where α^k is chosen by the minimization rule or the Armijo rule on the function f. (Such a method makes sense when $\nabla_x F$ is much easier to compute than $\nabla g \nabla_y F$.) Show that if there exists $\gamma \in (0,1)$ such that

$$\big\|\nabla g(x)\nabla_y F\big(x, g(x)\big)\big\| \leq \gamma \big\|\nabla_x F\big(x, g(x)\big)\big\|, \qquad \forall\, x \in \Re^n,$$

then the method is convergent in the sense that all limit points of the sequences that it generates are stationary points of f.

(b) Consider the constrained minimization problem

$$\begin{aligned} &\text{minimize} \quad f(x,y) \\ &\text{subject to} \quad h(x,y) = 0 \end{aligned}$$

where $f : \Re^{n+m} \to \Re$ and $h : \Re^{n+m} \to \Re^m$ are continuously differentiable functions of the two arguments $x \in \Re^n$ and $y \in \Re^m$. Consider also a method of the form

$$x^{k+1} = x^k - \alpha^k \nabla_x f\big(x^k, y^k\big),$$

where y^k is a solution of $h(x^k, y) = 0$, viewed as a system of m equations in the unknown vector y, and α^k is chosen by the minimization rule or the Armijo rule. Formulate conditions that guarantee that this method is convergent.

1.3 GRADIENT METHODS – RATE OF CONVERGENCE

The second major issue regarding gradient methods relates to the rate (or speed) of convergence of the generated sequences $\{x^k\}$. The mere fact that $\{x^k\}$ converges to a stationary point x^* will be of little practical value unless the points x^k are reasonably close to x^* after relatively few iterations. Thus, the study of the rate of convergence provides what are often the dominant criteria for selecting one algorithm in favor of others for solving a particular problem.

Approaches for Rate of Convergence Analysis

There are several approaches towards quantifying the rate of convergence of nonlinear programming algorithms. We will discuss briefly three possibilities and then concentrate on the third.

(a) *Computational complexity approach*: Here we try to estimate the number of elementary operations needed by a given method to find

le or the Armijo rule on the
when $\nabla_x F$ is much easier to
e exists $\gamma \in (0,1)$ such that

$(x))\|, \qquad \forall\, x \in \Re^n,$

ie that all limit points of the
points of f.

oblem

$= 0$

$\Re^m$ are continuously differen-
$\in \Re^n$ and $y \in \Re^m$. Consider

$v^k, y^k)$,

ved as a system of m equations
en by the minimization rule or
at guarantee that this method

CONVERGENCE

lient methods relates to the
ed sequences $\{x^k\}$. The mere
t x^* will be of little practical
ose to x^* after relatively few
nvergence provides what are
algorithm in favor of others

nalysis

quantifying the rate of conver-
We will discuss briefly three
rd.

Here we try to estimate the
ed by a given method to find

an optimal solution exactly or within an ϵ-tolerance. Usually, this approach provides worst-case estimates, that is, upper bounds on the number of required operations over a class of problems of given dimension and type (e.g. linear, convex, etc.). These estimates may also involve parameters such as the distance of the starting point from the optimal solution set, etc.

(b) *Informational complexity approach*: One difficulty with the computational complexity approach is that for a diverse class of problems, it is often difficult or meaningless to quantify the amount of computation needed for a single function or gradient evaluation. For example, in estimating the computational complexity of the gradient method applied to the entire class of differentiable convex functions, how are we to compare the overhead for finding the stepsize and for updating the x vector with the work needed to compute the cost function value and its gradient? The informational complexity approach, which is discussed in detail in [NeY83] and [TrW80], bypasses this difficulty by estimating the number of function (and possibly gradient) evaluations needed to find an exact or approximately optimal solution (as opposed to the number of necessary computational operations). In other respects, the informational and computational complexity approaches are similar.

(c) *Local analysis*: In this approach we focus on the local behavior of the method in a neighborhood of an optimal solution. Local analysis can describe quite accurately the behavior of a method near the solution by using Taylor series approximations, but ignores entirely the behavior of the method when far from the solution.

The main potential advantage of the computational and informational complexity approaches is that they provide information about the method's progress when far from the eventual limit. Unfortunately, however, this information is usually pessimistic as it accounts for the worst possible problem instance within the class considered. This has resulted in some striking discrepancies between the theoretical model predictions and practical real-world observations. For example, the most widely used linear programming method, the simplex method, is categorized as a "bad" method by worst-case complexity analysis, because it performs very poorly on some specially constructed examples, which, however, are highly unlikely in practice. On the other hand, the ellipsoid method of Khachiyan [Kha79] (see [BGT81] for a survey), which was the first linear programming method with a polynomial complexity bound, is categorized as much better than the simplex method by worst-case complexity analysis, even though it performs very poorly on most practical linear programs.

The computational complexity approach has received considerable attention in the context of interior point methods. These methods, discussed

in Sections 2.6, 4.2, and 4.4, were primarily motivated by Karmarkar's development of a linear programming algorithm with a polynomial complexity bound that was more favorable than the one of the ellipsoid method [Kar84]. It turned out, however, that the worst-case predictions for the required number of iterations of these methods were off by many orders of magnitude from the practically observed number of iterations. Furthermore, the interior point methods that perform best in practice have poor worst-case complexity, while the ones with the best complexity bounds are very slow in practice.

The local analysis approach, which will be adopted exclusively in this text, has enjoyed considerable success in predicting the behavior of various methods near nonsingular local minima where the cost function can be well approximated by a quadratic. However, the local analysis approach also has some important drawbacks, the most important of which is that it does not account for the rate of progress in the initial iterations. Nonetheless, in many practical situations this is not a serious omission because progress is fast in the initial iterations and slows down only in the limit (the reasons for this seem hard to understand; they are problem-dependent). Furthermore, often in practice, starting points that are near a solution are easily obtainable by a combination of heuristics and experience, in which case local analysis becomes more meaningful.

Local analysis is not very helpful for problems which either involve singular local minima or which are difficult in the sense that the principal methods take many iterations to get near their solution where local analysis applies. It may be said that at present there is little theory and experience to help a practitioner who is faced with such a problem.

1.3.1 The Local Analysis Approach

We now formalize the basic ingredients of our local rate of convergence analysis approach. These are:

(a) We restrict attention to sequences $\{x^k\}$ that converge to a unique limit point x^*.

(b) Rate of convergence is evaluated in terms of an *error function* $e : \Re^n \mapsto \Re$ satisfying $e(x) \geq 0$ for all $x \in \Re^n$ and $e(x^*) = 0$. Typical choices are the Euclidean distance

$$e(x) = \|x - x^*\|$$

and the cost difference

$$e(x) = |f(x) - f(x^*)|.$$

(c) Our analysis is asymptotic, that is, we look at the rate of convergence of the tail of the error sequence $\{e(x^k)\}$.

(d) The generated error sequence $\{e(x^k)\}$ is compared with some "standard" sequences. In our case, we compare $\{e(x^k)\}$ with the geometric progression

$$\beta^k, \qquad k = 0, 1, \ldots,$$

where $\beta \in (0,1)$ is some scalar. In particular, we say that $\{e(x^k)\}$ *converges linearly or geometrically*, if there exist $q > 0$ and $\beta \in (0,1)$ such that for all k

$$e(x^k) \leq q\beta^k.$$

It is possible to show that linear convergence is obtained if for some $\beta \in (0,1)$ we have

$$\limsup_{k\to\infty} \frac{e(x^{k+1})}{e(x^k)} \leq \beta,$$

that is, asymptotically, the error is dropping by a factor of at least β at each iteration (see Exercise 3.6, which gives several additional convergence rate characterizations). If for every $\beta \in (0,1)$, there exists q such that the condition $e(x^k) \leq q\beta^k$ holds for all k, we say that $\{e(x^k)\}$ converges *superlinearly*. This is true in particular, if

$$\limsup_{k\to\infty} \frac{e(x^{k+1})}{e(x^k)} = 0.$$

To quantify further the notion of superlinear convergence, we may compare $\{e(x^k)\}$ with the sequence

$$(\beta)^{p^k}, \qquad k = 0, 1, \ldots,$$

where $\beta \in (0,1)$, and $p > 1$ are some scalars. This sequence converges much faster than a geometric progression. We say that $\{e(x^k)\}$ *converges at least superlinearly with order* p, if there exist $q > 0$, $\beta \in (0,1)$, and $p > 1$ such that for all k

$$e(x^k) \leq q(\beta)^{p^k}.$$

The case where $p = 2$ is referred to as *quadratic convergence*. It is possible to show that superlinear convergence with order p is obtained if

$$\limsup_{k\to\infty} \frac{e(x^{k+1})}{e(x^k)^p} < \infty,$$

or equivalently, $e(x^{k+1}) = O\big(e(x^k)^p\big)$; see Exercise 3.7.

Most optimization algorithms that are of interest in practice produce sequences converging either linearly or superlinearly, at least when they converge to nonsingular local minima. Linear convergence is a fairly satisfactory rate of convergence for nonlinear programming algorithms, provided the factor β of the associated geometric progression is not too close

to unity. Several nonlinear programming algorithms converge superlinearly for particular classes of problems. Newton's method is an important example, as we will see in the present section and also in Section 1.4. For convergence to singular local minima, slower than linear convergence rate is quite common.

1.3.2 The Role of the Condition Number

Many of the important convergence rate characteristics of gradient methods reveal themselves when the cost function is quadratic. To see why, assume that a gradient method is applied to minimization of a twice continuously differentiable function function $f : \Re^n \mapsto \Re$, and it generates a sequence $\{x^k\}$ converging to a nonsingular local minimum x^*. By Taylor's theorem we have

$$f(x) = f(x^*) + \tfrac{1}{2}(x - x^*)'\nabla^2 f(x^*)(x - x^*) + o\big(\|x - x^*\|^2\big).$$

Therefore, since $\nabla^2 f(x^*)$ is positive definite, f can be accurately approximated near x^* by the quadratic function

$$f(x^*) + \tfrac{1}{2}(x - x^*)'\nabla^2 f(x^*)(x - x^*).$$

We thus expect that asymptotic convergence rate results obtained for the quadratic cost case have direct analogs for the general case. This conjecture can indeed be established by rigorous analysis and has been substantiated by extensive numerical experimentation. For this reason, we take the positive definite quadratic case as our point of departure. We subsequently discuss what happens when $\nabla^2 f(x^*)$ is not positive definite, in which case an analysis based on a quadratic model is inadequate.

Convergence Rate of Steepest Descent for Quadratic Functions

Suppose that the cost function f is quadratic with positive definite Hessian Q. We may assume without loss of generality that f is minimized at $x^* = 0$ and that $f(x^*) = 0$ [otherwise we can use the change of variables $y = x - x^*$ and subtract the constant $f(x^*)$ from $f(x)$]. Thus we have

$$f(x) = \tfrac{1}{2}x'Qx, \qquad \nabla f(x) = Qx, \qquad \nabla^2 f(x) = Q.$$

The steepest descent method takes the form

$$x^{k+1} = x^k - \alpha^k \nabla f(x^k) = (I - \alpha^k Q)x^k.$$

Therefore, we have

$$\|x^{k+1}\|^2 = x^{k\prime}(I - \alpha^k Q)^2 x^k.$$

Since by Prop. A.18(b) of Appendix A, we have for all $x \in \Re^n$

$$x'(I - \alpha^k Q)^2 x \leq \big(\text{maximum eigenvalue of } (I - \alpha^k Q)^2\big)\|x\|^2,$$

we obtain

$$\|x^{k+1}\|^2 \leq \big(\text{maximum eigenvalue of } (I - \alpha^k Q)^2\big)\|x^k\|^2.$$

Using Prop. A.13 of Appendix A, it can be seen that the eigenvalues of $(I - \alpha^k Q)^2$ are equal to $(1 - \alpha^k \lambda_i)^2$, where λ_i are the eigenvalues of Q. Therefore, we have

$$\text{maximum eigenvalue of } (I - \alpha^k Q)^2 = \max\big\{(1 - \alpha^k m)^2, (1 - \alpha^k M)^2\big\},$$

where

$$m: \text{ smallest eigenvalue of } Q,$$

$$M: \text{ largest eigenvalue of } Q.$$

It follows that for $x^k \neq 0$, we have

$$\frac{\|x^{k+1}\|}{\|x^k\|} \leq \max\big\{|1 - \alpha^k m|, |1 - \alpha^k M|\big\}. \tag{3.1}$$

It can be seen that if $|1 - \alpha^k m| \geq |1 - \alpha^k M|$, this inequality holds as an equation if x^k is proportional to an eigenvector corresponding to m. Otherwise, if $|1 - \alpha^k m| < |1 - \alpha^k M|$, the inequality holds as an equation if x^k is proportional to an eigenvector corresponding to M.

