Universitat Politècnica de Catalunya

Ciència i Enginyeria de Dades

Optimització Matemàtica: Apunts de Teoria

Gerard Comas Quiles

Feb 2022

Índice

1. Unconstrained Nonlinear Optimization **2**

1.1. Fundamentals . . . 2

1.1.1. Generic Unconstrained Optimization Algorithm (GUOA) . . . 2

1.1.2. Optimality conditions . . . 2

1.1.3. Convexity . . . 4

1.1.4. Direcció descendent . . . 4

1.1.5. Line Search . . . 4

1.1.6. Exact Line Search (quadratic f) . . . 4

1.1.7. Inexact Line Search . . . 5

1.1.8. Wolfe Condition . . . 5

1.1.9. Perfomance d'Optimitació d'Algoritmes . . . 6

1.2. First Derivative Method . . . 6

1.2.1. Gradient Method . . . 6

1.2.2. Conjugate Gradient Method . . . 7

1.2.3. Quasi-Newton Methods . . . 8

1.3. Second Derivative Methods . . . 9

1.3.1. Newton's Method . . . 9

1.3.2. Modified Newton's Method . . . 10

1.3.3. MNM-SD methods . . . 11

1.3.4. MNM bassed on the Cholesky Factorization . . . 11

1. Unconstrained Nonlinear Optimization

1.1. Fundamentals

1.1.1. Generic Unconstrained Optimization Algorithm (GUOA)

Els algoritmes UO (Unconstrained Optimization) generen una seqüència que convergeix en la solució óptima.

Algoritme GUOA: Generic Unconstrained Optimization Algorithm:

Fins que x^k no satisfaci les condicions óptimes fes:

I Troba una direcció descendent d^k

II Troba una longitud de pas α^k

III Actualitza les variables: $x^{k+1} \leftarrow x^k + \alpha^k d^k$

LLavors, en general, la última x serà la solució óptima

1.1.2. Optimality conditions

Def. (Local Minimizer): Un punt x^* és un local minimizer si hi ha un veïnat N de x^* tal que:

$$f(x^*) \leq f(x), x \in N \tag{1}$$

Direm que es un local minimizer fort si es compleix la desigualtat estricte ($<$)

Def. (Global Minimizer): Un punt x^* és un global minimizer si:

$$f(x^*) \leq f(x), \forall x \tag{2}$$

Direm que es un global minimizer fort si es compleix la desigualtat estricte ($<$)

Def. (Isolated Minimizer): Un punt x^* es un isolated minimizer si hi ha un veïnat N de x^* tal que x^* es l'únic local minimizer de N.

- Totes aquestes definicions s'apliquen anàlogament a màxims, simplement cal canviar les desigualtats de banda.
- Que hi hagi un Strict local minimizer no implica que aquest sigui isolated
- Estrictament parlant, x^* hauria de ser sempre un global minimizer, però els UOA son capaços, en general, de trobar un local minimizer.

Theorem. (Taylor's Theorem): Suponguem que tenim una funció continuament diferenciable i α i d pertanen als reals. Llavors tenim:

$$f(x+\alpha d) = f(x) + \alpha \nabla f(x)^T d + o(\alpha) \tag{3}$$

$$f(x+\alpha d) = f(x) + \nabla f(x+t\alpha d)^T d \tag{4}$$

i, si la funció f es doblement diferenciable ($f \in \mathcal{C}^2$) també tenim:

$$f(x+\alpha d) = f(x) + \alpha \nabla f(x)^T d + \frac{1}{2}\alpha^2 d^T \nabla^2 f(x) + o(\alpha^2) \tag{5}$$

$$f(x+\alpha d) = f(x) + \nabla f(x)^T d + \frac{1}{2} d^T \nabla^2 f(x+t\alpha d) d \tag{6}$$

per a alguna $t \in (0,1)$

Theorem. (First-Order Necessary Conditions, FONC): Si x^* es un local minimizer i la funció es continuament diferenciable en un veïnat obert de x^*, llavor $\nabla f(x^*) = 0$

