PRO 5970 Métodos de Otimização Não Linear

Celma de Oliveira Ribeiro

Aula 5 - Quasi-Newton methods - 2023

PPGEP - EPUSP

Newton Search

Advantages:

- Excellent performance of Newton search close to the optimum
- Less sensitive to numerical errors than steepest descent search

Disadvantage:

- Very sensitive to starting point x_0
- Can fail to converge when starting relatively far from a local optimum!
- Hessian matrix needed at each iteration, as well as solution of a linear system - Very burdensome task, especially for large-scale systems!

Need to mitigate these deficiencies!

Conjugate Direction Methods

The optimization methods considered usually find, at iteration k, a direction d_k, such that

$$x_{k+1} = x_k + \alpha_k d_k$$

For a given function f

- Steepest descent

$$d_k = -\nabla f(x_k)$$

- Newton

$$d_k = -H(x_k)^{-1} \nabla f(x_k)$$

Deflection Matrices

Idea: Blend first- and second-order methods so as to conserve their respective advantages:

- Use only first partial derivatives and guarantee convergence, as with steepest descent search
- Speed-up convergence with some higher-order information, as with Newton search

Deflection Matrices D_k produce modified gradient search direction:

$$d_k = -D_k \nabla f\left(x_k\right)$$

Special cases of deflection matrices: Steepest Descent: $D_k = I$, Newton: $D_k = H(x_k)^{-1}$

Quasi-Newton methods

Concepts

- For a general objective function, convergence from an arbitrary initial point to a solution cannot be assured in Newton's method
- Newton's method locally approximates the objective function by a quadratic function at every iteration.
- The point x_k, the minimizer for the quadratic approximation, is used as the starting point for the next iteration.

$$x_{k+1} = x_k - H(x_k)^{-1} \nabla f(x_k)$$

- The descent direction is obtained through $x_{k+1} = x_k - \alpha_k H(x_k)^{-1} \nabla f(x_k)$, choosing α_k to assure $f(x_{k+1}) < f(x_k)$
- Minimize $\phi(\alpha) = f(x_k - \alpha_k H(x_k)^{-1} \nabla f(x_k))$ may be difficult

Quasi-Newton methods

The Quasi-Newton method (QNM) is one of the important developments in the field of nonlinear optimization.

This method is used when Newton method is difficult to use or when computing Hessian is too expensive per iteration

Quasi-Newton methods

The Quasi-Newton method (QNM) is one of the important developments in the field of nonlinear optimization.

This method is used when Newton method is difficult to use or when computing Hessian is too expensive per iteration

Newton's method vs quasi-Newton

- Drawback of Newton's method: iteratively evaluate $H(x_k)^{-1}$ and solve the equation $H(x_k)d_k = -\nabla f(x_k)$
- The Quasi-Newton methods use an approximation to $H(x_k)^{-1}$ to avoid the true inverse calculation.

How to select $B_k \approx H(x_k)$ and $D_k \approx H(x_k)^{-1}$for fast convergence of algorithms?

Quasi-Newton methods

Quasi-Newton

- This method approximates Hessian by using only the gradient information.
- QNM can be applicable for both convex and nonconvex problems.
- The search directions are of the form $d_k = -D_k \nabla f(x_k)$ in lieu of $H(x_k)^{-1} \nabla f(x_k)$
- QNM iteration for minimizing the objective function, which is twice differentiable, is

  $$x_{k+1} = x_k - \alpha_k D_k \nabla f(x_k)$$

  where D_k is the symmetric positive definite approximation of $H(x_k)^{-1}$
- Instead of updating each iteration by computing second-order derivative computation, in QNM the sequence D_k is updated dynamically for each iteration

Quasi-Newton methods

The method works?

1. Idea
   - The Hessian $H(x_k)$ reflects the rate of change in gradient $\nabla f(x_k)$,

     $$\nabla f(x_{k+1}) - \nabla f(x_k) \approx H(x_k)\,(x_{k+1} - x_k)$$

   - Deflection matrices D_k approximate the inverse Hessian matrix $H(x_k)^{-1}$ by satisfying the quasi-Newton condition:

     $$D_k\,(\nabla f(x_{k+1}) - \nabla f(x_k)) = (x_{k+1} - x_k)$$

     Since the Hessian $H(x_k)$ ans $H(x_k)^{-1}$ are symmetric, it is natural to require that D_k, the approximation to $H(x_k)^{-1}$, be symmetric. Thus,

     $$D_k = (D_k)^t \quad \text{(symmetric)}$$

Quasi-Newton methods

The method works?