Figure 1.3.1 illustrates the convergence rate bound of Eq. (3.1) as a function of the stepsize α^k. It can be seen that the value of α^k that minimizes the bound is

$$\alpha^* = \frac{2}{M + m},$$

in which case

$$\frac{\|x^{k+1}\|}{\|x^k\|} \leq \frac{M - m}{M + m}.$$

This is the best convergence rate bound for steepest descent with constant stepsize.

There is another interesting convergence rate result, which holds when α^k is chosen by the line minimization rule. This result quantifies the rate at which the cost decreases and has the form

$$\frac{f(x^{k+1})}{f(x^k)} \leq \left(\frac{M - m}{M + m}\right)^2. \tag{3.2}$$

The above inequality is verified in Prop. 1.3.1, given in the next subsection, where we collect and prove the more formal results of this section. It can

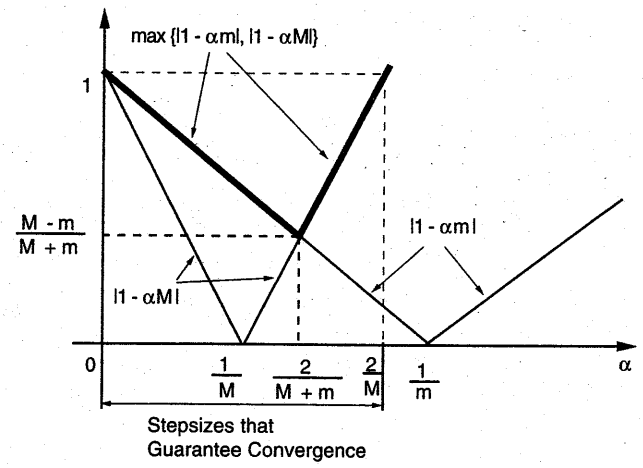


Figure 1.3.1. Illustration of the convergence rate bound $\|x^{k+1}\|/\|x^k\| \leq \max\{|1-\alpha m|, |1-\alpha M|\}$ for steepest descent. The bound is minimized for α such that $1-\alpha m = \alpha M - 1$, that is, for $\alpha = 2/(M+m)$.

be shown that the inequality is sharp in the sense that given any Q, there is a starting point x^0 such that this inequality holds as an equation for all k (see Fig. 1.3.2).

The ratio M/m is called the *condition number* of Q, and problems where M/m is large are referred as *ill-conditioned*. Such problems are characterized by very elongated elliptical level sets. The steepest descent method converges slowly for these problems as indicated by the convergence rate bounds of Eqs. (3.1) and (3.2), and as illustrated in Fig. 1.3.2.

Scaling and Steepest Descent

Consider now the more general method

$$x^{k+1} = x^k - \alpha^k D^k \nabla f(x^k), \tag{3.3}$$

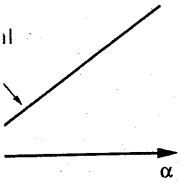
where D^k is positive definite and symmetric; most of the gradient methods of interest have this form as discussed in Section 1.2. It turns out that we may view this iteration as a *scaled version of steepest descent*. In particular, this iteration is just steepest descent applied in a different coordinate system, which depends on D^k.

Indeed, let

$$S = (D^k)^{1/2}$$

denote the positive definite square root of D^k (cf. Prop. A.21 in Appendix A), and consider a transformation of variables defined by

$$x = Sy.$$



d $\|x^{k+1}\|/\|x^k\| \leq \max\{|1-$
minimized for α such that

e that given any Q, there
)lds as an equation for all

mber of Q, and problems
ned. Such problems are
ets. The steepest descent
licated by the convergence
rated in Fig. 1.3.2.

), (3.3)
t of the gradient methods
1.2. It turns out that we
epest descent. In particu-
l in a different coordinate

. Prop. A.21 in Appendix
efined by

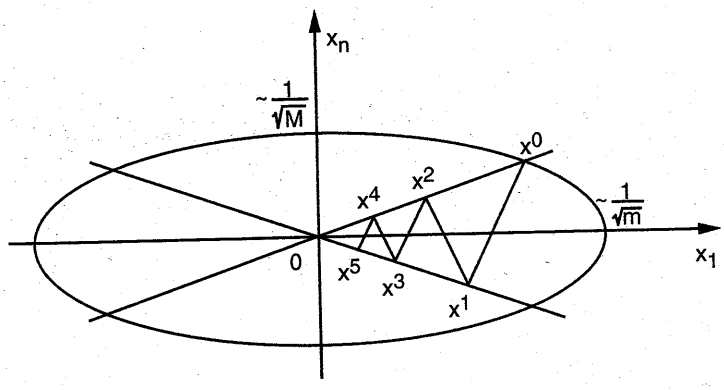


Figure 1.3.2. Example showing that the convergence rate bound

$$\frac{f(x^{k+1})}{f(x^k)} \leq \left(\frac{M-m}{M+m}\right)^2$$

is sharp for the steepest descent method with the line minimization rule. Consider the quadratic function

$$f(x) = \tfrac{1}{2}\sum_{i=1}^{n} \lambda_i x_i^2,$$

where $0 < m = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n = M$. Any positive definite quadratic function can be put into this form by transformation of variables. Consider the starting point

$$x^0 = \left(m^{-1}, 0, \ldots, 0, M^{-1}\right)'$$

and apply the steepest descent method $x^{k+1} = x^k - \alpha^k \nabla f(x^k)$ with α^k chosen by the line minimization rule. We have $\nabla f(x^0) = (1, 0, \ldots, 0, 1)'$ and it can be verified that the minimizing stepsize is $\alpha^0 = 2/(M+m)$. Thus we obtain $x_1^1 = 1/m - 2/(M+m)$, $x_n^1 = 1/M - 2/(M+m)$, $x_i^1 = 0$ for $i = 2, \ldots, n-1$. Therefore,

$$x^1 = \left(\frac{M-m}{M+m}\right)\left(m^{-1}, 0, \ldots, 0, -M^{-1}\right)'$$

and, we can verify by induction that for all k,

$$x^{2k} = \left(\frac{M-m}{M+m}\right)^{2k} x^0, \qquad x^{2k+1} = \left(\frac{M-m}{M+m}\right)^{2k} x^1.$$

Thus, there exist starting points on the plane of points x of the form $x = (\xi_1, 0, \ldots, 0, \xi_n)'$, $\xi_1 \in \Re$, $\xi_n \in \Re$, in fact two lines shown in the figure, for which steepest descent converges in a way that the inequality

$$\frac{f(x^{k+1})}{f(x^k)} \leq \left(\frac{M-m}{M+m}\right)^2$$

is satisfied as an equation at each iteration.

Then, in the space of y, the problem is written as

$$\begin{aligned} &\text{minimize} \quad h(y) \equiv f(Sy) \\ &\text{subject to} \quad y \in \Re^n. \end{aligned} \tag{3.4}$$

The steepest descent method for this problem takes the form

$$y^{k+1} = y^k - \alpha^k \nabla h(y^k). \tag{3.5}$$

Multiplying with S, we obtain

$$Sy^{k+1} = Sy^k - \alpha^k S \nabla h(y^k).$$

By passing back to the space of x, using the relations

$$x^k = Sy^k, \qquad \nabla h(y^k) = S \nabla f(x^k), \qquad S^2 = D^k, \tag{3.6}$$

we obtain

$$x^{k+1} = x^k - \alpha^k D^k \nabla f(x^k).$$

Thus the above gradient iteration is nothing but the steepest descent method (3.5) in the space of y.

We now apply the convergence rate results for steepest descent to the scaled iteration $y^{k+1} = y^k - \alpha^k \nabla h(y^k)$, obtaining

$$\frac{\|y^{k+1}\|}{\|y^k\|} \leq \max\{|1 - \alpha^k m^k|, |1 - \alpha^k M^k|\}$$

and

$$\frac{h(y^{k+1})}{h(y^k)} \leq \left(\frac{M^k - m^k}{M^k + m^k}\right)^2,$$

[cf. the convergence rate bounds (3.1) and (3.2), respectively], where m^k and M^k are the smallest and largest eigenvalues of the Hessian $\nabla^2 h(y)$, which is equal to $S \nabla^2 f(x) S = (D^k)^{1/2} Q (D^k)^{1/2}$. Using the equation $y^k = (D^k)^{-1/2} x^k$ to pass back to the space of x, we obtain the convergence rate bounds

$$\frac{x^{k+1\prime}(D^k)^{-1}x^{k+1}}{x^{k\prime}(D^k)^{-1}x^k} \leq \max\{(1 - \alpha^k m^k)^2, (1 - \alpha^k M^k)^2\} \tag{3.7}$$

and

$$\frac{f(x^{k+1})}{f(x^k)} \leq \left(\frac{M^k - m^k}{M^k + m^k}\right)^2, \tag{3.8}$$

where

m^k : smallest eigenvalue of $(D^k)^{1/2} Q (D^k)^{1/2}$,

M^k : largest eigenvalue of $(D^k)^{1/2}Q(D^k)^{1/2}$.

The stepsize that minimizes the right-hand side bound of Eq. (3.7) is

$$\frac{2}{M^k + m^k}. \tag{3.9}$$

The important point is that if M^k/m^k is much larger than unity, the convergence rate can be very slow, even if an optimal stepsize is used. Furthermore, we see that it is desirable to choose D^k as close as possible to Q^{-1}, so that $(D^k)^{1/2}$ is close to $Q^{-1/2}$ (cf. Prop. A.21 in Appendix A) and $M^k \approx m^k \approx 1$. Note that if D^k is so chosen, Eq. (3.9) shows that the stepsize $\alpha = 1$ is near optimal.

Diagonal Scaling

Many practical problems are ill-conditioned because of poor relative scaling of the optimization variables. By this we mean that the units in which the variables are expressed are incongruent, in the sense that single unit changes of different variables have disproportionate effects on the cost.

As an example, consider a financial problem with two variables, *investment* denoted x_1 and expressed in dollars, and *interest rate* denoted x_2 and expressed in percentage points. If the effect on the cost function f due to a million dollar increment of investment is comparable to the effect due to a percentage point increment of interest rate, then the condition number will be of the order of 10^{12}!! [This rough calculation is based on estimating the condition number by the ratio

$$\frac{\partial^2 f(x_1,x_2)}{(\partial x_2^2)} \Bigg/ \frac{\partial^2 f(x_1,x_2)}{(\partial x_1)^2},$$

approximating the second partial derivatives by the finite difference formulas

$$\frac{\partial^2 f(x_1,x_2)}{(\partial x_1)^2} \approx \frac{f(x_1+h_1,x_2)+f(x_1-h_1,x_2)-2f(x_1,x_2)}{h_1^2},$$

$$\frac{\partial^2 f(x_1,x_2)}{(\partial x_2)^2} \approx \frac{f(x_1,x_2+h_2)+f(x_1,x_2-h_2)-2f(x_1,x_2)}{h_2^2},$$

and using the relations $f(x_1+h_1,x_2) \approx f(x_1,x_2+h_2)$, $f(x_1-h_1,x_2) \approx f(x_1,x_2-h_2)$, and $h_1 = 10^6$, $h_2 = 1$, which express the comparability of the effects of a million dollar investment increment and an interest rate percentage point increment.]

