected to define $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k\mathbf{d}_k$, then $\mathbf{p}_k^T\mathbf{q}_k > 0$. (Refer to Sect. 10.3.)
10. Let $\{\mathbf{H}_k\}$, $k = 0, 1, 2 \ldots$ be the sequence of matrices generated by the Davidon-Fletcher-Powell method applied, without restarting, to a function f having continuous second partial derivatives. Assuming that there is $a > 0$, $A > 0$ such that for all k we have $\mathbf{H}_k - a\mathbf{I}$ and $A\mathbf{I} - \mathbf{H}_k$ positive definite and the corresponding sequence of $\mathbf{x}_k$'s is bounded, show that the method is globally convergent.
11. Verify Eq. (10.41).
12.

(a) Show that starting with the rank one update formula for $\mathbf{H}$, forming the complementary formula, and then taking the inverse restores the original formula.
(b) What value of ϕ in the Broyden class corresponds to the rank one formula?

13. Explain how the partial Davidon method can be implemented for $m < n/2$, with less storage than required by the full method.
14. Prove statements (10.1) and (10.2) below Eq. (10.46) in Sect. 10.6.
15. Consider using

$$\gamma k = \frac{\mathbf{p}_k^T\mathbf{H}_k^{-1}\mathbf{p}_k}{\mathbf{p}_k^T\mathbf{q}_k}$$

instead of (10.47).

(a) Show that this also serves as a suitable scale factor for a self-scaling quasi-Newton method.
(b) Extend part (a) to

$$\gamma_k = (1 - \phi)\frac{\mathbf{p}_k^T\mathbf{q}_k}{\mathbf{q}_k^T\mathbf{H}_k\mathbf{q}_k} + \phi\frac{\mathbf{p}_k^T\mathbf{H}_k^{-1}\mathbf{p}_k}{\mathbf{p}_k^T\mathbf{q}_k}$$

for $0 \leqslant \phi \leqslant 1$.

16. Prove global convergence of the combination of steepest descent and Newton's method.
17. Formulate a rate of convergence theorem for the application of the combination of steepest and Newton's method to nonquadratic problems.
18. Prove that if $\mathbf{Q}$ is positive definite

$$\frac{(\mathbf{p}^T\mathbf{p})}{\mathbf{p}^T\mathbf{Q}\mathbf{p}} \leqslant \frac{\mathbf{p}^T\mathbf{Q}^{-1}\mathbf{p}}{\mathbf{p}^T\mathbf{p}}$$

for any vector $\mathbf{p}$.

19. It is possible to combine Newton's method and the partial conjugate gradient method. Given a subspace $N \subset E^n$, $\mathbf{x}_{k+1}$ is generated from $\mathbf{x}_k$ by first finding $\mathbf{z}_k$ by taking a Newton step in the linear variety through $\mathbf{x}_k$ parallel to N, and then taking m conjugate gradient steps from $\mathbf{z}_k$. What is a bound on the rate of convergence of this method?
20. In this exercise we explore how the combined method of Sect. 10.7 can be updated as more information becomes available. Begin with $N_0 = \{0\}$. If N_k is represented by the corresponding matrix $\mathbf{B}_k$, define N_{k+1} by the corresponding $\mathbf{B}_{k+1} = [\mathbf{B}_k, \mathbf{p}_k]$, where $\mathbf{p}_k = \mathbf{x}_{k+1} - \mathbf{z}_k$.

(a) If $\mathbf{D}_k = \mathbf{B}_k(\mathbf{B}_k^T\mathbf{F}\mathbf{B}_k)^{-1}\mathbf{B}_k^T$ is known, show that

$$\mathbf{D}_{k+1} = \mathbf{D}_k = \frac{(\mathbf{p}_k - \mathbf{D}_k\mathbf{q}_k)(\mathbf{p}_k - \mathbf{D}_k\mathbf{q}_k)^T}{(\mathbf{p}_k - \mathbf{D}_k\mathbf{q}_k)^T\mathbf{q}_k},$$

where $\mathbf{q}_k = \mathbf{g}_{k+1} - \mathbf{g}_k$. (This is the rank one correction of Sect. 10.2.)

(b) Develop an algorithm that uses (a) in conjunction with the combined method of Sect. 10.8 and discuss its convergence properties.

References

10.1 An early analysis of this method was given by Crockett and Chenoff [C9].

10.2–10.3 The variable metric method was originally developed by Davidon [D12], and its relation to the conjugate gradient method was discovered by Fletcher and Powell [F11]. The rank one method was later developed by Davidon [D13] and Broyden [B24]. For an early general discussion of these methods, see Murtagh and Sargent [M10], and for an excellent recent review, see Dennis and Moré [D15].

10.4 The Broyden family was introduced in Broyden [B24]. The BFGS method was suggested independently by Broyden [B25], Fletcher [F6], Goldfarb [G9], and Shanno [S3]. The beautiful concept of complementarity, which leads easily to the BFGS update and definition of the Broyden class as presented in the text, is due to Fletcher. Another larger class was defined by Huang [H13]. A variational approach to deriving variable

metric methods was introduced by Greenstadt [G15]. Also see Dennis and Schnabel [D16]. Originally there was considerable effort devoted to searching for a best sequence of ϕ_k's in a Broyden method, but Dixon [D17] showed that all methods are identical in the case of exact linear search. There are a number of numerical analysis and implementation issues that arise in connection with quasi-Newton updating methods. From this viewpoint Gill and Murray [G6] have suggested working directly with $\mathbf{B}_k$, an approximation to the Hessian itself, and updating a triangular factorization at each step.

10.5 Under various assumptions on the criterion function, it has been shown that quasi- Newton methods converge globally and superlinearly, provided that accurate exact line search is used. See Powell [P8] and Dennis and Moré [D15]. With inexact line search, restarting is generally required to establish global convergence.

10.6 The lemma on interlocking eigenvalues is due to Loewner [L6]. An analysis of the one-by-one shift of the eigenvalues to unity is contained in Fletcher [F6]. The scaling concept, including the self-scaling algorithm, is due to Oren and Luenberger [O5]. Also see Oren [O4]. The two-parameter class of updates defined by the scaling procedure can be shown to be equivalent to the symmetric Huang class. Oren and Spedicato [O6] developed a procedure for selecting the scaling parameter so as to optimize the condition number of the update.

10.7 The idea of expressing conjugate gradient methods as update formulae is due to Perry [P3]. The development of the form presented here is due to Shanno [S4]. Preconditioning for conjugate gradient methods was suggested by Bertsekas [B9].

10.8 The combined method appears in Luenberger [L10].