1. Idea
    - The Hessian $H(x_k)$ reflects the rate of change in gradient $\nabla f(x_k)$,

    $$\nabla f(x_{k+1}) - \nabla f(x_k) \approx H(x_k)\,(x_{k+1} - x_k)$$

    - Deflection matrices D_k approximate the inverse Hessian matrix $H(x_k)^{-1}$ by satisfying the quasi-Newton condition:

    $$D_k\,(\nabla f(x_{k+1}) - \nabla f(x_k)) = (x_{k+1} - x_k)$$

    Since the Hessian $H(x_k)$ ans $H(x_k)^{-1}$ are symmetric, it is natural to require that D_k, the approximation to $H(x_k)^{-1}$, be symmetric. Thus,

    $$D_k = (D_k)^t \quad \text{(symmetric)}$$

2. Robustness - Guarantee Directions Improve
    - Consider $x_{k+1} = x_k - \alpha_k D_k \nabla f(x_k)$ For d_k to be improving (descent direction): $f(x_{k+1}) - f(x_k) \approx \nabla f(x_k)^t (x_{k+1} - x_k) = \nabla f(x_k)^t d_k < 0$

Quasi-Newton methods

The method works?

1. Idea
   - The Hessian $H(x_k)$ reflects the rate of change in gradient $\nabla f(x_k)$,

     $$\nabla f(x_{k+1}) - \nabla f(x_k) \approx H(x_k)\,(x_{k+1} - x_k)$$

   - Deflection matrices D_k approximate the inverse Hessian matrix $H(x_k)^{-1}$ by satisfying the quasi-Newton condition:

     $$D_k\,(\nabla f(x_{k+1}) - \nabla f(x_k)) = (x_{k+1} - x_k)$$

     Since the Hessian $H(x_k)$ ans $H(x_k)^{-1}$ are symmetric, it is natural to require that D_k, the approximation to $H(x_k)^{-1}$, be symmetric. Thus,

     $$D_k = (D_k)^t \quad \text{(symmetric)}$$

2. Robustness - Guarantee Directions Improve
   - Consider $x_{k+1} = x_k - \alpha_k D_k \nabla f(x_k)$ For d_k to be improving (descent direction): $f(x_{k+1}) - f(x_k) \approx \nabla f(x_k)^t (x_{k+1} - x_k) = \nabla f(x_k)^t d_k < 0$
   - With $d_k = -D_k \nabla f(x_k)$, one has: $\nabla f(x_k)^t D_k \nabla f(x_k) > 0$
   - Sufficient condition for direction improving is: D_k positive definite

We shall build D_k symmetric and positive definite.

Quasi-Newton methods

Observe:

First order approximation
$f\left(x_{k+1}\right) = f\left(x_k\right) + \nabla f(x_k)\left(x_{k+1} - x_k\right) + \underbrace{\mathbb{O}\left(\|\left(x_{k+1} - x_k\right)\|\right)}_{Error}$

As

$$x_{k+1} = x_k - \alpha D_k \nabla f(x_k)$$

it follows $f\left(x_{k+1}\right) = f\left(x_k\right) - \alpha \nabla f(x_k) D_k \nabla f(x_k) + \mathbb{O}\left(\|\left(x_{k+1} - x_k\right)\|\right)$

As $\alpha \to 0$, the second term dominates the third. To guarantee a decrease for small, we shall have $\alpha \nabla f(x_k) D_k \nabla f(x_k) > 0$

$$\boxed{\text{To ensure this, consider } D_k \text{ be positive definite.}}$$

$$\boxed{\text{Notation } \nabla f_k = \nabla f(x_k)}$$

Quasi-Newton methods

Different QNM such as symmetric rank one (SR1), Davidon-Fletcher-Powell (DFP), Broyden-Fletcher-Goldfarb-Shanno (BFGS), and Broyden class computes D_{k+1} from D_k differently.

