数值 ode

Yuexin Liao

June 2023

1 常微分方程

我们的主要目标是求解常微分方程的数值解。我们将关注以下形式的微分方程

$$\mathbf{y}'(t) = \mathbf{f}(t, \mathbf{y}(t))$$

其中 $0 \le t \le T$，初始条件为

$$\mathbf{y}(0) = \mathbf{y}_0$$

我们给出的条件是函数 $\mathbf{f} : \mathbb{R} \times \mathbb{R}^N \to \mathbb{R}^N$，结束时间 $T > 0$，以及初始条件 $\mathbf{y}_0 \in \mathbb{R}^N$。我们所寻求的是函数 $\mathbf{y} : [0, T] \to \mathbb{R}^N$。

当求解微分方程的数值解时，我们的目标是使我们的数值解尽可能接近真实解。这只有在“真实”解确实存在且唯一的情况下才有意义。我们接下来说明

命题 1. *如果 $\mathbf{f}$ 是 Lipschitz 的，那么 ODE 会有唯一解。*

定义 1 (Lipschitz 函数)**.** *如果函数 $\mathbf{f} : \mathbb{R} \times \mathbb{R}^N \to \mathbb{R}^N$ 满足以下条件：*

$$\|\mathbf{f}(t, \mathbf{x}) - \mathbf{f}(t, \hat{\mathbf{x}})\| \le \lambda \|\mathbf{x} - \hat{\mathbf{x}}\|$$

对于所有 $t \in [0, T]$ 和 $\mathbf{x}, \hat{\mathbf{x}} \in \mathbb{R}^N$，则称其为 Lipschitz 函数，Lipschitz 常数为 $\lambda \ge 0$。

如果函数对于某些 λ 是 Lipschitz 的，则称其为 Lipschitz 的。

证明. 首先，我们利用 Picard 迭代法在一个较小的时间区间内证明局部存在性和唯一性。

定义迭代序列 $\{\mathbf{y}_n(t)\}$：

$$\begin{cases} \mathbf{y}_0(t) = \mathbf{y}_0 \\ \mathbf{y}_{n+1}(t) = \mathbf{y}_0 + \int_0^t \mathbf{f}(s, \mathbf{y}_n(s))\, ds \end{cases}$$

我们希望序列 $\{\mathbf{y}_n(t)\}$ 在某个函数空间中收敛，并且极限函数是原初值问题的解。

令 $\|\cdot\|$ 表示在区间 $[0, T]$ 上的最大范数，即

$$\|g\| = \sup_{t \in [0,T]} \|g(t)\|$$

首先证明 $\{\mathbf{y}_n(t)\}$ 是 Cauchy 序列。计算 $\mathbf{y}_{n+1}(t)$ 和 $\mathbf{y}_n(t)$ 的差：

$$\mathbf{y}_{n+1}(t) - \mathbf{y}_n(t) = \int_0^t \left(\mathbf{f}(s, \mathbf{y}_n(s)) - \mathbf{f}(s, \mathbf{y}_{n-1}(s))\right) ds$$

取范数并使用 Lipschitz 条件，有：

$$\|\mathbf{y}_{n+1}(t) - \mathbf{y}_n(t)\| \leq \int_0^t \|\mathbf{f}(s, \mathbf{y}_n(s)) - \mathbf{f}(s, \mathbf{y}_{n-1}(s))\| \, ds \leq L \int_0^t \|\mathbf{y}_n(s) - \mathbf{y}_{n-1}(s)\| \, ds$$

对上式两边取最大范数，有：

$$\|\mathbf{y}_{n+1} - \mathbf{y}_n\| \leq L \int_0^T \|\mathbf{y}_n(s) - \mathbf{y}_{n-1}(s)\| \, ds$$

令 $M_n = \|\mathbf{y}_n - \mathbf{y}_{n-1}\|$，则

$$M_{n+1} \leq L \int_0^T M_n \, ds = LTM_n$$

为了保证 Picard 迭代法收敛，我们选择一个足够小的时间区间 $[0, \delta]$，使得 $L\delta < 1$。在这个区间上，递归关系表明 $M_n \to 0$ 随着 $n \to \infty$，即 $\{\mathbf{y}_n(t)\}$ 是 Cauchy 序列，在完备的函数空间中收敛到某个极限函数 $\mathbf{y}(t)$。根据极限的定义以及 Picard 迭代法的构造，极限函数 $\mathbf{y}(t)$ 满足原初值问题在 $[0, \delta]$ 上的解。

我们现在通过分段延展的方法，将局部解扩展到整个区间 $[0, T]$。

假设在区间 $[0, \delta]$ 上，存在唯一解 $\mathbf{y}(t)$。在 $t = \delta$ 处，定义新的初值问题：

$$\begin{cases} \mathbf{y}'(t) = \mathbf{f}(t, \mathbf{y}(t)), \\ \mathbf{y}(\delta) = \mathbf{y}_\delta, \end{cases}$$

其中 $\mathbf{y}_\delta = \mathbf{y}(\delta)$。在新的时间区间 $[\delta, 2\delta]$ 上，同样可以应用 Picard 迭代法证明存在唯一解。如此重复上述过程，我们可以在每个区间 $[(k-1)\delta, k\delta]$ 上构造解，使得每个解在接缝处保持连续。

通过这种分段延展的方法，我们可以在有限的步数内将解扩展到整个区间 $[0, T]$，因为 T 可以表示为 $N\delta$，其中 N 是正整数。

假设 $\mathbf{y}_1(t)$ 和 $\mathbf{y}_2(t)$ 是初值问题的两个解，我们需要证明 $\mathbf{y}_1(t) = \mathbf{y}_2(t)$ 对所有 $t \in [0, T]$ 成立。

定义 $\mathbf{z}(t) = \mathbf{y}_1(t) - \mathbf{y}_2(t)$，则：

$$\mathbf{z}'(t) = \mathbf{f}(t, \mathbf{y}_1(t)) - \mathbf{f}(t, \mathbf{y}_2(t))$$

取范数并使用 Lipschitz 条件，有：

$$\|\mathbf{z}'(t)\| \leq L\|\mathbf{y}_1(t) - \mathbf{y}_2(t)\| = L\|\mathbf{z}(t)\|$$

