# Chapter 6

# Numerical Ordinary Differential Equations - Boundary Value Problems

## 6.1 Ordinary Shooting Method — An Example

Consider a second-order linear 2-point boundary value problem (BVP)

$$\begin{aligned} -z'' + p(x)z' + q(x)z &= r(x) && (6.1)\\ z(a) &= \alpha && (6.2)\\ z(b) &= \beta && (6.3) \end{aligned}$$

where $p(x), q(x)$ and $r(x)$ are given. By defining $y(x) := [z(x), z'(x)]^T$, the problem can be changed into a first-order differential system

$$\begin{aligned} y' &= \begin{bmatrix} 0 & 1 \\ q(x) & p(x) \end{bmatrix} y + \begin{bmatrix} 0 \\ -r(x) \end{bmatrix} && (6.4)\\ y_1(a) - \alpha &= 0 && (6.5)\\ y_2(b) - \beta &= 0. && (6.6) \end{aligned}$$

**Remark.** In general, a linear 2-point BVP can be written as

$$\begin{aligned} y' &= A(x)y + \Phi(x) && (6.7)\\ g(y(a), y(b)) &= 0 && (6.8) \end{aligned}$$

where $y, \Phi$ and $g$ are $n$-dimensional vectors and $A(x) \in R^{n\times n}$.

Consider the following IVP associated with (6.4),

$$\begin{aligned} u' &= \begin{bmatrix} 0 & 1 \\ q(x) & p(x) \end{bmatrix} u + \begin{bmatrix} 0 \\ -r(x) \end{bmatrix} && (6.9)\\ u(a) &= \begin{bmatrix} \alpha \\ s \end{bmatrix}. && (6.10) \end{aligned}$$

Denote the corresponding solution of (6.9) as

$$u(x;s) = \begin{bmatrix} u_1(x;s) \\ u_2(x;s) \end{bmatrix}. \tag{6.11}$$

By a *shooting method* we mean to find an appropriate value of $s$ so that

$$G(s) := u(b;x) - \beta = 0. \tag{6.12}$$

Thus, a shooting method reduces a BVP into the problem of solving a nonlinear equation (6.12). Any variation of the Newton method, e.g.,

$$s_{n+1} = s_n - (G'(s_n))^{-1} G(s_n) \tag{6.13}$$

can be used to hlep to accomplish this goal. To carry out the shooting method, several additional numerical procedures are required:

1. One needs an initial guess on $s_0$.
2. One needs to numerically integrate (6.9) with initial condition $[\alpha, s_0]^T$ to the point $x = b$ to obtain the value of $G(s_0)$ as is defined in (6.12).
3. The derivative $G'(s_0)$ can be obtained from the difference approximation

$$G'(x(s_0)) \approx \frac{G(s_0 + \Delta s) - G(s_0)}{\Delta s}, \tag{6.14}$$

or from the variation equation associated with (6.9), i.e., if

$$\xi(x;s) := \frac{\partial u(x;s)}{\partial s}, \tag{6.15}$$

then

$$\begin{aligned} \frac{d\xi(x;s)}{dx} &= A(x)\xi(x;s) && (6.16) \\ \xi(a;s) &= \begin{bmatrix} 0 \\ 1 \end{bmatrix} && (6.17) \end{aligned}$$

and

$$G'(s) = \xi_1(b). \tag{6.18}$$

4. Apply one Newton step such as (6.13) to advance to $s_1$.
5. Repeat steps 2 to 4 with $s_0$ being replaced by the new $s_1$ until convergence.

