

Mathematics 502 Problem Set 1 Outline of Solutions

QUESTION 1.10. When the sum $\sum_{i=1}^d x_i y_i$ is evaluated in floating point arithmetic in the usual way, the result is

$$\text{fl} \left(\sum_{i=1}^d x_i y_i \right) = \sum_{i=1}^d \left(x_i y_i (1 + \rho_i) \prod_{j=i}^{d'} (1 + \rho'_j) \right).$$

The rounding error ρ_i occurs when the product $x_i y_i$ is computed; rounding errors ρ'_j occur as the term $x_j y_j$ is added in to the sum; and the prime on the product symbol indicates that when $i = 1$ the product starts at $j = 2$ instead of $j = 1$.

Each rounding error satisfies $|\rho_i| \leq \varepsilon$ or $|\rho'_j| \leq \varepsilon$ provided no underflow or overflow occurs. By the usual approximation, then, (second display, p. 16) the product of the roundoff terms multiplying $x_i y_i$ is bounded by $1 + d\varepsilon$.

To bound the roundoff in the matrix product AB , use the fact that the (i, j) entry of AB is the dot product of row i of A with column j of B .

QUESTION 1.16 We outline the proof of part 6. Part 7 is proved on page 23. First, denoting by $A_{,i}$ the i -th column vector of A , let j be the index of the column of A having the largest 1-norm. Then if e_j is the j -th column of the identity matrix, we have

$$\|A\|_1 \geq \|Ae_j\|_1 = \|A_{,j}\|_1$$

so the 1-norm of A is at least equal to the maximum column sum. On the other hand, if v is any vector with $\|v\|_1 = 1$ then we have

$$\|Av\|_1 = \left\| \sum_{i=1}^n v_i A_{,i} \right\|_1 \leq \sum_{i=1}^n |v_i| \|A_{,i}\|_1 \leq \|A_{,j}\|_1 \sum_{i=1}^n |v_i| = \|A_{,j}\|_1 \|v\|_1 = \|A_{,j}\|_1$$

so that also $\|A\|_1 \leq \|A_{,j}\|_1$, completing the proof.

QUESTION 2.3 The proof of inequality (2.1) occurs in three steps:

$$\begin{aligned} \|\delta x\|_2 &= \|A^{-1}(-\delta A \hat{x} + \delta b)\|_2 \\ &\leq \|A^{-1}\|_2 \cdot \|-\delta A \hat{x} + \delta b\|_2 \\ &\leq \|A^{-1}\|_2 (\|\delta A \hat{x}\|_2 + \|\delta b\|_2) \\ &\leq \|A^{-1}\|_2 (\|\delta A\|_2 \|\hat{x}\|_2 + \|\delta b\|_2). \end{aligned}$$

The first and third steps use Lemma 7.1; the middle step uses the triangle inequality. To achieve equality in (2.1), each step must be an equality. To this end, let u be a unit vector such that $v = A^{-1}u$ satisfies

$$\|v\|_2 = \|A^{-1}\|_2.$$

Then equality will be achieved in (2.1) provided

- (a) The vector $-\delta A\hat{x} + \delta b$ is a scalar multiple of u .
- (b) The vector $-\delta A\hat{x}$ is a positive scalar multiple of δb .
- (c) The matrix δA has the form $\delta A = \beta u\hat{x}^T$.

Let $\delta b = \alpha u$, with $\alpha > 0$ a small parameter to be determined. Let $\delta A = -\beta u(x + \gamma v)^T$, with $\beta > 0$ and γ to be determined. This takes care of (a) and (b). A short calculation reveals that

$$\delta x = A^{-1}(-\delta A(x + \delta x) + \delta b) = (\beta(x + \gamma v)^T(x + \delta x) + \alpha)v$$

is then a scalar multiple of v . Thus (c) is true if we make $\delta x = \gamma v$. This will be true if, given α and β we choose γ to be the root of the quadratic equation

$$\gamma = \beta\|x + \gamma v\|_2^2 + \alpha$$

that is near $\beta\|x\|_2^2 + \alpha$. Finally, we guarantee that $A + \delta A$ is invertible by choosing α and β small enough so that γ is also small and $\|A^{-1}\delta A\|_2 < \frac{1}{2}$, say.