Line Search and Gradient Methods

1 Line Search Methods

We are interested in the problem of minimizing an objective function $f : \mathbb{R} \to \mathbb{R}$ (i.e., a one-dimensional problem). The approach is to use an iterative search algorithm, also called a line-search method.

In an iterative algorithm, we start with an initial candidate solution x^0 and generate a sequence of iterates $x^1,\ x^2,\ \cdots$. For each iteration $k = 0, 1, 2, \cdots$, the next point x^{k+1} depends on x^k and the objective function f. The algorithm may use only the value of f at specific points, or perhaps its first derivative f', or even its second derivative f''.

2 Golden Section Search

The search methods allows us to determine the minimizer of an objective function $f : \mathbb{R} \to \mathbb{R}$ over a closed interval, say $[a_0, b_0]$. The only property that we assume of the objective function f is that it is **unimodal**, which means that f has only one local minimizer.

The method is based on evaluating the objective function at different points in the interval $[a_0, b_0]$. We choose these points in such a way that an approximation to the minimizer of f may be achieved in as few evaluations as possible. Our goal is to narrow the range progressively until the minimizer is "boxed in" with sufficient accuracy.

If we evaluate f at only one intermediate point of the interval, we cannot narrow the range within which we know the minimizer is located. We have to evaluate f at two intermediate points. We choose the intermediate points in such a way that the reduction in the range is symmetric, see Fig. 1, in the sense that

$$a_1 - a_0 = b_0 - b_1 = \rho(b_0 - a_0), \quad \rho < 1/2. \tag{1}$$

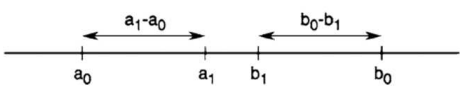

Figure 1: Evaluating the objective function at two intermediate points.

We then evaluate f at the intermediate points. If $f(a_1) < f(b_1)$, then the minimizer must lie in the range $[a_0, b_1]$. If, on the other hand, $f(a_1) \geq f(b_1)$, then the minimizer is located in the range $[a_1, b_0]$.

Starting with the reduced range of uncertainty, we can repeat the process and similarly find two new points, say a_2 and b_2, using the same value of $\rho < 1/2$ as before. However, we would like to minimize the number of objective function evaluations while reducing the width of the uncertainty interval.

Because a_1 is already in the uncertainty interval and $f(a_1)$ is already known, we can make a_1 coincide with b_2. Thus, only one new evaluation of f at a_2 would be necessary.

To find the value of ρ that results in only one new evaluation of f, see Figure 2. Without loss of generality, imagine that the original range $[a_0, b_0]$ is of unit length.

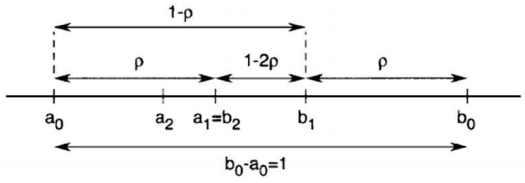

Figure 2: Finding value of ρ resulting in only one new evaluation of f.

Then, to have only one new evaluation of f it is enough to choose ρ so that

$$\rho(b_1 - a_0) = b_1 - b_2. \tag{2}$$

Because $b_1 - a_0 = 1 - \rho$ and $b_1 - b_2 = 1 - 2\rho$, we have

$$\rho(1 - \rho) = 1 - 2\rho. \tag{3}$$

We write the quadratic equation above as

$$\rho^2 - 3\rho + 1 = 0. \tag{4}$$

The solutions are

$$\rho_1 = \frac{3 + \sqrt{5}}{2}, \quad \rho_2 = \frac{3 - \sqrt{5}}{2}. \tag{5}$$

Because we require that $\rho < 1/2$, we take $\rho = (3 - \sqrt{5})/2 \approx 0.382$.

Observe that, $1 - \rho = (\sqrt{5} - 1)/2$, then

$$\frac{\rho}{1 - \rho} = \frac{3 - \sqrt{5}}{\sqrt{5} - 1}. \tag{6}$$

Multiply by $(\sqrt{5} - 1)/(\sqrt{5} - 1)$, noting that $(\sqrt{5} - 1)^2 = 2(3 - \sqrt{5})$ in the denominator, then

$$\frac{\rho}{1 - \rho} = \frac{3 - \sqrt{5}}{\sqrt{5} - 1} = \frac{\sqrt{5} - 1}{2} = \frac{1 - \rho}{1} = \frac{1 - \rho}{\rho + (1 - \rho)}. \tag{7}$$

Thus, dividing a range in the ratio of ρ to $1-\rho$ has the effect that the ratio of the shorter segment to the longer equals the ratio of the longer to the sum of the two. This rule was referred to by ancient Greek geometers as the golden section.

