BFGS ALGORITHM

SIEYE RYU

Suppose that n is a positive integer and $f : \mathbb{R}^n \to \mathbb{R}$ is of class C^2. We investigate the notion of BFGS algorithm in order to find the minimum of f. This will be done with basics from vector calculus such as Taylor's theorem and Lagrange multiplier method. Throughout the article, $x \in \mathbb{R}^n$ is regarded as a column vector.

1. Newton's Method

We recall the classical Newton's method. Suppose we want to find a root of a differentiable function $g : \mathbb{R} \to \mathbb{R}$. First, we choose an initial value $x_0 \in \mathbb{R}$. The tangent line through $(x_0, g(x_0))$ is given by

$$y = g'(x_0)(x - x_0) + g(x_0) \tag{1.1}$$

The root of (1.1) is taken as x_1, that is,

$$x_1 = x_0 - \frac{g(x_0)}{g'(x_0)}.$$

In general, the sequence $\{x_k\}_{k=1}^{\infty}$ is determined by the following recurrence relation:

$$x_{k+1} = x_k - \frac{g(x_k)}{g'(x_k)} \qquad (k = 0, 1, 2, \cdots).$$

If the initial value x_0 is close enough to a root of $g(x)$ and $g'(x_0) \neq 0$, then x_k converges to the root of $g(x)$ as k goes to the infinity.

Newton's method can be used to find a minimum or maximum of a C^2 function. In this case, we apply Newton's method to the derivative because the derivative is zero at a minimum or maximum. Suppose that we want to find a minimum of a C^2 function $g : \mathbb{R} \to \mathbb{R}$. The second order Taylor approximation of g is

$$g(x + d) \approx g(x) + g'(x)d + \frac{1}{2}g''(x)d^2.$$

This approximation is valid only if d is small enough. If we put

$$q(d) = g(x) + g'(x)d + \frac{1}{2}g''(x)d^2,$$

then $q(d)$ is minimized when

$$q'(d) = 0.$$

Since

$$q'(d) = g'(x) + g''(x)d,$$

it follows that $q(d)$ attains a minimum at

$$d = -\frac{g'(x)}{g''(x)}.$$

We note that $q'(d) = 0$ does not imply that q is minimized at d. It is well known that $g''(x)$ is positive at a local minimum. Since $q(d)$ is a quadratic function, positiveness of $g''(x)$ guarantees the existence of a minimum by the convexity of $q(d)$. We choose an initial value $x_0 \in \mathbb{R}$ and let

$$x_{k+1} = x_k - \frac{g'(x_k)}{g''(x_k)} \qquad (k = 0, 1, 2, \cdots).$$

The same argument can be applied when we want to find a maximum of $g(x)$ because a maximum of $g(x)$ is actually a minimum of $-g(x)$.

Now, we use Newton's method in order to find a minimum point of a C^2 function $f : \mathbb{R}^n \to \mathbb{R}$. Suppose $x, d \in \mathbb{R}^n$. The second order Taylor approximation of f is

$$f(x + d) \approx f(x) + \sum_{i=1}^{n} d_i \frac{\partial f}{\partial x_i}(x) + \frac{1}{2} \sum_{i,j=1}^{n} d_i d_j \frac{\partial^2 f}{\partial x_i \partial x_j}(x). \tag{1.2}$$

This approximation is valid only if d is sufficiently close to $0 \in \mathbb{R}^n$. If we denote the gradient of f and the Hessian of f by ∇f and Hf, respectively, then (1.2) can be written as follows:

$$f(x + d) \approx f(x) + d^\mathsf{T} \nabla f(x) + \frac{1}{2} d^\mathsf{T} Hf(x) d.$$