利用 Grönwall 不等式：

$$\|\mathbf{z}(t)\| \leq \|\mathbf{z}(0)\|e^{Lt} = 0 \cdot e^{Lt} = 0$$

因为 $\mathbf{z}(0) = \mathbf{y}_1(0) - \mathbf{y}_2(0) = \mathbf{y}_0 - \mathbf{y}_0 = 0$，所以 $\|\mathbf{z}(t)\| = 0$，即 $\mathbf{z}(t) = 0$，这表明 $\mathbf{y}_1(t) = \mathbf{y}_2(t)$。

□

Lipschitz 条件对于微分方程解的存在性和唯一性是充分条件，我们可以讨论我们的解是否收敛到这个唯一解。我们常常额外假设函数 $\mathbf{f}$ 可以展开为泰勒级数到任意高的阶数，因为这对于我们的分析很方便。

ODE 数值解包含什么呢？我们首先选择一个小的时间步长 $h > 0$，然后构建近似

$$\mathbf{y}_n \approx \mathbf{y}(t_n), \quad n = 1, 2, \cdots,$$

其中 $t_n = nh$。特别地，$t_n - t_{n-1} = h$ 并且始终是恒定的。在实践中，我们不会固定步长 $t_n - t_{n-1}$，而是允许其在每一步中变化。然而，这使得分析更加复杂，我们在本中不会考虑变化的时间步长。

如果我们使 h 变小，那么我们（可能）会得到更好的近似解。然而，这更具计算挑战性。因此，我们想研究数值方法的行为，以确定我们应该选择什么 h。

1.1 一步法

我们可以用很多方法对数值方法进行分类。一个重要的分类是一阶法和多阶法。在一阶法中，$\mathbf{y}_{n+1}$ 仅依赖于前一次迭代 t_n 和 $\mathbf{y}_n$。在多阶法中，我们可以回溯更长时间并使用更早的迭代结果。

定义 2 ((显式) 一步法)**.** *如果数值方法中的 $\mathbf{y}_{n+1}$ 仅依赖于 t_n 和 $\mathbf{y}_n$，即*

$$\mathbf{y}_{n+1} = \boldsymbol{\phi}_h(t_n, \mathbf{y}_n)$$

对于某个函数 $\boldsymbol{\phi}_h : \mathbb{R} \times \mathbb{R}^N \to \mathbb{R}^N$，则称该数值方法为 (显式) 一步法。

我们将理解 “显式” 的含义。

最简单的一步法为 *欧拉法*。

定义 3 (欧拉法)**.** *欧拉法 使用公式*

$$\mathbf{y}_{n+1} = \mathbf{y}_n + h\mathbf{f}(t_n, \mathbf{y}_n)$$

我们想要证明这种方法是 “收敛的”。首先，我们需要精确定义 “收敛” 的概念。Lipschitz 条件意味着微分方程有唯一解。因此，只要我们选择足够小的 h，我们希望数值解能够任意精确地逼近实际解。

定义 4 (数值方法的收敛性)**.** *对于每个 $h > 0$，我们可以生成一系列离散值 $\mathbf{y}_n$，其中 $n = 0, \cdots, [T/h]$，$[T/h]$ 是 T/h 的整数部分。如果随着 $h \to 0$ 且 $nh \to t$ (因此 $n \to \infty$)，我们得到*

$$\mathbf{y}_n \to \mathbf{y}(t),$$

其中 $\mathbf{y}$ 是微分方程的真实解。我们要求在 t 上的收敛是均匀的。

我们现在证明欧拉法是收敛的。我们将只对欧拉法进行严格证明，因为代数运算很快会变得繁琐且难以理解。然而，这个证明策略足够通用，可以适用于大多数其他方法。

定理 1 (欧拉法的收敛性)**.**

1. *对于所有 $t \in [0, T]$，我们有*
$$\lim_{\substack{h \to 0 \\ nh \to t}} \mathbf{y}_n - \mathbf{y}(t) = 0$$

2. *设 λ 为 f 的 Lipschitz 常数。则存在 $c \geq 0$，使得*
$$\|\mathbf{e}_n\| \leq ch \frac{e^{\lambda T} - 1}{\lambda}$$
对于所有 $0 \leq n \leq [T/h]$，其中 $\mathbf{e}_n = \mathbf{y}_n - \mathbf{y}(t_n)$。

注意，第二部分中的界是均匀的。因此这立即给出了定理的第一部分。

证明. 证明分为两个部分。我们首先看局部截断误差。这是假设我们前面的步骤都正确时每一步会得到的误差。更确切地说，我们写

$$\mathbf{y}(t_{n+1}) = \mathbf{y}(t_n) + h(\mathbf{f}, t_n, \mathbf{y}(t_n)) + \mathbf{R}_n$$

其中 $\mathbf{R}_n$ 是局部截断误差。对于欧拉法，很容易得到 $\mathbf{R}_n$，因为根据定义 $\mathbf{f}(t_n, \mathbf{y}(t_n)) = \mathbf{y}'(t_n)$。所以这只是 $\mathbf{y}$ 的泰勒级数展开。我们可以将 $\mathbf{R}_n$ 表示为泰勒级数的积分余项，

$$\mathbf{R}_n = \int_{t_n}^{t_{n+1}} (t_{n+1} - \theta) y''(\theta) \, \mathrm{d}\theta$$

我们不难得到

$$\|\mathbf{R}_n\|_\infty \leq ch^2$$

其中

$$c = \frac{1}{2}\|y''\|_\infty$$

这部分证明相对简单。一旦我们限定了局部截断误差，我们将它们拼接起来得到实际误差。我们可以写成

$$\begin{aligned}
\mathbf{e}_{n+1} &= \mathbf{y}_{n+1} - \mathbf{y}(t_{n+1}) \\
&= \mathbf{y}_n + h\mathbf{f}(t_n, \mathbf{y}_n) - \Big(\mathbf{y}(t_n) + h\mathbf{f}(t_n, \mathbf{y}(t_n)) + \mathbf{R}_n\Big) \\
&= (\mathbf{y}_n - \mathbf{y}(t_n)) + h\Big(\mathbf{f}(t_n, \mathbf{y}_n) - \mathbf{f}(t_n, \mathbf{y}(t_n))\Big) - \mathbf{R}_n
\end{aligned}$$

