Numerical Analysis
陈曦，HOME
Summer, 2024

数值分析

NODE

参考书目：

- Numerical Analysis（David Kincaid, Ward Cheney）
- Numerical Analysis（Timothy Sauer，2014）
- Finite Difference Methods for Ordinary and Partial Differential Equations（LeVeque，2007）

考虑常微分初值问题：

$$\begin{aligned} u'(t) &= f(u(t),t), \quad t\in[0,T], \\ u(0) &= u_0, \end{aligned} \tag{0.1}$$

本文给出一些基本的数值计算方法，并对一般的 LMM 给出其收敛性、相容性和稳定性判断方法。

1 数值方法

首先给出一些相对简单的数值方法：

1. 前向 Euler 法：
$$\frac{U^{n+1}-U^n}{h} = f(U^n, t_n).$$
2. 后向 Euler 法：
$$\frac{U^{n+1}-U^n}{h} = f(U^{n+1}, t_{n+1}).$$
3. 梯形法：
$$\frac{U^{n+1}-U^n}{2h} = \frac{1}{2}(f(U^n, t_n) + f(U^{n+1}, t_{n+1})).$$
4. 中点法（蛙跳法）：
$$\frac{U^{n+1}-U^{n-1}}{2h} = f(U^n, t_n).$$

从局部截断误差来看，前向 Euler 法和后向 Euler 法是一阶方法，梯形法和中点法是二阶方法；从显隐性上看，前向 Euler 法和中点法是显式方法，后向 Euler 法和梯形法是隐式方法；从步数上来看，前三种方法是单步法，中点法是两步法。除了使用中心差分或者混合前后向差分的方法可以得到二阶截断误差外，我们还可以使用单侧差分的方法来得到二阶截断误差，例如

$$\begin{aligned} \frac{3u(t)-4u(t-h)+u(t-2h)}{2h} &= u'(t)+O(h^2), \\ -\frac{u(t+2h)-4u(t+h)+3u(t)}{2h} &= u'(t)+O(h^2). \end{aligned}$$

稍后我们将对前向 Euler 进行改进使它达到二阶截断误差，这种方法本质上是二阶 Runge-Kutta 方法的特例，即

$$Y_1 = U^n, \quad Y_2 = U^n + hf(Y_1, t_n), \quad U^{n+1} = U^n + \frac{h}{2}(f(Y_1, t_n) + f(Y_2, t_{n+1})).$$

1.1 Runge-Kutta Methods

最具代表性的多阶段方法是 Runge-Kutta 法。一般的 r 阶段 Runge-Kutta 方法的形式为

$$\begin{aligned}
Y_1 &= U^n + h\sum_{j=1}^{r} a_{1j} f(Y_j, t_n + c_j h);\\
&\vdots\ \vdots\ \vdots\\
Y_i &= U^n + h\sum_{j=1}^{r} a_{ij} f(Y_j, t_n + c_j h);\\
&\vdots\ \vdots\ \vdots\\
Y_r &= U^n + h\sum_{j=1}^{r} a_{rj} f(Y_j, t_n + c_j h);\\
U^{n+1} &= U^n + h\sum_{j=1}^{r} b_j f(Y_j, t_n + c_j h).
\end{aligned} \tag{1.1}$$

其中各系数可以储存在如下的 Butcher 表中：

$$\begin{array}{c|ccc}
c_1 & a_{11} & \cdots & a_{1r}\\
\vdots & \vdots & \ddots & \vdots\\
c_r & a_{r1} & \cdots & a_{rr}\\
\hline
1 & b_1 & \cdots & b_r
\end{array} \tag{1.2}$$

当 $i \leqslant j$ 时 $a_{ij} = 0$，即上表中右上角矩阵是严格下三角的时，相应的方法为显式的，因此称为显式 Runge-Kutta 方法。

一般地，为了使局部截断误差达到某一阶数需要要求系数满足一定的条件，例如：

1. 一阶截断误差（相容）：
$$\sum_{j=1}^{r} b_j = 1, \quad \sum_{j=1}^{r} a_{ij} = c_i, \quad i = 1, 2, \cdots, r.$$
2. 二阶截断误差：除以上条件外还需
$$\sum_{j=1}^{r} b_j c_j = \frac{1}{2}.$$
3. 三阶截断误差：除以上条件外还需
$$\sum_{j=1}^{r} b_j c_j^2 = \frac{1}{3}, \quad \sum_{j=1}^{r} b_i a_{ij} c_j = \frac{1}{6}.$$

一种简单的二阶 Runge-Kutta 方法的 Butcher 表为

$$\begin{array}{c|cc}
0 & & \\
1 & 1 & \\
\hline
1 & \frac{1}{2} & \frac{1}{2}
\end{array} \tag{1.3}$$

该方法的局部截断误差为 $O(h^2)$，单步误差为 $O(h^3)$，具体格式为

$$\begin{aligned}
Y_1 &= U^n;\\
Y_2 &= U^n + hf(Y_1, t_n);\\
U^{n+1} &= U^n + h\left[\frac{1}{2} f(Y_1, t_n) + \frac{1}{2} f(Y_2, t_n + h)\right],
\end{aligned} \tag{1.4}$$

