6.7

Describe the behavior of the conjugate gradient method for a positive semidefinite quadratic function. Consider the case where there is no optimal solution and the case where there are infinitely many optimal solutions.

6.8

Let $f(x) = \frac{1}{2}x'Qx - b'x$, where Q is positive definite and symmetric. Suppose that x_1 and x_2 minimize f over linear manifolds that are parallel to subspaces S_1 and S_2, respectively. Show that if $x_1 \neq x_2$, then $x_1 - x_2$ is Q-conjugate to all vectors in the intersection of S_1 and S_2. Use this property to construct a conjugate direction method that does not evaluate gradients and uses only line minimizations.

1.7 QUASI-NEWTON METHODS

Quasi-Newton methods are gradient methods of the form

$$x^{k+1} = x^k + \alpha^k d^k, \tag{7.1}$$

$$d^k = -D^k \nabla f(x^k), \tag{7.2}$$

where D^k is a positive definite matrix, which may be adjusted from one iteration to the next so that the direction d^k tends to approximate the Newton direction. Some of these methods are quite popular because they typically converge fast, while avoiding the second derivative calculations associated with Newton's method. Their main drawback relative to the conjugate gradient method is that they require storage of the matrix D^k as well as the matrix-vector multiplication overhead associated with the calculation of the direction d^k (see the subsequent discussion).

An important idea for many quasi-Newton methods is that two successive iterates x^k, x^{k+1} together with the corresponding gradients $\nabla f(x^k)$, $\nabla f(x^{k+1})$, yield curvature information by means of the approximate relation

$$q^k \approx \nabla^2 f(x^{k+1}) p^k, \tag{7.3}$$

where

$$p^k = x^{k+1} - x^k, \tag{7.4}$$

$$q^k = \nabla f(x^{k+1}) - \nabla f(x^k). \tag{7.5}$$

In particular, given n linearly independent iteration increments $p^0, \ldots, p^{n-1}$ together with the corresponding gradient increments $q^0, \ldots, q^{n-1}$, we can obtain approximately the Hessian as

$$\nabla^2 f(x^n) \approx \left[q^0 \ \cdots \ q^{n-1} \right] \left[p^0 \ \cdots \ p^{n-1} \right]^{-1},$$

and the inverse Hessian as

$$\nabla^2 f(x^n)^{-1} \approx \left[p^0 \ \cdots \ p^{n-1} \right] \left[q^0 \ \cdots \ q^{n-1} \right]^{-1}.$$

When the cost is quadratic, this relation is exact. Many interesting quasi-Newton methods use similar but more sophisticated ways to build curvature information into the matrix D^k so that it progressively approaches the inverse Hessian.

In the most popular class of quasi-Newton methods, the matrix D^{k+1} is obtained from D^k, and the vectors p^k and q^k by means of the equation

$$D^{k+1} = D^k + \frac{p^k p^{k\prime}}{p^{k\prime} q^k} - \frac{D^k q^k q^{k\prime} D^k}{q^{k\prime} D^k q^k} + \xi^k \tau^k v^k v^{k\prime}, \tag{7.6}$$

where

$$v^k = \frac{p^k}{p^{k\prime} q^k} - \frac{D^k q^k}{\tau^k}, \tag{7.7}$$

$$\tau^k = q^{k\prime} D^k q^k, \tag{7.8}$$

the scalars ξ^k satisfy, for all k,

$$0 \leq \xi^k \leq 1, \tag{7.9}$$

and D^0 is an arbitrary positive definite matrix. The scalars ξ^k parameterize the method. If $\xi^k = 0$ for all k, we obtain the *Davidon-Fletcher-Powell (DFP) method*, which is historically the first quasi-Newton method. If $\xi^k = 1$ for all k, we obtain the *Broyden-Fletcher-Goldfarb-Shanno (BFGS) method*, for which there is substantial evidence that it is the best general purpose quasi-Newton method currently known.

