# DATA8004: Optimization for Statistical Learning

Subject Lecturer: Man-Chung YUE

Lecture 3
Unconstrained Optimization
Newton / Quasi-Newton methods

# Newton's method revisited

Can Theorem 2.5 be applied to guarantee convergence of the use of Newton direction $d^k = -[\nabla^2 f(x^k)]^{-1} \nabla f(x^k)$?

# Newton's method revisited

Can Theorem 2.5 be applied to guarantee convergence of the use of Newton direction $d^k = -[\nabla^2 f(x^k)]^{-1} \nabla f(x^k)$?

Problematic issues:

- $[\nabla^2 f(x^k)]^{-1}$ may not exist!
- $[\nabla^2 f(x^k)]^{-1}$ may not be (uniformly) positive definite!

# Newton's method revisited

Can Theorem 2.5 be applied to guarantee convergence of the use of Newton direction $d^k = -[\nabla^2 f(x^k)]^{-1}\nabla f(x^k)$?

Problematic issues:

- $[\nabla^2 f(x^k)]^{-1}$ may not exist!
- $[\nabla^2 f(x^k)]^{-1}$ may not be (uniformly) positive definite!

Modify Newton direction?

**Theorem 3.1**:
Let $A \in \mathbb{R}^{n\times n}$ be symmetric and $A = UDU^T$ be its eigenvalue decomposition. Define

$$A_+ := UD_+U^T,$$

where $D_+$ is the diagonal matrix with $(d_+)_{ii} = \max\{d_{ii}, 0\}$ for all $i$. Then $A_+$ is the unique solution of

$$\text{Minimize } \|Y - A\|_F \text{ subject to } Y \succeq 0.$$

# Examples: Projection onto $S_+^n$

1. Let $A = \begin{bmatrix} 3 & 0 \\ 0 & -1 \end{bmatrix}$. Then $A_+ = \begin{bmatrix} 3 & 0 \\ 0 & 0 \end{bmatrix}$.

# Examples: Projection onto $S^n_+$

1. Let $A = \begin{bmatrix} 3 & 0 \\ 0 & -1 \end{bmatrix}$. Then $A_+ = \begin{bmatrix} 3 & 0 \\ 0 & 0 \end{bmatrix}$.
2. If

$$A = \begin{bmatrix} 3 & -2 & 4 \\ -2 & 6 & 2 \\ 4 & 2 & 3 \end{bmatrix},$$

then $A = UDU^T$ with

$$U = \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{18}} & -\frac{2}{3} \\ 0 & \frac{4}{\sqrt{18}} & -\frac{1}{3} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{18}} & \frac{2}{3} \end{bmatrix}, \quad D = \begin{bmatrix} 7 & 0 & 0 \\ 0 & 7 & 0 \\ 0 & 0 & -2 \end{bmatrix}.$$

Hence,

$$A_+ = U \begin{bmatrix} 7 & 0 & 0 \\ 0 & 7 & 0 \\ 0 & 0 & 0 \end{bmatrix} U^T.$$

# Truncated Newton's method

For $f \in C^2(\mathbb{R}^n)$, we may consider the truncated Newton's method.

1. Pick $\sigma \in (0,1)$, $\beta \in (0,1)$, $\bar{\alpha}_k \equiv 1$, a small $\eta > 0$ and a huge $M > 0$. Initialize at $x^0 \in \mathbb{R}^n$.
2. For $k = 0, 1, 2, \ldots$, let $UDU^T$ be an eigenvalue decomposition of $\nabla^2 f(x^k)$. Let $\Lambda$ be diagonal with $\lambda_{ii} = \max\{\min\{M, d_{ii}\}, \eta\}$.
   Set $D_k := U\Lambda^{-1}U^T$ and $d^k := -D_k \nabla f(x^k)$. Update
   $$x^{k+1} = x^k + \alpha_k d^k,$$
   where $\alpha_k$ is obtained via the Armijo line search by backtracking.

**Theorem 3.2.**
Let $f \in C^2(\mathbb{R}^n)$ with $\inf f > -\infty$ and let $\{x^k\}$ be generated by the truncated Newton's method. Then any accumulation point of $\{x^k\}$ is a stationary point of $f$.

# Truncated Newton's method cont.

Computational concerns:

- In case $\{\nabla^2 f(x^k)\}$ is uniformly positive definite and bounded so that Newton directions are actually used, one should choose $\sigma \in (0, \frac{1}{2})$: For each nonstationary $x$ and $d = -[\nabla^2 f(x)]^{-1}\nabla f(x)$,

$$\begin{aligned} f(x+d) &= f(x) + \nabla f(x)^T d + \tfrac{1}{2} d^T \nabla^2 f(x) d + \underbrace{\tfrac{1}{2} d^T [\nabla^2 f(\xi) - \nabla^2 f(x)] d}_{\Delta_d} \\ &= f(x) + \tfrac{1}{2} \nabla f(x)^T d + \Delta_d \\ &= f(x) + \sigma \nabla f(x)^T d + \underbrace{(\tfrac{1}{2} - \sigma) \nabla f(x)^T d + \Delta_d}_{<\ 0 \text{ when } \|d\| \text{ is small}} . \end{aligned}$$

# Truncated Newton's method cont.

Computational concerns:

- In case $\{\nabla^2 f(x^k)\}$ is uniformly positive definite and bounded so that Newton directions are actually used, one should choose $\sigma \in (0, \frac{1}{2})$: For each nonstationary $x$ and $d = -[\nabla^2 f(x)]^{-1} \nabla f(x)$,

$$
\begin{aligned}
f(x+d) &= f(x) + \nabla f(x)^T d + \tfrac{1}{2} d^T \nabla^2 f(x) d + \underbrace{\tfrac{1}{2} d^T [\nabla^2 f(\xi) - \nabla^2 f(x)] d}_{\Delta_d} \\
&= f(x) + \tfrac{1}{2} \nabla f(x)^T d + \Delta_d \\
&= f(x) + \sigma \nabla f(x)^T d + \underbrace{(\tfrac{1}{2} - \sigma) \nabla f(x)^T d + \Delta_d}_{<\ 0 \text{ when } \|d\| \text{ is small}} .
\end{aligned}
$$

- Obtaining the $\nabla^2 f(x)$ can be difficult when $n$ is large. In addition, eigenvalue decomposition takes $O(n^3)$ flops, which can be prohibitively expensive when $n$ is huge.