**Example.** Consider the singular perturbed problem

$$\begin{aligned} \epsilon z'' + (1+\epsilon)z' + z &= 0 && (6.19) \\ z(0) &= 0 && (6.20) \\ z(1) &= 1 && (6.21) \end{aligned}$$

with $\epsilon \neq 0$. The exact solution is given by

$$z(x) = \begin{cases} \frac{e^{-x}-e^{-x/\epsilon}}{e^{-1}-e^{-1/\epsilon}} & \text{if } \epsilon \neq 1 \\ exe^{-x} & \text{if } \epsilon = 1. \end{cases} \tag{6.22}$$

The associated IVP is

$$u' = \begin{bmatrix} 0 & 1 \\ -\frac{1}{\epsilon} & -\frac{1+\epsilon}{\epsilon} \end{bmatrix}, \tag{6.23}$$

and the exact solution $u(x;s)$ for $\epsilon \neq 1$ is given by

$$u(x;s) = \begin{bmatrix} \frac{s}{1-1/\epsilon}(e^{-x/\epsilon} - e^{-x}) \\ \frac{s}{1-1/\epsilon}(-\frac{e^{-x/\epsilon}}{\epsilon} + e^{-x}) \end{bmatrix}. \tag{6.24}$$

So the desired $s$ is the the zero of

$$G(s) = \frac{s}{1-1/\epsilon}(e^{-1/\epsilon} - e^{-1}) - 1. \tag{6.25}$$

Denote $\delta(\epsilon) := \frac{1}{s}$. We see that $\lim_{\epsilon \to 0^-} = \infty$. Thus, for $-1 << \epsilon < 0$, the right endpoint is very sensitive to the variation of $s$.

## 6.2 Multiple Shooting Methods — The Set-up

Consider the problem

$$\begin{aligned} y' &= f(x,y) & (6.26) \\ Ay(a) + By(b) &= \alpha & (6.27) \end{aligned}$$

where $x \in [a,b]$, $A, B \in R^{n \times n}$, $\text{rank}[A,B] = n$, and $y, f, \alpha \in R^n$. Let the interval $[a,b]$ be partitioned into $a = x_0 < x_1 < \ldots < x_m = b$. Denote $y_j(x) := y(x; x_{j-1}, s_{j-1})$ as the solution of the IVP

$$\begin{aligned} y' &= f(x,y) & (6.28) \\ y(x_{j-1}) &= s_{j-1} & (6.29) \end{aligned}$$

in the interval $[x_{j-1}, x_j]$ for $j = 1, \ldots, m$. The idea of multiple shooting is to determine the $n \times m$ matrix

$$S := [s_0, s_1, \ldots, s_{m-1}] \tag{6.30}$$

so that the following conditions are satisfied.

- (Continuity condition)

$$y(x_j; x_{j-1}, s_{j-1}) = s_j, \text{ for } j = 1, \ldots m-1 \tag{6.31}$$

- (Boundary condition)

$$As_0 + By(b; x_{m-1}, s_{j-1}) = \alpha. \tag{6.32}$$

Let $\Delta_j = x_j - x_{j-1}$. Introduce the new variable

$$\tau := \frac{x - x_{j-1}}{\Delta_j} \tag{6.33}$$

and define

$$\hat{y}_j(\tau) := y_j(x_{j-1} + \tau\Delta_j) \tag{6.34}$$

over the interval $[x_{j-1}, x_j]$. Then (6.28) can be transformed into

$$\begin{aligned} \frac{d\hat{y}_j(\tau)}{d\tau} &= \Delta_j f(x_{j-1} + \tau\Delta_j, \hat{y}_j(\tau)) = f_j(\tau, \hat{y}_j) && (6.35)\\ \hat{y}_j(0) &= s_{j-1}. && (6.36) \end{aligned}$$

By this scaling, (6.31) and (6.32) become

$$\begin{aligned} \hat{y}_j(1) &= \hat{y}_{j+1}(0) \text{ for } j = 1, \dots m-1 && (6.37)\\ A\hat{y}_1(0) + B\hat{y}_m(1) &= \alpha. && (6.38) \end{aligned}$$

Let

$$\begin{aligned} Y(\tau) &:= [\hat{y}_1(\tau), \dots, \hat{y}_m(\tau)]^T\\ F(\tau, Y) &:= [f_1(\tau, \hat{y}_1), \dots, f_m(\tau, \hat{y}_m)]^T\\ \beta &:= [\alpha, 0, \dots 0]^T. \end{aligned}$$

