Numerical Optimization HW1

Sheng-Yen Chou

November 11, 2021

1 Problem 1.

Consider a function $f(x_1, x_2) = x_1^3 x_2 - 2x_1 x_2^2 + x_1 x_2^3$

1.1 (a)

Compute the gradient and Hessian of f.

The gradient

$$\nabla_x f = [\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}]$$

$$= \begin{bmatrix} 3x_2 x_1^2 - 2x_2^2 + x_2^3 \\ x_1^3 - 4x_1 x_2 + 3x_1 x_2^2 \end{bmatrix}$$

The Hessian

$$H_f = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 x_2} \\ \frac{\partial^2 f}{\partial x_1 x_2} & \frac{\partial f}{\partial x_2^2} \end{bmatrix}$$

$$= \begin{bmatrix} 6x_2 x_1 & 3x_1^2 - 4x_2 + 3x_2^2 \\ 3x_1^2 - 4x_2 + 3x_2^2 & -4x_1 + 6x_1 x_2 \end{bmatrix}$$

1.2 (b)

Gradient at $(1, 1)$

$$\nabla_x f(1, 1) = [3 \times 1 \times 1 - 2 \times 1 + 1, 1 - 4 \times 1 + 3 \times 1] = [2, 0]$$

Hessian at $(1, 1)$

$$H_f(1, 1) = \begin{bmatrix} 6 & 3 - 4 + 3 \\ 3 - 4 + 3 & -4 + 6 \end{bmatrix} = \begin{bmatrix} 6 & 2 \\ 2 & 2 \end{bmatrix}$$

1.3 (c)

The Taylor expansion of a function $f : \mathbb{R}^n \to \mathbb{R}^m$.

$$f(x + \Delta x) = f(x) + \nabla_x f(x)^T \Delta x$$

The steepest descent direction is the negative gradient since it can make the $f(x + \Delta x)$ decrease most. The detail is

$$\min_{\Delta x} f(x + \Delta x) \quad \text{subject to} \quad ||\Delta x|| = 1$$

$$= \min_{\Delta x} f(x) + \nabla_x f(x)^T \Delta x$$

Since $\nabla_x f(x)$ aren't related to Δx, so we can shorten the objective function

$$\min_{\Delta x} \nabla_x f(x)^T \Delta x = \min_{\Delta x} \langle \nabla_x f(x), \Delta x \rangle$$

We know that the inner product will reach its minimal while the 2 vector are opposite $-\nabla_x f(x)$ and we can normalize it $-\frac{\nabla_x f(x)}{||\nabla_x f(x)||_2}$. Let $\Delta x = -\frac{\nabla_x f(x)}{||\nabla_x f(x)||_2}$, thus,

$$f(x + \Delta x) = f(x) - \nabla_x f(x)^T \frac{\nabla_x f(x)}{||\nabla_x f(x)||_2}$$

And $-\frac{\nabla_x f(x)}{||\nabla_x f(x)||_2}$ is called the **steepest descent direction**.

$$-\frac{\nabla_x f(1,2)}{||\nabla_x f(1,2)||_2} = -\frac{[3 \times 2 \times 1 - 2 \times 2^2 + 2^3, 1^3 - 4 \times 1 \times 2 + 3 \times 1 \times 2^2]}{\sqrt{61}}$$

$$= -\frac{[6,5]}{\sqrt{61}}$$

1.4 (d)

A quadratic model $f : \mathbb{R}^N \to \mathbb{R}$ can be written as

$$f(x) = x^T A x + B^T x + C$$

where $x \in \mathbb{R}^N$, $A \in \mathbb{R}^{N \times N}$, $B \in \mathbb{R}^N$, and $C \in \mathbb{R}$.
Derive the gradient

$$g(x) = \frac{\partial f(x)}{\partial x} = Ax + B$$

$$H(x) = \frac{\partial^2 f(x)}{\partial x^2} = 2A = H(0)$$

Denote $H(0)$ as H_0 and $g(0)$ as g_0, Thus, we can rewrite the formula

$$f(x) = \frac{1}{2} x^T H_0 x + g_0^T x + f(0)$$

$$g(x) = H_0 x + g_0$$

If we want to take a step further in the quadratic model $f(x_k + \alpha p_k)$

$$f(x_k + \alpha p_k) = \frac{1}{2}(x_k + \alpha p_k)^T H_0 (x_k + \alpha p_k) + g_0^T (x_k + \alpha p_k) + f(0)$$