If we set

$$q(d) = f(x) + d^\mathsf{T} \nabla f(x) + \frac{1}{2} d^\mathsf{T} Hf(x) d, \tag{1.3}$$

then $q(d)$ is minimized when

$$\nabla q(d) = 0.$$

Since $\nabla q(d) = \nabla f(x) + Hf(x)d$, it follows that $q(d)$ attains a minimum at

$$d = -Hf^{-1}(x) \nabla f(x).$$

In the same manner as the one-dimensional case, $\nabla q(d) = 0$ does not guarantee that d is a minimum point of q. It is well known that the Hessian $Hf(x)$ is positive-definite at a local minimum x. Hence, if $Hf(x)$ is positive-definite, then the point $d \in \mathbb{R}^n$ satisfying $\nabla q(d) = 0$ is the minimum of q because q is a convex quadratic function. We choose an initial value $x_0 \in \mathbb{R}^n$ and let

$$\delta_k = -Hf^{-1}(x_k) \nabla f(x_k),$$

$$x_{k+1} = x_k + \delta_k \qquad (k = 0, 1, 2, \cdots).$$

If the Hessian $Hf(x_k)$ at x_k is positive definite, then so is $Hf^{-1}(x_k)$ and δ_k becomes the downhill direction from x_k. We discuss it in the rest of the section. For $x \in \mathbb{R}^n$, a direction $d \in \mathbb{R}^n$ is called a *downhill direction* if there is a positive real number α' such that

$$0 < \alpha < \alpha' \qquad \Rightarrow \qquad f(x + \alpha d) < f(x).$$

Lemma 1.1. *Suppose that $x, d \in \mathbb{R}^n$. If d satisfies*

$$d^\mathsf{T} \nabla f(x) < 0,$$

then d is a downhill direction. In particular, $-\nabla f(x)$ is a downhill direction.

Proof. By Taylor's theorem,

$$f(x+\alpha d) = f(x) + \alpha d^{\mathsf{T}} \nabla f(x) + R_2(x, \alpha d),$$

where

$$\frac{|R_2(x,\alpha d)|}{|\alpha|||d||} \to 0 \qquad \text{as} \qquad \alpha \to 0.$$

Hence, there is a positive real number α' such that

$$0 < \alpha < \alpha' \qquad \Rightarrow \qquad \frac{|R_2(x,\alpha d)|}{|\alpha|||d||} < |d^{\mathsf{T}} \nabla f(x)|.$$

This implies that

$$0 < \alpha < \alpha' \qquad \Rightarrow \qquad f(x+\alpha d) - f(x) = \alpha d^{\mathsf{T}} \nabla f(x) + R_2(x, \alpha d) < 0$$

because $d^{\mathsf{T}} \nabla f(x) < 0$. □

2. Quasi-Newton Method

If n is large, then Hessian is impractical to compute. Quasi-Newton method uses approximate Hessian H_k. Instead of (1.3), we apply the same argument to

$$q(d) = f(x) + d^{\mathsf{T}} \nabla f(x) + \frac{1}{2} d^{\mathsf{T}} H_k d.$$

If $x_0 \in \mathbb{R}^n$ is an initial value, then we have

$$\delta_k = -H_k^{-1} \nabla f(x_k)$$

and

$$x_{k+1} = x_k + \delta_k \qquad (k = 0, 1, 2, \cdots).$$

There are several methods to update H_k and in the next section we will investigate BFGS updates, which is one of quasi-Newton methods. In this section, we discuss four desired properties of H_k in quasi-Newton methods. We start with the secant condition. Since the approximation of $f(x_{k+1})$ is

$$f(x_k + \delta_k) \approx q(\delta_k) = f(x_k) + \delta_k^{\mathsf{T}} \nabla f(x_k) + \frac{1}{2} \delta_k^{\mathsf{T}} H_k \delta_k$$

and $\nabla q(d) = \nabla f(x) + H_k d$, it follows that

$$\begin{aligned} \nabla f(x_{k+1}) - \nabla f(x_k) &\approx \nabla q(\delta_k) - \nabla q(0) \\ &= \nabla f(x_k) + H_{k+1} \delta_k - \nabla f(x_k) \\ &= H_{k+1} \delta_k. \end{aligned}$$

For each $k = 0, 1, 2, \cdots$, we let

$$\gamma_k = \nabla f(x_{k+1}) - \nabla f(x_k).$$

We say that H_k satisfies *secant condition* if

$$\gamma_k = H_{k+1} \delta_k. \tag{2.1}$$

The first and the second desired properties of H_k are as follows:

(1) For a convex quadratic C^2 function $f : \mathbb{R}^n \to \mathbb{R}$, H_k converges to the Hessian Hf of f as k goes to the infinity.

