Boundary Value Problem

In fact, IVP is special case of BVP:

$\underline{y}'(t) = \underline{f}(t, \underline{y}(t))$, $t \in J = [a,b]$

$\underline{r}(\underline{y}(a), \underline{y}(b)) = 0$ with $f: I \times \mathbb{R}^d \to \mathbb{R}^d$, and

$\underline{y}: J \to \mathbb{R}^d$, $\underline{r}: \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$.

<u>Rmk</u>: i) Often we have linear boundary condition:

$B_a \underline{y}(a) + B_b \underline{y}(b) = \underline{g}$, $B_a, B_b \in \mathbb{R}^{d \times d}$, $\underline{g} \in \mathbb{R}^d$.

$\Rightarrow$ IVP take $B_a = I$, $B_b = 0$, $\underline{g} = \underline{y}_0$.

ii) We still need enough boundary condition to guarantee unique/existence.

iii) It's more difficult to solve BVP since now we may not be given initial value to start the algorithm.

(1) <u>First-order linear BVPs</u>:

Consider first-order linear inhomogeneous BVP:

$\underline{y}'(t) - A(t)\underline{y}(t) = \underline{f}(t)$, $t \in J = [a,b]$.

$B_a \underline{y}(a) + B_b \underline{y}(b) = \underline{\gamma}$, with $B_a, B_b \in \mathbb{R}^{n\times n}$, $\underline{\gamma} \in \mathbb{R}^n$

$A \in C(J, \mathbb{R}^{n\times n})$, $\underline{f} \in C(J, \mathbb{R}^n)$.

Next, we want to connect BVP to IVP:

First introduce $n+1$-dim IVPs:

$\underline{u}_0'(t) - A(t)\underline{u}_0(t) = \underline{f}(t)$, $\underline{u}_0(a) = \underline{0}$ $\quad t \in J$.

$\underline{u}_i'(t) - A(t)\underline{u}_i(t) = \underline{0}$, $\underline{u}_i(a) = \underline{e}_i$, $t \in J$, $1 \le i \le n$.

where $\underline{e}_i = (0, \cdots 0, \underset{i\text{th}}{1}, 0 \cdots 0)$.

Rmk: $\{\underline{e}_i\}$ can span $\mathbb{R}^n$. So that's why we

chose them to construct arbitrary one.

Def: Fundamental matrix $\phi(t) = (\underline{u}_1(t) \cdots \underline{u}_n(t))$

Rmk: $\phi(a) = I_n$.

For $\underline{s} \in \mathbb{R}^n$: $\underline{y}(t; \underline{s}) := \underline{u}_0(t) + \sum_1^n s_i \underline{u}_i(t)$.

$= \underline{u}_0(t) + \phi(t) \cdot \underline{s}$.

We check $\underline{y}(t;\underline{s})$ satisfies equation:

$\underline{y}'(t;\underline{s}) = \underline{f}(t) + A(t)\underline{y}(t;\underline{s})$.

So next we only need to find $\underline{s} \in \mathbb{R}^n$ to satisfies boundary conditions. i.e. we require:

$B_a \underline{y}(a; \underline{s}) + B_b \underline{y}(b; \underline{s}) = \underline{g}$

Rmk: $\underline{s}$ may not be solvable every time.

Qus: Given linear boundary cond. when it is possible to find $\underline{s}$ satisfy it?

Using $\underline{y}(t, \underline{s}) = \underline{u}_0(t) + \phi(t)\underline{s}$. We have:

$B_a(\overset{=0}{\underline{u}_0(a)} + \overset{=I}{\phi(a)}\underline{s}) + B_b(\underline{u}_0(b) + \phi(b)\underline{s}) = \underline{g}$

i.e. $(B_a + B_b\phi(b))\underline{s} = \underline{g} - B_b \underline{u}_0(b)$ (*)

Thm. The linear BVP has a unique solution $y(t)$ for any given $\underline{f}(t)$ and $\underline{g}$ $\Leftrightarrow$ the matrix $B_a + B_b\phi(b) \in \mathbb{R}^{n \times n}$ is invertible/regular.

Pf: For $A(t)$ is conti. $\Rightarrow$ $\phi(t)$ is invertible.

(It can be proved, and intuitively: $\phi(a) = I$, $|\phi(t)|$ is conti.)

$\Rightarrow$ Columns of $\phi(t)$ span $\mathbb{R}^n$.

$\Rightarrow$ Any s.l. to BVP can be written in $\underline{y}(t) = \underline{u}_0(t) + \phi(t)\underline{s}$. for some $\underline{s} \in \mathbb{R}^n$.

So we can still derive (*).

($\Leftarrow$) is trivial from solving $\underline{s}$ above.

($\Rightarrow$) Since it has a unique solution.

$\Rightarrow$ (*) must have unique sol. $\tilde{s}$.

So: $B_a + B_b \phi(b)$ is invertible

Rmk: We first solve $\{v_i(t)\}_0^k$. Then: we get $\phi(b)$. So we can find what B_a, B_b will imply BVP can be solved.

(2) <u>Single Shooting method:</u>

We summarize and improve the method from (1) to get single shooting algorithm:

i) Solve the $k+1$ system of IVP with N time steps and final time $T=b$.

Using one of method (RK, LMM...) to get approxi. $\underline{u}_{0,N}$ and ϕ_N for IVP to $t_N=b$.

ii) Set up $B_a + B_b \phi_N(t_N)$. If it's regular

$\Rightarrow$ solve $\hat{\underline{s}}_N = (B_a + B_b \phi_N(t_N))^{-1}(\underline{q} - B_b \underline{u}_{0,N}(t_N))$

iii) Discrete solution is:

a) $\underline{y}_n = \underline{u}_{0,N}(t_n) + \phi_N(t_n)\,\hat{\underline{s}}_N$. $n = 0, \dots, N$.

or b) Solve IVP:

$\underline{y}'(t) - A(t)\underline{y}(t) = \underline{f}(t)$. $t \in J = [a, b]$. $\underline{y}(a) = \hat{\underline{s}}_N$

Rmk: b) is from superposition property of sol. for linear IVP: $y(t; a) + y(t; b) = y(t; a+b)$. So the sol. of IVP b) is still $\hat{\underline{s}}_N$ - linear combination.

($y(t, \hat{s}_N)$ also satisfies the IVP.)

Thm. For suff. small step length $h > 0$. If $B_a + B_b\,\phi_N(t_N)$ is regular and we use the method of order m to compute $\underline{u}_{0,N}$. ϕ_N.

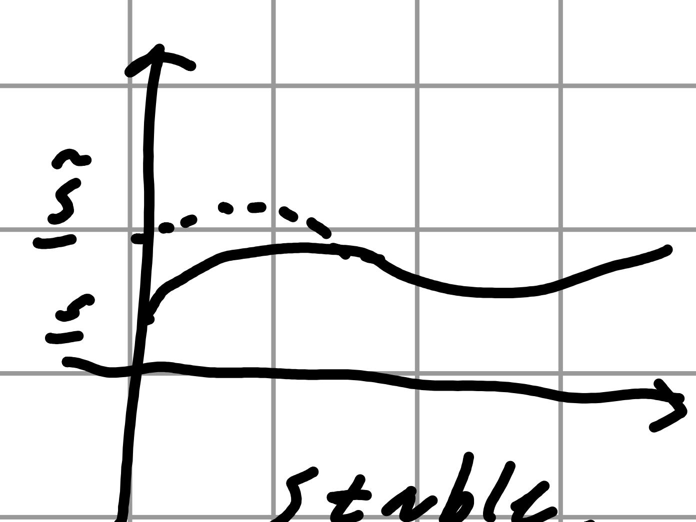
Then: $\max\limits_{1 \le n \le N} \|\underline{y}_n - \underline{y}(t_n)\| = O(h^m)\ (h \to 0)$.