即

$$U^{n+1} = U^n + \frac{h}{2}\left[f(U^n, t_n) + f(U^n + hf(U^n, t_n), t_{n+1})\right], \tag{1.5}$$

这种方法也称作改进 Euler 法。一般的显式次对角（仅 $a_{i+1,i}$ 非零）二阶 Runge-Kutta 方法的形式为

$$U^{n+1} = U^n + b_1 hf(U^n, t_n) + b_2 hf(U^n + a_2 hf(U^n, t_n), t_n + c_2 h),$$

为了具有二阶截断误差需要

$$b_1+b_2=1,\quad b_2c_2=\frac{1}{2},\quad b_2a_2=\frac{1}{2}.$$

最常使用的是四阶 Runge-Kutta 方法，即经典的 RK4 方法，其 Butcher 表为

$$\begin{array}{c|cccc} 0 & & & & \\ \frac{1}{2} & \frac{1}{2} & & & \\ \frac{1}{2} & 0 & \frac{1}{2} & & \\ 1 & 0 & 0 & 1 & \\ \hline 1 & \frac{1}{6} & \frac{1}{3} & \frac{1}{3} & \frac{1}{6} \end{array} \tag{1.6}$$

该方法的局部截断误差为 $O(h^4)$，单步误差为 $O(h^5)$。

在构造 Runge-Kutta 方法时，往往直接令除 $a_{i+1,i}$ 之外的 $a_{ij}=0$，然后通过相容性条件来确定 $a_{i+1,i}$ 的值，记 $a_{i+1,i}=a_{i+1}$ 于是这种情况下的 Runge-Kutta 方法形如

$$\begin{aligned} Y_1&=U^n;\\ Y_2&=U^n+ha_2f(Y_1,t_n+c_1h);\\ &\vdots\ \ \vdots\ \ \vdots\\ Y_r&=U^n+ha_rf(Y_{r-1},t_n+c_{r-1}h);\\ U^{n+1}&=U^n+h\sum_{j=1}^{r}b_jf(Y_j,t_n+c_jh). \end{aligned} \tag{1.7}$$

为了使该方法达到所需阶数需要使用 Taylor 展开，这里可能需要使用到二维的 Taylor 展开

$$\begin{aligned} f(x,y)&=f(x_0,y_0)+\nabla f(x_0,y_0)(x-x_0,y-y_0)\\ &\quad+\frac{1}{2}(x-x_0,y-y_0)H_f(x_0,y_0)\begin{pmatrix}x-x_0\\y-y_0\end{pmatrix}+O((x-x_0)^2+(y-y_0)^2)\\ &=f(x_0,y_0)+f_x(x_0,y_0)(x-x_0)+f_y(x_0,y_0)(y-y_0)\\ &\quad+\frac{1}{2}\left[(x-x_0)^2f_{xx}+2(x-x_0)(y-y_0)f_{xy}+(y-y_0)^2f_{yy}\right]+O((x-x_0)^2+(y-y_0)^2). \end{aligned} \tag{1.8}$$

1.2 Multistep Methods

线性多步法（LMM）涉及到不同时间层的信息，一般形式为

$$\sum_{j=0}^{k}\alpha_jU^{n+j}=h\sum_{j=0}^{k}\beta_jf(U^{n+j},t_{n+j}), \tag{1.9}$$

当 $\beta_j=0$ 时该方法是显式的，通常做归一化使得 $\alpha_k=1$。

最常用的 LMM 方法是 Adams 方法，该方法的数值格式形如

$$U^{n+k}=U^{n+k-1}+h\sum_{j=0}^{k}\beta_jf(U^{n+j},t_{n+j}), \tag{1.10}$$

即要求 $\alpha_k=-\alpha_{k-1}=1$，$\alpha_{k-2}=\cdots=\alpha_0=0$。当 $\beta_k=0$ 时该方法是显式的，借助 Taylor 展开，通过调整 $\beta_0,\cdots,\beta_{k-1}$ 可以使该方法的单步误差达到 k 阶，相应的方法称为显式 Adams-Bashforth 方法。当 $\beta_k\neq0$ 时，可以调整 $\beta_0,\cdots,\beta_k$ 使得单步误差阶达到 $k+1$，相应的方法称为隐式 Adams-Moulton 方法。

现在我们考察一般的 r 步 LMM 的局部截断误差，注意到

$$\tau(t_{n+r})=\frac{1}{h}\left(\sum_{j=0}^{r}\alpha_ju(t_{n+j})-h\sum_{j=0}^{r}\beta_ju'(t_{n+j})\right),$$

其中 u 是 $u'=f(u)$ 的精确解，$\tau(t_{n+r})$ 是局部截断误差。现在做 Taylor 展开可得

$$\begin{aligned} u(t_{n+j})&=u(t_n)+jhu'(t_n)+\frac{j^2}{2}h^2u''(t_n)+\cdots,\\ u'(t_{n+j})&=u'(t_n)+jhu''(t_n)+\frac{j^2}{2}h^2u'''(t_n)+\cdots, \end{aligned}$$

