ECSE 343 Numerical Methods in Engineering

Roni Khazaka

Dept. of Electrical and Computer Engineering

McGill University

Stationary Iterative Methods

$$(\boldsymbol{U}-\boldsymbol{P})(\boldsymbol{U}+\boldsymbol{P}+\boldsymbol{P}^2+\boldsymbol{P}^3+\cdots+\boldsymbol{P}^n)=$$

$$=(\boldsymbol{U}-\boldsymbol{P})\sum_{k=0}^{n}\boldsymbol{P}^k \quad =\sum_{k=0}^{n}\boldsymbol{P}^k-\sum_{k=1}^{n+1}\boldsymbol{P}^k$$

$$=(\boldsymbol{U}+\boldsymbol{P}+\boldsymbol{P}^2+\boldsymbol{P}^3+\cdots+\boldsymbol{P}^n)-(\boldsymbol{P}+\boldsymbol{P}^2+\boldsymbol{P}^3+\cdots+\boldsymbol{P}^{n+1})$$

$$=\boldsymbol{U}-\boldsymbol{P}^{n+1}$$

Stationary Iterative Methods

Assume that the spectral Radius $\rho(\boldsymbol{N})$ of $\boldsymbol{N}$ is such that $\rho(\boldsymbol{N}) < 1$

$$\Rightarrow \lim_{n\to\infty} \boldsymbol{P}^n = \boldsymbol{0}$$

$$(\boldsymbol{U} - \boldsymbol{P}) \sum_{k=0}^{n} \boldsymbol{P}^k = \boldsymbol{U} - \boldsymbol{P}^{n+1} = \boldsymbol{U}$$

$$\Rightarrow (\boldsymbol{U} - \boldsymbol{P})^{-1} = \sum_{k=0}^{\infty} \boldsymbol{P}^k$$

Solution Algorithm

$$(\boldsymbol{U} - \boldsymbol{P})\boldsymbol{x} = \boldsymbol{b}$$

$$\boldsymbol{x} = (\boldsymbol{U} - \boldsymbol{P})^{-1}\boldsymbol{b} = \sum_{k=0}^{\infty} \boldsymbol{P}^k\,\boldsymbol{b} \cong \sum_{k=0}^{n} \boldsymbol{P}^k\,\boldsymbol{b}$$

At iteration n: $\boldsymbol{x}^{(n)} = \sum_{k=0}^{n} \boldsymbol{P}^k\,\boldsymbol{b}$

At iteration $n+1$: $\boldsymbol{x}^{(n+1)} = \sum_{k=0}^{n+1} \boldsymbol{P}^k\,\boldsymbol{b} = \boldsymbol{b} + \sum_{k=1}^{n+1} \boldsymbol{P}^k\,\boldsymbol{b}$

Solution Algorithm

At iteration n: $\boldsymbol{x}^{(n)} = \sum_{k=0}^{n} \boldsymbol{P}^k \boldsymbol{b}$

At iteration $n+1$: $\boldsymbol{x}^{(n+1)} = \sum_{k=0}^{n+1} \boldsymbol{P}^k \boldsymbol{b} = \boldsymbol{b} + \sum_{k=1}^{n+1} \boldsymbol{P}^k \boldsymbol{b}$

$$= \boldsymbol{b} + \boldsymbol{P} \sum_{k=0}^{n} \boldsymbol{P}^k \boldsymbol{b} \quad = \boldsymbol{b} + \boldsymbol{P}\boldsymbol{x}^{(n)}$$

$$\boxed{\boldsymbol{x}^{(n+1)} \leftarrow \boldsymbol{b} + \boldsymbol{P}\boldsymbol{x}^{(n)}}$$

Solution of $\boldsymbol{Ax} = \boldsymbol{b}$

$$(\boldsymbol{U} - \boldsymbol{P})\boldsymbol{x} = \boldsymbol{b}$$

Split $\mathbf{A}$ such that: $\boldsymbol{A} = \boldsymbol{U} - \boldsymbol{P}$

$$\boldsymbol{x}^{(n+1)} \leftarrow \boldsymbol{b} + \boldsymbol{P}\boldsymbol{x}^{(n)}$$

General Splittings

$$\boldsymbol{Ax} = \boldsymbol{b}$$

Split $\mathbf{A}$ such that: $\boldsymbol{A} = \boldsymbol{M} - \boldsymbol{N}$

$\mathbf{M}$ is such that it is easy to “invert” or to solve: $\boldsymbol{Mx} = \boldsymbol{b}$

Split $\mathbf{A}$ such that: $\boldsymbol{A} = \boldsymbol{M} - \boldsymbol{N} = \boldsymbol{M}(\boldsymbol{U} - \boldsymbol{M}^{-1}\boldsymbol{N})$

$$\boldsymbol{Ax} = \boldsymbol{b} \quad \Rightarrow \quad \boldsymbol{M}(\boldsymbol{U} - \boldsymbol{M}^{-1}\boldsymbol{N})\boldsymbol{x} = \boldsymbol{b}$$

$$(\boldsymbol{U} - \boldsymbol{P})\boldsymbol{x} = \boldsymbol{f} \qquad \boldsymbol{P} = \boldsymbol{M}^{-1}\boldsymbol{N} \qquad \boldsymbol{f} = \boldsymbol{M}^{-1}\boldsymbol{b}$$