取无穷范数，我们得到

$$\begin{aligned}
\|\mathbf{e}_{n+1}\|_\infty &\leq \|\mathbf{y}_n - \mathbf{y}(t_n)\|_\infty + h\|\mathbf{f}(t_n, \mathbf{y}_n) - \mathbf{f}(t_n, \mathbf{y}(t_n))\|_\infty + \|\mathbf{R}_n\|_\infty \\
&\leq \|\mathbf{e}_n\|_\infty + h\lambda\|\mathbf{e}_n\|_\infty + ch^2 \\
&= (1 + \lambda h)\|\mathbf{e}_n\|_\infty + ch^2
\end{aligned}$$

这对于所有 $n \geq 0$ 都是成立的。我们还知道 $\|\mathbf{e}_0\| = 0$。做一些代数运算，我们得到

$$\|\mathbf{e}_n\|_\infty \leq ch^2 \sum_{j=0}^{n-1} (1 + h\lambda)^j \leq \frac{ch}{\lambda}\left((1 + h\lambda)^n - 1\right)$$

最后，我们有

$$(1 + h\lambda) \leq e^{\lambda h}$$

因为 $1 + \lambda h$ 是泰勒级数的前两项，其他项为正。因此

$$(1 + h\lambda)^n \leq e^{\lambda h n} \leq e^{\lambda T}$$

因此，我们获得了界

$$\|\mathbf{e}_n\|_\infty \leq ch\frac{e^{\lambda T} - 1}{\lambda}$$

然后当我们取 $h \to 0$ 时，这趋于 0。因此，该方法是收敛的。 □

只要 $\lambda \neq 0$，这都是有效的。然而，$\lambda = 0$ 是简单情况，因为它只是积分。我们可以直接验证这种情况，或者使用 $\frac{e^{\lambda T}-1}{\lambda} \to T$ 当 $\lambda \to 0$。

同样的证明策略适用于大多数数值方法，但代数计算会更复杂。我们不会对所有方法进行完整的证明，但会给出一些有用的术语:

定义 5 (局部截断误差). *对于一个一般的（多步）数值方法*

$$\mathbf{y}_{n+1} = \boldsymbol{\phi}(t_n, \mathbf{y}_0, \mathbf{y}_1, \cdots, \mathbf{y}_n)$$

局部截断误差 *是*

$$\boldsymbol{\eta}_{n+1} = \mathbf{y}(t_{n+1}) - \boldsymbol{\phi}_n(t_n, \mathbf{y}(t_0), \mathbf{y}(t_1), \cdots, \mathbf{y}(t_n))$$

这是我们在第 $(n + 1)$ 步时如果前 n 步都是准确值会产生的误差。

对于欧拉法，局部截断误差就是泰勒级数的余项。

定义 6 (阶数). *数值方法的阶数是使* $\boldsymbol{\eta}_{n+1} = O(h^{p+1})$ *的最大* $p \geq 1$。

欧拉方法是一阶的。注意，这比局部截断误差的阶数少一，因为当我们考虑全局误差时，会降一个阶数，只有 $\mathbf{e}_n \sim h$。我们来看看比欧拉法更高级的方法。

定义 7 (θ-方法). *对于* $\theta \in [0,1]$，θ-*方法 是*

$$\mathbf{y}_{n+1} = \mathbf{y}_n + h\Big(\theta \mathbf{f}(t_n, \mathbf{y}_n) + (1-\theta)\mathbf{f}(t_{n+1}, \mathbf{y}_{n+1})\Big)$$

如果我们取 $\theta = 1$，那么我们就得到了欧拉法。最常用的两个 θ 值是 $\theta = 0$（向后欧拉法）和 $\theta = \frac{1}{2}$（梯形法）。

注意，对于 $\theta \neq 1$，我们得到的是*隐式方法*。这是因为 $\mathbf{y}_{n+1}$ 不仅出现在等式的左边。我们的 $\mathbf{y}_{n+1}$ 公式中涉及 $\mathbf{y}_{n+1}$ 本身！这意味着，一般情况下，不像欧拉法，我们不能简单地写出 $\mathbf{y}_n$ 的值，必须将公式视为 N 个（一般情况下）非线性方程，并求解以找到 $\mathbf{y}_{n+1}$！

过去，人们不喜欢使用这种方法，因为他们没有计算机，或者计算机太慢。每一步都要求解这些方程很繁琐。如今，随着求解方程变得越来越容易，这些方法变得越来越流行，因为它们在理论上有很大的优势（但我们没有时间在此文中深入探讨这些优势）。

现在我们看一下 θ-方法的误差。我们有

$$\boldsymbol{\eta} = \mathbf{y}(t_{n+1}) - \mathbf{y}(t_n) - h\Big(\theta \mathbf{y}'(t_n) + (1-\theta)\mathbf{y}'(t_{n+1})\Big)$$

我们用泰勒级数展开所有关于 t_n 的项

$$= \left(\theta - \frac{1}{2}\right) h^2 \mathbf{y}''(t_n) + \left(\frac{1}{2}\theta - \frac{1}{3}\right) h^3 \mathbf{y}'''(t_n) + O(h^4)$$

我们看到 $\theta = \frac{1}{2}$ 给我们一个二阶方法。否则，我们得到一个一阶方法。

1.2 多步法

我们可以通过利用以前的 $\mathbf{y}_n$ 值而不仅仅是最近的一个值来使我们的方法更高效。一种常见的方法是 AB2 方法：

定义 8 (二步 Adams-Bashforth 方法). *二步 Adams-Bashforth (AB2) 方法 是*

$$\mathbf{y}_{n+2} = \mathbf{y}_{n+1} + \frac{1}{2}h\left(3\mathbf{f}(t_{n+1}, \mathbf{y}_{n+1}) - \mathbf{f}(t_n, \mathbf{y}_n)\right)$$

这是 Adams-Bashforth 方法的一种特例。

一般来说，多步法定义如下：

定义 9 (多步法). s-*步数值方法 由以下公式给出*

$$\sum_{\ell=0}^{s} \rho_\ell \mathbf{y}_{n+\ell} = h \sum_{\ell=0}^{s} \sigma_\ell \mathbf{f}(t_{n+\ell}, \mathbf{y}_{n+\ell})$$

该公式用于找到 $\mathbf{y}_{n+s}$ *的值。*

需要注意的一点是，如果我们将所有的常数 ρ_ℓ, σ_ℓ 乘以一个非零常数，我们得到相同的方法。按照惯例，我们将 $\rho_s = 1$ 进行标准化。然后我们可以改写为

$$\mathbf{y}_{n+s} = h \sum_{\ell=0}^{s} \sigma_\ell \mathbf{f}(t_{n+\ell}, \mathbf{y}_{n+\ell}) - \sum_{\ell=0}^{s-1} \rho_\ell \mathbf{y}_{n+\ell}$$

