MATH 4211/6211 – Optimization

Quasi-Newton Method

Xiaojing Ye
Department of Mathematics & Statistics
Georgia State University

Quasi-Newton Method

Motivation: Approximate the inverse Hessian $(\nabla^2 f(\boldsymbol{x}^{(k)}))^{-1}$ in the Newton's method by some $\boldsymbol{H}_k$:

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - \alpha_k \boldsymbol{H}_k \boldsymbol{g}^{(k)}$$

That is, the search direction is set to $\boldsymbol{d}^{(k)} = -\boldsymbol{H}_k \boldsymbol{g}^{(k)}$.

Based on $\boldsymbol{H}_k, \boldsymbol{x}^{(k)}, \boldsymbol{g}^{(k)}$, quasi-Newton generates the next $\boldsymbol{H}_{k+1}$, and so on.

Proposition. If $f \in \mathcal{C}^1$, $\boldsymbol{g}^{(k)} \neq \mathbf{0}$, and $\boldsymbol{H}_k \succ \mathbf{0}$, then $\boldsymbol{d}^{(k)} = -\boldsymbol{H}_k \boldsymbol{g}^{(k)}$ is a descent direction.

Proof. Let $\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - \alpha \boldsymbol{H}_k \boldsymbol{g}^{(k)}$ for some α, then by Taylor's expansion

$$f(\boldsymbol{x}^{(k+1)}) = f(\boldsymbol{x}^{(k)}) - \alpha {\boldsymbol{g}^{(k)}}^{\top} \boldsymbol{H}_k \boldsymbol{g}^{(k)} + o(\|\boldsymbol{H}_k \boldsymbol{g}^{(k)}\| \alpha) < f(\boldsymbol{x}^{(k)})$$

for α sufficiently small.

Recall that for quadratic functions with $\boldsymbol{Q} \succ \mathbf{0}$, the Hessian is $\boldsymbol{H}^{(k)} = \boldsymbol{Q}$ for all k, and

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

For notation simplicity, we denote

$$\Delta \boldsymbol{x}^{(k)} = \boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)} \quad \text{and} \quad \Delta \boldsymbol{g}^{(k)} = \boldsymbol{g}^{(k+1)} - \boldsymbol{g}^{(k)}$$

Then we can write the identity above as

$$\Delta \boldsymbol{g}^{(k)} = \boldsymbol{Q} \Delta \boldsymbol{x}^{(k)}$$

or equivalently

$$\boldsymbol{Q}^{-1} \Delta \boldsymbol{g}^{(k)} = \Delta \boldsymbol{x}^{(k)}$$

In quasi-Newton method, $\boldsymbol{H}_k$ is in the place of $\boldsymbol{Q}^{-1}$:

$$\textbf{Newton}: \quad \boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - \alpha_k \boldsymbol{Q}^{-1} \boldsymbol{g}^{(k)}$$

$$\textbf{Quasi-Newton}: \quad \boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - \alpha_k \boldsymbol{H}_k \boldsymbol{g}^{(k)}$$

Therefore we would like to have a sequence of $\boldsymbol{H}_k$ with same property of $\boldsymbol{Q}^{-1}$:

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

for all $k = 0, 1, 2, \dots$.

If this is true, then at iteration n, there are

$$\begin{aligned} \boldsymbol{H}_n \Delta \boldsymbol{g}^{(0)} &= \Delta \boldsymbol{x}^{(0)} \\ \boldsymbol{H}_n \Delta \boldsymbol{g}^{(1)} &= \Delta \boldsymbol{x}^{(1)} \\ &\vdots \\ \boldsymbol{H}_n \Delta \boldsymbol{g}^{(n-1)} &= \Delta \boldsymbol{x}^{(n-1)} \end{aligned}$$

or $\boldsymbol{H}_n[\Delta \boldsymbol{g}^{(0)}, \dots, \Delta \boldsymbol{g}^{(n-1)}] = [\Delta \boldsymbol{x}^{(0)}, \dots, \Delta \boldsymbol{x}^{(n-1)}]$.