General Splittings

$\boldsymbol{Ax} = \boldsymbol{b}$ Split $\mathbf{A}$ such that: $\boldsymbol{A} = \boldsymbol{M} - \boldsymbol{N}$

$(\boldsymbol{U} - \boldsymbol{P})\boldsymbol{x} = \boldsymbol{f}$ $\boldsymbol{P} = \boldsymbol{M}^{-1}\boldsymbol{N}$ $\boldsymbol{f} = \boldsymbol{M}^{-1}\boldsymbol{b}$

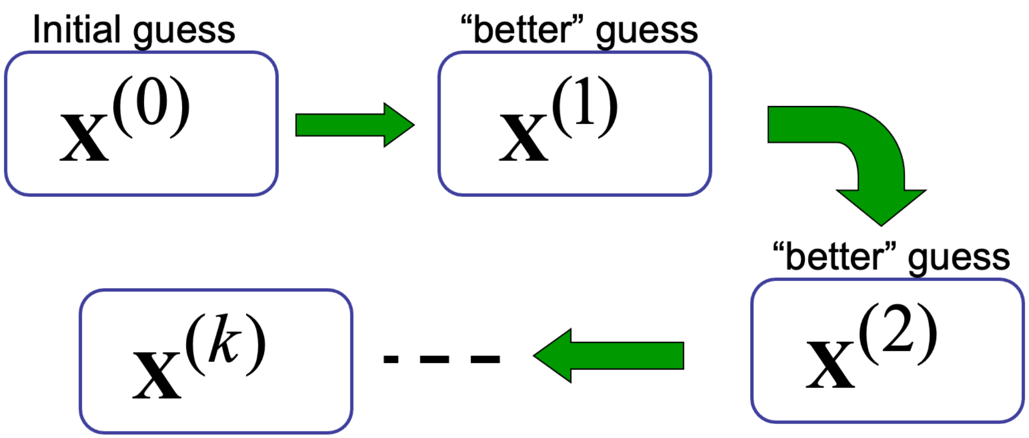
$$\boldsymbol{x}^{(n+1)} \leftarrow \boldsymbol{f} + \boldsymbol{P}\boldsymbol{x}^{(n)}$$

$$\boldsymbol{x}^{(n+1)} \leftarrow \boldsymbol{M}^{-1}\boldsymbol{b} + \boldsymbol{M}^{-1}\boldsymbol{N}\boldsymbol{x}^{(n)}$$

$$\boldsymbol{M}\boldsymbol{x}^{(n+1)} = \boldsymbol{b} + \boldsymbol{N}\boldsymbol{x}^{(n)}$$

Iterative Methods

$$\mathbf{AX} = \mathbf{b}$$



Jacobi Iteration

$$\mathbf{AX} = \mathbf{b} \qquad \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}$$

<u>Jacobi iteration:</u>

$$\mathbf{M}_J \mathbf{X}^{(k+1)} = \mathbf{N}_J \mathbf{X}^{(k)} + \mathbf{b}$$

Jacobi Iteration

<u>Jacobi iteration:</u> $$\mathbf{M}_J\mathbf{X}^{(k+1)} = \mathbf{N}_J\mathbf{X}^{(k)} + \mathbf{b}$$

$$\mathbf{L} = \begin{bmatrix} 0 & 0 & 0 \\ a_{21} & 0 & 0 \\ a_{31} & a_{32} & 0 \end{bmatrix} \qquad \mathbf{U} = \begin{bmatrix} 0 & a_{12} & a_{13} \\ 0 & 0 & a_{23} \\ 0 & 0 & 0 \end{bmatrix}$$

$$\mathbf{D} = \begin{bmatrix} a_{11} & 0 & 0 \\ 0 & a_{22} & 0 \\ 0 & 0 & a_{33} \end{bmatrix} \qquad \begin{cases} \mathbf{M}_J = \mathbf{D} \\ \mathbf{N}_J = -(\mathbf{L} + \mathbf{U}) \end{cases}$$

Jacobi Iteration

Jacobi iteration: $$\mathbf{M}_J \mathbf{X}^{(k+1)} = \mathbf{N}_J \mathbf{X}^{(k)} + \mathbf{b}$$

What does it mean?

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}$$

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 = b_1 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 = b_2 \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 = b_3 \end{cases}$$

Jacobi Iteration

Jacobi iteration: $$\mathbf{M}_J\mathbf{X}^{(k+1)} = \mathbf{N}_J\mathbf{X}^{(k)} + \mathbf{b}$$

What does it mean?