$$= \frac{1}{2}(x_k + \alpha p_k)^T H_0 (x_k + \alpha p_k) + g_0^T (x_k + \alpha p_k) + f(0)$$

$$= \frac{1}{2} x_k^T H_0 x_k + \frac{1}{2}\alpha^2 p_k^T H_0 p_k + \frac{1}{2}\alpha x_k^T H_0 p_k + \frac{1}{2}\alpha p_k^T H_0 x_k + g_0^T (x_k + \alpha p_k) + f(0)$$

$$= \frac{1}{2} x_k^T H_0 x_k + \frac{1}{2}\alpha^2 p_k^T H_0 p_k + \frac{1}{2}\alpha (p_k^T H_0 x_k)^T + \frac{1}{2}\alpha p_k^T H_0 x_k + g_0^T (x_k + \alpha p_k) + f(0)$$

Since $p_k^T H_0 x_k$ is a scalar, it's equal to its transpose $(H_0 x_k)^T p_k$

$$= \frac{1}{2} x_k^T H_0 x_k + \frac{1}{2}\alpha^2 p_k^T H_0 p_k + \frac{1}{2}\alpha (H_0 x_k)^T p_k + \frac{1}{2}\alpha (H_0 x_k)^T p_k + g_0^T x_k + \alpha g_0^T p_k + f(0)$$

$$= \frac{1}{2} x_k^T H_0 x_k + g_0^T x_k + f(0) + \alpha (H_0 x_k)^T p_k + \alpha g_0^T p_k + \frac{1}{2}\alpha^2 p_k^T H_0 p_k$$

$$= \frac{1}{2} x_k^T H_0 x_k + g_0^T x_k + f(0) + \alpha ((H_0 x_k) + g_0)^T p_k + \frac{1}{2}\alpha^2 p_k^T H_0 p_k$$

Since we've known that $f(x) = \frac{1}{2}x^T H_0 x + g_0^T x + f(0)$ and $g(x) = H_0 x + g_0$

$$= f(x_k) + \alpha g(x_k)^T p_k + \frac{1}{2}\alpha^2 p_k^T H_0 p_k$$

Derive the gradient of the quadratic model $\nabla_\alpha f(x_k + \alpha p_k)$ over α

$$\nabla_\alpha f(x_k + \alpha p_k) = g(x_k)^T p_k + \alpha p_k^T H_0 p_k$$

Step Length of The Steepest Descent Method
The step length of the steepest descent method

$$\nabla_\alpha f(x_k + \alpha p_k) = g(x_k)^T p_k + \alpha p_k^T H_0 p_k = 0$$

$$\alpha p_k^T H_0 p_k = -g(x_k)^T p_k$$

$$\alpha = -\frac{g(x_k)^T p_k}{p_k^T H_0 p_k}$$

Newton's Direction
The Newton's method use quadratic model to compute Newton's direction

$$f(x_k + p_k) = f(x_k) + g(x_k)^T p_k + \frac{1}{2} p_k^T H_0 p_k$$

Derive the gradient

$$\nabla_{p_k} f(x_k + p_k) = g(x_k) + H_0 p_k$$

Thus, the Newton's direction is

$$\nabla_{p_k} f(x_k + p_k) = g(x_k) + H_0 p_k = 0$$

$$H_0 p_k = -g(x_k)$$

$$p_k = -H_0^{-1} g(x_k)$$

Where $p_k = -H_0^{-1} g(x_k)$ is Newton's direction
As a result, the Newton's direction at $(1, 2)$ is

$$p_k = -H(1,2)^{-1} g_0(1,2)$$

$$= -\begin{bmatrix} 6*2*1 & 3-4*2+3*2^2 \\ 3*1-4*2+3*2^2 & -4*1+6*1*2 \end{bmatrix}^{-1} \begin{bmatrix} 3*2*1^2-2*2^2+2^3 \\ 1^3-4*1*2+3*1*2^2 \end{bmatrix}$$

$$= -\begin{bmatrix} 12 & 7 \\ 7 & 8 \end{bmatrix}^{-1} \begin{bmatrix} 6 \\ 5 \end{bmatrix}$$

$$= -\begin{bmatrix} \frac{8}{47} & -\frac{7}{47} \\ -\frac{7}{47} & \frac{12}{47} \end{bmatrix} \begin{bmatrix} 6 \\ 5 \end{bmatrix}$$

$$= -\begin{bmatrix} \frac{13}{47} \\ \frac{18}{47} \end{bmatrix}$$

1.5 (e)

