# TSTE18 Digital Arithmetic
# Seminar 10

Oscar Gustafsson

## Approximation of elementary functions

- **Errors**
- **Range reduction**
- **Tables**
- **(Piecewise) polynomial approximations**
- **Bi-/multi-partite tables**

- CORDIC
- Convergence-based approximation

## Errors

- Consider an approximation $\tilde{f}(X) \approx f(X)$ with an error $\epsilon(X) = \tilde{f}(X) - f(X)$
- $X$ is a $B$-bit number, so there are $2^B$ different inputs to consider
- Mean error and error variance are interesting statistical measures of the error
- Different error measures are possible
  - Maximum absolute error: $\max|\epsilon(X)|$
  - Mean square error: $\frac{1}{2^B}\sum_X \epsilon(X)^2$
  - Mean absolute error: $\frac{1}{2^B}\sum_X |\epsilon(X)|$
- Often we consider the maximum absolute error and state the number of correct fractional bits: $-\log_2 \max \epsilon(X)$
- Different applications may have different requirements
- A faithfully rounded approximation is correct within one ulp

## Range reduction

- For many functions it is possible to find a modified input which gives a simple relation to the original input and also for the required output
- Hence, the range of the arguments that must be evaluated can be reduced
- Examples:
  - $\sin(X) = \sin(X + 2\pi)$ – Can add/subtract an arbitrary integer multiple of $2\pi$
  - $\sin(X) = -\sin(-X)$ – Can change sign before evaluation and then again after
  - For $\sin, \cos, \tan$ it is enough to evaluate arguments between $0 \leq X \leq \pi/2$ (acutally $0 \leq X \leq \pi/4$)
  - $\sqrt{X} = \frac{\sqrt{4X}}{2}$ – enough to evaluate arguments between $k \leq X < 4k$ for some $k$
  - $\frac{1}{X} = \frac{1}{2^m}\frac{2^m}{X}$ – enough to evaluate arguments between $k \leq X < 2k$ for some $k$

## Range reduction

- This procedure is called range reduction
- Useful for evaluating $\sin(10^{20})$ etc
- More advanced techniques also available for e.g. log
- Note that when reducing $10^{20}$ to a value between $0 \leq X \leq \pi/2$ a rather large value is required ( $1.6 \times 10^{19}$) to be multiplied $2\pi$
- An error in $\pi$ will be amplified $10^{19}$ times...

## Tables

- The simplest way to approximate an arbitrary function is to use a table storing all the required values
- Table size assuming $B$ input bits and $W$ output bits: $W \times 2^B$ memory bits
- Approximation error is easily controlled to within 0.5 ulp by selecting the correct values
- Becomes very large for long input word lengths

## Polynomial approximation

- An alternative is to use a polynomial to approximate the value
- The Taylor-series expansion gives a good initial approximation
- The expansion of $f(X)$ about $X = a$ is

$$f(X) = \sum_{i=0}^{\infty} f^{(i)}(a)\frac{(X-a)^i}{i!} \tag{1}$$

with the error of omitting all terms of degree higher than $m$ is

$$f^{(m+1)}(a+\mu(X-a))\frac{(X-a)^{m+1}}{(m+1)!} \tag{2}$$

for some $0 < \mu < 1$

## Polynomial approximation

- The expansion of $f(X)$ about $X = 0$ is the Maclaurin-series expansion

$$f(X) = \sum_{i=0}^{\infty} f^{(i)}(0)\frac{(X-0)^i}{i!} \tag{3}$$

with the error of omitting all terms of degree higher than $m$ is

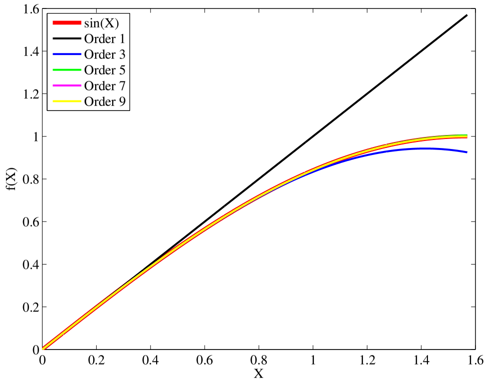
$$f^{(m+1)}(\mu X)\frac{(X)^{m+1}}{(m+1)!} \tag{4}$$

