Lecture 5

Zed

March 26, 2017

1 Newton's Method for Solving Nonlinear Equations

We want to find roots for nonlinear equation $g(x) = 0$.

Algo. **Newton's Method**

- Initialize a starting point x_0.
- Update x by $x_{n+1} \leftarrow x_n - \frac{g(x_n)}{g'(x_n)}$, suppose $g'(x_n) \neq 0$.
- Stop when $\|g(x_n)\| \leq thres$ (small), $\|x_{n+1} - x_n\| \leq thres$ (small) or $n \geq K$ (fail to converge).

The Newton's method only converges in a local sense. Suppose the real zero is α, s.t. $g(\alpha) = 0$, then using taylor expansion at x_n:

$$\begin{aligned} 0 = g(\alpha) &= g(x_n) + g'(x_n)(\alpha - x_n) + \frac{1}{2}g''(\xi_n)(\alpha - x_n)^2 \\ \Rightarrow 0 &= \left(\frac{g(x_n)}{g'(x_n)} - x_n\right) + \alpha + \frac{g''(\xi_n)}{2g'(x_n)}(\alpha - x_n)^2 \end{aligned} \tag{1}$$

Hence exist bound M such that

$$|\alpha - x_{n+1}| = \frac{g''(\xi_n)}{2g'(x_n)}|\alpha - x_n|^2 \leq M|\alpha - x_n|^2$$

That is, if the error at n-th iteration $|\alpha - x_n|$ is already small, the next error at $(n+1)$-th iteration will be the square of it. Which implies a (locally) quadratic convergence rate.

2 Implicit Methods

2.1 Implicit Runge-Kutta Method

We consider the Runge-Kutta method with table:

$$\begin{array}{c|cccc} c_1 & a_{11} & a_{12} & \dots & a_{1s} \\ c_2 & a_{21} & a_{22} & \dots & a_{2s} \\ c_3 & a_{31} & a_{32} & \dots & a_{3s} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c_s & a_{s1} & a_{s2} & \dots & a_{ss} \\ \hline & b_1 & b_2 & \dots & b_s \end{array} = \begin{array}{c|c} 1/2 & 1/2 \\ \hline & 1 \end{array}$$

A table of this kind with a full matrix of entries $\boldsymbol{A}$ (instead of entries only in the lowertriangular area) suggests that this is an *implicit* RK method.

For the listed example, we have

$$\begin{cases} y_{n+1} = y_n + hb_1k_1 \\ k_1 = f(t_n + c_1h, y_n + ha_{11}k_1) \end{cases} \tag{2}$$

And insert the values of coefficients:

$$y_{n+1} = y_n + h\left(f(t_n + \tfrac{1}{2}h, \tfrac{y_{n+1}}{2} + \tfrac{y_n}{2})\right)$$

Which is usually referred to as the *Implicit Midpoint Method*. This method is the simplest implicit RK method, the simplest *Gauss-Legendre* method. And a *Symplectic method*, which is energy-preserving.

The Midpoint method is of order 2 (see the first question in HW1).

Ex. (*Energy Preserving*) Consider the Hamilton system: $p' = -q, q' = p$; with $p(0) = q(0) = 1$. We can find a Hamilton function H such that $p' = -\frac{\partial H}{\partial q}$, $q' = \frac{\partial H}{\partial p}$. And hence $\frac{\partial H}{\partial t} = 0$. However, when we use the explicit RK method to solve the system (for example, the `ode45` in `Matlab`), we will find that the method does not preserve energy, i.e. $\frac{\partial H}{\partial t} \neq 0$.

3 Symplectic Methods

Our motivation to derive such a family of method is that it is natural to look into these discrete systems which preserves as much as possible the intrinsic properties of the continuous system. (Feng Kang, 1985; Ruth, 1983). We use the term *Symplectic* to refer to such systems, objects and methods.

