Constructor University Bremen

Dr. D. Sinden

Spring Semester 2024

JTMS-MAT-13: Numerical Methods

Summary from 21 March 2024

Notes from course, covering basic numerical methods.

This document can be downloaded from
`https://djps.github.io/courses/numericalmethods24/notes`

Note that the proofs for theorems marked with an * where presented in class.

Contents

Recommended Reading

- J. F. Epperson "*An Introduction to Numerical Methods and Analysis*", Wiley 2nd Edition (2013).
- R. L. Burden and J. D. Faires "*Numerical Analysis*", Brooks/Cole 9th Edition (2011).

1 Taylor Series

The Taylor series, or the Taylor expansion of a function, is defined as

Definition 1.1 (Taylor Series)**.** For a function $f : \mathbb{R} \mapsto \mathbb{R}$ which is infinitely differentiable at a point c, the Taylor series of $f(c)$ is given by

$$\sum_{k=0}^{\infty} \frac{f^{(k)}(c)}{k!} (x-c)^k .$$

This is a power series, which is convergent for some radius.

Theorem 1 (Taylor's Theorem). For a function $f \in C^{n+1}([a,b])$, i.e. f is $(n+1)$-times continuously differentiable in the interval $[a,b]$, then for some c in the interval, the function can be written as

$$f(x) = \sum_{k=0}^{n} \frac{f^{(k)}(c)}{k!} (x-c)^k + \frac{f^{(n+1)}(\xi)}{(n+1)!} (x-c)^{n+1}$$

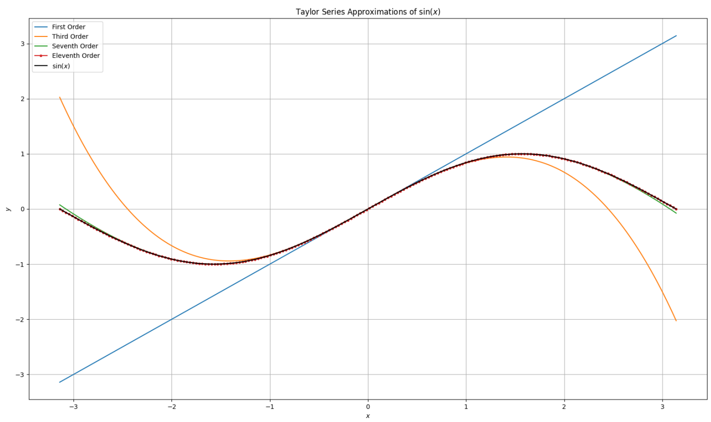
for some value $\xi \in [a,b]$ where

$$\lim_{\xi \to c} \frac{f^{(n+1)}(\xi)}{(n+1)!} (x-c)^{n+1} = 0.$$

Example. With $f(x) = \sin(x)$ around $c = 0$. Thus, as $f' = \cos(x)$, it can be shown that

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

Note that in this example, only odd powers of x contribute to the expansion.


Taylor Series Approximations of sin(x)
First Order
Third Order
Seventh Order
Eleventh Order
sin(x)
x
y

2 Errors

Definition 2.1 (Absolute and Relative Errors).
Let $\tilde{a}$ be an approximation to a, then the **absolute error** is given by

$$|\tilde{a} - a|.$$

If $|a| \neq 0$, the **relative error** may be given by

$$\left|\frac{\tilde{a} - a}{a}\right|.$$

The error bound is the magnitude of the admissible error.

Theorem 2. For both addition and subtraction the bounds for the absolute error are added. In division and multiplication the bounds for the relative errors are added.

Definition 2.2 (Linear Sensitivity to Uncertainties).
If $y(x)$ is a smooth function, i.e. is differentiable, then $|y'|$ can be interpreted as the **linear sensitivity** of $y(x)$ to uncertainties in x.

For functions of several variables, i.e. $f : \mathbb{R}^n \to \mathbb{R}$, then

$$|\Delta y| \leq \sum_{i=1}^{n} \left|\frac{\partial y}{\partial x_i}\right| |\Delta x_i|$$

where $|\Delta x_i| = |\tilde{x}_i - x_i|$ for an approximation $\tilde{x}_i$, thus $|\Delta y_i| = |\tilde{y}_i - y_i| = |f(\tilde{x}_i) - (x_i)|$.