Then the application of the multiple shooting method to the BVP (6.26) can be represented by a new system of 2-point BVP, i.e.,

$$\begin{aligned} \frac{dY}{d\tau} &= F(\tau, Y),\ 0 \le \tau \le 1 && (6.39)\\ PY(0) + QY(1) &= \beta && (6.40) \end{aligned}$$

where

$$P := \begin{bmatrix} A & 0 & \dots & 0\\ 0 & I & & 0\\ & & \ddots & \\ 0 & & & I \end{bmatrix}$$

and

$$Q := \begin{bmatrix} 0 & 0 & \dots & & B\\ -I & 0 & & 0 & \\ & \ddots & \ddots & & \\ 0 & & & -I & 0 \end{bmatrix}.$$

**Remark.** The 2-point BVP (6.39) incorporates both the original BVP (6.26) and the multiple shooting method together. If the boundary condition (6.40) is satisfied, then the continuity condition (6.31) and (6.32) required by the multiple shooting are automatically satisfied.

To solve (6.39) we now apply the ordinary shooting method again. That is, if we denote $U(\tau; S)$ to be the solution of the IVP

$$\begin{aligned} \frac{dU}{d\tau} &= F(\tau, U) \\ U(0) &= S, \end{aligned} \tag{6.41}$$

then we look for zeros of the equation

$$G(S) := PS + QU(1; S) - \beta = 0. \tag{6.42}$$

**Remark.** It can be shown that a multiple shooting method has larger domain of convergence than an ordinary shooting method.

**Remark.** The nonlinear equation $G(S) = 0$ resulted from a BVP by shooting is usually far more challenging and difficult because

1. The equation might have multiple solutions.

2. The close form of $G(S)$ generally in not explicitly known. Neither is its Jacobian matrix $\frac{\partial G}{\partial S}$.

3. The endpoint of a BVP can be quite sensitive to the values at the initial point, and thus makes a locally convergent method difficult to pick up an appropriate initial guess.

## 6.3 Solving Nonlinear Equations — Homotopy Method

The idea of a homotopy method is as follows. Given a system of algebraic equation, we start with a particular simple system whose solution is known. We then mathematically deform this simple system into the original, more difficult system. While deforming the systems, we carefully trace the corresponding deformation of the solution. Hopefully, the solution has also deformed from the obvious solution into the solution we are looking for. This deformation process is called a homotopy method.

Suppose we want to solve the system

$$F(x) = 0 \tag{6.43}$$

where $F : R^n \longrightarrow R^n$. Let $ER^n \longrightarrow R^n$ be another system so that $E(x^0) = 0$ for some known $x^0 \in R^n$. We introduce a homotopy function

$$H : R^n \times R \longrightarrow R^n \tag{6.44}$$

such that $H(x,0)=E(x)$ and $H(x,1)=F(x)$. Consider the set

$$H^{-1}(0) := \{(x,t \in R^n \times R | H(x,t)=0\}. \tag{6.45}$$

We want to construct $H$ so that

- $H^{-1}(0)$ is a smooth path in $R^n \times R$.
- The homotopy path connects $(x^0,0)$ to $(x^*,1)$.

Some of the choices of $H$ are:

1. (Convex Homotopy)
$$H(x,t) := (1-t)E(x) + tF(x). \tag{6.46}$$
2. (Fixed-point Homotopy)
$$H(x,t) := (1-t)(x-x^0) + tF(x). \tag{6.47}$$
3. (Newton Homotopy)
$$H(x,t) := (1-t)(F(x)-F(x^0)) + tF(x) = F(x)-(1-t)F(x^0) \tag{6.48}$$

**Remark.** If can be shown with the help of differential topology that $H^{-1}(0)$ is a 1-dimensional manifold under very mild assumptions on the functions $F$ and $E$.