# Truncated Newton's method cont.

Computational concerns:

- In case $\{\nabla^2 f(x^k)\}$ is uniformly positive definite and bounded so that Newton directions are actually used, one should choose $\sigma \in (0, \frac{1}{2})$: For each nonstationary $x$ and $d = -[\nabla^2 f(x)]^{-1} \nabla f(x)$,

$$\begin{aligned} f(x+d) &= f(x) + \nabla f(x)^T d + \tfrac{1}{2} d^T \nabla^2 f(x) d + \underbrace{\tfrac{1}{2} d^T [\nabla^2 f(\xi) - \nabla^2 f(x)] d}_{\Delta_d} \\ &= f(x) + \tfrac{1}{2} \nabla f(x)^T d + \Delta_d \\ &= f(x) + \sigma \nabla f(x)^T d + \underbrace{(\tfrac{1}{2} - \sigma) \nabla f(x)^T d + \Delta_d}_{< \ 0 \text{ when } \|d\| \text{ is small}} . \end{aligned}$$

- Obtaining the $\nabla^2 f(x)$ can be difficult when $n$ is large. In addition, eigenvalue decomposition takes $O(n^3)$ flops, which can be prohibitively expensive when $n$ is huge.
- Secant methods?

# Secant method

To solve $g(x) = 0$, where $g \in C^1(\mathbb{R})$:

Idea: Use finite difference to approximate $g'$ in Newton's method, i.e.,

$$x_{k+1} = x_k - g(x_k)\frac{x_k - x_{k-1}}{g(x_k) - g(x_{k-1})},$$

initialized at $x_0$ and $x_{-1}$ with $g(x_0) \neq g(x_{-1})$.

Note:

- The local convergence rate of the secant method is typically slower than Newton's method. However, the computational cost per iteration can be smaller when $g'$ is hard to compute compared with $g$.

# Example

Example: Find the square root of 2 using the secant method, starting at $x_{-1} = 1.4$ and $x_0 = 1.5$, up to 4 decimal places.

# Example

Example: Find the square root of 2 using the secant method, starting at $x_{-1} = 1.4$ and $x_0 = 1.5$, up to 4 decimal places.

Solution: Consider $g(x) = x^2 - 2$. The iterates of the secant method are

$$x_{k+1} = x_k - (x_k^2 - 2)\frac{x_k - x_{k-1}}{x_k^2 - x_{k-1}^2} = x_k - \frac{x_k^2 - 2}{x_k + x_{k-1}}.$$

# Example

Example: Find the square root of 2 using the secant method, starting at $x_{-1} = 1.4$ and $x_0 = 1.5$, up to 4 decimal places.

Solution: Consider $g(x) = x^2 - 2$. The iterates of the secant method are

$$x_{k+1} = x_k - (x_k^2 - 2)\frac{x_k - x_{k-1}}{x_k^2 - x_{k-1}^2} = x_k - \frac{x_k^2 - 2}{x_k + x_{k-1}}.$$

Starting at $x_0 = 1.5$ and $x_{-1} = 1.4$, we have (in 10 s.f.)

| | |
|---|---|
| $x_1$ | 1.413793103e+00 |
| $x_2$ | 1.414201183e+00 |
| $x_3$ | 1.414213564e+00 |
| $x_4$ | 1.414213562e+00 |
| $x_5$ | 1.414213562e+00 |

Thus, $x_* = 1.4142$, rounded to the nearest 4 decimal places.

The iterations in red are not needed in the answer.

# Secant equations

Idea: Let $f \in C^2(\mathbb{R}^n)$. Given $x^{k+1}$ and $x^k$, we would expect

$$\nabla^2 f(x^{k+1})(x^{k+1} - x^k) \approx \nabla f(x^{k+1}) - \nabla f(x^k).$$

# Secant equations

Idea: Let $f \in C^2(\mathbb{R}^n)$. Given $x^{k+1}$ and $x^k$, we would expect

$$\nabla^2 f(x^{k+1})(x^{k+1} - x^k) \approx \nabla f(x^{k+1}) - \nabla f(x^k).$$

Notation: $s^k := x^{k+1} - x^k$, $y^k := \nabla f(x^{k+1}) - \nabla f(x^k)$.

This motivates us to successively construct $B_{k+1}$ (resp., $H_{k+1}$) to approximate $\nabla^2 f(x^{k+1})$ (resp., $[\nabla^2 f(x^{k+1})]^{-1}$) so that

$$B_{k+1}s^k = y^k \quad \text{(resp., } H_{k+1}y^k = s^k\text{)}.$$

We refer to these equations as secant equations.

# Popular update formulae

Initialize $B_0$ (or $H_0$) at a positive definite matrix.

| Method | $B_{k+1} =$ | $H_{k+1} =$ |
|---|---|---|
| **DFP** | $\left(I - \frac{y^k {s^k}^T}{{y^k}^T s^k}\right) B_k \left(I - \frac{s^k {y^k}^T}{{y^k}^T s^k}\right) + \frac{y^k {y^k}^T}{{y^k}^T s^k}$ | $H_k + \frac{s^k {s^k}^T}{{y^k}^T s^k} - \frac{H_k y^k {y^k}^T H_k}{{y^k}^T H_k y^k}$ |
| **BFGS** | $B_k + \frac{y^k {y^k}^T}{{y^k}^T s^k} - \frac{B_k s^k {s^k}^T B_k}{{s^k}^T B_k s^k}$ | $\left(I - \frac{s^k {y^k}^T}{{y^k}^T s^k}\right) H_k \left(I - \frac{y^k {s^k}^T}{{y^k}^T s^k}\right) + \frac{s^k {s^k}^T}{{y^k}^T s^k}$ |
| **SR1** | $B_k + \frac{(y^k - B_k s^k)(y^k - B_k s^k)^T}{(y^k - B_k s^k)^T s^k}$ | $H_k + \frac{(s^k - H_k y^k)(s^k - H_k y^k)^T}{(s^k - H_k y^k)^T y^k}$ |

Remark:

- DFP and BFGS are rank-2 updates, while SR1 is rank-1 update.
- Since $B_0$ and $H_0$ were symmetric to start with, by induction, all $B_k$ and $H_k$ are symmetric.
- In practice, BFGS usually performs better.

