4.8 Quasi-Newton methods

Variants of the Newton method in which information about second order derivatives are extracted from the changes in $\nabla f(x)$ rather than using/inverting the Hessian matrix.

A sequence $\{H_k\}$ is generated, where $H_k$ is symmetric positive definite and an approximation of $[\nabla^2 f(x_k)]^{-1}$, and we take

$$x_{k+1} = x_k + \alpha_k d_k \quad \text{with} \quad d_k = -H_k \nabla f(x_k),$$

where $\alpha_k > 0$ minimizes $f(x)$ along $d_k$ or satisfies some inexact 1-D search conditions (e.g., Wolfe conditions).

**Advantages** with respect to Newton method:

- since the $H_k$'s are symmetric and positive definite, always well defined iteration and descent direction,
- only involves first order derivatives,
- matrix $H_k$ is constructed iteratively, each iteration has complexity $O(n^2)$.

**Disadvantages** with respect to conjugate direction methods: requires storage and handling of matrices.
**Idea:** Extract from $\nabla f(x_k)$ and $\nabla f(x_{k+1})$ information concerning the second order derivatives of $f(x)$.

Quadratic approximation of $f(x)$ around $x_k$:

$$f(x_k + \delta) \approx f(x_k) + \delta^t \nabla f(x_k) + \frac{1}{2} \delta^t \nabla^2 f(x_k) \delta.$$

Differentiating we obtain

$$\nabla f(x_k + \delta) \approx \nabla f(x_k) + \nabla^2 f(x_k) \delta.$$

Substituting $\delta$ with $\delta_k$ and setting $\delta_k = x_{k+1} - x_k$ e $\gamma_k = \nabla f(x_{k+1}) - \nabla f(x_k)$ we have

$$\gamma_k \approx \nabla^2 f(x_k) \delta_k,$$

namely

$$[\nabla^2 f(x_k)]^{-1} \gamma_k \approx \delta_k.$$

Since $\delta_k$ and $\gamma_k$ can only be determined after the 1-D search, we select $H_{k+1}$ symmetric and positive definite such that

$$H_{k+1} \gamma_k = \delta_k \quad \text{secant condition.} \quad (1)$$

This linear system does not define $H_{k+1}$ univocally. Many different ways to satisfy the conditions (1) since there are $n$ equations and $n(n+1)/2$ degrees of freedom.
One of the simplest way is to proceed by *successive updates*:

\[ H_{k+1} = H_k + a_k u u^t \]  

(2)

where \( u u^t \) is a symmetric matrix of rank 1 and \( a_k \) is a proportionality coefficient.

In order to satisfy (1) we must have

\[ H_k \gamma_k + a_k u u^t \gamma_k = \delta_k \]

and hence \( u \) and \( (\delta_k - H_k \gamma_k) \) must be collinear.

Since \( a_k \) accounts for the proportionality, we can set \( u = \delta_k - H_k \gamma_k \) and hence \( a_k u^t \gamma_k = 1 \).

**Rank one update formula:**

\[ H_{k+1} = H_k + \frac{(\delta_k - H_k \gamma_k)(\delta_k - H_k \gamma_k)^t}{(\delta_k - H_k \gamma_k)^t \gamma_k} \]  

(3)

**Properties**

1. For quadratic strictly convex functions, in at most \( n \) iterations we obtain \( H_n = Q^{-1} \), even with inexact 1-D search.
2. There is no guarantee that \( H_k \) is positive definite!
Rank two updates

\[ H_{k+1} = H_k + a_k u u^t + b_k v v^t \]  

(4)

turn out to be more interesting.

To satisfy (1) we have

\[ H_k \gamma_k + a_k u u^t \gamma_k + b_k v v^t \gamma_k = \delta_k \]

where \( u, v \) are not determined univocally.

Setting \( u = \delta_k \) and \( v = H_k \gamma_k \) we obtain the conditions \( a_k u^t \gamma_k = 1 \) and \( b_k v^t \gamma_k = -1 \) and hence the rank two update formula:

\[ H_{k+1} = H_k + \frac{\delta_k \delta_k^t}{\delta_k \gamma_k} - \frac{H_k \gamma_k \gamma_k^t H_k}{\gamma_k^t H_k \gamma_k} \]  

Davidon-Fletcher-Powell (DFP)  

(5)
Proposition: If 
\[ \delta_k \gamma_k > 0 \quad \forall k \quad \text{curvature condition}, \]
the DFP method preserves the positive definiteness of \( H_k \).

