

数值分析

李治平

北京大学
数学科学学院

Spring, 2013



- 误差: 模型、数据、方法 (离散)、计算 (舍入)。
- 稳定性:
 - ① 原始问题及数学模型的稳定性 (敏度分析, 病态、良态);
 - ② 算法稳定性 (精确计算时, 小数据误差 \Rightarrow 小结果误差);
 - ③ 数值稳定性 (舍入误差对结果的影响, 向后误差分析)。
- 要得到可靠的计算结果, 我们需要有相对
 - ① 比较良态的原始问题及数学模型;
 - ② 有较好的稳定性和数值稳定性的算法;
 - ③ 较小的数据误差和方法 (离散) 误差。
- 计算方法的评判标准—快、省、准 (实践是检验真理的唯一标准)。



截断误差与舍入误差

- **截断误差**：数学模型的真解通常无法通过有限次运算获得。为此，需要设计能够通过有限次运算得到结果的离散模型（算法）来获得近似解。由此产生的离散模型与连续模型间的误差称为截断误差。
- **舍入误差**：计算机的字长有限，数值计算结果一般只能按某种指定的方式舍入。例如在一定有效位后的四舍五入。
- 对误差的理解、分析和控制是计算科学最基本的内容之一。



四则运算结果的误差限与数据误差限的关系

两个数 a 和 b 的四则运算的误差满足：

- 运算结果的绝对误差： $e(a \pm b) \leq e(a) + e(b)$,
 $e(ab) \leq |b|e(a) + |a|e(b)$, $e(a/b) \leq e(a)/|b| + |a/b^2|e(b)$.

- 运算结果的相对误差：

$$e_r(a \pm b) \leq \frac{e(a) + e(b)}{|a \pm b|} \leq \frac{|a|e_r(a) + |b|e_r(b)}{|a \pm b|},$$

$$e_r(ab) \leq e_r(a) + e_r(b),$$

$$e_r(a/b) \leq e_r(a/b)/|a/b| \leq e_r(a) + e_r(b).$$

- 绝对误差控制：避免小除数。
- 相对误差控制：避免相近数相减。



在浮点数系统中有必要调整四则运算的顺序以减小误差

- 浮点数系统中只有有限个关于原点对称非均匀分布的数；
- 绝对值越大的两个相邻的浮点数间的间隔为也越大。
- 浮点数系中四则运算不满足实数系中的法则（如结合律）。
- 数学上等价的公式在浮点数系中的计算结果可能是不同的。
- 适当调整四则运算的顺序可以将误差控制在合理的范围内。



函数的多项式逼近问题的提法

- 插值逼近：找一 $(m+1)(k+1) - 1$ 次多项式 $p(x)$, 使 $p^{(i)}(x_j) = f^{(i)}(x_j)$, $i = 0, \dots, m, j = 0, \dots, k$.
- 最小二乘逼近：找一 $(m+1)(k+1) - 1$ 次多项式 $p(x)$, 使 $p^{(i)}(x_j) - f^{(i)}(x_j)$, $i = 0, \dots, m, j = 0, \dots, k+n$ 个点上的 l^2 -范数下的误差达到最小 (其中 $n \geq 1$).
- 最佳逼近：找一 k 次多项式使其在指定的范数下与 f 的误差达到最小 (常用的范数包括 L^∞, L^2 , 加权的 L^2 等).



函数的 Lagrange 型多项式插值方法

- Lagrange 插值法: $L_k(x) = \sum_{j=0}^k f(x_j)l_j(x)$, $l_j(x) = \frac{\prod_{i \neq j}(x-x_i)}{\prod_{i \neq j}(x_j-x_i)}$.
- Newton 插值: $N_k(x) = f[x_0] + \cdots + f[x_0, \cdots, x_k] \prod_{i=0}^{k-1}(x-x_i)$.
- 截断误差: $f[x_0, \cdots, x_k, x] \prod_{i=0}^k(x-x_i) = \frac{f^{(k+1)}(\xi)}{(k+1)!} \prod_{i=0}^k(x-x_i)$.
证明要点: $\phi(z) = f(z) - L_k(z) - \frac{f(x) - L_k(x)}{\prod_{i=0}^k(x-x_i)} \prod_{i=0}^k(z-x_i)$.
- Runge 现象与分段低次多项式插值.



函数的 Hermite 型多项式插值和样条插值方法

- 分段三次Hermite 插值法: $H_h(x) = \sum_{i=0}^n [y_i h_i(x) + m_i \hat{h}(x)].$

截断误差: $f(x) - H_h(x) = \frac{f^{(4)}(\xi)}{4!} (x - x_i)^2 (x - x_{i+1})^2.$

证明要点: $\phi(z) = f(z) - H_h(z) - \frac{f(x) - H_h(x)}{\prod_{i=0}^n (x - x_i)^2} \prod_{i=0}^n (z - x_i)^2.$

- 三次样条插值: $S_3(x) = \sum_{i=0}^n [y_i h_i(x) + m_i \hat{h}(x)],$ 取

m_i 得使 $S_3(x) \in C^2.$ (固支、自然、周期三种边界条件).

截断误差: $\|S_3^{(k)} - f^{(k)}\|_\infty \leq C_k \|f^{(4)}\|_\infty h^{4-k}, k = 0, 1, 2, 3.$

极小性质: $\int_a^b |g''(x)|^2 dx \geq \int_a^b |S_3''(x)|^2 dx$ (自然边条件, 或 $g'(x_0) = S_3'(x_0)$ 且 $g'(x_n) = S_3'(x_n)$ 时).



多项式逼近与样条逼近问题的提法

- 设 $f(x)$ 是定义在区间 $[a, b]$ 上的可微函数。
- 对任意给定的正整数 n , 找 $P_0 \in \mathbb{P}_n$, 使得

$$\|P_0(x) - f(x)\| = \min_{P \in \mathbb{P}_n} \|P(x) - f(x)\|,$$

L^∞ -范数(最佳一致逼近); L^2 -范数(最佳平方逼近)。

- 找 $S_m \in \mathbb{S}_m(x_0, x_1, \dots, x_n)$, 使得 S_m 在特定的意义下逼近 f , 其中 $\mathbb{S}_m(x_0, x_1, \dots, x_n)$ 为区间 $[a, b]$ 的剖分 $a = x_0 < x_1 < \dots < x_n = b$ 上定义的所有 m 次样条函数构成的线性函数空间. 例如, $S_m(x) = \sum_{j=-m}^{n-1} a_j B_{j,m}(x)$, 其中 $B_{j,m}(x)$ 为一组B-样条基底函数。



多项式函数的一致逼近性

Weierstrass 定理: 设 $f(x)$ 是闭区间 $[a, b]$ 上的连续函数。则对任给的 $\varepsilon > 0$, 存在 $N(\varepsilon) > 0$, 只要 $n \geq N(\varepsilon) > 0$, 就可以找到 n 次多项式 $P_n(x)$ 使得

$$\|f - P_n\|_\infty = \|f - P_n\|_{\infty, [a, b]} \triangleq \max_{a \leq x \leq b} |f(x) - P_n(x)| < \varepsilon.$$

令 $E_n \triangleq \inf_{P \in \mathbb{P}_n} \Delta(P)$, 称为 \mathbb{P}_n 对 $f(x)$ 的最小偏差或最佳逼近.

Borel 存在定理: 对任给的 $f(x) \in \mathbb{C}[a, b]$, 存在 $P^* \in \mathbb{P}_n$, 使得

$$\Delta(P^*) = E_n.$$

Chebyshev 定理: $f \in \mathbb{C}[a, b]$ 在 \mathbb{P}_n 中存在唯一的最佳一致逼近多项式, 且 $P(x) \in \mathbb{P}_n$ 为 f 的最佳一致逼近多项式的充要条件是 $\varepsilon(x) = P(x) - f(x)$ 在 $[a, b]$ 中个数不少于 $n+2$ 的点列 $x_1 < x_2 < \cdots < x_N$, ($N \geq n+2$) 上正负交错地取到 $\Delta(P)$.