On the other hand, $\boldsymbol{Q}^{-1}[\Delta \boldsymbol{g}^{(0)}, \dots, \Delta \boldsymbol{g}^{(n-1)}] = [\Delta \boldsymbol{x}^{(0)}, \dots, \Delta \boldsymbol{x}^{(n-1)}]$. If $[\Delta \boldsymbol{g}^{(0)}, \dots, \Delta \boldsymbol{g}^{(n-1)}]$ is invertible, then we have $\boldsymbol{H}_n = \boldsymbol{Q}^{-1}$.

Then at the iteration $n+1$, there is $\boldsymbol{x}^{(n+1)} = \boldsymbol{x}^{(n)} - \alpha_n \boldsymbol{H}_n \boldsymbol{g}^{(n)} = \boldsymbol{x}^*$ since this is the same as the Newton's update.

Hence for quadratic functions, quasi-Newton method would converge in at most n steps.

Quasi-Newton method

$$\begin{aligned} \boldsymbol{d}^{(k)} &= -\boldsymbol{H}_k \boldsymbol{g}^{(k)} \\ \alpha_k &= \operatorname*{arg\,min}_{\alpha \geq 0} f(\boldsymbol{x}^{(k)} + \alpha_k \boldsymbol{d}^{(k)}) \\ \boldsymbol{x}^{(k+1)} &= \boldsymbol{x}^{(k)} + \alpha_k \boldsymbol{d}^{(k)} \end{aligned}$$

where $\boldsymbol{H}_0, \boldsymbol{H}_1, \dots$ are symmetric.

Moreover, for quadratic functions of form $f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^\top \boldsymbol{Q}\boldsymbol{x} - \boldsymbol{b}^\top \boldsymbol{x}$, the matrices $\boldsymbol{H}_0, \boldsymbol{H}_1, \dots$ are required to satisfy

$$\boldsymbol{H}_{k+1} \Delta \boldsymbol{g}^{(i)} = \Delta \boldsymbol{x}^{(i)}, \quad 0 \leq i \leq k$$

Theorem. Consider a quasi-Newton algorithm applied to a quadratic function with symmetric $\boldsymbol{Q} \succ \mathbf{0}$, such that for all $k = 0, 1, \ldots, n-1$, there are

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

and $\boldsymbol{H}_k$ are all symmetric. If $\alpha_i \neq 0$ for $0 \le i \le k$, then $\boldsymbol{d}^{(0)}, \ldots, \boldsymbol{d}^{(n)}$ are $\boldsymbol{Q}$-conjugate.

Proof. We prove by induction. It is trivial to show ${\boldsymbol{g}^{(1)}}^\top \boldsymbol{d}^{(i)}$.

Assume the claim holds for some $k < n - 1$. We have for $i \leq k$ that

$$\begin{aligned}
{\boldsymbol{d}^{(k+1)}}^\top \boldsymbol{Q}\boldsymbol{d}^{(i)} &= -(\boldsymbol{H}_{k+1}\boldsymbol{g}^{(k+1)})^\top \boldsymbol{Q}\boldsymbol{d}^{(i)} \\
&= -{\boldsymbol{g}^{(k+1)}}^\top \boldsymbol{H}_{k+1} \frac{\boldsymbol{Q}\boldsymbol{\Delta x}^{(i)}}{\alpha_i} \\
&= -{\boldsymbol{g}^{(k+1)}}^\top \boldsymbol{H}_{k+1} \frac{\boldsymbol{\Delta g}^{(i)}}{\alpha_i} \\
&= -{\boldsymbol{g}^{(k+1)}}^\top \frac{\boldsymbol{\Delta x}^{(i)}}{\alpha_i} \\
&= -{\boldsymbol{g}^{(k+1)}}^\top \boldsymbol{d}^{(i)}
\end{aligned}$$

Since $\boldsymbol{d}^{(0)}, \dots, \boldsymbol{d}^{(k)}$ are $\boldsymbol{Q}$-conjugate, we know ${\boldsymbol{g}^{(k+1)}}^\top \boldsymbol{d}^{(i)} = 0$ for all $i \leq k$. Hence $\boldsymbol{d}^{(0)}, \dots, \boldsymbol{d}^{(k)}, \boldsymbol{d}^{(k+1)}$ are $\boldsymbol{Q}$-conjugate. By induction the claim holds.

The theorem above also shows that quasi-Newton method is a conjugate direction method, and hence converges in n steps for quadratic objective functions.

