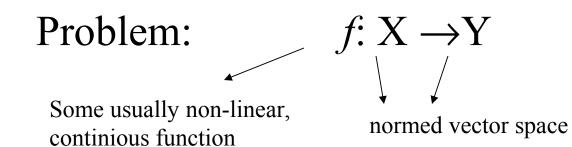
#### Lecture 12 Conditioning and Condition Numbers

NLA Reading Group Spring '13 by Can Kavaklıoğlu

## Outline

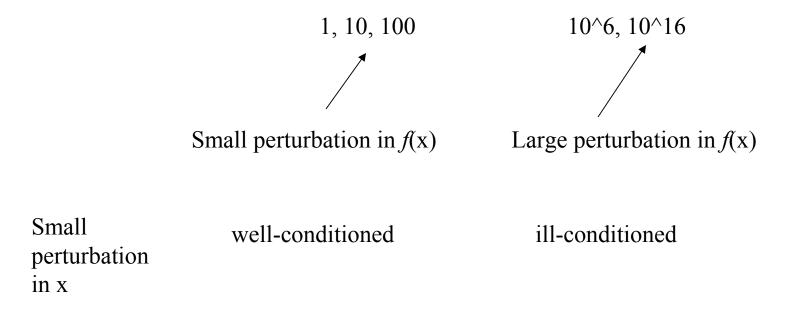
- Condition of a problem
- Absolute condition number
- Relative condition number
- Examples
- Condition of matrix-vector multiplication
- Condition number of a matrix
- Condition of system of equations

### Notation



Problem instance: combination of  $x \in X$  and f

## Problem Condition Types



#### Absolute Condition Number

Small perturbation in x 
$$\rightarrow \delta x$$

$$\delta f = f(x + \delta x) - f(x).$$

$$\hat{\kappa} = \lim_{\delta \to 0} \sup_{\|\delta x\| \le \delta} \frac{\|\delta f\|}{\|\delta x\|}$$

Assuming  $\delta x$  and  $\delta f$  are infinitesimal

$$\hat{\kappa} = \sup_{\delta x} \frac{\|\delta f\|}{\|\delta x\|}$$

### Absolute Condition Number

$$\hat{\kappa} = \sup_{\delta x} \frac{\|\delta f\|}{\|\delta x\|}$$

If f is differentiable, we can evaluate Jacobian of f at x

$$\delta f pprox J(x) \, \delta x_{
m c}$$
 with equality at limit  $\| \delta x \| o 0$ 

$$\hat{\kappa} = \|J(x)\|.$$

||J(x)|| represents norm of J(x) induced by norms of X and Y

### **Relative Condition Number**

$$\kappa = \kappa(x)$$

$$\kappa = \lim_{\delta \to 0} \sup_{\|\delta x\| \le \delta} \left( \frac{\|\delta f\|}{\|f(x)\|} / \frac{\|\delta x\|}{\|x\|} \right)$$

assuming  $\delta x$  and  $\delta f$  are infinitesimal,

$$\kappa = \sup_{\delta x} \left( \frac{\|\delta f\|}{\|f(x)\|} / \frac{\|\delta x\|}{\|x\|} \right)$$
  
if f is differentiable, 
$$\kappa = \frac{\|J(x)\|}{\|f(x)\| / \|x\|}$$

# Examples

#### Condition of Matrix-Vector Multiplication

$$\kappa = \sup_{\delta x} \left( \frac{\|\delta f\|}{\|f(x)\|} \left/ \frac{\|\delta x\|}{\|x\|} \right) \right)$$

Problem: compute Ax from input x with fixed  $A \in \mathbb{C}^{m \times n}$ 

$$\kappa = \sup_{\delta x} \left( \frac{\|A(x + \delta x) - Ax\|}{\|Ax\|} / \frac{\|\delta x\|}{\|x\|} \right)$$
$$= \sup_{\delta x} \frac{\|A\delta x\|}{\|\delta x\|} / \frac{\|Ax\|}{\|x\|}$$
$$\kappa = \|A\| \frac{\|x\|}{\|Ax\|}$$

#### Condition of Matrix-Vector Multiplication

$$\kappa = \sup_{\delta x} \left( \frac{\|\delta f\|}{\|f(x)\|} / \frac{\|\delta x\|}{\|x\|} \right)$$
$$\kappa = \|A\| \frac{\|x\|}{\|Ax\|}$$

If A is square and non-singular using  $||x||/||Ax|| \le ||A^{-1}||$ 

Loosen relative condition number to a bound independent of x

$$\kappa \le \|A\| \|A^{-1}\|$$

$$\kappa = \alpha \|A\| \|A^{-1}\| \quad \alpha = \frac{\|x\|}{\|Ax\|} / \|A^{-1}\|$$

If A is not square use pseudoinverse A<sup>+</sup>

#### Condition of Matrix-Vector Multiplication

**Theorem 12.1.** Let  $A \in \mathbb{C}^{m \times m}$  be nonsingular and consider the equation Ax = b. The problem of computing b, given x, has condition number

$$\kappa = \|A\| \frac{\|x\|}{\|b\|} \le \|A\| \|A^{-1}\|$$
(12.13)

with respect to perturbations of x. The problem of computing x, given b, has condition number

$$\kappa = \|A^{-1}\| \frac{\|b\|}{\|x\|} \le \|A\| \|A^{-1}\|$$
(12.14)

with respect to perturbations of b. If  $\|\cdot\| = \|\cdot\|_2$ , then equality holds in (12.13) if x is a multiple of a right singular vector of A corresponding to the minimal singular value  $\sigma_m$ , and equality holds in (12.14) if b is a multiple of a left singular vector of A corresponding to the maximal singular value  $\sigma_1$ .