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 = b_1 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 = b_2 \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 = b_3 \end{cases}$$

$$\begin{cases} x_1 = \left(b_1 - a_{12}x_2 - a_{13}x_3\right)/\,a_{11} \\ x_2 = \left(b_2 - a_{21}x_1 - a_{23}x_3\right)/\,a_{22} \\ x_3 = \left(b_3 - a_{31}x_1 - a_{32}x_2\right)/\,a_{33} \end{cases}$$

Jacobi Iteration

<u>Jacobi iteration:</u> $$\mathbf{M}_J \mathbf{X}^{(k+1)} = \mathbf{N}_J \mathbf{X}^{(k)} + \mathbf{b}$$

<u>What does it mean?</u>

$$\begin{cases} x_1 = \left(b_1 - a_{12}x_2 - a_{13}x_3\right)/\,a_{11} \\ x_2 = \left(b_2 - a_{21}x_1 - a_{23}x_3\right)/\,a_{22} \\ x_3 = \left(b_3 - a_{31}x_1 - a_{32}x_2\right)/\,a_{33} \end{cases}$$

$$\begin{cases} x_1^{(k+1)} = \left(b_1 - a_{12}x_2^{(k)} - a_{13}x_3^{(k)}\right)/\,a_{11} \\ x_2^{(k+1)} = \left(b_2 - a_{21}x_1^{(k)} - a_{23}x_3^{(k)}\right)/\,a_{22} \\ x_3^{(k+1)} = \left(b_3 - a_{31}x_1^{(k)} - a_{32}x_2^{(k)}\right)/\,a_{33} \end{cases}$$

Gauss-Seidel Iteration

Gauss-Seidel iteration: Use the most current estimate of x_i

$$\begin{cases} x_1 = \left(b_1 - a_{12}x_2 - a_{13}x_3\right)/\,a_{11} \\ x_2 = \left(b_2 - a_{21}x_1 - a_{23}x_3\right)/\,a_{22} \\ x_3 = \left(b_3 - a_{31}x_1 - a_{32}x_2\right)/\,a_{33} \end{cases}$$

$$\begin{cases} x_1^{(k+1)} = \left(b_1 - a_{12}x_2^{(k)} - a_{13}x_3^{(k)}\right)/\,a_{11} \\ x_2^{(k+1)} = \left(b_2 - a_{21}x_1^{(k+1)} - a_{23}x_3^{(k)}\right)/\,a_{22} \\ x_3^{(k+1)} = \left(b_3 - a_{31}x_1^{(k+1)} - a_{32}x_2^{(k+1)}\right)/\,a_{33} \end{cases}$$

Gauss Seidel Iteration

<u>Gauss-Seidel iteration</u>: Compact form

$$\mathbf{M}_G \mathbf{X}^{(k+1)} = \mathbf{N}_G \mathbf{X}^{(k)} + \mathbf{b}$$

$$\mathbf{M}_G = (\mathbf{D} + \mathbf{L})$$

$$\mathbf{N}_G = -\mathbf{U}$$

Successive Over Relaxation (SOR)

$$\begin{cases} x_1 = \left(b_1 - a_{12}x_2 - a_{13}x_3\right)/\,a_{11} \\ x_2 = \left(b_2 - a_{21}x_1 - a_{23}x_3\right)/\,a_{22} \\ x_3 = \left(b_3 - a_{31}x_1 - a_{32}x_2\right)/\,a_{33} \end{cases}$$

$$\begin{cases} x_1^{(k+1)} = \omega\left(b_1 - a_{12}x_2^{(k)} - a_{13}x_3^{(k)}\right)/\,a_{11} + (1-\omega)x_1^{(k)} \\ x_2^{(k+1)} = \omega\left(b_2 - a_{21}x_1^{(k+1)} - a_{23}x_3^{(k)}\right)/\,a_{22} + (1-\omega)x_2^{(k)} \\ x_3^{(k+1)} = \omega\left(b_3 - a_{31}x_1^{(k+1)} - a_{32}x_2^{(k+1)}\right)/\,a_{33} + (1-\omega)x_3^{(k)} \end{cases}$$

Successive Over Relaxation (SOR)

<u>Successive Over Relaxation (SOR)</u>: Compact form

$$\mathbf{M}_{\omega}\mathbf{X}^{(k+1)} = \mathbf{N}_{\omega}\mathbf{X}^{(k)} + \omega\mathbf{b}$$

$$\mathbf{M}_{\omega} = \left(\mathbf{D} + \omega\mathbf{L}\right)$$

$$\mathbf{N}_{\omega} = (1-\omega)\mathbf{D} - \omega\mathbf{U}$$

Symmetric Positive Definite Matrices

Matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is Symmetric Positive Definite iff:

$$\boldsymbol{A} = \boldsymbol{A}^T$$

$$\boldsymbol{v}^T\boldsymbol{A}\boldsymbol{v} > 0 \qquad \forall \boldsymbol{v} \in \mathbb{R}^n, \quad \boldsymbol{v} \neq \boldsymbol{0}$$