In practice, there are various ways to generate $\boldsymbol{H}_k$ such that

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

Now we learn three algorithms that produce such $\boldsymbol{H}_k$.

Rank one correction formula

Suppose we would like to update $\boldsymbol{H}_k$ to $\boldsymbol{H}_{k+1}$ by adding a rank one matrix $a_k \boldsymbol{z}^{(k)} \boldsymbol{z}^{(k)\top}$ for some $a_k \in \mathbb{R}$ and $\boldsymbol{z}^{(k)} \in \mathbb{R}^n$:

$$\boldsymbol{H}_{k+1} = \boldsymbol{H}_k + a_k \boldsymbol{z}^{(k)} \boldsymbol{z}^{(k)\top}$$

Now let us derive what this $a_k \boldsymbol{z}^{(k)} \boldsymbol{z}^{(k)\top}$ should be.

Since we need $\boldsymbol{H}_{k+1} \triangle \boldsymbol{g}^{(i)} = \triangle \boldsymbol{x}^{(i)}$ for $i \le k$, we at least need $\boldsymbol{H}_{k+1} \triangle \boldsymbol{g}^{(k)} = \triangle \boldsymbol{x}^{(k)}$. That is

$$\begin{aligned}
\triangle \boldsymbol{x}^{(k)} &= \boldsymbol{H}_{k+1} \triangle \boldsymbol{g}^{(k)} \\
&= (\boldsymbol{H}_k + a_k \boldsymbol{z}^{(k)} \boldsymbol{z}^{(k)\top}) \triangle \boldsymbol{g}^{(k)} \\
&= \boldsymbol{H}_k \triangle \boldsymbol{g}^{(k)} + a_k (\boldsymbol{z}^{(k)\top} \triangle \boldsymbol{g}^{(k)}) \boldsymbol{z}^{(k)}
\end{aligned}$$

Therefore

$$\boldsymbol{z}^{(k)} = \frac{\Delta \boldsymbol{x}^{(k)} - \boldsymbol{H}_k \Delta \boldsymbol{g}^{(k)}}{a_k(\boldsymbol{z}^{(k)}{}^\top \Delta \boldsymbol{g}^{(k)})}$$

and hence

$$\boldsymbol{H}_{k+1} = \boldsymbol{H}_k + \frac{(\Delta \boldsymbol{x}^{(k)} - \boldsymbol{H}_k \Delta \boldsymbol{g}^{(k)})(\Delta \boldsymbol{x}^{(k)} - \boldsymbol{H}_k \Delta \boldsymbol{g}^{(k)})^\top}{a_k(\boldsymbol{z}^{(k)}{}^\top \Delta \boldsymbol{g}^{(k)})^2}$$

On the other hand, multiplying $\Delta \boldsymbol{g}^{(k)}{}^\top$ on both sides of $\Delta \boldsymbol{x}^{(k)} - \boldsymbol{H}_k \boldsymbol{g}^{(k)} = a_k(\boldsymbol{z}^{(k)}{}^\top \Delta \boldsymbol{g}^{(k)}) \boldsymbol{z}^{(k)}$, we obtain

$$\Delta \boldsymbol{g}^{(k)}{}^\top (\Delta \boldsymbol{x}^{(k)} - \boldsymbol{H}_k \Delta \boldsymbol{g}^{(k)}) = a_k(\boldsymbol{z}^{(k)}{}^\top \Delta \boldsymbol{g}^{(k)})^2.$$

Hence

$$\boldsymbol{H}_{k+1} = \boldsymbol{H}_k + \frac{(\Delta \boldsymbol{x}^{(k)} - \boldsymbol{H}_k \Delta \boldsymbol{g}^{(k)})(\Delta \boldsymbol{x}^{(k)} - \boldsymbol{H}_k \Delta \boldsymbol{g}^{(k)})^\top}{\Delta \boldsymbol{g}^{(k)}{}^\top (\Delta \boldsymbol{x}^{(k)} - \boldsymbol{H}_k \Delta \boldsymbol{g}^{(k)})}$$

This is the **rank one correction** formula.