(2) H_k satisfies the secant condition.

Since f is a C^2 function, its Hessian is well-defined and symmetric at $x \in \mathbb{R}^n$. It is known that if x^* is a a local minimum point, then the Hessian $Hf(x^*)$ is positive-definite at x^*. The following is the third desired property of H_k:

(3) H_k is nonsingular, symmetric and positive definite.

We note that the third property guarantees that $\delta_k = -H_k^{-1}\nabla f(x_k)$ is the downhill direction from x_k.

The last property is as follows:

(4) We minimize the change in H_k. This is done by minimizing $||H_{k+1} - H_k||$ in some norm or alternatively minimizing $||H_{k+1}^{-1} - H_k^{-1}||$.

3. BFGS Updates

We start the section with the Frobenius norm. We need some notation. If n is a positive integer, we denote the set of $n \times n$ real matrix by $M(n)$. The Frobenius norm $||\cdot||_F$ is defined by

$$||A||_F = \sqrt{\operatorname{tr}(A^\mathsf{T} A)} = \sqrt{\sum_{i,j=1}^{n} |A(i,j)|^2} \qquad (A \in M(n)).$$

When $W \in M(n)$, a weighted Frobenius norm $||\cdot||_W$ is define by

$$||A||_W = \sqrt{\frac{1}{2}\operatorname{tr}(WAWA^\mathsf{T})} \qquad (A \in M(n)).$$

If $W = M^\mathsf{T} M$ for some $M \in M(n)$ and W is nonsingular, BFGS update is obtained by solving

$$\min_{H \in M(n)} ||H^{-1} - H_k^{-1}||_W,$$
$$\text{where} \qquad \delta_k = H^{-1}\gamma_k \qquad \text{and} \qquad H^\mathsf{T} = H. \tag{3.1}$$

For convenience, we drop the index k and write δ and γ instead of δ_k and γ_k. If we set

$$E = H^{-1} - H_k^{-1},$$

then (3.1) can be written as follows:

$$\min_{E \in M(n)} ||E||_W^2,$$
$$\text{where} \qquad E\gamma = \delta - H_k^{-1}\gamma \qquad \text{and} \qquad E^\mathsf{T} = E. \tag{3.2}$$

If we define a function $L(E, z, N) : M(n) \times \mathbb{R}^n \times M(n) \to \mathbb{R}$ by

$$L(E, z, N) = \operatorname{tr}\left[\frac{1}{2}WEWE^\mathsf{T} + z^\mathsf{T}(E\gamma - (\delta - H_k^{-1}\gamma)) + N(E - E^\mathsf{T})\right],$$

then there are $\lambda \in \mathbb{R}^n$ and $\Gamma \in M(n)$ such that

$$\frac{\partial L}{\partial E(s,t)}(E, \lambda, \Gamma) = 0 \tag{3.3}$$

for all $1 \le s, t \le n$. Lagrange multiplier method tells us that E satisfying (3.3) is a minimum point with respect to the weighted Frobenius norm $||\cdot||_W$ in the restrictions

$$E\gamma = \delta - H_k^{-1}\gamma \qquad \text{and} \qquad E^\mathsf{T} = E.$$

We first solve (3.3). Since

$$\frac{1}{2}\mathrm{tr}(WEWE^\mathsf{T}) = \frac{1}{2}\sum_{i,j,k,l=1}^{n} W(i,j)W(k,l)E(j,k)E(i,l),$$

it follows that

$$\frac{\partial}{\partial E(s,t)}\frac{1}{2}\mathrm{tr}(WEWE^\mathsf{T}) = \sum_{i,l=1}^{n} W(s,i)E(i,l)W(l,t) = WEW(s,t). \tag{3.4}$$

