Analysis of Newton's Method

Oscar Dalmau
dalmau@cimat.mx

Centro de Investigación en Matemáticas
CIMAT A.C. Mexico

February 2018

Outline

1. Steepest descent: Summary

2. Newton's Method

3. Newton's Method with Hessian modification

Steepest descent Method: Step size

Steepest descent: Step size

1. In the Steepest descent Method with exact line search two consecutive directions are orthogonal, i.e. $\boldsymbol{g}_k \perp \boldsymbol{g}_{k+1}$.
2. The solution trajectory of the steepest-descent method with exact line search follows a zig-zag pattern.
3. Cuadratic case (exact step size): $\alpha_k = \frac{\boldsymbol{g}_k^T \boldsymbol{g}_k}{\boldsymbol{g}_k^T \mathbf{Q} \boldsymbol{g}_k}$
4. General case (step size): $\alpha_k = \frac{\boldsymbol{g}_k^T \boldsymbol{g}_k}{\boldsymbol{g}_k^T \mathbf{H}_k \boldsymbol{g}_k}$
5. General case (step size approximation): $\alpha_k = \frac{\boldsymbol{g}_k^T \boldsymbol{g}_k \hat{\alpha}^2}{2(\hat{f} - f_k + \hat{\alpha} \boldsymbol{g}_k^T \boldsymbol{g}_k)}$ with $\hat{f} = f(x_k - \hat{\alpha} \boldsymbol{g}_k)$

Steepest descent Method: Cuadratic case

Steepest descent: Cuadratic case

1. In the steepest decent algorithm $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \alpha_k \boldsymbol{g}_k$, with exact line search, i.e. $\alpha_k = \frac{\boldsymbol{g}_k^T \boldsymbol{g}_k}{\boldsymbol{g}_k^T \mathbf{Q} \boldsymbol{g}_k}$, we have that $\boldsymbol{x}_k \rightarrow \boldsymbol{x}^*$ for any $\boldsymbol{x}_0$, i.e. converges globally.
2. In the steepest decent algorithm $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \alpha \boldsymbol{g}_k$, with with fixed step size, i.e. $\alpha_k = \alpha$ for all k, we have that $\boldsymbol{x}_k \rightarrow \boldsymbol{x}^*$ for any $\boldsymbol{x}_0$ iff $0 < \alpha < \frac{2}{\lambda_{max}(\mathbf{Q})}$, i.e. converges globally.
3. The method of steepest descent converges linearly with a ratio no greater than $1 - \frac{1}{\kappa}$.
4. The steepest descent algorithm with exact line search with $\boldsymbol{g}_k \neq \mathbf{0}$ converges in one iteration iff $\boldsymbol{g}_k$ is an eigenvector of $\mathbf{Q}$

Non-quadratic case

Theorem 1.1

Non-quadratic case Suppose f is defined on $\mathbb{R}^n$, has continuous second partial derivatives, and has a relative minimum at $\boldsymbol{x}^*$. Suppose further that the Hessian matrix of f, $\mathbf{H}(x^*)$, has smallest eigenvalue $a > 0$ and largest eigenvalue $A > 0$. If $\{\boldsymbol{x}_k\}$ is a sequence generated by the method of steepest descent that converges to $\boldsymbol{x}^*$, then the sequence of objective values $\{f(\boldsymbol{x}_k)\}$ converges to $f(\boldsymbol{x}^*)$ linearly with a convergence ratio no greater than $\left(\frac{A-a}{A+a}\right)^2$, i.e., for all k sufficiently large, we have

$$f(\boldsymbol{x}_{k+1}) - f(\boldsymbol{x}^*) \leq \left(\frac{A-a}{A+a}\right)^2 [f(\boldsymbol{x}_k) - f(\boldsymbol{x}^*)]$$

Newton's Method for Systems of NonLinear Equations

Problem

An unconstraint optimization problems, in general, yields the following System of NonLinear Equations

$$\boldsymbol{g}(\boldsymbol{x}) = 0$$

where, in our case, $\boldsymbol{g} \overset{def}{=} \nabla f : \mathbb{R}^n \to \mathbb{R}^n$ is assumed to be continuously differentiable.

