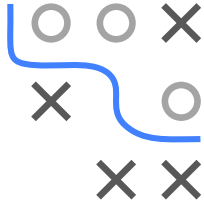
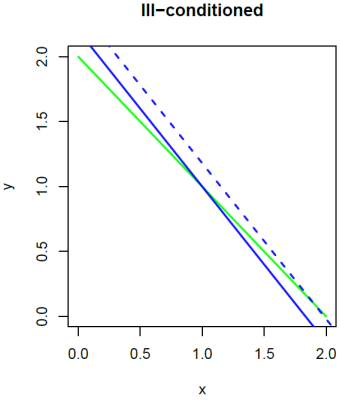
Algorithms and Data Structures



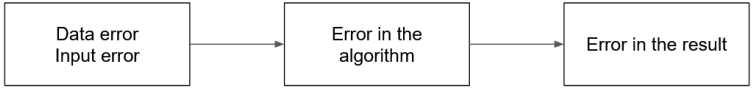
Numerics
Numerical Error & Conditioning

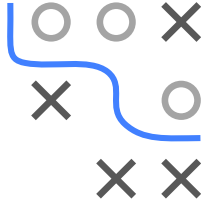


Learning goals

- Error caused by condition of the problem
- Error caused by stability of algorithms

NUMERICAL ERROR

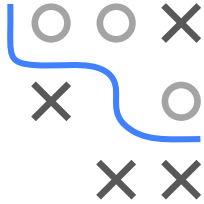




Errors in the result are caused by

- Data error or input error; often unavoidable, are part of the problem
  $\rightarrow$ **Condition of the problem**
- Error in the algorithm that can often be fixed by modification
  $\rightarrow$ **Stability of algorithms**

NUMERICAL ERROR / 2



Given: Approximate value $\tilde{x}$ for exact value x.

- Absolute error (this is signed!):

$$\Delta x = \tilde{x} - x$$

  If $|\Delta x| \leq \epsilon$, then ϵ is known as absolute error bound.

- Relative error:

$$\delta_x = \frac{|\tilde{x} - x|}{|x|} = \frac{|\Delta x|}{|x|}$$

  If $\delta_x \leq \rho$, then ρ is known as relative error bound.

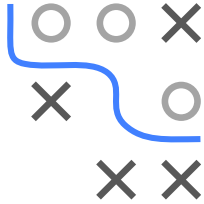
Comment:
Relative error is dimensionless, value in denominator may obviously not be too close to zero ($\rightarrow$ use absolute error instead).

NUMERICAL ERROR / 3

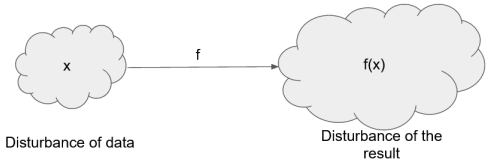
The following holds:

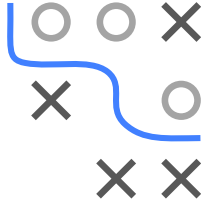
$$\tilde{x} \quad = \quad x\left(1 + \frac{\Delta x}{x}\right)$$

If only the first m digits of a number are determined in decimal representation, a possible relative error of 10^{-m} results (and vice versa).



CONDITION OF A PROBLEM





How sensitive is the result to a small disturbance of the data?

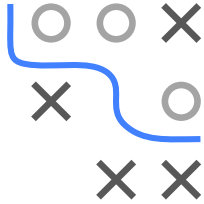
Well-conditioned: Result is a little sensitive

Ill-conditioned: Result is very sensitive

This is particularly important in statistics, since data is usually error-prone.

CONDITION OF A PROBLEM / 2

Example 1: System of equations



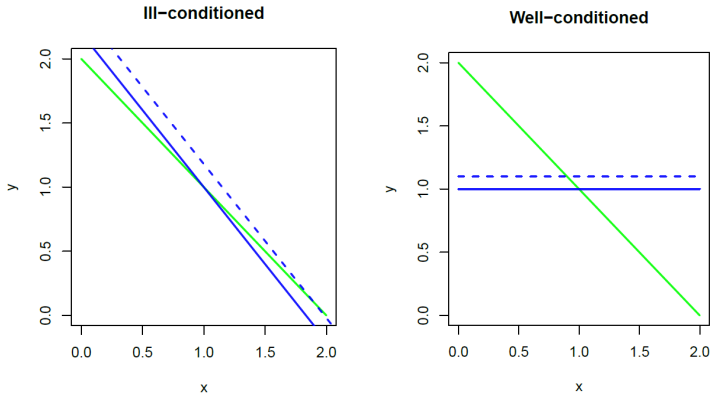
$$\begin{aligned} 0.835x + 0.667y &= 0.168 \\ 0.333x + 0.266y &= 0.067 \end{aligned}$$

The solution is $x = 1$ and $y = -1$.