Since

$$\frac{\partial}{\partial E(s,t)}\mathrm{tr}\left(\lambda^\mathsf{T}(E\gamma - (\delta - H_k^{-1}\gamma))\right) = \frac{\partial}{\partial E(s,t)}\mathrm{tr}\left(\lambda^\mathsf{T}E\gamma\right)$$

and

$$\lambda^\mathsf{T}E\gamma = \sum_{i,j=1}^{n}\lambda(i)E(i,j)\gamma(j),$$

it follows that

$$\frac{\partial}{\partial E(s,t)}\mathrm{tr}\left(\lambda^\mathsf{T}(E\gamma - (\delta - H_k^{-1}\gamma))\right) = \lambda(s)\gamma(t) = \lambda\gamma^\mathsf{T}(s,t). \tag{3.5}$$

Finally, from

$$\mathrm{tr}\left(\Gamma(E - E^\mathsf{T})\right) = \sum_{i,j=1}^{n}\Gamma(i,j)\left(E(j,i) - E(i,j)\right),$$

we obtain

$$\frac{\partial}{\partial E(s,t)}\mathrm{tr}\left(\Gamma(E - E^\mathsf{T})\right) = \Gamma(t,s) - \Gamma(s,t) = (\Gamma^\mathsf{T} - \Gamma)(s,t). \tag{3.6}$$

From (3.4), (3.5) and (3.6), (3.3) becomes

$$WEW + \lambda\gamma^\mathsf{T} + \Gamma^\mathsf{T} - \Gamma = 0$$

and (3.2) can be written as follows:

$$WEW + \lambda\gamma^\mathsf{T} + \Gamma^\mathsf{T} - \Gamma = 0, \tag{3.7}$$

$$E\gamma = \delta - H_k^{-1}\gamma \tag{3.8}$$

and

$$E^\mathsf{T} = E. \tag{3.9}$$

Since $W = M^\mathsf{T}M$ is nonsingular, (3.7) can be written as follows:

$$E = -W^{-1}(\lambda\gamma^\mathsf{T} + \Gamma^T - \Gamma)W^{-1}. \tag{3.10}$$

Since $W = M^\mathsf{T}M$ is symmetric, (3.9) implies

$$\lambda\gamma^\mathsf{T} + \Gamma^\mathsf{T} - \Gamma = \gamma\lambda^\mathsf{T} + \Gamma - \Gamma^\mathsf{T}$$

and this implies

$$\Gamma^\mathsf{T} - \Gamma = \frac{1}{2}(\gamma\lambda^\mathsf{T} - \lambda\gamma^\mathsf{T}).$$

Hence, (3.10) becomes

$$E = -\frac{1}{2}W^{-1}(\gamma\lambda^\mathsf{T} + \lambda\gamma^\mathsf{T})W^{-1}. \tag{3.11}$$

If we substitute it to (3.8), then we obtain

$$\gamma_k\lambda^\mathsf{T}W^{-1}\gamma + \lambda\gamma^\mathsf{T}W^{-1}\gamma = -2W(\delta - H_k^{-1}\gamma).$$

Since $\gamma^\mathsf{T} W^{-1}\gamma$ is a scalar, we have

$$\lambda = \frac{-2W\delta + 2WH_k^{-1}\gamma - \gamma\lambda^\mathsf{T} W^{-1}\gamma}{\gamma^\mathsf{T} W^{-1}\gamma}. \tag{3.12}$$

In order to represent λ in (3.12) with respect to W, H_k, δ and γ, we multiply both sides of (3.12) by $\gamma^\mathsf{T} W^{-1}$:

$$\gamma^\mathsf{T} W^{-1}\lambda = \frac{-2\gamma^\mathsf{T}\delta + 2\gamma^\mathsf{T} H_k^{-1}\gamma - \gamma^\mathsf{T} W^{-1}\gamma\lambda^\mathsf{T} W^{-1}\gamma}{\gamma^\mathsf{T} W^{-1}\gamma}. \tag{3.13}$$

