Computing the Frobenius norm using cr_hypot

Vedran Novaković

August 8, 2025

Theorem 1. *Let* $\mathbf{x} = \begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix}^T$ *be a vector of finite floating-point values, and* $\|\mathbf{x}\|_F$ *its Frobenius norm. If its approximation* $\underline{\|\mathbf{x}\|_F}$ *is computed as* $\underline{\|\mathbf{x}\|_F} = \underline{f_n}$*, where*

$$\underline{f_1} = f_1 = |x_1|, \qquad 2 \le i \le n \implies \underline{f_i} = \mathrm{hypot}(\underline{f_{i-1}}, x_i), \tag{1}$$

then, barring any overflow and inexact underflow, for all i *such that* $2 \le i \le n$ *it holds*

$$x_i \ne 0 \implies \underline{f_i} = f_i(1+\epsilon_i), \qquad 1+\epsilon_i = \sqrt{1 + \epsilon_{i-1}(2+\epsilon_{i-1})\frac{f_{i-1}^2}{f_i^2}}(1+\epsilon_i'), \tag{2}$$

where $|\epsilon_i'| \le \varepsilon$ *if* hypot *is* cr_hypot*, with* ε *being the machine precision, and* $\epsilon_1 = 0$*. If* $x_i = 0$ *then* $\underline{f_i} = \underline{f_{i-1}}$*. If a lower bound of* ϵ_{i-1} *is denoted by* ϵ_{i-1}^- *and an upper bound by* ϵ_{i-1}^+*, where* $\epsilon_1^- = \epsilon_1^+ = 0$*, then, when* hypot *is* cr_hypot *and* $0 \ge \epsilon_{i-1}^- \ge -1$*, the relative error factor* $1+\epsilon_i$ *from* (2) *can be bounded irrespectively of* $x_i \ne 0$ *and* f_i *as*

$$\begin{aligned} 1+\epsilon_i^- &= \sqrt{1+\epsilon_{i-1}^-(2+\epsilon_{i-1}^-)}(1-\varepsilon) \le 1+\epsilon_i \\ &\le \sqrt{1+\epsilon_{i-1}^+(2+\epsilon_{i-1}^+)}(1+\varepsilon) = 1+\epsilon_i^+. \end{aligned} \tag{3}$$

Proof. For $i = 1$ it holds $\epsilon_1 = 0$. Assuming that (2) holds for all $j < i$, where $2 \le i \le n$, and that $x_i \ne 0$, it follows

$$\underline{f_i} = \sqrt{\underline{f_{i-1}^2} + x_i^2}(1+\epsilon_i') = \sqrt{f_{i-1}^2(1+\epsilon_{i-1})^2 + x_i^2}(1+\epsilon_i'). \tag{4}$$

If the term under the square root on the right hand side of (4) is written as

$$f_{i-1}^2(1+\epsilon_{i-1})^2 + x_i^2 = (f_{i-1}^2 + x_i^2)(1+y), \tag{5}$$

then an easy algebraic manipulation gives

$$y = \epsilon_{i-1}(2 + \epsilon_{i-1}) \frac{f_{i-1}^2}{f_{i-1}^2 + x_i^2} = \epsilon_{i-1}(2 + \epsilon_{i-1}) \frac{f_{i-1}^2}{f_i^2}, \tag{6}$$

what, after taking the absolute value of both sides of (6), leads to

$$|y| \le |\epsilon_{i-1}|(2 + |\epsilon_{i-1}|) \frac{f_{i-1}^2}{f_i^2} \le |\epsilon_{i-1}|(2 + |\epsilon_{i-1}|). \tag{7}$$

Substituting (5) into (4) yields

$$\underline{f_i} = \sqrt{f_{i-1}^2 + x_i^2}\sqrt{1+y}(1 + \epsilon_i') = f_i\sqrt{1+y}(1 + \epsilon_i') = f_i(1 + \epsilon_i),$$

where $(1 + \epsilon_i) = \sqrt{1+y}(1 + \epsilon_i')$, as claimed in (2). The bounds (3) for $1 + \epsilon_i$ when $x_i \neq 0$ follow from (7) and the fact that the function $x \mapsto x(2 + x)$ is monotonically increasing for $x \ge -1$. $\square$

More practically, (3) can be further simplified as $-\epsilon_i^- < \epsilon_i^+ < i\varepsilon$ when

$$((\varepsilon = 2^{-24}) \wedge (3 \le i \le 5793)) \quad \vee \quad ((\varepsilon = 2^{-53}) \wedge (3 \le i \le 134217729)),$$

what follows from evaluating (3) iteratively over i using the MPFR library [1] with 2048 bits of precision.

A result similar to Theorem 1 can be obtained in the case of combining two partial norms as

$$\underline{f_k} = \mathrm{hypot}(\underline{f_i}, \underline{f_j}),$$

e.g., when OpenMP-reducing the partial, per-thread results. Bear in mind that the OpenMP standard [2, §7.6.7] leaves the reduction order unspecified.

References

[1] Laurent Fousse, Guillaume Hanrot, Vincent Lefèvre, Patrick Pélissier, and Paul Zimmermann. MPFR: A multiple-precision binary floating-point library with correct rounding. *ACM Trans. Math. Softw.*, 33(2):13, 2007.

[2] OpenMP Architecture Review Board. OpenMP API 6.0 Specification. online: `https://www.openmp.org/wp-content/uploads/OpenMP-API-Specification-6-0.pdf`, Nov 2024.