By a small change of the right side, for example $0.067 \rightarrow 0.066$, the solution changes to $x = -666$ and $y = 834$.

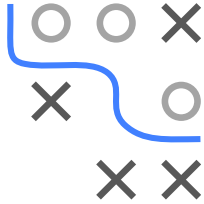
CONDITION OF A PROBLEM / 3

Example 2: Intersection of two straight lines



In the left case (ill-conditioned problem) a small disturbance has a big influence on the solution (position of the intersection). In the right case (well-conditioned) the disturbance has only a minor effect on the solution.

CONDITION (DEFINITION)



Consider a problem of form $y = f(x)$. Condition describes to what extent the solution changes with minor data disturbance ($\tilde{x} = x + \delta x$).

Condition number (at position x): the smallest $\kappa \geq 0$, so that

$$\frac{|f(x + \Delta x) - y|}{|y|} \leq \kappa \frac{|\Delta x|}{|x|} \text{ for } \Delta x \to 0.$$

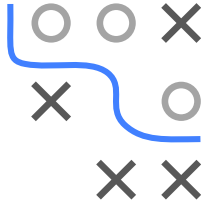
The condition therefore compares the relative error of input and output.
A problem is

- Well-conditioned if $\kappa \not\gg 1$
- Ill-conditioned if $\kappa \gg 1$
- Ill-posed when $\kappa = \infty$

Note that for machine numbers $\Delta x < \epsilon_m$ (approximation error when importing data)!

CONDITION FOR VECTORS

For vectors, errors are defined by means of a suitable norm.



Absolute error:

$$\Delta \mathbf{x} = \tilde{\mathbf{x}} - \mathbf{x}$$

Relative error:

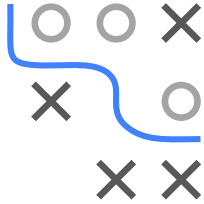
$$\delta = \frac{\|\tilde{\mathbf{x}} - \mathbf{x}\|}{\|\mathbf{x}\|}$$

Similarly, for $f : \mathbb{R}^n \to \mathbb{R}^d$ the **norm-wise** condition at the location $\mathbf{x}$ is defined by using the condition number.

Condition number: The smallest $\kappa \geq 0$, such that

$$\frac{\|f(\mathbf{x} + \Delta \mathbf{x}) - \mathbf{y}\|}{\|\mathbf{y}\|} \leq \kappa \frac{\|\Delta \mathbf{x}\|}{\|\mathbf{x}\|} \text{ for } \|\Delta \mathbf{x}\| \to 0.$$

CONDITION FOR VECTORS / 2



Let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable in point x, the condition can be specified using the derivative:

$$\kappa = \frac{\|\mathbf{x}\|}{\|f(\mathbf{x})\|}\|\nabla f(\mathbf{x})\|.$$

Proof sketch:
The definition of the condition can be written as

$$\kappa \quad = \quad \limsup_{\|\Delta\mathbf{x}\|\to 0} \frac{\|f(\mathbf{x}+\Delta\mathbf{x})-\mathbf{y}\|}{\|\mathbf{y}\|}\frac{\|\mathbf{x}\|}{\|\Delta\mathbf{x}\|} = \limsup_{\|\Delta\mathbf{x}\|\to 0} \frac{\|f(\mathbf{x}+\Delta\mathbf{x})-\mathbf{y}\|}{\|\Delta\mathbf{x}\|}\frac{\|\mathbf{x}\|}{\|f(\mathbf{x})\|}$$

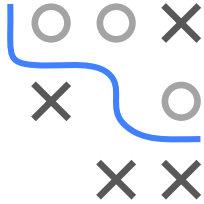
According to the mean value theorem the following applies: there is a $\boldsymbol{v} \in (\mathbf{0}, \Delta\mathbf{x})$ such that

$$\nabla f(\mathbf{x}+\boldsymbol{v}) = \frac{\|f(\mathbf{x}+\Delta\mathbf{x})-\mathbf{y}\|}{\|\Delta\mathbf{x}\|}.$$

CONDITION FOR VECTORS / 3

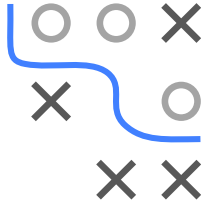
From $\|\Delta \mathbf{x}\| \to 0$ follows $\|\boldsymbol{v}\| \to 0$ and thus

$$\kappa \quad = \quad \limsup_{\|\Delta \mathbf{x}\| \to 0} \|\nabla f(\mathbf{x} + \boldsymbol{v})\| \frac{\|\mathbf{x}\|}{\|f(\mathbf{x})\|} = \|\nabla f(\mathbf{x})\| \frac{\|\mathbf{x}\|}{\|f(\mathbf{x})\|}$$



- $\nabla f(\mathbf{x})$ corresponds to the $1 \times n$ Jacobi matrix here. Thus, $\|\nabla f(\mathbf{x})\|$ is an induced *matrix norm*.
- The condition depends on the choice of the vector norm / induced matrix norm.