We transpose (3.13):

$$\begin{aligned}\lambda^\mathsf{T} W^{-1}\gamma &= \frac{-2\delta^\mathsf{T}\gamma + 2\gamma^\mathsf{T} H_k^{-1}\gamma - \gamma^\mathsf{T} W^{-1}\lambda\gamma^\mathsf{T} W^{-1}\gamma}{\gamma^\mathsf{T} W^{-1}\gamma}\\ &= \frac{-2\delta^\mathsf{T}\gamma + 2\gamma^\mathsf{T} H_k^{-1}\gamma}{\gamma^\mathsf{T} W^{-1}\gamma} - \lambda^\mathsf{T} W^{-1}\gamma\end{aligned}$$

and this implies

$$\lambda^\mathsf{T} W^{-1}\gamma = \frac{-\delta^\mathsf{T}\gamma + \gamma^\mathsf{T} H_k^{-1}\gamma}{\gamma^\mathsf{T} W^{-1}\gamma}. \tag{3.14}$$

We substitute (3.14) to (3.12):

$$\lambda = \frac{-2W\delta + 2WH_k^{-1}\gamma}{\gamma^\mathsf{T} W^{-1}\gamma} + \frac{\gamma(\delta^\mathsf{T} - \gamma^\mathsf{T} H_k^{-1})\gamma}{(\gamma^\mathsf{T} W^{-1}\gamma)^2}.$$

Again, we substitute this to (3.11):

$$\begin{aligned}E &= -\frac{1}{2}W^{-1}\left(\frac{-2\gamma\delta^\mathsf{T} W + 2\gamma\gamma^\mathsf{T} H_k^{-1}W}{\gamma^\mathsf{T} W^{-1}\gamma} + \frac{\gamma\gamma^\mathsf{T}(\delta - H_k^{-1}\gamma)\gamma^\mathsf{T}}{(\gamma^\mathsf{T} W^{-1}\gamma)^2}\right.\\ &\qquad \left. + \frac{-2W\delta\gamma^\mathsf{T} + 2WH_k^{-1}\gamma\gamma^\mathsf{T}}{\gamma^\mathsf{T} W^{-1}\gamma} + \frac{\gamma(\delta^\mathsf{T} - \gamma^\mathsf{T} H_k^{-1})\gamma\gamma^\mathsf{T}}{(\gamma^\mathsf{T} W^{-1}\gamma)^2}\right)W^{-1}.\end{aligned}$$

Since

$$\gamma(\delta^\mathsf{T} - \gamma^\mathsf{T} H_k^{-1})\gamma = \gamma\gamma^\mathsf{T}(\delta - H_k^{-1}\gamma),$$

we have

$$\begin{aligned}E &= \frac{W^{-1}\gamma(\delta^\mathsf{T} - \gamma^\mathsf{T} H_k^{-1})}{\gamma^\mathsf{T} W^{-1}\gamma} + \frac{(\delta - H_k^{-1}\gamma)\gamma^\mathsf{T} W^{-1}}{\gamma^\mathsf{T} W^{-1}\gamma}\\ &\quad - \frac{W^{-1}\gamma\gamma^\mathsf{T}(\delta - H_k^{-1}\gamma)\gamma^\mathsf{T} W^{-1}}{(\gamma^\mathsf{T} W^{-1}\gamma)^2}.\end{aligned} \tag{3.15}$$

We have not chosen the weight W yet. Different weights W lead to different updates. Here, we will put $W = H_{k+1}$. In the next proposition, we will see that this choice is possible. A matrix $A \in M(n)$ is said to have a *Cholesky factorization* if there is a lower triangular matrix L whose diagonal entries are all positive such that $X = LL^\mathsf{T}$.