Ex. We consider a simple system: $p' = -q, q' = p$, i.e. $H(p,q) = \frac{1}{2}p^2 + \frac{1}{2}q^2 = const.$ And there is a numerical method to solve this ode system. At n-th step we are at (p_n, q_n), and the numerical method will map this value to (p_{n+1}, q_{n+1}) in the next step.

We consider all the (p_n, q_n)'s within an infinitesimal area dV_n, and in next step it is mapped to dV_{n+1}. Furthermore, for all initial values (p_0, q_0) in the infinitesimal area dV_0, after n steps of numerial iteration, it will go to dV_n. We want a kind of *stability* such that initial values that are close to each other will yield numerical solutions that are also close to each other. In the setting of this problem we want the area of $dV_n \approx dV_0$. Intuitively, this suggests the same thing in the sense of stability. We investigate some methods.

1. *Explicit Euler*

$$\begin{cases} p_m = p_{m-1} - hq_{m-1} \\ q_m = q_{m-1} + hp_{m-1} \end{cases} \tag{3}$$

The Jacobian

$$\boldsymbol{J}_{ex} = \frac{\partial(p_m, q_m)}{\partial(p_{m-1}, q_{m-1})} = \begin{vmatrix} \frac{\partial p_m}{\partial p_{m-1}} & \frac{\partial p_m}{\partial q_{m-1}} \\ \frac{\partial q_m}{\partial p_{m-1}} & \frac{\partial q_m}{\partial q_{m-1}} \end{vmatrix} = \begin{vmatrix} 1 & -h \\ h & 1 \end{vmatrix} = 1 + h^2$$

So we will see $dV_m = (1+h^2)^m dV_0$. Any two initial values will evolve to be infinitely farwary.

2. *Implicit Euler*

$$\begin{cases} p_m = p_{m-1} - hq_m \\ q_m = q_{m-1} + hp_m \end{cases} \Rightarrow \begin{cases} (1+h^2)p_m = p_{m-1} - hq_{m-1} \\ (1+h^2)q_m = q_{m-1} + hp_{m-1} \end{cases} \tag{4}$$

The Jacobian

$$\boldsymbol{J}_{im} = \frac{\partial(p_m, q_m)}{\partial(p_{m-1}, q_{m-1})} = \begin{vmatrix} \frac{1}{1+h^2} & \frac{-h}{1+h^2} \\ \frac{h}{1+h^2} & \frac{1}{1+h^2} \end{vmatrix} = \frac{1}{1+h^2}$$

Therefore we will see $dV_m = (1+h^2)^{-m} dV_0$, the solution will eventually degenerate to one point.

3. *Symplectic Euler*

$$\begin{cases} p_m = p_{m-1} - hq_{m-1} \\ q_m = q_{m-1} + hp_m \end{cases} \Rightarrow \begin{cases} p_m = p_{m-1} - hq_{m-1} \\ q_m = q_{m-1} + h(p_{m-1} - hq_{m-1}) \end{cases} \tag{5}$$

$$\boldsymbol{J}_{symp} = \frac{\partial(p_m, q_m)}{\partial(p_{m-1}, q_{m-1})} = \begin{vmatrix} 1 & -h \\ h & 1-h^2 \end{vmatrix} = 1$$

4. *Implicit Midpoint*

$$\begin{cases} p_m = p_{m-1} - \frac{h}{2}q_{m-1} - \frac{h}{2}q_m \\ q_m = q_{m-1} + \frac{h}{2}p_{m-1} + \frac{h}{2}p_m \end{cases} \Rightarrow \begin{cases} (1+\frac{h^2}{4})p_m = (1-\frac{h^2}{4})p_{m-1} - hq_{m-1} \\ (1+\frac{h^2}{4})q_m = (1-\frac{h^2}{4})q_{m-1} + hp_{m-1} \end{cases} \tag{6}$$

$$\boldsymbol{J}_{mid} = \frac{\partial(p_m, q_m)}{\partial(p_{m-1}, q_{m-1})} = \begin{vmatrix} \frac{4-h^2}{4+h^2} & \frac{-4h}{4+h^2} \\ \frac{4h}{4+h^2} & \frac{4-h^2}{4+h^2} \end{vmatrix} = \frac{(4-h^2)^2 + 16h^2}{(4+h^2)^2} = 1$$