3 Number Representations

Definition 3.1 (Base Representation).
Let $b \in \mathbb{N} \backslash \{1\}$. Every number $x \in \mathbb{N}_0$ can be written as a unique expansion with respect to base b as

$$(x)_b = a_0 b^0 + a_1 b^1 + \ldots + a_n b^n = \sum_{i=0}^{n} a_i b^i$$

A number can be written in a nested form:

$$\begin{aligned}(x)_b &= a_0 b^0 + a_1 b^1 + \ldots a_n b^n \\ &= a_0 + b\left(a_1 + b\left(a_2 + b\left(a_3 + \ldots + b a_n\right) \ldots\right)\right.\end{aligned}$$

with $a_i < \mathbb{N}_0$ and $a_i < b$, i.e. $a_i \in \{0, \ldots, b-1\}$.
For a real number, $x \in \mathbb{R}$, write

$$\begin{aligned}x &= \sum_{i=0}^{n} a_i b^i + \sum_{i=1}^{\infty} \alpha_i b^{-i} \\ &= a_n \ldots a_0 \cdot \alpha_1 \alpha_2 \ldots\end{aligned}$$

Algorithm (Euclid).
Euclid's algorithm can convert number x in base 10, i.e. $(x)_{10}$ into another base, b, i.e. $(x)_b$.

1. Input $(x)_{10}$
2. Determine the smallest integer n such that $x < b^{n+1}$
3. Let $y = x$. Then for $i = n, \ldots, 0$

$$\begin{aligned}a_i &= y \text{ div } b^i \\ y &= y \text{ mod } b^i\end{aligned}$$

   which at each steps provides an a_i and updates y.
4. Output as $(x)_b = a_n a_{n-1} \cdots a_0$

Algorithm (Horner).

1. Input $(x)_{10}$
2. Set $i = 0$

3. Let $y = x$. Then while $y > 0$

$$\begin{aligned} a_i &= y \text{ div } b \\ y &= y \bmod b \\ i &= i + 1 \end{aligned}$$

which at each steps provides an a_i and updates y.

4. Output as $(x)_b = a_n a_{n-1} \cdots a_0$

Definition 3.2 (Normalized Floating Point Representations).
Normalized floating point representations with respect to some base b, store a number x as

$$x = 0 \cdot a_1 \dots a_k \times b^n$$

where the $a_i \in \{0, 1, \dots b-1\}$ are called the **digits**, k is the **precision** and n is the **exponent**. The set $a_1, \dots, a_k$ is called the **mantissa**. Impose that $a_1 \neq 0$, it makes the representation unique.

Theorem 3. Let x and y be two normalized floating point numbers with $x > y > 0$ and base $b = 2$. If there exists integers p and $q \in \mathbb{N}_0$ such that

$$2^{-p} \leq 1 - \frac{y}{x} \leq 2^{-q}$$

then, at most p and at least q significant bits (i.e. significant figures written in base 2) are lost during subtraction.

4 Linear Systems

Definition 4.1 (Systems of Linear Equations). A system of linear equations (or a linear system) is a collection of one or more linear equations involving the same variables. If there are m equations with n unknown variables to solve for, i.e.

$$\begin{aligned} a_{1,1}x_1 + a_{1,2}x_2 + \ldots + a_{1,n}x_n &= b_1 \\ a_{2,1}x_1 + a_{2,2}x_2 + \cdots + a_{2,n}x_n &= b_2 \\ &\vdots \\ a_{m,1}x_1 + a_{m,2}x_2 + \cdots + a_{m,n}x_n &= b_m \end{aligned}$$

then the system of linear equations can be written in matrix form $Ax = b$, where

$$A = \begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m,1} & a_{m,2} & \cdots & a_{m,n} \end{pmatrix}, \quad x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}, \quad \text{and} \quad b = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix},$$

so that $A \in \mathbb{R}^{m\times n}$, $x \in \mathbb{R}^n$ and $b \in \mathbb{R}^m$.

Definition 4.2 (Banded Systems). A **banded** matrix is a matrix whose non-zero entries are confined to a diagonal band, comprising the main diagonal and zero or more diagonals on either side.

Definition 4.3 (Symmetric Systems). A square matrix A is **symmetric** if $A = A^T$, that is, $a_{i,j} = a_{j,i}$ for all indices i and j.