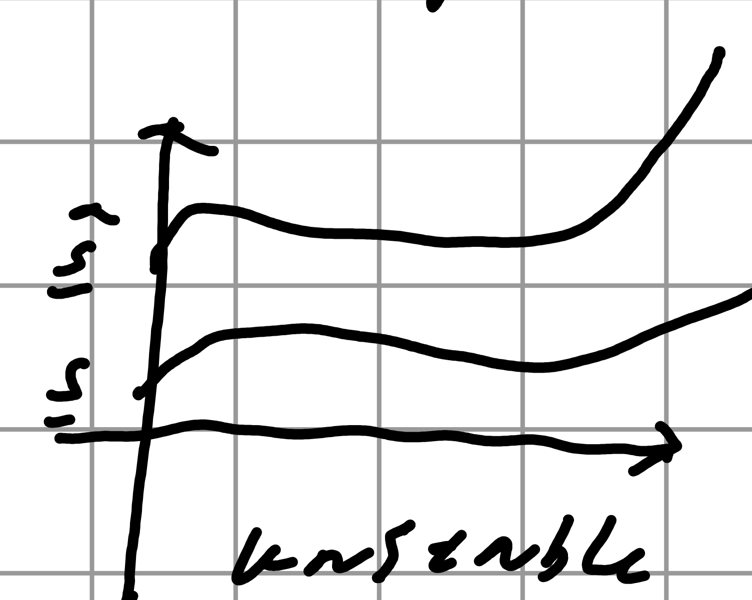
Rmk: (The name of single shooting)

Note step i) in algo. can be seen as solving the IVP with starting value $\underline{s}$ and use it to compute correct $\hat{s}$. And

We only shoot it once and get one trajectory (The whole algorithm can be seen as find correct initial conditions that is from boundary conditions.)

The main issue of the method is the potential lack of stability for integrating over long interval $[a,b]$. ($y(b;s)$ at end pt. will be very sensitive to perturb. at $t=a$)

a) 

b) 

<u>Rmk</u>: This kind of stability is property of the sol. of ODE, not of the integration method we use.

<u>e.g.</u> $y_1'(t) = y_2(t)$, $y_2'(t) = 110 y_1(t) + y_2(t)$.

$$\Rightarrow \underline{y}(t) = c_1 e^{-10t} \begin{pmatrix} 1 \\ -10 \end{pmatrix} + c_2 e^{11t} \begin{pmatrix} 1 \\ 11 \end{pmatrix}$$

For initial value $\underline{y}(0) = (s_1, s_2)^T$.

$$\Rightarrow c_1 = \frac{11 s_1 - s_2}{21}, \quad c_2 = \frac{10 s_1 + s_2}{21}$$

For BVP: $y_1(0) = 1$, $y_1(10) = 1$.

We solve $\underline{y}(t) =$

$$\frac{e^{110}-1}{e^{110}-e^{-100}} e^{-10t} \begin{pmatrix} 1 \\ -10 \end{pmatrix} + \frac{1-e^{-100}}{e^{110}-e^{-100}} e^{11t} \begin{pmatrix} 1 \\ 11 \end{pmatrix}.$$

$s_0 = y_2(0) \approx -10 + 3.5 \times 10^{-47}$

We're only able to set $\hat{\underline{s}} = (y_1(0),$ $y_2(0))^T \approx (1, -10+10^{-7})^T$ because of the limit of machine accuracy. But then $y_1(10, \hat{\underline{s}}) \sim 10^{37} >> 1$ which blows up.

<u>Rmk</u>: In the example, there's an exponential growth of the slight perturb. at initial value $\underline{s}$:

$$\|\underline{y}(t, \hat{\underline{s}}) - \underline{y}(t, \underline{s})\| = O(e^{11t}) \|\underline{s} - \hat{\underline{s}}\|$$

It can be deduced from:

$\underline{y}'(t) = \begin{pmatrix} 0 & 1 \\ 110 & 1 \end{pmatrix} \underline{y}(t)$ has Lip. const. $L \sim 11$. So by the stability Thm:

$$\|\underline{y}(t, \tilde{\underline{s}}) - \underline{y}(t, \underline{s})\| \leq e^{L|t-a|} \|\tilde{\underline{s}} - \underline{s}\|$$

So the problem is from the ODE but not the scheme.

(3) Multiple Shooting Method:

Idea: Reduce the time interval and so to reduce $e^{L|t-a|}$.

Split $[a, b]$ into $a = t_1 < \cdots < t_k < \cdots t_{K+1} = b$.

s.t. $e^{L(t_{k+1}-t_k)}$ isn't large. Then:

1) Apply shooting in each $[t_k, t_{k+1}]$

(But we need to know where to shoot)

2) Require continuity at each t_k to obtain the global solution.

Given $\underline{\sigma}_k \in \mathbb{R}^n$, $k = 1, \cdots K$. Let $\underline{y}(t, t_k, \underline{\sigma}_k)$ is

Sol. of $\underline{y}'(t) - A(t)\underline{y}(t) = \underline{f}(t)$. $\underline{y}(t_k) = \underline{\sigma}_k$.

(i.e. we preassume boundary value at $[t_k, t_{k+1}]$

and $\underline{\sigma}_k$ has role of $\underline{s}$ above)

Goal: Determine $\underline{\sigma}_k$, $k = 1, \cdots K$, s.t.

i) The combined Sol. $\underline{y}(t) : [a, b] \to \mathbb{R}^n$

given by $\underline{y}(t) = \underline{y}(t, t_k, \underline{\sigma}_k)$, $t \in [t_k, t_{k+1}]$ is conti. on $[a, b]$.

ii) It satisfies the boundary condition:

$$B_a \underline{y}(a) + B_b \underline{y}(b) = \underline{\gamma}.$$

Rmk: For $t \in [t_k, t_{k+1}]$. We have:

$$\underline{y}(t) \overset{TS}{=} \underline{y}(t) + \sum_{1}^{k-1} (\underline{y}(t_{i+1}) - \underline{y}(t_i)) + \underline{y}(a) - \underline{y}(t_k)$$

$$= \int_{t_k}^{t} \underline{y}'(t) + \sum_{1}^{k-1} \int_{t_i}^{t_{i+1}} \underline{y}'(t) + \underline{y}(a)$$

$$\overset{BVP}{=} \int_a^t (A(s)\underline{y}(s) + \underline{f}(s))\,ds + \underline{y}(a)$$

$\Rightarrow$ It $\in C'[a,b]$ actually

So require: $\underline{y}(t_{k+1}, t_k, \underline{\sigma}_k) = \underline{\sigma}_{k+1}$. $k = 1, \cdots R-1$.

$$B_a \underline{\sigma}_1 + B_b \underline{y}(b, t_R, \underline{\sigma}_R) = \underline{\gamma}.$$

Let $\underline{\eta}_k(t)$, $\Phi_k(t)$ are unique sol. of:

$\underline{\eta}_k'(t) = A(t)\underline{\eta}_k(t) + f(t)$. $\eta_k(t_k) = 0$. $t \in [t_k, t_{k+1}]$

$\Phi_k'(t) = A(t)\Phi_k(t)$. $\Phi_k(t_k) = I$. $t \in [t_k, t_{k+1}]$.

where Φ_k is fundamental matrix on $[t_k, t_{k+1}]$.

With $\underline{y}(t, t_k, \underline{\sigma}_k) = \underline{\eta}_k(t) + \Phi_k(t)\underline{\sigma}_k$. $k = 1, \cdots R$.

So we can rewrite the two conditions by plugging into $\underline{y}(t, t_k, \underline{\sigma}_k)$:

$B_a \underline{\sigma}_1 + B_b \Phi_R(b) \underline{\sigma}_R = \underline{q} - B_b \underline{\eta}_R(b)$

$-\Phi_k(t_{k+1}) \underline{\sigma}_k + \underline{\sigma}_{k+1} = \underline{\eta}_k(t_{k+1}), \quad k = 1 \cdots R-1.$

Determine $\underline{\sigma} = (\underline{\sigma}_1 \cdots \underline{\sigma}_R)^T$ as the sol. of :

$A_R \underline{\sigma} = \underline{rhs} = (\underline{q} - B_b \underline{\eta}_R(b), \underline{\eta}_1(t_2) \cdots \underline{\eta}_{R-1}(t_R))^T$

where $A_R = \begin{pmatrix} B_a & & & B_b \Phi_R(b) \\ -\Phi_1(t_2) & I & & \\ & \ddots & \ddots & \\ & & -\Phi_{R-1}(t_R) & I \end{pmatrix} \in \mathbb{R}^{nR \times nR}$

But is the matrix A_R invertible ?

Note that $A_R = KL \stackrel{\Delta}{=} \begin{pmatrix} \alpha_1 & \cdots & \alpha_R \\ & I & \\ & & \ddots \\ & & & I \end{pmatrix} \begin{pmatrix} I & & & \\ -\Phi_1(t_2) & I & & \\ & \ddots & \ddots & \\ & & -\Phi_{R-1}(t_R) & I \end{pmatrix}$

where $\alpha_i \in \mathbb{R}^{n \times n}$:

$\alpha_R = B_b \Phi_R(b).$

$\alpha_k = \alpha_{k+1} \Phi_k(t_{k+1}), \quad k = R-1, \cdots 2.$

$\alpha_1 = B_a + \alpha_2 \Phi_1(t_2) = B_a + \alpha_3 \Phi_2(t_3) \Phi_1(t_2)$

$= \cdots = B_a + B_b \Phi_R(b) \Phi_{R-1}(t_R) \cdots \Phi_1(t_2).$

$\Rightarrow$ We need to check when α_1 is regular.

