DATA 1010
In-class exercises
Samuel S. Watson
17 September 2018

Problem 1

Find the difference between the largest and second-largest representable `Float64` values.

Solution

The largest representable `Float64` is between 2^{1023} and 2^{1024}, and that interval is into 2^{52} equal-length intervals. Therefore, the gap between representable numbers in that interval is $(2^{1024} - 2^{1023}) \div 2^{52} = 2^{971}$.

Problem 2

Between which two consecutive powers of 2 are the representable numbers exactly the integers in that range?

Solution

The interval from 2^{52} to 2^{53} has length 2^{52} and is divided into 2^{52} equal-length intervals. Therefore, the representable numbers in this interval are exactly the integers.

Problem 3

Discuss the error in each of the following scenarios using the terms *roundoff error, truncation error, or statistical error*.

(i) We use the trapezoid rule with 1000 trapezoids to approximate $\int_0^{10} \frac{1}{4+x^4}\,\mathrm{d}x$.

(ii) We are trying to approximate $f'(5)$ for some function `f` that we can compute, and we attempt to do so by running `(f(5 + 0.5^100) - f(5))/0.5^100`. We fail to get a reasonable answer.

(iii) To approximate the minimum of a function $f : [0,1] \to \mathbb{R}$, we evaluate f at 100 randomly selected points in $[0,1]$ and return the smallest value obtained.

Solution

(i) The more trapezoids we use, the more accurate our answer will be. The difference between the exact answer and the value we get when we stop at 1000 trapezoids is truncation error.

(ii) The real problem here is roundoff error. `5 + 0.5^100` gets rounded off to 5.0, so the numerator will always evaluate to 0. However, even if we used a `BigFloat` version of each of these values, there would still be truncation error in this approximation, since the expression we used was obtained by cutting off the limit in the definition of the derivative at a small but positive increment size.

(iii) This is an example of statistical error, since the output of the algorithm depends on the randomness we use to select the points.

Problem 4

Consider a function $S : \mathbb{R} \to \mathbb{R}$. If the input changes from a to $a + \Delta a$ for some small value Δa, then the output changes to approximately $S(a) + \frac{\mathrm{d}}{\mathrm{d}a} S(a)\, \Delta a$. Calculate the ratio of the *relative change* in the output to the relative change in the input, and show that you get

$$\frac{a \frac{\mathrm{d}}{\mathrm{d}a} S(a)}{S(a)}.$$

Solution

The relative change in output is

$$\frac{\frac{\mathrm{d}}{\mathrm{d}a} S(a) \Delta a}{S(a)},$$

and the relative change in input is $\Delta a / a$. Dividing these two quantities gives

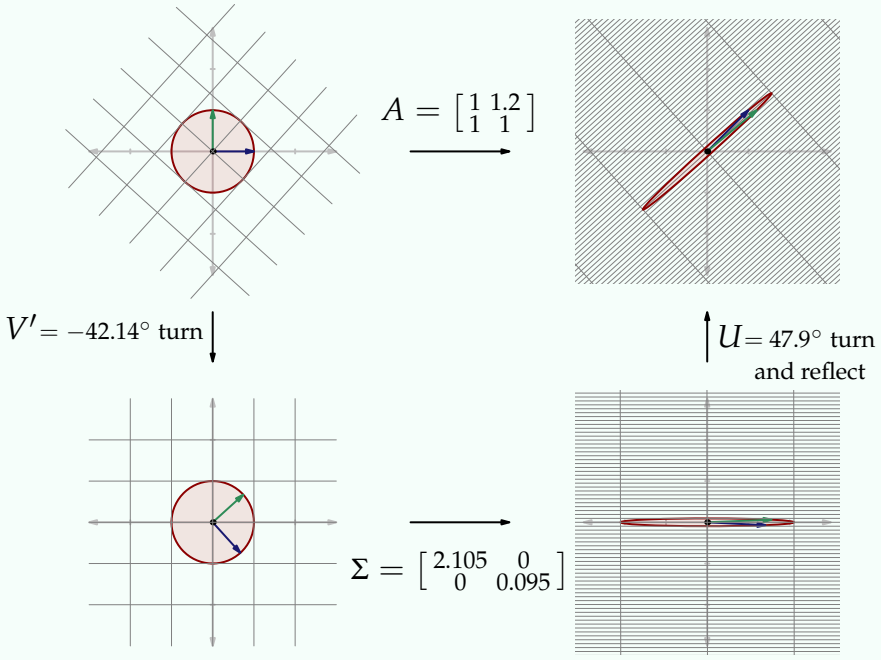
$$\frac{a \frac{\mathrm{d}}{\mathrm{d}a} S(a)}{S(a)},$$

as desired.

Problem 5

Show that the condition number of a matrix A is equal to the ratio of its largest and smallest singular values.

Interpret your resulting by explaining how to choose two vectors with small relative difference which are mapped to two vectors with large relative difference by A, assuming that A has a singular value which is many times larger than another. Use the figure below to help with the intuition.



Solution

The Jacobian of the transformation $\mathbf{x} \mapsto A\mathbf{x}$ is the matrix A itself, and the operator norm of A is equal to its largest singular value. Therefore, to maximize $\kappa(\mathbf{a}) = \frac{|\mathbf{a}|\|J(\mathbf{a})\|}{|\mathbf{S}(\mathbf{a})|}$, we minimize the ratio $|S(\mathbf{a})|/|\mathbf{a}|$. This ratio is minimized when $\mathbf{a}$ is the right singular vector with the least singular value. Therefore, the maximum possible value of κ is the ratio of the largest singular value of A to the smallest singular value of A.

Problem 6

The determinant of a 2×2 matrix can be close to zero either because both singular values are small or because one of the singular values is very small while the other is not.

Consider a matrix like

$$\begin{bmatrix} 1.01 & 1 \\ 1 & 1 \end{bmatrix}$$

This matrix has a small determinant (0.01). Are both singular values small? Is it possible that a 2×2 matrix has unit-length columns and two small singular values (having length less than or equal to 0.1, say)?

Solution

This is not possible. The largest singular value of A is equal to the maximum length of the image of a unit vector under A. Since the columns of A are the images under A of unit vectors (namely, the standard basis vectors), the columns have unit length implies that the largest singular value of A is no smaller than 1.