The ill-conditioning in such problems can be significantly alleviated by changing the units in which the optimization variables are expressed, which

amounts to diagonal scaling of the variables. By this, we mean working in a new coordinate system of a vector y related to x by a transformation,

$$x = Sy,$$

where S is a diagonal matrix. In the absence of further information, a reasonable choice of S is one that makes all the diagonal elements of the Hessian of the cost

$$S\nabla^2 f(x)S$$

in the y-coordinate system approximately equal to unity. For this, we must have

$$s_i \approx \left(\frac{\partial^2 f(x)}{(\partial x_i)^2}\right)^{-1/2},$$

where s_i is the ith diagonal element of S. As discussed earlier, we may express any gradient algorithm in the space of variables y as a gradient algorithm in the space of variables x. In particular, steepest descent in the y-coordinate system, when translated in the x-coordinate system, yields the *diagonally scaled steepest descent method*

$$x^{k+1} = x^k - \alpha^k D^k \nabla f(x^k),$$

where

$$D^k = \begin{pmatrix} d_1^k & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & d_2^k & 0 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \\ 0 & 0 & 0 & \cdots & 0 & d_{n-1}^k & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 & d_n^k \end{pmatrix},$$

and

$$d_i^k \approx \left(\frac{\partial^2 f(x^k)}{(\partial x_i)^2}\right)^{-1}.$$

This method is also valid for nonquadratic problems as long as d_i^k are chosen to be positive. It is not guaranteed to improve the convergence rate of steepest descent, but it is simple and often surprisingly effective in practice.

Nonquadratic Cost Functions

It is possible to show that our main conclusions on rate of convergence carry over to the nonquadratic case for sequences converging to nonsingular local minima.

Let f be twice continuously differentiable and consider the gradient method

$$x^{k+1} = x^k - \alpha^k D^k \nabla f(x^k), \qquad (3.10))$$

where D^k is positive definite and symmetric. Consider a generated sequence $\{x^k\}$, and assume that

$$x^k \to x^*, \qquad \nabla f(x^*) = 0, \qquad \nabla^2 f(x^*) : \text{ positive definite}, \tag{3.11}$$

and that $x^k \neq x^*$ for all k. Then, denoting

m^k : smallest eigenvalue of $(D^k)^{1/2}\nabla^2 f(x^k)(D^k)^{1/2}$,

M^k : largest eigenvalue of $(D^k)^{1/2}\nabla^2 f(x^k)(D^k)^{1/2}$,

it is possible to show the following:

(a) There holds

$$\limsup_{k\to\infty} \frac{(x^{k+1}-x^*)'(D^k)^{-1}(x^{k+1}-x^*)}{(x^k-x^*)'(D^k)^{-1}(x^k-x^*)} = \limsup_{k\to\infty} \max\left\{|1-\alpha^k m^k|^2, |1-\alpha^k M^k|^2\right\}.$$

(b) If α^k is chosen by the line minimization rule, there holds

$$\limsup_{k\to\infty} \frac{f(x^{k+1})-f(x^*)}{f(x^k)-f(x^*)} \leq \limsup_{k\to\infty} \left(\frac{M^k-m^k}{M^k+m^k}\right)^2. \tag{3.12}$$

The proof of these facts essentially involves a repetition of the proofs for the quadratic case. However, the details are complicated and tedious and will not be given.

From Eq. (3.12), we see that if D^k converges to some positive definite matrix as $x^k \to x^*$, the sequence $\{f(x^k)\}$ converges to $f(x^*)$ linearly. When

$$D^k \to \nabla^2 f(x^*)^{-1},$$

we have $\lim_{k\to\infty} M^k = \lim_{k\to\infty} m^k = 1$ and Eq. (3.12) shows that the convergence rate of $\{f(x^k)\}$ is superlinear. A somewhat more general version of this result for the case of the Armijo rule is given in the Prop. 1.3.2, which is given in the next subsection. In particular, it is shown that if the direction

$$d^k = -D^k \nabla f(x^k)$$

approaches asymptotically the Newton direction $-\left(\nabla^2 f(x^k)\right)^{-1}\nabla f(x^k)$ and the Armijo rule is used with initial stepsize equal to one, the rate of convergence is superlinear.

There is a consistent theme that emerges from our analysis, namely that to achieve asymptotically fast convergence of the gradient method

$$x^{k+1} = x^k - \alpha^k D^k \nabla f(x^k),$$

one should try to choose the matrices D^k as close as possible to $\left(\nabla^2 f(x^*)\right)^{-1}$ so that the maximum and minimum eigenvalues of $(D^k)^{1/2}\nabla^2 f(x^*)(D^k)^{1/2}$ satisfy $M^k \approx 1$ and $m^k \approx 1$. Furthermore, when D^k is so chosen, the initial stepsize $s = 1$ is a good choice for the Armijo rule and other related rules, or as a starting point for one-dimensional minimization procedures used in minimization stepsize rules. This finding has been supported by extensive numerical experience and is one of the most reliable guidelines for selecting and designing optimization algorithms for unconstrained problems. Note, however, that this guideline is valid only for problems where the cost function is twice differentiable and has positive definite Hessian near the points of interest. We discuss next problems where this condition is not satisfied.

Singular and Difficult Problems

We now consider problems where the Hessian matrix either does not exist or is not positive definite at or near local minima of interest. Expressed mathematically, there are local minima x^* and directions d such that the slope of f along d, which is $\nabla f(x^* + \alpha d)'d$, changes very slowly or very rapidly with α, that is, either

$$\lim_{\alpha \to 0} \frac{\nabla f(x^* + \alpha d)'d - \nabla f(x^*)'d}{\alpha} = 0, \tag{3.13}$$

or

$$\lim_{\alpha \to 0} \frac{\nabla f(x^* + \alpha d)'d - \nabla f(x^*)'d}{\alpha} = \infty. \tag{3.14}$$

The case of Eq. (3.13) is characterized by flatness of the cost along the direction d; large excursions from x^* along d produce small changes in cost. In the case of Eq. (3.14) the reverse is true; the cost rises steeply along d. An example is the function

$$f(x_1, x_2) = |x_1|^4 + |x_2|^{3/2},$$

where for the minimum $x^* = (0, 0)$, Eq. (3.13) holds along the direction $d = (1, 0)$ and Eq. (3.14) holds along the direction $d = (0, 1)$. Gradient methods that use directions that are comparable in size to the gradient may require very large stepsizes in the case of Eq. (3.13) and very small stepsizes in the case of Eq. (3.14). This suggests potential difficulties in the implementation of a good stepsize rule; certainly a constant stepsize does not look like an attractive possibility. Furthermore, in the Armijo rule, the initial stepsize should not be taken constant; it should be adjusted according to a suitable scheme, although designing such a scheme may not be easy.

From the point of view of speed of convergence one may view the cases of Eqs. (3.13) and (3.14) as corresponding to an "infinite condition number," thereby suggesting slower than linear convergence rate for the

ıe as possible to $\left(\nabla^2 f(x^*)\right)^{-1}$
s of $(D^k)^{1/2}\nabla^2 f(x^*)(D^k)^{1/2}$
ın D^k is so chosen, the initial
rule and other related rules,
mization procedures used in
ɔeen supported by extensive
iable guidelines for selecting
ɔnstrained problems. Note,
oblems where the cost func-
ınite Hessian near the points
ıis condition is not satisfied.

ssian matrix either does not
ɔal minima of interest. Ex-
ıa x^* and directions d such
$- \alpha d)'d$, changes very slowly

$$\frac{}{} = 0, \qquad (3.13)$$

$$\frac{}{} = \infty. \qquad (3.14)$$

ıtness of the cost along the
d produce small changes in
true; the cost rises steeply

$|^{3/2}$,

3) holds along the direction
ection $d = (0, 1)$. Gradient
able in size to the gradient
f Eq. (3.13) and very small
s potential difficulties in the
nly a constant stepsize does
ermore, in the Armijo rule,
tant; it should be adjusted
ning such a scheme may not

vergence one may view the
ng to an "infinite condition
ar convergence rate for the

method of steepest descent. Proposition 1.3.3 of the next subsection quantifies the rate of convergence of gradient methods for the case of a convex function whose gradient satisfies the Lipschitz condition

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|, \qquad (3.15)$$

for some L, and all x and y in a neighborhood of x^* [this assumption is consistent with the "flat" cost case of Eq. (3.13), but not with the "steep" cost case of Eq. (3.14)]. It is shown in particular that for a gradient method with the minimization rule, we have

$$f(x^k) - f(x^*) = o(1/k).$$

This type of estimate suggests that for many practical singular problems one may be unable to obtain a highly accurate approximation of an optimal solution. In the "steep" cost case where Eq. (3.14) holds for some directions d, computational examples suggest that the rate of convergence can be slower than linear for the method of steepest descent, although a formal analysis of this conjecture does not seem to have been published.

It should be noted that problems with singular local minima are not the only ones for which gradient methods may converge slowly. There are problems where a given method may have excellent asymptotic rate of convergence, but its progress when far from the eventual limit can be very slow. A prominent example is when the cost function is continuously differentiable but its Hessian matrix is discontinuous and possibly singular in some regions that are outside a small neighborhood of the solution; such functions arise for example in augmented Lagrangian methods for inequality constrained problems (see Section 4.2). Then the powerful Newton-like methods may require a very large number of iterations to get to the small neighborhood of the eventual limit where their convergence rate is favorable. What happens here is that these methods use second derivative information in sophisticated ways, but this information may be misleading due to the Hessian discontinuities.

Generally, there is a tendency to think that difficult problems should be addressed with sophisticated methods, such as Newton-like methods. This is often true, particularly for problems with nonsingular local minima that are poorly conditioned. However, it is important to realize that *often the reverse is true*, namely that for problems with "difficult" cost functions and singular local minima, it is best to use simple methods such as (perhaps diagonally scaled) steepest descent with simple stepsize rules such as a constant or a diminishing stepsize. The reason is that methods that use sophisticated descent directions and stepsize rules often rely on assumptions that are likely to be violated in difficult problems. We also note that for difficult problems, it may be helpful to supplement the steepest descent method with features that allow it to deal better with multiple local minima and peculiarities of the cost function. An often useful modification is the *heavy ball method*, discussed in Exercise 3.9.

1.3.3 Convergence Rate Results

We first derive the convergence rate of steepest descent with the minimization stepsize rule when the cost is quadratic.

Proposition 1.3.1: Consider the quadratic function

$$f(x) = \tfrac{1}{2}x'Qx, \tag{3.16}$$

where Q is positive definite and symmetric, and the method of steepest descent

$$x^{k+1} = x^k - \alpha^k \nabla f(x^k), \tag{3.17}$$

where the stepsize α^k is chosen according to the minimization rule

$$f\big(x^k - \alpha^k \nabla f(x^k)\big) = \min_{\alpha \geq 0} f\big(x^k - \alpha \nabla f(x^k)\big).$$

Then, for all k,

$$f(x^{k+1}) \leq \left(\frac{M-m}{M+m}\right)^2 f(x^k),$$

where M and m are the largest and smallest eigenvalues of Q, respectively.