The Hessian matrix H_f is

$$H_f(x_1, x_2) = \begin{bmatrix} 6x_2x_1 & 3x_1^2 - 4x_2 + 3x_2^2 \\ 3x_1^2 - 4x_2 + 3x_2^2 & -4x_1 + 6x_1x_2 \end{bmatrix}$$

$$H_f(1, 2) = \begin{bmatrix} 6 * 2 * 1 & 3 * 1^2 - 4 * 2 + 3 * 2^2 \\ 3 * 1^2 - 4 * 2 + 3 * 2^2 & -4 * 1 + 6 * 1 * 2 \end{bmatrix}$$

$$= \begin{bmatrix} 12 & 7 \\ 7 & 8 \end{bmatrix}$$

First of all, we need to compute the LU decomposition for $H_f(1, 2)$. Let $E_1 = \begin{bmatrix} 1 & 0 \\ -\frac{7}{12} & 1 \end{bmatrix}$, thus

$$U = E_1 H_f(1, 2) = \begin{bmatrix} 12 & 7 \\ 0 & -\frac{7}{12} * 7 + 1 * 8 \end{bmatrix}$$

$$= \begin{bmatrix} 12 & 7 \\ 0 & \frac{47}{12} \end{bmatrix}$$

Thus, $H_f(1, 2) = E_1^{-1}U$ and $L = E_1^{-1} = \begin{bmatrix} 1 & 0 \\ \frac{7}{12} & 1 \end{bmatrix}$. We get the LU decomposition

$$H_f(1, 2) = LU \quad L = \begin{bmatrix} 1 & 0 \\ \frac{7}{12} & 1 \end{bmatrix} \quad U = \begin{bmatrix} 12 & 7 \\ 0 & \frac{47}{12} \end{bmatrix}$$

We can move further to LDL decomposition

$$H_f(1, 2) = LDL^T \quad U = DL^T$$

Where D is a diagonal matrix and we can compute

$$D = \begin{bmatrix} 12 & 0 \\ 0 & \frac{47}{12} \end{bmatrix} \quad L^T = \begin{bmatrix} 1 & \frac{7}{12} \\ 0 & 1 \end{bmatrix}$$

Finally, we can get the LDL decomposition

$$H_f(1, 2) = LDL^T \quad D = \begin{bmatrix} 12 & 0 \\ 0 & \frac{47}{12} \end{bmatrix} \quad L = \begin{bmatrix} 1 & 0 \\ \frac{7}{12} & 1 \end{bmatrix}$$

1.6 (f)

The direction p is called descent direction of $f(x)$ at x if its directional derivative $D(f(x), p) < 0$. The directional derivative is defined as

$$D(f(x), p) = \lim_{h \to 0} \frac{f(x + hp) - f(x)}{h}$$

Then expand the directional derivative for 2 dimensional space

$$\begin{aligned} D(f(x), p) &= \lim_{h \to 0} \frac{f(x + hp) - f(x)}{h} \\ &= \lim_{h \to 0} \frac{f(a + hp_x, b + hp_y) - f(x)}{h} \\ &= \lim_{h \to 0} \frac{f(a + hp_x, b + hp_y) - f(a, b + hp_y)}{h} + \frac{f(a, b + hp_y) - f(x)}{h} \\ &= \lim_{h \to 0} p_x \frac{f(a + hp_x, b + hp_y) - f(a, b + hp_y)}{p_x h} + p_y \frac{f(a, b + hp_y) - f(x)}{p_y h} \\ &= p_x \frac{\partial f}{\partial x} + p_y \frac{\partial f}{\partial y} \end{aligned}$$

$$= \langle \begin{bmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \end{bmatrix}, \begin{bmatrix} p_x \\ p_y \end{bmatrix} \rangle$$

For the function $f(x_1, x_2)$

$$\langle \begin{bmatrix} 3x_2x_1^2 - 2x_2^2 + x_2^3 \\ x_1^3 - 4x_1x_2 + 3x_1x_2^2 \end{bmatrix}, p_k \rangle$$

$$= \langle \begin{bmatrix} 3 * 2 * 1^2 - 2 * 2^2 + 2^3 \\ 1^3 - 4 * 1 * 2 + 3 * 1 * 2^2 \end{bmatrix}, \begin{bmatrix} -\frac{13}{47} \\ -\frac{18}{47} \end{bmatrix} \rangle$$

$$= \langle \begin{bmatrix} 6 \\ 5 \end{bmatrix}, \begin{bmatrix} -\frac{13}{47} \\ -\frac{18}{47} \end{bmatrix} \rangle$$

$$= \frac{-78}{47} + \frac{-90}{47}$$

$$= \frac{-168}{47} < 0$$

Thus, the Newton's direction of f at $(x_1, x_2) = (1, 2)$ is a descent direction.