If the existence of a homotopy path has been established, the next issue is to following this path. Toward this end, we introduce the arc length $\sigma$ along the path as the parameter. Then the path is characterized as a solution to this differential equation

$$\frac{dH}{d\sigma} = \frac{\partial H}{\partial x}\frac{dx}{d\sigma} + \frac{\partial H}{\partial t}\frac{dt}{d\sigma} = 0. \tag{6.49}$$

More precisely, we have

$$\begin{aligned} \left[\frac{\partial H}{\partial x}, \frac{\partial H}{\partial t}\right]\left[\begin{array}{c} \frac{dx}{d\sigma} \\ \frac{dt}{d\sigma} \end{array}\right] &= 0 && (6.50) \\ \left[\begin{array}{c} x(0) \\ t(0) \end{array}\right] &= \left[\begin{array}{c} x^0 \\ 0 \end{array}\right] && (6.51) \\ \|\frac{dx}{d\sigma}\|^2 + (\frac{dt}{d\sigma})^2 &= 1. && (6.52) \end{aligned}$$

That is, the path $H^{-1}(0)$ is implicitly described by an initial value problem.

**Example.** Consider the Newton homotopy (6.48). We have

$$\frac{\partial F}{\partial x}\frac{dx}{d\sigma} + F(x^0)\frac{dt}{d\sigma} = 0. \tag{6.53}$$

Thus, we may write, if $\frac{\partial F}{\partial x}$ is invertible,

$$\dot{x} = -\frac{\dot{t}}{1-t}(\frac{\partial F}{\partial x})^{-1}F(x). \tag{6.54}$$

Obvious, with a suitable step size taken, one step of the Euler method applied to (6.54) is equivalent to the classical Newton method.

We now examine how the homotopy method should be implemented for a general 2-point BVP

$$\begin{aligned} y' &= f(x,y) \quad (6.55)\\ g(y(a), y(b)) &= 0. \quad (6.56) \end{aligned}$$

When a shooting method is applied to the BVP, we need to solve the nonlinear algebraic equation

$$g(s, u(b;s)) = 0 \tag{6.57}$$

where $u(x;s)$ is the solution to the differential equation (6.55) with initial value $u(a) = s$. Suppose we use the Newton homotopy, i.e.,

$$H(s,t) := g(s, u(b;s)) - (1-t)g(\gamma, u(b;\gamma)) = 0. \tag{6.58}$$

Then

$$\begin{aligned} \frac{\partial H}{\partial s} &= \frac{\partial g}{\partial y(a)} + \frac{\partial g}{\partial y(b)}\frac{\partial u(b;s)}{\partial s} \quad (6.59)\\ \frac{\partial H}{\partial t} &= g(\gamma, u(b;\gamma)). \quad (6.60) \end{aligned}$$

Again, the quantity $\frac{\partial u(b;s)}{\partial s}$ involved in (6.59) can be obtained from the value of $\xi(b)$ where $\xi(x)$ solves the variational equation

$$\begin{aligned} \frac{d\xi}{dx} &= \frac{\partial f(x,u)}{\partial u}\xi \quad (6.61)\\ \xi(a) &= I. \quad (6.62) \end{aligned}$$

One nice feature of the homotopy method is that if it works then it provides a globally convergent method.

## 6.4 Finite Difference Method

Consider a general 1-dimensional differential equation with the Dirichlet boundary conditions:

$$\begin{aligned} y'' &= f(x,y) \quad (6.63)\\ y(a) &= \alpha \quad (6.64)\\ y(b) &= \beta \quad (6.65) \end{aligned}$$

over the interval $[a, b]$. Let $x_i := a+ih, i = 0, 1, \ldots, n$ denote a uniformly spaced partition of $[a, b]$ with $h := (b - a)/n$. Let $y_i$ denote an approximation to $y(x_i$. Suppose we approximate the second-order derivative by the finite difference

$$y''(x_i) \approx \frac{y_{i+1} - 2y_i + y_{i-1}}{h^2}. \tag{6.66}$$

Then the differential equation over the interval can be discretized into the system

$$-y_{i-1} + 2y_i - y_{i+1} + h^2 f(x_i, y_i) = 0,\ i = 1, \ldots, n-1. \tag{6.67}$$