Using the golden section rule means that at every stage of the uncertainty range reduction (except the first), the objective function f need only be evaluated at one new point. The uncertainty range is reduced by the ratio $1-\rho \approx 0.618$ at every stage. Hence, N steps of reduction using the golden section method reduces the range by the factor

$$(1-\rho)^N \approx (0.618)^N. \tag{8}$$

3 Fibonacci Method

Suppose now that we are allowed to vary the value ρ from stage to stage, so that at the kth stage in the reduction process we use a value ρ_k, at the next stage we use a value ρ_{k+1}, and so on.

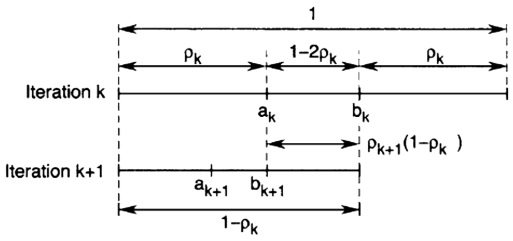

Figure 3: Selecting evaluation points.

From Figure 3, we see that it is sufficient to choose the ρ_k such that

$$\rho_{k+1}(1-\rho_k) = 1-2\rho_k. \tag{9}$$

Note that

$$\rho_{k+1} = \frac{1-2\rho_k}{1-\rho_k} = \frac{1-\rho_k-\rho_k}{1-\rho_k} = 1-\frac{\rho_k}{1-\rho_k}. \tag{10}$$

Suppose that we are given a sequence $\rho_1, \rho_2, \cdots$ satisfying the conditions above and we use this sequence in our search algorithm. Then, after N iterations of the algorithm, the uncertainty range is reduced by a factor of

$$(1-\rho_1)(1-\rho_2)\cdots(1-\rho_N). \tag{11}$$

The Fibonacci sequence F_1, F_2, F_3, $\cdots$ is defined as follows. First, let $F_{-1}=0$ and $F_0=1$ by convention. Then, for $k \geq 0$, $F_{k+1} = F_k + F_{k-1}$. Then we can have

$$\rho_1 = 1-\frac{F_N}{F_{N+1}}, \quad \rho_2 = 1-\frac{F_{N-1}}{F_N}, \quad \cdots, \quad \rho_N = 1-\frac{F_1}{F_2}, \tag{12}$$

where the F_k are the elements of the Fibonacci sequence. The resulting algorithm is called the Fibonacci search method.

In the Fibonacci search method, the uncertainty range is reduced by the factor

$$(1-\rho_1)(1-\rho_2)\cdots(1-\rho_N) = \frac{F_N}{F_{N+1}}\frac{F_{N-1}}{F_N}\cdots\frac{F_1}{F_2} = \frac{F_1}{F_{N+1}} = \frac{1}{F_{N+1}}. \quad (13)$$

Because the Fibonacci method uses the optimal values of ρ_1, ρ_2, $\cdots$, the reduction factor above is less than that of the golden section method. In other words, the Fibonacci method is better than the golden section method in that it gives a smaller final uncertainty range.

We point out that there is an anomaly in the final iteration of the Fibonacci search method, because

$$\rho_N = 1 - \frac{F_1}{F_2} = \frac{1}{2}. \quad (14)$$

Recall that we need two intermediate points at each stage, one that comes from a previous iteration and another that is a new evaluation point. However, with $\rho_N = 1/2$, the two intermediate points coincide in the middle of the uncertainty interval, and therefore we cannot further reduce the uncertainty range. To get around this problem, we perform the new evaluation for the last iteration using $\rho_N = 1/2 - \epsilon$, where ϵ is a small number. In other words, the new evaluation point is just to the left or right of the midpoint of the uncertainty interval.

4 Bisection Method

We consider finding the minimizer of an objective function $f : \mathbb{R} \to \mathbb{R}$ over an interval $[a_0, b_0]$. As before, we assume that the objective function f is unimodal. Further, suppose that f is continuously differentiate and that we can use values of the derivative f' as a basis for reducing the uncertainty interval. The bisection method is a simple algorithm for successively reducing the uncertainty interval based on evaluations of the derivative.

