Steepest Descent Method
Kefu Liu

Properties of Gradient Vector

The gradient vector of a scalar function $f(x_1, x_2, \cdots, x_n)$ is defined as a column vector

$$\nabla f = \begin{bmatrix} \dfrac{\partial f}{\partial x_1} & \dfrac{\partial f}{\partial x_2} & \cdots & \dfrac{\partial f}{\partial x_n} \end{bmatrix}^T = \mathbf{c}$$

For example

$$f(x_1, x_2) = 25x_1^2 + x_2^2$$

at the point $x_1^* = .6, x_2^* = 4$

$$\mathbf{c} = \nabla f = \begin{bmatrix} 2(25)x_1^* \\ 2x_2^* \end{bmatrix} = \begin{bmatrix} 2(25)(.6) \\ 2(4) \end{bmatrix} = \begin{bmatrix} 30 \\ 8 \end{bmatrix}$$

The normalized gradient vector

$$\overline{\mathbf{c}} = \frac{\mathbf{c}}{\sqrt{\mathbf{c}^T \mathbf{c}}}$$

For example, at the point $x_1^* = .6, x_2^* = 4$

$$\overline{\mathbf{c}} = \frac{1}{\sqrt{30^2 + 8^2}} \begin{bmatrix} 30 \\ 8 \end{bmatrix} = \begin{bmatrix} .96625 \\ .2577 \end{bmatrix}$$

Property 1. The gradient vector represents a direction of maximum rate of increase for the function $f(\mathbf{x})$ at $\mathbf{x}^*$. For example,

$$f(.6, 4) = 25(.6)^2 + 4^2 = 25$$

If we increase $\mathbf{x}$ in the direction $\overline{\mathbf{c}}$ by a step size of $\alpha = .5$

$$\mathbf{x}^{(1)} = \mathbf{x}^{(0)} + \alpha \overline{\mathbf{c}} = \begin{bmatrix} .6 \\ 4 \end{bmatrix} + .5 \begin{bmatrix} .96625 \\ .2577 \end{bmatrix} = \begin{bmatrix} 1.083125 \\ 4.12885 \end{bmatrix}$$

The function value becomes

$$f(\mathbf{x}^{(1)}) = 25(1.083125)^2 + (4.122885)^2 = 46.327$$

If we move in a direction $\begin{bmatrix} 1 & 0 \end{bmatrix}^T$

$$\mathbf{x}^{(1)} = \mathbf{x}^{(0)} + \alpha\bar{\mathbf{c}} = \begin{bmatrix} .6 \\ 4 \end{bmatrix} + .5\begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 1.1 \\ 4 \end{bmatrix}$$

The function value becomes

$$f(\mathbf{x}^{(1)}) = 25(1.1)^2 + (4)^2 = 46.25$$

If we move in a direction $\begin{bmatrix} 0 & 1 \end{bmatrix}^T$

$$\mathbf{x}^{(1)} = \mathbf{x}^{(0)} + \alpha\bar{\mathbf{c}} = \begin{bmatrix} .6 \\ 4 \end{bmatrix} + .5\begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} .6 \\ 4.5 \end{bmatrix}$$

The function value becomes

$$f(\mathbf{x}^{(1)}) = 25(.6)^2 + (4.5)^2 = 29.25$$

We can see that moving along the gradient direction results in the maximum increase in the function.

Property 2. The gradient vector $\mathbf{c}$ of $f(x_1, x_2, \cdots, x_n)$ at the point $\mathbf{x}^*$ is orthogonal (normal) to the tangent plane for the surface $f(\mathbf{x}^*) = \text{constant}$. For example, $f(x_1, x_2) = 25x_1^2 + x_2^2 = 25$ the slope at $x_1^* = .6, x_2^* = 4$ can be found to be

$$2(25)x_1 dx_1 + 2x_2 dx_2 = 0$$

$$\text{slope} = \frac{dx_2}{dx_1} = -\frac{25x_1}{x_2} = -\frac{25(.6)}{4} = -3.75$$

The direction of the tangent line is given by

$$\mathbf{t} = \begin{bmatrix} 1 \\ -3.75 \end{bmatrix}$$

$\mathbf{c}$ and $\mathbf{t}$ are normal each other as

$$\mathbf{c}^T\mathbf{t} = \begin{bmatrix} 30 & 8 \end{bmatrix}\begin{bmatrix} 1 \\ -3.75 \end{bmatrix} = 30 - 8(3.75) = 0$$