Proof:
Suppose \( H_0 \) is positive definite (p.d.) and proceed by induction.
Let us verify that if \( H_k \) is p.d. then \( z^t H_{k+1} z > 0 \quad \forall z \neq 0 \).
If \( H_k \) is p.d. it admits a Cholesky factorization \( H_k = L_k L_k^t \).
Eliminating the subscripts \( k \) and setting \( a = L^t z \) and \( b = L^t \gamma \) we have
\[
z^t (H - \frac{H \gamma \gamma^t H}{\gamma^t H \gamma}) z = a^t a - \frac{(a^t b)^2}{b^t b} \geq 0
\]
because \( |a^t b| \leq \|a\| \|b\| \) (Cauchy inequality).
Since \( z \neq 0 \), equality only holds if \( a \) and \( b \) collinear, namely if \( z \) and \( \gamma \) are collinear.
Since \( \delta^t \gamma > 0 \) we have that
\[
z^t \left( \frac{\delta \delta^t}{\delta^t \gamma} \right) z \geq 0
\]
which holds in the strict sense if \( z \) and \( \gamma \) are collinear.
It thus suffices to "develop" \( z^t H_{k+1} z \) and to apply these two inequalities.
**Fact:** The curvature condition $\delta^t_k \gamma_k > 0$ holds for every $k$ provided that the 1-D search satisfies (weak or strong) Wolfe conditions.

**Proof:**

For quadratic strictly convex functions, $\gamma_k = Q\delta_k$ implies $\delta^t_k Q\delta_k = \delta^t_k \gamma_k > 0$ because $Q$ is positive definite.

For arbitrary functions:

Weak Wolfe conditions

\[ f(\underline{x}_k + \alpha_k \underline{d}_k) \leq f(\underline{x}_k) + c_1 \alpha_k \nabla^t f(\underline{x}_k) \underline{d}_k \quad \text{(Armijo criterion)} \quad (6) \]

\[ \nabla^t f(\underline{x}_k + \alpha_k \underline{d}_k) \underline{d}_k \geq c_2 \nabla^t f(\underline{x}_k) \underline{d}_k \quad (7) \]

with $0 < c_1 < c_2 < 1$.

Since $\delta_k = \alpha_k \underline{d}_k$, (7) implies that

\[ \nabla^t f(\underline{x}_{k+1}) \delta_k \geq c_2 \nabla^t f(\underline{x}_k) \delta_k \]

which in turn implies that

\[ \frac{\gamma^t_k \delta_k}{\gamma_k} \geq (c_2 - 1) \alpha_k \nabla^t f(\underline{x}_k) \underline{d}_k \]

with $(c_2 - 1) < 0$, $\alpha_k > 0$, $\nabla^t f(\underline{x}_k) \underline{d}_k < 0$ because $\underline{d}_k$ is a descent direction. □
Properties

For quadratic strictly convex functions, the DFP method with exact 1-D search:

1. terminates in at most $n$ iterations with $H_n = Q^{-1}$;
2. generates $Q$–conjugate directions (starting from $H_0 = I$ it generates the conjugate gradient directions);
3. the secant condition is hereditary, that is, we have $H_i \gamma_j = \delta_j$ for $j = 0, \ldots, i - 1$.

For arbitrary functions:

4. if $\delta_k^t \gamma_k > 0$ (curvature condition), the $H_k$’s are positive definite if $H_0$ is positive definite (hence descent method);
5. each iteration has $O(n^2)$ complexity;
6. superlinear convergence rate (in general, only local convergence);
7. if the function $f(\mathbf{x})$ is convex, DFP method with exact 1-D search is globally convergent.
BFGS method

In a complementary way, we can construct an approximation of \( \nabla^2 f(x_k) \) instead of \( \left[ \nabla^2 f(x_k) \right]^{-1} \). Since we aim at \( B_k \approx \nabla^2 f(x_k) \), \( B_k \) must satisfy \( B_{k+1} \delta_k = \gamma_k \).

Taking \( B_{k+1} = B_k + a_k u u^t + b_k v v^t \), with similar manipulations, we have:

\[
B_{k+1} = B_k + \frac{\gamma_k \gamma_k^t}{\gamma_k^t \delta_k} - \frac{B_k \delta_k \delta_k^t B_k}{\delta_k^t B_k \delta_k}
\]  

(8)

which should be inverted at each iteration to obtain \( H_{k+1} \).