To begin, let

$$x^{(0)} = \frac{a_0 + b_0}{2} \quad (15)$$

be the midpoint of the initial uncertainty interval. Next, evaluate $f'(x^{(0)})$. If $f'(x^{(0)}) > 0$, then we deduce that the minimizer lies to the left of $x^{(0)}$. In other words, we reduce the uncertainty interval to $[a_0, x^{(0)}]$. On the other hand, if $f'(x^{(0)}) < 0$, then we deduce that the minimizer lies to the right of $x^{(0)}$. In this case, we reduce the uncertainty interval to $[x^{(0)}, b_0]$ Finally, if $f'(x^{(0)}) = 0$, then we declare $x^{(0)}$ to be the minimizer and terminate our search.

With the new uncertainty interval computed, we repeat the process iteratively. At each iteration k, we compute the midpoint of the uncertainty interval. Call this point $x^{(k)}$. Depending on the sign of $f'(x^{(k)})$ (assuming that it is nonzero), we reduce the uncertainty interval to the left or right of $x^{(k)}$. If at any

iteration k we find that $f'(x^{(k)}) = 0$, then we declare $x^{(k)}$ to be the minimizer and terminate our search.

Two salient features distinguish the bisection method from the golden section and Fibonacci methods. First, instead of using values of f, the bisection methods uses values of f'. Second, at each iteration, the length of the uncertainty interval is reduced by a factor of $1/2$. Hence, after N steps, the range is reduced by a factor of $(1/2)^N$. This factor is smaller than in the golden section and Fibonacci methods.

5 Newton's Method

Suppose again that we are confronted with the problem of minimizing a function f of a single real variable x. We assume now that at each measurement point $x^{(k)}$ we can determine $f(x^{(k)})$, $f'(x^{(k)})$, and $f''(x^{(k)})$. We can fit a quadratic function through $x^{(k)}$ that matches its first and second derivatives with that of the function f. This quadratic has the form

$$q(x) = f(x^{(k)}) + f'(x^{(k)})(x - x^{(k)}) + \frac{1}{2}f''(x^{(k)})(x - x^{(k)})^2. \tag{16}$$

Then, instead of minimizing $f(x)$, we minimize its approximation $q(x)$. The first order necessary condition for a minimizer of $q(x)$ yields

$$0 = q'(x) = f'(x^{(k)}) + f''(x^{(l)})(x - x^{(k)}). \tag{17}$$

Setting $x = x^{(k+1)}$ we obtain

$$x^{(k+1)} = x^{(k)} - \frac{f'(x^{(k)})}{f''(x^{(k)})}. \tag{18}$$

6 Secant Method

Newton's method for minimizing f uses derivatives of f, that is, $f'(x^{(k)})$ and $f''(x^{(k)})$. If the second derivative is not available, we may attempt to approximate it using first derivative information. In particular, we may approximate $f''(x^{(k)})$ with

$$f''(x^{(k)}) \approx \frac{f'(x^{(k)}) - f'(x^{(k-1)})}{x^{(k)} - x^{(k-1)}}. \tag{19}$$

Using the approximation of the second derivative, we obtain the algorithm

$$\begin{aligned} x^{(k+1)} &= x^{(k)} - \frac{x^{(k)} - x^{(k-1)}}{f'(x^{(k)}) - f'(x^{(k-1)})} f'(x^{(k)}) \\ &= \frac{f'(x^{(k)})x^{(k-1)} - f'(x^{(k-1)})x^{(k)}}{f'(x^{(k)}) - f'(x^{(k-1)})}. \end{aligned} \tag{20}$$

7 Bracketing

Many of the methods we have described rely on an initial interval in which the minimizer is known to lie. This interval is also called a bracket, and procedures for finding such a bracket are called bracketing methods.

To find a bracket $[a, b]$ containing the minimizer, assuming unimodality, it suffices to find three points $a < c < b$ such that $f(c) < f(a)$ and $f(c) < f(b)$. A simple bracketing procedure is as follows.

1. First, we pick three arbitrary points $x_0 < x_1 < x_2$. If $f(x_1) < f(x_0)$ and $f(x_1) < f(x_2)$, then we are done–the desired bracket is $[x_0, x_2]$.
2. If not, say $f(x_0) > f(x_1) > f(x_2)$, then we pick a point $x_3 > x_2$ and check if $f(x_2) < f(x_3)$. If it holds, then again we are done–the desired bracket is $[x_1, x_3]$.
3. Otherwise, we continue with this process until the function increases. Typically, each new point chosen involves an expansion in distance between successive test points.