A square matrix is said to be **Hermitian** if the matrix is equal to its conjugate transpose, i.e. $a_{i,j} = \overline{a_{j,i}}$ for all indices i and j. A Hermitian matrix is written as A^H.

Definition 4.4 (Positive Definite Matrices). A matrix, M, is said to be **positive definite** if it is symmetric (or Hermitian) and all its eigenvalues are real and positive.

Definition 4.5 (Nonsingular Matrices). A matrix is **non-singular** or **invertible** if there exists a matrix A^{-1} such that $A^{-1}A = AA^{-1} = I$, where I is the identity matrix.

Remarks (Properties of Nonsingular Matrices)**.** For a nonsingular matrix, the following all hold:

- Nonsingular matrix has full rank
- A square matrix is nonsingular if and only if the determinant of the matrix is non-zero.
- If a matrix is singular, both versions of Gaussian elimination will fail due to division by zero, yielding a floating exception error.

Definition 4.6. If $\tilde{x}$ is an approximate solution to the linear problem $Ax = b$, then the **residual** is defined as $r = A\tilde{x} - b$.

If the magnitude of the residual, $|r|$, is large due to rounding, the matrix is said to be **ill-conditioned**.

4.1 Direct Methods

Algorithm (Gaussian Elimination)**.** Gaussian elimination is a method to solve systems of linear equations based on forward elimination (a series of row-wise operations) to convert the matrix, A, to upper triangular form (echelon form), and then back-substitution to solve the system. The row operations are:

- row swapping
- row scaling, i.e. multiplying by a non-zero scalar
- row addition, i.e. adding a multiple of one row to another

For k=1 to $n-1$
  For $i = k+1$ to n
    For $j =$ k to n
      $a_{i,j} = a_{i,j} - \dfrac{a_{i,k}}{a_{k,k}} a_{k,j}$
    $b_i = b_i - \dfrac{a_{i,k}}{a_{k,k}} b_k$
Then back substitute,
$x_n = \dfrac{b_n}{a_{n,n}}$
For i =n-1 to 1
  $y = b_i$
  For $j = n$ to $i+1$
    $y = y - a_{i,j} x_y$
  $x_i = \dfrac{y}{a_{i,i}}$

Algorithm (Gaussian Elimination with Scaled Partial Pivoting). A pivot element is the element of a matrix which is selected first to do certain calculations. Pivoting helps reduce errors due to rounding.

Definition 4.7 (Upper and Lower Triangular Matrices). A square matrix is said to be a **lower triangular matrix** if all the elements above the main diagonal are zero and an **upper triangular** if all the entries below the main diagonal are zero.

Theorem 4 (LU-Decomposition). Let $A \in \mathbb{R}^{n\times n}$ be invertible. Then there exists a decomposition of A such that $A = LU$, where L is a lower triangular matrix and U is an upper triangular matrix, And

$$L = U_1^{-1}U_2^{-1}\cdots U_{n-1}^{-1}$$

where each matrix U_i is a matrix which describes the i^{th} step in forward elimination part of Gaussian elimination

$$U = U_{n-1}\cdots U_2U_1A$$

Definition 4.8 (Cholesky-Decomposition). A symmetric, positive definite matrix can be decomposed as $A = LL^T$.

Algorithm (Cholesky-Decomposition). For i=1 to n

For $j = 1$ to $i-1$
  $y = a_{i,j}$
  For $k = 1$ to j-1
    $y = y - l_{i,k}l_{j,k}$
  $l_{i,j} = y/l_{j,j}$
$y = a_{i,i}$
For k= 1 to $i-1$
  $y = y - l_{i,k}l_{i,k}$
if $y \leq 0$ there is no solution
else $l_{i,i} = \sqrt{y}$

5 Nonlinear Solvers

5.1 Bisection Method

Definition 5.1 (Bisection Method).
The bisection method, when applied in the interval $[a, b]$ to a function $f \in C^0([a, b])$ with $f(a)f(b) < 0$

Bisect the interval into two subintervals $[a, c]$ and $[c, b]$ such that $a < c < b$.