ARITHMETIC OPERATIONS



- **Multiplication:**
  We look at $f(x_1, x_2) = x_1 \cdot x_2$ and calculate κ using

$$\kappa = \frac{\|\nabla f(\mathbf{x})\|_1 \cdot \|\mathbf{x}\|_1}{|f(\mathbf{x})|}$$

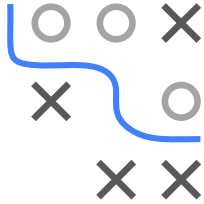
  Since

$$\begin{aligned} \|\nabla f(\mathbf{x})\|_1 &= \|(x_2, x_1)\|_1 \overset{*}{=} \max(|x_1|, |x_2|) \\ \|\mathbf{x}\|_1 &= \|(x_1, x_2)\|_1 = |x_1| + |x_2| \\ |f(\mathbf{x})| &= |x_1 x_2| \end{aligned}$$

  we get $\kappa = 1 + \frac{\max(|x_1|,|x_2|)}{\min(|x_1|,|x_2|)}$. So κ becomes very big for $x_2 \ll x_1$.

  * Attention: induced matrix norm (maximum absolute column sum norm).

ARITHMETIC OPERATIONS / 2



- **Addition / Subtraction:**
  We will consider $f(x_1, x_2) = x_1 \pm x_2$ and calculate κ using

$$\kappa = \frac{\|\nabla f(\mathbf{x})\|_1 \cdot \|\mathbf{x}\|_1}{|f(\mathbf{x})|}$$

  Since

$$\begin{aligned} \|\nabla f(\mathbf{x})\|_1 &= \|(1, \pm 1)\| \stackrel{*}{=} 1 \\ \|\mathbf{x}\|_1 &= \|(x_1, x_2)\|_1 = |x_1| + |x_2| \\ |f(\mathbf{x})| &= |x_1 \pm x_2| \end{aligned}$$

  we get $\kappa = \frac{|x_1|+|x_2|}{|x_1 \pm x_2|}$.

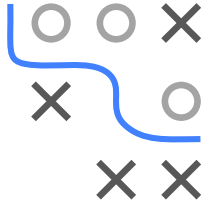
* Attention: induced matrix norm (maximum absolute column sum norm).

ARITHMETIC OPERATIONS / 3

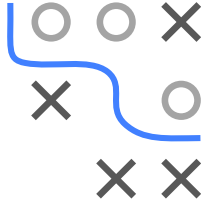
The addition of two positive numbers is well-conditioned with $\kappa = \frac{|x_1|+|x_2|}{|x_1+x_2|} = 1$.

In the case of $x_1 \approx x_2$, the subtraction of two positive numbers is very ill-conditioned ($\kappa = \frac{|x_1|+|x_2|}{|x_1+(-x_2)|}$ becomes arbitrarily large).

This problem is known as **loss of significance**.



LOSS OF SIGNIFICANCE (EXAMPLE)



```
x = 1.234567890123456
y = 1.234567890000000
```

Theoretically we should get a result of `1.23456e-10` for the difference $x - y$. But the computer returns

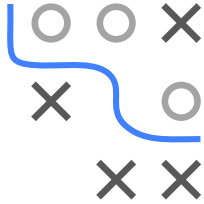
```
x - y

## [1] 1.234561342045026e-10
```

Why is this the case?

LOSS OF SIGNIFICANCE (EXAMPLE) / 2

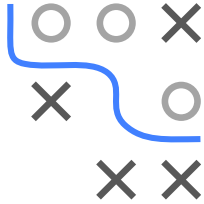
The numbers x, y cannot be represented exactly in floating point notation, but only as a rounded number with 16 significant digits. This inaccurate representation results in "uncertainties" from the 17th decimal place on.



```
sprintf("%.20f", x)
## [1] "1.23456789012345602430"
```

```
sprintf("%.20f", y)
## [1] "1.23456788999999989009"
```

LOSS OF SIGNIFICANCE (EXAMPLE) / 3



By subtracting the two numbers, the first 10 significant digits cancel each other out. The uncertainties from the 17th place now shift to the 7th significant place.

Due to the loss of significance we have lost $10 = 16 - 6$ digits (number of significant digits - number of deleted digits) in accuracy.

$$
\begin{aligned}
& 1.2345678901234\,56??? \ldots \\
-\,& 1.2345678900000\,000??? \ldots \\
\hline
& 0.0000000001234\,56??? \ldots \\
&= 1.2345\,6??? \ldots \times 10^{-10}
\end{aligned}
$$