We obtained the formula by requiring $\boldsymbol{H}_{k+1}\Delta\boldsymbol{g}^{(k)} = \Delta\boldsymbol{x}^{(k)}$. However, we also need $\boldsymbol{H}_{k+1}\Delta\boldsymbol{g}^{(i)} = \Delta\boldsymbol{x}^{(i)}$ for $i < k$. This turns out to be true automatically:

Theorem. For the rank one algorithm applied to quadratic functions with Hessian symmetric $\boldsymbol{Q}$, there are

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

for $k = 0, 1, \ldots, n-1$.

Proof. We have showed $\boldsymbol{H}_{k+1}\Delta\boldsymbol{g}^{(k)} = \Delta\boldsymbol{x}^{(k)}$ for all $k = 0, 1, 2, \cdots$. Assume the identities hold up to k, we use induction to show it's true for $k+1$. We here only need to show $\boldsymbol{H}_{k+1}\Delta\boldsymbol{g}^{(i)} = \Delta\boldsymbol{x}^{(i)}$ for $i < k$:

$$\begin{aligned}\boldsymbol{H}_{k+1}\Delta\boldsymbol{g}^{(i)} &= \left(\boldsymbol{H}_k + \frac{(\Delta\boldsymbol{x}^{(k)} - \boldsymbol{H}_k\Delta\boldsymbol{g}^{(k)})(\Delta\boldsymbol{x}^{(k)} - \boldsymbol{H}_k\Delta\boldsymbol{g}^{(k)})^\top}{\Delta\boldsymbol{g}^{(k)\top}(\Delta\boldsymbol{x}^{(k)} - \boldsymbol{H}_k\Delta\boldsymbol{g}^{(k)})}\right)\Delta\boldsymbol{g}^{(i)} \\ &= \Delta\boldsymbol{x}^{(i)} + \frac{(\Delta\boldsymbol{x}^{(k)} - \boldsymbol{H}_k\Delta\boldsymbol{g}^{(k)})(\Delta\boldsymbol{x}^{(k)} - \boldsymbol{H}_k\Delta\boldsymbol{g}^{(k)})^\top\Delta\boldsymbol{g}^{(i)}}{\Delta\boldsymbol{g}^{(k)\top}(\Delta\boldsymbol{x}^{(k)} - \boldsymbol{H}_k\Delta\boldsymbol{g}^{(k)})}\end{aligned}$$

Note that

$$\begin{aligned}(\boldsymbol{H}_k\Delta\boldsymbol{g}^{(k)})^\top\Delta\boldsymbol{g}^{(i)} &= \Delta\boldsymbol{g}^{(k)\top}\boldsymbol{H}_k\Delta\boldsymbol{g}^{(i)} = \Delta\boldsymbol{g}^{(k)\top}\Delta\boldsymbol{x}^{(i)} \\ &= \Delta\boldsymbol{x}^{(k)\top}\boldsymbol{Q}\Delta\boldsymbol{x}^{(i)} = \Delta\boldsymbol{x}^{(k)\top}\Delta\boldsymbol{g}^{(i)}\end{aligned}$$

Hence the second term on the right is zero, and we obtain

$$\boldsymbol{H}_k\Delta\boldsymbol{g}^{(i)} = \Delta\boldsymbol{x}^{(i)}$$

This completes the proof.

Issues with rank one correction formula:

- $\boldsymbol{H}_{k+1}$ may not be positive definite even if $\boldsymbol{H}_k$ is. Hence $-\boldsymbol{H}_k \boldsymbol{g}^{(k)}$ may not be a descent direction;
- the denominator in the rank one correction is $\Delta \boldsymbol{g}^{(k)\top}(\Delta \boldsymbol{x}^{(k)} - \boldsymbol{H}_k \Delta \boldsymbol{g}^{(k)})$, which can be close to 0 and makes computation unstable.

We now study the DFP algorithm which improves the rank one correction formula by ensuring positive definiteness of $\boldsymbol{H}_k$.

DFP algoirthm [Davidson 1959, Fletcher and Powell 1963]

$$\boldsymbol{H}_{k+1} = \boldsymbol{H}_k + \frac{\Delta \boldsymbol{x}^{(k)} \Delta \boldsymbol{x}^{(k)\top}}{\Delta \boldsymbol{x}^{(k)\top} \Delta \boldsymbol{g}^{(k)}} - \frac{(\boldsymbol{H}_k \Delta \boldsymbol{g}^{(k)})(\boldsymbol{H}_k \Delta \boldsymbol{g}^{(k)})^\top}{\Delta \boldsymbol{g}^{(k)\top} \boldsymbol{H}_k \Delta \boldsymbol{g}^{(k)}}$$

We first show that DFP is a quasi-Newton method.