Symmetric rank one - SR1

Consider the possibility of defining a recursion of the form $D_{k+1} = D_k + a_k z_k z_k^t$ where a_k is a constant and z_k a vector

In case of SR1 update, Hessian approximation is obtained by

$$B_{k+1} = B_k + \frac{(y_k - B_k s_k)(y_k - B_k s_k)^t}{(y_k - B_k s_k)^t s_k}$$

with $s_k = x_{k+1} - x_k$ and $y_k = \nabla f(x_{k+1}) - \nabla f(x_k)$

We obtain the corresponding update formula for the inverse Hessian approximation :

$$D_{k+1} = D_k + \frac{(s_k - D_k y_k)(s_k - D_k y_k)^t}{(s_k - D_k y_k)^t y_k}$$

Even if D_k is positive definite, D_{k+1} may not have the same property!

Quasi-Newton methods

Symmetric rank one - SR1

There are some difficulties with this simple rank one procedure.

1. The main drawback of SR1 updating is that the denominator can vanish.
   1.1 If $(y_k - B_k s_k)^t s_k \neq 0$ use SR1 update formula
   1.2 If $y_k = B_k s_k \Rightarrow B_{k+1} = B_k$
   1.3 If $(y_k \neq B_k s_k)$ and $(y_k - B_k s_k)^t s_k = 0$ Skip the update in this case $\Rightarrow B_{k+1} = B_k$
2. Second, the up- dating formula preserves positive definiteness only if $(y_k - B_k s_k)^t s_k > 0$ which cannot be guaranteed.
3. Also, even if $(y_k - B_k s_k)^t s_k$ is positive, it may be small, which can lead to numerical difficulties.

Thus, although an excellent simple example of how information gathered during the descent process can in principle be used to update an approximation to the inverse Hessian, the rank one method possesses some limitations.

Quasi-Newton methods

Davidon- Fletcher-Powell (DFP) Method

(Ref: Luemberger and Ye, 2016)

- Originally proposed by Davidon and later developed by Fletcher and Powell
- The earliest, and certainly one of the most clever schemes for constructing the inverse Hessian
- It has the fascinating and desirable property that, for a quadratic objective, it simultaneously generates the directions of the conjugate gradient method while constructing the inverse Hessian.
- At each step the inverse Hessian is updated by the sum of two symmetric rank one matrices, and this scheme is therefore often referred to as a rank two correction procedure.
- The method is also often referred to as the variable metric method, the name originally suggested by Davidon.

.

Quasi-Newton methods

Davidon- Fletcher-Powell (DFP) Method

- Method falls into the class of *quasi-Newton procedures*
- The gradient direction is deflected by premultiplying by $-D_k$, where D_k is a $n \times n$ positive definite symmetric matrix that approximates the inverse of the Hessian
- If the objective function is quadratic, the method yelds conjugate directions $d_k = -D_k \nabla f(x_k)$
- If the objective function is quadratic, after one complete iteration (that is, after search each of the conjugate directions created by the algorithm) the method stops with an optimal point

Quasi-Newton methods

Davidon- Fletcher-Powell (DFP) Method

1 Set $\epsilon > 0$ k:= 0; Let x_0 be the initial point. Select and a real symmetric positive definite D_0

2 if $\|\nabla f_k\| < \epsilon$ stop; else $d_k = -D_k \nabla f_k$

3 Let $\alpha_k > 0$ be an optimal solution of $\min_{\alpha \geq 0} f(x_k + \alpha d_k)$

4 $x_{k+1} = x_k + \alpha_k d_k$

5 Compute

- $s_k = x_{k+1} - x_k = \alpha_k d_k,$
- $y_k = \nabla f_{k+1} - \nabla f_k$
- $D_{k+1} = D_k + \frac{s_k s_k^{'}}{s_k^{'} y_k} - \frac{D_k y_k y_k^{'} D_k}{y_k^{'} D_k y_k}$

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

Quasi-Newton methods

Example 1 - DFP

Consider the following problem

$$min(x_1 - 2)^4 + (x_1 - 2x_2)^2$$

, Beginning with $D_1 = I$, consider $d_j - -D_j \nabla f(y_j)$. The computations for DFP are presented below:

Quasi-Newton Methods

This example is from Bazaraa (Example 8.8.4). (λ_k in Bazaraa in α_k in Nocedal)