<u>Lem.</u> Let $\alpha = B_a + B_b \Phi(b)$ matrix in single shooting $\Rightarrow \alpha_1 = \alpha$

Pf: For single shooting: $\phi'(t) = A(t)\phi(t)$.

multiple: $\phi_k'(t) = A(t)\phi_k(t)$ on $[t_k, t_{k+1}]$

and $\phi(a) = I$. $\phi_k(t_k) = I$

Note for $\underline{y}'(t) = A(t)\underline{y}(t)$. $y(a) = \underline{q}$

its sol. is $\underline{y}(t) = \phi(t)\underline{q}$. And $\phi_1(t)\underline{q}$

also satisfies this equation on $[t_1, t_2]$.

By uniqueness. $\underset{t=t_2}{\overset{\text{let}}{\Rightarrow}}$ $\phi(t_2)\underline{q} = \phi_1(t_2)\underline{q}$

consider $\underline{y}'(t) = A(t)\underline{y}(t)$. $y(t_2) = \phi(t_2)\underline{q}$.

Similarly. $\phi_2(t_3)\phi_1(t_2)\underline{q} = \phi(t_3)\underline{q}$.

$\Rightarrow$ By recursion: $\prod_1^k \phi_k(t_{k+1})\underline{q} = \phi(t_{k+1})\underline{q}$

Rmk: BVP has unique sol. $\Leftrightarrow$ $|Q_1| \neq 0$.

Note that we can also use the decompose

$A_R = UL$ to solve $\underline{\delta}$:

i) Solve $\underline{z}$ from $U\underline{z} = \underline{rhs}$ ii) Solve $L\underline{\delta} = \underline{z}$.

Rmk: The problem is: Q_1 is often ill-conditioned

$\Rightarrow$ better use Gauss elimination with pivoting in A_R.

Algorithm:

i) For $k = 1, \dots, R$. Solve:

$\underline{y}_k'(t) = A(t)\underline{y}_k(t) + \underline{f}(t)$. $\underline{y}_k(t_k) = 0$. $t \in [t_k, t_{k+1}]$.

$Q_k' = A(t) Q_k(t)$. $Q_k(t_k) = I$. $t \in [t_k, t_{k+1}]$.

$\Rightarrow$ obtain $\underline{y}_{k,N_k}$. Q_{k,N_k} are values of sol. and Q at ending pt on each $[t_k, t_{k+1}]$.

ii) Set up discrete version of: $\hat{A}_R \underline{\hat{\sigma}} = \underline{rhs}^{\wedge}$.

Solve $\hat{\sigma}$ from it.

iii) Sol. is given by $\underline{y}_n(t_n) = \underline{y}_{k,n} + Q_{k,n} \underline{\hat{\sigma}}_k$ where $t_n \in [t_k, t_{k+1}]$. $k = 1, \dots, R$.

Thm If IVPs in step i) are solved by one method of order m. Then:

$\max_n \| \underline{y}_n - \underline{y}(t_n) \| = O(h^m)$ $(h \to 0)$ for sol. of multiple shooting algorithm.

(4) Shooting for nonlinear BVPs:

Consider: $\underline{y}'(t) = f(t, y(t))$. $t \in [a, b] = I$.

with : $\underline{r}(\underline{y}(a), \underline{y}(b)) = 0$.

And assume it has locally unique solution.

① Single Shooting:

Consider $\underline{y}'(t) = \underline{f}(t, \underline{y}(t))$. $\underline{y}(a) = \underline{s}$. $t \in I$. $\Rightarrow$

We can solve $\underline{y}(t)$.

Next, we want to determine $\hat{\underline{s}}$. s.t.

$\underline{r}(\underline{y}(a, \hat{\underline{s}}), \underline{y}(b, \hat{\underline{s}})) = 0$ is satisfied.

i.e. We're looking for a root of $\underline{F}: \mathbb{R}^d \to \mathbb{R}^d$. $\underline{F}(\underline{s}) = \underline{r}(\underline{s}, \underline{y}(b, \underline{s}))$

Rmk: We need $\underline{y}(b, \underline{s})$, sol. for the IVP to compute $\underline{F}(\underline{s})$.

Next, we want to use Newton's to find roots of $\underline{F}(\underline{s})$. So we need to compute its Jacobian $J(\underline{s}) = \nabla_1 \underline{r}(\underline{s}, \underline{y}(b, \underline{s})) + \nabla_2 \underline{r}(\underline{s}, \underline{y}(b, \underline{s})) \cdot \partial \underline{y}(b, \underline{s}) / \partial \underline{s}$.

Rmk: We also need to know the deri. of

$\underline{y}(b, \underline{s})$ w.r.t initial data $\underline{s}$.

Thm. Consider IVP $\underline{y}'(t) = \underline{f}(t, \underline{y}(t))$, $\underline{y}(t_0) = \underline{y}_0$, $t \in [t_0, t_0+T]$. If it satisfies Picard-Lindelöf condition and $\nabla_{\underline{x}} \underline{f}(t, \underline{x})$ is conti. Then: the unique sol. of IVP depends conti. differentiable on initial value $\underline{y}_0 = (y_{0,1} \cdots y_{0,d})$. Besides the derivative matrix $G(t) = (\partial y_i(t) / \partial y_{0,j})_{d\times d}$ satisfies:

$dG(t)/dt = \nabla_{\underline{x}} \underline{f}(t, \underline{y}(t)) G(t)$, $G(t_0) = I_d$

Rmk: i) $G(t)$ is called Wronski / Sensitivity matrix. (How datum impact for $t \uparrow$)

ii) So we have $J(\underline{s}) = \nabla_1 \underline{r}(\underline{s}, \underline{y}(b, \underline{s})) + \nabla_2 \underline{r}(\underline{s}, \underline{y}(b, \underline{s})) G(b, \underline{s})$

e.g. Consider linear BVP on $I = [a, b]$:

$\underline{y}'(t) - A(t)\underline{y}(t) = \hat{\underline{f}}(t)$. $B_a \underline{y}(a) + B_b \underline{y}(b) = \underline{g}$.

$\nabla_{\underline{x}} f(t, \underline{y}(t)) = \nabla_{\underline{x}} (A(t)\underline{y}(t) + \hat{f}(t)) = A(t)$.

So: $G(t)$ is sol. of $G'(t) = A(t) G(t)$.

with $G(a) = I_n$ on $[a, b]$. And so we have

$J(\underline{s}) = B_a + B_b G(b)$.

<u>Rmk</u>: So the sensitivity matrix is exactly the fundamental matrix when we consider the linear case.

<u>Algorithm</u>.

For step $\underline{s}^{(i)} \to \underline{s}^{(i+1)}$. (Newton's is iteration)

i) Solve IVP: $\underline{y}'(t) = \underline{f}(t, \underline{y}(t))$ $\underline{y}(a) = \underline{s}^{(i)}$

$\Rightarrow$ evaluate $\underline{F}(\underline{s}^{(i)}) = \underline{r}(\underline{s}^{(i)}, \underline{y}(b, \underline{s}^{(i)}))$.

ii) Solve $G'(t, \underline{s}^{(i)}) = \nabla_{\underline{x}} \underline{f}(t, \underline{y}(t, \underline{s}^{(i)})) G(t, \underline{s}^{(i)})$

$G(a, \underline{s}^{(i)}) = I_n$ $t \in I$.

$\Rightarrow$ evaluate $J(\underline{s}^{(i)}) = \nabla_1 \underline{r}(\underline{s}^{(i)}, \underline{y}(b, \underline{s}^{(i)}))$

$+ \nabla_2 \underline{r}(\underline{s}^{(i)}, \underline{y}(b, \underline{s}^{(i)})) G(b, \underline{s}^{(i)})$.

iii) Solve $J(\underline{s}^{(i)}) \Delta \underline{s}^{(i)} = -\underline{F}(\underline{s}^{(i)})$.

Let $\underline{s}^{(i+1)} = \underline{s}^{(i)} + \Delta \underline{s}^{(i)}$.