1.7 (g)

We've known the LDL decomposition for the Hessian is $H = LDL^\top$. The Modified $LDL^\top$ decomposition would be $\hat{H} = L\hat{D}L^\top$ and $\hat{H}$ is the modified Hessian matrix and the elements of the diagonal matrix $\hat{D}$ are larger than a certain value ϵ . The Hessian matrix may be less ill condition and positive definite. The inverse matrix of the modified Hessian matrix is $\hat{H}^{-1} = L^{-\top}\hat{D}^{-1}L^{-1}$. Thus, we can derive the modified Newton's direction based on the $LDL^\top$ decomposition $\hat{p} = -g\hat{H}^{-1} = -gL^{-\top}\hat{D}^{-1}L^{-1}$. From subject (e) we've computed the LDL^T decomposition of the Hessian matrix $H_f(1, 2)$.

$$H_f(1, 2) = LDL^\top$$

where $H_f(1, 2) = \begin{bmatrix} 12 & 7 \\ 7 & 8 \end{bmatrix}, L = \begin{bmatrix} 1 & 0 \\ \frac{7}{12} & 1 \end{bmatrix}, D = \begin{bmatrix} 12 & 0 \\ 0 & \frac{47}{12} \end{bmatrix}$ Since the elements of the diagonal matrix D are all larger than 1, we don't need to modify the diagonal elements and $\hat{D} = D$. Thus, the modified Newton's direction is the same as the Newton's direction.

$$\hat{p} = -gL^{-\top}\hat{D}^{-1}L^{-1}$$

$$= -gL^{-\top}\hat{D}^{-1}L^{-1}$$

$$= -\begin{bmatrix} \frac{13}{47} \\ \frac{18}{47} \end{bmatrix}$$

1.8 (h)

Since the computing a Hessian matrix is too expensive, thus we can use first order derivative to approximate the Hessian matrix. It is also called the "secant method". For a Hessian matrix $\hat{H}_k$, it can be approximated by $\hat{H}_k = \frac{\nabla f(x_k) - \nabla f(x_{k-1})}{x_k - x_{k-1}}$. We can rewrite the formula as $\hat{H}_k(x_k - x_{k-1}) = \nabla f(x_k) - \nabla f(x_{k-1})$. As for Quasi-Newton, we use approximated Hessian matrix $\hat{H}_k$ to compute the Newton's direction instead of the original Hessian matrix H_k . Here we use SR1 update to compute the approximated Hessian matrix $\hat{H}_k$. The SR1 can be written as

$$\hat{H}_{k+1} = \hat{H}_k + \sigma_k uu^\top$$

Where $u \in \mathbb{R}^n$ and $\sigma_k \in \mathbb{R}$. Let $y_k = \nabla f(x_k + 1) - \nabla f(x_k)$ and $s_k = x_{k+1} - x_k$. Thus,

$$y_k = \hat{H}_{k+1}s_k$$

$$= (\hat{H}_k + \sigma_k uu^\top)s_k$$

$$= \hat{H}_k s_k + \sigma_k uu^\top s_k$$

Thus,

$$y_k - \hat{H}_k s_k = (\sigma_k u u^\top) s_k$$

Let $u = \delta^2 (y_k - H_k s_k)(y_k - H_k s_k)^\top$. Thus

$$H_{k+1} = H_k + \frac{(y_k - H_k s_k)(y_k - H_k s_k)^\top}{(y_k - \hat{H}_k s_k)^\top s_k}$$

Apply the Sherman-Morrison-Woodbury formula, we can derive the SR1 update as

$$\hat{H}_{k+1}^{-1} = \hat{H}_k^{-1} + \frac{(s_k - \hat{H}_k^{-1} y_k)(s_k - \hat{H}_k^{-1} y_k)^\top}{y_k^\top (s_k - \hat{H}_k^{-1} y_k)}$$

As for direction $p_1 = -\hat{H}_1^{-1} g_1 = -\hat{H}_1^{-1} \nabla_x f(x_1)$, $\hat{H}_0 = I$, $x_0 = \begin{bmatrix} x_{1,0} \\ x_{2,0} \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ and, $x_1 = \begin{bmatrix} x_{1,1} \\ x_{2,1} \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$. First we compute s_k and y_k .