We first show that under a mild assumption, the matrices D^k generated by Eq. (7.6) are positive definite. This is a very important property, since it guarantees that d^k is a descent direction.

Proposition 1.7.1: If D^k is positive definite and the stepsize α^k is chosen so that x^{k+1} satisfies

$$\nabla f(x^k)' d^k < \nabla f(x^{k+1})' d^k, \tag{7.10}$$

then D^{k+1} as given by Eq. (7.6) is positive definite.

Note: If x^k is not a stationary point, we have $\nabla f(x^k)'d^k < 0$, so in order to satisfy condition (7.10), it is sufficient to carry out the line search to a point where

$$|\nabla f(x^{k+1})'d^k| < |\nabla f(x^k)'d^k|.$$

In particular, if α^k is determined by the line minimization rule, then we have $\nabla f(x^{k+1})'d^k = 0$ and Eq. (7.10) is satisfied.

Proof: We first note that Eq. (7.10) implies that $\alpha^k \neq 0$, $q^k \neq 0$, and

$$p^{k\prime}q^k = \alpha^k d^{k\prime}\big(\nabla f(x^{k+1}) - \nabla f(x^k)\big) > 0. \tag{7.11}$$

Thus all denominator terms in Eqs. (7.6) and (7.7) are nonzero, and D^{k+1} is well defined.

Now for any $z \neq 0$ we have

$$z'D^{k+1}z = z'D^kz + \frac{(z'p^k)^2}{p^{k\prime}q^k} - \frac{(q^{k\prime}D^kz)^2}{q^{k\prime}D^kq^k} + \xi^k\tau^k(v^{k\prime}z)^2. \tag{7.12}$$

Using the notation $a = (D^k)^{1/2}z$, $b = (D^k)^{1/2}q^k$, this equation is written as

$$z'D^{k+1}z = \frac{\|a\|^2\|b\|^2 - (a'b)^2}{\|b\|^2} + \frac{(z'p^k)^2}{p^{k\prime}q^k} + \xi^k\tau^k(v^{k\prime}z)^2. \tag{7.13}$$

From Eqs. (7.8), and (7.11), and the Schwartz inequality [Eq. (A.2) in Appendix A], we have that all terms on the right-hand side of Eq. (7.13) are nonnegative. In order that $z'D^{k+1}z > 0$, it will suffice to show that we cannot have simultaneously

$$\|a\|^2\|b\|^2 = (a'b)^2 \qquad \text{and} \qquad z'p^k = 0.$$

Indeed if $\|a\|^2\|b\|^2 = (a'b)^2$, we must have $a = \lambda b$ or equivalently, $z = \lambda q^k$. Since $z \neq 0$, it follows that $\lambda \neq 0$, so if $z'p^k = 0$, we must have $q^{k\prime}p^k = 0$, which is impossible by Eq. (7.11). **Q.E.D.**

An important property of the algorithm is that when applied to the positive definite quadratic function $f(x) = \frac{1}{2}x'Qx - b'x$, with the stepsize α^k determined by line minimization, it generates a Q-conjugate direction sequence, while simultaneously constructing the inverse Hessian Q^{-1} after n iterations. This is the subject of the next proposition.

Proposition 1.7.2: Let $\{x^k\}$, $\{d^k\}$, and $\{D^k\}$ be sequences generated by the Quasi-Newton algorithm (7.1)-(7.2), (7.6)-(7.9), applied to minimization of the positive definite quadratic function

$$f(x) = \tfrac{1}{2}x'Qx - b'x,$$

with α^k chosen by

$$f(x^k + \alpha^k d^k) = \min_{\alpha} f(x^k + \alpha d^k). \tag{7.14}$$

Assume that none of the vectors $x^0, \ldots, x^{n-1}$ is optimal. Then:

(a) The vectors $d^0, \ldots, d^{n-1}$ are Q-conjugate.