Normal Equations

$$\boldsymbol{Ax} = \boldsymbol{b} \qquad \boldsymbol{A} \in \mathbb{R}^{m\times n}$$

If the Matrix $\boldsymbol{A}^T\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is non-singular then it is Symmetric Positive Definite:

$$\boldsymbol{A}^T\boldsymbol{A} = \boldsymbol{A}^T\boldsymbol{A}$$

$$\boldsymbol{v}^T\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{v} = \|\boldsymbol{A}\boldsymbol{v}\|^2 > 0 \qquad \forall \boldsymbol{v} \in \mathbb{R}^n \qquad \boldsymbol{v} \neq \boldsymbol{0}$$

Note: $\boldsymbol{A}$ is non-singular therefore $\boldsymbol{A}\boldsymbol{v} \neq \boldsymbol{0}$, for all $\boldsymbol{v} \neq \boldsymbol{0}$

Inner Product of Two Vectors

- $\langle \boldsymbol{u}, \boldsymbol{v} \rangle = \overline{\langle \boldsymbol{v}, \boldsymbol{u} \rangle}$
- $\langle \boldsymbol{u} + \boldsymbol{v}, \boldsymbol{w} \rangle = \langle \boldsymbol{u}, \boldsymbol{w} \rangle + \langle \boldsymbol{v}, \boldsymbol{w} \rangle$
- $\langle c\boldsymbol{u}, \boldsymbol{v} \rangle = \bar{c}\langle \boldsymbol{u}, \boldsymbol{v} \rangle$
- $\langle \boldsymbol{u}, c\boldsymbol{v} \rangle = c\langle \boldsymbol{u}, \boldsymbol{v} \rangle$

c is a scalar $c \in \mathbb{C}$

- $\langle \boldsymbol{u}, \boldsymbol{u} \rangle = \boldsymbol{0} \Leftrightarrow \boldsymbol{u} = \boldsymbol{0}$
- $\langle \boldsymbol{u}, \boldsymbol{u} \rangle > \boldsymbol{0}$ if $\boldsymbol{u} \neq \boldsymbol{0}$

Euclidian Inner Product

$\langle \boldsymbol{u}, \boldsymbol{v} \rangle = \boldsymbol{u}^T \boldsymbol{v} \qquad \boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^{\boldsymbol{n}}$

Generalize for complex vectors $\langle \boldsymbol{u}, \boldsymbol{v} \rangle = \boldsymbol{u}^* \boldsymbol{v} \qquad \boldsymbol{u}, \boldsymbol{v} \in \mathbb{C}^{\boldsymbol{n}}$

$\boldsymbol{u}^*$ is the conjugate transpose of $\boldsymbol{u}$.

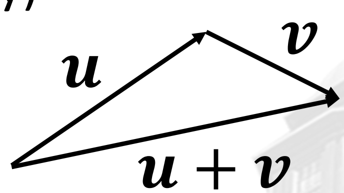
$$\boldsymbol{u} = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix} \qquad \boldsymbol{u}^* = [\bar{u}_1 \quad \bar{u}_2 \quad \cdots \quad \bar{u}_n]$$

$\bar{u}_n$ is the complex conjugate of u_n

Norm ("Size") of a Vector

Then Norm $||\boldsymbol{v}|| \in \mathbb{R}$ of a vector $\boldsymbol{v}$ is an indication of its size. It must be defined such that it obeys the following rules:

1. $||\boldsymbol{v}|| > 0$
2. $||a\boldsymbol{v}|| = |a|||\boldsymbol{v}||$
3. $||\boldsymbol{u} + \boldsymbol{v}|| \leq ||\boldsymbol{u}|| + ||\boldsymbol{v}||$ (Triangle Inequality)



Norm Based on Inner Product

For every inner product we can define a norm.

The Euclidean Norm is: $\|\boldsymbol{v}\|_2 = \sqrt{\langle \boldsymbol{v}, \boldsymbol{v} \rangle} = \sqrt{\boldsymbol{v}^T \boldsymbol{v}}$

Inner Product based on SPD Matrix $\boldsymbol{A}$

Matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ is Symmetric Positive.

Define the Inner Product: $\langle \boldsymbol{u}, \boldsymbol{v} \rangle_A \equiv \boldsymbol{u}^T \boldsymbol{A} \boldsymbol{v}$

We can verify that this definition satisfies all the properties of an inner product.

Note that the Euclidean Norm can be written as:

$$\langle \boldsymbol{u}, \boldsymbol{v} \rangle_U \equiv \boldsymbol{u}^T \boldsymbol{v}$$

Norm Based on Inner Product

For every inner product we can define a norm.

