

Dot Product Error Analysis

Let $x, y \in \mathbb{R}^n$ have floating point entries. Here we will try to bound the rounding errors in the computation of $x^T y$. Since *how* this is computed determines the result, we will analyze the most common algorithm:

$$s_1 = \text{fl}(x_1 y_1); \quad \text{then, for } k = 1, 2, \dots, n-1, \quad s_{k+1} = \text{fl}(s_k + \text{fl}(x_k y_k)).$$

Then, by the FAFA, we have $s_1 = x_1 y_1 (1 + \delta_1)$, with $|\delta_1| \leq \mu$.

The subsequent iterates are each computed with 2 rounding errors: the multiply (FAFA: $(1 + \delta_2)$) and then the add (FAFA: $(1 + \epsilon_2)$).

$$s_2 = (s_1 + x_2 y_2 (1 + \delta_2)) (1 + \epsilon_2) = x_1 y_1 (1 + \delta_1) (1 + \epsilon_2) + x_2 y_2 (1 + \delta_2) (1 + \epsilon_2).$$

The structure becomes clearer with s_3 : $s_3 = (s_2 + x_3 y_3 (1 + \delta_3)) (1 + \epsilon_3)$, or

$$s_3 = x_1 y_1 (1 + \delta_1) (1 + \epsilon_2) (1 + \epsilon_3) + x_2 y_2 (1 + \delta_2) (1 + \epsilon_2) (1 + \epsilon_3) + x_3 y_3 (1 + \delta_3) (1 + \epsilon_3).$$

I know this is rather ugly, but since each of the $|\delta_i|, |\epsilon_i| \leq \mu$, we can “simplify” a bit:

$$s_3 = x_1 y_1 (1 + \delta_1 + \epsilon_2 + \epsilon_3) + x_2 y_2 (1 + \delta_2 + \epsilon_2 + \epsilon_3) + x_3 y_3 (1 + \delta_3 + \epsilon_3) + O(\mu^2).$$

Each term above has one δ and some ϵ 's. The emerging pattern is

$$s_k = \sum_{i=1}^k x_i y_i (1 + (\text{up to } k \text{ rounding terms})) + O(\mu^2).$$

The number of ϵ terms goes down as i increases (the last term will only have 1).

Now we apply the triangle inequality (and $|\delta_i|, |\epsilon_i| \leq \mu$) to the difference between the computed and exact values:

$$\begin{aligned} |s_n - x^T y| &= \left| \sum_{i=1}^n x_i y_i (1 + (\text{up to } n \text{ rounding terms})) - \sum_{i=1}^n x_i y_i + O(\mu^2) \right| \\ &= \left| \sum_{i=1}^n x_i y_i (\text{up to } n \text{ rounding terms}) + O(\mu^2) \right| \\ &\leq \sum_{i=1}^n |x_i| |y_i| n |\text{rounding term bound}| + O(\mu^2) \\ &\leq \sum_{i=1}^n |x_i| |y_i| n \mu + O(\mu^2) \\ &= n \mu |x|^T |y| + O(\mu^2) \end{aligned}$$

As long as $x^T y \neq 0$, we can write this result as

$$\frac{|x^T y - \text{fl}(x^T y)|}{|x^T y|} \leq \mu n \frac{|x|^T |y|}{|x^T y|} + O(\mu^2).$$

In this form, the risk of cancellation is explicit. This little theorem (which can be given in other forms) is the error analysis workhorse in numerical linear algebra.

The standard algorithm we just analyzed is not the only one; see, e.g. Kahan summation or pairwise (cascade, divide-and-conquer) summation.