# Quasi-Newton method: Basic version

Given $f \in C^1(\mathbb{R}^n)$.

Quasi-Newton based on $B_k$:
Initialize at $x^0 \in \mathbb{R}^n$ and $B_0 \succ 0$.

For $k = 0, 1, 2, \ldots$

1. Find $d^k$ via $B_k d^k = -\nabla f(x^k)$.
2. Update $x^{k+1} = x^k + d^k$.
3. Set $y^k = \nabla f(x^{k+1}) - \nabla f(x^k)$ and $s^k = x^{k+1} - x^k$. Compute $B_{k+1}$.

Quasi-Newton based on $H_k$:
Initialize at $x^0 \in \mathbb{R}^n$ and $H_0 \succ 0$.

For $k = 0, 1, 2, \ldots$

1. Find $d^k$ via $d^k = -H_k \nabla f(x^k)$.
2. Update $x^{k+1} = x^k + d^k$.
3. Set $y^k = \nabla f(x^{k+1}) - \nabla f(x^k)$ and $s^k = x^{k+1} - x^k$. Compute $H_{k+1}$.

# Example 1

Example: Let $f(x) := x_1 x_2^2 + x_1^3 x_2 - x_1 x_2$. Perform 2 iterations of the BFGS method with $\alpha_k \equiv 1$, $x^0 = (1, -1)$ and $B_0 = I_2$. Write down $x^1$ and $x^2$. You may correct your answers to 4 d.p.

Solution: First, we note that

$$\nabla f(x) = \begin{bmatrix} x_2^2 + 3x_1^2 x_2 - x_2 \\ 2x_1 x_2 + x_1^3 - x_1 \end{bmatrix}.$$

Hence

$$\nabla f(x^0) = \begin{bmatrix} -1 \\ -2 \end{bmatrix} \quad \text{and} \quad B_0 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

# Example 1 cont.

Solution cont.: A direction computation then shows that

$$x^1 = x^0 - B_0^{-1}\nabla f(x^0) = \begin{bmatrix} 2 \\ 1 \end{bmatrix}.$$

Then

$$y^0 = \nabla f(x^1) - \nabla f(x^0) = \begin{bmatrix} 13 \\ 12 \end{bmatrix} \qquad s^0 = x^1 - x^0 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

$$B_1 = B_0 + \frac{y^0 {y^0}^T}{{y^0}^T s^0} - \frac{B_0 s^0 {s^0}^T B_0}{{s^0}^T B_0 s^0} = \begin{bmatrix} 5.3676 & 3.8162 \\ 3.8162 & 4.0919 \end{bmatrix}.$$

Then

$$x^2 = x^1 - B_1^{-1}\nabla f(x^1) = \begin{bmatrix} 0.5215 \\ -0.0650 \end{bmatrix}.$$

# Example 2

Example: Verify the secant equations for BFGS.

Solution: For the $B_{k+1}$, we have

$$B_{k+1}s^k = B_k s^k + \frac{y^k {y^k}^T s^k}{{y^k}^T s^k} - \frac{B_k s^k {s^k}^T B_k s^k}{{s^k}^T B_k s^k} = y^k.$$

For $H_{k+1}$, we have

$$\begin{aligned} H_{k+1}y^k &= \left(I - \frac{s^k {y^k}^T}{{y^k}^T s^k}\right) H_k \left(I - \frac{y^k {s^k}^T}{{y^k}^T s^k}\right) y^k + \frac{s^k {s^k}^T y^k}{{y^k}^T s^k} \\ &= \left(I - \frac{s^k {y^k}^T}{{y^k}^T s^k}\right) H_k \underbrace{\left(y^k - \frac{y^k {s^k}^T y^k}{{y^k}^T s^k}\right)}_{=0} + s^k = s^k. \end{aligned}$$

# Example 3

Example: Assuming that $H_k = B_k^{-1}$. Show that $H_{k+1} = B_{k+1}^{-1}$ for BFGS using the Sherman-Morrison-Woodbury formula:

$$(A + UCU^T)^{-1} = A^{-1} - A^{-1}U(C^{-1} + U^T A^{-1} U)^{-1} U^T A^{-1}.$$

Solution: First, rewrite $B_{k+1}$ as

$$B_k + \frac{y^k {y^k}^T}{{y^k}^T s^k} - \frac{B_k s^k {s^k}^T B_k}{{s^k}^T B_k s^k} = B_k + \begin{bmatrix} y^k & B_k s^k \end{bmatrix} \begin{bmatrix} \frac{1}{{y^k}^T s^k} & 0 \\ 0 & -\frac{1}{{s^k}^T B_k s^k} \end{bmatrix} \begin{bmatrix} y^k & B_k s^k \end{bmatrix}^T.$$

Now, apply the Sherman-Morrison-Woodbury formula with $A = B_k$,

$U = \begin{bmatrix} y^k & B_k s^k \end{bmatrix}$ and $C = \begin{bmatrix} \frac{1}{{y^k}^T s^k} & 0 \\ 0 & -\frac{1}{{s^k}^T B_k s^k} \end{bmatrix}$.