The Norm based on the inner product $\langle \, , \rangle_A$:

$$\|\boldsymbol{v}\|_A = \sqrt{\langle \boldsymbol{v}, \boldsymbol{v} \rangle_A} = \sqrt{\boldsymbol{v}^T \boldsymbol{A} \boldsymbol{v}}$$

Quadratic Function

Define the quadratic function $Q_A(\boldsymbol{u})$:

$$Q_A(\boldsymbol{u}) = \frac{1}{2}\langle \boldsymbol{u}, \boldsymbol{u} \rangle_A - \langle \boldsymbol{b}, \boldsymbol{u} \rangle_U$$

$$Q_A(\boldsymbol{u}) = \frac{1}{2}\boldsymbol{u}^T \boldsymbol{A}\boldsymbol{u} - \boldsymbol{b}^T \boldsymbol{u}$$

Minimize $Q_A(\boldsymbol{u})$

$$Q_A(\boldsymbol{u}) = \frac{1}{2}\boldsymbol{u}^T\boldsymbol{A}\boldsymbol{u} - \boldsymbol{b}^T\boldsymbol{u}$$

$$\frac{d}{d\boldsymbol{u}}Q_A(\boldsymbol{u}) = \boldsymbol{u}^T\boldsymbol{A} - \boldsymbol{b}^T = 0$$

$$\boldsymbol{A}\boldsymbol{u} = \boldsymbol{b}$$

Minimize $Q_A(\boldsymbol{u})$

$$Q_A(\boldsymbol{u}) = \frac{1}{2}\langle \boldsymbol{u}, \boldsymbol{u}\rangle_A - \langle \boldsymbol{b}, \boldsymbol{u}\rangle_U \qquad Q_A(\boldsymbol{v}) = \frac{1}{2}\boldsymbol{u}^T\boldsymbol{A}\boldsymbol{u} - \boldsymbol{b}^T\boldsymbol{u}$$

$$Q_A(\boldsymbol{u} + t\boldsymbol{v}) = \frac{1}{2}(\boldsymbol{u} + t\boldsymbol{v})^T\boldsymbol{A}(\boldsymbol{u} + t\boldsymbol{v}) - \boldsymbol{b}^T(\boldsymbol{u} + t\boldsymbol{v})$$

$$= \frac{1}{2}[\boldsymbol{u}^T\boldsymbol{A}\boldsymbol{u} + t\boldsymbol{u}^T\boldsymbol{A}\boldsymbol{v} + t\boldsymbol{v}^T\boldsymbol{A}\boldsymbol{u} + t^2\boldsymbol{v}^T\boldsymbol{A}\boldsymbol{v}] - \boldsymbol{b}^T(\boldsymbol{u} + t\boldsymbol{v})$$

$$= Q_A(\boldsymbol{u}) + t(\boldsymbol{u}^T\boldsymbol{A} - \boldsymbol{b}^T)\boldsymbol{v} + \frac{1}{2}t^2\langle \boldsymbol{v}, \boldsymbol{v}\rangle_A$$

Minimize $Q_A(\boldsymbol{u})$

$$Q_A(\boldsymbol{u}+t\boldsymbol{v}) = Q_A(\boldsymbol{u}) + t(\boldsymbol{u}^T\boldsymbol{A}-\boldsymbol{b}^T)\boldsymbol{v} + \frac{1}{2}t^2\langle \boldsymbol{v},\boldsymbol{v}\rangle_A$$

Note: $(\boldsymbol{u}^T\boldsymbol{A}-\boldsymbol{b}^T) = (\boldsymbol{A}\boldsymbol{u}-\boldsymbol{b})^T$

$$Q_A(\boldsymbol{u}+t\boldsymbol{v}) = Q_A(\boldsymbol{u}) + t\langle \boldsymbol{A}\boldsymbol{u}-\boldsymbol{b},\boldsymbol{v}\rangle_U + \frac{1}{2}t^2\langle \boldsymbol{v},\boldsymbol{v}\rangle_A$$

$$Q_A(\boldsymbol{u}+t\boldsymbol{v}) = \alpha + \beta t + \gamma t^2 \qquad \gamma > 0$$

Minimize $Q_A(\boldsymbol{u})$

$$Q_A(\boldsymbol{u} + t\boldsymbol{v}) = \alpha + \beta t + \gamma t^2 \qquad \gamma > 0$$

When $t = 0$: $\quad Q_A(\boldsymbol{u} + t\boldsymbol{v}) = Q_A(\boldsymbol{u})$

When does $Q_A(\boldsymbol{u} + t\boldsymbol{v})$ have a minimum at $t = 0$?

When does $\alpha + \beta t + \gamma t^2$ have a minimum at $t = 0$?