定理: 最小零偏差多项式为 $2^{1-n} T_n(x)$, 其中 $T_n(x)$ 为 Chebyshev 多项式

$$T_n(x) = \cos(n \arccos(x)).$$

最小二乘多项式拟合问题的提法与求解方法

给定数据 $\{(x_i, y_i)\}_{i=0}^n$,

- 选 \mathbb{P}_k 为有限维逼近函数空间, $k < n$ (一般小很多);
- 在 \mathbb{R}^{n+1} 的欧式距离意义下寻找最佳逼近函数;

即

$$\begin{cases} \text{找 } P(x) \in \mathbb{P}_k, \text{ s.t.} \\ I(P) \triangleq \sum_{i=0}^n [y_i - P(x_i)]^2 = \min_{P^* \in \mathbb{P}_k} I(P^*). \end{cases}$$

- 以 $\{x^i\}_{i=0}^k$ 为基底时, 最小二乘 k 次多项式拟合问题的法方程组:

$$\sum_{j=0}^k \left(\sum_{i=0}^n x_i^{j+l} \right) a_j = \sum_{i=0}^n y_i x_i^l, \quad l = 0, 1, \dots, k.$$

如果所选基底 $\{b_j(x)\}_{j=0}^k \subset \mathbb{P}_k$ 有以下形式的规范正交性 $\sum_{i=0}^n b_j(x_i) b_l(x_i) = \delta_{jl}$, 则法方程组的矩阵是单位阵。



最佳平方逼近多项式问题的提法与求解方法

- 设给定的数据点不是有限个，而是在区间 $[a, b]$ 上定义的平方可积函数 $f(x)$.
- 有限维逼近函数空间仍取作 \mathbb{P}_k .
- 在 $L^2[a, b]$ 的意义下寻找最佳逼近函数。

即

$$\begin{cases} \text{找 } P(x) \in \mathbb{P}_k, \text{ s.t.} \\ I(P) \triangleq \int_a^b |f(x) - P(x)|^2 dx = \min_{P^* \in \mathbb{P}_k} I(P^*). \end{cases}$$

- 类似于最小二乘多项式拟合问题，最佳平方逼近多项式问题可以导出法方程组。如果所选基底 $\{b_j(x)\}_{j=0}^k \subset \mathbb{P}_k$ 有以下形式的规范正交性 $\int_a^b b_j(x)b_l(x) dx = \delta_{jl}$ ，则法方程组的矩阵是单位阵。



常用的正交多项式系

- ① 区间 $[-1, 1]$, $\rho(x) \equiv 1$, 的正交多项式系, Legendre 多项式:

$$\begin{cases} \varphi_0(x) = 1, & \varphi_1(x) = x, \\ \varphi_{k+1}(x) = \frac{2k+1}{k+1}x\varphi_k(x) - \frac{k}{k+1}\varphi_{k-1}(x), & k = 1, 2, \dots \end{cases}$$

- ② 区间 $[-1, 1]$, $\rho(x) = \frac{1}{\sqrt{1-x^2}}$ 的 Chebyshev 多项式:

$$\begin{cases} T_0(x) = 1, & T_1(x) = x, \\ T_{k+1}(x) = 2xT_k(x) - T_{k-1}(x), & k = 1, 2, \dots \end{cases}$$

- ③ 区间 $[0, +\infty)$, $\rho(x) = \exp(-x)$ 的 Laguerre 多项式:

$$\begin{cases} Q_0(x) = 1, & Q_1(x) = 1 - x, \\ Q_{k+1}(x) = (1 + 2k - x)Q_k(x) - k^2Q_{k-1}(x), & k = 1, 2, \dots \end{cases}$$

- ④ 区间 $(-\infty, +\infty)$, $\rho(x) = \exp(-x^2)$ 的 Hermite 多项式:

$$\begin{cases} H_0(x) = 1, & H_1(x) = 2x, \\ H_{k+1}(x) = 2xH_k(x) - 2kH_{k-1}(x), & k = 1, 2, \dots \end{cases}$$

有理插值的定义和有理插值问题的提法与解法

定义： 给定 $m + n + 1$ 组数据 $(x_i, f(x_i))$, $i = 0, 1, \dots, m + n$, 若存在 $R_{m,n}(x) \in \mathbb{R}(m, n)$ 使得

$$R_{m,n}(x_i) = f(x_i), \quad i = 0, 1, \dots, m + n,$$

则称 $R_{m,n}(x)$ 为所给数据组 $(x_i, f(x_i))$, $i = 0, 1, \dots, m + n$ 的有理插值。

有理插值问题： 给定数据组 $\{(x_i, f(x_i))\}_{i=0}^{m+n}$, 在 $\mathbb{R}(m, n)$ 中找有理分式 $R_{m,n}(x) = \frac{P_m(x)}{Q_n(x)}$, 使得 $R_{m,n}(x_i) = f(x_i)$, $0 \leq i \leq m + n$.

混合连分式方法：

$$R_{k,n}(x) = v_0(x_0) + \frac{x - x_0}{v_1(x_1) + \frac{x - x_1}{\dots + \frac{x - x_{k-1}}{v_k(x_k) + \frac{x - x_k}{g_{n-k-1}(x)}}}}$$

注： $\mathbb{R}(m, n)$ 中的有理插值函数的存在性要求数据组满足一定的相容性条件. 若不特别限定 m 和 n , 则总可以通过混合连分式方法求解。



数值微分问题的典型提法

设 $f(x)$ 在区间 $[a, b]$ 上定义, $a \leq x_0 < x_1 < \cdots < x_n \leq b$ 是区间 $[a, b]$ 中的 $n+1$ 个节点。数值微分问题的典型提法是: 利用函数在所给节点上的值

$$f(x_0), f(x_1), f(x_2), \cdots, f(x_n),$$

求

$$f'(x_0), f'(x_1), f'(x_2), \cdots, f'(x_n),$$

$$f''(x_0), f''(x_1), f''(x_2), \cdots, f''(x_n),$$

或更一般地, 对 $1 \leq k \leq n$, 求

$$f^{(k)}(x_0), f^{(k)}(x_1), f^{(k)}(x_2), \cdots, f^{(k)}(x_n),$$

的近似值。



数值微分的典型算法

- Taylor 展开法: 例如 $f'(x_i) = \frac{f(x_{i+1}) - f(x_i)}{h} - \frac{h}{2} f''(\xi_{i1}^2)$.
- 差商近似法: 例如一阶向前差商: $f'(x_i) \approx \frac{f(x_{i+1}) - f(x_i)}{h}$.
- 隐式方法: 如 $f'(x_{i-1}) + 4f'(x_i) + f'(x_{i+1}) \approx \frac{3}{h}(f_{i+1} - f_{i-1})$.
- 插值型方法: 例如 $f'(x_i) = P'_n(x_i) + \frac{f^{(n+1)}(\xi(x_i))}{(n+1)!} \omega'_{n+1}(x_i)$.



截断误差、舍入误差及数值微分的理论极限

- 截断误差: 例如对一阶向前差商公式有 $T_h = -\frac{h}{2}f''(\xi_{i1}^2)$.
- 舍入误差: 设 $f(x_i)$ 和 $f(x_{i+1})$ 分别有舍入误差 ε_1 和 ε_2 , 令 $\varepsilon = \max\{|\varepsilon_1|, |\varepsilon_2|\}$, 令

$$f'_h(x_i) \triangleq \frac{f(x_{i+1}) - f(x_i)}{h}, \quad \tilde{f}'_h(x_i) \triangleq \frac{\tilde{f}(x_{i+1}) - \tilde{f}(x_i)}{h},$$

则由舍入误差带来的计算结果的误差限为

$$\delta(f'_h(x_i)) = |f'_h(x_i) - \tilde{f}'_h(x_i)| \leq \frac{|\varepsilon_1| + |\varepsilon_2|}{h} \leq \frac{2\varepsilon}{h}.$$

- 数值微分的理论极限: $e_h(f'(x_i)) = \frac{h}{2} \max_{x \in (x_i, x_{i+1})} |f''(x)| + \frac{2\varepsilon}{h}$ 当 $h = 2\sqrt{\varepsilon / \max_{x \in (x_i, x_{i+1})} |f''(x)|}$ 时最小 $2\sqrt{\varepsilon \max_{x \in (x_i, x_{i+1})} |f''(x)|}$.



数值积分问题的提法

设 $f(x)$ 在区间 $[a, b]$ 上定义, $a \leq x_0 < x_1 < \cdots < x_n \leq b$ 是区间 $[a, b]$ 中的 $n + 1$ 个已知或待定的节点。数值积分问题的典型提法是: 利用函数在所给节点上的值

$$f(x_0), f(x_1), f(x_2), \cdots, f(x_n),$$

求定积分

$$\int_a^b f(x) dx$$

的近似值。



中点公式、梯形公式、Simpson 公式和 Newton-Cotes 公式

- 求积公式 $I(f) = \int_a^b f(x) dx \approx I_n(f) = \sum_{k=0}^n A_k f(x_k)$, 其中 $A_k, x_k, k = 0, \dots, n$, 与 f 无关, 称为求积公式的权和节点;
 - 中点公式($n=0$): $\int_a^b f(x) dx \approx f\left(\frac{a+b}{2}\right)(b-a)$;
 - 梯形公式($n=1$): $\int_a^b f(x) dx \approx \frac{f(a)+f(b)}{2}(b-a)$;
 - Simpson 公式($n=2$): $\int_a^b f(x) dx \approx \frac{b-a}{6}\left[f(a)+4f\left(\frac{a+b}{2}\right)+f(b)\right]$.
 - Newton-Cotes 公式(等距): $A_k = \frac{(-1)^{n-k}(b-a)}{k!(n-k)!n} \int_0^n \prod_{j \neq k}^n (t-j) dt$.
- 记 $h = b-a$, 则当被积函数充分光滑时, 由插值误差估计可证
 - 中点公式的截断误差为 $O(h^3)$;
 - 梯形公式的截断误差为 $O(h^3)$;
 - Simpson 公式的截断误差为 $O(h^5)$.
- Newton-Cotes 公式的截断误差为 $O(h^{n+2})$ (n 奇), $O(h^{n+3})$ (n 偶)



插值型求积公式及其特征

定义： 如果 $E_n(f) = I(f) - I_n(f) = 0, \forall f \in \mathbb{P}_n$, 则称数值积分公式 $I_n(f)$ 为插值型的。

定理： 以下两个命题是等价的：

- ① $I_n(f)$ 为插值型的数值积分公式.
- ② $I_n(f)$ 可由对 x_0, x_1, \dots, x_n 的 n 次插值多项式的积分得到.

