Ch.11 Quasi-Newton Methods

11.1 Introduction

- Newton's method: for a general nonlinear objective function, convergence to a solution cannot be guaranteed from an arbitrary point $x^{(0)}$.
- If the initial point is not sufficiently close to the solution, then the algorithm may not possess the descent property (i.e., $f(x^{(k+1)}) < f(x^{(k)})$ for some k).
- Newton's recursive algorithm

$$x^{(k+1)} = x^{(k)} - F(x^{(k)})^{-1} g^{(k)}$$

- We modify the original algorithm as follows:

$$x^{(k+1)} = x^{(k)} - \alpha_k F(x^{(k)})^{-1} g^{(k)}$$

  where α_k is chosen to ensure that

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

- A computational drawback of Newton's method is the need to evaluate $F(x^{(k)})$ and solve the equation $F(x^{(k)})d^{(k)} = -g^{(k)}$.
- To avoid the computation of $F(x^{(k)})^{-1}$, the quasi-Newton methods use an approximation to

$F(x^{(k)})^{-1}$ in place of the true inverse.

$$x^{(k+1)} = x^{(k)} - \alpha H_k g^{(k)}$$

where H_k is an $n \times n$ real matrix, and α is a positive search parameter.

$$\begin{aligned} f(x^{(k+1)}) &= f(x^{(k)}) + g^{(k)T}(x^{(k+1)} - x^{(k)}) \\ &\quad + o(\|x^{(k+1)} - x^{(k)}\|) \\ &= f(x^{(k)}) - \alpha g^{(k)T} H_k g^{(k)} + o(\|H_k g^{(k)}\|\alpha) \end{aligned}$$

- To guarantee a decrease in f for small α, we have to have

$$g^{(k)T} H_k g^{(k)} > 0 \Rightarrow H_k \text{ be positive definite}$$

- **Proposition 11.1** Let $f \in C^1$, $x^{(k)} \in \Re^n$, $g^{(k)} = \nabla f(x^{(k)}) \neq 0$, and H_k an $n \times n$ real symmetric positive definite matrix. If we set $x^{(k+1)} = x^{(k)} - \alpha_k H_k g^{(k)}$, where $\alpha_k = \arg\min_{\alpha \geq 0} f(x^{(k)} - \alpha H_k g^{(k)})$, then $\alpha_k > 0$, and $f(x^{(k+1)}) < f(x^{(k)})$.

11.2 Approximating the Inverse Hessian

- Let $H_0.H_1, H_2, \cdots$ be successive approximations of the inverse $F(x^{(k)})^{-1}$ of the Hessian.

$$g^{(k+1)} - g^{(k)} = Q(x^{(k+1)} - x^{(k)})$$

- Let

$$\begin{aligned} \Delta g^{(k)} &\equiv g^{(k+1)} - g^{(k)} \\ \Delta x^{(k)} &\equiv x^{(k+1)} - x^{(k)} \\ \Delta g^{(k)} &= Q\Delta x^{(k)} \\ Q^{-1}\Delta g^{(i)} &= \Delta x^{(i)}, \;\; 0 \le i \le k \end{aligned}$$

- The requirement that the approximation H_{k+1} of the Hessian satisfy

$$H_{k+1}\Delta g^{(i)} = \Delta x^{(i)}, \;\; 0 \le i \le k$$

  Like this way,

  $H_n[\Delta g^{(0)}, \Delta g^{(1)}, \cdots, \Delta g^{(n-1)}] = [\Delta x^{(0)}, \Delta x^{(1)}, \cdots, \Delta x^{(n-1)}]$

- Therefore, if $[\Delta g^{(0)}, \Delta g^{(1)}, \cdots, \Delta g^{(n-1)}]$ is nonsingular, then Q^{-1} is determined uniquely after n steps, via

$$\begin{aligned} Q^{-1} = H_n &= [\Delta x^{(0)}, \Delta x^{(1)}, \cdots, \Delta x^{(n-1)}] \\ &\quad \times [\Delta g^{(0)}, \Delta g^{(1)}, \cdots, \Delta g^{(n-1)}]^{-1} \end{aligned}$$