# Example 3 cont.

Solution cont.: We obtain, upon noting $H_k = B_k^{-1}$, that,

$$\left(B_k + \frac{y^k {y^k}^T}{{y^k}^T s^k} - \frac{B_k s^k {s^k}^T B_k}{{s^k}^T B_k s^k}\right)^{-1}$$

$$= H_k - H_k \begin{bmatrix} y^k & B_k s^k \end{bmatrix} \left( \begin{bmatrix} {y^k}^T s^k & 0 \\ 0 & -{s^k}^T B_k s^k \end{bmatrix} + \begin{bmatrix} y^k & B_k s^k \end{bmatrix}^T H_k \begin{bmatrix} y^k & B_k s^k \end{bmatrix} \right)^{-1} \begin{bmatrix} y^k & B_k s^k \end{bmatrix}^T H_k$$

$$= H_k - H_k \begin{bmatrix} y^k & B_k s^k \end{bmatrix} \left( \begin{bmatrix} {y^k}^T s^k & 0 \\ 0 & -{s^k}^T B_k s^k \end{bmatrix} + \begin{bmatrix} {y^k}^T H_k y^k & {y^k}^T H_k B_k s^k \\ {s^k}^T B_k H_k y^k & {s^k}^T B_k H_k B_k s^k \end{bmatrix} \right)^{-1} \begin{bmatrix} y^k & B_k s^k \end{bmatrix}^T H_k$$

$$= H_k - H_k \begin{bmatrix} y^k & B_k s^k \end{bmatrix} \left( \begin{bmatrix} {y^k}^T s^k & 0 \\ 0 & -{s^k}^T B_k s^k \end{bmatrix} + \begin{bmatrix} {y^k}^T H_k y^k & {y^k}^T s^k \\ {s^k}^T y^k & {s^k}^T B_k s^k \end{bmatrix} \right)^{-1} \begin{bmatrix} y^k & B_k s^k \end{bmatrix}^T H_k$$

$$= H_k - \begin{bmatrix} H_k y^k & s^k \end{bmatrix} \left( \begin{bmatrix} {y^k}^T H_k y^k + {y^k}^T s^k & {y^k}^T s^k \\ {s^k}^T y^k & 0 \end{bmatrix} \right)^{-1} \begin{bmatrix} H_k y^k & s^k \end{bmatrix}^T$$

$$= H_k + \frac{1}{({y^k}^T s^k)^2} \begin{bmatrix} H_k y^k & s^k \end{bmatrix} \begin{bmatrix} 0 & -{y^k}^T s^k \\ -{y^k}^T s^k & {y^k}^T H_k y^k + {y^k}^T s^k \end{bmatrix} \begin{bmatrix} H_k y^k & s^k \end{bmatrix}^T$$

# Example 3 cont.

Solution cont.: Continuing, we have

$$\left(B_k + \frac{y^k {y^k}^T}{{y^k}^T s^k} - \frac{B_k s^k {s^k}^T B_k}{{s^k}^T B_k s^k}\right)^{-1}$$

$$= H_k + \frac{1}{({y^k}^T s^k)^2} \begin{bmatrix} H_k y^k & s^k \end{bmatrix} \begin{bmatrix} 0 & -{y^k}^T s^k \\ -{y^k}^T s^k & {y^k}^T H_k y^k + {y^k}^T s^k \end{bmatrix} \begin{bmatrix} {y^k}^T H_k \\ {s^k}^T \end{bmatrix}$$

$$= H_k + \frac{1}{({y^k}^T s^k)^2} \begin{bmatrix} H_k y^k & s^k \end{bmatrix} \begin{bmatrix} -{y^k}^T s^k {s^k}^T \\ -({y^k}^T s^k){y^k}^T H_k + {y^k}^T H_k y^k {s^k}^T + {y^k}^T s^k {s^k}^T \end{bmatrix}$$

$$= H_k + \frac{1}{({y^k}^T s^k)^2} [s^k {y^k}^T H_k y^k {s^k}^T + s^k({y^k}^T s^k){s^k}^T - H_k y^k ({y^k}^T s^k){s^k}^T - s^k({y^k}^T s^k){y^k}^T H_k]$$

$$= \left(I - \frac{s^k {y^k}^T}{{y^k}^T s^k}\right) H_k \left(I - \frac{y^k {s^k}^T}{{y^k}^T s_k}\right) + \frac{s^k {s^k}^T}{{y^k}^T s^k}$$

Note: Thus, if $H_0 = B_0^{-1}$, theoretically, one can stick to $H_k$ and generate the same sequence as if $B_k$ were used.

# Computational concerns

From now on, we focus on BFGS:

- Updating $B_k$ (resp., $H_k$) takes $O(n^2)$ flops. If $B_k$ is used, one also needs to compute $d^k$ by solving the linear system
$$B_k d^k = -\nabla f(x^k),$$
which takes $O(n^3)$ flops. Thus, let's stick to $H_k$!
- To obtain some convergence guarantee, it is tempting to apply line search and Theorem 2.5. However, $d^k$ is not necessarily a descent direction! Indeed, $H_k$ may not be positive definite. Thus, $\nabla f(x^k)^T d^k = -\nabla f(x^k)^T H_k \nabla f(x^k)$ can be positive.
- One get-around is to shift back to use $-\nabla f(x^k)$ when $d^k$ is not a descent direction.
- Alternatively, we would like to find conditions to guarantee $H_k \succ 0$.

# $H_k \succ 0$?

**Proposition 3.1**

Let $H_k \succ 0$ and ${y^k}^T s^k > 0$. Let $H_{k+1}$ be given by BFGS update. Then $H_{k+1} \succ 0$. The same conclusion holds if $H_k$ and $H_{k+1}$ are replaced by $B_k$ and $B_{k+1}$, respectively.