如果 $\sigma_s \neq 0$，则该方法是隐式的。否则，它是显式的。

注意，这个方法在某种意义上是线性的，因为系数 ρ_ℓ 和 σ_ℓ 出现在 $\mathbf{f}$ 之外的线性方程中。稍后我们将看到更多复杂的数值方法，其中这些系数出现在 $\mathbf{f}$ 的参数中。

对于多步方法，我们需要解决一个小问题。在一步法中，我们给定 $\mathbf{y}_0$，这允许我们立即应用一步法得到更高的 $\mathbf{y}_n$ 值。然而，对于 s-步方法，我们需要使用其他（可能是一阶）方法来获得 $\mathbf{y}_1, \cdots, \mathbf{y}_{s-1}$，然后才能开始。

幸运的是，即使当 $h \to 0$ 时，我们也只需要应用一次固定的、小步数的一步法。因此，开始时的一步法的精度并不太重要。

我们现在研究一般多步方法的性质。首先我们可以讨论阶数：

定理 2. *一个 s-步方法的阶数是 p $(p \geq 1)$ 当且仅当*

$$\sum_{\ell=0}^{s} \rho_\ell = 0$$

和

$$\sum_{\ell=0}^{s} \rho_\ell \ell^k = k \sum_{\ell=0}^{s} \sigma_\ell \ell^{k-1}$$

对于 $k = 1, \cdots, p$，其中 $0^0 = 1$。

这是一个直接从定义推导出来的技术结果。

证明. 局部截断误差是

$$\sum_{\ell=0}^{s} \rho_\ell \mathbf{y}(t_{n+\ell}) - h \sum_{\ell=0}^{s} \sigma_\ell \mathbf{y}'(t_{n+\ell})$$

我们现在展开 $\mathbf{y}$ 和 $\mathbf{y}'$ 关于 t_n 的泰勒级数，得到

$$\left(\sum_{\ell=0}^{s} \rho_\ell \right) \mathbf{y}(t_n) + \sum_{k=1}^{\infty} \frac{h^k}{k!} \left(\sum_{\ell=0}^{s} \rho_\ell \ell^k - k \sum_{\ell=0}^{s} \sigma_\ell \ell^{k-1} \right) \mathbf{y}^{(k)}(t_n)$$

在给定条件下，这是 $O(h^{p+1})$。 □

例 1.1 (AB2)**.** *在二步 Adams-Bashforth 方法中，我们看到条件对于 $p = 2$ 成立，但对于 $p = 3$ 不成立。因此阶数是 2。*

而不是直接处理 ρ_ℓ 和 σ_ℓ 系数，而是将它们组合成两个多项式。这涉及与数值方法相关的两个多项式。它们是

$$\rho(w) = \sum_{\ell=0}^{s} \rho_\ell w^\ell, \quad \sigma(w) = \sum_{\ell=0}^{s} \sigma_\ell w^\ell$$

然后我们可以用这个来重新陈述上述定理。

定理 3. *一个多步法具有阶数 p（$p \geq 1$）当且仅当*

$$\rho(e^x) - x\sigma(e^x) = O(x^{p+1})$$

当 $x \to 0$。

证明. 我们展开

$$\rho(e^x) - x\sigma(e^x) = \sum_{\ell=0}^{s} \rho_\ell e^{\ell x} - x \sum_{\ell=0}^{s} \sigma_\ell e^{\ell x}$$

我们现在将 $e^{\ell x}$ 在 $x = 0$ 处展开泰勒级数。结果是

$$\sum_{\ell=0}^{s} \rho_\ell + \sum_{k=1}^{\infty} \frac{1}{k!} \left(\sum_{\ell=0}^{s} \rho_\ell \ell^k - k \sum_{\ell=0}^{s} \sigma_\ell \ell^{k-1} \right) x^k$$

因此结果成立。 □

注意 $\sum_{\ell=0}^{s} \rho_\ell = 0$，这是该方法具有阶数的必要条件，可以表示为 $\rho(1) = 0$。

例 1.2 (AB2). *在二步 Adams-Bashforth 方法中，我们得到*

$$\rho(w) = w^2 - w, \quad \sigma(w) = \frac{3}{2}w - \frac{1}{2}$$

我们可以立即检查 $\rho(1) = 0$。我们还得到

$$\rho(e^x) - x\sigma(e^x) = \frac{5}{12}x^3 + O(x^4)$$

因此阶数是 2。

我们已经解决了多步法的阶数问题。接下来要检查的是收敛性。这是一步法和多步法的区别所在。对于一步法，我们只需要了解阶数即可理解收敛性。事实是，只要阶数 $p \geq 1$，一步法就会收敛。对于多步法，我们还需要额外的条件。

定义 10 (根条件). *如果 $\rho(w)$ 的所有零点都被 1 的模所限制，即所有根 w 满足 $|w| \leq 1$，则称 $\rho(w)$ 满足根条件。此外，模为 1 的任何零点必须是简单根。*

我们可以说这意味着最大根不能超过 1，并且我们不能有太多模为 1 的零点。

我们看到任何合理的多步法必须有 $\rho(1) = 0$。因此，特别的，1 必须是一个简单零点。

定理 4 (Dahlquist 等价定理). *多步法收敛当且仅当*

1. *阶数 p 至少为 1；并且*
2. *满足根条件。*

这个证明太过困难，这里省略。

例 1.3 (AB2). *再次考虑二步 Adams-Bashforth 方法。我们已经看到它的阶数 $p = 2 \geq 1$。因此，我们需要检查根条件。所以 $\rho(w) = w^2 - w = w(w-1)$。因此它满足根条件。*

让我们现在制定一个合理的策略来构建收敛的 s-步方法：

1. 选择一个 ρ，使得 $\rho(1) = 0$ 并满足根条件。
2. 选择 σ 以最大化阶数，即

$$\sigma = \frac{\rho(w)}{\log w} + \begin{cases} O(|w-1|^{s+1}) & \text{若隐式} \\ O(|w-1|^{s}) & \text{若显式} \end{cases}$$