Def. **Hamiltonian System** We investigate Hamiltonian system in higher space dimensions: generalized coordinate $\boldsymbol{q} = (q_1, ..., q_d)^\top$ and generalized momentum: $\boldsymbol{p} = (p_1, ..., p_d)^\top$,

$$\begin{cases} \dot{\boldsymbol{p}} = -\frac{\partial H}{\partial \boldsymbol{q}} \\ \dot{\boldsymbol{q}} = \frac{\partial H}{\partial \boldsymbol{p}} \end{cases} \tag{7}$$

It is easy to check that the *Hamiltonian Energy* $H = const$ over time.

$$\frac{d}{dt}H(\boldsymbol{p}, \boldsymbol{q}) = \left(\frac{\partial H}{\partial \boldsymbol{p}}\right)^\top \dot{\boldsymbol{p}} + \left(\frac{\partial H}{\partial \boldsymbol{q}}\right)^\top \dot{\boldsymbol{q}} = 0$$

We can write in another form, let $\boldsymbol{y} = (\boldsymbol{p}, \boldsymbol{q})^\top = (p_1, ..., p_d; q_1, ..., q_d)^\top$, $\boldsymbol{J} := \begin{pmatrix} \boldsymbol{O} & \boldsymbol{I}_d \\ -\boldsymbol{I}_d & \boldsymbol{O} \end{pmatrix}$, one can easily check that $\boldsymbol{J}^{-1} = -\boldsymbol{J}$. Then we can write the system as $\dot{\boldsymbol{y}} = \boldsymbol{J}^{-1}\nabla \boldsymbol{H}(\boldsymbol{y})$ $(*)$.

Proof. $\boldsymbol{J}^{-1} = -\boldsymbol{J}$ because:

$$\boldsymbol{J}\boldsymbol{J} = \begin{pmatrix} \boldsymbol{O} & \boldsymbol{I}_d \\ -\boldsymbol{I}_d & \boldsymbol{O} \end{pmatrix} \begin{pmatrix} \boldsymbol{O} & \boldsymbol{I}_d \\ -\boldsymbol{I}_d & \boldsymbol{O} \end{pmatrix} = \begin{pmatrix} -\boldsymbol{I}_d & \boldsymbol{O} \\ \boldsymbol{O} & -\boldsymbol{I}_d \end{pmatrix} = -\boldsymbol{I}_{2d}$$

And

$$RHS(*) = \begin{pmatrix} \boldsymbol{O} & -\boldsymbol{I}_d \\ \boldsymbol{I}_d & \boldsymbol{O} \end{pmatrix} \begin{pmatrix} \nabla_{\boldsymbol{p}} H \\ \nabla_{\boldsymbol{q}} H \end{pmatrix} = \begin{pmatrix} -\nabla_{\boldsymbol{q}} H \\ \nabla_{\boldsymbol{p}} H \end{pmatrix} = \begin{pmatrix} -\frac{\partial H}{\partial \boldsymbol{q}} \\ \frac{\partial H}{\partial \boldsymbol{p}} \end{pmatrix} = (\dot{\boldsymbol{p}}, \dot{\boldsymbol{q}})^\top = LHS$$

Def. **Flow Map**: Given an autonomous differential equation $\boldsymbol{y}' = \boldsymbol{f}(\boldsymbol{y})$, the *flow map* $\boldsymbol{\varphi}_t(\cdot) : \mathbb{R}^d \to \mathbb{R}^d$, such that $\boldsymbol{\varphi}_t(\boldsymbol{y}_0) = \boldsymbol{y}(t)$ maps the initial value to the solution at t. The definition can be extended to the flow map of *measurable sets* $\Omega \subseteq \mathbb{R}^d$:

$$\begin{aligned} &\boldsymbol{\varphi}_t : \mathcal{B}(\mathbb{R}^d) \to \mathcal{B}(\mathbb{R}^d) \\ &\boldsymbol{\varphi}_t(\Omega) = \{\boldsymbol{y}(t) : \boldsymbol{y}_0 \in \Omega\} \end{aligned} \tag{8}$$

where $\mathcal{B}(\mathbb{R}^d)$ is the Borel sigma-algebra on $\mathbb{R}^d$ (the collection of all measurable sets).