for some $0 < \mu < 1$

## Polynomial approximation

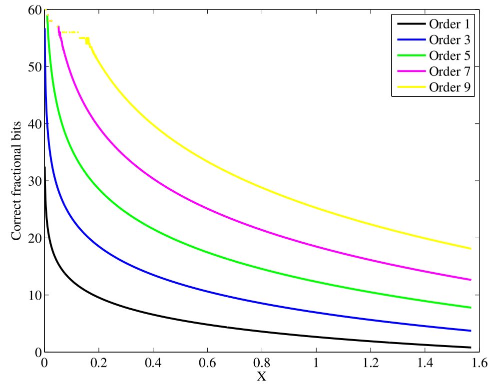
- Consider the function $\sin(X)$ with a Maclaurin expansion as

$$\sin(X) \approx X - \frac{X^3}{3!} + \frac{X^5}{5!} - \frac{X^7}{7!} + \dots \tag{5}$$




## Polynomial approximation

- Approximation error/correct fractional bits




## Polynomial approximation

- For a fixed-point representation it may make sense to scale the input argument in some cases
- For example, for $\sin(X)$ the input range $0 \leq X \leq \pi/2$ maps badly to the $0 \leq X \leq 1 - 2^{-B}$ range commonly used
- Better to approximate $\sin\left(\frac{\pi X}{2}\right)$

$$\sin\left(\frac{\pi X}{2}\right) \approx \frac{\pi X}{2} - \frac{\pi^3 X^3}{8 \times 3!} + \frac{\pi^5 X^5}{32 \times 5!} - \frac{\pi^7 X^7}{128 \times 7!} + \dots \tag{6}$$

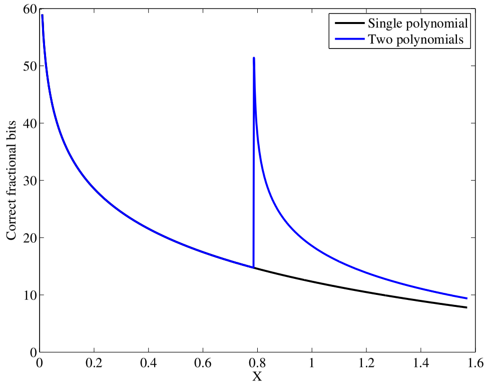
## Piecewise polynomial approximation

- Possible to use multiple polynomials for different segments
- Simplest way is to divide the range in similarly sized blocks
- Can use MSBs to decide block
- Compute Taylor expansion for $\sin\left(\frac{\pi X}{2}\right)$ about $X = 1/2$ ($\pi/4$ rad)

$$\begin{aligned}\tilde{f}(X) &= \frac{\sqrt{2}}{2} - \frac{\sqrt{2}\pi^2\left(X - \frac{1}{2}\right)^2}{16} - \frac{\sqrt{2}\pi^3\left(X - \frac{1}{2}\right)^3}{96} + \\ &\quad \frac{\sqrt{2}\pi^4\left(X - \frac{1}{2}\right)^4}{768} + \frac{\pi\sqrt{2}\left(X - \frac{1}{2}\right)}{4} \\ &\approx 0.1794X^4 - 0.8155X^3 + 0.0818X^2 + 1.551X + 0.00196\end{aligned} \tag{7}$$

## Piecewise polynomial approximation

- Approximation error using two polynomials

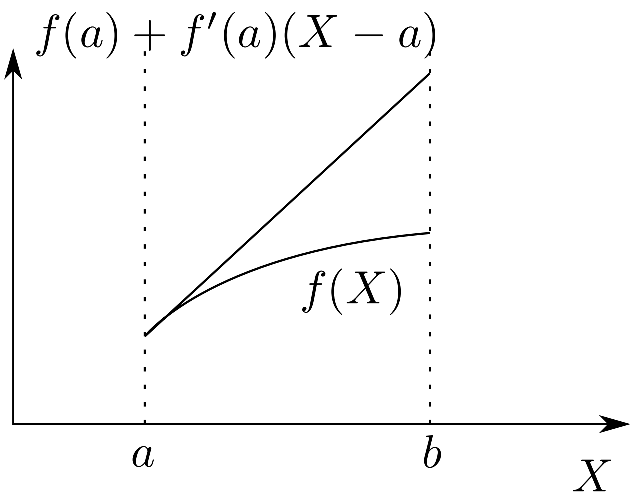



## Piecewise polynomial approximation

- Some observations:
  - In general there is a trade-off between number of segments (size of coefficients memories) and polynomial order (number of coefficient memories and arithmetic operations)
  - We have not yet considered finite word length coefficients
  - Higher resolution at the point where the Taylor series is expanded