Proof: Let $x \in \mathbb{R}^n$. Then we can write

$$x = \frac{x^T y^k}{{y^k}^T y^k} y^k + u$$

so that $y_k^T u = 0$. Then

$$x^T H_{k+1} x = x^T \left( I - \frac{s^k {y^k}^T}{{y^k}^T s^k} \right) H_k \left( I - \frac{y^k {s^k}^T}{{y^k}^T s^k} \right) x + x^T \frac{s^k {s^k}^T}{{y^k}^T s^k} x$$

$$= u^T \left( I - \frac{s^k {y^k}^T}{{y^k}^T s^k} \right) H_k \left( I - \frac{y^k {s^k}^T}{{y^k}^T s^k} \right) u + x^T \frac{s^k {s^k}^T}{{y^k}^T s^k} x$$

Since $H_k \succ 0$ and ${y^k}^T s^k > 0$, the above display is nonnegative. We need to show that it is zero only when $x = 0$.

# $H_k \succ 0$? cont.

Proof of Proposition 3.1 cont.: Now, suppose $x^T H_{k+1} x = 0$. Then

$$u^T \left( I - \frac{s^k {y^k}^T}{{y^k}^T s^k} \right) H_k \left( I - \frac{y^k {s^k}^T}{{y^k}^T s^k} \right) u = x^T \frac{s^k {s^k}^T}{{y^k}^T s^k} x = 0.$$

Since $H_k \succ 0$, we must then have $\left( I - \frac{y^k {s^k}^T}{{y^k}^T s^k} \right) u = 0$. Multiplying ${y^k}^T$ from the left and invoking ${y^k}^T u = 0$, we get ${s^k}^T u = 0$. Hence, $u = 0$ and we have $x = \frac{x^T y^k}{{y^k}^T y^k} y^k$. Then we see that

$$0 = x^T \frac{s^k {s^k}^T}{{y^k}^T s^k} x = \frac{x^T y^k}{{y^k}^T y^k} {y^k}^T \frac{s^k {s^k}^T}{{y^k}^T s^k} y^k \frac{x^T y^k}{{y^k}^T y^k} = \frac{(x^T y^k)^2 ({s^k}^T y^k)}{({y^k}^T y^k)^2}.$$

Thus, it holds that $x^T y^k = 0$. Consequently, $x = 0$.

# Wolfe conditions

In view of Proposition 3.1, it suffices to guarantee that $H_0 \succ 0$ and make sure that ${y^k}^T s^k > 0$ for each $k \geq 0$.

The latter can be guaranteed if line search is performed to guarantee the Wolfe conditions.

**Wolfe conditions**:
Let $0 < c_1 < c_2 < 1$, $x \in \mathbb{R}^n$ and $d \in \mathbb{R}^n$. Find $\alpha > 0$ so that

$$f(x + \alpha d) \leq f(x) + \alpha c_1 [\nabla f(x)]^T d,$$
$$-[\nabla f(x + \alpha d)]^T d \leq -c_2 [\nabla f(x)]^T d.$$

Remark:

- The first inequality in Wolfe conditions is the Armijo rule.
- The second relation is called curvature condition.

# Wolfe conditions cont.

**Theorem 3.3** (Wolfe conditions are not void)
Let $f \in C^1(\mathbb{R}^n)$ with $\inf f > -\infty$, $x \in \mathbb{R}^n$, and $d \in \mathbb{R}^n$ be a descent direction at $x$. Let $0 < c_1 < c_2 < 1$. Then there exists $\alpha > 0$ with

$$f(x + \alpha d) \leq f(x) + \alpha c_1 [\nabla f(x)]^T d,$$
$$-[\nabla f(x + \alpha d)]^T d \leq -c_2 [\nabla f(x)]^T d.$$

Proof: Since $[\nabla f(x)]^T d < 0$, we have $f(x + \alpha d) < f(x)$ for all sufficiently small $\alpha > 0$. Since $\inf f > -\infty$ and $c_1 \in (0, 1)$, there must be a smallest $\alpha_1 > 0$ so that $f(x + \alpha_1 d) = f(x) + \alpha_1 c_1 [\nabla f(x)]^T d$ and

$$f(x + \alpha d) \leq f(x) + \alpha c_1 [\nabla f(x)]^T d$$

whenever $\alpha \in [0, \alpha_1]$. Now, Taylor's theorem guarantees that there exists $\alpha' \in (0, \alpha_1)$ so that

$$f(x + \alpha_1 d) - f(x) = \alpha_1 [\nabla f(x + \alpha' d)]^T d.$$

Hence, $\alpha_1 [\nabla f(x + \alpha' d)]^T d = \alpha_1 c_1 [\nabla f(x)]^T d \geq \alpha_1 c_2 [\nabla f(x)]^T d$.

# Quasi-Newton method: Wolfe line search

Quasi-Newton using $H_k$ in BFGS for $f \in C^1(\mathbb{R}^n)$ with $\inf f > -\infty$:

Pick $0 < c_1 < c_2 < 1$, $x^0 \in \mathbb{R}^n$, $H_0 = \eta I$ for some $\eta > 0$.

For $k = 0, 1, 2, \dots$

1. Find $d^k$ via $d^k = -H_k \nabla f(x^k)$.
2. Compute $\alpha_k$ that satisfies the Wolfe conditions.
3. Update $x^{k+1} = x^k + \alpha_k d^k$.
4. Set $y^k = \nabla f(x^{k+1}) - \nabla f(x^k)$, $s^k = x^{k+1} - x^k$ and compute $H_{k+1}$ as in BFGS.

# Quasi-Newton method: Wolfe line search

Quasi-Newton using $H_k$ in BFGS for $f \in C^1(\mathbb{R}^n)$ with $\inf f > -\infty$:

Pick $0 < c_1 < c_2 < 1$, $x^0 \in \mathbb{R}^n$, $H_0 = \eta I$ for some $\eta > 0$.

For $k = 0, 1, 2, \ldots$

1. Find $d^k$ via $d^k = -H_k \nabla f(x^k)$.
2. Compute $\alpha_k$ that satisfies the Wolfe conditions.
3. Update $x^{k+1} = x^k + \alpha_k d^k$.
4. Set $y^k = \nabla f(x^{k+1}) - \nabla f(x^k)$, $s^k = x^{k+1} - x^k$ and compute $H_{k+1}$ as in BFGS.