- If $f(c) = 0$ or is sufficiently close, then c is a root
- else, if $f(c)f(a) < 0$ continue in the interval $[a, c]$
- else, if $f(c)f(b) < 0$ continue in the interval $[c, b]$

Theorem 5 (Bisection Method).
The bisection method, when applied in the interval $[a, b]$ to a function $f \in C^0([a, b])$ with $f(a)f(b) < 0$ will compute, after n steps, an approximation c_n of the root r with error

$$|r - c_n| < \frac{b - a}{2n}$$

5.2 Newton's Method

Definition 5.2.
Let a function $f \in C^1([a, b])$, then for an initial guess x_0, Newton's method is

$$x_{n+1} = x_n - \frac{f(x_n)}{f(x_n)}$$

Theorem 6.
When Newton's method converges, it converges to a root, r, of f, i.e. $f(r) = 0$.

Theorem 7.
Let $f \in C^1([a, b])$, with

1. $f(a)f(b) < 0$,
2. $f'(x) \neq 0$ for all $x \in (a, b)$,
3. $f''(x)$ exists, is continuous and either $f''(x) > 0$ or $f''(x) < 0$ for all $x \in (a, b)$.

Then $f(x) = 0$ has exactly one root, r, in the interval and the sequence generated by Newton iterations converges to the root when the initial guess is chosen according to

- if $f(a) < 0$ and $f''(a) < 0$ or $f(a) > 0$ and $f''(a) > 0$ then $x \in [a, r]$

or

- if $f(a) < 0$ and $f''(a) > 0$ or $f(a) > 0$ and $f''(a) < 0$ then $x \in [r, b]$

The iterate in the sequence satisfies

$$|x_n - r| < \frac{f(x_n)}{\min\limits_{x \in [a,b]} |f'(x)|}$$

Theorem 8.
Let $f \in C^1([a, b])$, with

1. $f(a)f(b) < 0$
2. $f'(x) \neq 0$ for all $x \in (a, b)$
3. $f''(x)$ exists and is continuous, i.e. $f(x) \in C^2([a, b])$

Then, if x_0 is close enough to the root r, Newton's method converges quadratically.

5.3 Secant Methods

Definition 5.3.
The secant method is defined as

$$x_{n+1} = x_n - f(x_n) \frac{x_{n-1} - x_n}{f(x_{n-1}) - f(x_n)}$$

Theorem 9.
Let $f \in C^2([a, b])$, and $r \in (a, b)$ such that $f(r) = 0$ and $f'(r) \neq 0$. Furthermore, let

$$x_{n+1} = x_n - f(x_n) \frac{x_{n-1} - x_n}{f(x_{n-1}) - f(x_n)}$$

Then there exists a $\delta > 0$ such that when $|r - x_0| < \delta$ and $|r - x_1| < \delta$, then the following holds:

1. $\lim\limits_{n\to\infty} |r - x_n| = 0 \Leftrightarrow \lim\limits_{n\to\infty} x_n = r$

2. $|r - x_{n+1}| \leq \mu |r - x_n|^\alpha$ with $\alpha = \dfrac{1+\sqrt{5}}{2}$

5.4 Convergence

Definition 5.4. If a sequence x_n converges to r as $n \to \infty$, then it is said to **converge linearly** if there exists a $\mu \in (0,1)$ such that

$$\lim_{n\to\infty} \frac{|x_{n+1} - r|}{|x_n - r|} = \mu$$

The sequences converges **super-linearly** if

$$\lim_{n\to\infty} \frac{|x_{n+1} - r|}{|x_n - r|} = 0$$

and **sub-linearly** if

$$\lim_{n\to\infty} \frac{|x_{n+1} - r|}{|x_n - r|} = 1$$

More generally, a sequence converges with order q if there exists a $\mu > 0$ such that

$$\lim_{n\to\infty} \frac{|x_{n+1} - r|}{|x_n - r|^q} = \mu$$

Thus a sequence is said to converge quadratically when $q = 2$ and exhibit cubic convergence when $q = 3$.