定理： 对 $k \geq 0$, 数值积分公式 $I_n(f) = \sum_{i=0}^n A_i f(x_i)$ 的代数精度为 $d = n + k$ 阶的充分必要条件是：

- ① $I_n(f)$ 为插值型的数值积分公式;
- ② $\omega_n(x) = \prod_{i=0}^n (x - x_i)$ 满足 $\int_a^b \omega_n(x) P(x) dx = 0, \forall P(x) \in \mathbb{P}_{k-1}$;
- ③ $\exists Q(x) \in \mathbb{P}_k, \text{ s.t. } \int_a^b \omega_n(x) Q(x) dx \neq 0$.

推论： 数值积分公式 $I_n(f) = \sum_{i=0}^n A_i f(x_i)$ 的代数精度最多为 $2n + 1$ 阶, 即 $0 \leq k \leq n + 1$, 因为总有 $\int_a^b \omega^2(x) dx > 0$.

Gauss 型求积公式

定义： 对给定的节点数 n , 代数精度达到最大值 $2n - 1$ 的求积公式称为 *Gauss* 型求积公式.

Gauss 型求积公式权和节点 $A_k, x_k, k = 0, \dots, n$ 的计算:

- 按定义列非线性方程组, 求解得。
- 设 $\{\Phi_k(x)\}_{k=0}^{\infty}$ 为正交多项式序列, 则 $\Phi_{n+1}(x)$ 的 $n+1$ 个零点即为相应 *Gauss* 型求积公式的节点。

定理： 设 $\{x_k\}_{k=0}^n$ 是区间 $[a, b]$ 上, 积分权为 $\rho(x)$ 的 $n+1$ 次正交多项式的零点. 记 $\omega_{n+1}(x) = \prod_{k=0}^n (x - x_k)$. 则当 $f(x) \in C^{2n+2}[a, b]$ 时, 存在 $\xi \in (a, b)$, 使得 *Gauss* 求积公式 $\int_a^b f(x) dx \approx \sum_{k=0}^n A_k f(x_k)$ 的截断误差可以表示为

$$E(f) = \frac{f^{(2n+2)}(\xi)}{(2n+2)!} \int_a^b \rho(x) [\omega_{n+1}(x)]^2 dx.$$



复合中点公式、复合梯形公式和复合 Simpson 公式

在 $[a, b]$ 上引进等距节点 $x_i = a + ih, i = 0, 1, \dots, n$,
 $h = (b - a)/n$, 并记 $x_{i+\frac{1}{2}} = a + (i + \frac{1}{2})h$, 在每个子区间 $[x_i, x_{i+1}]$
上分别使用中点公式、梯形公式和 Simpson 公式即得

$$\text{复合中点公式: } \int_a^b f(x) dx \approx h \sum_{i=0}^n f(x_{i+\frac{1}{2}}) \triangleq M(h);$$

$$\text{复合梯形公式: } \int_a^b f(x) dx \approx \frac{h}{2} \sum_{i=0}^n [f(x_i) + f(x_{i+1})] \triangleq T(h);$$

$$\text{复合 Simpson 公式: } \int_a^b f(x) dx \approx \frac{h}{6} \sum_{i=0}^n [f(x_i) + 4f(x_{i+\frac{1}{2}}) + f(x_{i+1})] \triangleq S(h).$$

注: M for Middle-point, T for Trapezoidal and S for Simpson.



复合梯形公式的后验误差估计与 Richardson 外推加速

$$\text{由} \quad \int_a^b f(x)dx - T(h) = -\frac{h^2}{12} \int_a^b f''(x)dx + O(h^3),$$

$$\int_a^b f(x)dx - T(h/2) = -\frac{(h/2)^2}{12} \int_a^b f''(x)dx + O(h^3),$$

$$\text{得} \quad \frac{T(h/2) - T(h)}{3} = -\frac{(h/2)^2}{12} \int_a^b f''(x)dx + O(h^3),$$

$$\text{即} \quad \int_a^b f(x)dx - T(h/2) = \frac{T(h/2) - T(h)}{3} + O(h^3).$$

$$\text{所以} \quad \int_a^b f(x)dx - \frac{4T(h/2) - T(h)}{3} = O(h^3).$$

特别的有：1. 基于 Euler-Maclaurin 公式和 Richardson 外推加速收敛技术的复合梯形公式的 Romberg 求积方法；2. 自适应求积方法。



积分方程数值解问题的导出与数值解的误差分析

积分方程的数值求解通常包括以下步骤

- 积分的离散化: 用适当数值积分代替积分;
- 函数空间的离散化: 用 $\{y(t_i)\}_{i=0}^n$ 代替 $y(t)$, $t \in [a, b]$;
- 积分方程的离散化: 将积分方程化为有限阶代数方程;
- 离散问题求解: 通过数值求解代数方程得到 $\{y(t_i)\}_{i=0}^n$;
- 给出积分方程近似解: 通过数据拟合 (如分段线性插值).

积分方程数值的误差分析:

- 离散格式的截断误差;
- 离散格式的稳定性: 离散格式的矩阵的逆的一致有界性;
- 离散格式的数值稳定性: 离散格式的矩阵的条件数的一致有界性.



非线性方程数值迭代求解问题

- 非线性方程：找 $x^* \in \mathbb{R}$ ，使得 $f(x^*) = 0$ ，其中 $f : \mathbb{R} \rightarrow \mathbb{R}$ 为非线性函数.
- 非线性方程组：找 $x^* \in \mathbb{R}^n$ ， $n > 1$ ，使得 $f(x^*) = 0$ ，其中 $f : \mathbb{R}^n \rightarrow \mathbb{R}^n$ 为非线性函数.

非线性方程和非线性方程组求解的迭代：给定初值 x_0 ，产生迭代序列

$$x_{k+1} = \varphi_k(x_k), \quad k = 0, 1, 2, \dots$$

判别算法优劣的准则包括收敛速度、计算量、稳定性和稳健性.

定义： 设序列 $\{x_k\}_{k=0}^{\infty}$ 收敛于 x^* . 记 $e_k = x_k - x^*$ ，如果正数 r 能够保证存在非负常数 C ，使得以下极限成立

$$\lim_{k \rightarrow \infty} \frac{\|e_{k+1}\|}{\|e_k\|^r} = C$$

则称序列 $\{x_k\}_{k=0}^{\infty}$ 收敛到 x^* 的速度是 r 阶的.



常用的迭代法及其收敛速度

- 线性收敛的稳健的二分法，缺点是难于推广至高维；
- 基于压缩映像原理的迭代法，要求 $|\varphi'(x^*)| < 1$. 若 $\varphi^{(m)}(x^*) = 0, 1 \leq m < r$, 但 $\varphi^{(r)}(x^*) \neq 0$, 且 $\varphi(x)$ 在 x^* 的邻域中是 \mathbb{C}^r 的, 则不动点迭代序列是 r 阶收敛的, 缺点是技巧性较强；
- 二阶收敛的 Newton 法(又名切线法), 缺点是需要算导数且常常收敛域较小；
- $\frac{1+\sqrt{5}}{2} \approx 1.618$ 阶收敛的割线法和 1.84 阶收敛的抛物线法, 缺点是难于推广至高维；



非线性方程组的常用迭代法

- 基于压缩映像原理的迭代法, 要求 $\|\nabla\varphi(x^*)\| < 1$.
- 一定条件下收敛的非线性 Jacobi 迭代、非线性 Gauss-Seidel 迭代、非线性 SOR 迭代, 缺点是一般收敛较慢;
- 二阶收敛的 Newton 法, 缺点是每步需要算梯度矩阵, 且常常收敛域较小;
- 一定条件下局部超线性收敛的拟 Newton 法, 如 Broyden 算法, 缺点是分析困难;
- 同伦算法, 缺点是技巧性强, 且分析困难.



离散 Fourier 变换 $\mathbf{c} = \hat{\mathbf{a}}$ 和逆变换 $\mathbf{a} = \check{\mathbf{c}}$

令 $\omega \triangleq e^{-i\frac{2\pi}{N}}$, 定义 Fourier 矩阵

$$\mathbf{F} = (f_{kj})_{k,j=0}^{N-1} \triangleq (\omega^{kj})_{k,j=0}^{N-1} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1 & \omega & \cdots & \omega^{N-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \omega^{N-1} & \cdots & \omega^{(N-1)^2} \end{pmatrix}.$$

则 $\mathbf{c} = \hat{\mathbf{a}} = \mathbf{F}\mathbf{a}$, $\mathbf{a} = \check{\mathbf{c}} = \mathbf{F}^{-1}\mathbf{c}$, 其中 $\mathbf{F}^{-1} = \mathbf{G}$ 为以下矩阵:

$$\mathbf{G} \triangleq \frac{1}{N}(\omega^{-jk})_{j,k=0}^{N-1} = \frac{1}{N} \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1 & \omega^{-1} & \cdots & \omega^{-(N-1)} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \omega^{-(N-1)} & \cdots & \omega^{-(N-1)^2} \end{pmatrix}.$$



离散 Fourier 变换和逆变换与多项式插值之间的关系

令 $P(x) = a_0 + a_1x + a_2x^2 + \cdots + a_{N-1}x^{N-1}$, 则由 $\mathbf{c} = \mathbf{F}\mathbf{a}$ 知

- 求向量 $\mathbf{a} = (a_0, a_1, \cdots, a_{N-1})^T$ 的离散 Fourier 变换 $\hat{\mathbf{a}} = (c_0, c_1, \cdots, c_{N-1})^T$ 相当于求多项式 $P(x)$ 在 $\omega^0, \omega^1, \cdots, \omega^{N-1}$ 这 N 个点上的值.

