

# Numerical Optimization

## Unit 2: Multivariable optimization problems

Che-Rung Lee

Department of Computer Science  
National Tsing Hua University

September 19, 2022

# Partial derivative of a two variable function

- Given a two variable function  $f(x_1, x_2)$ .
- The partial derivative of  $f$  with respect to  $x_i$  is

$$\begin{cases} \frac{\partial f}{\partial x_1} = \lim_{h \rightarrow 0} \frac{f(x_1 + h, x_2) - f(x_1, x_2)}{h} \\ \frac{\partial f}{\partial x_2} = \lim_{h \rightarrow 0} \frac{f(x_1, x_2 + h) - f(x_1, x_2)}{h} \end{cases}$$

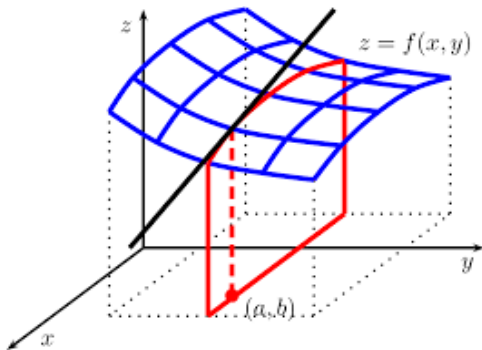
- The meaning of partial derivative: let  $F(x_1) = f(x_1, v)$  and  $G(x_2) = f(u, x_2)$ ,

$$\frac{\partial f}{\partial x_1}(x_1, v) = F'(x_1).$$

$$\frac{\partial f}{\partial x_2}(u, x_2) = G'(x_2).$$

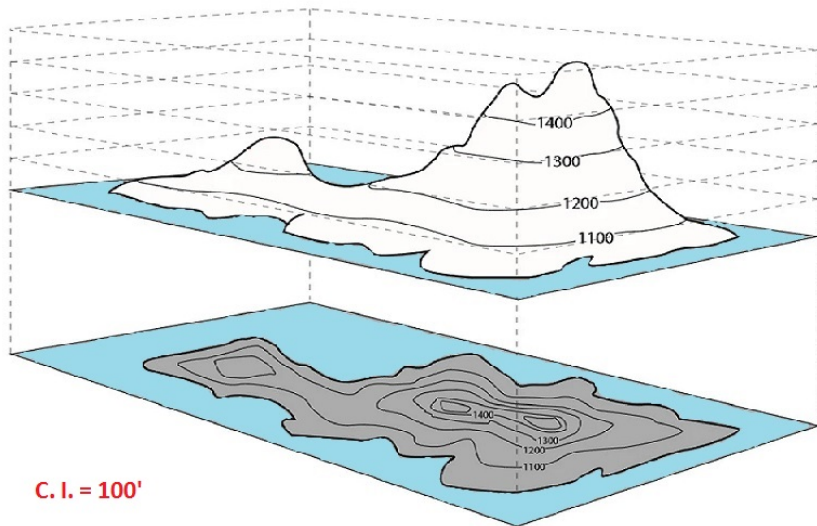
# Meaning of partial derivative

- Think the surface of  $z = f(x, y)$  as a cake. For  $\partial f / \partial x$ , think we cut the cake along the  $x$ -axis, and we look at the cut section, which is a curve in the  $x$ - $z$  plane, as the red lines in the figure.
- For a point  $(a, b)$ , the partial derivative  $\partial f / \partial x(a, b)$  is slope of the tangent line on the cutting plane.



(from <https://web.maths.unsw.edu.au/>)

# Level curve and contour plot



**C. I. = 100'**

(from <http://academic.brooklyn.cuny.edu/geology/grocha/mapcontour/>)

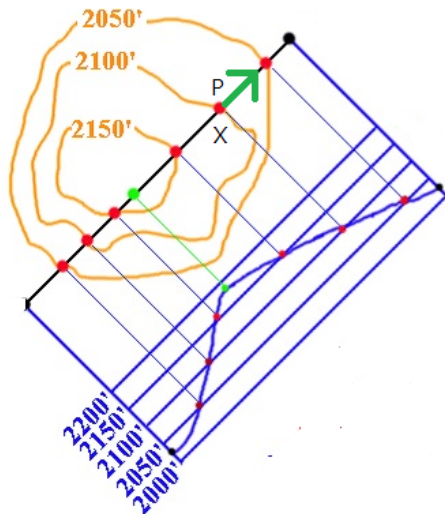
# Directional derivative

How about the directions other than  $x$  and  $y$ ?