于是带入 $\tau(t_{n+r})$ 可得

$$\tau(t_{n+r}) = \frac{1}{h}\left(\sum_{j=0}^{r}\alpha_j\right)u(t_n) + \left(\sum_{j=0}^{r}(j\alpha_j - \beta_j)\right)u'(t_n) + h\left(\sum_{j=0}^{r}(\frac{1}{2}j^2\alpha_j - j\beta_j)\right)u''(t_n) + \cdots + h^{q-1}\left(\sum_{j=0}^{r}(\frac{1}{q!}j^q\alpha_j - \frac{1}{(q-1)!}j^{q-1}\beta_j)\right)u^{(q)}(t_n) + \cdots,$$

为了保证相容性，至少需要令前两项为零，即

$$\sum_{j=0}^{r}\alpha_j = 0, \quad \sum_{j=0}^{r}j\alpha_j = \sum_{j=0}^{r}\beta_j.$$

如果定义 $d_0 = \sum_{j=0}^{r}\alpha_j$，且

$$d_i = \sum_{j=0}^{r}(\frac{j^i}{i!}\alpha_j - \frac{j^{i-1}}{(i-1)!}\beta_j), \quad i = 1, 2, \cdots, \tag{1.11}$$

则我们有如下结论：

Theorem 1. *对于 r 步 LMM，以下命题等价：*

1. $d_0 = d_1 = \cdots = d_m = 0$；
2. *该 LMM 的局部截断误差大小为 $O(h^m)$，单步误差大小为 $O(h^{m+1})$。*

为了进一步分析 LMM 的稳定性，我们定义 r 步 LMM 的特征多项式为

$$\rho(z) = \sum_{j=0}^{r}\alpha_j z^j \quad , \sigma(z) = \sum_{j=0}^{r}\beta_j z^j. \tag{1.12}$$

因此 LMM 是相容的必要条件是 $\rho(1) = 0$ 且 $\rho'(1) = \sigma(1)$。

2 收敛性

所谓收敛性是指当网格尺寸 $h \to 0$ 时数值解 U^n 逼近精确解 $u(t_n)$ 的性质，通常难以直接验证某一方法是否具有收敛性，常用的方法是使用如下定理将收敛性判定转化为相容性和稳定性判定：

Theorem 2. *使用 LMM 数值求解某常微分初值问题 $u'(t) = f(u(t), t)$ 时，该方法收敛当且仅当它具有相容性和稳定性。*

对于 LMM 而言，相容性要求 $\rho(1) = 0, \rho'(1) = \sigma(1)$，如果需要更高阶的相容性需要要求其他的 $d_i = 0$，相容性刻画了离散的差分方程与原方程之间的距离是否可以通过缩小步长来消除。另一方面，由于 LMM 作为一种时间推进法，某一次时间推进过程中会使用到前面逼近的信息，在这个过程中，稳定的数值方法可以保证之前的单步误差不会累积并被放大。稳定性有多种刻画，常用的包括零稳定性和绝对稳定性。

零稳定性要求方法在 $h \to 0$ 的极限是稳定的，对于 LMM 而言，零稳定性可以通过验证特征多项式 ρ 是否满足根条件来快速判断。

Definition. *考虑某 r 步 LMM，如果特征多项式 ρ 的所有根 ξ_i 满足如下根条件：*

1. $|\xi_i| \leqslant 1$ *对所有* $i = 1, \cdots, r$ *成立；*
2. *如果 ξ_i 是 ρ 的重根，那么 $|\xi_i| < 1$；*

则称该 LMM 是零稳定的。

零稳定性只能描述方法在 $h \to 0$ 时的稳定性，而在实际应用中我们更关心如何选取有限的步长 h 来使得方法是稳定的，为此需要考虑绝对稳定性。给定 LMM，要分析它的绝对稳定性需要将它用于模型问题：

$$u'(t) = \lambda u(t), \quad u(0) = 1, \tag{2.1}$$

当使用步长为 h 时，令 $z = \lambda h$。此时我们需要考察

$$\pi(\xi; z) = \rho(\xi) - z\sigma(\xi) \tag{2.2}$$

而非 ρ，对于 LMM 而言，绝对稳定性要求 $\pi(\xi; z)$ 的根满足根条件，即如下定义。

Definition. 给定任意 r 步 LMM，当 $\pi(\xi;z)$ 的所有根 ξ_i 满足如下根条件：

1. $|\xi_i| \leqslant 1$ 对所有 $i = 1, \cdots, r$ 成立；
2. 如果 ξ_i 是 π 的重根，那么 $|\xi_i| < 1$；

则称该 LMM 是绝对稳定的。因为 $\xi_i = \xi_i(z)$ 是 z 的函数，因此要求所有 π 的根满足根条件实际上给出了 z 的取值范围，这一范围称为绝对稳定区域。

不难看出，LMM 是零稳定的当且仅当 $z = 0$ 位于绝对稳定区域内。

更详细的介绍参见NOPDE。