Proof: Let us denote

$$g^k = \nabla f(x^k) = Qx^k. \tag{3.18}$$

The result clearly holds if $g^k = 0$, so we assume $g^k \neq 0$. We first compute the minimizing stepsize α^k. We have

$$\frac{d}{d\alpha} f(x^k - \alpha g^k) = -g^{k\prime} Q(x^k - \alpha g^k) = -g^{k\prime} g^k + \alpha g^{k\prime} Q g^k.$$

By setting this derivative equal to zero, we obtain

$$\alpha^k = \frac{g^{k\prime} g^k}{g^{k\prime} Q g^k}. \tag{3.19}$$

We have, using Eqs. (3.16)-(3.18),

$$\begin{aligned} f(x^{k+1}) &= \tfrac{1}{2}(x^k - \alpha^k g^k)' Q (x^k - \alpha^k g^k) \\ &= \tfrac{1}{2}\left(x^{k\prime} Q x^k - 2\alpha^k g^{k\prime} Q x^k + (\alpha^k)^2 g^{k\prime} Q g^k\right) \\ &= \tfrac{1}{2}\left(x^{k\prime} Q x^k - 2\alpha^k g^{k\prime} g^k + (\alpha^k)^2 g^{k\prime} Q g^k\right) \end{aligned}$$

and using Eq. (3.19),

$$f(x^{k+1}) = \tfrac{1}{2}\left(x^{k\prime}Qx^k - \frac{(g^{k\prime}g^k)^2}{g^{k\prime}Qg^k}\right).$$

Thus, using the fact $f(x^k) = \frac{1}{2}x^{k\prime}Qx^k = \frac{1}{2}g^{k\prime}Q^{-1}g^k$, we obtain

$$f(x^{k+1}) = \left(1 - \frac{(g^{k\prime}g^k)^2}{(g^{k\prime}Qg^k)(g^{k\prime}Q^{-1}g^k)}\right) f(x^k). \tag{3.20}$$

At this point we need the following lemma.

Lemma 3.1: (Kantorovich Inequality) Let Q be a positive definite and symmetric $n \times n$ matrix. Then for any vector $y \in \Re^n, y \neq 0$, there holds

$$\frac{(y'y)^2}{(y'Qy)(y'Q^{-1}y)} \geq \frac{4Mm}{(M+m)^2},$$

where M and m are the largest and smallest eigenvalues of Q, respectively.

Proof: Let $\lambda_1, \ldots, \lambda_n$ denote the eigenvalues of Q and assume that

$$0 < m = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n = M.$$

Let S be the matrix consisting of the n orthogonal eigenvectors of Q, normalized so that they have unit norm (cf. Prop. A.27 in Appendix A). Then, it can be seen that $S'QS$ is diagonal with diagonal elements $\lambda_1, \ldots, \lambda_n$. By using if necessary a transformation of the coordinate system that replaces y by Sx, we may assume that Q is diagonal and that its diagonal elements are $\lambda_1, \ldots, \lambda_n$. We have for $y = (y_1, \ldots, y_n)' \neq 0$

$$\frac{(y'y)^2}{(y'Qy)(y'Q^{-1}y)} = \frac{\left(\sum_{i=1}^n y_i^2\right)^2}{\left(\sum_{i=1}^n \lambda_i y_i^2\right)\left(\sum_{i=1}^n \frac{y_i^2}{\lambda_i}\right)}.$$

By letting

$$\xi_j = \frac{y_j^2}{\sum_{i=1}^n y_i^2}$$

and by defining

$$\phi(\xi) = \frac{1}{\sum_{i=1}^n \xi_i \lambda_i}, \qquad \psi(\xi) = \sum_{i=1}^n \frac{\xi_i}{\lambda_i},$$

we obtain

$$\frac{(y'y)^2}{(y'Qy)(y'Q^{-1}y)} = \frac{\phi(\xi)}{\psi(\xi)}.$$

Figure 1.3.3 shows that we have

$$\frac{\phi(\xi)}{\psi(\xi)} \geq \frac{4\lambda_1\lambda_n}{(\lambda_1+\lambda_n)^2},$$

which proves the desired inequality. **Q.E.D.**

Returning to the proof of Prop. 1.3.1, we have by using the Kantorovich inequality in Eq. (3.20)

$$f(x^{k+1}) \leq \left(1 - \frac{4Mm}{(M+m)^2}\right) f(x^k) = \left(\frac{M-m}{M+m}\right)^2 f(x^k).$$

Q.E.D.

The following proposition shows superlinear convergence for methods where d^k approaches the Newton direction $-\left(\nabla^2 f(x^*)\right)^{-1}\nabla f(x^k)$ and the Armijo rule is used.

Proposition 1.3.2: (Superlinear Convergence of Newton-Like Methods) Let f be twice continuously differentiable. Consider a sequence $\{x^k\}$ generated by the gradient method $x^{k+1} = x^k + \alpha^k d^k$ and suppose that

$$x^k \to x^*, \qquad \nabla f(x^*) = 0, \qquad \nabla^2 f(x^*): \text{ positive definite.} \tag{3.21}$$

Assume further that $\nabla f(x^k) \neq 0$ for all k and

$$\lim_{k\to\infty} \frac{\|d^k + \left(\nabla^2 f(x^*)\right)^{-1}\nabla f(x^k)\|}{\|\nabla f(x^k)\|} = 0. \tag{3.22}$$

Then, if α^k is chosen by means of the Armijo rule with initial stepsize $s = 1$ and $\sigma < 1/2$, we have

$$\lim_{k\to\infty} \frac{\|x^{k+1} - x^*\|}{\|x^k - x^*\|} = 0. \tag{3.23}$$

Furthermore, there exists an integer $\overline{k} \geq 0$ such that $\alpha^k = 1$ for all $k \geq \overline{k}$ (i.e., eventually no reduction of the initial stepsize will be taking place).

e have by using the Kan-

$$\left(\frac{M-m}{M+m}\right)^2 f(x^k).$$

ır convergence for methods
$\nabla^2 f(x^*))^{-1}\nabla f(x^k)$ and the

gence of Newton-Like
ferentiable. Consider a
thod $x^{k+1} = x^k + \alpha^k d^k$

positive definite. (3.21)

d

$\nabla^k)\|$
$= 0.$ (3.22)

rule with initial stepsize

). (3.23)

such that $\alpha^k = 1$ for all
ial stepsize will be taking

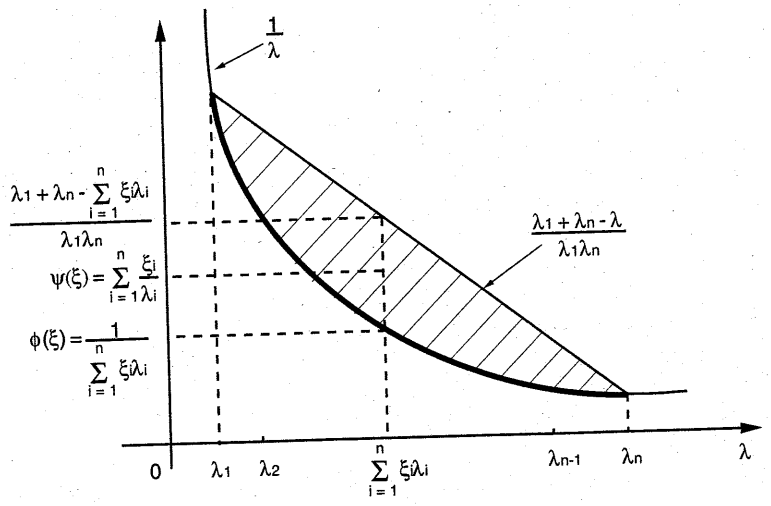


Figure 1.3.3. Proof of the Kantorovich inequality. Consider the function $1/\lambda$. The scalar $\sum_{i=1}^n \xi_i\lambda_i$ represents, for any $\xi = (\xi_1, \ldots, \xi_n)$ with $\xi_i \geq 0$, $\sum_{i=1}^n \xi_i = 1$, a point in the line segment $[\lambda_1, \lambda_n]$. Thus, the values $\phi(\xi) = 1/\sum_{i=1}^n \xi_i\lambda_i$ correspond to the thick part of the curve $1/\lambda$. On the other hand, the value $\psi(\xi) = \sum_{i=1}^n (\xi_i/\lambda_i)$ is a convex combination of $1/\lambda_1, \ldots, 1/\lambda_n$ and hence corresponds to a point in the shaded area in the figure. For the same vector ξ, both $\phi(\xi)$ and $\psi(\xi)$ are represented by points on the same vertical line. Hence,

$$\frac{\phi(\xi)}{\psi(\xi)} \geq \min_{\lambda_1 \leq \lambda \leq \lambda_n} \frac{\frac{1}{\lambda}}{\frac{\lambda_1+\lambda_n-\lambda}{\lambda_1\lambda_n}}.$$

The minimum is attained for $\lambda = (\lambda_1 + \lambda_n)/2$ and we obtain

$$\frac{\phi(\xi)}{\psi(\xi)} \geq \frac{4\lambda_1\lambda_n}{(\lambda_1+\lambda_n)^2},$$

which is used to show the result.

Proof: We first prove that there exists a $\overline{k} \geq 0$ such that for all $k \geq \overline{k}$,

$$f(x^k + d^k) - f(x^k) \leq \sigma \nabla f(x^k)' d^k,$$

that is, the unity initial stepsize passes the test of the Armijo rule. By the mean value theorem, we have

$$f(x^k + d^k) - f(x^k) = \nabla f(x^k)' d^k + \tfrac{1}{2} d^{k\prime} \nabla^2 f(\overline{x}^k) d^k,$$

where $\overline{x}^k$ is a point on the line segment joining x^k and $x^k + d^k$. Thus, it will be sufficient to show that for k sufficiently large, we have

$$\nabla f(x^k)' d^k + \tfrac{1}{2} d^k \nabla^2 f(\overline{x}^k) d^k \leq \sigma \nabla f(x^k)' d^k.$$

By defining

$$p^k = \frac{\nabla f(x^k)}{\|\nabla f(x^k)\|}, \qquad q^k = \frac{d^k}{\|\nabla f(x^k)\|},$$

this condition is written

$$(1-\sigma)p^{k\prime}q^k + \tfrac{1}{2}q^{k\prime}\nabla^2 f(\overline{x}^k)q^k \leq 0. \tag{3.24}$$

From Eq. (3.22), we have $q^k - (\nabla^2 f(x^*))^{-1}p^k \to 0$. Since $\nabla^2 f(x^*)$ is positive definite and $\|p^k\| = 1$, it follows that $\{q^k\}$ is a bounded sequence, and in view of $q^k = d^k/\|\nabla f(x^k)\|$ and $\nabla f(x^k) \to 0$, we obtain $d^k \to 0$. Hence, $x^k + d^k \to x^*$, and it follows that $\overline{x}^k \to x^*$ and $\nabla^2 f(\overline{x}^k) \to \nabla^2 f(x^*)$. We now write Eq. (3.22) as

$$q^k = -\big(\nabla^2 f(x^*)\big)^{-1}p^k + \beta^k,$$

where $\{\beta^k\}$ denotes a vector sequence with $\beta^k \to 0$. By using the above relation and the fact $\nabla^2 f(\overline{x}^k) \to \nabla^2 f(x^*)$, we may write Eq. (3.24) as

$$(1-\sigma)p^{k\prime}\big(\nabla^2 f(x^*)\big)^{-1}p^k - \tfrac{1}{2}p^{k\prime}\big(\nabla^2 f(x^*)\big)^{-1}p^k \geq \gamma^k,$$

where $\{\gamma^k\}$ is some scalar sequence with $\gamma^k \to 0$. Thus Eq. (3.24) is equivalent to

$$\left(\tfrac{1}{2} - \sigma\right)p^{k\prime}\big(\nabla^2 f(x^*)\big)^{-1}p^k \geq \gamma^k.$$

Since $1/2 > \sigma$, $\|p^k\| = 1$, and $\nabla^2 f(x^*)$ is positive definite, the above relation holds for sufficiently large k. Thus, the unity initial stepsize is acceptable for sufficiently large k, as desired.

