Backwards stability of recursive summation

Steven G. Johnson, MIT course 18.335 notes, Fall 2012

September 19, 2013

Consider the summation function $f(x) = \sum_{i=1}^{n} x_i$ for vectors $x \in \mathbb{F}^n$ (n floating-point numbers). (The case of real inputs rounded to $\mathbb{F}$ just increases the error coefficient slightly, and is handled below.) We wish to demonstrate the backwards stability of the function $\tilde{f}(x)$ which computes the summation "in order" in floating-point arithmetic ($\oplus$), that is:

$$\tilde{f}(x) = ((x_1 \oplus x_2) \oplus x_3) \oplus \cdots),$$

which can be defined via the recursion:

$$\tilde{s}_1 = x_1,$$

$$\tilde{s}_i = \tilde{s}_{i-1} \oplus x_i,$$

$$f(\tilde{x}) = \tilde{s}_n,$$

and this arrangement is sometimes called "recursive summation" (independent of whether it is implemented via recursion or a loop in a computer language; the key is the *order* of the operations).

To be backwards stable, we must find a vector $\tilde{x}$ such that $\tilde{f}(x) = f(\tilde{x})$, and also $\tilde{x}$ is "close" to x in the sense that $\|\tilde{x} - x\| = \|\tilde{x}\| O(\epsilon_{\text{mach}})$ in some norm $\|\cdot\|$. We do this in two steps. First, we construct $\tilde{x}$ such that $\tilde{f}(x) = f(\tilde{x})$, and then we show that it is close to x.

To construct $\tilde{x}$ is easy. We define $\tilde{x}_1 = x_1$, and then define $\tilde{x}_i$ for $i > 1$ such that $\tilde{s}_i = \tilde{s}_{i-1} \oplus x_i = \tilde{s}_{i-1} + \tilde{x}_i$. It follows by induction that $\tilde{s}_i = \sum_{k=1}^{i} \tilde{x}_i$, and hence $\tilde{f}(x) = \tilde{s}_n = f(\tilde{x})$ as desired. That is:

$$\tilde{x}_i = \tilde{s}_{i-1} \oplus x_i - \tilde{s}_{i-1} = (\tilde{s}_{i-1} + x_i) \cdot (1 + \epsilon_i) - \tilde{s}_{i-1}$$

where $|\epsilon_i| \leq \epsilon_{\text{mach}}$, by definition of $\oplus$.

Now, we need to show that $\|\tilde{x} - x\|$ is "small" in the sense above. As we shall shortly see in 18.335, it turns out that we can choose any norm that we wish for proving stability (stability in one norm implies stability in *every* norm), and in this problem it is convenient to choose the L_1 norm $\|x\|_1 = \sum_{i=1}^{n} |x_i|$. First, consider $|\tilde{x}_i - x_i|$, using the formula above for $\tilde{x}_i$:

$$\begin{aligned} |\tilde{x}_i - x_i| &= |x_i + \tilde{s}_{i-1}| \cdot |\epsilon_i| = \left| x_i + \sum_{k=1}^{i-1} \tilde{x}_k \right| \cdot |\epsilon_i| \\ &\leq (|x_i| + \|\tilde{x}\|_1) \cdot |\epsilon_i|, \end{aligned}$$

where we have used the fact that $\tilde{s}_{i-1}$, by construction, is equal to the *exact* sum of the $\tilde{x}_k$ for $k < i$, which in turn is $\leq$ the L_1 norm of $\tilde{x}$. It now follows that

$$\begin{aligned}\|\tilde{x} - x\|_1 \leq \sum_{i=1}^{n} (|x_i| + \|\tilde{x}\|_1) \cdot |\epsilon_i| &\leq \left[\left(\sum_{i=1}^{n} |x_i| \right) + n\|\tilde{x}\|_1 \right] \max_i |\epsilon_i| \\ &= \|x\|_1 O(\epsilon_{\text{mach}}) + \|\tilde{x}\|_1 O(\epsilon_{\text{mach}}),\end{aligned}$$

where the second $O(\epsilon_{\text{mach}})$ has a factor of n in its coefficient. (This doesn't matter: we only require that the constants hidden inside the O be independent of x, not of n.) But we can easily convert $\|x\|$ to $\|\tilde{x}\|$ (or vice versa) since x and $\tilde{x}$ are close. In particular, by the triangle inequality, $\|x\| = \|\tilde{x} + (x - \tilde{x})\| \leq \|\tilde{x}\| + \|\tilde{x} - x\|$ for any norm, and substituting this into the equation above and solving for $\|\tilde{x} - x\|_1$ we find:

$$\|\tilde{x} - x\|_1 = \frac{\|\tilde{x}\|_1 O(\epsilon_{\text{mach}})}{1 - O(\epsilon_{\text{mach}})} = \|\tilde{x}\|_1 O(\epsilon_{\text{mach}}),$$

since (as you show more explicitly in pset 2) we can Taylor expand $\frac{1}{1-O(\epsilon)} = 1 + O(\epsilon) + O(\epsilon^2)$ for small ϵ.

Regarding $\|x\|$ versus $\|\tilde{x}\|$ in the denominator

Note that this last point means that it doesn't matter whether we use $\|x\|$ or $\|\tilde{x}\|$ on the right-hand side (or in the denominator) for the definition of backwards stability (or stability), since by the same argument one can show:

$$\|\tilde{x} - x\| = \|\tilde{x}\| O(\epsilon_{\text{mach}}) \Longleftrightarrow \|\tilde{x} - x\| = \|x\| O(\epsilon_{\text{mach}})$$

in any norm.

Regarding inputs in $\mathbb{R}$ versus $\mathbb{F}$

In the beginning, we assumed that x was in $\mathbb{F}^n$, i.e. that the inputs are already floating point numbers. This was merely a convenience, and almost the same proof applies if x is in $\mathbb{R}^n$ and we first compute $\text{fl}(x)$ (rounding x to the nearest floating-point values) before summing. The reason is that, for any $x_i \in \mathbb{R}$ (neglecting the cases of overflow or underflow as usual), $\text{fl}(x_i) = x_i(1 + \epsilon_i')$ for $|\epsilon_i'| \leq \epsilon_{\text{mach}}$, and so it follows that

$$\tilde{s}_{i-1} \oplus \text{fl}(x_i) = [\tilde{s}_{i-1} + x_i(1 + \epsilon_i')] (1 + \epsilon_i) = (\tilde{s}_{i-1} + x_i)(1 + \epsilon_i) + x_i \epsilon_i' + O(\epsilon_{\text{mach}}{}^2),$$

where $|\epsilon_i| \leq \epsilon_{\text{mach}}$. However, if we plug the new $x_i \epsilon_i'$ term into the above proof, it just gives another $\|x\|_1 O(\epsilon_{\text{mach}})$ term in $\|\tilde{x} - x\|_1$, which doesn't change anything.