我们有两种不同的条件，因为对于隐式方法，我们有更多的系数可以调整，因此可以获得更高的阶数。

这个 $\frac{1}{\log w}$ 从何而来？我们尝试代入 $w = e^x$（注意 $e^x - 1 \sim x$）。然后公式表示

$$\sigma(e^x) = \frac{1}{x}\rho(e^x) + \begin{cases} O(x^{s+1}) & \text{若隐式} \\ O(x^{s}) & \text{若显式} \end{cases}$$

重新排列得到

$$\rho(e^x) - x\sigma(e^x) = \begin{cases} O(x^{s+2}) & \text{若隐式} \\ O(x^{s+1}) & \text{若显式} \end{cases}$$

这就是我们的阶数条件。因此，给定任意的 ρ，只有一种合理的方式来选择 σ。所以关键在于选择一个足够好的 ρ。

根条件“最好”通过选择 $\rho(w) = w^{s-1}(w-1)$ 来满足，即除一个根之外的所有根都是 0。然后我们有

$$\mathbf{y}_{n+s} - \mathbf{y}_{n+s-1} = h\sum_{\ell=0}^{s} \sigma_\ell \mathbf{f}(t_{n+\ell}, \mathbf{y}_{n+\ell})$$

其中 σ 是通过最大化阶数来选择的。

定义 11 (Adams 方法). Adams 方法 是一个 $\rho(w) = w^{s-1}(w-1)$ 的多步数值方法。

这些方法可以是显式的或隐式的。在不同情况下，我们得到不同的名称。

定义 12 (Adams-Bashforth 方法). Adams-Bashforth 方法是一个显式 *Adams* 方法。

定义 13 (Adams-Moulton 方法). Adams-Moulton 方法是一个隐式 *Adams* 方法。

例 1.4. 我们来看一个两步三阶 *Adams-Moulton* 方法。这由以下公式给出

$$\mathbf{y}_{n+2} - \mathbf{y}_{n+1} = h\left(-\frac{1}{2}\mathbf{f}(t_n, \mathbf{y}_n) + \frac{2}{3}\mathbf{f}(t_{n+1}, \mathbf{y}_{n+1}) + \frac{5}{12}\mathbf{f}(t_{n+1}, \mathbf{y}_{n+2})\right)$$

这些系数从何而来？我们首先要将 $\frac{\rho(w)}{\log w}$ 展开在 $w-1$ 处：

$$\frac{\rho(w)}{\log w} = \frac{w(w-1)}{\log w} = 1 + \frac{3}{2}(w-1) + \frac{5}{12}(w-1)^2 + O(|w-1|^3)$$

这些不是我们 σ 的系数，因为我们需要重新排列前面三项，以便用 w 表示。因此我们有

$$\frac{\rho(w)}{\log w} = -\frac{1}{12} + \frac{2}{3}w + \frac{5}{12}w^2 + O(|w-1|^3)$$

另一类重要的多步方法是以相反的方式构建的——我们选择一个特定的 σ，然后找到最优的 ρ。

定义 14 (反向微分法). 反向微分法具有 $\sigma(w) = \sigma_s w^s$，即

$$\sum_{\ell=0}^{s} \rho_\ell \mathbf{y}_{n+\ell} = \sigma_s \mathbf{f}(t_{n+s}, \mathbf{y}_{n+s})$$

这是一步反向欧拉法的推广。

给定这个 σ，我们需要选择合适的 ρ。幸运的是，这可以很容易地完成。

引理 1. 通过选择

$$\rho(w) = \sigma_s \sum_{\ell=1}^{s} \frac{1}{\ell} w^{s-\ell}(w-1)^\ell$$

可以得到阶数为 s 的 s-步反向微分法，其中 σ_s 选择为使 $\rho_s = 1$，即

$$\sigma_s = \left(\sum_{\ell=1}^{s} \frac{1}{\ell}\right)^{-1}$$

证明. 我们需要构建 ρ 使得

$$\rho(w) = \sigma_s w^s \log w + O(|w-1|^{s+1})$$

这很容易，如果我们写成

$$\begin{aligned}\log w &= -\log\left(\frac{1}{w}\right) \\ &= -\log\left(1 - \frac{w-1}{w}\right) \\ &= \sum_{\ell=1}^{\infty} \frac{1}{\ell}\left(\frac{w-1}{w}\right)^\ell\end{aligned}$$

乘以 $\sigma_s w^s$ 得到所需结果。 □

对于这种方法要收敛，我们需要确保它确实满足根条件。事实证明，根条件仅在 $s \leq 6$ 时满足。这并不直观，但我们可以验证这一点。

1.3 Runge-Kutta 方法

最后，我们来看 Runge-Kutta 方法。这些方法非常复杂，分析起来相当繁琐。它们在很长一段时间内被忽视了，直到更强大的计算机出现，使这些方法更加实用。由于它们具有许多优良特性，现在被广泛使用。

Runge-Kutta 方法可以通过高斯求积来引出，但我们不会深入探讨这种联系。相反，我们直接进入该方法的讨论

定义 15 (Runge-Kutta 方法)**.** *一般（隐式）ν-阶段 Runge-Kutta (RK) 方法 具有以下形式*

$$\mathbf{y}_{n+1} = \mathbf{y}_n + h \sum_{\ell=1}^{\nu} b_\ell \mathbf{k}_\ell$$

其中

$$\mathbf{k}_\ell = \mathbf{f}\left(t_n + c_n h, \mathbf{y}_n + h \sum_{j=1}^{\nu} a_{\ell j} \mathbf{k}_j \right)$$

对于 $\ell = 1, \cdots, \nu$。

我们有很多参数需要选择。我们需要选择

$$\{b_\ell\}_{\ell=1}^{\nu}, \quad \{c\}_{\ell=1}^{\nu}, \quad \{a_{\ell j}\}_{\ell,j=1}^{\nu}$$

注意，通常 $\{\mathbf{k}_\ell\}_{\ell=1}^{\nu}$ 需要被解出来，因为它们是互相定义的。然而，对于某些参数选择，我们可以使其成为显式方法。这使得计算更容易，但我们会失去一些精度和灵活性。

与我们之前看到的所有方法不同，参数出现在 $\mathbf{f}$ 之内。它们非线性地出现在函数内部。这使得方法更加复杂，用泰勒级数分析非常困难。然而，一旦我们正确地完成了这一点，这些方法具有许多优良特性。局限于笔者的时间，没时间去具体讨论

注意这是一种一步法。因此，一旦我们得到阶数 $p \geq 1$，我们就会有收敛性。那么我们需要什么条件才能获得一个合适的阶数呢?