Proposition 3.1. *A matrix $A \in M(n)$ has a Cholesky factorization $A = LL^\mathsf{T}$ if and only if A is symmetric and positive definite. Moreover, if the Cholesky factorization exists, then it is unique.*

Proof. We only prove that every symmetric positive definite matrix has a Cholesky factorization. The proof will be done by mathematical induction on n. When $n = 1$ and $A = [a]$, obviously $L = [\sqrt{a}]$. We assume the proposition is true for n and prove a symmetric positive matrix $A \in M(n+1)$ has a Cholesky factorization. Let

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n+1} \\ a_{21} & a_{22} & \cdots & a_{2n+1} \\ \vdots & \vdots & & \vdots \\ a_{n+11} & a_{n+12} & \cdots & a_{n+1n+1} \end{bmatrix} \quad \text{and} \quad A' = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}.$$

By assumption A' has a Cholesky factorization $A' = LL^\intercal$. When

$$L = \begin{bmatrix} l_{11} & 0 & \cdots & 0 \\ l_{21} & l_{22} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ l_{n1} & l_{n2} & \cdots & l_{nn} \end{bmatrix},$$

we can find $m_1, m_2, \cdots, m_{n+1}$ such that

$$M = \begin{bmatrix} l_{11} & 0 & \cdots & 0 & 0 \\ l_{21} & l_{22} & \cdots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ l_{n1} & la_{n2} & \cdots & l_{nn} & 0 \\ m_1 & m_2 & \cdots & m_n & m_{n+1} \end{bmatrix}$$

has the requiring property. $\square$

If $W = H_{k+1}$, then $W^{-1} = H_{k+1}^{-1} = E + H_k^{-1}$. By (3.15) and the secant condition (2.1), we have

$$\begin{aligned} H_{k+1}^{-1} &= H_k^{-1} + \frac{\delta(\delta^\intercal - \gamma^\intercal H_k^{-1})}{\gamma^\intercal\delta} + \frac{(\delta - H_k^{-1}\gamma)\delta^\intercal}{\gamma^\intercal\delta} - \frac{\delta\gamma^\intercal(\delta - H_k^{-1}\gamma)\delta^\intercal}{(\gamma^\intercal\delta)^2} \\ &= H_k^{-1} - \frac{\delta\gamma^\intercal H_k^{-1}}{\gamma^\intercal\delta} - \frac{H_k^{-1}\gamma\delta^\intercal}{\gamma^\intercal\delta} + \frac{\delta\delta^\intercal}{\gamma^\intercal\delta} + \frac{(\gamma^\intercal H_k^{-1}\gamma)\delta\delta^\intercal}{(\gamma^\intercal\delta)^2} \\ &= H_k^{-1} - \frac{\delta\gamma^\intercal H_k^{-1}}{\gamma^\intercal\delta} - \frac{H_k^{-1}\gamma\delta^\intercal}{\gamma^\intercal\delta} + \frac{1}{\gamma^\intercal\delta}\left(1 + \frac{\gamma^\intercal H_k^{-1}\gamma}{\gamma^\intercal\delta}\right)\delta\delta^\intercal. \end{aligned} \tag{3.16}$$

In order to drive the update of H_k from that of H_k^{-1}, we introduce the Sherman-Morrison formula:

Lemma 3.2. *Suppose that $A \in M(n)$ is nonsingular and that $u, v \in \mathbb{R}^n$. Then $A + uv^\intercal$ is nonsingular if and only if $1 + v^\intercal A^{-1}u$ is nonzero. If $A + uv^\intercal$ is nonsingular, its inverse is*

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

Proof. Direct computations show that

$$(A + uv^\intercal)\left(A^{-1} - \frac{A^{-1}uv^\intercal A^{-1}}{1 + v^\intercal A^{-1}u}\right) = I_n$$

and that

$$\left(A^{-1} - \frac{A^{-1}uv^\intercal A^{-1}}{1 + v^\intercal A^{-1}u}\right)(A + uv^\intercal) = I_n.$$