By applying twice Sherman–Morrison identity

\[
(A + a b^t)^{-1} = A^{-1} - \frac{A^{-1} a b^t A^{-1}}{1 + b^t A^{-1} a}, \quad A \in \mathbb{R}^{n \times n} \text{ non singular, } a, b \in \mathbb{R}^n, \text{ denominator } \neq 0,
\]

we obtain the Broyden Fletcher Goldfarb and Shanno (BFGS) update formula:

\[
H_{k+1} = H_k + \left( 1 + \frac{\gamma_k^t H_k \gamma_k}{\delta_k^t \gamma_k} \right) \frac{\delta_k \delta_k^t}{\delta_k^t \gamma_k} - \frac{H_k \gamma_k \delta_k^t + \delta_k \gamma_k^t H_k}{\delta_k^t \gamma_k}
\]  

(9)

which does not require many more operations than the DFP update.

It is easily verified that \( B_{k+1} H_{k+1} = I \) if \( B_k H_k = I \).
The BFGS method has the same properties 1 to 5 as the DFP method.

In practice, it is more robust (with respect to rounding errors and inexact 1-D search).

BFGS and DFP are two extreme cases of the unique **Broyden family** of update formulae:

\[ H_{k+1} = (1 - \phi)H_{k+1}^{\text{DFP}} + \phi H_{k+1}^{\text{BFGS}} \]

with \( 0 \leq \phi \leq 1 \).

**Properties:** (Broyden family)

- \( H_{k+1} \) satisfy the secant condition and is positive definite if \( \delta_k^t \gamma_k > 0 \).
- Invariant with respect to affine transformations of the variables.
- If \( f(x) \) is quadratic strictly convex, the method with exact 1-D search finds \( x^* \) in at most \( n \) iterations (\( H_n = Q^{-1} \)) and the generated directions are \( Q \)-conjugate.
- Quasi-Newton methods are much less "sensitive" to the inexact 1-D search than conjugate direction methods (appropriate for highly nonlinear functions).
Convergence of quasi-Newton methods

Complex analysis because the approximation of the Hessian matrix (or of its inverse) is updated at each iteration.

Speed of convergence for any \( \{B_k\} \) or \( \{H_k\} \) with inexact 1-D search (Wolfe conditions) in which the step length \( \alpha_k = 1 \) is tried first:

**Theorem:** (Dennis and Moré)

Consider \( f \in C^3 \) and quasi-Newton method with \( B_k \) positive definite and \( \alpha_k = 1 \) for each \( k \). If \( \{x_k\} \) converges to \( x^* \) with \( \nabla f(x^*) = 0 \) and \( \nabla^2 f(x^*) \) is definite positive, \( \{x_k\} \) converges superlinearly if and only if

\[
\lim_{k \to \infty} \frac{\| (B_k - \nabla^2 f(x^*)) d_k \|}{\| d_k \|} = 0. \tag{10}
\]

\( B_k = H_k^{-1} \approx \nabla^2 f(x_k) \)

If a quasi-Newton direction \( d_k \) approximates the Newton direction well enough, \( \alpha_k = 1 \) satisfies the Wolfe conditions when \( x_k \) converges to \( x^* \).

**Observation:** No need that \( B_k \to \nabla^2 f(x^*) \), it suffices that the \( B_k \)'s become increasingly accurate approximations of \( \nabla^2 f(x^*) \) along the directions \( d_k \)!
The condition (10), necessary and sufficient for superlinear convergence, is satisfied by the quasi-Newton methods such as BFGS and DFP.

**Comparison of the convergence rates of the gradient, Newton and BFGS methods:**


**Global convergence:**

Under some assumptions, there exist global convergence results for arbitrary functions when the 1-D search is inexact.

"Classical" globalization techniques (restart or trust region) are in general not used for quasi-Newton methods because in practice no examples of non convergence are known.
Limited memory variants

The most widely used quasi-Newton methods are based on the BFGS and DFP updating formulae and 1-D search procedures satisfying Wolfe conditions.

Since for large values of $n$ it is too costly to explicitly construct $B_k$ or $H_k$, limited memory variants have been proposed where the Hessian approximations can be stored compactly in just a few vectors of dimension $n$ ($\delta_k$ and $\gamma_k$) related to the last iterations.

The *limited memory* BFGS method (cf. exercise 7.5) is quite efficient for large scale optimization problems.