## Definition

The directional derivative of a function  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  in the direction  $\vec{p}$  is defined as

$$D(f(\vec{x}), \vec{p}) = \lim_{h \rightarrow 0} \frac{f(\vec{x} + h\vec{p}) - f(\vec{x})}{h}$$



(from <https://www.geogebra.org/m/bxhwxr2x>)

# How to compute the directional derivative?

Suppose  $\vec{x} = (a, b)$  and  $\vec{p} = (p_x, p_y)$ . Also, we assume  $\|\vec{p}\| = 1$ .

$$\begin{aligned} D(f(\vec{x}), \vec{p}) &= \lim_{h \rightarrow 0} \frac{f(\vec{x} + h\vec{p}) - f(\vec{x})}{h} \\ &= \lim_{h \rightarrow 0} \frac{f(a + hp_x, b + hp_y) - f(a, b)}{h} \\ &= \lim_{h \rightarrow 0} \frac{f(a + hp_x, b + hp_y) - f(a + hp_x, b)}{h} + \frac{f(a + hp_x, b) - f(a, b)}{h} \\ &= \lim_{h \rightarrow 0} p_y \frac{f(a + hp_x, b + hp_y) - f(a + hp_x, b)}{p_y h} + p_x \frac{f(a + hp_x, b) - f(a, b)}{p_x h} \\ &= p_y \frac{\partial f}{\partial y} + p_x \frac{\partial f}{\partial x} = \left\langle \begin{bmatrix} \partial f / \partial x \\ \partial f / \partial y \end{bmatrix}, \begin{bmatrix} p_x \\ p_y \end{bmatrix} \right\rangle \end{aligned}$$

which is the inner product of  $(\partial f / \partial x, \partial f / \partial y)$  and  $\vec{p}$ .

## Definition

The gradient of a function  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  is a **vector** in  $\mathbb{R}^n$  defined as

$$\vec{g} = \nabla f(\vec{x}) = \begin{pmatrix} \partial f / \partial x_1 \\ \vdots \\ \partial f / \partial x_n \end{pmatrix}, \text{ where } \vec{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$$

## Remark

If  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  is continuously differentiable in a neighborhood of  $\vec{x}$ ,

$$D(f(\vec{x}), \vec{p}) = \nabla f(\vec{x})^T \vec{p},$$

for any vector  $\vec{p}$ .

# Gradient and tangent plane

- Let  $z = f(x, y)$  be the surface of  $f(x, y)$ . We can rewrite it as

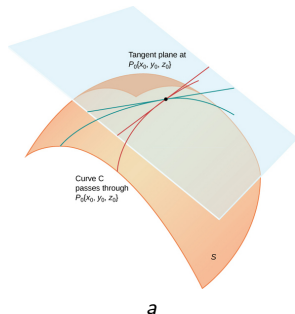
$$F(x, y, z) = f(x, y) - z = 0.$$

- At a point  $(x_0, y_0, z_0)$ , the tangent plane of  $F$  is the plane passing  $(x_0, y_0, z_0)$  and has the same normal vector as  $F$ .
- The normal vector of  $F$  at  $(x_0, y_0, z_0)$  is  $(\partial F / \partial x, \partial F / \partial y, \partial F / \partial z)$ . The plane equation is

$$\frac{\partial F}{\partial x}(x - x_0) + \frac{\partial F}{\partial y}(y - y_0) + \frac{\partial F}{\partial z}(z - z_0) = 0$$

or

$$z = f(x_0, y_0) + \frac{\partial F}{\partial x}(x - x_0) + \frac{\partial F}{\partial y}(y - y_0). \quad (1)$$