$$s_0 = x_1 - x_0 = \begin{bmatrix} 1 \\ 2 \end{bmatrix} - \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

$$\begin{aligned}
y_0 &= \nabla_x f_1 - \nabla_x f_0 = \nabla_x f(x_1) - \nabla_x f(x_0) \\
&= \begin{bmatrix} 3x_{2,1}x_{1,1}^2 - 2x_{2,1}^2 + x_{2,1}^3 \\ x_{1,1}^3 - 4x_{1,1}x_{2,1} + 3x_{1,1}x_{2,1}^2 \end{bmatrix} - \begin{bmatrix} 3x_{2,0}x_{1,0}^2 - 2x_{2,0}^2 + x_{2,0}^3 \\ x_{1,0}^3 - 4x_{1,0}x_{2,0} + 3x_{1,0}x_{2,0}^2 \end{bmatrix} \\
&= \begin{bmatrix} 3*2*1^2 - 2*2^2 + 2^3 \\ 1^3 - 4*1*2 + 3*1*2^2 \end{bmatrix} - \begin{bmatrix} 3*1*1^2 - 2*1^2 + 1^3 \\ 1^3 - 4*1*1 + 3*1*1^2 \end{bmatrix} \\
&= \begin{bmatrix} 6 \\ 5 \end{bmatrix} - \begin{bmatrix} 2 \\ 0 \end{bmatrix} \\
&= \begin{bmatrix} 4 \\ 5 \end{bmatrix}
\end{aligned}$$

Compute the SR1 update

$$\begin{aligned}
B_1^{-1} &= B_0^{-1} + \frac{(s_0 - \hat{H}_0^{-1} y_0)(s_0 - \hat{H}_0^{-1} y_0)^\top}{y_0^\top (s_0 - \hat{H}_0^{-1} y_0)} \\
&= I + \frac{(\begin{bmatrix} 0 \\ 1 \end{bmatrix} - I \begin{bmatrix} 4 \\ 5 \end{bmatrix})(\begin{bmatrix} 0 \\ 1 \end{bmatrix} - I \begin{bmatrix} 4 \\ 5 \end{bmatrix})^\top}{\begin{bmatrix} 4 & 5 \end{bmatrix} (\begin{bmatrix} 0 \\ 1 \end{bmatrix} - I \begin{bmatrix} 4 \\ 5 \end{bmatrix})} \\
&= I + \frac{(\begin{bmatrix} -4 \\ -4 \end{bmatrix})(\begin{bmatrix} -4 \\ -4 \end{bmatrix})^\top}{\begin{bmatrix} 4 & 5 \end{bmatrix} (\begin{bmatrix} -4 \\ -4 \end{bmatrix})} \\
&= I + \frac{(\begin{bmatrix} 16 & 16 \\ 16 & 16 \end{bmatrix})}{(-36)} \\
&= I + \begin{bmatrix} -\frac{4}{9} & -\frac{4}{9} \\ -\frac{4}{9} & -\frac{4}{9} \end{bmatrix} \\
&= \begin{bmatrix} \frac{5}{9} & -\frac{4}{9} \\ -\frac{4}{9} & \frac{5}{9} \end{bmatrix}
\end{aligned}$$

Thus, we can compute the Newton's direction

$$\begin{aligned}
p_1 &= -\hat{H}_1^{-1} \nabla_x f(x_1) \\
&= -\begin{bmatrix} \frac{5}{9} & -\frac{4}{9} \\ -\frac{4}{9} & \frac{5}{9} \end{bmatrix} \begin{bmatrix} 6 \\ 5 \end{bmatrix} \\
&= \begin{bmatrix} -\frac{10}{9} \\ -\frac{1}{9} \end{bmatrix}
\end{aligned}$$

1.9 (i)

As for direction $p_1 = -\hat{H}_1^{-1} g_1 = -\hat{H}_1^{-1} \nabla_x f(x_1)$, $\hat{H}_0 = I$, $x_0 = \begin{bmatrix} x_{1,0} \\ x_{2,0} \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ and, $x_1 = \begin{bmatrix} x_{1,1} \\ x_{2,1} \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$. First we compute s_k and y_k

$$s_0 = x_1 - x_0 = \begin{bmatrix} 1 \\ 2 \end{bmatrix} - \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

$$y_0 = \nabla_x f_1 - \nabla_x f_0 = \nabla_x f(x_1) - \nabla_x f(x_0) = \begin{bmatrix} 4 \\ 5 \end{bmatrix}$$