令 $Q(x) = \frac{1}{N}(c_0 + c_1x + \cdots + c_{N-1}x^{N-1})$, 则由 $\mathbf{a} = \mathbf{F}^{-1}\mathbf{c}$ 知

- 求向量 $(c_0, c_1, \cdots, c_{N-1})^T$ 的离散 Fourier 逆变换 $\mathbf{a} = \check{\mathbf{c}}$, 相当于求多项式 $Q(x)$ 在 $\omega^0, \omega^{-1}, \cdots, \omega^{-(N-1)}$ 这 N 个点上的值.



快速 Fourier 变换的基本思想

我们以 $N = 2^m$ 的情形为例来揭示快速 Fourier 变换的基本思想。

记 $P(x) = a_0 + a_1x^1 + \cdots + a_{N-1}x^{N-1}$, 注意到

$$\begin{aligned} P(x) &= (a_0 + a_2x^2 + \cdots + a_{N-2}x^{N-2}) + x(a_1 + a_3x^2 + \cdots + a_{N-1}x^{N-2}) \\ &= P_e(x^2) + xP_o(x^2). \end{aligned}$$

于是对 $j = 0, 1, \dots, \frac{N}{2} - 1$, 有

$$\begin{cases} c_j = P(\omega_N^j) = P_e(\omega_N^{2j}) + \omega_N^j P_o(\omega_N^{2j}), \\ c_{\frac{N}{2}+j} = P(\omega_N^{\frac{N}{2}+j}) = P_e(\omega_N^{2(\frac{N}{2}+j)}) + \omega_N^{\frac{N}{2}+j} P_o(\omega_N^{2(\frac{N}{2}+j)}). \end{cases}$$



快速 Fourier 变换的基本思想（续）

由 $P_e(\omega_N^{2j}) = P_e(\omega_{\frac{N}{2}}^j)$, $P_o(\omega_N^{2j}) = P_o(\omega_{\frac{N}{2}}^j)$, $j = 0, 1, \dots, \frac{N}{2} - 1$, 有

$$\begin{cases} c_j = P(\omega_N^j) = P_e(\omega_{\frac{N}{2}}^j) + \omega_N^j P_o(\omega_{\frac{N}{2}}^j), \\ c_{\frac{N}{2}+j} = P(\omega_N^{\frac{N}{2}+j}) = P_e(\omega_{\frac{N}{2}}^j) - \omega_N^j P_o(\omega_{\frac{N}{2}}^j), \end{cases} \quad j = 0, 1, \dots, \frac{N}{2} - 1.$$

又 $P_e(x) = P_{ee}(x^2) + xP_{oe}(x^2)$, $P_o(x) = P_{eo}(x^2) + xP_{oo}(x^2)$, 于是

$$\begin{cases} P_e(\omega_{\frac{N}{2}}^j) = P_{ee}(\omega_{\frac{N}{4}}^j) + \omega_{\frac{N}{2}}^j P_{oe}(\omega_{\frac{N}{4}}^j), \\ P_o(\omega_{\frac{N}{2}}^j) = P_{eo}(\omega_{\frac{N}{4}}^j) - \omega_{\frac{N}{2}}^j P_{oo}(\omega_{\frac{N}{4}}^j), \end{cases} \quad j = 0, 1, \dots, \frac{N}{4} - 1.$$

如此递推至 $N = 2^m$ 个 0 次多项式在 1 点的取值. 设计算 $\frac{N}{2}$ 的 Fourier 变换总共需要 $M_{\frac{N}{2}}$ 次乘法和 $A_{\frac{N}{2}}$ 次加法, 则 $M_N = 2M_{\frac{N}{2}} + \frac{N}{2}$, $A_N = 2A_{\frac{N}{2}} + N$. 归纳得 $M_N = 2^m M_{\frac{N}{2^m}} + m \frac{N}{2} = m \frac{N}{2} = \frac{1}{2} N \log_2 N$, $A_N = 2^m A_{\frac{N}{2^m}} + mN = mN = N \log_2 N$.



快速 Fourier 变换基本算法的实现

设 $N = 2^m$, 由以上分析知, FFT 的实现主要分两个步骤:

- ① **分割**: 将第 m 层的向量 \mathbf{a} 分割成 $\mathbf{a}_e, \mathbf{a}_o$, 奇、偶两个第 $m-1$ 层的向量, 再将它们每一个分解成奇、偶两个第 $m-2$ 层的向量 (共 2^2 个), \dots , 直至第 0 层分割成按逆序 (重排序) 的 2^m 个数 (一维向量);
- ② **组装**: 设第 k 层的一个向量 \mathbf{a}_δ^k 分解成 $\mathbf{a}_{e\delta}^{k-1}$ 和 $\mathbf{a}_{o\delta}^{k-1}$ 两个 $k-1$ 层的向量, 记它们的 Fourier 变换为 $\mathbf{c}_{e\delta}^{k-1}$ 和 $\mathbf{c}_{o\delta}^{k-1}$, 记 $K = 2^k, \omega_K = e^{-i\frac{2\pi}{K}}, \boldsymbol{\omega}_K = (\omega_K^0, \dots, \omega_K^{\frac{K}{2}-1})^T$, 则有

$$\mathbf{c}_\delta^k = \begin{pmatrix} \mathbf{c}_{e\delta}^{k-1} + \boldsymbol{\omega}_{2^k} \circ \mathbf{c}_{o\delta}^{k-1} \\ \mathbf{c}_{e\delta}^{k-1} - \boldsymbol{\omega}_{2^k} \circ \mathbf{c}_{o\delta}^{k-1} \end{pmatrix}, \quad \forall \delta, k = 1, \dots, m,$$

其中 $\circ : \mathbb{R}^{\frac{K}{2}} \times \mathbb{R}^{\frac{K}{2}} \rightarrow \mathbb{R}^{\frac{K}{2}}, \mathbf{x} \circ \mathbf{y} = (x_0 y_0, x_1 y_1, \dots, x_{\frac{K}{2}-1} y_{\frac{K}{2}-1})^T$



快速 Fourier 变换基本算法的实现

由组装的方式可以看出, FFT 算法实现的关键在于将 $N = 2^m$ 个数据给出一个第 0 层的排序方法, 使得:

- 顺序两两(2^0 维向量) 将所排数据顺序两两成对组装得到第 1 层的 $N/2 = 2^{m-1}$ 个顺序排列的二维 (2^1) 向量;
- 将顺序排列的 2^{m-1} 个二维 (2^1) 向量顺序两两成对组装得到第 2 层顺序排列的 2^{m-2} 个四维 (2^2) 向量;
- ;
- 将顺序排列的两个 2^{m-1} 维向量组装得到第 m 层顺序排列的 2^m 维向量, 而最终得到的向量正是我们要求的所给数据的 Fourier 变换。



快速 Fourier 变换基本算法的实现

组装过程是从 \mathbf{a} 的逆序按最新指标的偶 0 奇 1(前面的指标相同) 配对计算产生的, 每组装一次, 子向量个数减少一半, 维数增一倍, 组装 m 次后得到 Fourier 变换 \mathbf{c} (一个 $N = 2^m$ 维的向量).

$$\begin{array}{rccccccc}
 \mathbf{a}_{000}; & & & & & & & \\
 \mathbf{a}_{100}; & & \mathbf{c}_{00}; & & & & & \\
 \mathbf{a}_{010}; & & & & \mathbf{c}_0; & & & \\
 \mathbf{a}_{110}; & \pm & \mathbf{c}_{10}; & \pm & & \pm & & \\
 & & \omega_{2^0} & & \omega_{4^0} & & \omega_{8^0} & \mathbf{c}. \\
 \mathbf{a}_{001}; & \longrightarrow & \mathbf{c}_{01}; & \longrightarrow & & \longrightarrow & & \\
 \mathbf{a}_{101}; & & & & \mathbf{c}_1; & & & \\
 \mathbf{a}_{011}; & & \mathbf{c}_{11}; & & & & & \\
 \mathbf{a}_{111}; & & & & & & &
 \end{array}$$



快速 Fourier 变换的应用

- 应用快速 Fourier 变换计算卷积.
- 应用快速 Fourier 变换求解循环矩阵的矩阵特征值问题.
- 应用快速 Fourier 变换求解循环矩阵线性代数方程组.
- 应用快速 Fourier 变换求解微分方程边值问题.
- 实数组的快速 Fourier 变换, 快速正弦和余弦变换.