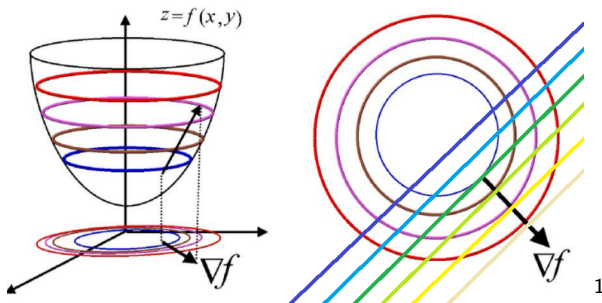


<sup>a</sup><https://math.libretexts.org/Bookshelves/C>



# Gradient on the contour plot

- The level curves of the tangent plane are straight lines.
- At  $(x_0, y_0)$ , the level curve of tangent plane is tangent to the level of  $f(x)$ .
- The gradient is orthogonal to the level curves of the tangent plane.



<sup>1</sup><https://slidesplayer.com/slide/14873112/>

# The descent directions

- A direction  $\vec{p}$  is called a **descent direction** of  $f(\vec{x})$  at  $\vec{x}$  if  $D(f(\vec{x}_0), \vec{p}) < 0$ .
- If  $f$  is smooth enough,  $\vec{p}$  is a descent direction if  $\nabla f(\vec{x}_0)^T \vec{p} < 0$ .
- Which direction  $\vec{p}$ ,  $\|\vec{p}\| = 1$ , makes  $f(\vec{x}_0 + \vec{p})$  decreasing most?
  - We can use the tangent plane at  $x_0$  to approximate  $f(\vec{x})$  at  $x_0$ .
  - The generalization of (1) gives the tangent plane equation:

$$f(\vec{x}_0 + \vec{p}) = f(\vec{x}_0) + \nabla f(\vec{x}_0)^T \vec{p} \quad (2)$$

- Consider the meaning of inner product.
- When  $\vec{p} = -\nabla f(\vec{x}_0)/\|\nabla f(\vec{x}_0)\|$ ,  $f(\vec{x}_0 + \vec{p})$  has the smallest value.

$$f(\vec{x}_0 + \vec{p}) = f(\vec{x}_0) - \nabla f(\vec{x}_0)^T \nabla f(\vec{x}_0) / \|\nabla f(\vec{x}_0)\|$$

- The direction  $-\nabla f(\vec{x}_0)$  is called the steepest descent direction.

# The steepest descent algorithm

## The steepest descent algorithm

For  $k = 1, 2, \dots$  until convergence

Compute  $\vec{p}_k = -\nabla f(\vec{x}_k)$

Find  $\alpha_k \in (0, 1)$  s.t,  $F(\alpha_k) = f(\vec{x}_k + \alpha_k \vec{p}_k)$  is minimized.

$\vec{x}_{k+1} = \vec{x}_k + \alpha_k \vec{p}_k$

- You can use any single variable optimization techniques to compute  $\alpha_k$ .
- If  $F(\alpha_k) = f(\vec{x}_k + \alpha_k \vec{p}_k)$  is a quadratic function,  $\alpha_k$  has a theoretical formula. (will be derived in next slides.)
- If  $F(\alpha_k) = f(\vec{x}_k + \alpha_k \vec{p}_k)$  is more than a quadratic function, we may approximate it by a quadratic model and use the formula to solve  $\alpha_k$ .
- Higher order polynomial approximation will be mentioned in the line search algorithm.