In the procedure described above, when the bracketing process terminates, we have three points x_{k-2}, x_{k-1}, and x_k such that $f(x_{k-1}) < f(x_{k-2})$ and $f(x_{k-1}) < f(x_k)$. The desired bracket is then $[x_{k-2}, x_k]$, which we can then use to initialize any of a number of search methods, including the golden section, Fibonacci, and bisection methods.

8 Line Search in Multidimensional Optimization

One-dimensional search methods play an important role in multidimensional optimization problems. In particular, iterative algorithms for solving such optimization problems typically involve a line search at every iteration.

To be specific, let $f : \mathbb{R}^n \to \mathbb{R}$ be a function that we wish to minimize. Iterative algorithms for finding a minimizer of f are of the form

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha_k \mathbf{d}^{(k)}, \tag{21}$$

where $\mathbf{x}^{(0)}$ is a given initial point and $\alpha_k > 0$ is chosen to minimize

$$\phi_k(\alpha) = f(\mathbf{x}^{(k+1)}) = f(\mathbf{x}^{(k)} + \alpha_k \mathbf{d}^{(k)}). \tag{22}$$

The vector $\mathbf{d}^{(k)}$ is called the search direction and α_k is called the step size. Note that choice of α_k involves a one-dimensional minimization. This choice ensures that under appropriate conditions,

$$f(\mathbf{x}^{(k+1)}) < f(\mathbf{x}^{(k)}). \tag{23}$$

9 Gradient Methods

A level set of a function $f : \mathbb{R}^n \to \mathbb{R}$ is the set of points $\mathbf{x}$ satisfying $f(\mathbf{x}) = c$ for some constant c. Thus, a point $\mathbf{x}_0 \in \mathbb{R}^n$ is on the level set corresponding to level c if $f(\mathbf{x}_0) = c$.

The gradient of f at $\mathbf{x}_0$, denoted $\nabla f(\mathbf{x}_0)$, if it is not a zero vector, is orthogonal to the tangent vector to an arbitrary smooth curve passing through $\mathbf{x}_0$ on the level set $f(\mathbf{x}_0) = c$.

Thus, the direction of maximum rate of increase of a real-valued differentiable function at a point is orthogonal to the level set of the function through that point. In other words, the gradient acts in such a direction that for a given small displacement, the function f increases more in the direction of the gradient than in any other direction.

Recall that $\langle \nabla f(\mathbf{x}), \mathbf{d} \rangle$, $||\mathbf{d}|| = 1$, is the rate of increase of f in the direction $\mathbf{d}$ at the point $\mathbf{x}$. By the Cauchy-Schwarz inequality,

$$\langle \nabla f(\mathbf{x}), \mathbf{d} \rangle \leq ||\nabla f(\mathbf{x})||, \tag{24}$$

because $||\mathbf{d}|| = 1$. But if $\mathbf{d} = \nabla f(\mathbf{x})/||\nabla f(\mathbf{x})||$, then

$$\left\langle \nabla f(\mathbf{x}), \frac{\nabla f(\mathbf{x})}{||\nabla f(\mathbf{x})||} \right\rangle = ||\nabla f(\mathbf{x})||. \tag{25}$$

Thus, the direction in which $\nabla f(\mathbf{x})$ points is the direction of maximum rate of increase of f at $\mathbf{x}$. The direction in which $-\nabla f(\mathbf{x})$ points is the direction of maximum rate of decrease of f at $\mathbf{x}$. Hence, the direction of negative gradient is a good direction to search if we want to find a function minimizer.

Let $\mathbf{x}^0$ be a starting point, and consider the point $\mathbf{x}^0 - \alpha \nabla f(\mathbf{x}^0)$. Then, by Taylor's theorem, we obtain

$$f(\mathbf{x}^0 - \alpha \nabla f(\mathbf{x}^0)) = f(\mathbf{x}^0) - \alpha ||\nabla f(\mathbf{x}^0)||^2 + \cdots . \tag{26}$$

This means that the point $\mathbf{x}^0 - \alpha \nabla f(\mathbf{x}^0)$ is an improvement over the point $\mathbf{x}^0$ if we are searching for a minimizer.