Def. (Stationary and Saddle points: Un punt Stationary o Critical es qualsevol x^* que satisfà FONC Un punt Stationary que no es ni mínim ni màxim és un Saddle point

- FONC només es pot utilitzar per a comprobar que un punt no és óptim

- Per a saber distingit si ens trobem en un local maximizer or a saddle point necessitem la segona derivada

Theorem. (Second-Order Necessary Conditions, SONC): Si x^* es un local minimizer i f i $\nabla^2 f$ son continuas en un veïnat obert de x^*, llavors $\nabla f(x^*) = 0$ i $\nabla^2 f(x^*)$ es semidefinida positiva

Theorem. (Second-Order Sufficient Conditions, SOSC): Suposem que $\nabla^2 f$ es continua en un veïnat obert de x^* i que $\nabla f(x^*) = 0$ i $\nabla^2 f(x^*)$ es semidefinida positiva. Llavors x^* és un strict local minimizer de f

Considerem que $H(x) = \nabla^2 f(x)$ es la matriu Hessiana de f en x. **Def. (Symmetric matrix):**

1. **definida positiva:** si $x^T H x > 0, \forall x \neq 0$
2. **definida negativa:** si $x^T H x < 0, \forall x \neq 0$
3. **indefinida** si $x^T H x > 0$, per algun x i $x^T H x < 0$, per algun altre x
4. **semidefinida positiva:** si $x^T H x \geq 0, \forall x$, però $x^t H x = 0$ per algun $x \neq 0$
5. **semidefinida negativa:** si $x^T H x \leq 0, \forall x$, però $x^t H x = 0$ per algun $x \neq 0$

1.1.3. Convexity

Def. (Convex Set): $\mathcal{S} \subset \mathbb{R}^n$ és un Convex Set si $\forall x, y \in \mathcal{S} : \alpha x + (1-\alpha)y \in \mathcal{S}, \forall \alpha \in [0,1]$

Def. (Convex Function): $f : \mathcal{S} \subseteq \mathbb{R}^n \to \mathbb{R}, \mathcal{S}$ covex, es una funció convexa si $\forall x, y \in \mathcal{S}$:

$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha)f(y), \forall \alpha \in [0,1] \tag{7}$$

Proposition. (Combinació de Convex Functions):

1. Siguin f_1 i f_2 dos funcions convexes sobre $\mathcal{S} \Rightarrow f_1 + f_2$ es convex sobre $\mathcal{S}$
2. Sigui f convex sobre $\mathcal{S}$ convex i $a > 0 \Rightarrow af$ es convex sobre $\mathcal{S}$

Proposition. (Propietats de Diferenciació de funcions convexes):

1. Sigui $f \in \mathcal{C}^1$, llavors f és convexa sobre $\mathcal{S} \Leftrightarrow$

$$f(y) >= f(x) + \nabla f(x)^T (y-x), \forall x, y \in \mathcal{S} \tag{8}$$

2. Sigui $f \in \mathcal{C}^2$, llavors f és convexa(estricament convexa) sobre el conjunt convex $\mathcal{S}$ contenint un punt interior $\Leftrightarrow \nabla^2 f(x)$ + semidef

Theorem. (Optimització-Convexitat):

1. Si f es convex, llavors qualsevol mínim local x^* es un mínim global de f
2. Si f es convex i diferenciable, llavors qualsevol punt estacionari x^* es un mínim global de f

1.1.4. Direcció descendent

Def. (Descent Direction: $d \in \mathcal{R}^n$ es una direcció descendent de f en x si existeix algun escalar $\hat{\alpha} > 0$ tal que $\forall \alpha \in (0, \hat{\alpha}) : f(x + \alpha d) < f(x)$

Si $\nabla f(x) \neq 0$, llavors podem assumir la següent condició suficient:

Proposition: $\nabla f(X)^t$ůd $< 0 \to d$ es una direcció descendent de f en x.