# Quadratic model

- If  $f(\vec{x})$  is a quadratic function, we can write it as

$$f(x, y) = ax^2 + bxy + cy^2 + dx + ey + f(0, 0).$$

- If  $f$  is smooth, the derivatives of  $f$  are

$$\frac{\partial f}{\partial x} = 2ax + by + d, \quad \frac{\partial f}{\partial y} = 2cy + bx + e$$

$$\frac{\partial^2 f}{\partial x^2} = 2a, \quad \frac{\partial^2 f}{\partial y^2} = 2c, \quad \frac{\partial^2 f}{\partial x \partial y} = \frac{\partial^2 f}{\partial y \partial x} = b.$$

- Let  $\vec{x} = \begin{pmatrix} x \\ y \end{pmatrix}$ ,  $f(\vec{x})$  can be expressed as

$$f(\vec{x}) = \frac{1}{2} \vec{x}^T \begin{pmatrix} 2a & b \\ b & 2c \end{pmatrix} \vec{x} + \vec{x}^T \begin{pmatrix} d \\ e \end{pmatrix} + f(\vec{0}).$$

# Gradient and Hessian

- The gradient of  $f$ , as defined before, is

$$\vec{g}(\vec{x}) = \nabla f(\vec{x}) = \begin{pmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \end{pmatrix} = \begin{pmatrix} 2a & b \\ b & 2c \end{pmatrix} \vec{x} + \begin{pmatrix} d \\ e \end{pmatrix}$$

- The second derivative, which is a matrix called **Hessian**, is

$$\nabla^2 f(\vec{x}) = H(\vec{x}) = \begin{pmatrix} \frac{\partial^2 f}{\partial x^2} & \frac{\partial^2 f}{\partial x \partial y} \\ \frac{\partial^2 f}{\partial y \partial x} & \frac{\partial^2 f}{\partial y^2} \end{pmatrix} = \begin{pmatrix} 2a & b \\ b & 2c \end{pmatrix}$$

- Therefore,  $f(\vec{x}) = 1/2 \vec{x}^T H(\vec{0}) \vec{x} + \vec{g}(\vec{0})^T \vec{x} + f(\vec{0})$ ,

$$\nabla f(\vec{x}) = H\vec{x} + \vec{g}, \text{ and } \nabla^2 f = H$$

- In the following lectures, we assume  $H$  is symmetric. Thus,  $H = H^T$ .

# Optimal $\alpha_k$ for quadratic model

- We denote  $H_k = H(\vec{x}_k)$ ,  $\vec{g}_k = g(\vec{x}_k)$ , and  $f_k = f(\vec{x}_k)$ .
- Also,  $H = H(\vec{0})$ ,  $\vec{g} = g(\vec{0})$ , and  $f = f(\vec{0})$ .

$$\begin{aligned} F(\alpha) &= f(\vec{x}_k + \alpha \vec{p}_k) \\ &= \frac{1}{2}(\vec{x}_k + \alpha \vec{p}_k)^T H(\vec{x}_k + \alpha \vec{p}_k) + \vec{g}^T(\vec{x}_k + \alpha \vec{p}_k) + f(\vec{0}) \\ &= \frac{1}{2}\vec{x}_k^T H \vec{x}_k + \vec{g}^T \vec{x}_k + f(\vec{0}) + \alpha(H\vec{x}_k + \vec{g})^T \vec{p}_k + \frac{\alpha^2}{2}\vec{p}_k^T H \vec{p}_k \\ &= f_k + \alpha \vec{g}_k^T \vec{p}_k + \frac{\alpha^2}{2}\vec{p}_k^T H \vec{p}_k \\ F'(\alpha) &= \vec{g}_k^T \vec{p}_k + \alpha \vec{p}_k^T H \vec{p}_k \end{aligned}$$

The optimal solution of  $\alpha_k$  is at  $F'(\alpha) = 0$ , which is  $\alpha_k = \frac{-\vec{g}_k^T \vec{p}_k}{\vec{p}_k^T H \vec{p}_k}$

# Convergence of the steepest descent method

## Theorem (Convergence theorem of the steepest descent method)

*If the steepest descent method converges to a local minimizer  $\vec{x}^*$ , where  $\nabla^2 f(\vec{x})$  is positive definite, and  $e_{\max}$  and  $e_{\min}$  are the largest and the smallest eigenvalue of  $\nabla^2 f(\vec{x})$ , then*

$$\lim_{k \rightarrow \infty} \frac{\|\vec{x}_{k+1} - \vec{x}^*\|}{\|\vec{x}_k - \vec{x}^*\|} \leq \left( \frac{e_{\max} - e_{\min}}{e_{\max} + e_{\min}} \right)$$

## Definition

For a scalar  $\lambda$  and an unit vector  $\vec{v}$ ,  $(\lambda, \vec{v})$  is an eigenpair of of a matrix  $H$  if  $H\vec{v} = \lambda\vec{v}$ . The scalar  $\lambda$  is called an eigenvalue of  $H$ , and  $\vec{v}$  is called an eigenvector.