By the uniqueness of the inverse, the result follows. $\square$

If $A + uv^{\mathsf T} + vu^{\mathsf T}$ is nonsingular, then we can obtain its inverse by applying the Sherman-Morrison formula to

$$(A + uv^{\mathsf T})^{-1} \quad \text{and} \quad \big((A + uv^{\mathsf T}) + vu^{\mathsf T}\big)^{-1}.$$

Direct computations show that

$$\begin{aligned}
&\big(A + uv^{\mathsf T} + vu^{\mathsf T}\big)^{-1} \\
&\quad = A^{-1} + \frac{(u^{\mathsf T}A^{-1}u)A^{-1}vv^{\mathsf T}A^{-1} + (v^{\mathsf T}A^{-1}v)A^{-1}uu^{\mathsf T}}{(1+v^{\mathsf T}A^{-1}u)(1+u^{\mathsf T}A^{-1}v) - (u^{\mathsf T}A^{-1}u)(v^{\mathsf T}A^{-1}v)} \\
&\quad\quad + \frac{-(1+v^{\mathsf T}A^{-1}u)A^{-1}vu^{\mathsf T}A^{-1} - (1+u^{\mathsf T}A^{-1}v)A^{-1}uv^{\mathsf T}A^{-1}}{(1+v^{\mathsf T}A^{-1}u)(1+u^{\mathsf T}A^{-1}v) - (u^{\mathsf T}A^{-1}u)(v^{\mathsf T}A^{-1}v)}.
\end{aligned} \tag{3.17}$$

By (3.17), we obtain

$$\begin{aligned}
&\left(H_k^{-1} - \frac{H_k^{-1}\gamma\delta^{\mathsf T}}{\gamma^{\mathsf T}\delta} - \frac{\delta\gamma^{\mathsf T}H_k^{-1}}{\gamma^{\mathsf T}\delta}\right)^{-1} \\
&= H_k - \frac{(\gamma^{\mathsf T}H_k^{-1}\gamma)H_k\delta\delta^{\mathsf T}H_k}{(\gamma^{\mathsf T}H_k^{-1}\gamma)(\delta^{\mathsf T}H_k\delta)} - \frac{(\delta^{\mathsf T}H_k\delta)\gamma\gamma^{\mathsf T}}{(\gamma^{\mathsf T}H_k^{-1}\gamma)(\delta^{\mathsf T}H_k\delta)}.
\end{aligned}$$

Applying the Sherman-Morrison formula once again, we have

$$\begin{aligned}
H_{k+1} &= \left(H_k^{-1} - \frac{\delta\gamma^{\mathsf T}H_k^{-1}}{\gamma^{\mathsf T}\delta} - \frac{H_k^{-1}\gamma\delta^{\mathsf T}}{\gamma^{\mathsf T}\delta} + \frac{1}{\gamma^{\mathsf T}\delta}\left(1 + \frac{\gamma^{\mathsf T}H_k^{-1}\gamma}{\gamma^{\mathsf T}\delta}\right)\delta\delta^{\mathsf T}\right)^{-1} \\
&= H_k + \frac{\gamma\gamma^{\mathsf T}}{\gamma^{\mathsf T}\delta} - \frac{H_k\delta\delta^{\mathsf T}H_k}{\delta^{\mathsf T}H_k\delta}.
\end{aligned} \tag{3.18}$$

If $u, v \in \mathbb{R}^n$, then the rank of $u \times v^{\mathsf T} \in M(n)$ is 1. The update of the form

$$H_{k+1} = H_k + uv^{\mathsf T}$$

is called a rank 1 update. Our BFGS update is expressed as a sum of rank 1 updates.