Newton's Method for Systems of NonLinear Equations

Algorithm

Given $\boldsymbol{g} : \mathbb{R}^n \to \mathbb{R}^n$ continuously differentiable and $\boldsymbol{x}_0 \in \mathbb{R}^n$: at each iteration k, (Note: in our case $\boldsymbol{g} = \nabla f$ and $D\boldsymbol{g} = \nabla^2 f = \mathbf{H}$)

1. Solve $D\boldsymbol{g}(\boldsymbol{x}_k)\boldsymbol{d}_k = -\boldsymbol{g}(\boldsymbol{x}_k)$
2. Update $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \boldsymbol{d}_k$

The derivative of $\boldsymbol{g} = \nabla f(\cdot)$ at $\boldsymbol{x}$ is the Jacobian (matrix) of $\boldsymbol{g} = \nabla f(\cdot)$ at $\boldsymbol{x}$, denoted here as $\mathbf{J}(\boldsymbol{x}_k) = D\boldsymbol{g}(\boldsymbol{x}_k)$, or the Hessian of f, ie, $\mathbf{H}(\boldsymbol{x}_k) = D\boldsymbol{g}(\boldsymbol{x}_k) = \nabla^2 f(\boldsymbol{x}_k)$.

Advantages of Newton's method

1. Quadratically convergence from good starting guesses if $\mathbf{H}(\boldsymbol{x}^*) = \nabla^2 f(\boldsymbol{x}^*)$ is nonsingular.
2. Exact solution in one iteration for an affine ∇f (exact at each iteration for any affine component functions of ∇f), i.e., for quadratic problem.

Disadvantages of Newton's method

1. In general, it is not globally convergent.
2. Requires the computation of $\mathbf{H}(\boldsymbol{x}_k) = \nabla^2 f(\boldsymbol{x}_k)$ at each iteration.
3. It requires, at each iteration, the solution of a system of linear equations that may be singular or ill-conditioned.
4. $\boldsymbol{d}_k = -\mathbf{H}(\boldsymbol{x}_k)^{-1}\boldsymbol{g}(\boldsymbol{x}_k)$ could not be a descent direction.

Local convergence of Newton's Method

Theorem 2.1

Suppose that $f \in \mathcal{C}^3$,and $\boldsymbol{x}^* \in \mathbb{R}^n$ is a point such that $\nabla f(\boldsymbol{x}^*) = 0$ and $\mathbf{H}(\boldsymbol{x}^*)$ is invertible. Then, for all $\boldsymbol{x}_0$ sufficiently close to $\boldsymbol{x}^*$, Newton's method is well defined for all k, and converges to $\boldsymbol{x}^*$ with order of convergence at least 2.

Local convergence of Newton's Method

proof

Using Taylor expansion of ∇f around $\boldsymbol{x}_0$

$$\nabla f(\boldsymbol{x}) \quad = \quad \nabla f(\boldsymbol{x}_0) + \mathbf{H}(\boldsymbol{x}_0)(\boldsymbol{x} - \boldsymbol{x}_0) + O(\|\boldsymbol{x} - \boldsymbol{x}_0\|^2)$$

then by definition of $O(\cdot)$, big-Oh, there exist constants c_1, c_2 such that, if $\boldsymbol{x}_0, \boldsymbol{x} \in B(\boldsymbol{x}^*, \varepsilon)$ then

$$\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{x}_0) - \mathbf{H}(\boldsymbol{x}_0)(\boldsymbol{x} - \boldsymbol{x}_0)\| \quad \leq \quad c_1\|\boldsymbol{x} - \boldsymbol{x}_0\|^2$$

and by Lemma 3.2 lemma, for $\boldsymbol{x} \in B(\boldsymbol{x}^*, \varepsilon)$

$$\|\mathbf{H}(\boldsymbol{x})^{-1}\| \quad \leq \quad c_2$$

Local convergence of Newton's Method

proof

Assuming $\boldsymbol{x}_0 \in B(\boldsymbol{x}^*, \varepsilon)$ and $\boldsymbol{x} = \boldsymbol{x}^*$, and as $\nabla f(\boldsymbol{x}^*) = 0$

$$\begin{aligned}
\|\boldsymbol{x}_1 - \boldsymbol{x}^*\| &= \|\boldsymbol{x}_0 - \mathbf{H}(\boldsymbol{x}_0)^{-1}\nabla f(\boldsymbol{x}_0) - \boldsymbol{x}^*\| \\
&= \|\mathbf{H}(\boldsymbol{x}_0)^{-1}[\mathbf{H}(\boldsymbol{x}_0)(\boldsymbol{x}_0 - \boldsymbol{x}^*) - \nabla f(\boldsymbol{x}_0)]\| \\
&= \|\mathbf{H}(\boldsymbol{x}_0)^{-1}[\nabla f(\boldsymbol{x}^*) - \nabla f(\boldsymbol{x}_0) - \mathbf{H}(\boldsymbol{x}_0)(\boldsymbol{x}^* - \boldsymbol{x}_0)]\| \\
&\leq \|\mathbf{H}(\boldsymbol{x}_0)^{-1}\|\|[\nabla f(\boldsymbol{x}^*) - \nabla f(\boldsymbol{x}_0) - \mathbf{H}(\boldsymbol{x}_0)(\boldsymbol{x}^* - \boldsymbol{x}_0)]\| \\
&\leq c_1 c_2 \|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2
\end{aligned}$$