The Rank2 Update is

$$\hat{H}_{k+1} = \hat{H}_k - \frac{\hat{H}_k s_k s_k^\top \hat{H}_k}{s_k^\top \hat{H}_k s_k} + \frac{y_k y_k^\top}{y_k^\top s_k}$$

The inverse Hessian matrix approximated by BFGS can be derived as

$$\begin{aligned}
\hat{H}_{k+1}^{-1} &= (I - \rho_k s_k y_k^\top) \hat{H}_k^{-1} (I - \rho y_k s_k^\top) + \rho_k s_k s_k^\top, \text{ where } \rho_k = \frac{1}{y_k^\top s_k} \\
&= (I - \frac{1}{5} \begin{bmatrix} 0 \\ 1 \end{bmatrix} \begin{bmatrix} 4 & 5 \end{bmatrix}) I (I - \frac{1}{5} \begin{bmatrix} 4 \\ 5 \end{bmatrix} \begin{bmatrix} 0 & 1 \end{bmatrix}) + \frac{1}{5} \begin{bmatrix} 0 \\ 1 \end{bmatrix} \begin{bmatrix} 0 & 1 \end{bmatrix} \\
&= (I - \begin{bmatrix} 0 & 0 \\ \frac{4}{5} & 1 \end{bmatrix}) I (I - \begin{bmatrix} 0 & \frac{4}{5} \\ 0 & 1 \end{bmatrix}) + \begin{bmatrix} 0 & 0 \\ 0 & \frac{1}{5} \end{bmatrix} \\
&= \begin{bmatrix} 1 & 0 \\ -\frac{4}{5} & 0 \end{bmatrix} I \begin{bmatrix} 1 & -\frac{4}{5} \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & \frac{1}{5} \end{bmatrix} \\
&= \begin{bmatrix} 1 & -\frac{4}{5} \\ -\frac{4}{5} & \frac{16}{25} \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & \frac{1}{5} \end{bmatrix} \\
&= \begin{bmatrix} 1 & -\frac{4}{5} \\ -\frac{4}{5} & \frac{21}{25} \end{bmatrix}
\end{aligned}$$

$$\rho_k = \frac{1}{\begin{bmatrix} 4 & 5 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix}} = \frac{1}{5}$$

Thus, we can compute the Quasi-Newton direction using BFGS

$$\begin{aligned}
p_1 &= -\hat{H}_1^{-1} \nabla_x f(x_1) \\
&= -\begin{bmatrix} 1 & -\frac{4}{5} \\ -\frac{4}{5} & \frac{21}{25} \end{bmatrix} \begin{bmatrix} 6 \\ 5 \end{bmatrix} \\
&= \begin{bmatrix} -2 \\ \frac{3}{5} \end{bmatrix}
\end{aligned}$$

2 Problem 2.

2.1 (a)

Prove by contradiction

Assume that for a convex set $S \subseteq \mathbb{R}^n$ and a convex function $f : S \to \mathbb{R}^n$, exist a local minimum $\hat{x}$ and a global minimum x^* individually. That is, $\hat{x} \neq x^*$. According to the definition of the local minimum, $\exists \hat{\epsilon} > 0 \ s.t. \ \forall x_1, x_2 \in S \ ||x_1 - \hat{x}||_2 \leq \hat{\epsilon} \ and \ ||x_2 - \hat{x}||_2 \leq \hat{\epsilon}$ Thus, $f(\alpha x_1 + (1-\alpha) x_2) \leq \alpha f(x_1) + (1-\alpha) f(x_2) \ \forall \alpha \in [0,1]$ Similarly, according to the definition of the global minimum, $\exists \epsilon^* > 0 \ s.t. \ \forall x_1, x_2 \in S \ ||x_1 - x^*||_2 \leq \ \epsilon^* \ and \ ||x_2 - x^*||_2 \leq \epsilon^*$ Thus, $f(\beta x_1 + (1-\beta) x_2) \leq \beta f(x_1) + (1-\beta) f(x_2) \ \forall \beta \in [0,1]$ Consider a line cross the local minimum and the global minimum. Let $\hat{x}_1$ be near to the local minimum $\hat{x}$, $||\hat{x}_1 - \hat{x}||_2 \leq \hat{\epsilon}$. Let x_1^* be near to the global minimum x^*, $||x_1^* - x^*||_2 \leq \epsilon^*$. Trivially, $\exists \delta \in [0,1] \ s.t. \ f(\delta \hat{x}_1 + (1 - \delta x_1^*)) \geq \delta f(\hat{x}_1) + (1-\delta) f(x_1^*)$ and it contradicts with the definition of the convex function f which requires $\forall \delta \in [0,1] \ s.t. \ f(\delta \hat{x}_1 + (1 - \delta x_1^*)) \leq \delta f(\hat{x}_1) + (1-\delta) f(x_1^*)$. As a result, the local minimum must be the same as the global minimum.