Remark:

- If $x^k$ is not stationary, then
  $${y^k}^T s^k = \alpha_k (\nabla f(x^{k+1}) - \nabla f(x^k))^T d^k \geq \alpha_k (c_2 - 1) \nabla f(x^k)^T d^k > 0.$$
- Wolfe conditions cannot be satisfied by simply backtracking. One needs a special root-finding procedure. See §3.4 in Ref 4.

  Unlike Armijo line search by backtracking, it computes additional $\nabla f$ and is more expensive.

# Convergence under Wolfe conditions

**Theorem 3.4**: (Zoutendijk's theorem)
Let $f \in C^1(\mathbb{R}^n)$ with $\inf f > -\infty$, $x^0 \in \mathbb{R}^n$ and $\exists \ell > 0$ so that

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq \ell \|x - y\|_2$$

whenever $\max\{f(x), f(y)\} \leq f(x^0)$. Let $\{x^k\}$ be a sequence of non-stationary points generated as

$$x^{k+1} = x^k + \alpha_k d^k,$$

with $d^k$ being a descent direction and $\alpha_k$ satisfying the Wolfe conditions. Then it holds that

$$\sum_{k=0}^{\infty} \cos^2 \theta_k \|\nabla f(x^k)\|_2^2 < \infty,$$

where $\cos \theta_k := \frac{-[\nabla f(x^k)]^T d^k}{\|\nabla f(x^k)\|_2 \|d^k\|_2}$.

# Convergence under Wolfe conditions cont.

Proof of Theorem 3.4: Since $\{x^k\} \subset \{x : \; f(x) \le f(x^0)\}$ by the Armijo rule, we have

$$\|\nabla f(x^{k+1}) - \nabla f(x^k)\|_2 \le \ell \|x^{k+1} - x^k\|_2.$$

Combining this with the curvature condition, we have

$$\begin{aligned}(c_2 - 1)[\nabla f(x^k)]^T d^k &\le [\nabla f(x^{k+1}) - \nabla f(x^k)]^T d^k \\ &\le \ell \|x^{k+1} - x^k\|_2 \|d^k\|_2 = \ell \alpha_k \|d^k\|_2^2.\end{aligned}$$

Thus, we have a lower bound on $\alpha_k$:

$$\alpha_k \ge \frac{(c_2 - 1)[\nabla f(x^k)]^T d^k}{\ell \|d^k\|_2^2} > 0.$$

# Convergence under Wolfe conditions cont.

Proof of Theorem 3.4 cont.: Substituting the bound on $\{\alpha_k\}$ into Armijo rule, we obtain

$$
\begin{aligned}
f(x^{k+1}) &\leq f(x^k) - c_1 \frac{(1-c_2)([\nabla f(x^k)]^T d^k)^2}{\ell \|d^k\|^2} \\
&= f(x^k) - \frac{c_1(1-c_2)}{\ell} \cos^2 \theta_k \|\nabla f(x^k)\|_2^2.
\end{aligned}
$$

Rearranging terms and summing from $k = 0$ to $\infty$, we see that

$$
\frac{c_1(1-c_2)}{\ell} \sum_{k=0}^{\infty} \cos^2 \theta_k \|\nabla f(x^k)\|_2^2 \leq f(x^0) - \inf f < \infty.
$$

# Convergence under Wolfe conditions cont.

Remark: According to Theorem 3.4:

- If there exists $\delta > 0$ so that $\cos\theta_k \geq \delta$ for all $k$, then $\lim_{k\to\infty} \|\nabla f(x^k)\|_2 = 0$. Hence, any accumulation point of $\{x^k\}$ is stationary.
- For BFGS, if there exists $M > 0$ so that

$$\|H_k\|_2 \|H_k^{-1}\|_2 \leq M \quad \forall k,$$

then $\lim_{k\to\infty} \|\nabla f(x^k)\|_2 = 0$. Indeed, in this case,

$$\cos\theta_k = \frac{{d^k}^T H_k^{-1} d^k}{\|H_k^{-1} d^k\|_2 \|d^k\|_2} \geq \frac{{d^k}^T H_k^{-1} d^k}{\|H_k^{-1}\|_2 \|d^k\|_2^2} \geq \frac{\lambda_{\min}(H_k^{-1})}{\|H_k^{-1}\|_2}$$
$$= \frac{1}{\lambda_{\max}(H_k) \|H_k^{-1}\|_2} = \frac{1}{\|H_k^{-1}\|_2 \|H_k\|_2} \geq \frac{1}{M}.$$

See Ref 4 for more checkable conditions.

# Limited-memory BFGS

- If BFGS is used, it takes $O(n^2)$ of memory to store each $H_k$.

# Limited-memory BFGS

- If BFGS is used, it takes $O(n^2)$ of memory to store each $H_k$.
- Unfoiling a BFGS update backward by $m < k$ steps: Let $\rho_k := 1/({y^k}^T s^k)$ and $V_k := I - \rho_k y^k {s^k}^T$. Then

$$
\begin{aligned}
H_k &= V_{k-1}^T H_{k-1} V_{k-1} + \rho_{k-1} s^{k-1} {s^{k-1}}^T \\
&= V_{k-1}^T (V_{k-2}^T H_{k-2} V_{k-2} + \rho_{k-2} s^{k-2} {s^{k-2}}^T) V_{k-1} + \rho_{k-1} s^{k-1} {s^{k-1}}^T \\
&\vdots \\
&= (V_{k-1}^T \cdots V_{k-m}^T) H_k^0 (V_{k-m} \cdots V_{k-1}) \\
&\quad + \rho_{k-m} (V_{k-1}^T \cdots V_{k-m+1}^T) s^{k-m} {s^{k-m}}^T (V_{k-m+1} \cdots V_{k-1}) \\
&\quad + \cdots \\
&\quad + \rho_{k-1} s^{k-1} {s^{k-1}}^T.
\end{aligned}
$$