(b) There holds

$$D^n = Q^{-1}.$$

Proof: We will show that for all k

$$d^{i\prime}Qd^j = 0, \qquad 0 \leq i < j \leq k, \tag{7.15}$$

$$D^{k+1}Qp^i = p^i, \qquad 0 \leq i \leq k. \tag{7.16}$$

Equation (7.15) proves part (a) and it can be shown that Eq. (7.16) proves part (b). Indeed, since for $i < n$, none of the vectors x^i is optimal and d^i is a descent direction [cf. Eq. (7.2) and Prop. 1.7.1], we have that $p^i \neq 0$. Since $p^i = \alpha^i d^i$ and $d^0, \ldots, d^{n-1}$ are Q-conjugate, it follows that $p^0, \ldots, p^{n-1}$ are linearly independent and therefore, Eq. (7.16) implies that D^nQ is equal to the identity matrix.

We first verify that

$$D^{k+1}Qp^k = p^k, \qquad \forall\ k. \tag{7.17}$$

From the equation $Qp^k = q^k$ and the updating formula (7.6), we have

$$\begin{aligned} D^{k+1}Qp^k = D^{k+1}q^k &= D^kq^k + \frac{p^kp^{k\prime}q^k}{p^{k\prime}q^k} - \frac{D^kq^kq^{k\prime}D^kq^k}{q^{k\prime}D^kq^k} + \xi^k\tau^kv^kv^{k\prime}q^k \\ &= p^k + \xi^k\tau^kv^kv_k{}'q^k. \end{aligned}$$

From Eqs. (7.7) and (7.8), we have that $v^{k\prime}q^k = 0$ and Eq. (7.17) follows.

We now show Eqs. (7.15) and (7.16) simultaneously by induction. For $k = 0$ there is nothing to show for Eq. (7.15), while Eq. (7.16) holds in view of Eq. (7.17). Assuming that Eqs. (7.15) and (7.16) hold for k, we prove them for $k+1$. We have, for $i < k$,

$$\nabla f(x^{k+1}) = \nabla f(x^{i+1}) + Q(p^{i+1} + \cdots + p^k). \tag{7.18}$$

The vector p^i is orthogonal to each vector in the right-hand side of this equation; it is orthogonal to $Qp^{i+1},\ldots,Qp^k$ because $p^0,\ldots,p^k$ are Q-conjugate (since $p^i=\alpha^i d^i$) and it is orthogonal to $\nabla f(x^{i+1})$ because of the line minimization property of the stepsize [cf. Eq. (7.14)]. Therefore from Eq. (7.18) we obtain

$$p^{i\prime}\nabla f(x^{k+1})=0, \qquad 0\leq i<k.$$

From this equation and Eq. (7.16),

$$p^{i\prime}QD^{k+1}\nabla f(x^{k+1})=0, \qquad 0\leq i\leq k,$$

and since $p^i=\alpha^i d^i$ and $d^{k+1}=-D^{k+1}\nabla f(x^{k+1})$, we obtain

$$d^{i\prime}Qd^{k+1}=0, \qquad 0\leq i\leq k. \tag{7.19}$$

This proves Eq. (7.15) for $k+1$.

From the induction hypothesis (7.16) and Eq. (7.19), we have for all i with $0\leq i\leq k$,

$$q^{k+1\prime}D^{k+1}Qp^i=q^{k+1\prime}p^i=p^{k+1\prime}Qp^i=\alpha^{k+1}\alpha^i d^{k+1\prime}Qd^i=0. \tag{7.20}$$

From Eq. (7.6), we have, for $0\leq i\leq k$,

$$\begin{aligned}D^{k+2}q^i=&D^{k+1}q^i+\frac{p^{k+1}p^{k+1\prime}q^i}{p^{k+1\prime}q^{k+1}}-\frac{D^{k+1}q^{k+1}q^{k+1\prime}D^{k+1}q^i}{q^{k+1\prime}D^{k+1}q^{k+1}}\\&+\xi^{k+1}\tau^{k+1}v^{k+1}v^{k+1\prime}q^i.\end{aligned} \tag{7.21}$$