Def. **Symplecticity**: For the Hamiltonian system $(\boldsymbol{p}, \boldsymbol{q}) \in \mathbb{R}^d \times \mathbb{R}^d$:

1. If p, q has only one dimension, i.e. $d = 1$, the system is *symplectic* if the area is preserved in phase space $\mathbb{R} \times \mathbb{R}$ under the flow map $\boldsymbol{\varphi}_t$.
2. $d \geq 2$, $(\boldsymbol{p}, \boldsymbol{q}) = (p_1, ..., p_d; q_1, ..., q_d)$ the system is symplectic if the quantity

$$w^2 = \sum_{i=1}^{d} dp_i \wedge dq_i$$

is preserved under the flow map $\boldsymbol{\varphi}_t$. Where $\wedge$ is the wedge product.

Prop. The equivalent definition of symplecticity: a dynamic system is symplectic $\iff$

$$\boldsymbol{J} = \boldsymbol{\Phi}_t^\top \boldsymbol{J} \boldsymbol{\Phi}_t$$

where $\boldsymbol{J} = \begin{pmatrix} \boldsymbol{O} & \boldsymbol{I}_d \\ -\boldsymbol{I}_d & \boldsymbol{O} \end{pmatrix}$, $\boldsymbol{\Phi}_t = \boldsymbol{\Phi}_t(\boldsymbol{y}) = \partial \boldsymbol{\varphi}_t(\boldsymbol{y}) / \partial \boldsymbol{y}$ is the Jacobian of the flow map.

*Proof.**

$$\begin{aligned} w^2 = \sum_{i=1}^{d} dp_i \wedge dq_i &= \frac{1}{2}\left(\sum_{i=1}^{d} dp_i \wedge dq_i - \sum_{i=1}^{d} dq_i \wedge dp_i \right) \\ &= \frac{1}{2} \boldsymbol{J}^{-1} d\boldsymbol{\varphi}_t \wedge d\boldsymbol{\varphi}_t = \frac{1}{2} \boldsymbol{J}^{-1} \boldsymbol{\Phi}_t d\boldsymbol{y} \wedge \boldsymbol{\Phi}_t d\boldsymbol{y} \\ &= \frac{1}{2} \boldsymbol{\Phi}_t^\top \boldsymbol{J}^{-1} \boldsymbol{\Phi}_t d\boldsymbol{y} \wedge d\boldsymbol{y} \end{aligned} \tag{9}$$

And we require $w^2(t) = w^2(0)$, knowing $\boldsymbol{\Phi}_t(0) = \boldsymbol{I} \Rightarrow \frac{1}{2} \boldsymbol{\Phi}_t^\top \boldsymbol{J}^{-1} \boldsymbol{\Phi}_t d\boldsymbol{y} \wedge d\boldsymbol{y} = \frac{1}{2} \boldsymbol{J}^{-1} d\boldsymbol{y} \wedge d\boldsymbol{y}$. So we have $\boldsymbol{\Phi}_t^\top \boldsymbol{J} \boldsymbol{\Phi}_t = \boldsymbol{J}$. $\square$

Thm. (**Poincaré**): Hamiltonian system $\iff$ Symplectic system.