if $\|\boldsymbol{x}_0 - \boldsymbol{x}^*\| \leq \frac{\alpha}{c_1 c_2}$ with $\alpha \in (0, 1)$ then

$$\|\boldsymbol{x}_1 - \boldsymbol{x}^*\| \quad \leq \quad \alpha\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|$$

Local convergence of Newton's Method

proof

by induction

$$\|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\| \;\; \leq \;\; \alpha\|\boldsymbol{x}_k - \boldsymbol{x}^*\|$$

which implies that

$$\lim_{k\to\infty} \|\boldsymbol{x}_k - \boldsymbol{x}^*\| \;\; = \;\; 0$$

and therefore $\boldsymbol{x}_k$ converges to $\boldsymbol{x}^*$.
(About convergence order): as

$$\|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\| \;\; \leq \;\; c_1c_2\|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2$$

then $\|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\| = O(\|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2)$ then the order of convergence at least 2.

Comments

1. Newton's method has superior convergence properties if the starting point is near the solution.
2. However, the method is not guaranteed to converge to the solution if we start far away from it (in fact, it may not even be well defined because the Hessian may be singular).
3. The method may not be a descent method; that is, it is possible that $f(\boldsymbol{x}_{k+1}) > f(\boldsymbol{x}_k)$.
4. It is possible to modify the algorithm such that the descent property holds.

Theorem 2.2

Let $\{\boldsymbol{x}_k\}$ *a sequence generated by Newton's method for minimizing a function* $f(\boldsymbol{x})$. *If the Hessian* $\mathbf{H}(\boldsymbol{x}_k) \succ 0$ *and* $\boldsymbol{g}_k = \nabla f(\boldsymbol{x}_k) \neq \mathbf{0}$ *then, the direction*

$$\boldsymbol{d}_k \;\; = \;\; -\mathbf{H}(\boldsymbol{x}_k)^{-1}\boldsymbol{g}(\boldsymbol{x}_k) = \boldsymbol{x}_{k+1} - \boldsymbol{x}_k$$

is a descent direction

Proof.

It is straightforward,

$$\boldsymbol{g}_k^T \boldsymbol{d}_k \;=\; -\boldsymbol{g}_k^T \mathbf{H}(\boldsymbol{x}_k)^{-1} \boldsymbol{g}_k < 0$$

□

Recall: Newton's Method with line search

1. Compute $\boldsymbol{d}_k = -\mathbf{H}(\boldsymbol{x}_k)^{-1}\boldsymbol{g}_k$
2. Find $\alpha_k = \arg\min_{\alpha>0} f(\boldsymbol{x}_k + \alpha\boldsymbol{d}_k)$
3. Update $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k\boldsymbol{d}_k$

According to Theorem 2.2 if $\mathbf{H}(\boldsymbol{x}_k) \succ 0$ then the previous algorithm guarantees the descent property, i.e.,

$$f(\boldsymbol{x}_{k+1}) < f(\boldsymbol{x}_k)$$

whenever $\boldsymbol{g}_k \neq \mathbf{0}$.

Kantorovich Theorem

Theorem 2.3

Let $r > 0$, $\boldsymbol{x}_0 \in \mathbb{R}^n$, $F : \mathbb{R}^n \to \mathbb{R}^n$, *and assume that* F *is continuously differentiable in* $B(\boldsymbol{x}_0, r)$. *Assume for a vector norm and the induced operator norm that* $\mathbf{H}(\boldsymbol{x}) = \nabla \boldsymbol{g}(\boldsymbol{x})$ *is Lipschitz with constant* γ *in* $B(\boldsymbol{x}_0, r)$ *with* $\mathbf{H}(\boldsymbol{x}_0)$ *nonsingular, and there exist constants* β, η *such that*

$$\|\mathbf{H}(\boldsymbol{x}_0)\| \leq \beta, \ \|\mathbf{H}(\boldsymbol{x}_0)^{-1} \boldsymbol{g}(\boldsymbol{x}_0)\| \leq \eta,$$