数值格式是对微分方程的离散逼近

考虑适定的一阶常微分方程初值问题:

$$\begin{cases} \frac{dy}{dx} = f(x, y), & \forall x \in (a, b), \\ y(a) = y_0, \end{cases}$$

的单步法 (即在求 y_{n+1} 时, 不需用到 n 步之前的信息):

$$y_{n+1} = y_n + h\varphi(x_n, x_{n+1}, y_n, y_{n+1}, h),$$

其中 φ 与 f 有关. 记微分算子 $Dy(x) \triangleq y'(x) - f(x, y(x))$, 记离散算子 $D_h y_{n+1} \triangleq \frac{y_{n+1} - y_n}{h} - \varphi(x_n, x_{n+1}, y_n, y_{n+1}, h)$, 则可以认为算法正是用离散算子 D_h 近似替代微分算子 D 之后得出的.

$(D_h - D)y_{n+1}$ 反映了数值格式对微分方程的逼近的精度. 注意, 若 $y(x)$ 是微分方程的精确解, 则有 $Dy(x) \equiv 0$, 因此, 引出以下定义.



数值方法的局部截断误差和局部截断误差主项

定义： 设 $y(x)$ 是微分方程的精确解, 即 $Dy(x) \equiv 0$, 则称

$$T_{n+1} = y(x_{n+1}) - y(x_n) - h\varphi(x_n, x_{n+1}, y(x_n), y(x_{n+1}), h)$$

为单步法的局部截断误差.

注： 按定义有 $T_{n+1} = h \cdot (D_h - D)y(x_{n+1})$.

定义： 如果一个数值方法的局部截断误差 $T_{n+1} = O(h^{p+1})$, 其中 p 为自然数, 则称该方法是 p 阶的, 或具有 p 阶精度. 若有

$$T_{n+1} = g(x_n, y(x_n))h^{p+1} + O(h^{p+2}), \quad g(x_n, y(x_n)) \neq 0,$$

则称 $g(x_n, y(x_n))h^{p+1}$ 是该方法的局部截断误差主项.



单步法的误差 inequality、格式的稳定性与整体误差的阶

由截断误差的定义有, 记 $e_n = y_n - y(x_n)$, 则有

$$|e_{n+1}| \leq |e_n| + h(L_1|e_n| + L_2|e_{n+1}|) + |T_{n+1}|, \quad n \geq 0,$$

其中 L_1, L_2 分别是 φ 关于 y_n 和 y_{n+1} 的 Lipschitz 常数. 于是有

$$(1 - L_2h)|e_{n+1}| \leq (1 + L_1h)|e_n| + |T_{n+1}|, \quad n \geq 0.$$

不妨设 $2L_2h < 1$, 令 $L = 2(L_1 + L_2)$, 递推得误差 inequality

$$\begin{aligned} |e_{n+1}| &\leq (1 + Lh)^{n+1}|e_0| + \sum_{k=0}^n (1 + Lh)^k |T_{n+1-k}| \\ &\leq (1 + Lh)^{n+1}|e_0| + \frac{(1 + Lh)^{n+1} - 1}{Lh} \max_{0 \leq k \leq n} |T_{k+1}|. \end{aligned}$$

注: 误差 inequality 也说明了格式关于初值和右端项的稳定性, 并给出了格式整体误差的阶为 $O(|e_0| + h^{-1}|T_{k+1}|)$.



常微分方程数值格式的绝对稳定性— 数值稳定性判据

定义： 若一个常微分方程数值方法在以定步长 h 求解试验方程

$$y' = \lambda y, \quad \lambda \in \mathbb{C},$$

的非零初值问题时, 所给出的是一线性差分方程, 且得到的离散解 y_n 满足 $\lim_{n \rightarrow \infty} y_n = 0$, 则称该数值方法关于 λh 为绝对稳定的; 否则, 称该方法关于 λh 为非绝对稳定, 或计算不稳定的.

定义： 将一个常微分方程数值方法应用于求解试验方程

$$y' = \lambda y, \quad \lambda \in \mathbb{C},$$

时, 若存在复平面中的区域 $R \subset \mathbb{C}$, 使得当且仅当步长 h 满足

$$\mu = \lambda h \in R \subset \mathbb{C}$$

时, 该数值方法对步长 h 为绝对稳定的, 则称区域 R 为该方法的绝对稳定区域.



Runge-Kutta 方法是通过利用函数 $f(x, y)$ 在 (x, y) 邻域中的某些点选择适当的系数做组合, 来构造单步法的一种系统性的方法. 其几何意义是: 用若干点的加权平均梯度近似真解的平均梯度. 例如

设 m 是一个正整数, 代表使用函数值 f 的次数. 令

$$\begin{aligned} K_1 &= f(x_n, y_n), \\ K_2 &= f(x_n + a_2 h, y_n + h b_{21} K_1), \\ &\dots\dots\dots \\ K_m &= f(x_n + a_m h, y_n + h \sum_{i=1}^{m-1} b_{mi} K_i), \end{aligned}$$

选待定系数 $a_i, b_{ij}, i = 2, \dots, m, j = 1, \dots, i-1$, 和 $c_i, i = 1, \dots, m$, 构造 m 级显式 Runge-Kutta 格式

$$y_{n+1} = y_n + h(c_1 K_1 + \dots + c_m K_m).$$

系数的选择原则是使所得到的方法有尽可能高的精度阶并有较大的绝对稳定区域, 前一目标通常可以通过使方法的 Taylor 展开式与精确解的 Taylor 展开式有尽可能多的重合项来获得.



m 级显式 Runge-Kutta 格式的阶数 $\leq m$ ($\leq m-1$, $m=5, 6, 7$; $\leq m-2$, $m \geq 8$). 而隐式 Runge-Kutta 格式的阶数却可以达到 $2m$ 阶. 且有更大的绝对稳定区域.

给定 $m \geq 1$, 设 $a_i, b_{ij}, c_i \in \mathbb{R}^1, i, j = 1, 2, \dots, m$, 为待定系数. 则称任何由以下方式给出的常微分方程数值格式

$$\begin{cases} K_i = f\left(x_n + a_i h, y_n + h \sum_{j=1}^m b_{ij} K_j\right), & i = 1, \dots, m, \\ y_{n+1} = y_n + h \sum_{i=1}^m c_i K_i, \end{cases}$$

为一个 m 级隐式 Runge-Kutta 格式(方法).

Runge-Kutta 格式的构造方法:

- 将精确解 $y(x_{n+1})$ 和格式在 $x_n, y_n := y(x_n)$ 处做 Taylor 展开, 通过比较展开式系数确定待定系数 c_i, a_i, b_{ij} ;
- 将微分方程化为等价的积分方程, 选取阶数较高的 Gauss 积分公式计算积分, 并由此确定待定系数 $c_i, a_i, i=1, \dots, m$; 然后, 将格式在 $x_n, y_n := y(x_n)$ 处做 Taylor 展开, 通过比较与精确解 $y(x_{n+1})$ 展开式相应项的系数确定待定系数 b_{ij} .



线性多步法及其相容性

线性多步法一般可以表示为

$$\sum_{i=0}^k \alpha_i y_{n+i} = h \sum_{i=0}^k \beta_i f(x_{n+i}, y_{n+i}),$$

其中 $\alpha_k = 1$, α_0 和 β_0 不同时为零. 当 $\beta_k = 0$ 时, 方法称为显式的, $\beta_k \neq 0$ 时, 方法则称为隐式的.

定义: 对线性多步法, 如果存在 $p > 0$, 使得其局部截断误差

$$T_{n+k} \triangleq \sum_{i=0}^k \alpha_i y(x_{n+i}) - h \sum_{i=0}^k \beta_i f(x_{n+i}, y(x_{n+i})) = O(h^{p+1}),$$

其中 $y(x)$ 是相应微分方程的精确解, 则称格式是 (p 阶) 相容的.

注: 可证线性多步法相容的充要条件是 $\rho(1) = 0$, $\rho'(1) = \sigma(1)$ (利用 Taylor 展开), 其中 $\rho(\xi) \triangleq \sum_{i=0}^k \alpha_i \xi^i$ 和 $\sigma(\xi) \triangleq \sum_{i=0}^k \beta_i \xi^i$ 分别称为格式的第一和第二特征多项式.



线性多步法的稳定性与绝对稳定性

对线性多步法, 定义残量算子 \mathcal{R}_h 为

$$(\mathcal{R}_h \mathbf{u})_n = \frac{1}{h} \left(\sum_{i=0}^k \alpha_i u_{n+i} \right) - \sum_{i=0}^k \beta_i f(x_{n+i}, u_{n+i}).$$

定义: 如果存在常数 $K > 0$ 和 $h_0 > 0$, 使得对任意网格尺度满足 $h \leq h_0$ 的网格节点上取值的向量 \mathbf{v}, \mathbf{w} 都有

$$\|\mathbf{v} - \mathbf{w}\|_\infty \leq K \left(\sum_{i=0}^{k-1} \|\mathbf{v}_i - \mathbf{w}_i\|_\infty + \|\mathcal{R}_h \mathbf{v} - \mathcal{R}_h \mathbf{w}\|_\infty \right),$$

则称相应线性多步法格式是稳定的.