# Optimal condition

## Theorem (Necessary and sufficient condition of optimality)

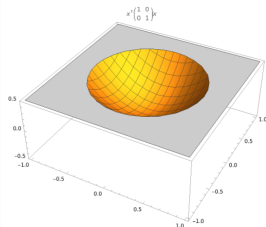
- Let  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  be continuously differentiable in  $D$ . If  $\vec{x}^* \in D$  is a local minimizer,  $\nabla f(\vec{x}^*) = 0$  and  $\nabla^2 f(\vec{x}^*)$  is **positive semidefinite**.
- If  $\nabla f(\vec{x}^*) = 0$  and  $\nabla^2 f(\vec{x}^*)$  is **positive definite**, then  $\vec{x}^*$  is a local minimizer.

## Definition

- A matrix  $H$  is called **positive definite** if for any nonzero vector  $\vec{v} \in \mathbb{R}^n$ ,  $\vec{v}^\top H \vec{v} > 0$ .
- $H$  is called **positive semidefinite** if  $\vec{v}^\top H \vec{v} \geq 0$  for all  $\vec{v} \in \mathbb{R}^n$ .
- $H$  is **negative definite** or **negative semidefinite** if  $-H$  is positive definite or positive semidefinite.
- $H$  is **indefinite** if it is neither positive semidefinite nor negative semidefinite.



# Quadratic forms for 2D



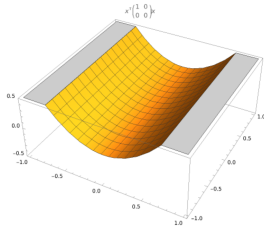
positive definite

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$\lambda_1 = 1, \lambda_2 = 1$$



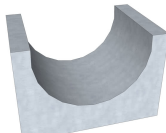
bowl



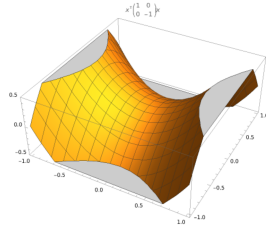
positive semidefinite

$$\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$$

$$\lambda_1 = 1, \lambda_2 = 0$$



half pipe



indefinite

$$\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$$

$$\lambda_1 = 1, \lambda_2 = -1$$



potato chip

<https://demonstrations.wolfram.com/EigenvaluesCurvatureAndQuadraticForms/>

# Newton's method

- We use the quadratic model to find the step length  $\alpha_k$ . Can we use the quadratic model to find the search direction  $\vec{p}_k$ ?
- Yes, we can. Recall the quadratic model (now  $\vec{p}$  is the variable.)

$$f(\vec{x}_k + \vec{p}) \approx \frac{1}{2} \vec{p}^T H_k \vec{p} + \vec{p}^T \vec{g}_k + f_k$$

- Compute the gradient  $\nabla_{\vec{p}} f(\vec{x}_k + \vec{p}) = H_k \vec{p} + \vec{g}_k$
- The solution of  $\nabla_{\vec{p}} f(\vec{x}_k + \vec{p}) = 0$  is  $\vec{p}_k = -H_k^{-1} \vec{g}_k$ .
- Newton's method uses  $\vec{p}_k$  as the search direction

## Newton's method

- 1 Given an initial guess  $\vec{x}_0$
- 2 For  $k = 0, 1, 2, \dots$  until converge

$$\vec{x}_{k+1} = \vec{x}_k - H_k^{-1} \vec{g}_k.$$

# Descent direction

- The direction  $\vec{p}_k = -H_k^{-1} \vec{g}_k$  is called Newton's direction
- Is  $\vec{p}_k$  a descent direction? (what's the definition of descent directions?)
- We only need to check if  $\vec{g}_k^T \vec{p}_k < 0$ .

$$\vec{g}_k^T \vec{p}_k = -\vec{g}_k^T H_k^{-1} \vec{g}_k.$$

Thus,  $\vec{p}_k$  is a descent direction if  $H^{-1}$  is positive definite.