<u>Rmk</u>: For linear BVPs, it reduces to the algo.

in (2) Since $G(t) = g(t)$ and $J(\underline{s}) =$

$B_a + B_b G(b)$ by example above.

Note ii) require to solve for G. Which is a $k\times k$-system. So the computation can be expensive.

Since G is computed to get Jacobi J.

Alternatively, we can approxi. the deri. by using finite difference approxi.:

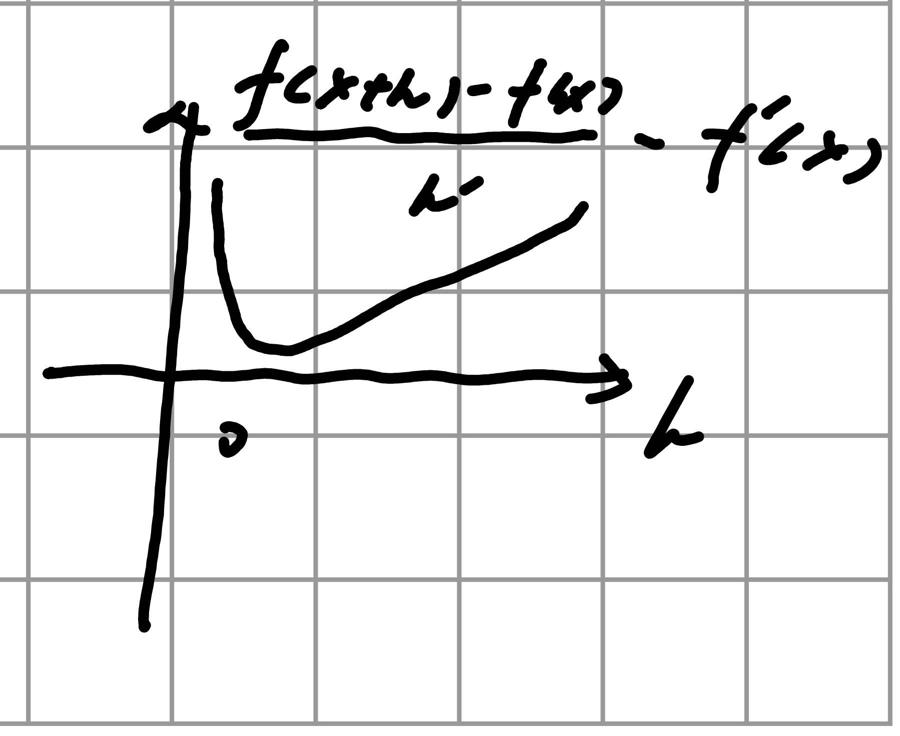
$$\left(F(s_1,\cdots s_j+\delta s_j,\cdots s_k) - F(s_1,\cdots s_k)\right)/\delta s_j \to \partial_j F(\underline{s})$$

<u>Rmk</u>: Evaluate of F at $\underline{s}$, $\underline{s}+\delta s_j$ can be done in Algo i) by computing $\underline{y}$ (The IVP system only has size k)

But how to choose δs_j?

a) δs_j need to be small enough to approxi.

b) δs_j can't be too small to avoid rounding error to dominate.



Rule of thumb:

For func evaluated with machine accuracy ε:

$\delta s_j \sim \sqrt{\varepsilon}$.

But eval. of F depends on integration accuracy 10^{-TOL}. So we use $\delta s_j \sim \sqrt{10^{-TOL}}$.

Advantage of single shooting:

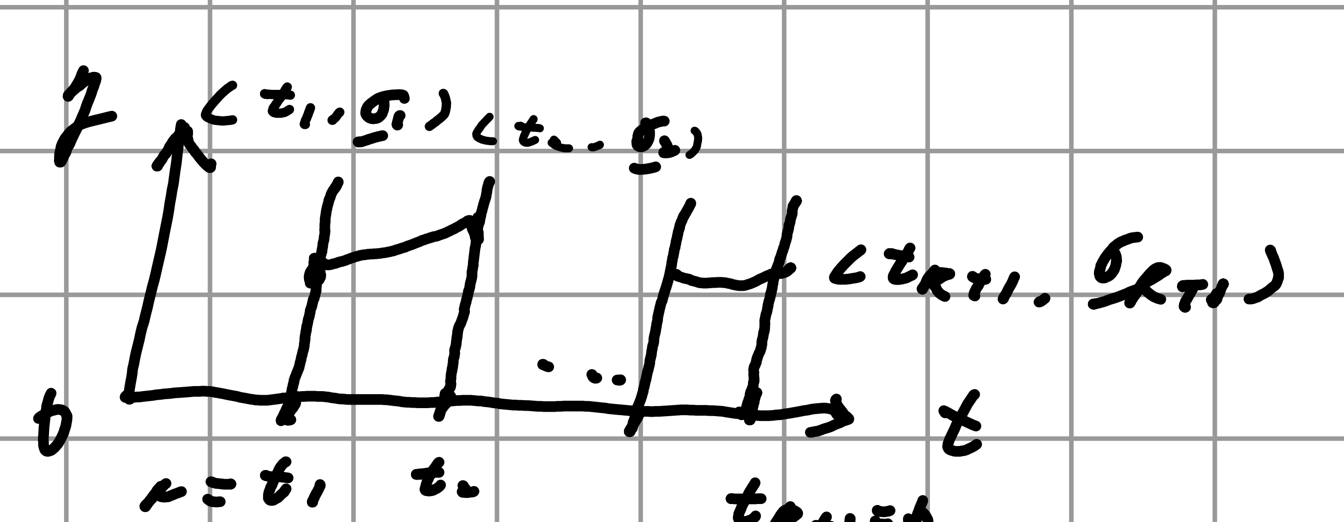
i) Simple concepts ii) Easy to implement.

Downside:

i) Long integration time can lead to instab.

ii) For bad datum $\underline{s}^{(k)}$. $\Rightarrow$ No guarantee the sol. will exist on the whole $[a,b]$.

② Multiple Shooting:



Rmk: We also introduce $\underline{s}_{K+1} = y(t_{K+1}, \underline{s}_K)$ here

For $(\underline{s}_i)_1^{K+1} \subset \mathbb{R}^n$. unknown. Recall the requirement of continuity and boundary cond:

$$\underline{F}(\underline{\sigma}) = \begin{pmatrix} \underline{r}(\underline{\sigma}_1, \underline{\sigma}_{R+1}) \\ \underline{y}(t_2, t_1, \underline{\sigma}_1) - \underline{\sigma}_2 \\ \vdots \\ \underline{y}(t_{R+1}, t_R, \underline{\sigma}_R) - \underline{\sigma}_{R+1} \end{pmatrix} = 0.$$

use Newton's : $J(\underline{\sigma}^{(i)}) \Delta \underline{\sigma}^{(i)} = -\underline{F}(\underline{\sigma}^{(i)})$.

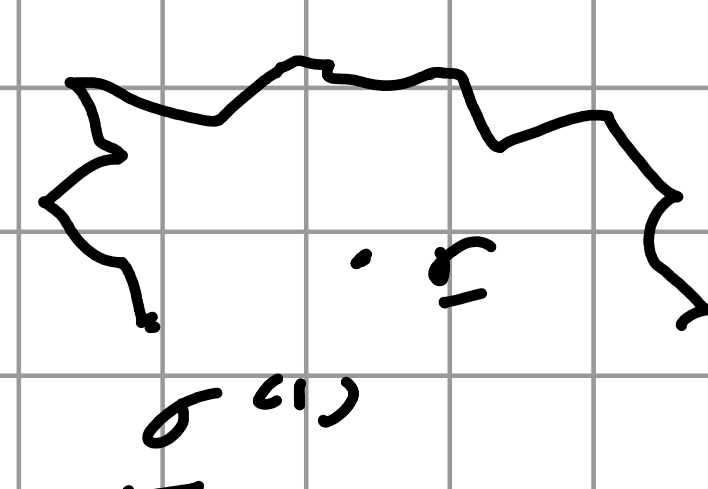
Let $\underline{\sigma}^{(i+1)} = \underline{\sigma}^{(i)} + \Delta \underline{\sigma}^{(i)}$.