1.1.5. Line Search

Def. (Line Search): És el procediment per a trobar el pas de longitud òptim.

$$\alpha^* = argmin_{\alpha > 0} \{\phi(\alpha) = f(x(\alpha)) = f(x + \alpha d)\} \tag{9}$$

1.1.6. Exact Line Search (quadratic f)

Proposition. (Optimal step length for a quadratic function): Considerem $f(x) = \frac{1}{2} x^T Q x - b^T x$ una funció quadràtica convexa. La longitud de pas òptima associada a x y la direcció descendent d, amb $d^T Q d > 0$ es:

$$\alpha^x = -\frac{(Qx - b)^T d}{d^T Q d} \tag{10}$$

1.1.7. Inexact Line Search

El problema és que fora de les funcions quadràtiques el pas de longitud òptim α no es pot trobar analiticament. En aquest cas s'han de acceptar les condicions d'acceptabilitat seguents:

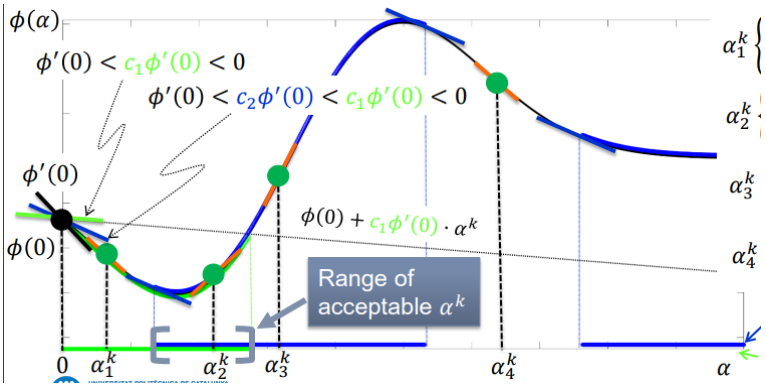
1. Que α^k ha de fer variar suficientment la f
2. Que α^k ha de ser prou gran com per a no quedar-se estancada en una x concreta
3. Que α^k ha de ser prou petit com per a que no es perdi la conergència

1.1.8. Wolfe Condition

Def. (Wolfe Condition):

Sufficient Decrease: $f(x^k + \alpha^k d^k) \leq f^k + c_1 \nabla f^{k^T} d^k \alpha^k$

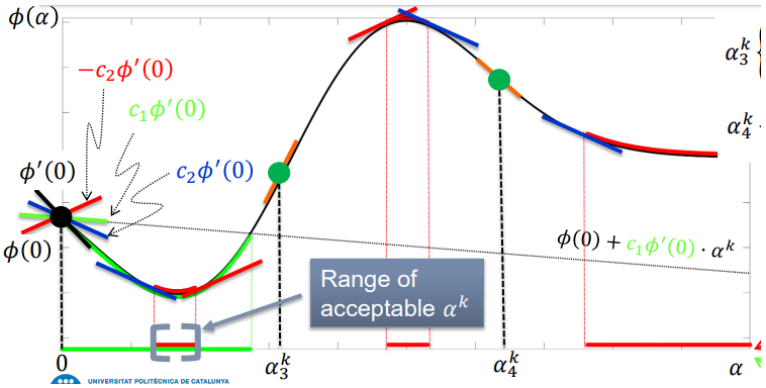
Curvature Condition: $\nabla f(x^k + \alpha^k d^k)^T d^k \geq c_2 \nabla f^{k^T} d^k$



Def. (Strong Wolfe Condition):

Sufficient Decrease: $f(x^k + \alpha^k d^k) \leq f^k + c_1 \nabla f^{k^T} d^k \alpha^k$