Method	Regularity	Proximity to r	Initial points	Function calls	Convergence
Bisection	$\mathcal{C}^0$	No	2	1	Linear
Newton	$\mathcal{C}^2$	Yes	1	2	Quadratic
Secant	$\mathcal{C}^2$	Yes	1	1	Superlinear

5.5 Systems of Nonlinear Equation

Definition 5.5 (Multi-Dimensional Newton Method)**.** For a vector-valued function $f : \mathbb{R}^n \to \mathbb{R}^n$, which takes as an argument the vector

$$x = (x_1, x_2 \ldots x_n) \,,$$

the **Jacobian** matrix is defined as

$$J = \begin{pmatrix} \dfrac{\partial f_1}{\partial x_1} & \dfrac{\partial f_1}{\partial x_2} & \cdots & \dfrac{\partial f_1}{\partial x_n} \\ \dfrac{\partial f_2}{\partial x_1} & & & \\ \vdots & & & \vdots \\ \dfrac{\partial f_n}{\partial x_1} & \dfrac{\partial f_n}{\partial x_2} & \cdots & \dfrac{\partial f_n}{\partial x_n} \end{pmatrix}$$

where f_1 is the first component of the vector-valued function f.

If the derivatives are evaluated at the vector x, the Jacobian matrix can be parameterised as $J\left(x\right)$. Newton's method can then be written as a vector equation,

$$x_{m+1} = x_m - J^{-1}\left(x_m\right) f\left(x_n\right)$$

where $J^{-1}\left(x_n\right)$ is the inverse of the Jacobian matrix evaluated at the m-iterate of the vector approximation vector which is denoted by x_m.

In practice, as matrix inversion can be computationally expensive, the system

$$J\left(x_n\right)\left(x_{n+1} - x_n\right) = -f\left(x_n\right)$$

is solved for the unknown vector $x_{m+1} - x_m$.

6 Interpolation

Definition 6.1. Given as set of points $p_0, \ldots, p_n \in \mathbb{R}$ and corresponding nodes $u_0, \ldots, u_n \in \mathbb{R}$, a function $f : \mathbb{R} \to \mathbb{R}$ with $f(u_i) = p_i$ is an **interpolating function**.

This can be generalised to higher dimensions, i.e. $f : \mathbb{R} \to \mathbb{R}^N$.

Definition 6.2. If the interpolating function is a polynomial, write can be written as

$$p(u) = \sum_{i=0}^{n} \alpha_i \varphi_i(u)$$

So that for every j, the polynomial satisfies $p(u_j) = \sum_{i=0}^{n} \alpha_i \varphi_i(u_j)$, thus the α_i lead to a linear system of the form

$$\Phi\alpha = p$$

where p is the vector defined the polynomial evaluated at the node points, i.e. $p = p(u_j)$ and Φ is the **collocation matrix**, given by

$$\Phi = \begin{pmatrix} \varphi_0(u_0) & \varphi_1(u_1) & \cdots & \varphi_n(u_n) \\ \vdots & & & \vdots \\ \varphi_0(u_n) & \cdots & \cdots & \varphi_n(u_n) \end{pmatrix}$$

Thus $\alpha = \Phi^{-1}p$.

The collocation matrix is invertible if and only if the set of functions φ are linearly independent.

Definition 6.3. If

$$p(u) = \sum_{i=0}^{n} \alpha_i \varphi_i(u)$$

So that for every j, $p(u_j) = \sum_{i=0}^{n} \alpha_i \varphi_i(u_j)$, thus the α_i lead to a linear system of the form

$$\Phi\alpha = p$$

where Φ is the **Vandermonde matrix**.

Definition 6.4 (Lagrange Polynomials)**.** The **Lagrange form of an interpolating polynomial** is given by

$$p(x) = \sum_{i=0}^{n} \alpha_i l_i(x)$$

where $l_i \in \mathbb{P}\_n$ are such that $l_i(x_j) = \delta_{ij}$. The polynomials $l_i(x) \in \mathbb{P}\_n$ for $i = 0, \ldots, n$, are called **characteristic polynomials** and are given by

$$l_i(x) = \prod_{j=0, j\neq i}^{n} \frac{x - x_j}{x_i - x_j}.$$

Algorithm (Aitken's Algorithm)**.** Aitken's algorithm is an iterative process for evaluating Lagrange interpolation polynomials without explicitly constructing them.

$$p(u) = \sum_{i=0}^{n} p_i^n l_i^n(u)$$

this can be written as

$$\begin{array}{llll}
l(u) = p_0 & & & \\
 & p_0^1(u) & & \\
l(u) = p_1 & & l_0^1(u) & \\
 & p(u) & & l_0^1(u) \\
l(u) = p_2 & & l(u) & \\
 & p(u) & & \\
l(u) = p_3 & & &
\end{array}$$

where the coefficients are evaluated from left to right.