- Quasi-Newton algorithm have the form

$$\begin{aligned} d^{(k)} &= -H_k g^{(k)} \\ \alpha_k &= \arg\min f(x^{(k)} + \alpha d^{(k)}) \\ x^{(k+1)} &= x^{(k)} + \alpha_k d^{(k)} \end{aligned}$$

  where the matrices $H_0, H_1, \cdots$ are symmetric.

- In the quadratic case, the above matrices are required to satisfy

$$H_{k+1}\Delta g^{(i)} = \Delta x^{(i)}, \quad 0 \le i \le k,$$

where $\Delta x^{(i)} = x^{(i+1)} - x^{(i)} = \alpha_i d^{(i)}$, and $\Delta g^{(i)} = g^{(i+1)} - g^{(i)} = Q\Delta x^{(i)}$.

- **Theorem 11.1** Consider a quasi-Newton algorithm applied to a quadratic function with Hessian $Q = Q^T$, such that for $0 \le k \le n-1$,

$$H_{k+1}\Delta g^{(i)} = \Delta x^{(i)}, \quad 0 \le i \le k$$

where $H_{k+1} = H_{k+1}^T$. If $\alpha \ne 0, 0 \le i \le k+1$, then $d^{(0)}, \cdots, d^{(k+1)}$ are Q-conjugate.

11.3 Rank One Correction Formula

- Update equation on H_k.

$$H_{k+1} = H_k + \alpha_k z^{(k)} z^{(k)T}$$
$$\text{rank } z^{(k)} z^{(k)T} = 1$$

- Our goal is to determine α_k and $z^{(k)}$, given $H_k, \Delta g^{(k)}, \Delta x^{(k)}$, so that $H_{k+1}\Delta g^{(i)} = \Delta x^{(i)}$, $i = 1, \cdots, k$.

- Rank One Algorithm
  1. Set k:=0; select $x^{(0)}$, and a real symmetric positive definite H_0.

2. If $g^{(k)} = 0$, stop; else $d^{(k)} = -H_k g^{(k)}$.
3. Compute

$$\begin{aligned} \alpha_k &= \arg\min_{\alpha \geq 0} f(x^{(k)} + \alpha d^{(k)}) \\ x^{(k+1)} &= x^{(k)} + \alpha_k d^{(k)} \end{aligned}$$

4. Compute

$$\begin{aligned} \Delta x^{(k)} &= \alpha_k d^{(k)} \\ g^{(k+1)} &= Qx^{(k)} \\ \Delta g^{(k)} &= g^{(k+1)} - g^{(k)} \\ H_{k+1} &= H_k + \frac{(\Delta x^{(k)} - H_k \Delta g^{(k)})(\Delta x^{(k)} - H_k \Delta g^{(k)})^T}{\Delta g^{(k)T}(\Delta x^{(k)} - H_k \Delta g^{(k)})} \end{aligned}$$

5. Set $k{:=}k+1$; go to step 2.

- The rank one algorithm is based on satisfying the equation

$$H_{k+1} \Delta g^{(k)} = \Delta x^{(k)}$$

- **Theorem 11.2** For the rank one algorithm applied to the quadratic with Hessian $Q = Q^T$, we have $H_{k+1} \Delta g^{(i)} = \Delta x^{(i)}$, $0 \leq i \leq k$.
- **Ex.11.1** Let

$$f(x_1, x_2) = x_1^2 + \frac{1}{2} x_2^2 + 3$$

Apply the rank one correction algorithm to minimize f.