Since $p^{k+1\prime}q^i=p^{k+1\prime}Qp^i=\alpha^{k+1}\alpha^i d^{k+1\prime}Qd^i=0$, we see that the second term in the right-hand side of Eq. (7.21) is zero. Similarly, Eq. (7.16) implies that $q^{k+1\prime}D^{k+1}q^i=q^{k+1\prime}D^{k+1}Qp^i=q^{k+1\prime}p^i=p^{k+1\prime}Qp^i=0$ and we see that the third term in the right-hand side of Eq. (7.21) is zero. Finally, a similar argument using the definition (7.7) of v^{k+1}, shows that the fourth term in the right-hand side of Eq. (7.21) is zero as well. Therefore, Eqs. (7.21) and (7.16) yield

$$D^{k+2}Qp^i=D^{k+2}q^i=D^{k+1}q^i=D^{k+1}Qp^i=p^i, \qquad 0\leq i\leq k.$$

Taking into account also Eq. (7.17), this proves Eq. (7.16) for $k+1$. **Q.E.D.**

It is also interesting to note that *the sequence $\{x^k\}$ in Prop. 1.7.2 is identical to the one that would be generated by the preconditioned conjugate gradient method with scaling matrix $H=D^0$*; i.e., for $k=0,1,\ldots,n-1$, the vector x^{k+1} minimizes f over the linear manifold

$$M^k=\{z\mid z=x^0+\gamma^0D^0\nabla f(x^0)+\cdots+\gamma^kD^0\nabla f(x^k),\ \gamma^0,\ldots,\gamma^k\in\Re\}.$$

This can be proved for the case where $D^0 = I$ by verifying through induction that for all k there exist scalars β_{ij}^k such that

$$D^k = I + \sum_{i=0}^{k} \sum_{j=0}^{k} \beta_{ij}^k \nabla f(x^i) \nabla f(x^j)'.$$

Therefore, for some scalars b_i^k and all k, we have

$$d^k = -D^k \nabla f(x^k) = \sum_{i=0}^{k} b_i^k \nabla f(x^i).$$

Hence, for all i, x^{i+1} lies on the manifold

$$M^i = \left\{ z \mid z = x^0 + \gamma^0 \nabla f(x^0) + \cdots + \gamma^i \nabla f(x^i),\ \gamma^0, \ldots, \gamma^i \in \Re \right\},$$

and since, by Prop. 1.7.2, the algorithm is a conjugate direction method, x^{i+1} minimizes f over M^i based on the results of the preceding section [cf. Eqs. (5.7) and (5.8)]. Thus, when $D^0 = I$, the algorithm satisfies the defining property of the conjugate gradient method (for all i, x^{i+1} is the unique minimum of f over M^i).

For the case where $D^0 \neq I$, the proof follows by making a transformation of variables so that in the transformed space the initial matrix is the identity. A consequence of this result is that if line minimization is used and the cost is quadratic, the generated iterates do not depend on the values of the scalar ξ^k. It turns out that this is also true even when the cost is nonquadratic ([Dix72a], [Dix72b]), which is a rather surprising result. Thus the choice of ξ^k makes a difference only if the line minimization is inaccurate.

Finally, we note that multiplying the initial matrix D^0 by a positive scaling factor can have a significant beneficial effect on the behavior of the algorithm in the initial iterations of an n-iteration cycle, and also more generally in the case of a nonquadratic problem. A popular choice is to compute

$$\tilde{D}^0 = \frac{p^{0\prime} q^0}{q^{0\prime} D^0 q^0} D^0, \tag{7.22}$$

once the vector x^1 (and hence also p^0 and q^0) has been obtained, and use $\tilde{D}^0$ in place of D^0 in computing D^1. Sometimes it is beneficial to scale D^k even after the first iteration by multiplication with $p^{k\prime} q^k / q^{k\prime} D^k q^k$; see [OrL74], [Ore73], where it is shown that such scaling can improve the convergence rate.