To complete the proof, we note that from Eq. (3.22), we have

$$d^k + \big(\nabla^2 f(x^*)\big)^{-1}\nabla f(x^k) = \|\nabla f(x^k)\|\delta^k, \tag{3.25}$$

where δ^k is some vector sequence with $\delta^k \to 0$. From Taylor's theorem we obtain

$$\nabla f(x^k) = \nabla^2 f(x^*)(x^k - x^*) + o(\|x^k - x^*\|),$$

from which

$$\big(\nabla^2 f(x^*)\big)^{-1}\nabla f(x^k) = x^k - x^* + o(\|x^k - x^*\|),$$

$$\|\nabla f(x^k)\| = O(\|x^k - x^*\|).$$

Using the above two relations in Eq. (3.25), we obtain

$$d^k + x^k - x^* = o(\|x^k - x^*\|). \tag{3.26}$$

Since for sufficiently large k we have $d^k + x^k = x^{k+1}$, Eq. (3.26) yields

$$x^{k+1} - x^* = o(\|x^k - x^*\|),$$

$$\frac{d^k}{f(x^k)\|},$$

$$l^k \leq 0. \qquad (3.24)$$

$\to 0$. Since $\nabla^2 f(x^*)$ is
$^k\}$ is a bounded sequence,
$\to 0$, we obtain $d^k \to 0$.
and $\nabla^2 f(\overline{x}^k) \to \nabla^2 f(x^*)$.

β^k,

$\to 0$. By using the above
ıay write Eq. (3.24) as

$$(x^*))^{-1} p^k \geq \gamma^k,$$

$\to 0$. Thus Eq. (3.24) is

$\geq \gamma^k$.

ositive definite, the above
he unity initial stepsize is

Eq. (3.22), we have

$$^7 f(x^k)\|\delta^k, \qquad (3.25)$$

From Taylor's theorem we

$(\|x^k - x^*\|)$,

$o(\|x^k - x^*\|)$,

$^*\|)$.

obtain

$$x^*\|). \qquad (3.26)$$

x^{k+1}, Eq. (3.26) yields

$^*\|)$,

from which

$$\lim_{k\to\infty} \frac{\|x^{k+1} - x^*\|}{\|x^k - x^*\|} = \lim_{k\to\infty} \frac{o(\|x^k - x^*\|)}{\|x^k - x^*\|} = 0.$$

Q.E.D.

Note that the equation $\lim_{k\to\infty}\big(\|x^{k+1} - x^*\|/\|x^k - x^*\|\big) = 0$ implies that $\{\|x^k - x^*\|\}$ converges superlinearly (see Exercise 3.6). In particular, we see that Newton's method, combined with the Armijo rule with unity initial stepsize, converges superlinearly when it converges to a local minimum x^* such that $\nabla^2 f(x^*)$ is positive definite. The capture theorem (Prop. 1.2.5) together with the preceding proposition suggest that Newton-like methods with the Armijo rule and a unity initial stepsize converge to a local minimum x^* such that $\nabla^2 f(x^*)$ is positive definite, whenever they are started sufficiently close to such a local minimum. The proof of this is left as Exercise 3.2 for the reader.

We finally consider the convergence rate of gradient methods for singular problems whose cost is sufficiently flat for a Lipschitz condition on the gradient to hold.

Proposition 1.3.3: (Convergence Rate of Gradient Methods for Singular Problems) Suppose that the cost function f is convex and its gradient satisfies for some L the Lipschitz condition

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|, \qquad \forall\, x, y \in \Re^n. \qquad (3.27)$$

Consider a gradient method $x^{k+1} = x^k + \alpha^k d^k$ where α^k is chosen by the minimization rule, and the angle between d^k and $\nabla f(x^k)$ is bounded away from 90 degrees, that is, for some $c > 0$ and all k we have

$$\nabla f(x^k)' d^k \leq -c\|\nabla f(x^k)\|\,\|d^k\|. \qquad (3.28)$$

Suppose that the set of global minima X^* of f is nonempty and bounded. Then

$$f(x^k) - f^* = o(1/k),$$

where $f^* = \min_x f(x)$ is the optimal value.

Proof: We assume that $\nabla f(x^k) \neq 0$ and therefore also $d^k \neq 0$ for all k; otherwise the method terminates finitely at a global minimum and the result holds trivially. Let

$$\tilde{\alpha}^k = \frac{|\nabla f(x^k)' d^k|}{L\|d^k\|^2}, \qquad (3.29)$$

$$\tilde{x}^k = x^k + \tilde{\alpha}^k d^k. \tag{3.30}$$

By using the descent lemma (Prop. A.24 in Appendix A), and Eqs. (3.28) and (3.29), we have

$$\begin{aligned} f(\tilde{x}^k) - f(x^k) &\leq -\tilde{\alpha}^k |\nabla f(x^k)' d^k| + \tfrac{1}{2}(\tilde{\alpha}^k)^2 L \|d^k\|^2 \\ &= \tilde{\alpha}^k \left(-|\nabla f(x^k)' d^k| + \tfrac{1}{2}|\nabla f(x^k)' d^k|\right) \\ &= -\frac{\tilde{\alpha}^k}{2}|\nabla f(x^k)' d^k| \\ &= -\frac{|\nabla f(x^k)' d^k|^2}{2L\|d^k\|^2} \\ &\leq -\frac{c^2 \|\nabla f(x^k)\|^2}{2L}. \end{aligned}$$

Using this relation together with the fact $f(x^{k+1}) \leq f(\tilde{x}^k)$, we obtain

$$f(x^{k+1}) \leq f(x^k) - \frac{c^2 \|\nabla f(x^k)\|^2}{2L}. \tag{3.31}$$

Since X^*, the set of global minima of f, is nonempty and compact, all the level sets of f are compact (Prop. B.9 in Appendix B). Thus, $\{x^k\}$ is bounded, and by Prop. 1.2.1, all limit points of $\{x^k\}$ belong to X^*, and the distance of x^k from X^*, defined by

$$d(x^k, X^*) = \min_{x^* \in X^*} \|x^k - x^*\|,$$

converges to 0. Using the convexity of f, we also have for every global minimum x^*

$$f(x^k) - f(x^*) \leq \nabla f(x^k)'(x^k - x^*) \leq \|\nabla f(x^k)\| \cdot \|x^k - x^*\|,$$

from which, by minimizing over $x^* \in X^*$,

$$f(x^k) - f^* \leq \|\nabla f(x^k)\|\, d(x^k, X^*). \tag{3.32}$$

Let us denote for all k

$$e^k = f(x^k) - f^*.$$

Combining Eqs. (3.31) and (3.32), we obtain

$$e^{k+1} \leq e^k - \frac{c^2 (e^k)^2}{2L d(x^k, X^*)^2}, \qquad \forall\, k, \tag{3.33}$$

where we assume without loss of generality that $d(x^k, X^*) \neq 0$.

We will show that Eq. (3.33) implies that $e^k = o(1/k)$. Indeed we have

$$0 < e^{k+1} \leq e^k \left(1 - \frac{c^2 e^k}{2L d(x^k, X^*)^2}\right),$$

$$0 < 1 - \frac{c^2 e^k}{2Ld(x^k, X^*)^2},$$

from which

$$(e^{k+1})^{-1} \geq (e^k)^{-1}\left(1 - \frac{c^2 e^k}{2Ld(x^k, X^*)^2}\right)^{-1} \geq (e^k)^{-1}\left(1 + \frac{c^2 e^k}{2Ld(x^k, X^*)^2}\right)$$
$$= (e^k)^{-1} + \frac{c^2}{2Ld(x^k, X^*)^2}.$$

Summing this inequality over all k, we obtain

$$e^k \leq \left((e^0)^{-1} + \frac{c^2}{2L}\sum_{i=0}^{k-1} d(x^i, X^*)^{-2}\right)^{-1},$$

or

$$ke^k \leq \left(\frac{1}{ke^0} + \frac{c^2}{2Lk}\sum_{i=0}^{k-1} d(x^i, X^*)^{-2}\right)^{-1}. \tag{3.34}$$

Since $d(x^i, X^*) \to 0$, we have $d(x^i, X^*)^{-2} \to \infty$ and

$$\frac{c^2}{2Lk}\sum_{i=0}^{k-1} d(x^i, X^*)^{-2} \to \infty.$$

Therefore the right-hand side of Eq. (3.34) tends to 0, implying that $e^k = o(1/k)$. **Q.E.D.**

Note that the preceding proof can be modified to cover the case where the Lipschitz condition (3.27) holds within the set $\{x \mid f(x) \leq f(x^0)\}$. Furthermore, the proof goes through for any stepsize rule for which a relation of the form $f(x^{k+1}) \leq f(x^k) - \gamma\|\nabla f(x^k)\|^2$ can be established for some $\gamma > 0$ [cf. Eq. (3.31)]; see Exercise 3.8.

With additional assumptions on the structure of the function f some more precise convergence rate results can be obtained. In particular, if f is convex, has a unique minimum x^*, and satisfies the following growth condition

$$f(x) - f(x^*) \geq q\|x - x^*\|^\beta, \qquad \forall\, x \text{ such that } f(x) \leq f(x^0),$$

for some scalars $q > 0$ and $\beta > 2$, it can be shown [Dun81] that for the method of steepest descent with the Armijo rule we have

$$f(x^k) - f(x^*) = O\left(\frac{1}{k^{\frac{\beta}{\beta-2}}}\right).$$

EXERCISES

3.1

Estimate the rate of convergence of steepest descent with the line minimization rule when applied to the function of two variables $f(x,y) = x^2 + 1.999xy + y^2$. Find a starting point for which this estimate is sharp (cf. Fig. 1.3.2).

3.2

Let f be twice continuously differentiable. Consider a sequence $\{x^k\}$ generated by the gradient method $x^{k+1} = x^k + \alpha^k d^k$ and suppose that x^* is a nonsingular local minimum. Assume that, for all k, $\nabla f(x^k) \neq 0$ and $d^k = d(x^k)$, where $d(\cdot)$ is a continuous function of x with

$$\lim_{x \to x^*,\ \nabla f(x) \neq 0} \frac{\left\| d(x) + \left(\nabla^2 f(x)\right)^{-1} \nabla f(x) \right\|}{\left\| \nabla f(x) \right\|} = 0.$$

Furthermore, α^k is chosen by means of the Armijo rule with initial stepsize $s = 1$ and $\sigma < 1/2$. Show that there exists an $\epsilon > 0$ such that if $\|x^0 - x^*\| < \epsilon$, then:

(a) $\{x^k\}$ converges to x^*.

(b) $\alpha^k = 1$ for all k.

(c) $\lim_{k \to \infty} \left(\|x^{k+1} - x^*\| / \|x^k - x^*\| \right) = 0$.

3.3

Consider a positive definite quadratic problem with Hessian matrix Q. Suppose we use scaling with the diagonal matrix whose ith diagonal element is q_{ii}^{-1}, where q_{ii} is the ith diagonal element of Q. Show that if Q is 2×2, this diagonal scaling improves the condition number of the problem and the convergence rate of steepest descent. (*Note*: This need not be true for dimensions higher than 2.)