2.2 (b)

For a function $f : S \to \mathbb{R}^n$ $f(x) = x^\top Qx$, Q is a symmetric positive semi-definite matrix. We want to prove that $S \subseteq \mathbb{R}^n$ is a convex set. For a convex function $h : K \to \mathbb{R}^n$, $\forall a, b \in S, \forall \alpha \in [0, 1]$, $h(\alpha a + (1-\alpha)b) \le \alpha h(a) + (1-\alpha)h(b)$. Thus, we expect f should be $f(\beta x + (1-\beta)y) - \beta f(x) - (1-\beta)f(y) \le 0$, $x, y \in S$, $\beta \in [0, 1]$.

$$
\begin{gathered}
f(\beta x + (1-\beta)y) - \beta f(x) - (1-\beta)f(y) \\
= (\beta x + (1-\beta)y)^\top Q(\beta x + (1-\beta)y) - \beta x^\top Qx - (1-\beta)y^\top Qy \\
= \beta^2 x^\top Qx + (1-\beta)^2 y^\top Qy + \beta(1-\beta)(x^\top Qy + y^\top Qx) - \beta x^\top Qx - (1-\beta)y^\top Qy \\
= \beta(\beta-1)x^\top Qx + \beta(\beta-1)y^\top Qy + \beta(1-\beta)(x^\top Qy + y^\top Qx) \\
= \beta(\beta-1)x^\top Qx + \beta(\beta-1)y^\top Qy - \beta(\beta-1)(x^\top Qy + y^\top Qx) \\
= \beta(\beta-1)x^\top Q(x-y) + \beta(\beta-1)y^\top Q(y-x) \\
= \beta(\beta-1)x^\top Q(x-y) - \beta(\beta-1)y^\top Q(x-y) \\
= \beta(\beta-1)(x-y)^\top Q(x-y)
\end{gathered}
$$

Since Q is semi-definite, $(x-y)^\top Q(x-y) \ge 0$. On the other hand, $\beta - 1 < 0$. As a result, we prove that $\beta(\beta-1)(x-y)^\top Q(x-y) \le 0$. That is $f(x)$ is a convex function.

3 Problem 3

3.1 (a)

Let $\phi(\alpha) = (\alpha-1)^2$ and the sufficient decrease condition is $\phi(\alpha) \le \phi(0) + c_1\alpha\phi'(0)$, $\alpha \in [0, \infty)$. Suppose $c_1 = 0.1$. Derive the derivative of ϕ

$$\phi'(\alpha) = 2(\alpha - 1)$$

Thus

$$
\begin{gathered}
\phi(\alpha) = (\alpha-1)^2 \le (0-1)^2 + 0.1 * \alpha * 2 * (0-1) = \phi'(0) \\
(\alpha-1)^2 \le 1 - 0.2\alpha \\
\alpha^2 - 2\alpha + 1 \le 1 - 0.2\alpha \\
\alpha^2 - 1.8\alpha \le 0 \\
\alpha^2 - 1.8\alpha + 0.81 \le 0.81 \\
(\alpha - 0.9)^2 \le 0.81
\end{gathered}
$$

Thus, the feasible region for α

$$0 \le \alpha \le 1.8$$

3.2 (b)

Let $\phi(\alpha) = (\alpha-1)^2$ and the curvature condition is $\phi'(\alpha) \ge c_2\phi'(0)$, $\alpha \in [0, \infty)$. Suppose $c_2 = 0.9$.

$$
\begin{gathered}
\phi'(\alpha) = 2(\alpha-1) \ge 0.9 * 2(0-1) = c_2\phi'(0) \\
2\alpha - 2 \ge -1.8
\end{gathered}
$$

The feasible region for α is

$$\alpha \ge 0.1$$

4 Problem 4.

4.1 (1)

Show that $\alpha_k = \frac{\vec{p}_k^\top \vec{r}_k}{\vec{p}_k^\top A\vec{p}_k} = \frac{\vec{r}_k^\top \vec{r}_k}{\vec{p}_k^\top A\vec{p}_k}$ To simplify the goal, we know if we can show $\vec{p}_k^\top \vec{r}_k = \vec{r}_k^\top \vec{r}_k$, then $\frac{\vec{p}_k^\top \vec{r}_k}{\vec{p}_k^\top A\vec{p}_k} = \frac{\vec{r}_k^\top \vec{r}_k}{\vec{p}_k^\top A\vec{p}_k}$ will hold.