If H_k is symmetric, then so is H_{k+1}. Now, we want to see that if H_k is positive definite, then so is H_{k+1}. For any $x \in \mathbb{R}^n \setminus \{0\}$, we have

$$\begin{aligned}
x^{\mathsf T}H_{k+1}x &= x^{\mathsf T}H_kx + x^{\mathsf T}\frac{\gamma\gamma^{\mathsf T}}{\gamma^{\mathsf T}\delta}x - x^{\mathsf T}\frac{H_k\delta\delta^{\mathsf T}H_k}{\delta^{\mathsf T}H_k\delta}x \\
&= \frac{(x^{\mathsf T}H_kx)(\delta^{\mathsf T}H_k\delta) - (x^{\mathsf T}H_k\delta)(\delta^{\mathsf T}H_kx)}{\delta^{\mathsf T}H_k\delta} + \frac{(x^{\mathsf T}\gamma)^2}{\gamma^{\mathsf T}\delta}.
\end{aligned}$$

If we define a map $\langle\ ,\ \rangle_{H_k} : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ by

$$\langle x, y\rangle_{H_k} = x^{\mathsf T}H_ky,$$

then $\langle\ ,\ \rangle_{H_k}$ is an inner product since H_k is positive definite and symmetric. By Cauchy-Schwarz inequality, we have

$$(x^{\mathsf T}H_kx)(\delta^{\mathsf T}H_k\delta) - (x^{\mathsf T}H_k\delta)(\delta^{\mathsf T}H_kx) = \langle x, x\rangle_{H_k}\langle \delta, \delta\rangle_{H_k} - \langle x, \delta\rangle_{H_k}^2 \geq 0.$$

Thus, positiveness of H_k implies

$$\frac{(x^{\mathsf T}H_kx)(\delta^{\mathsf T}H_k\delta) - (x^{\mathsf T}H_k\delta)(\delta^{\mathsf T}H_kx)}{\delta^{\mathsf T}H_k\delta} > 0.$$

It remains to show that
$$\frac{(x^{\mathsf{T}}\gamma)^2}{\gamma^{\mathsf{T}}\delta} > 0.$$
In (1.2), if the length of the vector d is large, then the Taylor approximation is not accurate. Exact line search and inexact line search are common methods to fix this problem. Exact line search is a method to choose α which minimizes $f(x_k+\alpha\delta)$ but it is not considered cost effective. Inexact line search is a method to try different α until the step length α is acceptable (not to be too long or not to be too short) so that x_k would converge to some value well. Wolfe conditions provide upper and lower bound on the admissible step length values when we perform inexact line search. We say α satisfies the Wolfe conditions if the following two inequalities hold:

(1) $f(x_k + \alpha\delta) \le f(x_k) + c_1\alpha\delta^{\mathsf{T}}\nabla f(x_k) \quad (0 < c_1 < 1)$.

(2) $-\delta^{\mathsf{T}}\nabla f(x_k + \alpha\delta) \le -c_2\delta^{\mathsf{T}}\nabla f(x_k) \quad (c_1 < c_2 < 1)$.

The first inequality (1) ensures that the step length decreases f sufficiently and the second inequality (2) ensures that the slop has been reduced sufficiently.

If we perform exact line search, then we have
$$\frac{d}{d\alpha} f(x_k + \alpha\delta) = 0.$$
Since
$$\frac{d}{d\alpha} f(x_k + \alpha\delta) = \delta^{\mathsf{T}}\nabla f(x_{k+1}),$$
we have
$$\gamma^{\mathsf{T}}\delta = \delta^{\mathsf{T}}\gamma = \delta^{\mathsf{T}}(\nabla f(x_{k+1}) - \nabla f(x_k)) = -\delta^{\mathsf{T}}\nabla f(x_k) > 0$$
by Lemma 1.1.

If we perform inexact line search, then we have
$$\gamma^{\mathsf{T}}\delta = \delta^{\mathsf{T}}\gamma = \delta^{\mathsf{T}}(\nabla f(x_{k+1}) - \nabla f(x_k)) = -\delta^{\mathsf{T}}\nabla f(x_k) > 0$$
by Lemma 1.1 and the second Wolfe condition.

References

[1] Quasi-Newton optimization: Origin of the BFGS update Steven G. Johnson, notes for 18.335 at MIT April 25, 2019

[2] Continuous optimisation: Quasi-Newton methods, Honour School of Mathematics, Oxford University Hilary Term 2005, Dr. Raphael Hauser

[3] Wolfe Conditions, Wikipedia