If $\beta\gamma\eta \leq \frac{1}{2}$ *and* $r \geq r_0 \equiv (1 - \sqrt{1 - 2\alpha})/(\beta\gamma)$, *then the sequence* $\{\boldsymbol{x}_k\}$ *produced by Newton's is well defined and converges to* $\boldsymbol{x}^*$, *a unique zero of F in the closure of* $B(\boldsymbol{x}_0, r_0)$. *If* $\beta\gamma\eta < \frac{1}{2}$ *, then* $\boldsymbol{x}^*$ *is the unique zero of* F *in* $B(\boldsymbol{x}_0, r_1)$, *where* $r_1 = \min(r, (1 + \sqrt{1 - 2\alpha})/(\beta\gamma))$ *and*

$$\|\boldsymbol{x}_k - \boldsymbol{x}^*\| \leq (2\alpha)^{2^k} \eta/\alpha, \ k = 1, 2, \cdots$$

1. The Kantorovich theorem is a second convergence result for Newton's method (different and powerful result).
2. Its assumptions and method of proof are related to a classical result on the convergence of iterative algorithms called the contractive mapping theorem.
3. The Kantorovich theorem differs from Theorem 2.1 mainly in that it makes no assumption about the existence of a solution to $\boldsymbol{g}(\boldsymbol{x}^*) = 0$.
4. It shows that if $\mathbf{H}(\boldsymbol{x}_0)$ is nonsingular, $\mathbf{H}$ is Lipschitz continuous in a region containing $\boldsymbol{x}_0$, and the first step of Newton's method is sufficiently small relative to the nonlinearity of $\boldsymbol{g}(\cdot)$, then there must be a root in this region, and furthermore it is unique.

Contractive Mapping Theorem

Theorem 2.4

Let $\boldsymbol{h} : D \to D$, D *a closed subset of* $\mathbb{R}^n$. *If for some norm* $\|\cdot\|$ *there exists* $L \in [0,1)$ *such that* Lipschitz

$$\|\boldsymbol{h}(\boldsymbol{x}) - \boldsymbol{h}(\boldsymbol{y})\| \quad \leq \quad L\|\boldsymbol{x} - \boldsymbol{y}\|$$

then

1. *there exits a unique* $\boldsymbol{x}^* \in D$ *such that* $\boldsymbol{h}(\boldsymbol{x}^*) = \boldsymbol{x}^*$
2. *for any* $\boldsymbol{x}_0 \in D$ *the sequence generated by* $\boldsymbol{x}_{k+1} = \boldsymbol{h}(\boldsymbol{x}_k)$, *remains in* D *and converges linearly to* $\boldsymbol{x}^*$ *with constant* α
3. *for any* $\eta \geq \|\boldsymbol{h}(\boldsymbol{x}_0) - \boldsymbol{x}_0\|$, *(with* $\boldsymbol{h}(\boldsymbol{x}_0) = \boldsymbol{x}_1$*)*

$$\|\boldsymbol{x}_k - \boldsymbol{x}^*\| \leq \frac{\eta L^k}{1-L}, \; k = 0, 1, \cdots$$

Hessian modification

- If the Hessian matrix $\nabla^2 f(\boldsymbol{x})$ is not positive definite, the Newton direction $\boldsymbol{d}_k^N$

$$\nabla^2 f(\boldsymbol{x}_k)\boldsymbol{d}_k^N \quad = \quad -\nabla f(\boldsymbol{x}_k)$$

  may not be a descent direction.

- One alternative to solve the previous problem is to modify the Hessian, ie,

$$\mathbf{B}_k \quad = \quad \nabla^2 f(\boldsymbol{x}_k) + \mathbf{E}_k$$

  such that $\mathbf{B}_k \succ 0$ and the new direction

$$\mathbf{B}_k \boldsymbol{d}_k \quad = \quad -\nabla f(\boldsymbol{x}_k)$$

  is a descent direction

Line Search Newton with Modification

```
Require: $\boldsymbol{x}_0$
k = 0
while $\|\nabla f_k\| > \tau_g$ do
    Factorize the matrix $B_k = \nabla^2 f(\boldsymbol{x}_k) + \mathbf{E}_k$ where $\mathbf{E}_k = 0$ if $\nabla^2 f(\boldsymbol{x}_k)$ is sufficiently positive definite; otherwise, $\mathbf{E}_k$ is chosen to ensure that $\mathbf{B}_k$ is sufficiently positive definite
    Solve $\mathbf{B}_k \boldsymbol{d}_k = -\nabla f(\boldsymbol{x}_k)$
    Compute $\alpha_k$ (Wolfe, Goldstein, or Armijo conditions)
    Set $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k \boldsymbol{d}_k$
    $k = k + 1$
end while
```