Together with the boundary conditions, we may rewrite the BVP in the matrix form:

$$\begin{bmatrix} 2 & -1 & 0 & \ldots & & 0 \\ -1 & 2 & -1 & & & 0 \\ & \ddots & \ddots & \ddots & & \\ & & & & & \\ & & & & 2 & -1 \\ & & & & -1 & 2 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ \\ y_{n-2} \\ y_{n-1} \end{bmatrix} + h^2 \begin{bmatrix} f(x_1, y_1) \\ f(x_2, y_2) \\ \vdots \\ \\ f(x_{n-2}, y_{n-2}) \\ f(x_{n-1}, y_{n-1}) \end{bmatrix} = \begin{bmatrix} \alpha \\ 0 \\ \vdots \\ \\ 0 \\ \beta \end{bmatrix}. \tag{6.68}$$

Depending upon the function $f(x, y)$, the system (6.68) might be linear or non-linear. Obviously, the step size $h$ must be smaller when higher accuracy is required for the solution of the BVP, which results in a larger (but often structured) algebraic system (6.68).

One may consider to replace (6.63) by a more general finite difference scheme:

$$a_k y_{n+k} + \ldots + a_0 y_n - h^2\{b_k f_{n+k} + \ldots + b_0 f_n\} = 0 \tag{6.69}$$

where $a_i, b_i$ are some suitably selected coefficients. In doing so, however, one concern is that (6.69) may involve values of $y_j$ which are not available if $k$ is too large. In practice, one usually considers only a 3-term scheme

$$-y_{i-1} + 2y_i - y_{i+1} + h^2\{b_0 f_{i-1} + b_1 f_i + b_{i+1} f_{i+1}\} = 0 \tag{6.70}$$

such as

$$\begin{array}{lll} -y_{i-1} + 2y_i - y_{i+1} & +h^2 f_i & = 0 \text{ (order} = 2) \qquad (6.71) \\ -y_{i-1} + 2y_i - y_{i+1} & +\frac{h^2}{\{} f_{i-1} + 10 f_i + f_{i+1}\} & = 0 \text{ (order} = 4). \qquad (6.72) \end{array}$$

Similar to (6.68), (6.70) can be rewritten in the form

$$JY + h^2 BF(Y) = A \tag{6.73}$$

where $J$ is the same tridiagonal matrix as in (6.68),

$$B := \begin{bmatrix} b_1 & b_2 & 0 & \ldots & & 0 \\ b_0 & b_1 & b_2 & & & 0 \\ & \ddots & \ddots & \ddots & & \\ & & & & & \\ & & & & b_1 & b_2 \\ & & & & b_0 & b_1 \end{bmatrix},$$

$Y := [y_1, \ldots, y_{n-1}]^T$, $F(Y) := [f(x_1, y_1), \ldots, f(x_{n-1}, y_{n-1})]^T$, and $A = [\alpha - b_0h^2f(x_0, \alpha), 0, \ldots, 0, \beta - b_2h^2f(x_n, \beta)]^T$. To solve (6.73) by the Newton method, one needs to calculate the Jacobian of the left-hand side of (6.73) which is easily seen to be

$$J + h^2B\text{diag}\{f'(x_1, y_1), \ldots, f'(x_{n-1}, y_{n-1})\} \tag{6.74}$$

where $f'$ means $\frac{\partial f}{\partial y}$. Thus the resulting linear system is a tridiagonal matrix which can be solved much easier than expected.

The above discussion can be generalized to system of differential equations and can be customized for more general boundary conditions.

## 6.5 Finite Element Methods

In this section we shall explore the basic ideas of the finite element method by studying its application to the 1-dimensional linear BVP:

$$\begin{aligned} -u'' &= f(x) - ku(x) && (6.75)\\ u(0) &= 0 && (6.76)\\ u(1) &= 1 && (6.77) \end{aligned}$$

where $u(x)$ stands for the transverse deflection of a taut string, fixed at its ends, under an applied transverse, distributed load $f(x)$. The total energy at the current deflected state is half as much as

$$J(u) := \int_0^1 (u')^2dx + \int_0^1 ku^2dx - 2\int_0^1 fudx. \tag{6.78}$$

The first term represents the strain energy of the string ($= \frac{(u')^2}{2}$.) The second term represents the energy stored in the string ($= -\int_0^u(-ky)dy = \frac{ku^2}{2}$.) The third term represents the work done by the load $f$ to take string from its original configuration to its current deflected state ($\int_0^1 fudx$.) As a rule of the nature (The Principle of Least Action,) the minimizer of $J$ should be the solution to the BVP (6.75). Indeed, we can prove the following theorem.