# Limited-memory BFGS

- If BFGS is used, it takes $O(n^2)$ of memory to store each $H_k$.
- Unfoiling a BFGS update backward by $m < k$ steps: Let $\rho_k := 1/({y^k}^T s^k)$ and $V_k := I - \rho_k y^k {s^k}^T$. Then

$$
\begin{aligned}
H_k &= V_{k-1}^T H_{k-1} V_{k-1} + \rho_{k-1} s^{k-1} {s^{k-1}}^T \\
&= V_{k-1}^T ( V_{k-2}^T H_{k-2} V_{k-2} + \rho_{k-2} s^{k-2} {s^{k-2}}^T ) V_{k-1} + \rho_{k-1} s^{k-1} {s^{k-1}}^T \\
&\;\;\vdots \\
&= (V_{k-1}^T \cdots V_{k-m}^T) H_k^0 (V_{k-m} \cdots V_{k-1}) \\
&\quad + \rho_{k-m} (V_{k-1}^T \cdots V_{k-m+1}^T) s^{k-m} {s^{k-m}}^T (V_{k-m+1} \cdots V_{k-1}) \\
&\quad + \cdots \\
&\quad + \rho_{k-1} s^{k-1} {s^{k-1}}^T.
\end{aligned}
$$

- Ideas:
  - ⋆ Keep $m$ small (Restart);
  - ⋆ Only need $-H_k \nabla f(x^k)$ — NEVER form $H_k$!

# Limited-memory BFGS cont.

- Choose $m$ moderately (usually 5 in practice).
- At iteration $k \geq m$, keep $\{s^{k-m}, \cdots, s^{k-1}\}$, $\{y^{k-m}, \cdots, y^{k-1}\}$ and $\{\rho_{k-m}, \cdots, \rho_{k-1}\}$ in the memory: $m(2n+1)$ numbers saved.
- To compute $H_k \nabla f(x^k)$ with the choice of $H_k^0$ and $m \leq k$.

**L-BFGS two-loop recursion**

Initialize with $q \leftarrow \nabla f(x^k)$.

1. For $i = k-1, k-2, \ldots, k-m$,
   Update $\alpha_i \leftarrow \rho_i {s^i}^T q$ and then $q \leftarrow q - \alpha_i y^i$.
2. Set $r = H_k^0 q$;
3. For $i = k-m, k-m+1, \ldots, k-1$,
   Update $\beta \leftarrow \rho_i {y^i}^T r$ and then $r \leftarrow r + (\alpha_i - \beta) s^i$.

Outputs $r = H_k \nabla f(x^k)$.

# Choice of $H_k^0$

- When $k > 1$, one can choose $H_k^0$ to be a multiple of identity that "best" verifies the secant equations.
- Based on $H_k^0 = \gamma_k I$ and $H_k^0 y^{k-1} \approx s^{k-1}$: This means $\gamma_k y^{k-1} \approx s^{k-1}$.
- This naturally gives rise to two possible ways of defining $\gamma_k$:

$$\gamma_k = \frac{{s^{k-1}}^T s^{k-1}}{{y^{k-1}}^T s^{k-1}} \quad \text{or} \quad \gamma_k = \frac{{s^{k-1}}^T y^{k-1}}{{y^{k-1}}^T y^{k-1}}.$$

# Choice of $H_k^0$

- When $k > 1$, one can choose $H_k^0$ to be a multiple of identity that "best" verifies the secant equations.
- Based on $H_k^0 = \gamma_k I$ and $H_k^0 y^{k-1} \approx s^{k-1}$: This means $\gamma_k y^{k-1} \approx s^{k-1}$.
- This naturally gives rise to two possible ways of defining $\gamma_k$:

$$\gamma_k = \frac{{s^{k-1}}^T s^{k-1}}{{y^{k-1}}^T s^{k-1}} \quad \text{or} \quad \gamma_k = \frac{{s^{k-1}}^T y^{k-1}}{{y^{k-1}}^T y^{k-1}}.$$

- (Digression) In the setting of Theorem 2.5 with $D^k \equiv I$ (steepest descent direction), one can choose

$$\bar{\alpha}_k = \max\{\min\{M, \gamma_k\}, \rho\}$$

for some $M \gg \rho > 0$. This is called the Barzilai-Borwein stepsize. Empirically, in many problems, the Armijo rule is usually satisfied without backtracking (or at most 1 or 2) when this $\bar{\alpha}_k$ is used.

# Example

Example: Consider the function $f(x) = \frac{1}{2}\|Ax - b\|_2^2 + \mu\|x\|_2^2$, where $A \in \mathbb{R}^{m\times n}$ ($m \ll n$), $b \in \mathbb{R}^m \backslash \{0\}$ and $\mu > 0$. Consider an iterate of the following form:

$$x^{k+1} = x^k + \alpha_k d^k.$$

1. Show that at any nonstationary point, the Newton direction $-[\nabla^2 f(x)]^{-1}\nabla f(x)$ is a descent direction.
2. Let $d^k$ be the Newton direction and $\alpha_k$ be chosen to satisfy the Wolfe's condition. Show that the sequence $\{x^k\}$ is bounded and any accumulation point is stationary.
3. Let $d^k$ be the Newton direction and $\alpha_k$ be chosen using Armijo line search by backtracking with $\bar{\alpha}_k \equiv 1$. Show that the sequence $\{x^k\}$ is bounded and any accumulation point is stationary.

# Example cont.

Remark: We first recall the following notation and properties: For $n \times n$ symmetric matrices $B$, $C$ and $D$,

- We write $C \succeq D$ to mean $C - D \succeq 0$.
- If $B \succeq C$ and $C \succeq D$, then $B \succeq D$. This is known as transitivity.
- If $B \succeq C$, then $\lambda_{\max}(B) \geq \lambda_{\max}(C)$ and $\lambda_{\min}(B) \geq \lambda_{\min}(C)$.