Comments

- The global convergence can be stablished is the sequence $\{\mathbf{B}_k\}$ have bounded condition number (bounded modified factorization property) whenever the sequence of Hessians $\{\nabla^2 f(\boldsymbol{x}_k)\}$ is bounded; ie,

$$\kappa(\mathbf{B}_k) \quad = \quad \|\mathbf{B}_k\| \|\mathbf{B}_k^{-1}\| \leq C$$

  for $C > 0$, with $\|\cdot\| = \|\cdot\|_2$
- The global convergence follows from Zoutendijk's Theorem, becuase if $\kappa(\mathbf{B}_k) \leq C$ then $\cos\theta_k \geq 1/C$ which guarantees that

$$\lim_{k\to\infty} \|\nabla f(\boldsymbol{x}_k)\| \quad \to \quad 0$$

Comments

Using the following inequalities

$$
\begin{aligned}
\frac{\|\mathbf{A}^{-1}(\mathbf{A}\boldsymbol{x})\|}{\|\mathbf{A}\boldsymbol{x}\|} &\leq \|\mathbf{A}^{-1}\| \Longrightarrow \|\boldsymbol{x}\| \leq \|\mathbf{A}^{-1}\|\|\mathbf{A}\boldsymbol{x}\| \\
\frac{\boldsymbol{x}^T\mathbf{A}^{-1}\boldsymbol{x}}{\|\boldsymbol{x}\|^2} &\geq \frac{1}{\lambda_M(\mathbf{A})} = \frac{1}{\|\mathbf{A}\|_2}
\end{aligned}
$$

If $\mathbf{A} \succ 0$ is symmetric then $\lambda_M(\mathbf{A}) = \|\mathbf{A}\|_2$. Using $\mathbf{B}_k\boldsymbol{d}_k = -\boldsymbol{g}_k$ and $\mathbf{A} = \mathbf{B}_k$ we obtain

$$
\begin{aligned}
\cos\theta_k &= \frac{-\boldsymbol{d}_k^T\boldsymbol{g}_k}{\|\boldsymbol{d}_k\|\|\boldsymbol{g}_k\|} \geq \frac{-\boldsymbol{d}_k^T\boldsymbol{g}_k}{\|\mathbf{B}_k^{-1}\|\|\mathbf{B}_k\boldsymbol{d}_k\|\|\boldsymbol{g}_k\|} \\
&= \frac{\boldsymbol{g}_k^T\mathbf{B}_k^{-1}\boldsymbol{g}_k}{\|\boldsymbol{g}_k\|^2}\frac{1}{\|\mathbf{B}_k^{-1}\|} \geq \frac{1}{\|\mathbf{B}_k\|}\frac{1}{\|\mathbf{B}_k^{-1}\|} \geq \frac{1}{C}
\end{aligned}
$$

Theorem

Theorem 3.1

Let f be twice continuously differentiable on an open set D, and assume that the starting point x_0 of the previous Algorithm is such that the level set $L = \{\boldsymbol{x} \in D : f(\boldsymbol{x}) \leq f(\boldsymbol{x}_0)\}$ is compact. Then if the bounded modified factorization property holds, we have that

$$\lim_{k\to\infty} \|\nabla f(\boldsymbol{x}_k)\| \quad \to \quad 0$$

Cholesky with Added Multiple of the Identity

- We can simply select $\mathbf{E_k} = \tau_k \mathbf{I}$ then

  $$\mathbf{B}_k \;=\; \nabla^2 f(\boldsymbol{x}_k) + \tau_k \mathbf{I}$$

  with $\tau_k \geq 0$ such that it ensures that $\mathbf{B}_k$ is sufficiently positive definite.
- We can compute τ_k based on the smallest eigenvalue of the Hessian $\nabla^2 f(\boldsymbol{x}_k)$, but this is not always possible or it is computationally expensive.

Cholesky with Added Multiple of the Identity

```
Choose $\beta > 0$, ie, $\beta = 1e-3$
if $\min_i(a_{ii}) > 0$ then
    $\tau_0 = 0$
else
    $\tau_0 = -\min_i(a_{ii}) + \beta$
end if
for  $k = 0, 1, \ldots$ do
    Attempt to apply the Cholesky to obtain $\mathbf{LL}^T = \mathbf{A} + \tau_k \mathbf{I}$
    if The decomposition is successful  then
        Return $\mathbf{L}$
    else
        $\tau_{k+1} = \max(2\tau_k, \beta)$
    end if
end for
```

Modified Cholesky factorization

- Another approach for modifying a Hessian matrix is to increase the diagonal elements encountered during the factorization to ensure that they are sufficiently positive.
- For this, we can use the modified Cholesky.
- A symmetric positive definite matrix $\mathbf{A}$ can be written as

$$\mathbf{A} = \mathbf{L}\mathbf{D}\mathbf{L}^T$$

  where $\mathbf{L}$ is a lower triangular matrix with unit diagonal elements and $\mathbf{D}$ is a diagonal matrix with positive elements on the diagonal.
- Recall: The Cholesky decomposition of a positive-definite matrix $\mathbf{A}$ is a decomposition of the form $\mathbf{A} = \mathbf{L}\mathbf{L}^T$, where $\mathbf{L}$ is a lower triangular matrix with real and positive diagonal entries.