To formulate an algorithm that implements this idea, suppose that we are given a point $\mathbf{x}^k$. To find the next point $\mathbf{x}^{k+1}$, we start at $\mathbf{x}^k$ and move by an amount $-\alpha_k \nabla f(\mathbf{x}^k)$, where α_k is a positive scalar called the *step size*. This procedure leads to the following iterative algorithm:

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \alpha_k \nabla f(\mathbf{x}^k). \tag{27}$$

We refer to this as a gradient descent algorithm (or simply a gradient algorithm).

The gradient varies as the search proceeds, tending to zero as we approach the minimizer. We have the option of either taking very small steps and reevaluating the gradient at every step, or we can take large steps each time. The first approach results in a laborious method of reaching the minimizer, whereas the second approach may result in a more zigzag path to the minimizer. The advantage of the second approach is possibly fewer gradient evaluations. Among many different methods that use this philosophy the most popular is the method of steepest descent.

10 Method of Steepest Descent

The method of steepest descent is a gradient algorithm where the step size α_k is chosen to achieve the maximum amount of decrease of the objective function at each individual step.

Specifically, α_k is chosen to minimize $\phi(\alpha_k) \triangleq f(\mathbf{x}^k - \alpha \nabla f(\mathbf{x}^k))$. In other words,

$$\alpha_k = \underset{\alpha \geq 0}{\operatorname{argmin}}\ f(\mathbf{x}^k - \alpha_k \nabla f(\mathbf{x}^k)). \tag{28}$$

At each step, starting from the point $\mathbf{x}^k$, we conduct a line search in the direction $-\nabla f(\mathbf{x}^k)$ until a minimizer, $\mathbf{x}^{k+1}$ is found.

Proposition 1. If $\{\mathbf{x}^k\}_{k=0}^{\infty}$ is a steepest descent sequence for a given function $f : \mathbb{R}^n \to \mathbb{R}$, then for each k the vector $\mathbf{x}^{k+1} - \mathbf{x}^k$ is orthogonal to the vector $\mathbf{x}^{k+2} - \mathbf{x}^{k+1}$.

Proof. From the iterative formula of the method of steepest descent it follows that

$$\langle \mathbf{x}^{k+1} - \mathbf{x}^k,\ \mathbf{x}^{k+2} - \mathbf{x}^{k+1} \rangle = \alpha_k \alpha_{k+1} \langle \nabla f(\mathbf{x}^k), \nabla f(\mathbf{x}^{k+1}) \rangle. \tag{29}$$

To complete the proof it is enough to show that $\langle \nabla f(\mathbf{x}^k), \nabla f(\mathbf{x}^{k+1}) \rangle = 0$. Observe that α_k is a nonnegative scalar that minimizes $\phi(\alpha_k) \triangleq f(\mathbf{x}^k - \alpha \nabla f(\mathbf{x}^k))$. Hence, using the first order necessary condition (FONC) and the chain rule gives us

$$\begin{aligned} 0 = \phi_k'(\alpha_k) &= \frac{d}{d\alpha} \phi_k(\alpha_k) \\ &= \nabla f(\mathbf{x}^k - \alpha_k \nabla f(\mathbf{x}^k))^\top (-\nabla f(\mathbf{x}^k)) \\ &= -\langle \nabla f(\mathbf{x}^{k+1}), \nabla f(\mathbf{x}^k) \rangle, \end{aligned} \tag{30}$$

which completes the proof. □

The proposition above implies that $\nabla f(\mathbf{x}^k)$ is parallel to the tangent plane to the level set $\{f(\mathbf{x}) = f(\mathbf{x}^{k+1})\}$ at $\mathbf{x}^{k+1}$. Note that as each new point is generated by the steepest descent algorithm, the corresponding value of the function f decreases in value.

Proposition 2. If $\{\mathbf{x}^k\}_{k=0}^{\infty}$ is the steepest descent sequence for $f : \mathbb{R}^n \to \mathbb{R}$ and if $\nabla f(\mathbf{x}^k) \neq 0$, then $f(\mathbf{x}^{k+1}) < f(\mathbf{x}^k)$.