3.4 (Steepest Descent with Errors)

Consider the steepest descent method

$$x^{k+1} = x^k - s\left(\nabla f(x^k) + e^k\right),$$

where s is a constant stepsize, e^k is an error satisfying $\|e^k\| \leq \delta$ for all k, and f is the positive definite quadratic function

$$f(x) = \tfrac{1}{2}(x - x^*)' Q (x - x^*).$$

cent with the line minimiza-
bles $f(x,y) = x^2+1.999xy+$
is sharp (cf. Fig. 1.3.2).

ider a sequence $\{x^k\}$ gener-
nd suppose that x^* is a non-
$\nabla f(x^k) \neq 0$ and $d^k = d(x^k)$,

$\frac{\nabla f(x)\|}{} = 0.$

ijo rule with initial stepsize
> 0 such that if $\|x^0 - x^*\| < \epsilon$,

with Hessian matrix Q. Sup-
hose ith diagonal element is
Show that if Q is 2×2, this
of the problem and the con-
ed not be true for dimensions

$e^k)$,

isfying $\|e^k\| \leq \delta$ for all k, and

$x^*)$.

Let

$$q = \max\{|1 - sm|, |1 - sM|\},$$

where

m : smallest eigenvalue of Q, M : largest eigenvalue of Q,

and assume that $q < 1$. Show that for all k, we have

$$\|x^k - x^*\| \leq \frac{s\delta}{1-q} + q^k\|x^0 - x^*\|.$$

3.5

Consider the positive definite quadratic function $f(x) = \frac{1}{2}x'Qx$ and the steepest descent method with the stepsize α^k chosen by the Goldstein rule. Show that for all k,

$$f(x^{k+1}) \leq \left(1 - \frac{16\sigma(1-\sigma)Mm}{(M+m)^2}\right) f(x^k).$$

Explain why when $\sigma = 1/2$ this relation yields the result of Prop. 1.3.1. *Hint*: Use the result of Exercise 2.9.

3.6 [Ber82a]

Consider a scalar sequence $\{e^k\}$ with $e^k \geq 0$ for all k, and $e^k \to 0$. We say that $\{e^k\}$ converges *faster than linearly with convergence ratio* β, where $0 < \beta < 1$, if for every $\overline{\beta} \in (\beta, 1)$ and $q > 0$, there exists $\overline{k}$ such that

$$e^k \leq q\overline{\beta}^k, \qquad \forall\, k \geq \overline{k}.$$

We say that $\{e^k\}$ converges *slower than linearly with convergence ratio* β, where $0 < \beta < 1$, if for every $\overline{\beta} \in (\beta, 1)$ and $q > 0$, there exists $\overline{k}$ such that

$$q\overline{\beta}^k \leq e^k, \qquad \forall\, k \geq \overline{k}.$$

We say that $\{e^k\}$ converges *linearly with convergence ratio* β if it converges both faster and slower than linearly with convergence ratio β. Show that:

(a) $\{e^k\}$ converges faster than linearly with convergence ratio β if and only if

$$\limsup_{k\to\infty} (e^k)^{1/k} \leq \beta.$$

$\{e^k\}$ converges slower than linearly with convergence ratio β if and only if

$$\liminf_{k\to\infty} (e^k)^{1/k} \geq \beta.$$

$\{e^k\}$ converges linearly with convergence ratio β if and only if

$$\lim_{k\to\infty} (e^k)^{1/k} = \beta.$$

(b) Assume that $e^k \neq 0$ for all k, and denote

$$\beta_1 = \liminf_{k\to\infty} \frac{e^{k+1}}{e^k}, \qquad \beta_2 = \limsup_{k\to\infty} \frac{e^{k+1}}{e^k}.$$

Show that if $0 < \beta_1 \leq \beta_2 < 1$, then $\{e^k\}$ converges faster than linearly with convergence ratio β_2 and slower than linearly with convergence ratio β_1. Furthermore, if $\beta_1 = \beta_2 = 0$, then $\{e^k\}$ converges superlinearly.

3.7

Consider a scalar sequence $\{e^k\}$ with $e^k > 0$ for all k, and $e^k \to 0$. Show that $\{e^k\}$ converges superlinearly with order p if

$$\limsup_{k\to\infty} \frac{e^{k+1}}{(e^k)^p} < \infty.$$

3.8

Prove the result of Prop. 1.3.3 for the steepest descent case $[d^k = -\nabla f(x^k)]$, and assuming that the stepsize is not chosen by the line minimization rule but is instead a sufficiently small constant.

3.9 (The Heavy Ball Method [Pol64])

Consider the following variant of the steepest descent method:

$$x^{k+1} = x^k - \alpha \nabla f(x^k) + \beta(x^k - x^{k-1}), \qquad k = 1, 2, \ldots,$$

where α is a constant positive stepsize and β is a scalar with $0 < \beta < 1$.

(a) Let f be the quadratic function $f(x) = (1/2)x'Qx + c'x$, where Q is positive definite and symmetric, and let m and M be the minimum and the maximum eigenvalues of Q, respectively. Show that the method converges linearly to the unique solution if $0 < \alpha < 2(1+\beta)/M$. Show that with optimal choices of α and β, the ratio of linear convergence is

$$\frac{\sqrt{M} - \sqrt{m}}{\sqrt{M} + \sqrt{m}},$$

atio β if and only if

$$\limsup_{k\to\infty} \frac{e^{k+1}}{e^k}.$$

onverges faster than linearly
linearly with convergence ra-
$\{e^k\}$ converges superlinearly.

all k, and $e^k \to 0$. Show that

lescent case $[d^k = -\nabla f(x^k)]$,
y the line minimization rule

escent method:

$$k = 1, 2, \ldots,$$

a scalar with $0 < \beta < 1$.

$(1/2)x'Qx + c'x$, where Q is
and M be the minimum and
rely. Show that the method
f $0 < \alpha < 2(1+\beta)/M$. Show
ratio of linear convergence is

which, if $m < M$, is faster than the corresponding ratio of the steepest descent method where $\beta = 0$ and α is chosen optimally [cf. Eq. (3.1)]. *Hint*: Consider the iteration

$$\begin{pmatrix} x^{k+1} \\ x^k \end{pmatrix} = \begin{pmatrix} (1+\beta)I - \alpha Q & -\beta I \\ I & 0 \end{pmatrix} \begin{pmatrix} x^k \\ x^{k-1} \end{pmatrix}$$

and show that v is an eigenvalue of the matrix in the above equation if and only if $v + \beta/v$ is equal to $1 + \beta - \alpha\lambda$ where λ is an eigenvalue of Q.

(b) It is generally conjectured that in comparison to steepest descent, the method is less prone to getting trapped at "shallow" local minima, and tends to behave better for difficult problems where the cost function is alternatively very flat and very steep. Argue for or against this conjecture.

(c) In support of your answer in (b), write a computer program to test the method with $\beta = 0$ and $\beta > 0$ with one-dimensional cost functions of the form

$$f(x) = \tfrac{1}{2}x^2\big(1 + \gamma\cos(x)\big),$$

where $\gamma \in (0, 1)$, and

$$f(x) = \tfrac{1}{2}\sum_{i=1}^{m} |z_i - \tanh(xy_i)|^2,$$

where z_i and y_i are given scalars.

1.4 NEWTON'S METHOD AND VARIATIONS

In the last two sections we emphasized a basic tradeoff in gradient methods: implementation simplicity versus fast convergence. We have already discussed steepest descent, one of the simplest but also one of the slowest methods. We now consider its opposite extreme, Newton's method, which is arguably the most complex and also the fastest of the gradient methods (under appropriate conditions).

Newton's method consists of the iteration

$$x^{k+1} = x^k - \alpha^k\big(\nabla^2 f(x^k)\big)^{-1}\nabla f(x^k), \tag{4.1}$$

assuming that the Newton direction

$$d^k = -\big(\nabla^2 f(x^k)\big)^{-1}\nabla f(x^k) \tag{4.2}$$

is defined and is a direction of descent [i.e., $d_k'\nabla f(x^k) < 0$]. As explained in the preceding section, one may view this iteration as a scaled version of steepest descent where the "optimal" scaling matrix $\left(\nabla^2 f(x^k)\right)^{-1}$ is used. It is worth mentioning in this connection that *Newton's method is "scale-free"*, in the sense that it cannot be affected by a change in coordinate system as is true for steepest descent (see Exercise 4.1).

When the Armijo rule is used with initial stepsize $s = 1$, then no reduction of the stepsize will be necessary near a nonsingular minimum (positive definite Hessian), as shown in Prop. 1.3.2. Thus, near convergence the method takes the form

$$x^{k+1} = x^k - \left(\nabla^2 f(x^k)\right)^{-1}\nabla f(x^k), \tag{4.3}$$

which will be referred to as the *pure form of Newton's method*. On the other hand, far from such a local minimum, the Hessian matrix may be singular or the Newton direction of Eq. (4.2) may not be a direction of descent because the Hessian $\nabla^2 f(x^k)$ is not positive definite. Thus the analysis of Newton's method has two principal aspects:

(a) Local convergence, dealing with the behavior of the pure form of the method near a nonsingular local minimum.

(b) Global convergence, addressing the modifications that are necessary to ensure that the method is valid and is likely to converge to a local minimum when started far from all local minima.

We consider these issues in this section and we also discuss some variations of Newton's method, which are aimed at reducing the overhead for computing the Newton direction.

Local Convergence

The local convergence result for gradient methods (Prop. 1.2.5) together with the superlinear convergence result for Newton-like methods (Prop. 1.3.2) suggest that the pure form of Newton's method converges superlinearly when started close enough to a nonsingular local minimum. Results of this type hold for a more general form of Newton's method, that can be used to solve the system of n equations with n unknowns

$$g(x) = 0, \tag{4.4}$$

where $g : \Re^n \mapsto \Re^n$ is a continuously differentiable function. This method has the form

$$x^{k+1} = x^k - \left(\nabla g(x^k)'\right)^{-1} g(x^k), \tag{4.5}$$

and for the special case where $g(x)$ is equal to the gradient $\nabla f(x)$, it yields the pure form of Eq. (4.3). [A continuously differentiable function $g : \Re^n \mapsto$

$\nabla f(x^k) < 0$]. As explained
ıtion as a scaled version of
ıtrix $\left(\nabla^2 f(x^k)\right)^{-1}$ is used.
Newton's method is "scale-
ɔy a change in coordinate
ɔise 4.1).
ıl stepsize $s = 1$, then no
ır a nonsingular minimum
.2. Thus, near convergence

$'f(x^k)$, (4.3)

Newton's method. On the
he Hessian matrix may be
may not be a direction of
ɔositive definite. Thus the
aspects:

ɾior of the pure form of the
ı.

ications that are necessary
likely to converge to a local
minima.

and we also discuss some
d at reducing the overhead

; methods (Prop. 1.2.5) to-
t for Newton-like methods
Newton's method converges
ıonsingular local minimum.
n of Newton's method, that
with n unknowns

(4.4)

able function. This method

$g(x^k)$, (4.5)

he gradient $\nabla f(x)$, it yields
ɛrentiable function $g : \Re^n \mapsto$

$\Re^n$ need not be equal to the gradient of some function. In particular, $g(x) = \nabla f(x)$ for some $f : \Re^n \mapsto \Re$, if and only if the $n \times n$ matrix $\nabla g(x)$ is symmetric for all x ([OrR70], p. 95). Thus, the equation version of Newton's method (4.5) is more broadly applicable than the optimization version of Eq. (4.3).]