Modified Cholesky factorization

For example, we can write a 4-by-4 symmetric matrix as follows:

$$\begin{bmatrix} a_{11} & a_{21} & a_{31} & a_{41} \\ a_{21} & a_{22} & a_{32} & a_{42} \\ a_{31} & a_{32} & a_{33} & a_{43} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ l_{21} & 1 & 0 & 0 \\ l_{31} & l_{32} & 1 & 0 \\ l_{41} & l_{42} & l_{43} & 1 \end{bmatrix} \begin{bmatrix} d_1 & 0 & 0 & 0 \\ 0 & d_2 & 0 & 0 \\ 0 & 0 & d_3 & 0 \\ 0 & 0 & 0 & d_4 \end{bmatrix} \begin{bmatrix} 1 & l_{21} & l_{31} & l_{41} \\ 0 & 1 & l_{32} & l_{42} \\ 0 & 0 & 1 & l_{43} \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Computing the matrix products of the left side and equating the elements of both sides of the equality, we obtain (next slide)

Modified Cholesky factorization

First we compare the diagonal elements

$$\begin{bmatrix} a_{11} & a_{21} & a_{31} & a_{41} \\ a_{21} & a_{22} & a_{32} & a_{42} \\ a_{31} & a_{32} & a_{33} & a_{43} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix}$$

Modified Cholesky factorization

Secondly, we compare the entries below the diagonal

$$\begin{bmatrix} a_{11} & a_{21} & a_{31} & a_{41} \\ a_{21} & a_{22} & a_{32} & a_{42} \\ a_{31} & a_{32} & a_{33} & a_{43} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix}$$

Modified Cholesky factorization

We can compute the elements d_j of the diagonal matrix $\mathbf{D}$

$$\begin{aligned} a_{11} &= d_1 \\ a_{22} &= d_1 l_{21}^2 + d_2 \\ a_{33} &= d_1 l_{31}^2 + d_2 l_{32}^2 + d_3 \end{aligned}$$

$$a_{44} = d_1 l_{41}^2 + d_2 l_{42}^2 + d_3 l_{43}^2 + d_4$$

$$d_4 = a_{44} - d_1 l_{41}^2 - d_2 l_{42}^2 - d_3 l_{43}^2$$

Then, in general

$$\boxed{d_j = a_{jj} - \sum_{s=1}^{j-1} d_s l_{js}^2}$$

for $j = 1, 2, ..., n$

Modified Cholesky factorization

We also can compute the entries l_{ij}, $i > j$ of the $\mathbf{L}$ matrix

$$\begin{aligned} a_{21} &= d_1 l_{21} \\ a_{31} &= d_1 l_{31};\ a_{32} = d_1 l_{31} l_{21} + d_2 l_{32} \\ a_{41} &= d_1 l_{41};\ a_{42} = d_1 l_{41} l_{21} + d_2 l_{42};\ a_{43} = d_1 l_{41} l_{31} + d_2 l_{42} l_{32} + d_3 l_{43} \end{aligned}$$

$$l_{43} = \frac{1}{d_3}[a_{43} - d_1 l_{41} l_{31} - d_2 l_{42} l_{32}]$$

Then, in general

$$\boxed{l_{ij} = \frac{1}{d_j}[a_{ij} - \sum_{s=1}^{j-1} d_s l_{is} l_{js}]}$$

for $j = 1, 2, ..., n$ and $i = j + 1, ..., n$

Cholesky Factorization, $\mathbf{LDL}^T$ Form

```
for $j = 0, 1, ..., n$ do
    $d_j = a_j - \sum_{s=1}^{j-1} d_s l_{js}^2$
    for $i = j+1, ..., n$ do
        $l_{ij} = \frac{1}{d_j}[a_{ij} - \sum_{s=1}^{j-1} d_s l_{is} l_{js}]$
    end for
end for
```