Minimize $Q_A(\boldsymbol{u})$

When does $Q_A(\boldsymbol{u} + t\boldsymbol{v}) = \alpha + \beta t + \gamma t^2$ have a minimum at $t = 0$?

$$\frac{\partial}{\partial t} Q_A(\boldsymbol{u} + t\boldsymbol{v}) = \beta + 2\gamma t$$

Minimum occurs at $t = \dfrac{\beta}{2\gamma}$

Minimum occurs at $t = 0$ iff $\beta = 0$

Minimize $Q_A(\boldsymbol{u})$

$$Q_A(\boldsymbol{u}+t\boldsymbol{v}) = \alpha + \beta t + \gamma t^2 \qquad \gamma > 0$$

When $t = 0$: $\quad Q_A(\boldsymbol{u}+t\boldsymbol{v}) = Q_A(\boldsymbol{u})$

$Q_A(\boldsymbol{u}+t\boldsymbol{v})$ has a minimum at $t = 0$ when $\beta = 0$

$$Q_A(\boldsymbol{u}+t\boldsymbol{v}) = Q_A(\boldsymbol{u}) + t\langle \boldsymbol{Au}-\boldsymbol{b},\boldsymbol{v}\rangle_U + \frac{1}{2}t^2\langle \boldsymbol{v},\boldsymbol{v}\rangle_A$$

$$\langle \boldsymbol{Au}-\boldsymbol{b},\boldsymbol{v}\rangle_U = 0 \qquad \forall \boldsymbol{v} \in \mathbb{R}^n$$

Minimize $Q_A(\boldsymbol{u})$

$$\langle \boldsymbol{Au} - \boldsymbol{b}, \boldsymbol{v} \rangle_U = 0 \qquad \forall \boldsymbol{v} \in \mathbb{R}^n$$

$$\boldsymbol{Au} = \boldsymbol{b}$$

Solving $\boldsymbol{Au} = \boldsymbol{b}$ is equivalent to minimizing:

$$Q_A(\boldsymbol{u}) = \frac{1}{2} \langle \boldsymbol{u}, \boldsymbol{u} \rangle_A - \langle \boldsymbol{b}, \boldsymbol{u} \rangle_U$$

Minimize $Q_A(\boldsymbol{u})$: Line Search

$$Q_A(\boldsymbol{u}) = \frac{1}{2}\langle \boldsymbol{u}, \boldsymbol{u}\rangle_A - \langle \boldsymbol{b}, \boldsymbol{u}\rangle_U$$

Start with initial guess $\boldsymbol{u}_0$ then search in direction $\boldsymbol{s}$

That is minimize:

$$Q_A(\boldsymbol{u}_0 + t\boldsymbol{s}) = Q_A(\boldsymbol{u}_0) + t\langle \boldsymbol{A}\boldsymbol{u}_0 - \boldsymbol{b}, \boldsymbol{s}\rangle_U + \frac{1}{2}t^2\langle \boldsymbol{s}, \boldsymbol{s}\rangle_A$$

Minimize $Q_A(\boldsymbol{u})$: Line Search

$$Q_A(\boldsymbol{u}_0 + t\boldsymbol{s}) = Q_A(\boldsymbol{u}_0) + t\langle \boldsymbol{A}\boldsymbol{u}_0 - \boldsymbol{b}, \boldsymbol{s}\rangle_U + \frac{1}{2}t^2\langle \boldsymbol{s}, \boldsymbol{s}\rangle_A$$

$$\frac{\partial}{\partial t}Q_A(\boldsymbol{u}_0 + t\boldsymbol{s}) = \langle \boldsymbol{A}\boldsymbol{u}_0 - \boldsymbol{b}, \boldsymbol{s}\rangle_U + \langle \boldsymbol{s}, \boldsymbol{s}\rangle_A t$$

$$\frac{\partial}{\partial t}Q_A(\boldsymbol{u}_0 + t\boldsymbol{s}) = 0 \quad \Leftrightarrow \quad \boxed{t = -\frac{\langle \boldsymbol{A}\boldsymbol{u}_0 - \boldsymbol{b}, \boldsymbol{s}\rangle_U}{\langle \boldsymbol{s}, \boldsymbol{s}\rangle_A}}$$

Note: $\boldsymbol{r} = \boldsymbol{A}\boldsymbol{u}_0 - \boldsymbol{b}$ is the residual at the guess $\boldsymbol{u}_0$

Search Direction $\boldsymbol{s}$: Gradient Descent

$$Q_A(\boldsymbol{u}_0 + t\boldsymbol{s}) \cong Q_A(\boldsymbol{u}_0 + t\boldsymbol{s}) + t\left.\frac{dQ_A}{d\boldsymbol{u}}\right|_{\boldsymbol{u}_0} \boldsymbol{s}$$

Define the Gradient: $\nabla Q_A(\boldsymbol{u}_0) \equiv \left.\frac{dQ_A}{d\boldsymbol{u}}\right|_{\boldsymbol{u}_0}^{T}$

For a given small and positive t, the change in Q_A due to $t\boldsymbol{s}$ is maximum when $s = \nabla Q_A(\boldsymbol{u}_0)$