这通常非常复杂。然而，我们可以得到一些必要条件。我们可以考虑 $\mathbf{f}$ 是常数的情况。然后 $\mathbf{k}_e ll$ 总是那个常数。因此我们必须有

$$\sum_{\ell=1}^{\nu} b_\ell = 1$$

事实证明，我们还需要，对于 $\ell = 1, \cdots, \nu$，

$$c_\ell = \sum_{j=1}^{\nu} a_{\ell j}$$

虽然这些是必要条件，但它们不是充分条件。我们还需要其他条件，我们稍后会看到。事实是，最佳的 ν-阶段 Runge-Kutta 方法的阶数是 2ν。

为了描述 Runge-Kutta 方法，标准的记法是将系数放在 *Butcher* 表中:

$$\begin{array}{c|ccc} c_1 & a_{11} & \cdots & a_{1\nu} \\ \vdots & \vdots & \ddots & \vdots \\ c_\nu & a_{\nu 1} & \cdots & a_{\nu\nu} \\ \hline & b_1 & \cdots & b_\nu \end{array}$$

有时我们更简洁地写成

$$\begin{array}{c|c} \mathbf{c} & A \\ \hline & \mathbf{v}^T \end{array}$$

这个表适用于一般的隐式方法。最初，显式方法首先出现，因为它们计算起来要容易得多。在这种情况下，矩阵 A 是严格下三角的，即 $\ell \leq j$ 时 $a_{\ell j} = 0$。

例 1.5. *最著名的显式 Runge-Kutta 方法是四阶段四阶方法，通常称为经典* Runge-Kutta 方法。公式可以明确表示为

$$\mathbf{y}_{n+1} = \mathbf{y}_n + \frac{h}{6}(\mathbf{k}_1 + 2\mathbf{k}_2 + 2\mathbf{k}_3 + \mathbf{k}_4)$$

其中

$$\begin{aligned}
\mathbf{k}_1 &= \mathbf{f}(x_n, \mathbf{y}_n) \\
\mathbf{k}_2 &= \mathbf{f}\left(x_n + \frac{1}{2}h, \mathbf{y}_n + \frac{1}{2}h\mathbf{k}_1\right) \\
\mathbf{k}_3 &= \mathbf{f}\left(x_n + \frac{1}{2}h, \mathbf{y}_n + \frac{1}{2}h\mathbf{k}_2\right) \\
\mathbf{k}_4 &= \mathbf{f}(x_n + h, \mathbf{y}_n + h\mathbf{k}_3)
\end{aligned}$$

我们看到这是一个显式方法。我们不需要解任何方程。

选择 Runge-Kutta 方法的参数以最大化阶数非常困难。考虑最简单的情况，即两阶段显式方法。一般公式是

$$\mathbf{y}_{n+1} = \mathbf{y}_n + h(b_1\mathbf{k}_1 + b_2\mathbf{k}_2)$$

其中

$$\begin{aligned}
\mathbf{k}_1 &= \mathbf{f}(x_n, \mathbf{y}_n) \\
\mathbf{k}_2 &= \mathbf{f}(x_n + c_2 h, \mathbf{y}_n + r_2 h\mathbf{k}_1)
\end{aligned}$$

为了分析这个，我们将真实解插入方法中。首先，我们需要将 ODE 的真实解插入 $\mathbf{k}$ 中。我们得到

$$\begin{aligned}
\mathbf{k}_1 &= \mathbf{y}'(t_n) \\
\mathbf{k}_2 &= \mathbf{f}(t_n + c_2 h, \mathbf{y}(t_n) + c_2 h\mathbf{y}'(t_n)) \\
&= \mathbf{y}'(t_n) + c_2 h\left(\frac{\partial \mathbf{f}}{\partial t}(t_n, \mathbf{y}(t_n)) + \nabla \mathbf{f}(t_n, \mathbf{y}(t_n))\mathbf{y}'(t_n)\right) + O(h^2)
\end{aligned}$$

幸运的是，括号内的内容只是 $\mathbf{y}''(t_n)$。所以这只是

$$= \mathbf{y}'(t_n) + c_2 h\mathbf{y}''(t_n) + O(h^2)$$

因此，Runge-Kutta 方法的局部截断误差是

$$\begin{aligned}
&\mathbf{y}(t_{n+1}) - \mathbf{y}(t_n) - h(b_1\mathbf{k}_1 + b_2\mathbf{k}_2) \\
&\qquad = (1 - b_1 - b_2)h\mathbf{y}'(t_n) + \left(\frac{1}{2} - b_2 c_2\right)h^2\mathbf{y}''(t_n) + O(h^3)
\end{aligned}$$

现在我们看到 Runge-Kutta 方法分析起来为什么困难。系数在这个表达式中非线性地出现。虽然在这种情况下仍然可以以明显的方式解决，但对于更高级的方法，这变得更加复杂。

在这种情况下，我们有一个阶数为 2 的方法族，满足

$$b_1 + b_2 = 1, \quad b_2 c_2 = \frac{1}{2}$$

很容易检查使用简单的方程 $y' = \lambda h$ 不可能得到更高阶的方法。因此，只要我们选择的 b_1 和 b_2 满足这个方程，我们就会得到一个阶数为 2 的好方法。

正如我们所看到的，Runge-Kutta 方法非常复杂，即使在最简单的情况下。然而，它们有太多优良特性，现在变得非常流行。

2 刚性方程

最初，人们在开发数值方法时，主要关注的是阶数和精度等定量特性。然后我们开发了许多不同的方法，如多步法和 Runge-Kutta 方法。

最近，人们开始关注结构特性。通常，方程具有一些特殊特性。例如，描述粒子运动的微分方程很可能能量守恒。当我们用数值方法近似时，我们希望数值近似也能保持能量守恒。这是近年来的发展的方向——我们希望研究数值方法是否保留某些好的性质。

由于复杂度，我们在这里不会研究能量守恒。相反，我们来看以下问题。假设我们有一个 $0 \leq t \leq T$ 的 ODE 系统：

$$\begin{aligned}\mathbf{y}'(t) &= \mathbf{f}(t, \mathbf{y}(t)) \\ \mathbf{y}(0) &= \mathbf{y}_0\end{aligned}$$

假设 $T > 0$ 是任意的，并且

$$\lim_{t\to\infty} \mathbf{y}(t) = \mathbf{0}$$

为了使数值方法满足 $\lim_{n\to\infty} \mathbf{y}_n = \mathbf{0}$，对 h 有什么限制?