线性多步法的绝对稳定区域 $R = \{\mu \in \mathbb{C} : |\xi_i(\mu)| < 1, i = 1, \dots, k\}$, 其中 $\xi_i(\mu), i = 1, \dots, k$ 为线性多步法的特征方程 $\rho(\xi) - \mu\sigma(\xi) = 0$ 的根 (含重数).



线性多步法的零稳定性

定义： 如果存在常数 $K > 0$ 和 $h_0 > 0$, 使得对任意网格尺度满足 $h \leq h_0$ 的网格上线性多步法格式的任何两个解 \mathbf{v}, \mathbf{w} 都有

$$\|\mathbf{v} - \mathbf{w}\|_{\infty} \leq K \sum_{i=0}^{k-1} \|\mathbf{v}_i - \mathbf{w}_i\|_{\infty},$$

则称相应线性多步法格式是零稳定的。

注 1： 可以证明, 线性多步法格式是零稳定的充要条件是其第一特征多项式 $\rho(\xi) \triangleq \sum_{i=0}^k \alpha_i \xi^i$ 满足根条件, 即其所有根都在单位圆内或单位圆上, 且在单位圆上的根只能是单根。

注 2： 可以证明, 若一线性多步法格式是零稳定的, 且 f 满足 Lipschitz 条件, 则当 h 充分小时该格式是稳定的。

注 3： 零稳定等价于格式 (f 满足 Lipschitz 条件) $h \rightarrow 0$ 时的稳定性。



刚性方程组的定义及其数值求解的困难所在

定义： 对于线性常微分方程组 $\mathbf{y}' = \mathbf{A}\mathbf{y} + \phi(t)$, 如果 \mathbf{A} 的特征值 λ_i , 满足 $\operatorname{Re}(\lambda_i) < 0, i = 1, \dots, m$, 且 $\max_i |\operatorname{Re}(\lambda_i)| \gg \min_i |\operatorname{Re}(\lambda_i)|$, 则称该方程组为刚性方程组, 称

$$S = \frac{\max_i |\operatorname{Re}(\lambda_i)|}{\min_i |\operatorname{Re}(\lambda_i)|}$$

为刚性比. 对于非线性常微分方程组 $\mathbf{y}'(t) = \mathbf{f}(t, \mathbf{y}(t))$, 设 $\bar{\mathbf{y}}$ 为其精确解, 令 $\mathbf{J}(t) = \frac{\partial \mathbf{f}}{\partial \mathbf{y}}(t, \bar{\mathbf{y}})$, 若线性方程组 $\mathbf{z}' = \mathbf{J}(t)\mathbf{z}$ 是刚性的, 则称该非线性方程组为刚性的.

困境： 刚性系统中内含差别很大的时间尺度, 时间步长一般由最小的时间尺度 (系统变化最快的子过程) 和格式的绝对稳定区域决定, 而弛豫时间则由最大的时间尺度 (系统变化最慢的子过程) 决定.



A 稳定的数值方法

定义： 如果一个数值格式的绝对稳定区域包含了整个左半复平面 $\text{Re}(\lambda h) < 0$, 则称该格式是 A -稳定的.

显然, 只要数值格式是 A -稳定的, 则不论方程组的刚性有多大, 格式的绝对稳定性都不会限制步长 h 的选取, 因此, 步长 h 的选取完全由精度要求决定.

定理： 关于 A -稳定格式有以下结果:

- ① 任何显式线性多步法 (包括显式 *Runge-Kutta* 方法) 都不是 A -稳定的.
- ② A -稳定的隐式线性多步法的精度不超过二阶.
- ③ 具有最小误差常数的二阶 A -稳定隐式线性多步法是梯形法.
- ④ *Gauss* 型的 m 级 $2m$ 阶 (及某些 m 级 $2m - 1$ 和 $2m - 2$ 阶) 的隐式 *Runge-Kutta* 方法是 A -稳定的.



$A(\alpha)$ -稳定性和刚性稳定性

定义： 如果存在常数 $\alpha > 0$, 使得一个数值格式的绝对稳定区域 R 满足

$$R \supset \left\{ \mu \in \mathbb{C} : \operatorname{Re}(\mu) < 0, \arctan \frac{|\operatorname{Im}(\mu)|}{|\operatorname{Re}(\mu)|} \leq \alpha \right\},$$

则称该格式是 $A(\alpha)$ -稳定的.

定义： 如果存在常数 $a > 0, c > 0$, 使得一个数值格式的绝对稳定区域 R 满足

$$R \supset \left\{ \mu \in \mathbb{C} : \operatorname{Re}(\mu) \leq -a, \text{ 或 } \operatorname{Re}(\mu) < 0, |\operatorname{Im}(\mu)| \leq c \right\},$$

则称该格式是刚性稳定的.

注： 若一个数值格式是刚性稳定的, 令 $\alpha = \arctan \frac{c}{a}$, 则该格式是 $A(\alpha)$ -稳定的.



$A(\alpha)$ -稳定性和刚性稳定性数值方法的例 — Gear 方法

Gear 方法是一种 k 步 k 阶隐式线性多步法. 其形式为:

$$\mathbf{y}_{n+k} + \sum_{j=1}^k \alpha_{k,j} \mathbf{y}_{n+k-j} = h\beta_k \mathbf{f}(x_{n+k}, \mathbf{y}_{n+k}).$$

k	$\alpha_{k,1}$	$\alpha_{k,2}$	$\alpha_{k,3}$	$\alpha_{k,4}$	$\alpha_{k,5}$	$\alpha_{k,6}$	β_k
1	-1						1
2	$-\frac{4}{3}$	$\frac{1}{3}$					$\frac{2}{3}$
3	$-\frac{18}{11}$	$\frac{9}{11}$	$-\frac{2}{11}$				$\frac{6}{11}$
4	$-\frac{48}{25}$	$\frac{36}{25}$	$-\frac{16}{25}$	$\frac{3}{25}$			$\frac{12}{25}$
5	$-\frac{300}{137}$	$\frac{300}{137}$	$-\frac{200}{137}$	$\frac{75}{137}$	$-\frac{12}{137}$		$\frac{60}{137}$
6	$-\frac{360}{147}$	$\frac{450}{147}$	$-\frac{400}{147}$	$\frac{225}{147}$	$-\frac{72}{147}$	$\frac{10}{147}$	$\frac{60}{147}$



Hamilton 系统的相空间是一个辛空间

Hamilton 系统一般可表示为

$$\begin{cases} \frac{d\mathbf{p}}{dt} = -H_{\mathbf{q}}(\mathbf{p}, \mathbf{q}), \\ \frac{d\mathbf{q}}{dt} = H_{\mathbf{p}}(\mathbf{p}, \mathbf{q}). \end{cases}$$

其中 $\mathbf{p} = (p_1, \dots, p_n)$ 称为广义动量, $\mathbf{q} = (q_1, \dots, q_n)$ 称为广义坐标, 可微函数 $H(\mathbf{p}, \mathbf{q})$ 称为系统的 Hamilton 函数. 若记

$$\mathbf{z} = \begin{pmatrix} \mathbf{p} \\ \mathbf{q} \end{pmatrix}, \quad \mathbf{J}_{2n} = \begin{pmatrix} \mathbf{0} & \mathbf{I}_n \\ -\mathbf{I}_n & \mathbf{0} \end{pmatrix},$$

这里 \mathbf{I}_n 为 n 阶单位矩阵, 则 Hamilton 系统又可表示为

$$\frac{d\mathbf{z}}{dt} = \mathbf{J}_{2n}^{-1} H_{\mathbf{z}} = \mathbf{J}_{2n}^{-1} \begin{pmatrix} H_{\mathbf{p}} \\ H_{\mathbf{q}} \end{pmatrix}.$$

Hamilton 系统的相空间在由 \mathbf{J}_{2n} 定义的辛内积 $[\mathbf{x}, \mathbf{y}] = \mathbf{x}^T \mathbf{J}_{2n} \mathbf{y}$ 下是一个辛空间.



定义： 如果一个 $2n$ 阶矩阵 \mathbf{S} 是保辛内积的, 即 $[\mathbf{S}\mathbf{x}, \mathbf{S}\mathbf{y}] = [\mathbf{x}, \mathbf{y}]$, 则称其为一个辛矩阵. 所有辛矩阵组成一个群, 称为辛群, 记为 $Sp(2n)$.

Hamilton 系统的解有一个重要的性质, 即它给出了 Hamilton 系统相空间上的一族单参数可微映射: $(\mathbf{p}(t_0), \mathbf{q}(t_0)) \mapsto (\mathbf{p}(t), \mathbf{q}(t))$, 且该族映射是保辛结构的(映射的 Jacobi 矩阵是辛矩阵).