Property 3. The maximum rate of change of $f(\mathbf{x})$ at any point $\mathbf{x}^*$ is the magnitude of the gradient vector given by

$$\|\mathbf{c}\| = \sqrt{\mathbf{c}^T\mathbf{c}}$$

Steepest descent direction. Let $f(\mathbf{x})$ be a differentiable function with respect to $\mathbf{x}$. The direction of steepest descent for $f(\mathbf{x})$ at any point is

$$\mathbf{d} = -\mathbf{c} \text{ or } \overline{\mathbf{d}} = -\overline{\mathbf{c}}$$

Example. Use the steepest descent direction to search for the minimum for $f(x_1, x_2) = 25x_1^2 + x_2^2$ starting at $\mathbf{x}^{(0)} = \begin{bmatrix} 1 & 3 \end{bmatrix}^T$ with a step size of $\alpha = .5$. The function value at the starting point is

$$f(\mathbf{x}^{(0)}) = 25(1)^2 + 3^2 = 34$$

An analytical solution revealsthat the minimum point is at $\mathbf{x}^* = \begin{bmatrix} 0 & 0 \end{bmatrix}^T$ and $f(\mathbf{x}^*) = 0$. Let us start the process of iterations.

$$\mathbf{c}^{(0)} = \begin{bmatrix} 2(25)(1) \\ 2(3) \end{bmatrix} = \begin{bmatrix} 50 \\ 6 \end{bmatrix}, \ \overline{\mathbf{c}}^{(0)} = \frac{1}{\sqrt{50^2 + 6^2}}\begin{bmatrix} 50 \\ 6 \end{bmatrix} = \begin{bmatrix} .9929 \\ .1191 \end{bmatrix}$$

$$\mathbf{x}^{(1)} = \mathbf{x}^{(0)} - .5\overline{\mathbf{c}}^{(0)} = \begin{bmatrix} 1 \\ 3 \end{bmatrix} - .5\begin{bmatrix} .9929 \\ .1191 \end{bmatrix} = \begin{bmatrix} .5035 \\ 2.9404 \end{bmatrix}$$

$$f(\mathbf{x}^{(1)}) = 25(.5035)^2 + (2.9404)^2 = 14.984$$

$$\mathbf{c}^{(1)} = \begin{bmatrix} 2(25)(.5035) \\ 2(2.9404) \end{bmatrix} = \begin{bmatrix} 25.175 \\ 5.8808 \end{bmatrix}, \ \overline{\mathbf{c}}^{(1)} = \begin{bmatrix} .9738 \\ .2275 \end{bmatrix}$$

$$\mathbf{x}^{(2)} = \mathbf{x}^{(1)} - .5\overline{\mathbf{c}}^{(1)} = \begin{bmatrix} .5035 \\ 2.9404 \end{bmatrix} - .5\begin{bmatrix} .9738 \\ .2275 \end{bmatrix} = \begin{bmatrix} .0166 \\ 2.8267 \end{bmatrix}$$

$$f(\mathbf{x}^{(2)}) = 25(.0166)^2 + (2.8267)^2 = 7.997$$

$$\mathbf{c}^{(2)} = \begin{bmatrix} .83 \\ 5.6534 \end{bmatrix}, \ \overline{\mathbf{c}}^{(2)} = \begin{bmatrix} .1453 \\ .9894 \end{bmatrix}$$

$$\mathbf{x}^{(3)} = \mathbf{x}^{(2)} - .5\overline{\mathbf{c}}^{(2)} = \begin{bmatrix} .0166 \\ 2.8267 \end{bmatrix} - .5\begin{bmatrix} .1453 \\ .9894 \end{bmatrix} = \begin{bmatrix} -.0561 \\ 2.332 \end{bmatrix}$$

$$f(\mathbf{x}^{(3)}) = 25(-.0561)^2 + (2.332)^2 = 5.5169$$

$$\mathbf{c}^{(3)} = \begin{bmatrix} -2.805 \\ 4.664 \end{bmatrix}, \ \overline{\mathbf{c}}^{(3)} = \begin{bmatrix} -.5154 \\ .8570 \end{bmatrix}$$

$$\mathbf{x}^{(4)} = \mathbf{x}^{(3)} - .5\overline{\mathbf{c}}^{(3)} = \begin{bmatrix} -.0561 \\ 2.332 \end{bmatrix} - .5\begin{bmatrix} -.5154 \\ .857 \end{bmatrix} = \begin{bmatrix} .2016 \\ 1.9035 \end{bmatrix}$$