- For a symmetric matrix  $H$ , the following conditions are equivalent
- $H$  is positive definite.
- $H^{-1}$  is positive definite.
- All the eigenvalues of  $H$  are positive.

## Example of Steepest Descent and Newton's Method

Let  $f(x, y) = \frac{1}{2}x^2 + \frac{9}{2}y^2$ . The gradient and the Hessian matrix are

$$\vec{g}_1 = \nabla f(x, y) = \begin{pmatrix} x \\ 9y \end{pmatrix}, H = \begin{pmatrix} 1 & 0 \\ 0 & 9 \end{pmatrix}.$$

The initial guess is  $\vec{x}_1 = \begin{pmatrix} 9 \\ 1 \end{pmatrix}$ . For the steepest descent method,

$$\vec{p}_1 = -\nabla f(9, 1) = \begin{pmatrix} -9 \\ -9 \end{pmatrix}, \alpha_1 = \frac{-\vec{g}_1^T \vec{p}_1}{\vec{p}_1^T H \vec{p}_1} = \frac{162}{9^3 + 9^2} = 0.2$$

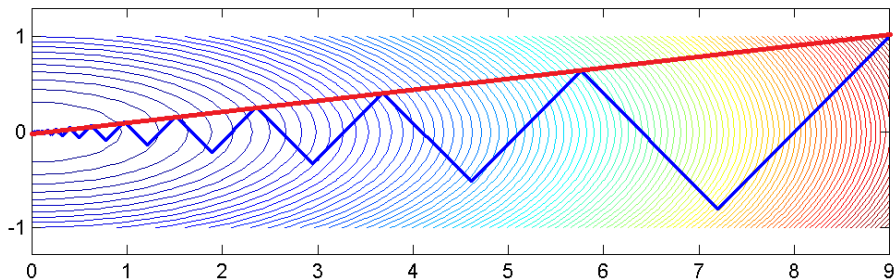
$$\vec{x}_2 = \begin{pmatrix} 9 \\ 1 \end{pmatrix} + \frac{1}{5} \begin{pmatrix} -9 \\ -9 \end{pmatrix} = \begin{pmatrix} 7.2 \\ -0.8 \end{pmatrix}$$

## Example—continue

For Newton's method,

$$\vec{p}_1 = -H^{-1}\vec{g}_1 = -\begin{pmatrix} 1 & 0 \\ 0 & 1/9 \end{pmatrix} \begin{pmatrix} 9 \\ 9 \end{pmatrix} = \begin{pmatrix} -9 \\ -1 \end{pmatrix},$$

So  $\vec{x}_2 = \begin{pmatrix} 9 \\ 1 \end{pmatrix} + \begin{pmatrix} -9 \\ -1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ , which is the optimal solution.



Blue line: Steepest Descent Method

Red line: Newton's Method

# Some properties of eigenvalues/eigenvectors

- A symmetric matrix  $H$ , of order  $n$  has  $n$  real eigenvalues and  $n$  real and linearly independent (orthogonal) eigenvectors

$$H\vec{v}_1 = \lambda_1 \vec{v}_1, \quad H\vec{v}_2 = \lambda_2 \vec{v}_2, \quad \dots, \quad H\vec{v}_n = \lambda_n \vec{v}_n$$

- Let  $V = [\vec{v}_1 \ \vec{v}_2 \ \dots \ \vec{v}_n]$ ,  $\Lambda = \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{bmatrix}$ ,  $HV = V\Lambda$ .