Proof. We only show ($\Rightarrow$), i.e. showing $\boldsymbol{\Phi}_t^\top \boldsymbol{J} \boldsymbol{\Phi}_t = \boldsymbol{J}$ from Hamiltonian.
We use $\dot{\boldsymbol{y}} = \boldsymbol{J}^{-1} \nabla H(\boldsymbol{y})$, $\boldsymbol{\varphi}_t(\boldsymbol{y})$ is flow map. Since $\boldsymbol{\varphi}_t(\boldsymbol{y}) = \boldsymbol{y}(t)$, it must satisfy the ode when we *regard t as its varable* (in the time dimension). That is

$$\frac{d\boldsymbol{\varphi}_t(\boldsymbol{y})}{dt} = \boldsymbol{J}^{-1} \nabla H(\boldsymbol{\varphi}_t(\boldsymbol{y}))$$

$$\frac{d\boldsymbol{\Phi}_t(\boldsymbol{y})}{dt} = \frac{\partial}{\partial \boldsymbol{y}} \frac{d\boldsymbol{\varphi}_t(\boldsymbol{y})}{dt} = \boldsymbol{J}^{-1} \nabla^2 H(\boldsymbol{\varphi}_t(\boldsymbol{y})) \frac{\partial \boldsymbol{\varphi}_t(\boldsymbol{y})}{\partial \boldsymbol{y}} = \boldsymbol{J}^{-1} \nabla^2 H(\boldsymbol{\varphi}_t(\boldsymbol{y})) \boldsymbol{\Phi}_t(\boldsymbol{y})$$

where $\nabla^2 H$ is the hessian. Therefore

$$\begin{aligned} \frac{d}{dt}\left(\boldsymbol{\Phi}_t^\top \boldsymbol{J} \boldsymbol{\Phi}_t \right) &= \left(\frac{d\boldsymbol{\Phi}_t}{dt} \right)^\top \boldsymbol{J} \boldsymbol{\Phi}_t + \boldsymbol{\Phi}_t^\top \boldsymbol{J} \frac{d\boldsymbol{\Phi}_t}{dt} \\ &= (\boldsymbol{J}^{-1} \nabla^2 H(\boldsymbol{\varphi}_t(\boldsymbol{y}) \boldsymbol{\Phi}_t)^\top \boldsymbol{J} \boldsymbol{\Phi}_t + \boldsymbol{\Phi}_t^\top \boldsymbol{J} \boldsymbol{J}^{-1} \nabla^2 H(\boldsymbol{\varphi}_t(\boldsymbol{y})) \boldsymbol{\Phi}_t \\ &= \boldsymbol{\Phi}_t^\top \nabla^2 H(\boldsymbol{\varphi}_t(\boldsymbol{y})) \boldsymbol{J}^{-\top} \boldsymbol{J} \boldsymbol{\Phi}_t + \boldsymbol{\Phi}_t^\top \boldsymbol{J} \boldsymbol{J}^{-1} \nabla^2 H(\boldsymbol{\varphi}_t(\boldsymbol{y})) \boldsymbol{\Phi}_t \\ &= -\boldsymbol{\Phi}_t^\top \nabla^2 H(\boldsymbol{\varphi}_t(\boldsymbol{y})) \boldsymbol{\Phi}_t + \boldsymbol{\Phi}_t^\top \nabla^2 H(\boldsymbol{\varphi}_t(\boldsymbol{y})) \boldsymbol{\Phi}_t \\ &= 0 \end{aligned} \tag{10}$$

Using the fact that $\boldsymbol{J}^{-\top} = (-\boldsymbol{J})^\top = \begin{pmatrix} \boldsymbol{O} & -\boldsymbol{I}_d \\ \boldsymbol{I}_d & \boldsymbol{O} \end{pmatrix}^\top = \boldsymbol{J}$, and $\boldsymbol{J}\boldsymbol{J} = -\boldsymbol{I}$. Therefore $\boldsymbol{\Phi}_t^\top \boldsymbol{J} \boldsymbol{\Phi}_t = \boldsymbol{\Phi}_t^\top \boldsymbol{J} \boldsymbol{\Phi}_t \mid_{t=0}$. Note that $\boldsymbol{\Phi}_t \mid_{t=0} = \det(\boldsymbol{I})$. Hence we have $\boldsymbol{\Phi}_t^\top \boldsymbol{J} \boldsymbol{\Phi}_t = \boldsymbol{J}$. $\square$