Comparison of Quasi-Newton Methods with Other Methods

Let us now consider a nonquadratic problem, and compare the Quasi-Newton method of Eqs. (7.1)-(7.2), (7.6)-(7.9) with the conjugate gradient method. One advantage of the quasi-Newton method is that when line search is accurate, the algorithm not only tends to generate conjugate directions but also constructs an approximation to the inverse Hessian matrix. As a result, near convergence to a local minimum with positive definite Hessian, it tends to approximate Newton's method thereby attaining a fast convergence rate. It is significant that this property does not depend on the starting matrix D^0, and as a result it is not usually necessary to periodically restart the method with a steepest descent-type step, which is something that is essential for the conjugate gradient method.

A second advantage is that the quasi-Newton method is not as sensitive to accuracy in the line search as the conjugate gradient method. This has been verified by extensive computational experience and can be substantiated to some extent by analysis. A partial explanation is that, under essentially no restriction on the line search accuracy, the method generates positive definite matrices D^k and hence directions of descent (Prop. 1.7.1).

To compare further the conjugate gradient method and the quasi-Newton method, we consider their computational requirements per iteration when n is large. The kth iteration of the conjugate gradient method requires computation of the cost function and its gradient (perhaps several times in the course of the line minimization) together with $O(n)$ operations to compute the conjugate direction d^k and the next point x^{k+1}. The quasi-Newton method requires roughly the same amount of computation for function and gradient evaluations together with $O(n^2)$ operations to compute the matrix D^k and the next point x^{k+1}. If the computation needed for a function and gradient evaluation is larger or comparable to $O(n^2)$ operations, the quasi-Newton method requires only slightly more computation per iteration than the conjugate gradient method and holds the edge in view of its other advantages mentioned earlier. In problems where a function and gradient evaluation requires computation time much less than $O(n^2)$ operations, the conjugate gradient method is typically preferable. As an example, we will see in Section 1.9, that in optimal control problems where typically n is very large, a function and a gradient evaluation typically requires $O(n)$ operations. For this reason the conjugate gradient method is typically preferable for these problems.

In general, both the conjugate gradient method and the quasi-Newton algorithm require less computation per iteration than Newton's method, which requires a function, gradient, and Hessian evaluation, as well as $O(n^3)$ operations at each step for computing the Newton direction. This is counterbalanced by the faster speed of convergence of Newton's method. Furthermore, in some cases, special structure can be exploited to compute the Newton direction efficiently. For example in optimal control problems,

Newton's method typically requires $O(n)$ operations per iteration versus $O(n^2)$ operations for the quasi-Newton method (see Section 1.9).

EXERCISES

7.1 (Rank One Quasi-Newton Methods)

Suppose that D^k is updated according to the formula

$$D^{k+1} = D^k + \frac{(p^k - D^k q^k)y^{k\prime}}{q^{k\prime}y^k},$$

where y^k is any vector such that $q^{k\prime}y^k \neq 0$. Show that we have

$$D^{k+1}q^k = p^k.$$

Conclude that for a positive definite quadratic problem, after n steps for which n linearly independent increments $q^0, \ldots, q^{n-1}$ are obtained, D^n is equal to the inverse Hessian.

7.2 (Limited Memory BFGS Method [Noc80])

A major drawback of Quasi-Newton methods for large problems is the large storage requirement. This motivates methods that construct the Quasi-Newton direction $d^k = -D^k \nabla f(x^k)$ using only a limited number of the vectors p^i and q^i (for example, the last m). This exercise shows one way to do this.

(a) Show that the BFGS updating formula can be written as

$$D^{k+1} = V^{k\prime}D^k V^k + \rho^k p^k p^{k\prime},$$

where

$$\rho^k = \frac{1}{q^k p^{k\prime}}, \qquad V^k = I - \rho^k q^k p^{k\prime}.$$

(b) Show how to calculate the direction $d^k = -D^k \nabla f(x^k)$ using D^0 and the past vectors p^i, q^i, $i = 0, 1, \ldots, k-1$.