定义： $2n$ 阶矩阵 \mathbf{B} 称为是无穷小辛矩阵, 如果

$$\mathbf{B}^T \mathbf{J}_{2n} + \mathbf{J}_{2n} \mathbf{B} = \mathbf{0}.$$

无穷小辛矩阵组成的线性空间与对易运算 $[\mathbf{A}, \mathbf{B}] = \mathbf{AB} - \mathbf{BA}$ 一起定义了一个李代数, 记作 $sp(2n)$.

引理： 矩阵 \mathbf{B} 是无穷小辛矩阵当且仅当 $\mathbf{B} = \mathbf{J}_{2n} \mathbf{A}$, 其中 \mathbf{A} 是对称矩阵.

定理： 如果 $\mathbf{B} \in sp(2n)$, 则 $\exp(t\mathbf{B}) \in Sp(2n)$, $\forall t \in \mathbb{R}$. 若还有 $|\mathbf{I}_{2n} + \mathbf{B}| \neq 0$, 则有 $\mathbf{F} \triangleq (\mathbf{I}_{2n} + \mathbf{B})^{-1}(\mathbf{I}_{2n} - \mathbf{B}) \in Sp(2n)$.

定理： 设函数 $\psi(\lambda)$ 在 $\lambda = 0$ 处能展成幂级数, $\psi(\lambda)\psi(-\lambda) = 1$, 且 $\psi'(0) \neq 0$, $\psi(0) = 1$. 如果 $\mathbf{B} \in sp(2n)$, 则 $\psi(h\mathbf{B}) \in Sp(2n)$.



线性 Hamilton 系统及其解的无穷小辛矩阵表示

考虑线性 Hamilton 系统, 即 Hamilton 函数 H 是 \mathbf{z} 的二次型:

$$H(\mathbf{z}) = \frac{1}{2} \mathbf{z}^T \mathbf{S} \mathbf{z}, \quad \mathbf{S}^T = \mathbf{S}.$$

在这种情况下, Hamilton 系统可以写成

$$\frac{d\mathbf{z}}{dt} = \mathbf{B} \mathbf{z}, \quad \mathbf{B} = \mathbf{J}_{2n}^{-1} \mathbf{S}.$$

由于 $\mathbf{B} \in sp(2n)$, 所以 $\exp(t\mathbf{B}) \in Sp(2n)$. 此时, Hamilton 系统初值问题的解可以表示为 $\mathbf{z}(t) = \exp(t\mathbf{B})\mathbf{z}(0)$, 容易看出这是一个由无穷小辛矩阵的指数变换定义的相空间上保辛结构映射.



Hamilton 系统的辛格式

Hamilton 系统的辛结构有重要的物理意义. 例如, 对线性可分 Hamilton 系统来说, 保辛结构等价于能量守恒.

线性 Hamilton 系统的差分格式称为是辛的, 如果差分格式定义了从 \mathbf{z}_m 到 \mathbf{z}_{m+1} 一个辛变换 (即相应的变换矩阵为辛矩阵).

令 $\mathbf{S}(t) = \frac{\partial \mathbf{z}(t)}{\partial \mathbf{z}(t_0)}$, 则非线性 Hamilton 系统的保辛结构性质是由 $\mathbf{S}(t) \in Sp(2n)$, 即 $\mathbf{S}(t)^T \mathbf{J}_{2n} \mathbf{S}(t) = \mathbf{J}_{2n}, \forall t$, 刻画的.

对于非线性 Hamilton 系统的单步隐式数值格式, 令 $\mathbf{S}_m = \frac{\partial \mathbf{z}_{m+1}}{\partial \mathbf{z}_m}$, 如果 $\mathbf{S}_m \in Sp(2n)$, 即 $\mathbf{S}_m^T \mathbf{J}_{2n} \mathbf{S}_m = \mathbf{J}_{2n}, \forall m \geq 0$, 则称该格式是辛的.



线性 Hamilton 系统辛格式的例

定理： 设 $P_k(x)$ 是 e^x 的有如下形式的有理逼近：

$$e^x = \frac{P_k(x)}{P_k(-x)} + O(|x|^{2k+1}),$$

则线性 *Hamilton* 系统的差分格式

$$\mathbf{z}_{m+1} = \frac{P_k(h\mathbf{B})}{P_k(-h\mathbf{B})} \mathbf{z}_m, \quad k = 1, 2, \dots$$

是具有 $2k$ 阶精度的辛格式, 且有 $\frac{1}{2} \mathbf{z}_{m+1}^T \mathbf{S} \mathbf{z}_{m+1} = \frac{1}{2} \mathbf{z}_m^T \mathbf{S} \mathbf{z}_m$.

基于 Gauss-Legendre 数值积分公式的 s 级 $2s$ 阶隐式 Runge-Kutta 格式是辛格式.



随机方法的例 — 数值积分的 Monte Carlo 方法

对独立同分布的随机变量 $X_i, i = 1, 2, \dots$, 如果 $E|X_i| < +\infty$, 记 $S_N = \sum_{i=1}^N X_i$, 则由弱大数定律, 当 $N \rightarrow \infty$ 时, S_N/N 依概率收敛于 EX_1 .

因此, 为了计算 $I(f) = Ef(X)$, 取相互独立且同服从区域 Ω 上均匀分布的随机变量 $X_i, i = 1, 2, \dots, N$, 构造数值积分公式

$$I(f) = \int_{\Omega} f(x)dx \approx I_N(f) \triangleq \frac{1}{N} \sum_1^N f(X_i).$$

这就是数值积分的 Monte Carlo 方法. 其误差 $e_N = |I_N(f) - I(f)|$ 满足

$$E|e_N| \leq \sqrt{E|e_N|^2} = \sqrt{\frac{\text{Var}(f(X))}{N}}.$$

由此知产生独立同分布随机数序列和减小方差 $\text{Var}(f(X))$ 至关重要.



- 产生 $U[0, 1]$ 伪随机数的线性同余法
 $X_{n+1} = aX_n + b \pmod{m}$, 其中 a, b, m 满足
 - ① b 与 m 互素;
 - ② $(a - 1)$ 是 m 的任一奇数因子的倍数;
 - ③ 若 4 是 m 的因子, 则 4 也必定是 $(a - 1)$ 的因子,则该伪随机数发生器的最大循环长度为 m , 即满长度.
- Lewis, Goodman 和 Miller 提出的最大循环长度达 2.1×10^9 的 $U[0, 1]$ 伪随机数发生器: $X_{n+1} = aX_n \pmod{m}$, 其中 $a = 7^5 = 16807$, $m = 2^{31} - 1 = 2147483647$.
- 基于变换法的一般分布的伪随机数发生器. (如果随机变量 $X \sim U[0, 1]$, 则随机变量 $Y = F^{-1}(X)$ 的分布函数为 $F(y)$.)
- 概率密度为 $p(x)$ 的随机变量的产生方法 — 舍选法:
 - ① 按分布 $U[a, b]$ 生成随机数 X_i ;
 - ② 按分布 $U[0, d]$ 生成随机数 Y_i , 称为决策变量;
 - ③ 决策: 如果 $0 \leq Y_i < p(X_i)$, 则接受 X_i , 否则拒绝;
 - ④ 转步 1.



- 重要性抽样法: $I(f) \approx \frac{1}{N} \sum_{i=1}^N \frac{f(Y_i)}{p(Y_i)}$, 其中 $Y_i, i = 1, \dots, N$, 是概率密度为 $p(x)$ (满足 $\int_0^1 \frac{f^2(x)}{p(x)} - f^2(x) dx < 0$) 的 i.i.d. 随机变量.
- 控制变量法: $I_N(f) = \frac{1}{N} \sum_{i=1}^N [f(X_i) - g(X_i)] + I(g)$, 其中 $X_i, i = 1, \dots, N, i.i.d. \sim \mathcal{U}[0, 1], I(g)$ 已知, $\text{Var}(f-g) < \text{Var}(f)$.

- 分层抽样法: 将区域 $\Omega = [0, 1]$ 分为 M 等份:

$\Omega_k = [\frac{k-1}{M}, \frac{k}{M}]$, $k = 1, \dots, M$; 取 i.i.d. $X_i^{(k)} \sim \mathcal{U}(\Omega_k)$, $i = 1, \dots, n, k = 1, \dots, M$, 共 $N = nM$ 个随机变量; 令

$$I_{nM}(f) = \sum_{k=1}^M \frac{1}{nM} \sum_{i=1}^n f(X_i^{(k)}) = \frac{1}{N} \sum_{k=1}^M \sum_{i=1}^n f(X_i^{(k)}).$$

- 对偶变量法: 如果 $f(x)$ 是单调的, $X \sim \mathcal{U}[0, 1]$, 则 $I_N \triangleq \frac{1}{2N} \sum_{i=1}^N [f(X_i) + f(1 - X_i)]$, 有 $E I_N = I(f)$, 且 $\text{Var}(I_N) = E |I_N - I(f)|^2 \leq \frac{1}{2N} \text{Var}(f)$.