Use $x^{(0)} = [1, 2]^T$ and $H_0 = I_2$.

$$\begin{aligned} f(x) &= \frac{1}{2} x^T \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix} x + 3 \\ g^{(k)} &= \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix} x^{(k)} \end{aligned}$$

1-st iteration

$$\begin{aligned} d^{(0)} &= -g^{(0)} = [-2, -2]^T \\ \alpha_0 &= \arg\min_{\alpha \geq 0} f(x^{(0)} + \alpha d^{(0)}) \\ &= -\frac{g^{(0)T} d^{(0)}}{d^{(0)T} Q d^{(0)}} = \frac{2}{3} \\ x^{(1)} &= x^{(0)} + \alpha_0 d^{(0)} = [-1/3, 2/3]^T \\ \Delta x^{(0)} &= \alpha_0 d^{(0)} = [-4/3, -4/3]^T \\ g^{(1)} &= Q x^{(1)} = [-2/3, 2/3]^T \\ \Delta g^{(0)} &= g^{(1)} - g^{(0)} = [-8/3, -4/3]^T \end{aligned}$$

$$\Delta g^{(0)T} (\Delta x^{(0)} - H_0 \Delta g^{(0)}) = [-8/3, -4/3] \begin{bmatrix} 4/3 \\ 0 \end{bmatrix}$$

$$H_1 = H_0 + \frac{(\Delta x^{(0)} - H_0 \Delta g^{(0)})(\Delta x^{(0)} - H_0 \Delta g^{(0)})^T}{\Delta g^{(0)T} (\Delta x^{(0)} - H_0 \Delta g^{(0)})}$$

$$= \begin{bmatrix} 1/2 & 0 \\ 0 & 1 \end{bmatrix}$$

$$
\begin{aligned}
d^{(1)} &= -H_1 g^{(1)} = [1/3, -2/3]^T \\
\alpha_1 &= -\frac{g^{(1)T} d^{(1)}}{d^{(1)T} Q d^{(1)}} = 1 \\
x^{(2)} &= x^{(1)} + \alpha_1 d^{(1)} = [0, 0]^T
\end{aligned}
$$

Note that $g^{(2)} = 0$, and therefore $x^{(2)} = x^*$. The algorithm solves the problem in two steps. The direction $d^{(0)}$ and $d^{(1)}$ are Q-conjugate.

- For nonquadratic case, the rank one correction algorithm is not satisfactory. (d^{k+1} may not be a descent direction).

11.4 DFP Algorithm

- DFP Algorithm
  1. Set k:=0; select $x^{(0)}$, and a real symmetric positive definite H_0.
  2. If $g^{(k)} = 0$, stop; else $d^{(k)} = -H_k g^{(k)}$.
  3. Compute

$$
\begin{aligned}
\alpha_k &= \arg\min{}_{\alpha \geq 0} f(x^{(k)} + \alpha d^{(k)}) \\
x^{(k+1)} &= x^{(k)} + \alpha_k d^{(k)}
\end{aligned}
$$

  4. Compute

$$
\begin{aligned}
\Delta x^{(k)} &= \alpha_k d^{(k)} \\
\Delta g^{(k)} &= g^{(k+1)} - g^{(k)}
\end{aligned}
$$

$$H_{k+1} = H_k + \frac{\Delta x^{(k)} \Delta x^{(k)T}}{\Delta x^{(k)T} \Delta g^{(k)}} - \frac{[H_k \Delta g^{(k)}][H_k \Delta g^{(k)}]^T}{\Delta g^{(k)T} H_k \Delta g^{(k)}}$$

5. Set k:=$k+1$; go to step 2.

- **Theorem 11.3** *In the DFP algorithm applied to the quadratic with Hessian $Q = Q^T$, we have $H_{k+1}\Delta g^{(i)} = \Delta x^{(i)}, 0 \le i \le k$.*

- **Ex. 11.3** Locate the minimizer of

$$f(x) = \frac{1}{2} x^T \begin{bmatrix} 4 & 2 \\ 2 & 2 \end{bmatrix} x - x^T \begin{bmatrix} -1 \\ 1 \end{bmatrix}, x \in \Re^2$$