Proof. Recall that

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \alpha_k \nabla f(\mathbf{x}^k), \tag{31}$$

where α_k is the minimizer of

$$\phi_k(\alpha) = f(\mathbf{x}^k - \alpha_k \nabla f(\mathbf{x}^k)), \tag{32}$$

over all $\alpha \geq 0$. Thus, for $\alpha \geq 0$, we have $\phi_k(\alpha_k) \leq \phi_k(\alpha)$. By the chain rule, then

$$\phi_k'(0) = \frac{d\phi_k(0)}{d\alpha} = -(\nabla f(\mathbf{x}^k - 0 \times \nabla f(\mathbf{x}^k)))^\top \nabla f(\mathbf{x}^k) = -||\nabla f(\mathbf{x}^k)||^2 < 0, \tag{33}$$

because $\nabla f(\mathbf{x}^k) \neq 0$ by assumption. Thus, $\phi_k'(0) < 0$ and this implies that there is an $\bar{\alpha} > 0$ such that $\phi_k(0) > \phi_k(\alpha)$ for all $\alpha \in (0, \bar{\alpha}]$. Hence

$$f(\mathbf{x}^{k+1}) = \phi_k(\alpha_k) \leq \phi_k(\bar{\alpha}_k) < \phi_k(0) = f(\mathbf{x}^k), \tag{34}$$

which completes the proof. □

We have proved that the algorithm possesses the descent property: $f(\mathbf{x}^{k+1}) < f(\mathbf{x}^k)$ if $\nabla f(\mathbf{x}^k) \neq 0$. If for some k, we have $\nabla f(\mathbf{x}^k) = 0$, then the point x^k satisfies the FONC. In this case, $\mathbf{x}^{k+1} = \mathbf{x}^k$. We can use the above as the basis for a stopping (termination) criterion for the algorithm.

11 Stopping Criteria

A practical stopping criterion is to check if the norm $||\nabla f(\mathbf{x}^{k+1})||$ of the gradient is less than a prespecified threshold, in which case we stop.

Alternatively, we may compute the absolute difference $|f(\mathbf{x}^{k+1}) - f(\mathbf{x}^k)|$ between objective function values for every two successive iterations, and if the difference is less than some prespecified threshold, then we stop; that is, we stop when $|f(\mathbf{x}^{k+1}) - f(\mathbf{x}^k)| < \epsilon$.

Yet another alternative is to compute the norm $||\mathbf{x}^{k+1} - \mathbf{x}^k||$ of the difference between two successive iterates, and we stop if the norm is less than a prespecified threshold: $||\mathbf{x}^{k+1} - \mathbf{x}^k|| < \epsilon$.

Alternatively, we may check 'relative' values of the quantities above; for example,

$$\frac{|f(\mathbf{x}^{k+1}) - f(\mathbf{x}^k)|}{|f(\mathbf{x}^k)|} < \epsilon, \text{ or } \frac{|\mathbf{x}^{k+1} - \mathbf{x}^k|}{|\mathbf{x}^k|} < \epsilon \tag{35}$$

The two 'relative' stopping criteria above are preferable to the previous 'absolute' criteria because the relative criteria are scale-independent.

Example 1. Apply the Method of Steepest Descent to the function $f(x, y) = 4x^2 - 4xy + 2y^2$, with initial point $x^0 = [2, 3]^\top$. We first compute the steepest descent direction from

$$\nabla f(x, y) = [8x - 4y, 4y - 4x]^\top, \quad \nabla f(\mathbf{x}^0) = \nabla f(2, 3) = [4, 4]^\top. \tag{36}$$

We then minimize the function $\phi(\alpha) = f((2, 3) - \alpha(4, 4)) = f(2 - 4\alpha, 3 - 4\alpha)$ by setting $x \leftarrow 2 - 4\alpha$, $y \leftarrow 3 - 4\alpha$ in $[8x - 4y, 4y - 4x]$, then

$$\phi'(\alpha) = -\nabla f(2 - 4\alpha, 3 - 4\alpha) \begin{bmatrix} 4 \\ 4 \end{bmatrix} = 64\alpha - 32 = 0, \rightarrow \alpha = 0.5. \tag{37}$$

Then

$$\mathbf{x}^1 = \mathbf{x}^0 - 0.5\nabla f(\mathbf{x}^0) = [2, 3]^\top - 0.5[4, 4]^\top = [0, 1]^\top. \tag{38}$$

Continuing the process, we have

$$\nabla f(\mathbf{x}^1) = \nabla f([0, 1]^\top) = [-4, 4]^\top, \tag{39}$$

and by defining

$$\phi(\alpha) = f([0,1]^\top - \alpha[-4,4]^\top) = f([4\alpha, 1-4\alpha]^\top) \tag{40}$$

we obtain

$$\phi'(\alpha) = -[8(4\alpha) - 4(1-4\alpha), 4(1-4\alpha) - 4(4\alpha)][-4,4]^\top = 320\alpha - 32 \rightarrow \alpha = 0.1. \tag{41}$$

We therefore set

$$\mathbf{x}^2 = \mathbf{x}^1 - 0.1\nabla f(\mathbf{x}^1) = [0,1]^\top - 0.1[-4,4]^\top = [0.4, 0.6]^\top. \tag{42}$$

Repeating this process, gives $\mathbf{x}^* = [0,0]^\top$.