- If  $\lambda_1, \lambda_2, \dots, \lambda_n$  are nonzero, since  $H = V\Lambda V^{-1}$ ,

$$H^{-1} = V\Lambda^{-1}V^{-1}, \quad \Lambda^{-1} = \begin{bmatrix} 1/\lambda_1 & & & \\ & 1/\lambda_2 & & \\ & & \ddots & \\ & & & 1/\lambda_n \end{bmatrix}$$

The eigenvalues of  $H^{-1}$  are  $\frac{1}{\lambda_1}, \frac{1}{\lambda_2}, \dots, \frac{1}{\lambda_n}$ .

# How to solve $H\vec{p} = -\vec{g}$ ?

- For a symmetric positive definite matrix  $H$ ,  $H\vec{p} = -\vec{g}$  can be solved by Cholesky decomposition, which is similar to LU decomposition, but is only half computational cost of LU decomposition.

- Let  $H = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix}$ , where  $h_{12} = h_{21}$ ,  $h_{13} = h_{31}$ ,  $h_{23} = h_{32}$ .

Cholesky decomposition makes  $H = LL^T$ , where  $L$  is a lower

triangular matrix,  $L = \begin{bmatrix} \ell_{11} & & \\ \ell_{21} & \ell_{22} & \\ \ell_{31} & \ell_{32} & \ell_{33} \end{bmatrix}$

- Using Cholesky decomposition,  $H\vec{p} = -\vec{g}$  can be solved by
  - 1 Compute  $H = LL^T$
  - 2  $\vec{p} = -(L^T)^{-1}L^{-1}\vec{g}$
- In Matlab, use  $p = -H \setminus g$ . Don't use  $\text{inv}(H)$ .

# The Cholesky decomposition

For  $i = 1, 2, \dots, n$

$$\ell_{ii} = \sqrt{h_{ii}}$$

For  $j = i + 1, i + 2, \dots, n$

$$\ell_{ji} = \frac{h_{ji}}{\ell_{ii}}$$

For  $k = i + 1, i + 2, \dots, j$

$$h_{jk} = h_{jk} - \ell_{ji}\ell_{ki}$$

$$\begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix} = LL^T = \begin{bmatrix} \ell_{11}^2 & \ell_{11}\ell_{21} & \ell_{11}\ell_{31} \\ \ell_{11}\ell_{21} & \ell_{21}^2 + \ell_{22}^2 & \ell_{21}\ell_{31} + \ell_{22}\ell_{32} \\ \ell_{11}\ell_{31} & \ell_{21}\ell_{31} + \ell_{22}\ell_{32} & \ell_{31}^2 + \ell_{32}^2 + \ell_{33}^2 \end{bmatrix}$$

$$\ell_{11} = \sqrt{h_{11}} \quad h_{22}^{(2)} = h_{22} - \ell_{21}\ell_{21}$$

$$\ell_{21} = h_{21}/\ell_{11} \quad h_{32}^{(2)} = h_{32} - \ell_{21}\ell_{31}$$

$$\ell_{31} = h_{31}/\ell_{11} \quad h_{33}^{(2)} = h_{33} - \ell_{31}\ell_{31}$$

$$\ell_{22} = \sqrt{h_{22}^{(2)}}$$

$$\ell_{32} = h_{32}^{(2)}/\ell_{22}$$

$$\ell_{33} = \sqrt{h_{33}^{(2)} - \ell_{32}\ell_{32}}$$



# Convergence of Newton's method

## Theorem

*Suppose  $f$  is twice differentiable.  $\nabla^2 f$  is continuous in a neighborhood of  $\vec{x}^*$  and  $\nabla^2 f(\vec{x}^*)$  is positive definite, and if  $\vec{x}_0$  is sufficiently close to  $\vec{x}^*$ , the sequence converges to  $\vec{x}^*$  quadratically.*

## Three problems of Newton's method

- ①  $H$  may not be positive definite  $\Rightarrow$  Modified Newton's method + Line search.
- ②  $H$  is expensive to compute  $\Rightarrow$  Quasi-Newton.
- ③  $H^{-1}$  is expensive to compute  $\Rightarrow$  Conjugate gradient.