Use the initial point $x^{(0)} = [0, 0]^T$ and $H_0 = I_2$.

$$\begin{aligned} g^{(k)} &= \begin{bmatrix} 4 & 2 \\ 2 & 2 \end{bmatrix} x^{(k)} - \begin{bmatrix} -1 \\ 1 \end{bmatrix} \\ g^{(0)} &= [1, -1]^T, \\ d^{(0)} &= -H_0 g^{(0)} = \begin{bmatrix} -1 \\ 1 \end{bmatrix} \end{aligned}$$

Because f is a quadratic function

$$\begin{aligned} \alpha_0 &= \arg\min_{\alpha \ge 0} f(x^{(0)} + \alpha d^{(0)}) = -\frac{g^{(0)T} d^{(0)}}{d^{(0)T} Q d^{(0)}} = 1 \\ x^{(1)} &= x^{(0)} + \alpha_0 d^{(0)} = [-1, 1]^T \\ \Delta x^{(0)} &= x^{(1)} - x^{(0)} = [-1, 1]^T \end{aligned}$$

$$
\begin{aligned}
g^{(1)} &= Qx - b = \begin{bmatrix} -1 \\ -1 \end{bmatrix} \\
\Delta g^{(0)} &= g^{(1)} - g^{(0)} = [-2, 0]^T
\end{aligned}
$$

We now compute H_1 as

$$
\begin{aligned}
H_1 &= H_0 + \frac{\Delta x^{(0)} \Delta x^{(0)T}}{\Delta x^{(0)T} \Delta g^{(0)}} - \frac{(H_0 \Delta g^{(0)})(H_0 \Delta g^{(0)})^T}{\Delta g^{(0)T} H_0 \Delta g^{(0)}} \\
&= \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{3}{2} \end{bmatrix} \\
d^{(1)} &= -H_1 g^{(1)} = [0, 1]^T \\
\alpha_1 &= \arg\min_{\alpha \geq 0} f(x^{(1)} + \alpha d^{(1)}) = -\frac{g^{(1)T} d^{(1)}}{d^{(1)T} Q d^{(1)}} = \frac{1}{2} \\
x^{(2)} &= x^{(1)} + \alpha_1 d^{(1)} = [-1, 3/2]^T = x^*
\end{aligned}
$$

Because f is a quadratic function of two variables.

- **Theorem 11.4** *Suppose that $g^{(k)} \neq 0$. In the DFP algorith, if H_k is positive definite, then so is H_{k+1}.*
- DFP alogorithm is superior to the rank one algorithm in that it preserves the positive definiteness of H_k. However, in case of larger nonquadratic problems, the algorithm has the tendency of sometimes getting "stuck".

11.5 BFGS Algorithm

- Updating formulas for the approximation of the inverse of the Hessian matrix were based on satisfying the equations

$$H_{k+1}\Delta g^{(i)} = \Delta x^{(i)}, \quad 0 \le i \le k$$

- Let B_k be the estimate of Q at the kth step.

$$\Delta g^{(i)} = B_{k+1}\Delta x^{(i)}, \quad 0 \le i \le k$$

- The roles of $\Delta x^{(i)}$ and $\Delta g^{(i)}$ are interchanged.
- DFP update for the approximation H_k of the inverse Hessian is

$$H_{k+1}^{DFP} = H_k + \frac{\Delta x^{(k)}\Delta x^{(k)T}}{\Delta x^{(k)T}\Delta g^{(k)}} - \frac{H_k\Delta g^{(k)}\Delta g^{(k)T}H_k}{\Delta g^{(k)T}H_k\Delta g^{(k)}}$$

- The Approximation B_k of the Hessian:

$$B_{k+1} = B_k + \frac{\Delta g^{(k)}\Delta g^{(k)T}}{\Delta g^{(k)T}\Delta x^{(k)}} - \frac{B_k\Delta x^{(k)}\Delta x^{(k)T}B_k}{\Delta x^{(k)T}B_k\Delta x^{(k)}}$$