Curvature Condition: $|\nabla f(x^k + \alpha^k d^k)^T d^k| \leq c_2 |\nabla f^{k^T} d^k|$



Theorem (Condicions de l'existència de solucion WC i SWC): Suposem una funció f smooth, llavors si $0 < c_1 < c_2 < 1$, existeixen intervals pels quals es compleixen WC i SWC

1.1.9. Perfomance d'Optimitació d'Algoritmes

Per mesurar la performance ens mirem la conergència global i la local. La convergència global es preocupa de trobar quan un algoritme pot trobar una solució. La convergència local es preocupa de trobar com de ràpid convergeix a la solució.

Def. (Global Convergence): Els algoritmes son globalment convergents si:

$$lim||\nabla f(x^k)|| = 0 \tag{11}$$

Theorem. (Zoutendijk's Theorem):

Serà convergent globalment si es compleix:

$$\sum_{k\geq 0} cos^2\theta^k ||\nabla f^k||^2 < \infty \tag{12}$$

Def. (Local Convergence): la convergència local o velocitat de convergència es l'ordre de convergència de la sèrie

Def (Linear and Superlineal Order of Convergence):

Sigui $\{x^k\}$ una sequència en R^n que convergeix a x^*. La convergència serà lineal si existeix una constant $r \in (0, 1)$ tal que:

$$\frac{||x^{k+1} - x^*||}{||x^k - x^*||} \leq r \tag{13}$$

Per una K suficientment gran. Considerarem una convergència superlineal si:

$$lim\frac{||x^{k+1} - x^*||}{||x^k - x^*||} = 0 \tag{14}$$

Def (Quadratic Order of Convergence):

Sigui $\{x^k\}$ una sequència en R^n que convergeix a x^*. La convergència serà quadràtica si existeix una constant $M > 0$ tal que:

$$\frac{||x^{k+1} - x^*||}{||x^k - x^*||^2} \leq M \tag{15}$$

1.2. First Derivative Method

1.2.1. Gradient Method

La direcció de busqueda ve definida per $d_G^k = -\nabla f^k$

Té les següents propietats:

Global convergence: el métode GM es globalment convergent:

1. Cada d_G^k és una direcció descendent: $\nabla f^{k^T} d_G^k = -||\nabla f^k||^2 < 0, \forall k : \nabla f^k \neq 0$

2. Cada d_G^k satisfà la condició d'angle de convergència: $\theta^k = 0 \rightarrow cos\theta^k = 1, \forall k$

Local convergence for quadratic f:

1. L'error de GM aplicat a una funció quadràtica f satisfà:

$$(f^{k+1} - f^*) = [1 - \frac{(\nabla {f^k}^T \nabla f^k)^2}{(\nabla {f^k}^T Q \nabla f^k)(\nabla {f^k}^T Q^{-1} \nabla f^k)}](f^k - f^*) \tag{16}$$

2. Quan volem minimitza una funció convexa quadràtica l'error satisfà:

$$f^{k+1} - f^* \leq (\frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1})^2 (f^k - f^*) \tag{17}$$

on $0 < \lambda_1 \leq ... \leq \lambda_n$ son eigenvalues de Q

1.2.2. Conjugate Gradient Method

La direcció de busqueda ve definida per $d_G^k = -\nabla f^k + \beta^k d_{CG}^{k-1}$

La idea del mètode CG és agafar com a direcció una combinació lineal entre el gradient i la direcció descendent de la darrera iteració. Per a triar els coeficients β^k s'utilitzen les següents fórmules:

1. **Fletcher-Reeves:**

$$\beta_{FR}^k = \frac{\nabla {f^k}^T \nabla f^k}{||\nabla f^{k-1}||^2} \tag{18}$$

2. **Polak-Ribière:**

$$\beta_{PR}^k = \frac{\nabla {f^k}^T (\nabla f^k - \nabla f^{k-1})}{||\nabla f^{k-1}||^2} \tag{19}$$

