Notes week 1.
Math 270B: Applied Numerical Linear Algebra

1 Relative condition number of \( f : \mathbb{R}^n \rightarrow \mathbb{R}^m \)

The relative condition number of \( f \) at \( x \in \mathbb{R}^n \) \( \kappa(x) \) is defined as

\[
\kappa(x) = \lim_{\delta \to 0} \kappa_\delta(x), \quad \kappa_\delta(x) = \sup_{|\delta x| \leq \delta} \left| \frac{\delta f}{\delta x} \right| / \frac{|f(x)|}{|x|}
\]

It is also useful to define

\[
c_\delta(x) = \kappa_\delta(x) - \kappa(x) \geq 0
\]

and then of course

\[
\lim_{\delta \to 0} c_\delta(x) = 0.
\]

2 Backward stability of \( \tilde{f} : \mathbb{R}^n \rightarrow \mathbb{R}^m, \tilde{f} \approx f \)

An algorithm \( \tilde{f} : \mathbb{R}^n \rightarrow \mathbb{R}^m \) is said to be backward stable at \( x \in \mathbb{R} \) if \( \exists \tilde{x} \) such that

\[
\frac{|\tilde{x} - x|}{|x|} \leq C(x)\epsilon_{machine}, \quad \text{and} \quad \tilde{f}(x) = f(\tilde{x}).
\]

3 Relative error bound theorem

If \( f \) has relative condition number \( \kappa(x) \) at \( x \) and \( \tilde{f} \) is backward stable at \( x \) with constant \( C(x) \), then

\[
\frac{|\tilde{f}(x) - f(x)|}{|f(x)|} \leq C(x)(\kappa(x) + c_\delta(x))\epsilon_{machine}
\]

4 IEEE floating point

The IEEE standard parameterizes the real line in terms of 32 bit (single precision) or 64 bit (double precision) representations. Specifically, the bits are organized as

<table>
<thead>
<tr>
<th>s (1 bit)</th>
<th>e (( e_{\max} ) bits)</th>
<th>( f ) (( f_{\max} ) bits)</th>
</tr>
</thead>
</table>

where

\[
e = \sum_{i=0}^{e_{\max}-1} e_i 2^i, \quad f = \sum_{j=1}^{f_{\max}} f_j 2^{-j}
\]

with bits \( e_i = 0 \) or 1, \( f_j = 0 \) or 1 and \( s = 0 \) or 1. The parameterization creates the subset \( \mathcal{F} \subset \mathbb{R} \), \( \mathcal{F} = \{ x \in \mathbb{R} | x = -1^s 2^{-\text{bias}} (1 + f) \} \). Our best approximation to an arbitrary real number is to represent it with the closest element in \( \mathcal{F} \). In other words, we can only approximate the function \( f : \mathbb{R} \rightarrow \mathbb{R} \) as \( \tilde{f} : \mathbb{R} \rightarrow \mathbb{R} \) in the sense that given \( x \in \mathbb{R} \) \( \tilde{f} \) returns the closest \( y \in \mathcal{F} \). We will usually refer to this as

\[
\tilde{f}(x) = \text{fl}(x) = y \in \mathcal{F} \text{ such that } |x - y| \text{ is minimal}
\]
5 Machine precision

We will define the number

$$\epsilon_{\text{machine}} = 2^{-f_{\text{max}}} - 1.$$

It represents the relative error in approximating an arbitrary real number with a floating point number. Specifically, it can be shown that

$$\frac{|\tilde{f}(x) - x|}{|x|} \leq \epsilon_{\text{machine}}.$$