Theorem. The DFP algorithm applied to quadratic functions satisfies

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

for all k.

Proof. We prove this by induction. It is trivial for $k = 0$.

Assume the claim is true for k, i.e., $\boldsymbol{H}_k \Delta \boldsymbol{g}^{(i)} = \Delta \boldsymbol{x}^{(i)}$ for all $i \le k - 1$.

Now we first have $\boldsymbol{H}_{k+1} \Delta \boldsymbol{g}^{(i)} = \Delta \boldsymbol{x}^{(i)}$ for $i = k$ by direct computation. For $i < k$, there is

$$\begin{aligned} \boldsymbol{H}_{k+1} \Delta \boldsymbol{g}^{(i)} = \boldsymbol{H}_k \Delta \boldsymbol{g}^{(i)} &+ \frac{\Delta \boldsymbol{x}^{(k)} (\Delta \boldsymbol{x}^{(k)\top} \Delta \boldsymbol{g}^{(i)})}{\Delta \boldsymbol{x}^{(k)\top} \Delta \boldsymbol{g}^{(k)}} \\ &- \frac{(\boldsymbol{H}_k \Delta \boldsymbol{g}^{(k)})(\boldsymbol{H}_k \Delta \boldsymbol{g}^{(k)})^\top \Delta \boldsymbol{g}^{(i)}}{\Delta \boldsymbol{g}^{(k)\top} \boldsymbol{H}_k \Delta \boldsymbol{g}^{(k)}} \end{aligned}$$

Note that due to assumption $\boldsymbol{d}^{(0)}, \dots, \boldsymbol{d}^{(k)}$ are $\boldsymbol{Q}$-conjugate, and hence

$$\Delta \boldsymbol{x}^{(k)\top} \Delta \boldsymbol{g}^{(i)} = \Delta \boldsymbol{x}^{(k)\top} \boldsymbol{Q} \Delta \boldsymbol{x}^{(i)} = \alpha_k \alpha_i \boldsymbol{d}^{(k)\top} \boldsymbol{Q} \boldsymbol{d}^{(i)} = 0$$

similarly $\Delta \boldsymbol{g}^{(k)\top} \boldsymbol{H}_k \Delta \boldsymbol{g}^{(i)} = \Delta \boldsymbol{g}^{(k)\top} \Delta \boldsymbol{x}^{(i)} = 0$. This completes the proof.

Next we show that $\boldsymbol{H}_{k+1}$ inherits positive definiteness of $\boldsymbol{H}_k$ in DFP algorithm.

Theorem. Suppose $\boldsymbol{g}^{(k)} \neq \boldsymbol{0}$, then $\boldsymbol{H}_k \succ \boldsymbol{0}$ implies $\boldsymbol{H}_{k+1} \succ \boldsymbol{0}$ in DFP.

Proof. For any $\boldsymbol{x} \in \mathbb{R}^n$, there is

$$\boldsymbol{x}^\top \boldsymbol{H}_{k+1} \boldsymbol{x} = \boldsymbol{x}^\top \boldsymbol{H}_k \boldsymbol{x} + \frac{(\boldsymbol{x}^\top \Delta \boldsymbol{x}^{(k)})^2}{\Delta \boldsymbol{x}^{(k)\top} \Delta \boldsymbol{g}^{(k)}} - \frac{(\boldsymbol{x}^\top \boldsymbol{H}_k \Delta \boldsymbol{g}^{(k)})^2}{\Delta \boldsymbol{g}^{(k)\top} \boldsymbol{H}_k \Delta \boldsymbol{g}^{(k)}}$$

For notation simplicity, we denote

$$\boldsymbol{a} = \boldsymbol{H}_k^{1/2} \boldsymbol{x} \quad \text{and} \quad \boldsymbol{b} = \boldsymbol{H}_k^{1/2} \Delta \boldsymbol{g}^{(k)}$$

where $\boldsymbol{H}_k = \boldsymbol{H}_k^{1/2} \boldsymbol{H}_k^{1/2}$ (we know $\boldsymbol{H}_k^{1/2}$ exists since $\boldsymbol{H}_k$ is SPD).