12 Quadratic Functions

Consider a quadratic function of the form

$$f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top \mathbf{Q}\mathbf{x} - \mathbf{b}^\top \mathbf{x}, \tag{43}$$

where $\mathbf{Q} \in \mathbb{R}^{n\times n}$ is a symmetric positive definite matrix, $\mathbf{b} \in \mathbb{R}^n$, and $\mathbf{x} \in \mathbb{R}^n$. The unique minimizer of f can be found by setting the gradient of f to zero, where

$$\nabla f(\mathbf{x}) = \mathbf{Q}\mathbf{x} - \mathbf{b}, \tag{44}$$

There is no loss of generality in assuming $\mathbf{Q}$ to be a symmetric matrix. For if we are given a quadratic form $\mathbf{x}^\top \mathbf{A}\mathbf{x}$ and $\mathbf{A} \neq \mathbf{A}^\top$, then because the transposition of a scalar equals itself, we obtain $(\mathbf{x}^\top \mathbf{A}\mathbf{x})^\top = \mathbf{x}^\top \mathbf{A}^\top \mathbf{x} = \mathbf{x}^\top \mathbf{A}\mathbf{x}$. Hence,

$$\mathbf{x}^\top \mathbf{A}\mathbf{x} = \frac{1}{2}\mathbf{x}^\top \mathbf{A}\mathbf{x} + \frac{1}{2}\mathbf{x}^\top \mathbf{A}^\top \mathbf{x} = \frac{1}{2}\mathbf{x}^\top (\mathbf{A} + \mathbf{A}^\top)\mathbf{x} \triangleq \frac{1}{2}\mathbf{x}^\top \mathbf{Q}\mathbf{x}. \tag{45}$$

The Hessian of f is $\mathbf{F}(\mathbf{x}) = \mathbf{Q} = \mathbf{Q}^\top > 0$. To simplify the notation we write $\mathbf{g}^k = \nabla f(x^k)$. Then, the steepest descent algorithm for the quadratic function can be represented as

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \alpha_k \mathbf{g}^k, \tag{46}$$

where $\alpha_k = \underset{\alpha \geq 0}{\operatorname{argmin}}\ f(\mathbf{x}^k - \alpha \mathbf{g}^k)$.

Assume that $\mathbf{g}^k \neq 0$. Because $\alpha_k \geq 0$ is a minimizer of $\phi_k(\alpha) = f(\mathbf{x}^k - \alpha \mathbf{g}^k)$, we apply the FONC to $\phi_k(\alpha)$ to obtain

$$\phi'(\alpha_k) = (\mathbf{x}^k - \alpha \mathbf{g}^k)^\top \mathbf{Q}(-\mathbf{g}^k) - \mathbf{b}^\top (-\mathbf{g}^k). \tag{47}$$

Therefore, $\phi'(\alpha_k) = 0$ if $\alpha \mathbf{g}^{kT}\mathbf{Q}\mathbf{g}^k = (\mathbf{x}^{kT}\mathbf{Q} - \mathbf{b}^\top)\mathbf{g}^k$. But $\mathbf{x}^{kT}\mathbf{Q} - \mathbf{b}^\top = \mathbf{g}^{kT}$, hence

$$\alpha_k = \frac{\mathbf{g}^{kT}\mathbf{g}^k}{\mathbf{g}^{kT}\mathbf{Q}\mathbf{g}^k}. \tag{48}$$

In summary, the method of steepest descent for the quadratic takes the form

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \frac{\mathbf{g}^{kT}\mathbf{g}^k}{\mathbf{g}^{kT}\mathbf{Q}\mathbf{g}^k}\mathbf{g}^k, \tag{49}$$

where $\mathbf{g}^k = \nabla f(\mathbf{x}^k) = \mathbf{Q}\mathbf{x}^k - \mathbf{b}$.