There is a simple argument that shows the fast convergence of Newton's method (4.5). Suppose that the method generates a sequence $\{x^k\}$ that converges to a vector x^* such that $g(x^*) = 0$ and $\nabla g(x^*)$ is invertible. Let us use Taylor's theorem to write

$$0 = g(x^*) = g(x^k) + \nabla g(x^k)'(x^* - x^k) + o\big(\|x^k - x^*\|\big).$$

By multiplying this relation with $\left(\nabla g(x^k)'\right)^{-1}$ we have

$$x^k - x^* - \left(\nabla g(x^k)'\right)^{-1} g(x^k) = o\big(\|x^k - x^*\|\big),$$

so for the pure Newton iteration $x^{k+1} = x^k - \left(\nabla g(x^k)'\right)^{-1} g(x^k)$ we obtain

$$x^{k+1} - x^* = o\big(\|x^k - x^*\|\big),$$

or, for $x^k \neq x^*$,

$$\lim_{k\to\infty} \frac{\|x^{k+1} - x^*\|}{\|x^k - x^*\|} = \lim_{k\to\infty} \frac{o\big(\|x^k - x^*\|\big)}{\|x^k - x^*\|} = 0,$$

implying superlinear convergence. This argument can also be used to show convergence to x^* if the initial vector x^0 is sufficiently close to x^*. The following proposition proves a more detailed result.

Proposition 1.4.1: Consider a function $g : \Re^n \mapsto \Re^n$, and a vector x^* such that $g(x^*) = 0$. For $\delta > 0$, let S_δ denote the sphere $\{x \mid \|x - x^*\| \leq \delta\}$. Assume that g is continuously differentiable within some sphere $S_{\bar{\delta}}$ and that $\nabla g(x^*)$ is invertible.

(a) There exists $\delta > 0$ such that if $x^0 \in S_\delta$, the sequence $\{x^k\}$ generated by the iteration

$$x^{k+1} = x^k - \left(\nabla g(x^k)'\right)^{-1} g(x^k)$$

is defined, belongs to S_δ, and converges to x^*. Furthermore, $\{\|x^k - x^*\|\}$ converges superlinearly.

(b) If for some $L > 0$, $M > 0$, $\delta > 0$, and for all x and y in S_δ,

$$\|\nabla g(x) - \nabla g(y)\| \leq L\|x - y\|, \qquad \left\|\left(\nabla g(x)'\right)^{-1}\right\| \leq M, \quad (4.6)$$

then, if $x^0 \in S_\delta$, we have

$$\|x^{k+1} - x^*\| \leq \frac{LM}{2}\|x^k - x^*\|^2, \qquad \forall\ k = 0, 1, \ldots,$$

so $\{\|x^k - x^*\|\}$ converges superlinearly with order at least two.

Proof: (a) Let $\delta > 0$ be such that $\left(\nabla g(x)'\right)^{-1}$ exists within S_δ and let $M > 0$ be such that

$$\left\|\left(\nabla g(x)'\right)^{-1}\right\| \leq M, \qquad \forall\ x \in S_\delta.$$

Assuming $x \in S_\delta$, and using the relation

$$g(x^k) = \int_0^1 \nabla g\big(x^* + t(x^k - x^*)\big)' dt (x^k - x^*),$$

we estimate $\|x^{k+1} - x^*\|$ as

$$\begin{aligned}
\|x^{k+1} - x^*\| &= \left\|x^k - x^* - \left(\nabla g(x^k)'\right)^{-1} g(x^k)\right\| \\
&= \left\|\left(\nabla g(x^k)'\right)^{-1}\left(\nabla g(x^k)'(x^k - x^*) - g(x^k)\right)\right\| \qquad (4.7) \\
&= \left\|\left(\nabla g(x^k)'\right)^{-1}\left(\nabla g(x^k)' - \int_0^1 \nabla g\big(x^* + t(x^k - x^*)\big)' dt\right)(x^k - x^*)\right\| \\
&= \left\|\left(\nabla g(x^k)'\right)^{-1}\left(\int_0^1 \left[\nabla g(x^k)' - \nabla g\big(x^* + t(x^k - x^*)\big)'\right] dt\right)(x^k - x^*)\right\| \\
&\leq M\left(\int_0^1 \left\|\nabla g(x^k) - \nabla g\big(x^* + t(x^k - x^*)\big)\right\| dt\right)\|x^k - x^*\|.
\end{aligned}$$

By continuity of ∇g, we can take δ sufficiently small to ensure that the term under the integral sign is arbitrarily small. The convergence $x^k \to x^*$ and the superlinear convergence of $\|x^k - x^*\|$ follow.

(b) If the condition (4.6) holds, Eq. (4.7) yields

$$\|x^{k+1} - x^*\| \leq M\left(\int_0^1 Lt\|x^k - x^*\| dt\right)\|x^k - x^*\| = \frac{LM}{2}\|x^k - x^*\|^2.$$

Q.E.D.

all x and y in S_δ,

$$\left\| \left(\nabla g(x)'\right)^{-1} \right\| \leq M, \quad (4.6)$$

$\forall\ k = 0, 1, \ldots,$

ith order at least two.

exists within S_δ and let

$\in S_\delta$.

$lt(x^k - x^*)$,

(4.7)

$$\left. - x^*))' dt \right) (x^k - x^*) \Big\|$$

$$\left. {}^k - x^*))' \right] dt \Big) (x^k - x^*) \Big\|$$

$t \Big) \|x^k - x^*\|.$

small to ensure that the The convergence $x^k \to x^*$ llow.

$$\|x^{k+1} - x^*\| = \frac{LM}{2} \|x^k - x^*\|^2.$$

A related result is the following. Its proof is left for the reader.

Proposition 1.4.2: Under the assumptions of Prop. 1.4.1(a), given any $r > 0$, there exists a $\delta > 0$ such that if $\|x^k - x^*\| < \delta$, then

$$\|x^{k+1} - x^*\| \leq r\|x^k - x^*\|, \qquad \|g(x^{k+1})\| \leq r\|g(x^k)\|.$$

Thus, once it gets "near" a solution x^* where $\nabla g(x^*)$ is invertible, the pure form of Newton's method converges extremely fast, typically taking a handful of iterations to achieve very high solution accuracy; see Fig. 1.4.1. Unfortunately, it is typically difficult to predict whether a given starting point is sufficiently near to a solution for the fast convergence rate of Newton's method to become effective right away. Thus, in practice one can only expect that *eventually* the fast convergence rate of Newton's method will become effective. Figure 1.4.2 illustrates how the method can fail to converge when started far from a solution.

Global Convergence

Newton's method in its pure form for unconstrained minimization of f has several serious drawbacks.

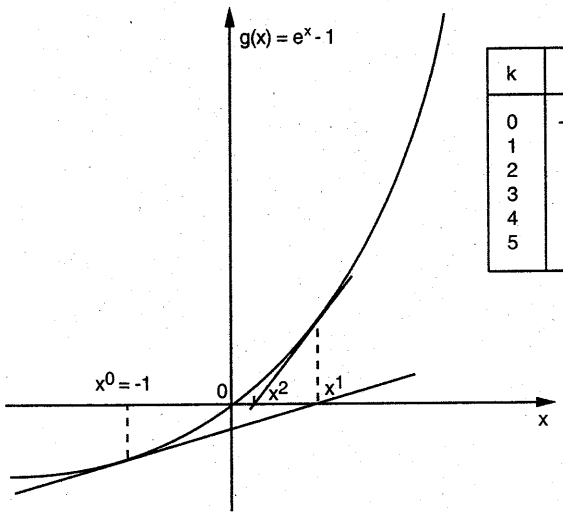
(a) The inverse $\left(\nabla^2 f(x^k)\right)^{-1}$ may fail to exist, in which case the method breaks down. This will happen, for example, in regions where f is linear ($\nabla^2 f = 0$).

(b) The pure form is not a descent method; it may happen that $f(x^{k+1}) > f(x^k)$.

(c) The pure form tends to be attracted by local maxima just as much as it is attracted by local minima. It just tries to solve the system of equations $\nabla f(x) = 0$.

For these reasons, it is necessary to modify the pure form of Newton's method to turn it into a reliable minimization algorithm. There are several schemes that accomplish this by converting the pure form into a gradient method with a gradient related direction sequence. Simultaneously the modifications are such that, near a nonsingular local minimum, the algorithm assumes the pure form of Newton's method (4.3) and achieves the attendant fast convergence rate.

A simple possibility is to replace the Newton direction by the steepest descent direction (possibly after diagonal scaling), whenever the Newton



k	x^k	$g(x^k)$
0	- 1.00000	- 0.63212
1	0.71828	1.05091
2	0.20587	0.22859
3	0.01981	0.02000
4	0.00019	0.00019
5	0.00000	0.00000

Figure 1.4.1. Fast convergence of Newton's method for solving the equation $e^x - 1 = 0$.

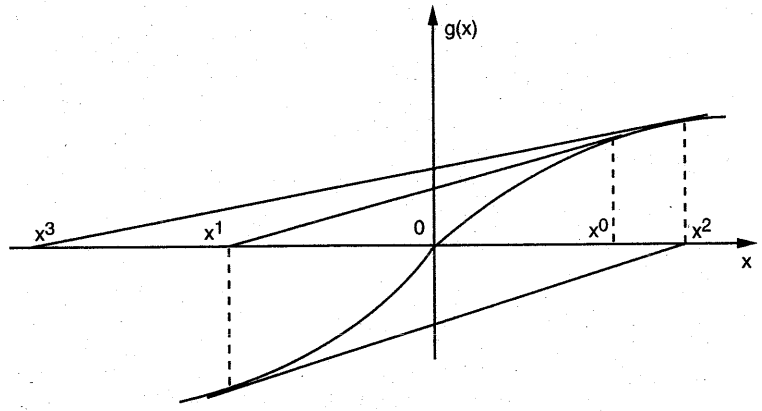


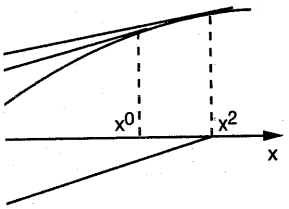
Figure 1.4.2. Divergence of Newton's method for solving an equation $g(x) = 0$ of a single variable x, when the starting point is far from the solution. This phenomenon typically happens if $\|\nabla g(x)\|$ tends to decrease as $\|x\| \to \infty$.

direction is either not defined or is not a descent direction.† With proper

† Interestingly, this motivated the development of steepest descent by M. Augustin Cauchy. In his original paper [Cau47], Cauchy states as motivation for the steepest descent method its capability to obtain a close approximation to the solution, in which case "... one can obtain new approximations very rapidly with

k	x^k	$g(x^k)$
0	- 1.00000	- 0.63212
1	0.71828	1.05091
2	0.20587	0.22859
3	0.01981	0.02000
4	0.00019	0.00019
5	0.00000	0.00000

ethod for solving the equation



r solving an equation $g(x) = 0$
s far from the solution. This
o decrease as $\|x\| \to \infty$.

cent direction.† With proper

ient of steepest descent by M.
Cauchy states as motivation for
in a close approximation to the
proximations very rapidly with

safeguards, such a method has appropriate convergence and asymptotic rate of convergence properties (see Exercise 4.3 and for a related method, see Exercise 4.4). However, its performance at the early iterations may be quite slow, whether the Newton direction or the steepest descent direction is used in these iterations.

Generally, no modification of Newton's method can be guaranteed to converge fast in the early iterations, but there are schemes that can use second derivative information effectively, even when the Hessian is not positive definite. These schemes are based on making diagonal modifications to the Hessian; that is, they obtain the direction d^k by solving a system of the form

$$\big(\nabla^2 f(x^k) + \Delta^k\big)d^k = -\nabla f(x^k),$$

whenever the Newton direction does not exist or is not a descent direction. Here Δ^k is a diagonal matrix such that

$$\nabla^2 f(x^k) + \Delta^k : \text{ positive definite.}$$

We outline some possibilities.