### Condition Number of a Matrix

Condition number of A relative to norm  $\|\bullet\| \quad \kappa(A) = \|A\| \|A^{-1}\|$ If A is singular  $\kappa(A) = \infty$ .

if  $\|\cdot\| = \|\cdot\|_2$ , then  $\|A\| = \sigma_1$  and  $\|A^{-1}\| = 1/\sigma_m$ . Thus  $\kappa(A) = \frac{\sigma_1}{\sigma_m}$  in the 2-norm

 $A \in \mathbb{C}^{m \times n}$  of full rank,  $m \ge n$ 

$$\kappa(A) = \|A\| \|A^+\|$$
  

$$\kappa(A) = \frac{\sigma_1}{\sigma_n} \text{ in the 2-norm}$$

Condition of a System of Equations

Fix b and perturb A, in problem:  $A \mapsto x = A^{-1}b$ 

$$(A + \delta A)(x + \delta x) = b$$

$$\frac{\|\delta x\|}{\|x\|} \left/ \frac{\|\delta A\|}{\|A\|} \le \|A^{-1}\| \|A\| = \kappa(A).$$

Equality in this bound will hold whenever  $\delta A$  is such that

$$||A^{-1}(\delta A)x|| = ||A^{-1}|| ||\delta A|| ||x||,$$

### Condition of a System of Equations

**Theorem 12.2.** Let b be fixed and consider the problem of computing  $x = A^{-1}b$ , where A is square and nonsingular. The condition number of this problem with respect to perturbations in A is

$$\kappa = \|A\| \|A^{-1}\| = \kappa(A).$$
(12.18)

#### Lecture 13 Floating Point Arithmetic

NLA Reading Group Spring '13 by Can Kavaklıoğlu

### Outline

- Limitations of Digital Representations
- Floating Point Number
- Machine Epsilon
- Floating Point Arithmetic
- Complex Floating Point Arithmetic

## Limitations of Digital Representations

Finite number of bits — Finite subset of real/complex numbers Two limitations

- Precision: IEEE double between 1.79 x 10<sup>3</sup>08 and 2.23 x 10<sup>-308</sup>
- Overflow / underflow
- Interval representation: IEEE interval [1 2]:

1, 
$$1 + 2^{-52}$$
,  $1 + 2 \times 2^{-52}$ ,  $1 + 3 \times 2^{-52}$ , ..., 2

interval [2 4]:

2, 
$$2+2^{-51}$$
,  $2+2\times2^{-51}$ ,  $2+3\times2^{-51}$ , ..., 4

gap size:

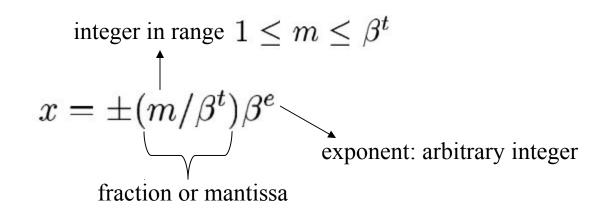
$$2^{-52} \approx 2.22 \times 10^{-16}$$

## Floating Point Number

F: subset of real numbers, including 0

 $\beta$ : base/radix

t: precision (23 single, 53 double precision - IEEE)



Idelized system: ignores underflow and overflow. F is a countably infinite set and it is self similar:  $F = \beta F$ 

Machine Epsilon

Resolution of F:  $\epsilon_{\text{machine}} = \frac{1}{2}\beta^{1-t}$ IEEE single IEEE double  $2^{-24} \approx 5.96 \times 10^{-8}$   $2^{-53} \approx 1.11 \times 10^{-16}$ 

For all  $x \in \mathbb{R}$ , there exists  $x' \in \mathbf{F}$  such that  $|x - x'| \leq \epsilon_{\text{machine}} |x|$ 

Rounding:

For all  $x \in \mathbb{R}$ , there exists  $\epsilon$  with  $|\epsilon| \leq \epsilon_{\text{machine}}$ such that  $fl(x) = x(1 + \epsilon)$ .

### Floating Point Arithmetic

$$x \circledast y = \mathrm{fl}(x \ast y)$$

### **Fundamental Axiom of Floating Point Arithmetic**

For all  $x, y \in \mathbf{F}$ , there exists  $\epsilon$  with  $|\epsilon| \leq \epsilon_{\text{machine}}$  such that  $x \circledast y = (x \ast y)(1 + \epsilon).$ 

Every operation of floating point arithmetic is exact up to a relative error of size at most machine epsilon

## Different Machine Epsilon and Complex Floating Point Arithmetic

- Some (very old) hardware may not support IEEE machine epsilon
- It may be possible to use a larger machine epsilon value
- Complex arithmetic is performed using two floating point numbers
- Machine epsilon needs to be adjusted

### The end

# thanks