Modified Cholesky factorization

Step 1: Compute d_1

$$\begin{aligned} a_{11} &= d_1 \\ a_{22} &= d_1 l_{21}^2 + d_2 \\ a_{33} &= d_1 l_{31}^2 + d_2 l_{32}^2 + d_3 \\ a_{44} &= d_1 l_{41}^2 + d_2 l_{42}^2 + d_3 l_{43}^2 + d_4 \end{aligned}$$

Modified Cholesky factorization

Step 2: Compute l_{21}, l_{31}, l_{41}

$$\begin{aligned} a_{11} &= d_1 \\ a_{22} &= d_1 {l_{21}}^2 + d_2 \\ a_{33} &= d_1 {l_{31}}^2 + d_2 l_{32}^2 + d_3 \\ a_{44} &= d_1 {l_{41}}^2 + d_2 l_{42}^2 + d_3 l_{43}^2 + d_4 \end{aligned}$$

Modified Cholesky factorization

Step 3: Compute d_2

$$\begin{aligned} a_{11} &= d_1 \\ a_{22} &= d_1 {l_{21}}^2 + d_2 \\ a_{33} &= d_1 {l_{31}}^2 + d_2 l_{32}^2 + d_3 \\ a_{44} &= d_1 {l_{41}}^2 + d_2 l_{42}^2 + d_3 l_{43}^2 + d_4 \end{aligned}$$

Modified Cholesky factorization

Step 4: Compute l_{32}, l_{42}

$$\begin{aligned} a_{11} &= d_1 \\ a_{22} &= d_1 {l_{21}}^2 + d_2 \\ a_{33} &= d_1 {l_{31}}^2 + d_2 {l_{32}}^2 + d_3 \\ a_{44} &= d_1 {l_{41}}^2 + d_2 {l_{42}}^2 + d_3 l_{43}^2 + d_4 \end{aligned}$$

Modified Cholesky factorization

Step 5: Compute d_3

$$
\begin{aligned}
a_{11} &= d_1 \\
a_{22} &= d_1 {l_{21}}^2 + d_2 \\
a_{33} &= d_1 {l_{31}}^2 + d_2 {l_{32}}^2 + d_3 \\
a_{44} &= d_1 {l_{41}}^2 + d_2 {l_{42}}^2 + d_3 l_{43}^2 + d_4
\end{aligned}
$$

Modified Cholesky factorization

Step :6 Compute l_{43}

$$\begin{aligned} a_{11} &= d_1 \\ a_{22} &= d_1 {l_{21}}^2 + d_2 \\ a_{33} &= d_1 {l_{31}}^2 + d_2 {l_{32}}^2 + d_3 \\ a_{44} &= d_1 {l_{41}}^2 + d_2 {l_{42}}^2 + d_3 {l_{43}}^2 + d_4 \end{aligned}$$

Modified Cholesky factorization

Step :7 Compute d_4

$$\begin{aligned} a_{11} &= d_1 \\ a_{22} &= d_1 {l_{21}}^2 + d_2 \\ a_{33} &= d_1 {l_{31}}^2 + d_2 {l_{32}}^2 + d_3 \\ a_{44} &= d_1 {l_{41}}^2 + d_2 {l_{42}}^2 + d_3 {l_{43}}^2 + d_4 \end{aligned}$$

Cholesky Factorization, $\mathbf{LDL}^T$ Form

- If $\mathbf{A}$ is indefinite, the factorization $\mathbf{A} = \mathbf{LDL}^T$ may not exist. Even if it does exist, the algorithm is numerically unstable.

Cholesky Factorization, $\mathbf{LDL}^T$ Form

```
1: for $j = 0, 1, ..., n$ do
2:     $d_j = a_{jj} - \sum_{s=1}^{j-1} d_s l_{js}^2$
3:     for $i = j+1, ..., n$ do
4:         $l_{ij} = \frac{1}{d_j}[a_{ij} - \sum_{s=1}^{j-1} d_s l_{is} l_{js}]$
5:     end for
6: end for
```

Note: we can change **step 2** to guarantee $d_j \geq \delta > 0$. For example, it could be

$$d_j = \max(\delta, a_{jj} - \sum_{s=1}^{j-1} d_s l_{js}^2)$$

Cholesky Factorization, $\mathbf{LDL}^T$ Form

- However, in order to control the quality of the modification, two parameters δ and β are selected
- The following bounds should be satisfied:

$$d_j \geq \delta,\ |m_{ij}| \leq \beta,\ i = j+1, j+2, ..., n$$

with $m_{ij} = l_{ij}\sqrt{d_j}$

- The diagonal elements d_j are computed with

$$\boxed{d_j = \max(\delta, a_{jj} - \textstyle\sum_{s=1}^{j-1} d_s l_{js}^2, \left(\frac{\theta_j}{\beta}\right)^2)} \qquad (1)$$

with $\theta_j = \max_{j<i\leq n} |a_{ij} - \sum_{s=1}^{j-1} d_s l_{is} l_{js}|$.