*Modified Cholesky Factorization**

It can be shown that every positive definite matrix Q has a unique factorization of the form

$$Q = LL',$$

where L is a lower triangular matrix; this is known as the *Cholesky factorization of* Q (see Appendix D). Systems of equations of the form $Qx = b$ can be solved by first solving for y the triangular system $Ly = b$, and then by solving for x the triangular system $L'x = y$. These triangular systems can be solved easily [in $O(n^2)$ operations as opposed to general systems, which require $O(n^3)$ operations; see Appendix D]. Since calculation of the Newton direction involves solution of the system

$$\nabla^2 f(x^k)d^k = -\nabla f(x^k),$$

it is natural to compute d^k by attempting to form the Cholesky factorization of $\nabla^2 f(x^k)$. During this process, one can detect whether $\nabla^2 f(x^k)$ is either nonpositive definite or nearly singular, in which case some of the diagonal elements of $\nabla^2 f(x^k)$ are suitably increased to ensure that the resulting matrix is positive definite. This is done sequentially during the factorization process, so in the end we obtain

$$L^k L^{k'} = \nabla^2 f(x^k) + \Delta^k,$$

where L^k is lower triangular and nonsingular, and Δ^k is diagonal.

---

the aid of the linear or Newton's method ..." (Note the attribution to Newton by Cauchy.)

As an illustration, consider the 2-dimensional case (for the general case, see Appendix D). Let

$$\nabla^2 f(x^k) = \begin{pmatrix} h_{11} & h_{12} \\ h_{21} & h_{22} \end{pmatrix}$$

and let the desired factorization be of the form

$$LL' = \begin{pmatrix} \alpha & 0 \\ \gamma & \beta \end{pmatrix} \cdot \begin{pmatrix} \alpha & \gamma \\ 0 & \beta \end{pmatrix}.$$

We choose α, β, and γ, so that $\nabla^2 f(x^k) = LL'$ if $\nabla^2 f(x^k)$ is positive definite, and we appropriately modify h_{11} and h_{22} otherwise. This determines the first diagonal element α according to the relation

$$\alpha = \begin{cases} \sqrt{h_{11}} & \text{if } h_{11} > 0 \\ \sqrt{h_{11} + \delta_1} & \text{otherwise} \end{cases}$$

where δ_1 is such that $h_{11} + \delta_1 > 0$. Given α, we can calculate γ by equating the corresponding elements of $\nabla^2 f(x^k)$ and LL'. We obtain $\gamma\alpha = h_{12}$ or

$$\gamma = \frac{h_{12}}{\alpha}.$$

We can now calculate the second diagonal element β by equating the corresponding elements of $\nabla^2 f(x^k)$ and LL', after appropriately modifying h_{22} if necessary,

$$\beta = \begin{cases} \sqrt{h_{22} - \gamma^2} & \text{if } h_{22} > \gamma^2, \\ \sqrt{h_{22} - \gamma^2 + \delta_2} & \text{otherwise,} \end{cases}$$

where δ_2 is such that $h_{22} - \gamma^2 + \delta_2 > 0$. The method for choosing the increments δ_1 and δ_2 is largely heuristic. One possibility is discussed in Appendix D, which also describes more sophisticated versions of the above procedure where a positive increment is added to the diagonal elements of the Hessian even when the corresponding diagonal elements of the factorization are positive but very close to zero.

Given the $L^k L^{k'}$ factorization, the direction d^k is obtained by solving the system

$$L^k L^{k'} d^k = -\nabla f(x^k).$$

The next iterate is

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

where α^k is chosen according to the Armijo rule or one of the other stepsize rules we have discussed.

To guarantee convergence, the increments added to the diagonal elements of the Hessian can be chosen so that $\{d^k\}$ is gradient related (cf. Prop. 1.2.1). Also, these increments can be chosen to be zero near a nonsingular local minimum. In particular, with proper safeguards, near such a point, the method becomes identical to the pure form of Newton's method and achieves the corresponding superlinear convergence rate (see Appendix D).

*Trust Region Methods**

As explained in Section 1.2, the pure Newton step is obtained by minimizing over d the second order Taylor series approximation of f around x^k, given by

$$f^k(d) = f(x^k) + \nabla f(x^k)'d + \tfrac{1}{2}d'\nabla^2 f(x^k)d.$$

We know that $f^k(d)$ is a good approximation of $f(x^k + d)$ when d is in a small neighborhood of zero, but the difficulty is that with unconstrained minimization of $f^k(d)$ one may obtain a step that lies outside this neighborhood. It therefore makes sense to consider a *restricted Newton step* d^k obtained by minimizing $f^k(d)$ over a suitably small neighborhood of zero, called the *trust region*:

$$d^k = \arg \min_{||d|| \le \gamma^k} f^k(d), \tag{4.8}$$

where γ^k is some positive scalar. [It can be shown that the restricted Newton step d^k also solves a system of the form $\big(\nabla^2 f(x^k) + \delta^k I\big)d = -\nabla f(x^k)$, where I is the identity matrix and δ^k is a nonnegative scalar (a Lagrange multiplier in the terminology of Chapter 3), so the preceding method of determining d^k fits the general framework of using a correction of the Hessian matrix by a positive semidefinite matrix.] An approximate solution of the constrained minimization problem of Eq. (4.8) can be obtained quickly using the fact that it has only one constraint (see [MoS83]).

An important observation here is that even if $\nabla^2 f(x^k)$ is not positive definite or, more generally, even if the pure Newton direction is not a descent direction, the restricted Newton step d^k improves the cost, provided $\nabla f(x^k) \neq 0$ and γ^k is sufficiently small. To see this, note that we have for all d with $||d|| \le \gamma^k$

$$f(x^k + d) = f^k(d) + o\big((\gamma^k)^2\big),$$

so that

$$\begin{aligned} f(x^k + d^k) &= f^k(d^k) + o\big((\gamma^k)^2\big) \\ &= f(x^k) + \min_{||d|| \le \gamma^k} \Big\{\nabla f(x^k)'d + \tfrac{1}{2}d'\nabla^2 f(x^k)d\Big\} + o\big((\gamma^k)^2\big). \end{aligned}$$

Therefore, denoting

$$\tilde{d}^k = -\frac{\nabla f(x^k)}{||\nabla f(x^k)||}\gamma^k,$$

we have

$$\begin{aligned} f(x^k + d^k) \le & f(x^k) + \nabla f(x^k)'\tilde{d}^k + \tfrac{1}{2}\tilde{d}^{k\prime}\nabla^2 f(x^k)\tilde{d}^k + o\big((\gamma^k)^2\big) \\ = & f(x^k) - \gamma^k||\nabla f(x^k)|| + \frac{(\gamma^k)^2}{2||\nabla f(x^k)||^2}\nabla f(x^k)'\nabla^2 f(x^k)\nabla f(x^k) \\ & + o\big((\gamma^k)^2\big). \end{aligned}$$

For γ^k sufficiently small, the negative term $-\gamma^k||\nabla f(x^k)||$ dominates the last two terms on the right-hand side above, showing that

$$f(x^{k+1}) < f(x^k).$$

It can be seen in fact from the preceding relations that a cost improvement is possible even when $\nabla f(x^k) = 0$, provided γ^k is sufficiently small and f has a direction of negative curvature at x^k, that is, $\nabla^2 f(x^k)$ is not positive semidefinite. Thus the preceding procedure will fail to improve the cost only if $\nabla f(x^k) = 0$ and $\nabla^2 f(x^k)$ is positive semidefinite, that is, x^k satisfies the first and second order necessary conditions. In particular, one can typically make progress even if x^k is a stationary point that is not a local minimum.

We are thus motivated to consider a method of the form

$$x^{k+1} = x^k + d^k,$$

where d^k is the restricted Newton step corresponding to a suitably chosen scalar γ^k as per Eq. (4.8). Here, for a given x^k, γ^k should be small enough so that there is cost improvement; one possibility is to start from an initial trial γ^k and successively reduce γ^k by a certain factor as many times as necessary until a cost reduction occurs $[f(x^{k+1}) < f(x^k)]$. The choice of the initial trial value for γ^k is crucial here; if it is chosen too large, a large number of reductions may be necessary before a cost improvement occurs; if it is chosen too small the convergence rate may be poor. In particular, to maintain the superlinear convergence rate of Newton's method, as x^k approaches a nonsingular local minimum, one should select the initial trial value of γ^k sufficiently large so that the restricted Newton step and the pure Newton step coincide.

A reasonable way to adjust the initial trial value for γ^k is to increase this value when the method appears to be progressing well and to decrease this value otherwise. One can measure progress by using the ratio of actual over predicted cost improvement [based on the approximation $f^k(d)$]

$$r^k = \frac{f(x^k) - f(x^{k+1})}{f(x^k) - f^k(d^k)}.$$

In particular, it makes sense to increase the initial trial value for γ ($\gamma^{k+1} > \gamma^k$) if this ratio is close to or above unity, and decrease γ otherwise. The following algorithm is a typical example of such a method. Given x^k and an initial trial value γ^k, it determines x^{k+1} and an initial trial value γ^{k+1} by using two threshold values σ_1, σ_2 with $0 < \sigma_1 \leq \sigma_2 \leq 1$ and two factors β_1, β_2 with $0 < \beta_1 < 1 < \beta_2$ (typical values are $\sigma_1 = 0.2$, $\sigma_2 = 0.8$, $\beta_1 = 0.25$, $\beta_2 = 2$).

Step 1: Find

$$d^k = \arg \min_{\|d\| \leq \gamma^k} f^k(d), \tag{4.9}$$

If $f^k(d^k) = f(x^k)$ stop (x^k satisfies the first and second order necessary conditions for a local minimum); else go to Step 2.

Step 2: If $f(x^k + d^k) < f(x^k)$ set

$$x^{k+1} = x^k + d^k \tag{4.10}$$

calculate

$$r^k = \frac{f(x^k) - f(x^{k+1})}{f(x^k) - f^k(d^k)} \tag{4.11}$$

and go to Step 3; else set $\gamma^k := \beta_1 \|d^k\|$ and go to Step 1.

Step 3: Set

$$\gamma^{k+1} = \begin{cases} \beta_1 \|d^k\| & \text{if } r^k < \sigma_1, \\ \beta_2 \gamma^k & \text{if } \sigma_2 \leq r_k \text{ and } \|d^k\| = \gamma^k, \\ \gamma^k & \text{otherwise.} \end{cases} \tag{4.12}$$

Go to the next iteration.

Assuming that f is twice continuously differentiable, it is possible to show that the above algorithm is convergent in the sense that if $\{x^k\}$ is a bounded sequence, there exists a limit point of $\{x^k\}$ that satisfies the first and the second order necessary conditions for optimality. Furthermore, if $\{x^k\}$ converges to a nonsingular local minimum x^*, then asymptotically, the method is identical to the pure form of Newton's method, thereby attaining a superlinear convergence rate; see the references given at the end of the chapter for proofs of these and other related results for trust region methods.

Newton's Method with Periodic Reevaluation of the Hessian

A variation of Newton's method is obtained if the Hessian matrix $\nabla^2 f$ is recomputed every $p > 1$ iterations rather than at every iteration. In particular, this method, in unmodified form, is given by

$$x^{k+1} = x^k - \alpha^k D^k \nabla f(x^k),$$

where

$$D^{ip+j} = \left(\nabla^2 f(x^{ip})\right)^{-1}, \qquad j = 0, 1, ..., p-1, \ i = 0, 1, ...$$

The idea here is to save the computation and the inversion (or factorization) of the Hessian for the iterations where $j \neq 0$. This reduction in overhead is achieved at the expense of what is usually a small degradation in speed of convergence.

Truncated Newton Methods*

We have so far implicitly assumed that the system $\nabla^2 f(x^k) d^k = -\nabla f(x^k)$ will be solved for the direction d^k by Cholesky factorization or Gaussian elimination, which require a finite number of arithmetic operations $[O(n^3)]$. When the dimension n is large, the calculation required for exact solution of the system may be prohibitive, and one may have to be satisfied with only an approximate solution. Such an approximation may be obtained by using an iterative method. This approach is often used for solving very large linear systems of equations, arising in the solution of partial differential equations, where an adequate approximation to the