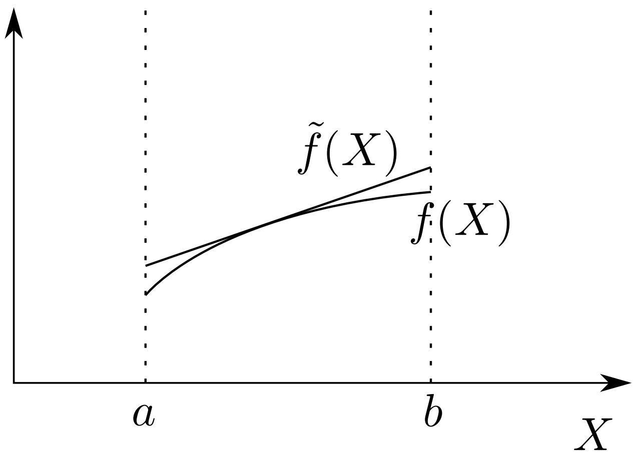
## Piecewise polynomial approximation

- Consider a first order approximation of a segment



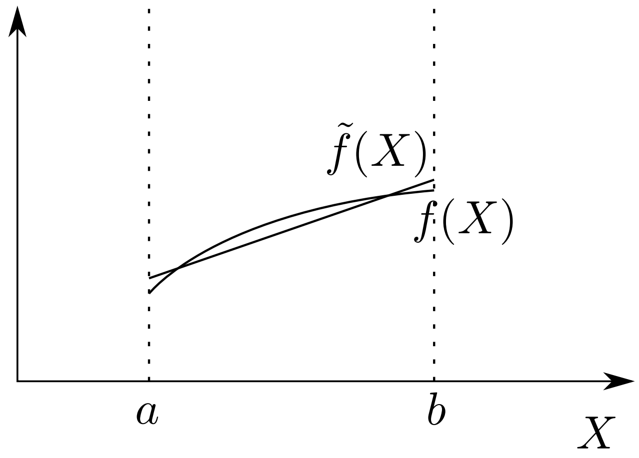

## Piecewise polynomial approximation

- Possible to obtain better approximation expanding around the mid-point



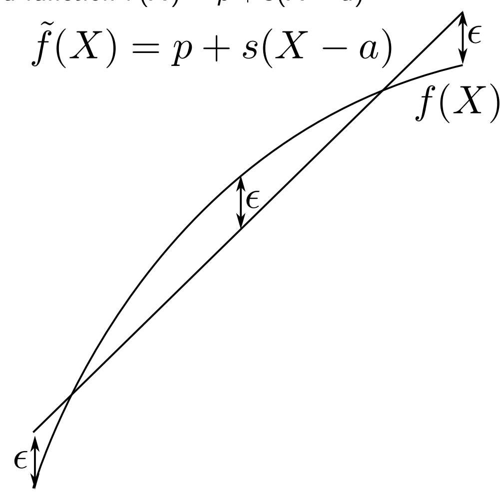
## Piecewise polynomial approximation

- Possible to obtain an even better approximation




## Piecewise polynomial approximation

- Assume a function $\tilde{f}(X) = p + s(X - a)$




- Should have the same approximation error $\epsilon$ at the end points and at the worst case in the middle

## Piecewise polynomial approximation

- Using the observation that the errors are the same gives

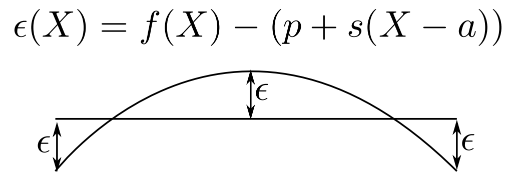
$$\begin{aligned} f(a) - (p + s(a - a)) &= -\epsilon \\ f(m) - (p + s(m - a)) &= \epsilon \\ f(b) - (p + s(b - a)) &= -\epsilon \end{aligned}$$

$$\begin{bmatrix} 1 & 0 & -1 \\ 1 & m - a & 1 \\ 1 & b - a & -1 \end{bmatrix} \begin{bmatrix} p \\ s \\ \epsilon \end{bmatrix} = \begin{bmatrix} f(a) \\ f(m) \\ f(b) \end{bmatrix} \qquad (8)$$

- But where is $m$?