Note that let $A = \nabla_1 \underline{r}(\underline{\sigma}_1, \underline{\sigma}_{R+1})$, $B = \nabla_2 \underline{r}(\underline{\sigma}_1, \underline{\sigma}_{R+1})$

and $G_i = \dfrac{\partial \underline{y}(t_{i+1}, t_i, \underline{\sigma}_i)}{\partial \underline{\sigma}_i} \Rightarrow J(\underline{\sigma}) = \begin{pmatrix} A & 0 & \cdots & & B \\ G_1 & -I & & & \vdots \\ & G_2 & \ddots & \ddots & \vdots \\ & & & G_R & -I \end{pmatrix}$

Rmk: i) A, B depends on $\underline{\sigma}$. $\Rightarrow$ They will keep updating.

ii) Newton's method may not converge if starting pt $\underline{\sigma}^{(0)}$ is far from the real. $\underline{\sigma}$:



We'll use globalized Newton's.

(5) <u>Finite Difference Method:</u>

We consider linear BVP in the following one

Assume it has unique sol. $\Leftrightarrow$ $(B_a + B_b \varphi(b)) \neq 0$)

Idea: Use equidistant mesh on $[a,b]$.

$t_n = a + nh$. $h = (b-a)/N$. $0 \le n \le N$. and try to find approxi. sol. y_n at t_n.

Rmk: i) It's different from RK. LMM. We solve y_n. $\forall n$ at the same time. (Not time-stepping method.)

ii) We can extend to non-equidistant setting. But it'll lead to reduced accuracy.

Approach: Approxi. derivative by finite differences

Rmk: Many time-stepping method can be reinterpreted like this. (see example below.)

Def: To approxi. first derivative: (diff. quotient.)

$D_h^+ u(x) = (u(x+h) - u(x))/h$. (forward diff.)

$D_h^- u(x) = (u(x) - u(x-h))/h$. (backward diff.)

$D_h^c u(x) = (u(x+h) - u(x-h))/2h$. (central diff.)

To approxi. Second derivative:

$D_h^2 u(x) = (u(x+h) + u(x-h) - 2u(x))/h^2$.

<u>Def</u>: A finite diff. operator D_h is consistent of order p with the q^{th} deri. if for $u \in C^{q+p}(a,b)$, $x \in (a+h, b-h)$, $\exists\, C>0$ indept of h. s.t. $|D_h^q u(x) - u^{(q)}(x)| \le Ch^p$.

<u>Rmk</u>: i) It's in sense of h small enough.

ii) D_h^+, D_h^- are order 1 for 1^{st} deri.

D_h^c is order 2 for 1^{st} deri.

D_h^2 is order 2 for 2^{nd} deri.

<u>Pf</u>: By Taylor expansion on $u(x+h)$ and $u(x-h)$ at $t=x$.

<u>e.g.</u> For D_h^c:

$u(x+h) - u(x-h) = 2hu'(x) + h^3(u^{(3)}(\xi^+) + u^{(3)}(\xi^-))/6$.

iii) Note we'd like h as small as possible to get good approxi. above

But there's a optimal h^* in fact

(i.e. $h \downarrow$ doesn't mean approxi. better)

e.g. i) For scalar IVP: $y'(t) = a(t)y(t) + f(t)$

$y(a) = \hat{y}$. $t \in I = [a, b]$.

Use FDE D_h^+ for example:

For $n = 0, 1, \cdots N-1$. approxi. $y'(t_n)$ by $\frac{y_{n+1} - y_n}{h}$

So: $y_{n+1} - y_n = h\,a(t_n)y_n + h f(t_n)$. $y_0 = \hat{y}$.

Denote $a_n = a(t_n)$. $f_n = f(t_n)$.

$$\Rightarrow \begin{pmatrix} 1 & & & 0 \\ -(1+ha_0) & 1 & & \\ & \ddots & \ddots & \\ 0 & & -(1+ha_{N-1}) & 1 \end{pmatrix} \begin{pmatrix} y_0 \\ \vdots \\ \vdots \\ y_N \end{pmatrix} = \begin{pmatrix} \hat{y} \\ hf_0 \\ \vdots \\ hf_{N-1} \end{pmatrix}.$$

$\Rightarrow$ We solve $(y_n)_{n=0}^{N}$ by getting y_0. then $y_1, y_2, \cdots y_N$. (It's just expl. Euler!)

Rmk: For using D_h^- $\Rightarrow$ it's just impl. Euler

ii) For $\underline{y}'(t) - A(t)\underline{y}(t) = \underline{f}(t)$. $B_a\underline{y}(a) + B_b\underline{y}(b) = \underline{\gamma}$

Combine D_h^+, D_h^- to get trapezoidal rule:

$(\underline{y}_n - \underline{y}_{n-1})/h = \frac{1}{2}(A_n\underline{y}_n + \underline{f}_n + A_{n-1}\underline{y}_{n-1} + \underline{f}_{n-1})$

Where $A_n = A(t_n)$. $\underline{f}_n = \underline{f}(t_n)$

$$\Rightarrow \begin{pmatrix} B_a & 0 & - & - & - & - & B_b \\ -(I+\frac{hA_0}{2}) & (I-\frac{hA_1}{2}) & & & & & \\ & \ddots & \ddots & & & & \\ & & & \ddots & \ddots & & \\ & & & & & -(I+\frac{hA_{n-1}}{2}) & (I-\frac{hA_n}{2}) \end{pmatrix} \begin{pmatrix} \underline{y}_0 \\ \vdots \\ \vdots \\ \underline{y}_n \end{pmatrix} = \begin{pmatrix} \underline{\gamma} \\ h(\underline{f}_0+\underline{f}_1)/2 \\ \vdots \\ h(\underline{f}_{n-1}+\underline{f}_n)/2 \end{pmatrix}$$

Denote the matrix on LHS by A_h. Since we don't know whether it's regular $\Rightarrow$ We don't know whether the linear system is solvable.

Def: i) $(L\underline{y})(t) := \underline{y}'(t) - A(t)\underline{y}(t) \; (= \underline{f}(t))$, $t \in I$.

$R\underline{y} = B_a\underline{y}(a) + B_b\underline{y}(b) \; (= \underline{\gamma})$.

We approxi. L, R by discrete version. e.g. in the example above.

$L_h\underline{y}_n = (\underline{y}_n - \underline{y}_{n-1})/h - \frac{1}{2}(A_n\underline{y}_n + A_{n-1}\underline{y}_{n-1})$

$= \frac{1}{2}(\underline{f}_n + \underline{f}_{n-1}) =: F_n(h, \underline{f})$.

$R_h(\underline{y}_n)_0^n = B_a\underline{y}_0 + B_b\underline{y}_n = \underline{\gamma}$.

ii) The truncation error is $\underline{\tau}_n^h = (L_h\underline{y}^h)_n - \underline{F}_n(h, \underline{f})$ where $\underline{y}^h = (\underline{y}(t_1), \dots, \underline{y}(t_n))$

Rmk: As before, we insert true sol. value in term $L_h\underline{y}^h$.

iii) The method is consistent of order m if

$$\| R_h y^h - \underline{g} \| + \max_{1 \le n \le N} \| (L_h y^h)_n - F_n(h, \underline{f}) \| = O(h^m)$$

iv) The method is stable if $\exists K > 0$ for sufficient small h: (provided $\forall (y_n)$)

$$\max_{0 \le n \le N} \| \underline{y}_n \| \le K (\| R_h (\underline{y}_n)_0^N \| + \max_{1 \le n \le N} \| (L_h (\underline{y}_n)_0^N)_n \|)$$

<u>Rmk:</u> First term on RHS is coding of $\underline{g}$ (i.e. boundary cond.) and second term is coding of $\underline{f}$ from F.

Note that the stability in iv) strongly depends on matrix A_h where A_h is the matrix from solving linear system of $\underline{y}$ when applying the scheme. (Note $\underline{y} = A_h^{-1} \cdot \square \Rightarrow K > 0$ will come from $|A_h|^{-1}$).

<u>Denote:</u> $\|A\|_\infty = \sup_{z^h \in \mathbb{R}^{k(N+1)}} \|A z^h\|_\infty / \|z^h\|_\infty$ induced matrix norm by $\|(\underline{y}_n)_0^N\|_\infty = \max_{0 \le n \le N} \|\underline{y}_n\|_2$. $\underline{y}_n \in \mathbb{R}^k$.

<u>Lem.</u> The scheme is stable $\Longleftrightarrow$ A_h is regular with $\sup_{h > 0} \|A_h^{-1}\|_\infty < \infty$ (uniform in h).

Thm. If the difference scheme is consistent of order $m \geq 1$ and stable. Then: it's convergent of order m: $\max_{0 \leq n \leq N} \| y_n - y(t_n) \| = O(h^m)$

Thm. The difference scheme is consistent (of order m) and stable for the linear BVP $\iff$ it's consistent (of order m) and stable for the corresponding IVP.