这个问题对于我们来说仍然太复杂了。它只能容易地解决线性问题，即以下形式的 ODE

$$\mathbf{y}'(t) = A\mathbf{y}(t)$$

其中 $A \in \mathbb{R}^{N\times N}$。

首先，对于哪些 A，我们有 $\mathbf{y}(t) \to 0$ 当 $t \to \infty$? 根据一些基本线性代数知识，我们知道仅当 A 的所有特征值 λ 满足 $\Re(\lambda) < 0$ 时，这才成立。为了进一步简化，我们考虑

$$y'(t) = \lambda y(t) \quad \Re(\lambda) < 0$$

显然，如果 A 是对角化的，那么可以化简为这种情况的多个实例。否则，我们需要做一些额外工作，但我们在本文中不会涉及。这种简化足够了。

2.1 线性稳定性

我们只考虑问题 $y' = \lambda y$。无论数值方法多么复杂，当应用于这个问题时，通常会变得非常简单。

定义 16 (线性稳定域). 如果我们将数值方法应用于

$$y'(t) = \lambda y(t)$$

且 $y(0) = 1$，$\lambda \in \mathbb{C}$，那么其线性稳定域是

$$D = \left\{ z = h\lambda : \lim_{n\to\infty} y_n = 0 \right\}$$

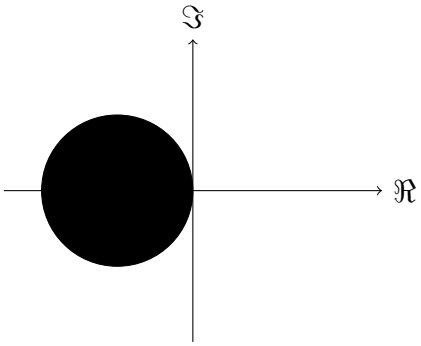
例 2.1. 考虑欧拉法。离散解为

$$y_n = (y_n = (1 + h\lambda)^n$$

因此我们得到

$$D = \{z \in \mathbb{C} : |1 + z| < 1\}$$

我们可以在复平面上将其可视化为单位圆内部的区域：



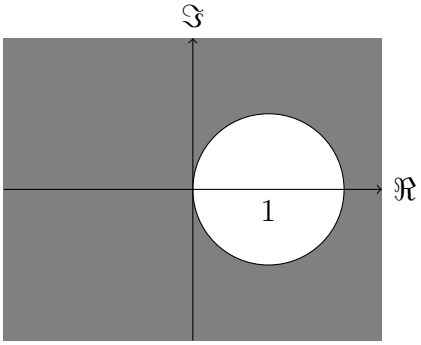
例 2.2. 反向欧拉法的稳定域如何？这个方法是隐式的，但由于我们的问题简单，我们可以找到第 n 步的方法。结果是

$$y_n = (1 - \lambda h)^{-n}$$

然后我们得到

$$D = \{z \in \mathbb{C} : |1 - z| > 1\}$$

我们可以将其可视化为：



我们做如下定义：

定义 17 (A-稳定性). 如果数值方法满足

$$\mathbb{C}^- = \{z \in \mathbb{C} : \Re(z) < 0\} \subseteq D$$

则称该方法是 A-稳定的。

特别地，对于 $\Re(z) < 0$，A-稳定性意味着无论 h 多大，y_n 都会趋向于 0。

因此反向欧拉法是 A-稳定的，而欧拉法不是。

A-稳定性是一个非常强的要求。要获得 A-稳定性非常困难。特别地，Dahlquist 证明没有阶数 $p \geq 3$ 的多步方法是 A-稳定的。此外，没有显式 Runge-Kutta 方法可以是 A-稳定的。

让我们看看一些其他的隐式方法。

例 2.3 (梯形法). 再次考虑 $y'(t) = \lambda y$，使用梯形法。然后我们可以找到

$$y_n = \left(\frac{1 + h\lambda/2}{1 - h\lambda/2}\right)^n$$

因此，线性稳定域是

$$D = \left\{z \in \mathbb{C} : \left|\frac{2 + z}{2 - z}\right| < 1\right\}$$

这意味着 z 必须比 -2 更接近 -2。换句话说，D 正好是 $\mathbb{C}^-$。

一般来说，测试数值方法的 A-稳定性时，复分析是有帮助的。通常，当将数值方法应用于问题 $y' = \lambda y$ 时，我们得到

$$y_n = [r(h\lambda)]^n$$

其中 r 是某个有理函数。因此

$$D = \{z \in \mathbb{C} : |r(z)| < 1\}$$

我们想知道 D 是否包含左半平面。对于更复杂的 r 表达式，如梯形法的情况，这并不明显。幸运的是，我们有最大值原理：

定理 5 (最大值原理)**.** *设 g 在开集 $\Omega \subseteq \mathbb{C}$ 中解析且非常数。则 $|g|$ 在 Ω 中没有最大值。*

由于 $|g|$ 需要在 Ω 的闭包中达到最大值，因此最大值必须出现在边界上。因此，要证明 $|g| \leq 1$ 在区域 Ω 上成立，我们只需证明该不等式在边界 $\partial\Omega$ 上成立。

我们试试 $\Omega = \mathbb{C}^-$. 技巧是首先检查 g 在 Ω 中是否解析，然后检查在左半平面的边界上会发生什么。这个技术在以下例子中清楚地说明了：

例 2.4. *考虑*

$$r(z) = \frac{6 - 2z}{6 - 4z + z^2}$$

这仍然相对简单，但可以说明如何使用最大值原理。

我们首先检查它是否解析。这个函数肯定有一些极点，但它们是 $2 \pm \sqrt{2}i$，在右半平面。因此它在 $\mathbb{C}^-$ 中解析。

接下来，在左半平面的边界上会发生什么？首先，当 $|z| \to \infty$ 时，我们发现 $r(z) \to 0$，因为分母有 z^2 项。接下来，检查 z 在虚轴上，即 $z = it$，其中 $t \in \mathbb{R}$。然后我们可以通过一些复杂的代数证明

$$|r(it)| \leq 1$$

对于 $t \in \mathbb{R}$。因此，根据最大值原理，我们必须有 $|r(z)| \leq 1$ 对所有 $z \in \mathbb{C}^-$ 成立。

3 ODE 方法的实现

我们刚刚讨论了许多数值方法的理论。为了结束这一节，我们将探讨一些 ODE 方法的实践方面。

3.1 局部误差估计

我们想要解决的第一个问题是选择什么 h。通常，当使用数值分析软件时，你会被要求给出误差容限，然后软件会自动计算我们需要的 h。这是怎么做到的?

