Appendix F

Trace precision

F.1 The problem

The trace value λ_j is computed by the FOX algorithm:

$$\lambda_j = \sum_{i=1}^{j} e_i \, C \, A^i \tag{F.1}$$

This value is used to compute the quantity

$$\Delta = (\lambda_{i_2} - \lambda_{i_1 - 1}) A^{-i_1} \qquad (i_2 > i_1) \tag{F.2}$$

The exponentially decreasing A^i values causes λ_j to quickly converge to a steady value for large j. As time (i_1 and i_2) increases, equation F.2 multiples an exponentially increasing value (A^{-i_1}) by an exponentially decreasing difference ($\lambda_{i_2} - \lambda_{i_1 - 1}$). Thus Δ quickly loses precision and can cause numerical errors in the FOX algorithm.

This appendix will quantify the amount of precision lost. The worst case precision loss is calculated, so it will be assumed that $e_i = 1$, and that $i_2 = \infty$ and $i_1 = j + 1$ (with $j \gg 1$).

F.2 The scalar case

First the precision lost when computing $(\lambda_\infty - \lambda_j)\vartheta^{-(j+1)}$ will be calculated, where

$$\lambda_j = \sum_{i=1}^{j} \vartheta^i \qquad , |\vartheta| < 1 \tag{F.3}$$

$$= \frac{\vartheta(\vartheta^j - 1)}{\vartheta - 1} \tag{F.4}$$

$$\text{and} \qquad \lambda_\infty = \frac{\vartheta}{1 - \vartheta} \tag{F.5}$$

Now, λ_∞ and λ_j have a similar magnitude because $j \gg 1$, so the decimal precision lost in the difference $\lambda_\infty - \lambda_j$ is (referring to figure F.1):

$$\text{precision lost} = k_1 - k_2 \tag{F.6}$$

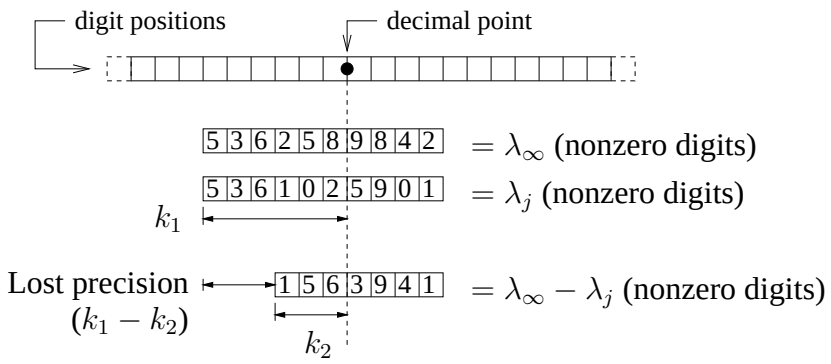

Figure F.1: *Calculating the precision lost in the difference $\lambda_\infty - \lambda_j$.*

$$\begin{aligned}
&\approx \log_{10}(\lambda_\infty) - \log_{10}(\lambda_\infty - \lambda_j) && \text{(F.7)}\\
&= \log_{10}\left(\frac{\lambda_\infty}{\lambda_\infty - \lambda_j}\right) && \text{(F.8)}\\
&= \log_{10}\left(\frac{1}{\vartheta^{j+1}}\right) && \text{(F.9)}\\
&= -(j+1)\log_{10}\vartheta && \text{(F.10)}\\
&\approx j(-\log_{10}\vartheta) && \text{(F.11)}
\end{aligned}$$

When all of the difference's precision has been lost, numerical error will dominate the result.

F.3 The matrix case

The matrix A can be decomposed into a diagonal matrix Λ of eigenvalues and a matrix S of eigenvectors as follows:

$$A = S\,\Lambda\,S^{-1} \qquad \text{(F.12)}$$

Thus

$$\begin{aligned}
\lambda_j &= \sum_{i=1}^{j} C\,A^i && \text{(F.13)}\\
&= \sum_{i=1}^{j} C\,(S\,\Lambda^i\,S^{-1}) && \text{(F.14)}\\
&= C\,S\left[\sum_{i=1}^{j}\Lambda^i\right]S^{-1} && \text{(F.15)}\\
&= C\,S\begin{bmatrix} \sum_{i=1}^{j}\vartheta_1^i & & \cdots & 0 \\ & \sum_{i=1}^{j}\vartheta_2^i & & \\ \vdots & & \ddots & \\ 0 & & & \sum_{i=1}^{j}\vartheta_n^i \end{bmatrix}S^{-1} && \text{(F.16)}
\end{aligned}$$

where $\vartheta_1 \ldots \vartheta_n$ are the eigenvalues of A (they should all have magnitude $|\vartheta| < 1$ for the FOX algorithm). Each diagonal in the matrix is equivalent to an instance of equation F.3. From this it is apparent that the first diagonal element to lose precision will cause the entire value of Δ to lose precision. From equation F.11 this first element will be the one with the smallest (i.e. least magnitude) eigenvalue. It is curious that this smallest eigenvalue has the shortest lived effect on the eligibility profile, but the largest effect on numerical accuracy.

F.4 Trace buffer size

Let ϑ_s be the smallest (in magnitude) eigenvalue of A. The trace buffer size σ is selected from a parameter p such that, at most, no more than p decimal digits of precision are lost. From equation F.11:

$$p = \sigma\left(-\log_{10}\vartheta_s\right) \tag{F.17}$$

so

$$\text{maximum } \sigma = -\frac{p}{\log_{10}\vartheta_s} \tag{F.18}$$

$$= \frac{-p\ \ln 10}{\ln\vartheta_s} \tag{F.19}$$

The maximum eligibility decay (relative to its initial value) at the end of the trace buffer is (assuming that all the eigenvalues are "close" to ϑ_s) :

$$\text{maximum decay} = \vartheta_s^{\sigma} \tag{F.20}$$

$$= \left(10^{-p/\sigma}\right)^{\sigma} \tag{F.21}$$

$$= \frac{1}{10^p} \tag{F.22}$$

For example a good compromise is to select $p = 2$ (two digits of precision lost and a maximum decay of 0.01). Then:

$$\text{maximum } \sigma = \frac{-4.605}{\ln\vartheta_s} \tag{F.23}$$

For reference, the IEEE single precision floating point data type has 7.2 decimal digits of precision (24 bits of mantissa) so the loss of two digits of precision is no problem.