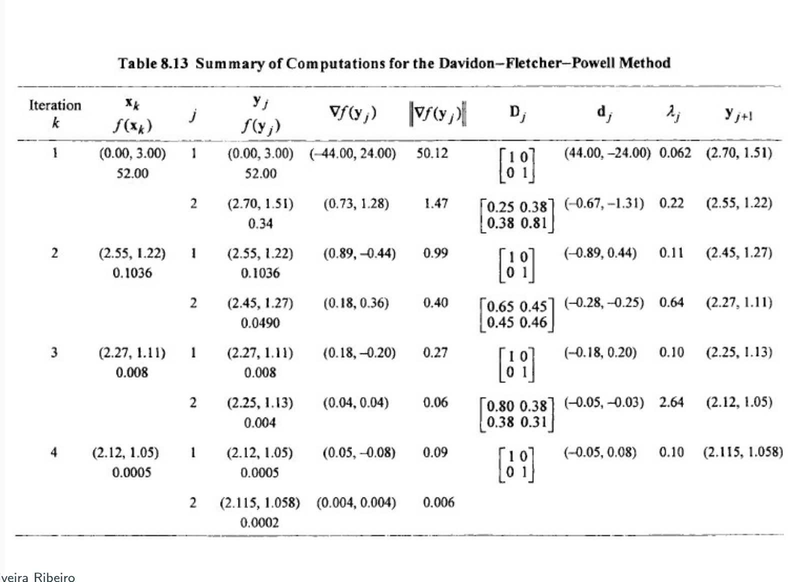

Table 8.13 Summary of Computations for the Davidon–Fletcher–Powell Method

Iteration k	$\mathbf{x}_k$ $f(\mathbf{x}_k)$	j	$\mathbf{y}_j$ $f(\mathbf{y}_j)$	$\nabla f(\mathbf{y}_j)$	$\lVert\nabla f(\mathbf{y}_j)\rVert$	$\mathbf{D}_j$	$\mathbf{d}_j$	λ_j	$\mathbf{y}_{j+1}$
1	(0.00, 3.00) 52.00	1	(0.00, 3.00) 52.00	(–44.00, 24.00)	50.12	$\begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}$	(44.00, –24.00)	0.062	(2.70, 1.51)
		2	(2.70, 1.51) 0.34	(0.73, 1.28)	1.47	$\begin{bmatrix}0.25 & 0.38\\0.38 & 0.81\end{bmatrix}$	(–0.67, –1.31)	0.22	(2.55, 1.22)
2	(2.55, 1.22) 0.1036	1	(2.55, 1.22) 0.1036	(0.89, –0.44)	0.99	$\begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}$	(–0.89, 0.44)	0.11	(2.45, 1.27)
		2	(2.45, 1.27) 0.0490	(0.18, 0.36)	0.40	$\begin{bmatrix}0.65 & 0.45\\0.45 & 0.46\end{bmatrix}$	(–0.28, –0.25)	0.64	(2.27, 1.11)
3	(2.27, 1.11) 0.008	1	(2.27, 1.11) 0.008	(0.18, –0.20)	0.27	$\begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}$	(–0.18, 0.20)	0.10	(2.25, 1.13)
		2	(2.25, 1.13) 0.004	(0.04, 0.04)	0.06	$\begin{bmatrix}0.80 & 0.38\\0.38 & 0.31\end{bmatrix}$	(–0.05, –0.03)	2.64	(2.12, 1.05)
4	(2.12, 1.05) 0.0005	1	(2.12, 1.05) 0.0005	(0.05, –0.08)	0.09	$\begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}$	(–0.05, 0.08)	0.10	(2.115, 1.058)
		2	(2.115, 1.058) 0.0002	(0.004, 0.004)	0.006				

Quasi-Newton Methods

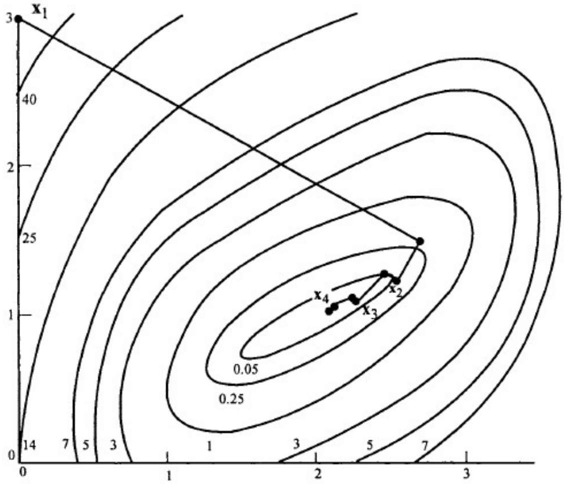

Figure 8.22 Davidon–Fletcher–Powell method.