# Example cont.

Remark: We first recall the following notation and properties: For $n \times n$ symmetric matrices $B$, $C$ and $D$,

- We write $C \succeq D$ to mean $C - D \succeq 0$.
- If $B \succeq C$ and $C \succeq D$, then $B \succeq D$. This is known as transitivity.
- If $B \succeq C$, then $\lambda_{\max}(B) \geq \lambda_{\max}(C)$ and $\lambda_{\min}(B) \geq \lambda_{\min}(C)$.

Solution:

1. A direct computation shows that $\nabla^2 f(x) = A^T A + 2\mu I$. Thus, $\nabla^2 f(x) \succeq 2\mu I \succ 0$, meaning that $[\nabla^2 f(x)]^{-1} \succ 0$.

   Thus, at a nonstationary point (so that $\nabla f(x) \neq 0$), we have

   $$[\nabla f(x)]^T \left(-[\nabla^2 f(x)]^{-1} \nabla f(x)\right) = -[\nabla f(x)]^T [\nabla^2 f(x)]^{-1} \nabla f(x) < 0.$$

# Example cont.

Solution cont.:

2. If any $x^k$ is stationary, then $\nabla f(x^k) = 0$ and $x^l = x^k$ for all $l \geq k$.

   We now consider the case that $\{x^k\}$ is a sequence of nonstationary points. From the Armijo rule, we have for any $k \geq 1$ that

$$\begin{aligned}\mu\|x^k\|_2^2 \leq f(x^k) &\leq f(x^{k-1}) + c_1\alpha_{k-1}[\nabla f(x^{k-1})]^T d^{k-1} \\ &\leq f(x^{k-1}) \leq \cdots \leq f(x^0).\end{aligned}$$

   where the third inequality holds because Newton direction is a descent direction. Thus,

$$\|x^k\|_2 \leq \sqrt{f(x^0)/\mu} \quad \forall k \geq 1,$$

   meaning that $\{x^k\}$ is bounded.

# Example cont.

Solution cont.:

2. We next apply Zoutendijk's theorem.

   First, clearly, $f \in C^1(\mathbb{R}^n)$ and $\inf f \geq 0$.

   Also, $\nabla f(x) = A^T(Ax - b) + 2\mu x$. Thus,

$$\begin{aligned} \|\nabla f(x) - \nabla f(y)\|_2 &= \|A^T(Ax - Ay) + 2\mu(x - y)\|_2 \\ &\leq \|A^T(Ax - Ay)\|_2 + 2\mu\|x - y\|_2 \\ &\leq (\|A^T A\|_2 + 2\mu)\|x - y\|_2. \end{aligned}$$

   One can take $\ell = \|A^T A\|_2 + 2\mu$ in Zoutendijk's theorem. Since Newton direction is a descent direction and $\alpha_k$ satisfies the Wolfe conditions, we conclude that

$$\sum_{k=0}^{\infty} \cos^2\theta_k \|\nabla f(x^k)\|_2^2 < \infty. \tag{1}$$

# Example cont.

Solution cont.:

2. Moreover, notice that $\nabla^2 f(x) = A^T A + 2\mu I \succeq 2\mu I \succ 0$. Hence

$$
\begin{aligned}
\cos\theta_k &= \frac{[\nabla f(x^k)]^T [\nabla^2 f(x^k)]^{-1} \nabla f(x^k)}{\|\nabla f(x^k)\|_2 \|[\nabla^2 f(x^k)]^{-1} \nabla f(x^k)\|_2} \\
&\geq \frac{[\nabla f(x^k)]^T [\nabla^2 f(x^k)]^{-1} \nabla f(x^k)}{\|\nabla f(x^k)\|_2^2 \|[\nabla^2 f(x^k)]^{-1}\|_2} \\
&\geq \frac{\lambda_{\min}([\nabla^2 f(x^k)]^{-1})}{\|[\nabla^2 f(x^k)]^{-1}\|_2} = \frac{\lambda_{\min}([\nabla^2 f(x^k)]^{-1})}{\lambda_{\max}([\nabla^2 f(x^k)]^{-1})} \\
&= \frac{\lambda_{\min}(\nabla^2 f(x^k))}{\lambda_{\max}(\nabla^2 f(x^k))} = \frac{\lambda_{\min}(A^T A + 2\mu I)}{\lambda_{\max}(A^T A + 2\mu I)} > 0.
\end{aligned}
$$

This together with (1) shows that $\lim_{k \to \infty} \|\nabla f(x^k)\|_2 = 0$. Hence, any accumulation point of $\{x^k\}$ is stationary.

# Example cont.

Solution cont.:

3. From the Armijo rule, we have for any $k \geq 1$ that

$$\begin{aligned}\mu\|x^k\|_2^2 \leq f(x^k) &\leq f(x^{k-1}) + c_1\alpha_{k-1}[\nabla f(x^{k-1})]^T d^{k-1} \\ &\leq f(x^{k-1}) \leq \cdots \leq f(x^0).\end{aligned}$$

where the third inequality holds because

- ⋆ Newton direction is a descent direction when $x^{k-1}$ is nonstationary; and
- ⋆ the relation holds as an equality when $x^{k-1}$ is stationary.

Thus,

$$\|x^k\|_2 \leq \sqrt{f(x^0)/\mu} \quad \forall k \geq 1,$$

meaning that $\{x^k\}$ is bounded.

# Example cont.

Solution cont.:

3. We next apply Theorem 2.5.

   First, clearly, $f \in C^1(\mathbb{R}^n)$ with $\inf f \geq 0$ and $\inf_k \bar{\alpha}_k = 1 > 0$.

   Also, $D_k = (A^T A + 2\mu I)^{-1}$ for all $k$. Thus

   $$\lambda_{\min}(D_k) = \frac{1}{\lambda_{\max}(A^T A + 2\mu I)} > 0.$$

   One can take $\delta = \frac{1}{\lambda_{\max}(A^T A + 2\mu I)}$ in Theorem 2.5. Then we conclude that any accumulation point of $\{x^k\}$ is stationary.