Rmk: Convergent difference method for IVP can be used to solve BVP and have same order of convergent.

(But classical time stepping e.g. RK. LMM is cheaper for IVPs since we don't need construct linear system)

Comparisons of FD and Shooting for BVP:

i) FD avoids issue of finding good datum.

ii) FD can be more robust for ill-cond. problem.

iii) FD allows for incorporating more complex forms of boundary conditions.

iv) Adaptive length control is easier to apply on shooting.

(*) only store y_N. not $\{y_n\}_1^N$

v) Shooting is cheaper (They both cost $O(d^3N)$.

But memory consump are diff.: $O(d^2)^{(*)}$ v.s $O(d^2N)$)

vi) FD is easier to use in high-order ODE.

(6) <u>Sturm-Liouville Problems:</u>

SL problems are 2^{nd}-order linear ODEs:

$(p(t)y'(t))' + q(t)y(t) = -\lambda w(t)y(t)$ with some boundary cond. where $p(t), q(t), w(t)$ are given func.

Value of λ for nontrivial sol. exist are called eigenvalues. Next, we consider variant:

$-(p(t)y'(t))' + q(t)y'(t) + r(t)y(t) = f(t),\ t\in[a,b]$

$\alpha_1 y'(a) + \alpha_0 y(a) = \gamma_1,\ \beta_1 y'(b) + \beta_0 y(b) = \gamma_2.$

Assume $p, q \in C^1[a,b]$. $p(t) \geq \delta > 0$. $r, f \in C[a,b]$.

$\alpha_0, \alpha_1, \beta_0, \beta_1, \gamma_1, \gamma_2 \in \mathbb{R}^1$.

Rewrite in system of first order:

$y_1'(t) = y_2(t)$. $y(t) = y_1(t)$

$-(p(t)y_2(t))' + q(t)y_2(t) + r(t)y_1(t) = f(t).$

with $\alpha_1 y_2(a) + \alpha_0 y_1(a) = \gamma_1$. $\beta_1 y_2(b) + \beta_0 y_1(b) = \gamma_2$.

$$\Rightarrow \begin{pmatrix} y_1(t) \\ y_2(t) \end{pmatrix}' = \begin{pmatrix} 0 & 1 \\ r(t)/p(t) & \frac{q(t)-p'(t)}{p(t)} \end{pmatrix} \begin{pmatrix} y_1(t) \\ y_2(t) \end{pmatrix} + \begin{pmatrix} 0 \\ \frac{-f(t)}{p(t)} \end{pmatrix}$$

$$\text{with } \begin{pmatrix} \alpha_0 & \alpha_1 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} y_1(a) \\ y_2(a) \end{pmatrix} + \begin{pmatrix} \beta_0 & \beta_1 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} y_1(b) \\ y_2(b) \end{pmatrix} = \begin{pmatrix} \gamma_1 \\ \gamma_2 \end{pmatrix}.$$

Which's reduced to form of linear BVP.

Rmk: For $\alpha_0 = \beta_0 = 1$. $\alpha_1 = \beta_1 = 0$. i.e. $\begin{cases} y_1(a) = \gamma_1 \\ y_1(b) = \gamma_2 \end{cases}$

It's called Dirichlet condition.

Thm. Under the assumptions above and assume

$\xi + (b-a)^{-2} \min_{[a,b]} (r(t) - \frac{1}{2} q'(t)) > 0$. Then:

the variant of SL problem with Dirichlet b.c. has unique sol. $y(t) \in C^2[a,b]$.

Pf: Note it has unique sol. $\Leftrightarrow$

$B_a + B_b \phi(b)$ is regular matrix.

Since for $Cx = b$ linear system can have 0. 1. ∞ sol. To prove $C = B_a + B_b \phi(b)$ can have unique sol. is equi to show $CX = 0$ has only $X = 0$ to solve it.

i.e. show the homo. problem with $f(t)=0$

$g_1(t)=g_2(t)=0$ has only sol. $y(t)=0$.

which can be showed by condition.

<u>FD Approxi.:</u>

Consider $Ly(t) = -(p(t)y'(t))' + q(t)y'(t) + r(t)y(t)$

$y(a)=g_1,\ y(b)=g_2.$

Assume $p, q \in C^3[a,b]$. $p(t) \geq \zeta > 0$. $g_1, g_2 \in \mathbb{R}^1$.

$r, f \in C^2[a,b]$. ($\Rightarrow y \in C^4$. to use Taylor)

$\zeta + (b-a)^2 \min\limits_{[a,b]} (r(t) - \frac{1}{2} q'(t)) > 0$.

We use equidistant mesh with $N+2$ pts:

$h = (b-a)/N+1$. $t_k = kh$. $0 \leq k \leq N+1$.

Discretize: $L_h y_n = -\tilde{D}^1 (p_n \overset{\leftarrow}{D}_1 y_n) + q_n \widehat{D}^1 y_n + r_n y_n$

$= f_n$. $1 \leq n \leq N$.

$$y_0 = g_1,\ y_{N+1} = g_2.$$

where p_n, q_n, r_n, f_n are value of $p(t), q(t)$, $r(t), f(t)$ at $t = t_n$.

Note D_h^c has order 2. We prefer to use it

$\Rightarrow$ let $\tilde{D}_h^1 = D_h^c$. But if $\widehat{D}_h^1 = D_h^c$:

When $p \equiv 1$, then: $(-p(t)y'(t))' = -y''(t)$.

But $D_h^c(D_h^c y_n) = (y_{n+2} - 2y_n + y_{n-2})/4h^2 = D^2_{2h} y_n$

which extends the distance of mesh.

So let $\hat{D}_h^c = D^c_{h/2}$. We get for $1 \le n \le N$:

$-h^{-2}(p_{n+\frac{1}{2}} y_{n+1} - (p_{n+\frac{1}{2}} + p_{n-\frac{1}{2}}) y_n + p_{n-\frac{1}{2}} y_{n-1}) +$

$(2h)^{-1} q_n (y_{n+1} - y_{n-1}) + r_n y_n = f_n$. $y_0 = g_1$, $y_{N+1} = g_2$

We get $N \times N$ system: $A_h Y = b_h$. $Y = (y_1 \cdots y_N)^T$

where $A_h = \frac{1}{h^2}\begin{pmatrix} p_{1/2} + p_{3/2} + h^2 r_1 & -p_{3/2} + \frac{h}{2} q_1 & & \\ & \ddots & \ddots & \\ -p_{n-1/2} - \frac{h}{2} q_n & p_{n+1/2} + p_{n-1/2} + h^2 r_n & -p_{n+1/2} + \frac{h}{2} q_n & \\ & & \ddots & \ddots \\ & & -p_{N-1/2} - \frac{h}{2} q_N & p_{N+1/2} + p_{N-1/2} + h^2 r_N \end{pmatrix}$

and $b_h = \begin{pmatrix} f_1 + \underbrace{\left(\frac{1}{h^2} p_{1/2} + \frac{1}{2h} q_1\right) g_1}_{\text{from } y_0} \\ f_2 \\ \vdots \\ f_{N-1} \\ f_N + \underbrace{\left(\frac{1}{h^2} p_{N+1/2} - \frac{1}{2h} q_N\right) g_2}_{\text{from } y_{N+1}} \end{pmatrix}$

Next, we consider the approxi. accuracy:

$$\underline{z}_n^h = (L_h y_h)_n - f_n. \quad y_h = (y(t_0), \dots, y(t_{N+1}))^T.$$

$$D_h^c u(t) \overset{\text{Taylor}}{=} u'(t) + \frac{h^2}{12}\underbrace{(u^{(3)}(\xi^+) + u^{(3)}(\xi^-))}_{:= h^2 R_\xi(u)} \quad (\Delta)$$

So $(L_h y_h)_n - f_n$

$$\overset{(\Delta)}{=} \underbrace{-(py')'(t_n) + q_n y'(t_n) + r_n y(t_n) - f_n}_{=0} + O(h^2)$$

$= O(h^2) \Rightarrow$ Consistency of order 2.

To get convergence, we need stability:

$\max_{1\le n\le N} |y_n| \le K(|y_0| + |y_{N+1}| + \max_{1\le i\le N} |(L_h(y_n)_{n=0}^{N+1})_i|)$ with

$K > 0$ const. indept of h. $\forall$ grid func. $(y_n)_0^{N+1}$.