## Piecewise polynomial approximation

- Consider the error function $\epsilon(X) = f(X) - (p + s(X - a))$




- $m$ is where the first derivative of $\epsilon(X)$ is zero

$$f'(m) - s = 0 \qquad (9)$$

- Need to determine $s$
- Slope, $s$

$$s = \frac{f(b) - f(a)}{b - a} \qquad (10)$$

- Solve for $m$ by using $f'(m) = s$

## Piecewise polynomial approximation

- Now, since $s$ and $m$ are known

$$\begin{bmatrix} 1 & -1 \\ 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} p \\ \epsilon \end{bmatrix} = \begin{bmatrix} f(a) \\ f(m) - s(m-a) \\ f(b) - s(b-a) \end{bmatrix} \quad (11)$$

- As the first and last row of the matrix are linearly dependent, remove one of them

$$\begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} p \\ \epsilon \end{bmatrix} = \begin{bmatrix} f(a) \\ f(m) - s(m-a) \end{bmatrix} \quad (12)$$

and solve the equation system

$$\begin{aligned} \begin{bmatrix} p \\ \epsilon \end{bmatrix} &= \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}^{-1} \begin{bmatrix} f(a) \\ f(m) - s(m-a) \end{bmatrix} && (13) \\ &= \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} \end{bmatrix} \begin{bmatrix} f(a) \\ f(m) - s(m-a) \end{bmatrix} && (14) \end{aligned}$$

## Piecewise polynomial approximation

- Possible to find better polynomial coefficients for higher-order polynomials as well
- Example: compute a two-segment linear approximation for $\sin\left(\frac{\pi X}{2}\right)$ using a Taylor expansion and the method earlier described

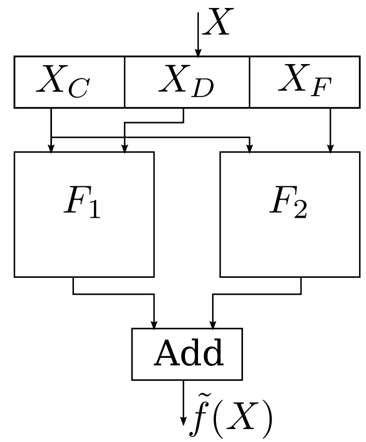
## Bi-partite approximation

- The polynomial schemes use multipliers which may be costly
- Consider $\sin(X)$ where $X$ is split into three different parts $X = (X_C X_D X_F)$

$$\begin{aligned} \sin(X) &= \sin(X_C + X_D + X_F) \\ &= \sin(X_C + X_D) \cdot \cos(X_F) + \cos(X_C + X_D) \cdot \sin(x_F) \\ &= \sin(X_C + X_D) \cdot \cos(X_F) + \cos(X_C) \cdot \cos(X_D) \cdot \sin(X_F) - \\ &\quad \sin(X_C) \cdot \sin(X_D) \cdot \sin(X_F) \\ &\approx \sin(X_C + X_D) + \cos(X_C) \cdot \sin(X_F) \\ &= F_1(X_C, X_D) + F_2(X_C, X_F) \end{aligned}$$

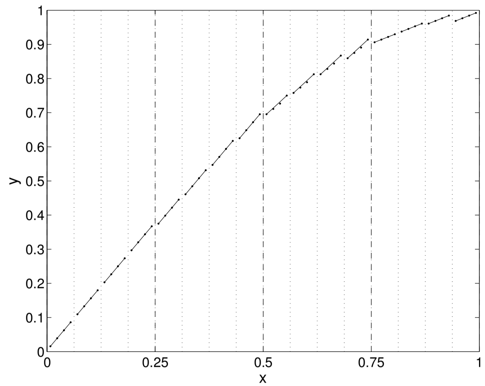
## Bi-partite approximation

- Assuming equal partitions, a total of $2^{\frac{2B}{3}+1}$ words are required compared to $2^B$



## Bi-partite approximation

- Can be seen as a piecewise linear approximation where each slope line is tabulated and used in several segments




## Bi-partite approximation

- In the general case

$$F_1(X_C, X_D) = f(X_C X_D)$$
$$F_2(X_C, X_F) = f'(X_C) \cdot X_F$$

- Can be improved by
  - Using the mid-point derivative
  - Starting from the mid-point and add/subtract the table value (will also reduce the number of words with a factor 2)
  - Split into several tables (multi-partite) corresponding to different derivatives