Search Direction $\boldsymbol{s}$: Gradient Descent

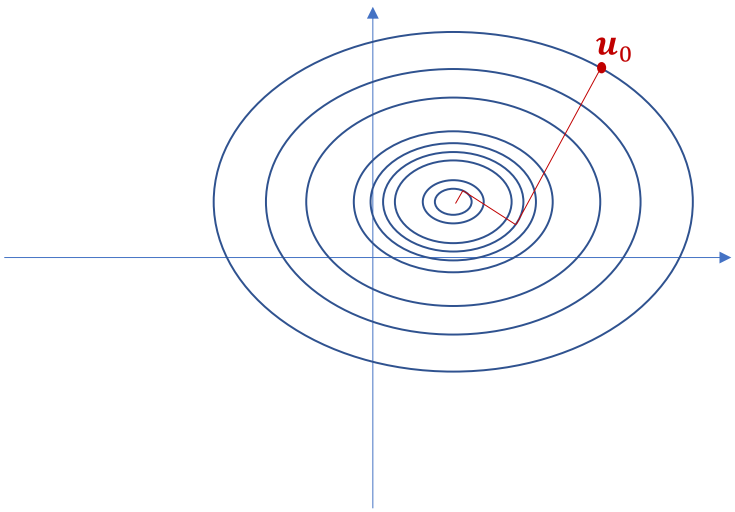
Use the Gradient as a search direction $\nabla Q_A(\boldsymbol{u}_0) \equiv \left.\frac{dQ_A}{d\boldsymbol{u}}\right|_{\boldsymbol{u}_0}^{T}$

$$Q_A(\boldsymbol{u}) = \frac{1}{2}\boldsymbol{u}^T\boldsymbol{A}\boldsymbol{u} - \boldsymbol{b}^T\boldsymbol{u}$$

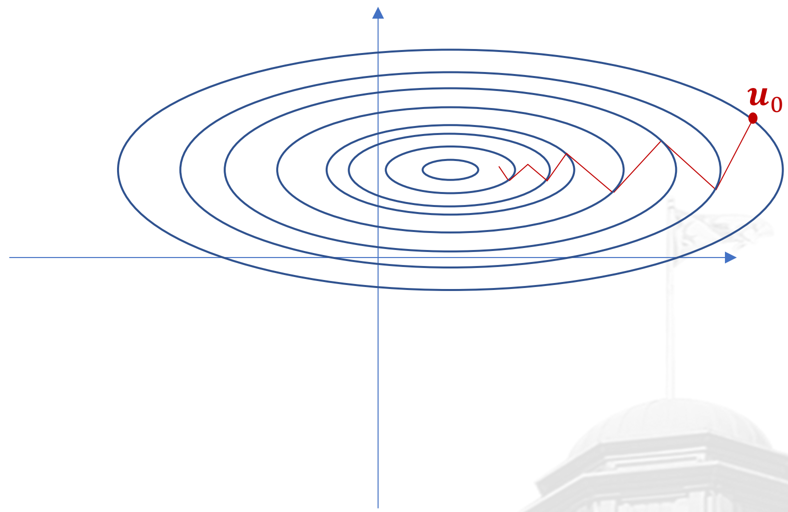
$$\frac{d}{d\boldsymbol{u}}Q_A(\boldsymbol{u}) = \boldsymbol{u}^T\boldsymbol{A} - \boldsymbol{b}^T$$

$$\nabla Q_A(\boldsymbol{u}_0) = \overbrace{\boldsymbol{A}\boldsymbol{u}_0 - \boldsymbol{b}}^{\text{Residual } \boldsymbol{r}_0}$$

Gradient Descent



Well conditioned Matrix



Ill-conditioned Matrix

Gradient Descent Method

- Choose Initial Guess $\boldsymbol{u}_0$
- First search direction: $\boldsymbol{s}_0 = \boldsymbol{r}_0 = \boldsymbol{A}\boldsymbol{u}_0 - \boldsymbol{b}$
- $t_0 = \frac{\langle \boldsymbol{s}_0, \boldsymbol{s}_0 \rangle_U}{\langle \boldsymbol{s}_0, \boldsymbol{s}_0 \rangle_A}$
- $\boldsymbol{u}_1 = \boldsymbol{u}_0 - t_0 \boldsymbol{s}_0$
- Second search direction: $\boldsymbol{s}_1 = \boldsymbol{r}_1 = \boldsymbol{A}\boldsymbol{u}_1 - \boldsymbol{b}$
- $t_1 = \frac{\langle \boldsymbol{s}_1, \boldsymbol{s}_1 \rangle_U}{\langle \boldsymbol{s}_1, \boldsymbol{s}_1 \rangle_A}$
- $\boldsymbol{u}_2 = \boldsymbol{u}_1 - t_1 \boldsymbol{s}_1$
- …