Cholesky Factorization, $\mathbf{LDL}^T$ Form

Note that, the previous selection guarantees that $|m_{ij}| \leq \beta$, ie, as $l_{ij} = [a_{ij} - \sum_{s=1}^{j-1} d_s l_{is} l_{js}]/d_j$ and $d_j \geq \left(\frac{\theta_j}{\beta}\right)^2$ due to (1)

$$\begin{aligned} |m_{ij}| &\leq |l_{ij}\sqrt{d_j}| = \frac{|a_{ij} - \sum_{s=1}^{j-1} d_s l_{is} l_{js}|}{\sqrt{d_j}} \\ &\leq \frac{|a_{ij} - \sum_{s=1}^{j-1} d_s l_{is} l_{js}|}{\theta_j}\beta \leq \beta \end{aligned}$$

due to $\theta_j \geq |a_{ij} - \sum_{s=1}^{j-1} d_s l_{is} l_{js}|$

Algorithm Cholesky Factorization, $\mathbf{LDL}^T$ Form

```
for $j = 0, 1, ..., n$ do
    $d_j = \max(\delta, a_{jj} - \sum_{s=1}^{j-1} d_s l_{js}^2, \left(\frac{\theta_j}{\beta}\right)^2)$ with
    $\theta_j = \max_{j<i\leq n} |a_{ij} - \sum_{s=1}^{j-1} d_s l_{is} l_{js}|$.
    for $i = j + 1, ..., n$ do
        $l_{ij} = \frac{1}{d_j}[a_{ij} - \sum_{s=1}^{j-1} d_s l_{is} l_{js}]$
    end for
end for
```

Lemma 3.2

Let $\mathbf{F} : \mathbb{R}^n \to \mathbb{R}^{n\times n}$ be an $n \times n$ matrix-valued function that is continuous at $\boldsymbol{x}_0$. If $\mathbf{F}(\boldsymbol{x}_0)^{-1}$ exists, then for $\mathbf{F}(\boldsymbol{x})^{-1}$ exists for $\boldsymbol{x}$ sufficiently close to $\boldsymbol{x}_0$, and $\mathbf{F}(\cdot)^{-1}$ is continuous at $\boldsymbol{x}_0$.

As $\mathbf{F}(\cdot)^{-1}$ is continuous at $\boldsymbol{x}_0$, we can select a ball $B(\boldsymbol{x}_0, r) \overset{def}{=} \{\boldsymbol{x} | \, \|\boldsymbol{x} - \boldsymbol{x}_0\| < r\}$ for which $\|\mathbf{F}(\cdot)^{-1}\|$ is bounded, i.e. there exists c such that

$$\|\mathbf{F}(\boldsymbol{x})^{-1}\| \leq c$$

for all $\boldsymbol{x} \in B(\boldsymbol{x}_0, r)$.
(By continuity definition) For any $\varepsilon > 0$ there exists $\delta > 0$ such that, $\|\boldsymbol{x} - \boldsymbol{x}_0\| < \delta$ implies $\|f(\boldsymbol{x}) - f(\boldsymbol{x}_0)\| < \varepsilon$, then $\|f(\boldsymbol{x})\| < \|f(\boldsymbol{x}_0)\| + \|f(\boldsymbol{x}) - f(\boldsymbol{x}_0)\| < \|f(\boldsymbol{x}_0)\| + \varepsilon$. Hence $\|f(\boldsymbol{x})\|$ is bounded for all $\boldsymbol{x} \in B(\boldsymbol{x}_0, \delta)$. back

Definition 3.3

A function $f : \mathbb{R}^n \to \mathbb{R}^m$ is Lipschitz continuous with constant $L \geq 0$ if for every $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$,

$$\|f(\boldsymbol{x}) - f(\boldsymbol{y})\| \leq L\|\boldsymbol{x} - \boldsymbol{y}\|$$

Any such L is referred to as a *Lipschitz constant* for the function f.

1. A function is called *locally Lipschitz continuous* if for every $\boldsymbol{x} \in \mathbb{R}^n$ there exists a neighborhood U of $\boldsymbol{x}$ such that f restricted to U is Lipschitz continuous.
2. If $0 \leq L < 1$ the function f is called a contraction.
3. The smallest constant is called the *best Lipschitz constant*. However, in many cases, it is not required to know this constant. back