Quasi-Newton Methods

Example 2 - DFP Entregar

Consider the quadratic function $f(x) = \frac{1}{2}x^{'}Ax - b^{'}x$ with

$$A = \begin{bmatrix} 4 & 2 \\ 2 & 2 \end{bmatrix}, b = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

Find the minimizer using the DFP algorithm. Starting point $x_0 = \begin{bmatrix} 0 & 0 \end{bmatrix}^{'}$
Assume $D_0 = I$

Quasi-Newton methods

Comments about DFP

- If f quadratic, after n steps $D_n = H^{-1}$
- If $D_0 = I$ we have the Conjugate Gradient method.
- The DFP algorithm preserves the positive definiteness of the matrices.
- For larger nonquadratic problems the algorithm has the tendency of sometimes getting stuck.

Lemma

Let $x_0 \in \mathbb{R}$ and D_1 be a positive definite symmetric matrix.

If $\nabla f_k \neq 0 \ \forall k$, then $D_1, D_2, \ldots, D_n$ are symmetric positive definite and $d_1, d_2, \ldots, d_n$ are descent directions

Quasi-Newton Methods

BFGS Method

One of the most popular methods, known as the BFGS method. The name is an acronym of the algorithm's creators: Broyden, Fletcher, Goldfarb, and Shanno, who each came up with the algorithm independently in 1970.

Figure 2. From left to right: Broyden, Fletcher, Goldfarb, and Shanno.

Quasi-Newton Methods

BFGS Method

In this case BFGS update procedure is

$$D_{k+1} = \left(I - \rho_k s_k y_k^t\right) D_k \left(I - \rho_k y_k s_k^t\right) + \rho_k s_k s_k^t$$

with $\quad y_k = \nabla f_{k+1} - \nabla f_k \qquad s_k = x_{k+1} - x_k \qquad \rho_k = \frac{1}{y_k^t s_k}$

D_k is the approximation of the inverse of the Hessian

Quasi-Newton Methods

BFGS Method

```
1 Set $\epsilon > 0$ k:= 0; Let $x_0$ be the initial point. Select an inverse Hessian approximation $D_0$
2 if $\|\nabla f_k\| < \epsilon$ stop; else $d_k = -D_k \nabla f_k$
3 Let $\lambda_k > 0$ be an optimal solution of $\min_{\lambda \geq 0} f(x_k + \lambda d_k)$
4 $x_{k+1} = x_k + \alpha_k d_k$
5 Compute
    • $s_k = x_{k+1} - x_k$,
    • $y_k = \nabla f_{k+1} - \nabla f_k$
    • $D_{k+1} = \left(I - \rho_k s_k y_k^t\right) D_k \left(I - \rho_k y_k s_k^t\right) + \rho_k s_k s_k^t$ with $\rho_k = \frac{1}{y_k^t s_k}$
5 Set k := k+1; go to step 2.
```

In the BFGS method, the positive definiteness (and thus nonsingularity) of all Hessian approximations is guaranteed.

Quasi-Newton Methods

Advantages and disdvantages of BFGS Method

1. BFGS update preserve positive definiteness under appropriate conditions and has low computation cost.
2. BFGS update has local superlinear rate of convergence without the need to solve linear systems.
3. Even when the Hessian matrix are sparse, updated inverse Hessian approximation yields dense matrix. This problem restricts BFGS method to use for small scale and midscale data set.

Quasi-Newton Methods

Example 3 - BFGS Entregar

Consider the quadratic function $f(x) = \frac{1}{2}x^{'}Ax - b^{'}x$ with

$$A = \begin{bmatrix} 4 & 2 \\ 2 & 2 \end{bmatrix}, b = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

Find the minimizer using the DFP algorithm. Starting point $x_0 = \begin{bmatrix} 0 & 0 \end{bmatrix}^{'}$
Assume $D_0 = I$

Exercice (24/07, 23:59)

Use the four algorithms (Conjugate gradient, SR1, DFP, BFGS) to find a minimum for Rosenbrock's Functions (in $\mathbb{R}^2$). Use the library

https://www.sfu.ca/˜ssurjano/optimization.html

Discuss the methods considering implementation, convergence, and efficiency