Relation Between Search Directions

- Current Guess $\boldsymbol{u}_k$
- Search direction: $\boldsymbol{s}_k = \boldsymbol{r}_k = \boldsymbol{A}\boldsymbol{u}_k - \boldsymbol{b}$
- $t_k = \frac{\langle \boldsymbol{s}_k, \boldsymbol{s}_k \rangle_U}{\langle \boldsymbol{s}_k, \boldsymbol{s}_k \rangle_A}$
- $\boldsymbol{u}_{k+1} = \boldsymbol{u}_k - t_k \boldsymbol{s}_k$
- Second search direction: $\boldsymbol{s}_{k+1} = \boldsymbol{r}_{k+1} = \boldsymbol{A}\boldsymbol{u}_{k+1} - \boldsymbol{b}$

$$\boldsymbol{s}_{k+1} = \boldsymbol{A}(\boldsymbol{u}_k - t_k \boldsymbol{s}_k) - \boldsymbol{b} = \boldsymbol{A}\boldsymbol{u}_k - \boldsymbol{b} - t_k \boldsymbol{A}\boldsymbol{s}_k$$

$$\boldsymbol{s}_{k+1} = \boldsymbol{s}_k - t_k \boldsymbol{A}\boldsymbol{s}_k$$

Relation Between Search Directions

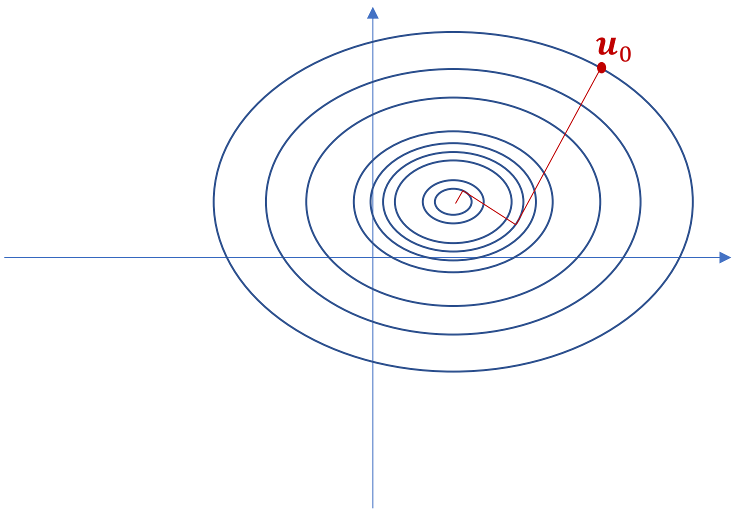
$$\boldsymbol{s}_{k+1} = \boldsymbol{s}_k - t_k \boldsymbol{A}\boldsymbol{s}_k$$

$$\boldsymbol{s}_k^T \boldsymbol{s}_{k+1} = \boldsymbol{s}_k^T \boldsymbol{s}_k - t_k \boldsymbol{s}_k^T \boldsymbol{A}\boldsymbol{s}_k$$

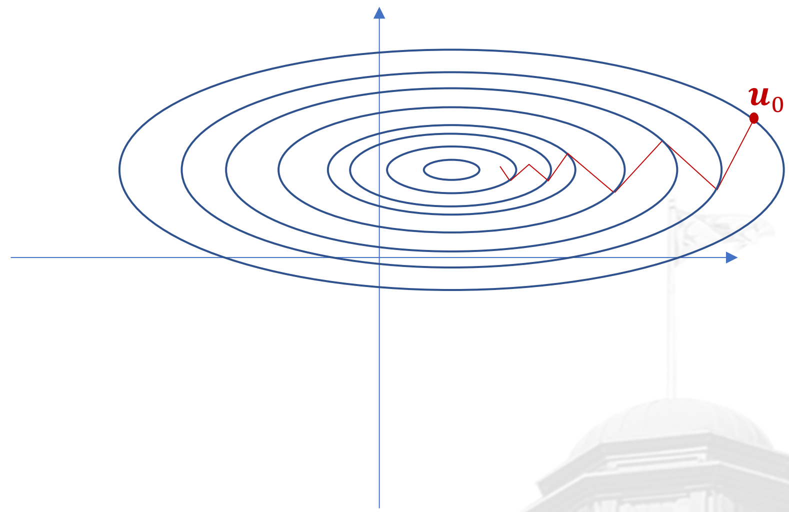
$$\boldsymbol{s}_k^T \boldsymbol{s}_{k+1} = \langle \boldsymbol{s}_k, \boldsymbol{s}_k \rangle_U - \frac{\langle \boldsymbol{s}_k, \boldsymbol{s}_k \rangle_U}{\langle \boldsymbol{s}_k, \boldsymbol{s}_k \rangle_A} \langle \boldsymbol{s}_k, \boldsymbol{s}_k \rangle_A = 0$$

$$\boxed{\boldsymbol{s}_{k+1} \perp \boldsymbol{s}_k}$$

Gradient Descent: Zigzagging



Well conditioned Matrix



Ill-conditioned Matrix

Conjugate Gradients

- Instead of choosing the direction of steepest descent, choose new directions that are orthogonal to all previous search directions.
- Avoids zigzagging.
- Converges in a maximum of n iterations.
- Proof in Solomon section 11.2