Metropolis 算法 — 一种马氏链 Monte Carlo 方法

Metropolis 算法要求转移概率矩阵 \mathbf{P} 具有如下性质:

- ① 转移概率矩阵 \mathbf{P} 满足: (i) $p_{ij} \geq 0, \forall i, j = 1, \dots, N_t$; (ii) $\sum_{j=1}^{N_t} p_{ij} = 1, \forall i = 1, \dots, N_t$.
- ② 转移概率矩阵 \mathbf{P} 是本原矩阵. 此时以 \mathbf{P} 为转移概率矩阵的马氏链是本原的 (遍历的).
- ③ $\sigma^{(i)}$ 的出现概率 $\frac{1}{Z_M} \exp\{\beta H(\sigma^{(i)})\}$ 排成的 N_t 维行向量 $\boldsymbol{\pi}$ 是转移概率矩阵 \mathbf{P} 的不变分布.
- ④ 由转移概率矩阵 \mathbf{P} 定义的马氏链满足细致平衡条件:

$$\pi(\sigma)P(\sigma \rightarrow \sigma') = \pi(\sigma')P(\sigma' \rightarrow \sigma).$$

定理: 设马氏链具有性质 (1)-(4), 设函数 $g(\sigma)$ 满足 $E|g(\sigma)| = \sum_{\sigma \in \Omega} \pi(\sigma)|g(\sigma)| < \infty$. 则对由该马氏链生成的从任一初始状态 $\sigma^{(0)}$ 出发的状态序列 $\sigma^{(1)}, \sigma^{(2)}, \dots, \sigma^{(n)}, \dots$ 都有

$$\frac{1}{n} \sum_{i=1}^n g(\sigma^{(i)}) \rightarrow \sum_{\sigma \in \Omega} \pi(\sigma)g(\sigma), \text{ 当 } n \rightarrow \infty, \text{ a.s.}$$



Metropolis 算法的流程

Metropolis 算法:

- ① 设定初始态 $\sigma^{(1)}$ 和总计算步数 N .
- ② 根据一定的预选规则由 $\sigma^{(n)}$ 产生预选态 σ' .
- ③ 计算 $\Delta H = H(\sigma') - H(\sigma)$, 计算

$$A = \min\{1, R\} = \begin{cases} 1, & \Delta H \leq 0, \\ \exp\{-\beta\Delta H\}, & \text{其它.} \end{cases}$$

- ④ 生成一均匀分布随机数 $r \sim \mathcal{U}[0, 1]$;
- ⑤ 如果 $r \leq A$, 则令 $\sigma^{(n+1)} = \sigma'$; 否则, 令 $\sigma^{(n+1)} = \sigma^{(n)}$.
- ⑥ 如果 $n + 1 < N$, 转步 2; 否则计算 $\frac{1}{N} \sum_{i=1}^N H(\sigma^{(i)})$.

注: 若在第三步中取 $A = (1 + \exp\{-\beta\Delta H\})$, 则相应的算法称为 Glauber 算法. 该算法也产生满足细致平衡条件的本原马氏链.



模拟退火算法所依据的基本事实

对于优化问题

$$\mathbf{x} = \arg \min_{\mathbf{x} \in X} H(\mathbf{x}),$$

定义 $H(\mathbf{x})$ 的全局极小点集 $M = \{\mathbf{x}_0 : H(\mathbf{x}_0) = \min_{\mathbf{x} \in X} H(\mathbf{x})\}$,
引入参数 β , 定义概率密度函数

$$\Pi^\beta(\mathbf{x}) = \frac{1}{Z_\beta} \exp\{-\beta H(\mathbf{x})\}, \quad Z_\beta = \sum_{\mathbf{x} \in X} \exp\{-\beta H(\mathbf{x})\},$$

相应分布称为 $\Pi^\beta(\mathbf{x})$ 分布.

定理: 记 $|M|$ 为集合 M 中元素的个数, 则 $\Pi^\beta(\mathbf{x})$ 有性质

$$\lim_{\beta \rightarrow +\infty} \Pi^\beta(\mathbf{x}) = \Pi^\infty(\mathbf{x}) \triangleq \begin{cases} \frac{1}{|M|}, & \mathbf{x} \in M, \\ 0, & \mathbf{x} \notin M, \end{cases}$$

且当 β 充分大时, 对任一 $\mathbf{x} \in M$, $\Pi^\beta(\mathbf{x})$ 作为 β 的函数单调增;
对任一 $\mathbf{x} \notin M$, $\Pi^\beta(\mathbf{x})$ 作为 β 的函数单调减.



模拟退火算法的基本流程

设给定正单调增序列 $\{\beta_n\}_{n=0}^{\infty}$, $\lim_{n \rightarrow \infty} \beta_n = +\infty$.

模拟退火算法:

- ① 设定初始值 $n = 0$, \mathbf{x}_0 , $\beta = \beta_0$, 内循环步数 K 和迭代步数 N .
- ② 根据一定的预选规则由 \mathbf{x}_k 产生预选态 \mathbf{x}' .
- ③ 计算 $\Delta H = H(\mathbf{x}') - H(\mathbf{x}_k)$, 计算

$$A = \min\{1, R\} = \begin{cases} 1, & \Delta H \leq 0, \\ \exp\{-\beta\Delta H\}, & \text{其它.} \end{cases}$$

- ④ 生成一均匀分布随机数 $r \sim \mathcal{U}[0, 1]$;
- ⑤ 如果 $r \leq A$, 则令 $\mathbf{x}_{k+1} = \mathbf{x}'$; 否则, 令 $\mathbf{x}_{k+1} = \mathbf{x}_k$.
- ⑥ 如果 $k + 1 < K$, 转步 2; 否则, $\mathbf{x}_0 := \mathbf{x}_K$, $k := 0$.
- ⑦ 如果 $n < N$, $n := n + 1$, $\beta = \beta_n$, 转步 2; 否则, 输出结果.



模拟退火算法收敛性的基本定理

为简单起见, 内循环次数 K 甚至可以取 1. 此时有收敛性结果:

定理: 设 X 为一有限集, $H(\mathbf{x})$ 为 X 上的非常数函数, \mathbf{G} 是以 $\mathbf{G}(\mathbf{x}, \mathbf{y})$ 为元素的对称不可约预选阵, \mathbf{P}^β 为由 \mathbf{G} 和相应的参数为 β 的满足细致平衡条件的规则所定义的转移概率矩阵. 如果退火速度 $\beta_n \leq C \ln n$, 其中 $C = C(\mathbf{G}, H)$ 为常数, 则对任意初始分布 ν , 有

$$\lim_{n \rightarrow \infty} \|\nu \mathbf{P}^{\beta_1} \dots \mathbf{P}^{\beta_n} - \Pi^\infty\|_1 = 0.$$

注: 退火速度 $\beta_n \leq C \ln n$, 意味着若要 β_n 达到指定的 $N_0 \gg 1$ 从而使计算满足指定的精度要求, 则必须取 $n \sim \exp(N_0)$. 这通常是无法忍受的. 实际计算时, 常选 $\beta_n \sim p^{-n}$ ($p \lesssim 1$ 例如 0.999), 往往也可得到不错的计算效果(不得已而求其次).



拟 Monte Carlo 方法理论基础 — Koksma-Hlawka 定理

定理： 对任何序列 $\{\mathbf{x}_n\}_{n=1}^N$ 及任何有界变差函数 f , 数值积分误差满足 *Koksma-Hlawka* 不等式：

$$\mathcal{E}[f] \triangleq |I(f) - I_N(f)| = \left| \int_{I^d} f(x) dx - \frac{1}{N} \sum_{n=1}^N f(x_n) \right| \leq V[f] D_N^*,$$

其中 $V[f]$ 是 Hardy 和 Krause 意义下函数 $f(x)$ 的变差, D_N^* 是拟随机数序列

注 1： 由 Koksma-Hlawka 不等式知, 拟 Monte Carlo 积分的关键在于生成尽可能小的 D_N^* 的拟随机数序列的差异 (discrepancy) .

注 2： 已有若干种方法, 能够产生 $D_N \leq C_d (\ln N)^{k_d} N^{-1}$ 的拟随机数序列, 其中 C_d, k_d 是与 d 有关的常数.



能够产生 $D_N \leq C_d(\ln N)^d N^{-1}$ 的拟随机数序列方法的例

- Van der Corput 序列 ($d = 1$): x_n 的产生方法:
 - ① 将 n 表示为 2 进制数 $n = a_m a_{m-1} \cdots a_1 a_0$;
 - ② 产生 2 进制数 $x_n = 0.a_0 a_1 \cdots a_{m-1} a_m$.
- Halton 序列 ($d > 1$): $\mathbf{x}_n = (x_n^1, \cdots, x_n^d)$ 的产生方法:
 - ① 将 n 表示为 p_k 进制数 $n = a_{m_k}^k a_{m_k-1}^k \cdots a_1^k a_0^k$, $1 \leq k \leq d$, 其中 p_k 为第 k 个素数;
 - ② 产生 p_k 进制数 $x_n^k = 0.a_0^k a_1^k \cdots a_{m_k-1}^k a_{m_k}^k$.

注 1: 可以证明 Halton 序列满足 $D_N \leq C_d(\ln N)^d N^{-1}$.

注 2 (拟 Monte Carlo 方法的局限性): 一般最适合于空间维数中等、被积函数光滑、盒形区域上的积分。



Thank You!