Prove by induction Basis: According to the step (1), we've known that $\vec{r}_0 = \vec{p}_0$. Thus, $\vec{p}_0^\top \vec{r}_0 = \vec{r}_0^\top \vec{r}_0$. Induction: According to the step (7), we've known that $\vec{p}_k = \vec{r}_k + \beta_{k-1}\vec{p}_{k-1}$. Thus, plugin $\vec{p}_k$ into the formula $\vec{p}_k^\top \vec{r}_k$

$$\vec{p}_k^\top \vec{r}_k = (\vec{r}_k + \beta_{k-1}\vec{p}_{k-1})\vec{r}_k$$

$$= (\vec{r}_k\vec{r}_k) + \beta_{k+1}\vec{p}_{k-1}\vec{r}_k$$

According to the assumption, we can get $\vec{p}_{k-1} = \sum_{i=1}^{k-1} \gamma_i \vec{r}_i$. Thus,

$$= (\vec{r}_k\vec{r}_k) + \beta_{k+1}(\sum_{i=1}^{k-1} \gamma_i \vec{r}_i)\vec{r}_k$$

According to the property (a) $\vec{r}_i^\top \vec{r}_j = 0,\ i \neq j$, we can eliminate $\beta_{k+1}\vec{r}_{k-1}\vec{r}_k = 0$

$$= \vec{r}_k\vec{r}_k$$

Thus, we can argue that $\alpha_k = \frac{\vec{p}_k^\top \vec{r}_k}{\vec{p}_k^\top A\vec{p}_k} = \frac{\vec{r}_k^\top \vec{r}_k}{\vec{p}_k^\top A\vec{p}_k}$ will hold.

4.2 (2)

Show that $\beta_k = \frac{\vec{r}_{k+1}^\top \vec{r}_{k+1}}{\vec{r}_k^\top \vec{r}_k} = -\frac{\vec{p}_k^\top A\vec{r}_{k+1}}{\vec{p}_k^\top A\vec{p}_k}$

$$\beta_k = \frac{\vec{r}_{k+1}^\top \vec{r}_{k+1}}{\vec{r}_k^\top \vec{r}_k}$$

From step (3), we know that $\vec{r}_{k+1} = \vec{r}_k - \alpha_k A\vec{p}_k$. Also, from the previous proof, we've known that $\alpha_k = \frac{\vec{r}_k^\top \vec{r}_k}{\vec{p}_k^\top A\vec{p}_k}$. Thus, we can derive $\vec{r}_k = \vec{r}_{k+1} + \alpha_k A\vec{p}_k$, $\vec{r}_k^\top \vec{r}_k = \alpha_k \vec{p}_k^\top A\vec{p}_k$ and plug into the formula.

$$= \frac{(\vec{r}_k - \alpha_k A\vec{p}_k)^\top \vec{r}_{k+1}}{\alpha_k \vec{p}_k^\top A\vec{p}_k}$$

$$= \frac{(\vec{r}_{k+1}^\top \vec{r}_k) - (\alpha_k \vec{p}_k^\top A\vec{r}_{k+1})}{\alpha_k \vec{p}_k^\top A\vec{p}_k}$$

$$= -\frac{\alpha_k \vec{p}_k^\top A\vec{r}_{k+1}}{\alpha_k \vec{p}_k^\top A\vec{p}_k}$$

$$= -\frac{\vec{p}_k^\top A\vec{r}_{k+1}}{\vec{p}_k^\top A\vec{p}_k}$$

Thus, $\beta_k = \frac{\vec{r}_{k+1}^\top \vec{r}_{k+1}}{\vec{r}_k^\top \vec{r}_k} = -\frac{\vec{p}_k^\top A\vec{r}_{k+1}}{\vec{p}_k^\top A\vec{p}_k}$

5 Reference

References