__Def__: $e_n^h = y(t_n) - y_n$. $\underline{e}^h := (\underline{e}_0^h \cdots \underline{e}_{N+1}^h)$.

__Rmk__: $\underline{e}_0^h = \underline{e}_{N+1}^h = 0$ by boundary condition

$$(L_h \underline{e}^h)_n = (L_h y^h)_n - f_n = \underline{z}_n^h.$$

By stability, let $(y_n) = \underline{e}^h$:

$$\max_{1\le n\le N} |e_n^h| \le K(0 + 0 + \max_{1\le n\le N} |z_n^h|) = O(h^2).$$

Next, we want to show stability exists:

1°) $q \equiv 0$, $r \ge 0$: $-(p(t)y'(t))' + r(t)y(t) = f(t)$

$$y(a) = \gamma_1, \quad y(b) = \gamma_2, \quad t \in [a, b].$$

So:

$$A_h = \frac{1}{h^2}\begin{pmatrix} p_{1/2}+p_{3/2}+h^2 r_1 & -p_{3/2} & & & \\ & \ddots & \ddots & & \\ & -p_{n-1/2} & p_{n+1/2}+p_{n-1/2}+h^2 r_n & -p_{n+1/2} & \\ & & \ddots & \ddots & \\ & & & -p_{N-1/2} & p_{N+1/2}+p_{N-1/2}+h^2 r_N \end{pmatrix}$$

A_h is:

a) symmetric b) tridiag: $a_{ii} > 0$. $a_{i\pm1,i} < 0$. $\forall i$.

c) diagonally dominant. ($i = 1, N$ holds strictly)

d) irreducible. (No permutation matrix P. st.

PA_hP^T is block upper triangular matrix)

<u>Lem.</u> For $A \in \mathbb{R}^{n\times n}$ st.

i) diagonal dominant and at least one

a_i holds strictly ii) irreducible.

$\Rightarrow$ i) A is invertible

ii) If $A = A^T$. $a_{ii} > 0$. Then: we have

A is positive definite

iii) If $a_{ii} > 0$. $a_{ij} \le 0$. $\forall j \ne i$. $\forall i$. Then:

A is M-matrix. i.e. $A^{-1} = (c_{ij})_{n\times n}$

st. $c_{ij} \ge 0$. $\forall i, j$

Using this Lem. We can show the stab. estimate will hold (intuitively by $y = A_h^{-1} b_h$)

2) $q \neq 0$, $r \geq 0$: A_h isn't symmetric.

$$A_h = \frac{1}{h^2} \begin{pmatrix} p_{1/2} + p_{3/2} + h^2 r_1 & -p_{3/2} + \frac{h}{2} q_1 & & \\ & \ddots & & \\ -p_{n-1/2} - \frac{h}{2} q_n & p_{n+1/2} + p_{n-1/2} + h^2 r_n & -p_{n+1/2} + \frac{h}{2} q_n & \\ & & \ddots & \\ & & -p_{N-1/2} - \frac{h}{2} q_N & p_{N+1/2} + p_{N-1/2} + h^2 r_N \end{pmatrix}$$

$\Rightarrow$ A_h is:

a) irreducible if $-p_{n-\frac{1}{2}} - \frac{h}{2} q_n \neq 0$
$-p_{n+\frac{1}{2}} + \frac{h}{2} q_n \neq 0$ $\quad \forall n = 1, \dots N$.

b) diagonally dominant:

Note if $\frac{h}{2} |q_n| \leq p_{n-\frac{1}{2}} \wedge p_{n+\frac{1}{2}}$. Then:

$$\sum_{k \neq n} |a_{nk}| = |p_{n-\frac{1}{2}} + \frac{h}{2} q_n| + |p_{n+\frac{1}{2}} - \frac{h}{2} q_n|$$
$$= p_{n-\frac{1}{2}} + \frac{h}{2} q_n + p_{n+\frac{1}{2}} - \frac{h}{2} q_n$$
$$= p_{n-\frac{1}{2}} + p_{n+\frac{1}{2}} \leq |a_{nn}| \Rightarrow \text{diag. dom.}$$

So we need to choose h small s.t.

$h < 2 \min\limits_{1 \leq n \leq N} p_{n-\frac{1}{2}} \wedge p_{n+\frac{1}{2}} / |q_n|$ (h-condition)

Rmk: i) It also guarantees A_h is irred.

So we can show stability afterward

ii) No difference on "<" and "≤" in numerical. But it will be unstable if $h \gg 2 \min_{1\le n\le N} p_{n-\frac{1}{2}} \wedge p_{n+\frac{1}{2}} / |q_n|$ which happens when $|q_n| \gg |p_n|$ and h is not small enough.

c) $a_{ii} > 0$ and $a_{i\pm1,i} < 0$, $\forall i$ under h-condition

Potential Fix: (upwind discretization)

The troubling term involves q: $q(t)y'(t)$

instead of using $q_n D_h^c y_n$. We use $D_h^{\pm}$ by

$$q_n D_h^{\uparrow} y_n = \begin{cases} q_n D_h^{-} y_n & \text{if } q_n \ge 0 \\ q_n D_h^{+} y_n & \text{if } q_n < 0. \end{cases}$$

$\Rightarrow \hat{A}_h$ will be M-matrix without any requirement on h.

Rmk: Downside is we only have order 1 by using $D_h^{\pm}$. i.e. $\|\tau^h\| = O(h)$.

(7) <u>Finite Element Method:</u>

Next, assume $r = q = 0$ on Sturm-Liouville eq.:

$-(p(t)y'(t))' = f(t)$, $y(a) = g_1$, $y(b) = g_2$, $t \in I := [a, b]$

and $p \in C^1(I)$, $p(t) \geq \xi > 0$, $f \in C(I)$, $g_i \in \mathbb{R}^1$

$\Rightarrow \exists$ unique sol. $y \in C^2(I)$.

<u>Def</u>: $w(t) = y(t) - \ell(t)$, $\ell(t) = g_1 \frac{b-t}{b-a} + g_2 \frac{t-a}{b-a}$

i.e. a linear func. s.t. $\ell(a) = g_1$, $\ell(b) = g_2$

(So: $w(a) = w(b) = 0$)

$\Rightarrow -(pw')' = -(py')' + (p\ell')'$

$= f + (p\ell')' =: \hat{f}$

So to solve y, we can solve this homo. boundary cond. ($g_1 = g_2 = 0$) problem first.

Then: We can recover y from w.

So next, we'll focus on case $g_i = 0$, $i = 1, 2$

<u>Ideas:</u>

a) Write the problem in weak form.

b) Approxi. the weak form to get discrete sol.

a) Set $V = \{ v \in C[a,b] \mid v'$ is bdd and piece-wise conti. on $[a,b]$. $v(a) = v(b) = 0 \}$.

Rmk: We can also consider $\tilde{V} = H_0^1(a,b) \subsetneq V$.

$\Rightarrow \int_a^b p(t) \eta'(t) v'(t) dt = \int_a^b f(t) v(t) dt \quad \forall v \in V$.

follows from integration by part.

Write in $a(\eta, v) = \ell(v)$. $\forall v \in V$. $(*)$

Lem The problem $(*)$ has unique sol. in V.

Pf: Uniqueness: $a(\eta_1 - \eta_2, v) = 0$. $\forall v \in V$.

let $v = \eta_1 - \eta_2 \Rightarrow \int_a^b p(\eta_1' - \eta_2')^2 = 0$.

But $p(t) \geq \delta > 0$. $\Rightarrow \eta_1' = \eta_2'$.

With boundary cond. So: $\eta_1 = \eta_2$.

Existence is from def of W.

Rmk: For the case of $\tilde{V} = H_0^1(a,b)$. Note

BLO $a(v,v) \geq \delta \|v\|_{1,2}^2$. $\Rightarrow$ Apply Lax-Milgram

b) Next we do finite element approxi.:

$V \subset C[a,b]$ is separable. Set basis (ϕ_n).

$V_n = \text{span}\{\phi_1, \cdots \phi_n\} \subset V$. finite dimension.

Next, we want to find $y_n \in V_n$ s.t.