Milne 方法是一种估计多步法的局部截断误差并因此改变步长 h 的方法 (对于 Runge-Kutta 方法有类似的技术，但更复杂)。这使用两个相同阶数的多步法。

为了简化，我们考虑二步 Adams-Bashforth 方法。回忆一下，这是

$$\mathbf{y}_{n+1} = \mathbf{y}_n + \frac{h}{2}(2\mathbf{f}(t_n, \mathbf{y}_n) - \mathbf{f}(t_{n-1}, \mathbf{y}_{n-1}))$$

这是一个二阶误差方法，具有

$$\boldsymbol{\eta}_{n+1} = \frac{5}{12}h^3\mathbf{y}'''(t_n) + O(h^4)$$

另一个二阶多步法是我们所熟悉的梯形法。这是一个隐式方法

$$\mathbf{y}_{n+1} = \mathbf{y}_n + \frac{h}{2}(\mathbf{f}(t_n, \mathbf{y}_n)\ \mathbf{f}(t_{n+1}, \mathbf{y}_{n+1}))$$

它的局部截断误差是

$$\boldsymbol{\eta}_{n+1} = -\frac{1}{12}h^3\mathbf{y}'''(t_n) + O(h^4)$$

Milne 方法的关键是 $h^3\mathbf{y}'''(t_n)$ 的系数，即

$$c_{\text{AB}} = \frac{5}{12}, \quad c_{\text{TR}} = -\frac{1}{12}$$

由于这些是两种不同的方法，我们得到不同的 $\mathbf{y}_{n+1}$。我们用上标区分这些，并且有

$$\mathbf{y}(t_{n+1}) - \mathbf{y}_{n+1}^{\mathrm{AB}} \simeq c_{\mathrm{AB}} h^3 \mathbf{y}'''(t_n)$$
$$\mathbf{y}(t_{n+1}) - \mathbf{y}_{n+1}^{\mathrm{TR}} \simeq c_{\mathrm{TR}} h^3 \mathbf{y}'''(t_n)$$

我们现在可以消去一些项，得到

$$\mathbf{y}(t_{n+1}) - \mathbf{y}_{n+1}^{\mathrm{TR}} \simeq \frac{-c_{\mathrm{TR}}}{c_{\mathrm{AB}} - c_{\mathrm{TR}}} (\mathbf{y}_{n+1}^{\mathrm{AB}} - \mathbf{y}_{n+1}^{\mathrm{TR}})$$

在这个例子中，常数是 $\frac{1}{6}$。因此我们可以估计梯形法的局部截断误差，而不需要知道 $\mathbf{y}'''$ 的值。然后我们可以使用这个来相应地调整 h。

我们需要做的额外工作是使用两种方法计算数值近似值。通常，当我们想要估计一个更复杂但更好的方法的误差时，我们会使用一个简单的显式方法，如 Adams-Bashforth 方法，作为第二种方法。

3.2 解隐式方法

隐式方法通常更可能保留诸如能量守恒等好的性质。由于我们现在拥有更多的计算能力，使用这些更复杂的方法往往是更可取的。当使用这些隐式方法时，我们必须想出一些方法来求解涉及的方程。

例如，我们考虑反向欧拉法

$$\mathbf{y}_{n+1} = \mathbf{y}_n + h\mathbf{f}(t_{n+1}, \mathbf{y}_{n+1})$$

有两种方法可以求解 $\mathbf{y}_{n+1}$。最简单的方法是函数迭代。顾名思义，这种方法是迭代的。因此我们使用上标表示迭代。在这种情况下，我们使用公式

$$\mathbf{y}_{n+1}^{(k+1)} = \mathbf{y}_n + h\mathbf{f}(t_{n+1}, \mathbf{y}_{n+1}^{(k)})$$

为了进行这个迭代，我们需要一个初始值 $\mathbf{y}_{n+1}^{(0)}$。通常，我们将 $\mathbf{y}_{n+1}^{(0)} = \mathbf{y}_n$。更好的方法是使用一些更简单的显式方法来获得 $\mathbf{y}_{n+1}^{(0)}$ 的初始猜测。

问题是，这个迭代是否收敛？幸运的是，如果 $\mathbf{f}$ 的 Lipschitz 常数 λ 满足 λh 足够小，则该迭代收敛到局部唯一解。对于反向欧拉法，我们需要 $\lambda h < 1$。这需要用到收缩映射定理。

这重要吗？有时是的。通常，我们根据精度考虑选择 h，选择最大可能的 h 来满足所需的精度。然而，如果我们使用这种方法，我们可能需要选择一个更小的 h 以使其工作。这将需要我们计算更多的迭代，并可能花费很多时间。

另一种选择是牛顿法。公式为

$$(I - hJ^{(k)})\mathbf{z}^{(k)} = \mathbf{y}_{n+1}^{(k)} - (\mathbf{y}_n + h\mathbf{f}(t_{n+1}, \mathbf{y}_{n+1}^{(k)}))$$
$$\mathbf{y}_{n+1}^{(k+1)} = \mathbf{y}_{n+1}^{(k)} - \mathbf{z}^{(k)}$$

其中 $J^{(k)}$ 是雅可比矩阵

$$J^{(k)} = \nabla\mathbf{f}(t_{n+1}, \mathbf{y}_{n+1}^{(k)}) \in \mathbb{R}^{N\times N}$$

这需要我们在第一个方程中求解 $\mathbf{z}$，但这是一个线性系统，我们有一些有效的方法来求解。

牛顿法有几种变体。这是全牛顿法，我们在每次迭代中重新计算雅可比矩阵。也可以使用相同的雅可比矩阵反复计算。这在求解方程时有一些速度上的提升，但我们需要更多的迭代才能得到 $\mathbf{y}_{n+1}$。