Global Convergence:

- Descent condition: $\nabla {f^k}^T d_{CG}^k = -||\nabla f^k||^2 + \beta^k \nabla {f^k}^T d_{CG}^{k-1} < 0$

1. **Exact Line Search:** Si $\alpha^k = \alpha^*$ entonces $\nabla {f^k}^T d_{CG}^{k-1}$

2. **Inexact Line Search:** en aquest cas s'han de complir unes condicions a α^k:

   a) **FR:** α ha de complir SWC i $C_2 < 1/2$

   b) **PR:** per garantir que la direcció és de descens calen dos condicions:

      1) S'han d'evitar valors de β_{PR}^k negatius

      2) S'han de complir les WC i la condició suficient de descens:

$$\nabla {f^k}^T d_{CG}^k = -||\nabla f^k||^2 + \beta_{PR}^k \nabla {f^k}^T d_{CG}^{k-1} \leq -c_3 ||\nabla f^k||^2, 0 < c_3 < 1 \tag{20}$$

- Convergence Angle Condition:

Theorem. (Global Convergence of the CG algorithm):

Suposant una funció f que satisfà les condicions de Zoutendijk, llavors:

$$lim\ inf ||\nabla f^k|| = 0 \tag{21}$$

Local Convergence:

- Cada k iteracions es reinicia la pasa de gradient, $d^k = -\nabla f^k$:

1. Sense reinici: $k = +\infty$
2. Amb reinici cada n iterancions: $k = i \times n, i \in 0, 1, 2, ...$
3. Amb reinici cada cop que els dos gradients estan lluny de l'ortogonalitat:

$$\frac{|\nabla f^{k^T} \nabla f^{k-1}|}{||\nabla f^k||^2} \geq \nu \tag{22}$$

A la pràctica, si n és prou gran, la segona condició pot no ocurrir mai. Per tant utilitzem la tercera, amb una $\nu = 0{,}1$

1.2.3. Quasi-Newton Methods

Donada una iteració x^k, el mètode de Newton itera de la següent forma: $x^{k+1} = x^k + \alpha^k d_N^k$, on d_N^k és un minimitzador de $f_Q^k(d)$, siguent l'aproximació quadràtica de Taylor per f(x) al voltant del punt x^k:

$$f(x^k + d) \approx f_Q^k(d) = f^k + \nabla f^{k^T} d + \frac{1}{2} d^T \nabla^2 f^k d \tag{23}$$

$$\nabla f_Q^k(d)|_{d=d_N^k} = 0 \rightarrow d_N^k = -\nabla^2 f^{k^{-1}} \nabla f^k \tag{24}$$

El que farem serà aproximar la Matriu Hessiana $\nabla^2 f^k$ per una matriu B^k que no utilitzi segones derivades. La forma més habitual per aproximar-la es Broyden-Fletcher-Goldfarb-Shanno (BFGS):

1. Primer impossem que f_{QN}^{k+1} tingui les primeres derivades que f en x^k:

$$\nabla f_{QN}^{k+1}(-\alpha^k d_{QN}^k) = \nabla f^{k+1} - \alpha^k B^{k+1} d_{QN}^k = \nabla f^k \tag{25}$$

$$B^{k+1} \alpha^k d_{QN}^k = \nabla f^{k+1} - \nabla f^k \tag{26}$$

2. Definim $s^k = x^{k+1} - x^k$, $y^k = \nabla f^{k+1} - \nabla f^k$ i $H^k = B^{k^{-1}}$ obtenint així l'equació de la secant:

$$H^{k+1} y^k = s^k \tag{27}$$

3. S'ha de complir la següent condició de curvatura:

$$y^{k^T} s^k > 0 \tag{28}$$

4. La equació de la Secant té $\frac{n(n+1)}{2}$ elements desconeguts i n equacions, llavors té un infinit nombre de solucions.