The corresponding method for Newton's form of the interpolating polynomial is called Neville's algorithm.

6.1 Piecewise Polynomial Interpolation

Definition 6.5. A function $s(u)$ is called a **spline** of degree k on the domain $[a, b]$ if $s \in C^{k-1}([a, b])$ and there exists nodes $a = u_0 < u_1 < \ldots < u_m = b$ such that s is a polynomial of degree k for $i = 0, \ldots m - 1$.

Definition 6.6 (B-Splines)**.** A spline is said to be a **b-spline** if it is of the form

$$s\left(u\right) = \sum_{i=0}^{m} \alpha_i \mathcal{N}\_i^{n}\left(u\right)$$

where $\mathcal{N}^n$ are the **basis spline functions** of degree n with minimal support. (That is they are positive in the domain and zero outside). The functions are defined recursively. Let u_i be the set of nodes $u_0, u_1, \ldots, u_m$, then

$$\mathcal{N}\_i^{0}\left(u\right) = \left\{ \begin{array}{ll} 1 & \text{for} \quad u_i \leq u \leq u_{i+1} \\ 0 & \text{else.} \end{array} \right.$$

and

$$\mathcal{N}\_i^{n}\left(u\right) = \alpha_i^{n-1}\left(u\right) \mathcal{N}\_i^{n-1}\left(u\right) + \left(1 - \alpha_{i+1}^{n-1}\left(u\right)\right) \mathcal{N}\_i + 1^{n-1}\left(u\right)$$

where

$$\alpha_i^{n-1}\left(u\right) = \frac{u - u_i}{u_{i+n} - u_i}$$

is a local parameter.

Given data with nodes u_i and values p_i, to interpolate with splines, of order n, requires solving

$$\text{Find} \quad s = \sum_{i=0}^{m} \alpha_i \mathcal{N}_i^n\left(u\right) \quad \text{such that} \quad s\left(u_i\right) = p_i \quad \text{for} \quad i = 0, \ldots, m$$

which is matrix form is $\Phi \alpha = p$, where the collocation matrix, $\Phi \in \mathbb{R}^{(m+1)\times(m+1)}$ is given by

$$\Phi = \begin{pmatrix} \mathcal{N}\_0^{n}\left(u_0\right) & \cdots & \mathcal{N}\_m^{n}\left(u_0\right) \\ \vdots & & \vdots \\ \mathcal{N}\_0^{n}\left(u_m\right) & \cdots & \mathcal{N}\_m^{n}\left(u_m\right) \end{pmatrix}$$

6.2 Least-Squares Approximation

Definition 6.7 (Least-Squares Approximation). Given a set of points $y = (y_0, y_1, \ldots y_n)$ at nodes x_i, seek a continuous function of x, with a given form characterized by m parameters $\beta = (\beta_0, \beta_1, \ldots, \beta_m)$, i.e. $f\left(x, \beta\right)$, which approximates the points while minimizing the error, defined by the sum of the squares

$$E = \sum_{i=0}^{n} \left(y - f\left(x_i, \beta\right)\right)^2 .$$

The minimum is found when

$$\frac{\partial E}{\partial \beta_j} = 0 \quad \text{for all} \quad j = 1, \ldots m$$

i.e.

$$-2\sum_{i=0}^{n}\left(y_i - f\left(x_i, \beta_j\right)\right)\frac{\partial f\left(x_i, \beta\right)}{\partial \beta_j} = 0 \quad \text{for all} \quad j = 1, \ldots m.$$

Definition 6.8 (Linear Least-Squares Approximation)**.** If the function f is a function of the form

$$y = \sum_{j=1}^{m} \beta_j \varphi_j\left(x\right)$$

then the least squares problem can be expressed as

$$\frac{\partial E}{\partial \beta_j} = \sum_{j=1}^{m}\left(\sum_{i=1}^{n} \varphi_j\left(x_i\right)\varphi_k\left(x_i\right)\right).$$

Thus, the weights β can be determined by solving the linear system,

$$\Phi\Phi^T\beta = \Phi y,$$

i.e. $\beta = \left(\Phi\Phi^T\right)^{-1}\Phi y$, where Φ is the collocation matrix.