**Theorem 6.5.1** *Let $D := \{u \in C^2(0,1) | u(0) = u(1) = 0\}$ denote the set of admissible functions. Then*

1. *Let $u$ be a solution of the BVP (6.75). Then $J(u) \leq J(v)$ for all $v \in D$.*

2. *Let $u$ minimize $J$ among all $v \in D$. Then $u$ is a classical solution to the BVP (6.75).*

(pf:) Define the operator

$$Au := -u'' + ku. \tag{6.79}$$

Then for any $v \in D$, we have

$$< Av, v >= \int_0^1 (-v'' + kv)vdx = \int_0^1 [(v')^2 + kv^2]dx. \tag{6.80}$$

Hence we may rewrite

$$J(v) =< Av, v > -2 < f, u > . \tag{6.81}$$

For any given two elements $v, w \in D$, define $h := v - w$. Then we have

$$\begin{aligned} J(v) - j(w) &= < A(w+h), w+h > - < Aw, w > -2 < f, h > \\ &= 2 < Aw - f, h > + < Ah, h > . \end{aligned} \tag{6.82}$$

Suppose $w$ is a solution of (6.75). Then $< Aw - f, h >= 0$. It is clear from (6.80) that $< Ah, h >\geq 0$. We thus conclude that $J(v) \geq J(w)$.

Suppose now that $w$ is a minimizer of $J$. Then from (6.82) we find that is is necessary to have $2 < Aw - f, h > + < Ah, h >\geq 0$ for every $h \in D$. Choose $h$ to be of the form $h = \epsilon\eta$ for some $\eta \in D$. Then we observe that

$$2 < Aw - f, \eta > +\epsilon < A\eta, \eta >\geq 0 \quad \text{for } \epsilon > 0 \tag{6.83}$$

$$2 < Aw - f, \eta > +\epsilon < A\eta, \eta >\leq 0 \quad \text{for } \epsilon < 0. \tag{6.84}$$

This is possible only if $< Aw - f, \eta >= 0$ for every $\eta \in D$. It follows that $Aw = f$.

**Remark.** The energy formulation (6.78) works only for the problem (6.75). For general problem, it is not always possible how the energy integral should be set up. In many cases, however, one may consider the so called *weak formulation* directly without referring to tany physical meaning at all. In the above it turns out that the energy formulation is equivalent to the weak formulation.

Motivated by the above discussion, for a general differential equation

$$Au = f \tag{6.85}$$

where $A$ stands for some linear differential operator, we attempt to find a solution $u$ so that the equation

$$< Au, v >=< f, v > \tag{6.86}$$

is true for all $v$ from a certain set $D$ of functions. The idea of a finite element method is to limit $v$ to a finite-dimensional subspace, say, $D = \text{span}\{e_1, \ldots, e_n\}$, and to consider a solution $u^* \in D$ of (6.86) as an approximaton to the true solution $u$. Write

$$u^* := \sum_{j=1}^{n} c_j e_j \tag{6.87}$$

for a certain suitable coefficients $c_j$. Then (6.86) suggests that $c_j$ can be determined from the system

$$\sum_{j=1}^{n} < Ae_j, e_i >=< f, e_i >, \; i = 1, \ldots, n. \tag{6.88}$$

The equation (6.88) is known as the Galerkin equation. A finite element method, therefore, involves

1. The section of basis elements $\{e_j\}$ and the characterization of the admissible space $D$.

2. The calculation of $< Ae_j, e_i >$ and $< f, e_i >$, which usually involves numerical integrations. Such a process is called the assembling.

3. The solution of a large, sparse, and usually structured linear algebraic equation.