- **Lemma 11.1** *Let A be a nonsingular matrix. Let u and v be column vectors such that $1 + v^TA^{-1}u \neq 0$. Then, $A + uv^T$ is nonsingular, and*

$$(A + uv^T)^{-1} = A^{-1} - \frac{(A^{-1}u)(v^TA^{-1})}{1 + v^TA^{-1}u}$$

- Using the result of the above Lemma,

$$H_{k+1}^{BFGS} = (B_{k+1})^{-1}$$

$$= H_k + \left(1 + \frac{\Delta g^{(k)T} H_k \Delta g^{(k)}}{\Delta g^{(k)T} \Delta x^{(k)}}\right) \frac{\Delta x^{(k)} \Delta x^{(k)T}}{\Delta x^{(k)T} \Delta g^{(k)}} - \frac{H_k \Delta g^{(k)} \Delta x^{(k)T} + (H_k \Delta g^{(k)} \Delta x^{(k)T})^T}{\Delta g^{(k)T} \Delta x^{(k)}}$$

- **Ex. 11.4** Use the BFGS method to minimize

$$f(x) = \frac{1}{2} x^T Q x - x^T b + log(\pi)$$

$$Q = \begin{bmatrix} 5 & -3 \\ -3 & 2 \end{bmatrix}, \; b = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

Take $H_0 = I_2$, and $x^{(0)} = [0, 0]^T$.
1-iteration

$$d^{(0)} = -g^{(0)} = -(Qx^{(0)} - b) = b = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

The objective function is a quadratic, and hence

$$\alpha_0 = -\frac{g^{(0)T} d^{(0)}}{d^{(0)T} Q d^{(0)}} = \frac{1}{2}$$

$$x^{(1)} = x^{(0)} + \alpha_0 d^{(0)} = \begin{bmatrix} 0 \\ 1/2 \end{bmatrix}$$

$$\Delta x^{(0)} = x^{(1)} - x^{(0)} = \begin{bmatrix} 0 \\ 1/2 \end{bmatrix}$$

$$
\begin{aligned}
g^{(1)} &= Qx^{(1)} - b = \begin{bmatrix} -3/2 \\ 0 \end{bmatrix} \\
\Delta g^{(0)} &= g^{(1)} - g^{(0)} = \begin{bmatrix} -3/2 \\ 1 \end{bmatrix} \\
H_1 &= H_0 + \left(1 + \frac{\Delta g^{(0)T} H_0 \Delta g^{(0)}}{\Delta g^{(0)T} \Delta x^{(0)}}\right) \frac{\Delta x^{(0)} \Delta x^{(0)T}}{\Delta x^{(0)T} \Delta g^{(0)}} \\
&\quad - \frac{\Delta x^{(0)} \Delta g^{(0)T} H_0 + H_0 \Delta g^{(0)} \Delta x^{(0)T}}{\Delta g^{(0)T} \Delta x^{(0)}} \\
&= \begin{bmatrix} 1 & 3/2 \\ 3/2 & 11/4 \end{bmatrix}
\end{aligned}
$$

2-iteration

$$
\begin{aligned}
d^{(1)} &= -H_1 g^{(1)} = \begin{bmatrix} 3/2 \\ 9/4 \end{bmatrix} \\
\alpha_1 &= -\frac{g^{(1)T} d^{(1)}}{d^{(1)T} Q d^{(1)}} = 2 \\
x^{(2)} &= x^{(1)} + \alpha_1 d^{(1)} = \begin{bmatrix} 3 \\ 5 \end{bmatrix}
\end{aligned}
$$

$x^{(2)}$ is the minimizer. And the gradient at $x^{(2)}$ is 0; that is, $g^{(2)} = 0$.

- For nonlinear quadratic problems, quasi-Newton

algorithm will not usuall converge in n step. Instead, some modifications may be necessary.(Stopping conditions)