$$f(\mathbf{x}^{(4)}) = 25(.2016)^2 + (1.9035)^2 = 4.6394$$

$$\mathbf{c}^{(4)} = \begin{bmatrix} 10.08 \\ 3.807 \end{bmatrix}, \ \overline{\mathbf{c}}^{(4)} = \begin{bmatrix} .9355 \\ .3533 \end{bmatrix}$$

$$\mathbf{x}^{(5)} = \mathbf{x}^{(4)} - .5\overline{\mathbf{c}}^{(4)} = \begin{bmatrix} .2016 \\ 1.9035 \end{bmatrix} - .5\begin{bmatrix} .9355 \\ .3533 \end{bmatrix} = \begin{bmatrix} -.2662 \\ 1.7269 \end{bmatrix}$$

$$f(\mathbf{x}^{(5)}) = 25(-.2662)^2 + (1.7269)^2 = 4.7537$$

It is noted that the function values start to oscillate, i.e., not monotonically reduce. This is caused by the constant step size. When the search is near the minimum, a smaller step size should be used. Otherwise, an “overshoot” will occur. Overshoot means that we move along the steepest direction more than needed. As a matter of fact, we are supposed to find the best step size at each iteration by conducting a one-D optimization in the steepest descent direction. For example, the new point can be expressed as a function of step size α, i.e.,

$$\mathbf{x}^{(1)} = \mathbf{x}^{(0)} - \alpha\overline{\mathbf{c}}^{(0)} = \begin{bmatrix} 1 \\ 3 \end{bmatrix} - \alpha\begin{bmatrix} .9929 \\ .1191 \end{bmatrix} = \begin{bmatrix} 1-.9929\alpha \\ 3-.1191\alpha \end{bmatrix}$$

$$f(\mathbf{x}^{(1)}) = 25(1-.9929\alpha)^2 + (3-.1191\alpha)^2$$

$f(\mathbf{x}^{(1)})$ is a function of α. Using the analytical approach, we get

$$\frac{df(\alpha)}{d\alpha} = 2(25)(1-.9929\alpha^{(0)})(-.9929) + 2(3-.1191\alpha^{(0)})(-.1191) = 0$$

$$\alpha^{(0)} = \frac{25(.9929) + 3(.1191)}{25(.9929^2) + .1191^2} = 1.0211$$

$$\mathbf{x}^{(1)} = \mathbf{x}^{(0)} - \alpha^{(0)}\overline{\mathbf{c}}^{(0)} = \begin{bmatrix} 1 \\ 3 \end{bmatrix} - 1.0211 \begin{bmatrix} .9929 \\ .1191 \end{bmatrix} = \begin{bmatrix} \text{-}.0139 \\ 2.8784 \end{bmatrix}$$

$$f(\mathbf{x}^{(1)}) = 25(-.0139)^2 + 2.8784^2 = 8.29$$

$f = 8.29$ is the minimum value we can find at this iteration.

Steepest descent algorithm

Step 1. Estimate a starting design $\mathbf{x}^{(0)}$ and set the iteration counter $k = 0$. Select a convergence parameter $\varepsilon > 0$.

Step 2. Calculate the gradient of $f(\mathbf{x})$ at the point $\mathbf{x}^{(k)}$ as $\mathbf{c}^{(k)} = \nabla f(\mathbf{x}^{(k)})$. Calculate $\|\mathbf{c}\| = \sqrt{\mathbf{c}^T\mathbf{c}}$. If $\|\mathbf{c}\| < \varepsilon$, then stop the iteration process as $\mathbf{x}^* = \mathbf{x}^{(k)}$ is a minimum point. Otherwise, go to Step 3.

Step 3. Let the search direction at the current point $\mathbf{x}^{(k)}$ as $\mathbf{d}^{(k)} = -\mathbf{c}^{(k)}$.

Step 4. Calculate a step size $\alpha^{(k)}$ to minimize $f(\mathbf{x}^{(k)} + \alpha^{(k)}\mathbf{d}^{(k)})$. A one-dimensional search is used to determine $\alpha^{(k)}$.

Step 5. Update the design as $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha^{(k)}\mathbf{d}^{(k)}$. Set $k = k + 1$ and go to Step 2.