5. $H_{BFGS}^{k+1} = argmin||H - H^k||_W | H = H^T, Hy^k = s^k$

6. Es pot demostrar que la única solució del problema segueix la formula següent:

$$H_{BFGS}^{k+1} = (I - \rho^k s^k y^{k^T}) H_{BFGS}^k (I - \rho^k s^k y^{k^T}) + \rho^k s^k s^{k^T},\ \rho^k = \frac{1}{y^{k^T} s^k} \tag{29}$$

Proposition. (Propietats de la BFGS): Sigui H_{BFGS}^k una matriu simètrica definida positiva:

1. H_{BFGS}^{k+1} és simétrica

2. H_{BFGS}^{k+1} satisfà l'equació secant $H^{k+1} y^k = s^k$

3. H_{BFGS}^{k+1} és definida positiva si α^k satisfà WC o SWC

Finalment, la direcció de busqueda del quasi-Newton BFGS és:

$$d_{BFGS}^k = -H_{BFGS}^k \nabla f^k \tag{30}$$

Global Convergence:

- Descent Condition:

Es compleix la condició de descens:

$$\nabla f^{k^T} d_{BFGS}^k = -\nabla f^{k^T} H_{BFGS}^k \nabla f^k < 0 \tag{31}$$

- Convergent angle condition:

Proposition (CAC para la fórmula BFGS):

Si la matriu H^k:

1. es definida positiva, amb eigenvalues $0 < \lambda_1 \leq ... \leq \lambda_n$

2. té un límit inferior que es:

$$k(H^k) = ||H^k||\ ||H^{k^{-1}}|| = \frac{\lambda_n}{\lambda_1} \leq C, \forall k \tag{32}$$

llavors, $cos\theta^k \geq 1/C > 0$

Local convergence:

Theorem. (Ordre de convergència Superlineal):

Suposem $\{x^k\}_{k=0}^{\infty}$ els iterats generats per l'argoritme BFGS, llavors aquesta sèrie tendeix a x^* amb ordre de convergència superlineal.

1.3. Second Derivative Methods

1.3.1. Newton's Method

Global convergence:

- Descent condition: $\nabla f^{k^T} p_N^k = -\nabla f^{k^T} \nabla^2 f^{k^{-1}} \nabla f^k < 0,\ \nabla f^k \neq 0$

Només podem assegurar que convergeixi globalment si $\nabla^2 f^k$ és definida positiva per a tot k.

Local convergence:

Theorem. (Quadratic Order Of Convergence of the Newton's method)

Considerem la seqüència $\{x^k\}_{k=0}^{\infty}$. En el punt inicial x^0 està suficientment a prop de x^* llavors:

1. $\{x^k\} \to x^*$
2. L'ordre de convergència de $\{x^k\}_{k=0}^{\infty}$ és quadràtic.
3. La seqüència de normes de gradients $\{||\nabla f^k||\}_{k=0}^{\infty}$ convergeix quadràticament a zero.

1.3.2. Modified Newton's Method

Motivation:

El que busquem és que quan x^k estigui proper a x^* segueixi convergint quadràticament i per tant que la B^k s'assembli a la Hessiana. Però en el cas que x^k estigui lluny de x^*, haurem d'assegurar que B^k sigui definida positiva i així guanyarem la convergència global.

Conditioning of B^k:

La Hessiana modifica B^k ha de ser lo suficientment definida positiva. Això està relacionat amb la condició numèrica de B^k, $\kappa(B^k)$.

Def. Sigui A una matriu simmètrica de n x n definida positiva amb eigenvalues $0 < \lambda_1 \leq \lambda_2 \leq ... \leq \lambda_n$. La condició numèrica de A es $\kappa(A) = ||A||_2||A^{-1}||_2 = \lambda_n/\lambda_1$

Global Convergence:

La convergència global dels algoritmes MNM depen, entre altres coses, de la condició numèrica $\kappa(B^K)$.