$a(y_n, v_n) = \ell(v_n)$, $\forall v_n \in V_n$

WLOG. let $V_n = \langle \phi_1, \cdots, \phi_n \rangle$. Denote $y_n = \sum_1^n \xi_j \phi_j$

$$\Rightarrow \sum_{j=1}^{n} a(\phi_j, \phi_i)\,\xi_j = \ell(\phi_i), \quad i = 1, \cdots, n.$$

Solve linear system: $A\xi = b$ defined in

$$(a(\phi_i, \phi_j))_{n\times n} \cdot \begin{pmatrix} \xi_1 \\ \vdots \\ \xi_n \end{pmatrix} = (\ell(\phi_1) \cdots \ell(\phi_n))^T.$$

Rmk: i) A is regular. ($Az = 0 \Rightarrow 0 = a(\sum z_i \phi_i) = 0 \Rightarrow z = 0$)

ii) We want to choose V_n to get a "nice" matrix A. (e.g. sparse matrix having lots of zero entries)

Def: mesh $h = (b-a)/N+1$. $t_i = a + ih$. $i = 0 \dots N+1$

$V_h := \{ v \in C[a,b] : v|_{(t_{i-1}, t_i)}$ is a linear polynomial. $v(a) = v(b) = 0 \}$.

We let basis of V_h is $\phi_i(t) = \frac{t - t_{i-1}}{h} I_{[t_{i-1}, t_i]} + \frac{t_{i+1} - t}{h} I_{[t_i, t_{i+1}]}$. hat func. $1 \le i \le N$.

e.g. $A_h = h^{-1} \begin{pmatrix} 2 & -1 & & 0 \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ 0 & & -1 & 2 \end{pmatrix}$ for $p \equiv 1$.

Rmk: i) We only save $O(N)$ rather $O(N^2)$ in this case.

ii) A_h is irred. diagonally dominant $\Rightarrow$ A_h is also positive definite. M-matrix.

Next, we want to compute a_{ij} of A_h:

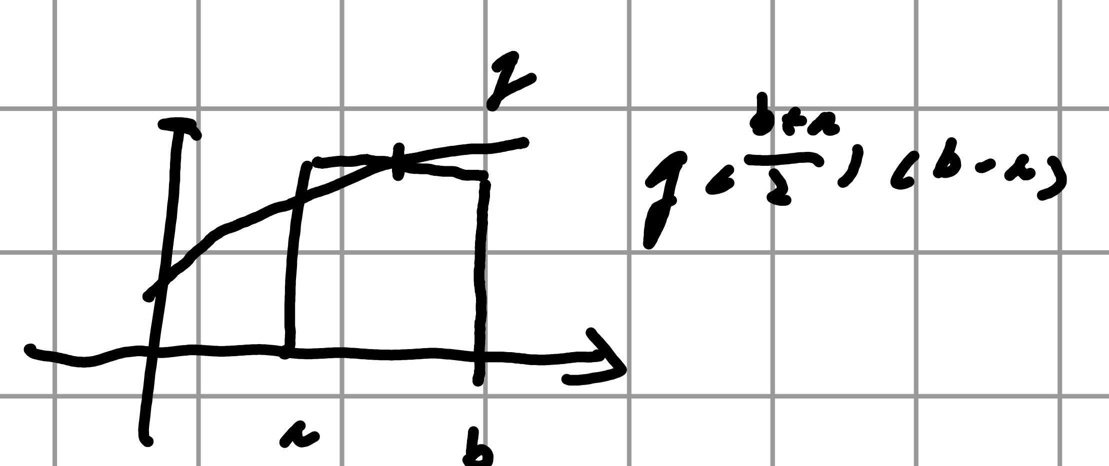
i) $a_{ij} = 0$ for $|i-j| > 1$.

ii) $a_{ij} = \int_a^b \varphi_i' \varphi_j' p \, dx$ $\Rightarrow$ we use quadrature to get the estimate.

Rmk: We will use the method of order $(k-1)^2$. k is degree of poly. (φ_j)

And $\ell(\varphi_i) = \int_a^b f \varphi_i \, dt = \int_{t_{i-1}}^{t_{i+1}} f \varphi_i \, dt$.

We let $\ell(\varphi_i) \approx h f(t_{i-\frac{1}{2}}) \varphi(t_{i-\frac{1}{2}}) +$



$$h f(t_{i+\frac{1}{2}}) \varphi(t_{i+\frac{1}{2}})$$

$$= \frac{h}{2} \left(f(t_{i-\frac{1}{2}}) + f(t_{i+\frac{1}{2}}) \right)$$

i.e. using mid-pt rule on (t_{i-1}, t_i), (t_i, t_{i+1})

Comparison to 2^{nd} diff quo.:

Use $D_h^2 u(t) = (u(t+h) + u(t-h) - 2u(t))/h^2$

Plug into $-y''(t) = f(t)$ on $[a,b]$. $y(a) = y(b) = 0$

$\Rightarrow A\underline{Y} = b$. $b = (f(t_1), \cdots f(t_N))^T$, $\underline{Y} = (y_1, \cdots y_N)^T$

$$A = h^{-2}\begin{pmatrix} 2 & -1 & & \\ -1 & 2 & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 2 \end{pmatrix}$$. So very similar LS.

<u>Rmk</u>: i) FEM and FD are more different for higher order approxi.

ii) Analysis for FEM / FD is different.

FD requires C^k space.

FEM requires Sobolev space.

<u>Thm.</u> (Error estimate for linear FE)

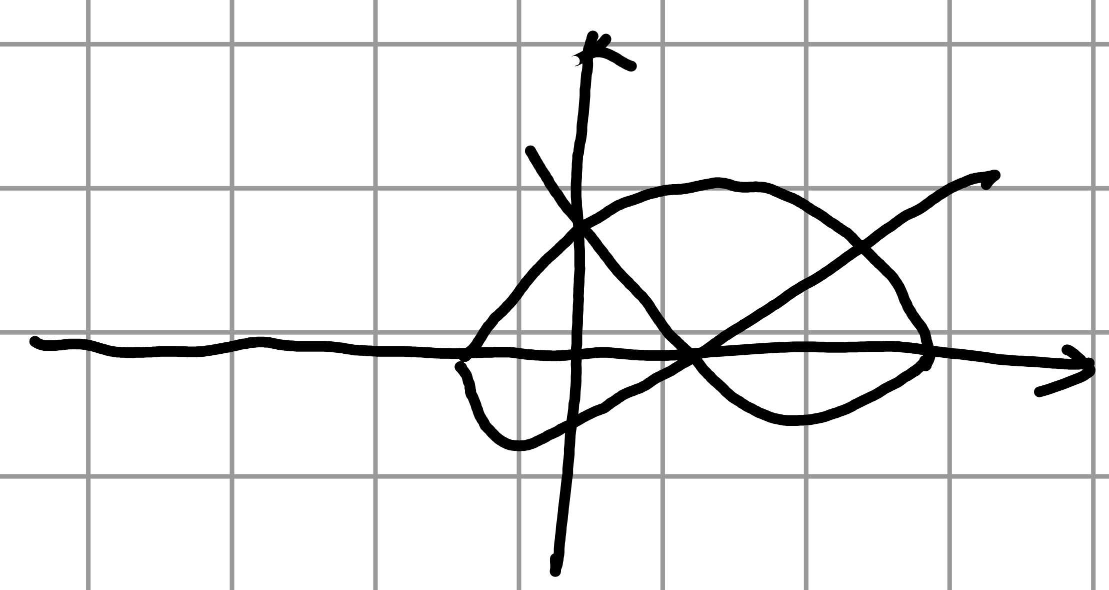
$$\max_{[a,b]} |y - y_h| \le Ch^2(|\log(h)| + 1) \max_{[a,b]} |y''|$$

<u>Rmk</u>: For order 2: it only needs C^2 while FD needs C^4. Typically, the regularity assumption for FD are higher than FEM

<u>Quadratic FE:</u>

$V_h := \{ v \in C[a,b] : v|_{(t_{i-1}, t_i)}$ is quadratic poly. and $v(a) = v(b) = 0 \}$.

We have 3 basis func. on (t_i, t_{i+1}):



Require : i) $\varphi_i(t_j) = \delta_{ij}$.

ii) $\sum \varphi_i = 1$. $\forall x$.

iii) φ_i are quadratic poly.

(8) <u>Numerical comparison for BVP:</u>

When coping with ill-conditioned problem (<u>e.g.</u> $y'(t) = A y(t)$. A has large eigenvalue $\Rightarrow$ has a large Lip. const.) :

① <u>Coarse step length:</u>

i) Multiple shooting is more expensive than single one. FDM (use D_h^2) is cheapest.

ii) Single shooting by solving IVP with $y(0)=s$ has large error. It's from property of ODE since Lip. const. is large.

② <u>Finer step length:</u>

multiple shooting has highest accuracy but it cost a lot. FDM can't increase its accuracy since it's just order 2. (even rounding error ↑)