Proof (cont). Now we have

$$\boldsymbol{x}^\top \boldsymbol{H}_{k+1}\boldsymbol{x} = \frac{\|\boldsymbol{a}\|^2\|\boldsymbol{b}\|^2 - (\boldsymbol{a}^\top \boldsymbol{b})^2}{\|\boldsymbol{b}\|^2} + \frac{(\boldsymbol{x}^\top \Delta \boldsymbol{x}^{(k)})^2}{\Delta \boldsymbol{x}^{(k)^\top} \Delta \boldsymbol{g}^{(k)}}$$

Note also that $\Delta \boldsymbol{x}^{(k)} = \alpha_k \boldsymbol{d}^{(k)} = -\alpha_k \boldsymbol{H}_k \boldsymbol{g}^{(k)}$, therefore

$$\Delta \boldsymbol{x}^{(k)^\top} \Delta \boldsymbol{g}^{(k)} = \Delta \boldsymbol{x}^{(k)^\top} (\boldsymbol{g}^{(k+1)} - \boldsymbol{g}^{(k)}) = -\Delta \boldsymbol{x}^{(k)^\top} \boldsymbol{g}^{(k)} = \alpha_k \boldsymbol{g}^{(k)^\top} \boldsymbol{H}_k \boldsymbol{g}^{(k)}$$

where we used $\boldsymbol{d}^{(k)^\top} \boldsymbol{g}^{(k+1)} = 0$ due to $\boldsymbol{Q}$-conjugacy of $\boldsymbol{d}^{(k)}$ in the second equality. Hence $\boldsymbol{x}^\top \boldsymbol{H}_{k+1} \boldsymbol{x} \geq 0$ since both terms on the right side are nonnegative.

Proof (cont). Now we need to show that the two terms cannot be 0 simultaneously.

Suppose the first term is 0, then $\boldsymbol{a} = \beta \boldsymbol{b}$ for some scalar $\beta > 0$. That is $\boldsymbol{H}_k^{1/2}\boldsymbol{x} = \beta \boldsymbol{H}_k^{1/2}\Delta \boldsymbol{g}^{(k)}$, or $\boldsymbol{x} = \beta \Delta \boldsymbol{g}^{(k)}$.

In this case, there is

$$\begin{aligned}(\boldsymbol{x}^\top \Delta \boldsymbol{x}^{(k)})^2 &= (\beta \Delta \boldsymbol{g}^{(k)\top} \Delta \boldsymbol{x}^{(k)})^2 = \alpha_k^2 \beta^2 (\Delta \boldsymbol{g}^{(k)\top} \boldsymbol{d}^{(k)})^2 \\ &= \alpha_k^2 \beta^2 (\boldsymbol{g}^{(k)\top} \boldsymbol{d}^{(k)})^2 = (\alpha_k \beta)^2 (\boldsymbol{g}^{(k)\top} \boldsymbol{H}_k \boldsymbol{g}^{(k)})^2 > 0\end{aligned}$$

and hence the second term is positive.

This completes the proof.

BFGS algorithm (named after Broyden, Fletcher, Goldfarb, Shannon)

Instead of directly finding $\boldsymbol{H}_k$ such that $\boldsymbol{H}_{k+1}\Delta\boldsymbol{g}^{(i)} = \Delta\boldsymbol{x}^{(i)}$ for $0 \le i \le k$, the BFGS first find $\boldsymbol{B}_k$ such that

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

Then replacing $\boldsymbol{H}_k$ by $\boldsymbol{B}_k$ and swapping $\Delta\boldsymbol{x}^{(k)}$ and $\Delta\boldsymbol{g}^{(k)}$ in DFP yield

$$\boldsymbol{B}_{k+1} = \boldsymbol{B}_k + \frac{\Delta\boldsymbol{g}^{(k)}\Delta\boldsymbol{g}^{(k)\top}}{\Delta\boldsymbol{x}^{(k)\top}\Delta\boldsymbol{g}^{(k)}} - \frac{(\boldsymbol{B}_k\Delta\boldsymbol{x}^{(k)})(\boldsymbol{B}_k\Delta\boldsymbol{x}^{(k)})^\top}{\Delta\boldsymbol{x}^{(k)\top}\boldsymbol{B}_k\Delta\boldsymbol{x}^{(k)}}$$