Theorem. (Global Convergence dels MNM):

Es pot demostrar que si la funció compleix certes propietats i es compleix:

$$\kappa(B^k) \leq C,\ C > 0,\ \forall k \tag{33}$$

Llavors l'agoritme convergeix en un punt estacionàri que és:

$$lim_{k\to\infty} \nabla f^k = 0 \tag{34}$$

Això no ens asegura arribar a un mínim, però si afegimla hipòtesi que $E^* = 0$ llavors l'algoritme convergeix a un mínim local estricte.

Local Convergence:

Theorem. (Unit Step Length for MNM):

Si la seqüència $\{x^k\}_{k=0}^{\infty}$ convergeix a un punt x^* tal que $\nabla^2 f^*$ i a més:

$$lim_{k\to\infty} \frac{||\nabla f^k + \nabla^2 f^k d^k||}{||d^k||} = 0 \tag{35}$$

Llavors existeix un index $k_0 \geq 0$ tal que $\alpha^k = 1$ és admissible per $k \geq k_0$. A més, la convergència a x^* és superlineal. A partir dels anteriors teoremes es pot demostrar que els algoritmes MNM tenen convergència

quadràtica, si $E^k = 0$ per a k molt grans. Aquí trobem un conflicte, ja que també volem que E^k sigui lo suficientment gran per a que se sostingui la convergència global.

Theorem. (Spectral Theorem For Symmetric Matrices):
Si $A \in \mathcal{R}^{n\ x\ n}$, symmetric, then:

1. A té n valors propis $\lambda_1 \leq \lambda_2 \leq ... \leq \lambda_n$
2. Existeix una base ortonormal de vectors propis $Q = [q_1, q_2, ..., q_n]$
3. A diagonalitza:

$$Q^T A Q = \Lambda = diag(\lambda_1, \lambda_2, ..., \lambda_n) \tag{36}$$

o equivalentment (Spectral or Symmetric Schur descomposition):

$$A = Q \Lambda Q^T \tag{37}$$

1.3.3. MNM-SD methods

Es centra en canviar els valors propis λ_i negatius o molt propers a zero per nombres positius δ:

$$B^k_{MNM-SD} = Q \hat{\Lambda} Q^T \tag{38}$$

on,

$$\hat{\Lambda} = diag(max(\delta, \lambda_i)) = \Lambda + diag(max(0, \delta - \lambda_i)) \tag{39}$$

Així que podem garantir que la Hessiana modificada és definida positiva:

$$B^k_{MN-SD} = Q \Lambda Q^T + Q(\Delta \Lambda) Q^T \tag{40}$$

$$d^k_{MN-SD} = -B^{k^{-1}}_{MN-SD} \nabla f^k = -Q(diag(\frac{1}{max(\delta,_i)})) Q^T \nabla f^K \tag{41}$$

és una direcció descendent

1.3.4. MNM bassed on the Cholesky Factorization

Theorem. (Cholesky Factorization):
Sigui $A \in \mathcal{R}^{n\ x\ n}$, és simètrica i definida positiva. Llavors existeix una única matriu triangular R amb una diagonal positiva tal que $A = R^T R$
Existeixen una gran varitetat de MNM basats en la factorització de Cholesky de la Hessiana $\nabla^2 f^k$. Tots aquests han de tenir en compte:

1. La factorització de Cholesky pot no existir per una Hessiana que no sigui definida positiva.
2. Encara que la Hessiana sigui definida positiva, si està mal condicionada (gran condició numèrica), la computació de la factorització pot ser innestable.

Els algoritmes MNM apliquen l'anomenat modified Cholesky factorization, una variant que garanteix l'existència de la factorització, fins i tot quan la Hessiana no és definida positiva. Per simplicitat, introduirem una aproximació, $B^k = \nabla^2 f^k + \tau I$, amb una $\tau \geq 0$ estrictament creixent, fins que per una τ s'arribi a la factorització.