Then the actual $\boldsymbol{H}_k = \boldsymbol{B}_k{}^{-1}$ and hence

$$\begin{aligned}
\boldsymbol{H}_{k+1} &= \Big(\boldsymbol{B}_k + \frac{\Delta \boldsymbol{g}^{(k)} \Delta \boldsymbol{g}^{(k)\top}}{\Delta \boldsymbol{x}^{(k)\top} \Delta \boldsymbol{g}^{(k)}} - \frac{(\boldsymbol{B}_k \Delta \boldsymbol{x}^{(k)})(\boldsymbol{B}_k \Delta \boldsymbol{x}^{(k)})^\top}{\Delta \boldsymbol{x}^{(k)\top} \boldsymbol{B}_k \Delta \boldsymbol{x}^{(k)}}\Big)^{-1} \\
&= \boldsymbol{H}_k + \Big(1 + \frac{\Delta \boldsymbol{g}^{(k)\top} \boldsymbol{H}_k \Delta \boldsymbol{g}^{(k)}}{\Delta \boldsymbol{g}^{(k)\top} \Delta \boldsymbol{x}^{(k)}}\Big) \frac{\Delta \boldsymbol{x}^{(k)} \Delta \boldsymbol{x}^{(k)\top}}{\Delta \boldsymbol{x}^{(k)\top} \Delta \boldsymbol{g}^{(k)}} \\
&\quad - \frac{\boldsymbol{H}_k \Delta \boldsymbol{g}^{(k)} \Delta \boldsymbol{x}^{(k)\top} + (\boldsymbol{H}_k \Delta \boldsymbol{g}^{(k)} \Delta \boldsymbol{x}^{(k)\top})^\top}{\Delta \boldsymbol{g}^{(k)\top} \Delta \boldsymbol{x}^{(k)}}
\end{aligned}$$

This is the update rule of $\boldsymbol{H}_k$ in BFGS algorithm

The inverse was obtained by applying the following result:

Lemma. [Sherman-Morrison formula] Let $\boldsymbol{A}$ be a nonsingular matrix, and $\boldsymbol{u}$ and $\boldsymbol{v}$ are column vectors such that $1 + \boldsymbol{v}^\top \boldsymbol{A}^{-1}\boldsymbol{u} \neq 0$, then $\boldsymbol{A} + \boldsymbol{u}\boldsymbol{v}^\top$ is nonsingular, and

$$(\boldsymbol{A} + \boldsymbol{u}\boldsymbol{v}^\top)^{-1} = \boldsymbol{A}^{-1} - \frac{(\boldsymbol{A}^{-1}\boldsymbol{u})(\boldsymbol{v}^\top \boldsymbol{A}^{-1})}{1 + \boldsymbol{v}^\top \boldsymbol{A}^{-1}\boldsymbol{u}}$$

Proof. Direct computation.

BFGS algorithm:

1. Set $k = 0$; select $\boldsymbol{x}^{(0)}$ and SPD $\boldsymbol{H}_0$, and compute $\boldsymbol{g}^{(0)} = \nabla f(\boldsymbol{x}^{(0)})$.

2. Repeat:

$$\begin{aligned}
\boldsymbol{d}^{(k)} &= -\boldsymbol{H}_k \boldsymbol{g}^{(k)} \\
\alpha_k &= \operatorname*{arg\,min}_{\alpha \geq 0} f(\boldsymbol{x}^{(k)} + \alpha \boldsymbol{d}^{(k)}) \\
\boldsymbol{x}^{(k+1)} &= \boldsymbol{x}^{(k)} + \alpha_k \boldsymbol{d}^{(k)} \\
\boldsymbol{g}^{(k+1)} &= \nabla f(\boldsymbol{x}^{(k+1)}) \\
\boldsymbol{H}_{k+1} &= \boldsymbol{H}_k + \cdots \quad \text{(Compute the BFGS update of } \boldsymbol{H}_k\text{)} \\
k &\leftarrow k + 1
\end{aligned}$$

   Until $\boldsymbol{g}^{(k)} = \boldsymbol{0}$.