数值分析内部课程习题解答

第一章

SCAU DataHub

2025 年 3 月 22 日

目录

1 Floating-point numbers

Solution 1.1. 参考书中的 Example 以及题目的 Hint.

(a) 在区间 $[1/2, 4]$ 共有 49 个浮点数.

参考 Example 1.1.2 区间 $[1/2, 4]$ 可以写为 $[2^{-1}, 2^2]$，进一步划分为

$$[2^{-1}, 2^0) \cup [2^0, 2^1) \cup [2^1, 2^2) \cup \{2^2\},$$

其中每个区间共有 2^d 个浮点数，题设 $d = 4$，即每个区间有 $2^4 = 16$ 个浮点数，所以在整个 $[1/2, 4]$ 上共有 $16 \times 3 + 1 = 49$ 个浮点数.

(b) 浮点数为 $(1 + \sum_{i=1}^{4} b_i 2^{-i}) \times 10^{-4}$，其中 $(b_1, b_2, b_3, b_4) = (1, 0, 1, 0)$.

想找到 $\mathbb{F}$ 中最接近 $1/10$ 的元素，首先就是定位哪个区间 $[2^n, 2^{n+1})$ 能包住 $1/10$，显然取 $n = -4$ 就可以. 书中给的 Hint 枚举出 $[2^{-4}, 2^{-3})$ 的所有元素，当然也可以，我们就按照浮点数的定义写出来

$$\min_{b_i, i=1,\cdots,4} \left| \left(1 + \sum_{i=1}^{d} b_i 2^{-i}\right) \times 2^n - \frac{1}{10} \right|, \quad (d = 4, n = -4)$$

不过手算的话有点麻烦.

换个角度，我们的目的无非是确定系数 b_i，所以可以把问题转化为

$$\min_{b_i, i=1,\cdots,4} \left| \left(1 + \sum_{i=1}^{4} b_i 2^{-i}\right) - \frac{16}{10} \right| \Leftrightarrow \min_{b_i, i=1,\cdots,4} \left| \left(1 + b_1 \frac{1}{2} + b_2 \frac{1}{4} + b_3 \frac{1}{8} + b_4 \frac{1}{16}\right) - \frac{16}{10} \right|$$

于是确认出系数 $(b_1, b_2, b_3, b_4) = (1, 0, 1, 0)$，进而表示出浮点数.

(c) 33.

根据题设 $d = 4$，找出 n 使得浮点数 $\left(1 + \sum_{i=1}^{d} b_i 2^{-i}\right) \times 2^n$ 的最小变化幅度大于 1 即可. 最小变化就是最低位的变化，即 2^{-4+n}，当 $n = 5$ 时，最小变化就是 2，比 1 大了. 此时最小整数为 32，下一个整数 33 就不在 $\mathbb{F}$ 了.

或者我们可以换个角度思考. 给定精度 $d = 4$，表示一个二进制的浮点数数时，形式上就是 $1.b_1b_2b_3b_4$，其中 $b_i \in \{0, 1\}, i = 1, \cdots, 4$，然后 n 的作用就是让小数点左右移. 例如，当 $n = 1$，浮点数就形如 $1b_1.b_2b_3b_4$，这个时候取 $b_2, b_3, b_4 = 0$，是能通过 b_1 表示二进制整数 10, 11 的，即十进制整数 2, 3。

而当 $n > d$ 时，例如 $n = 5, d = 4$，二进制浮点数形如 $1b_1b_2b_3b_40$，此时最小的整数就是 100000，让最低位的 b_4 变动一下就是 100010，可见中间损失掉了一个整数 100001. 也就是说 $d = 4$ 精度时，无法表示整数 100001. 换算成十进制就是 $2^5 + 1 = 33$.

Solution 1.2. 根据

$$\frac{|\mathrm{fl}(x)-x|}{|x|} \leq \frac{2^{n-d-1}}{2^n} = \frac{1}{2}\epsilon_{\mathrm{mach}},$$

我们把最左边的分母挪到最右边，有

$$|\mathrm{fl}(x)-x| \leq \frac{1}{2}|x|\epsilon_{\mathrm{mach}},$$

拆掉左边的绝对值，然后一步步推

$$\Rightarrow -\frac{1}{2}|x|\epsilon_{\mathrm{mach}} \leq \mathrm{fl}(x)-x \leq \frac{1}{2}|x|\epsilon_{\mathrm{mach}} \tag{1}$$

$$\Rightarrow x-\frac{1}{2}|x|\epsilon_{\mathrm{mach}} \leq \mathrm{fl}(x) \leq x+\frac{1}{2}|x|\epsilon_{\mathrm{mach}} \tag{2}$$

$$\Rightarrow -|x|-\frac{1}{2}|x|\epsilon_{\mathrm{mach}} \leq \mathrm{fl}(x) \leq |x|+\frac{1}{2}|x|\epsilon_{\mathrm{mach}} \tag{3}$$

$$\Rightarrow -|x|(1+\frac{1}{2}\epsilon_{\mathrm{mach}}) \leq \mathrm{fl}(x) \leq |x|(1+\frac{1}{2}\epsilon_{\mathrm{mach}}), \tag{4}$$

即可推出

$$\mathrm{fl}(x) = x(1+\epsilon) \quad \text{for some } |\epsilon| \leq \frac{1}{2}\epsilon_{\mathrm{mach}}.$$

然后反过来的推导只需要把 $\mathrm{fl}(x)$ 代进去验证就可以了.

Solution 1.3. 代码参考 Gitee 仓库.

Solution 1.4. 参考 1 (c) 的思路.

(a) $1+\frac{1}{2^{23}}$.

(b) $2^{24}+1$. 找最低位变化大于 1 的时候，n 的最小取值. 即 $2^{-23+n}>1$，当 $n=24$ 时满足，所以最小的无法被表示的正整数是 $2^{24}+1$.

Solution 1.5. 关键在于理解 Inf 这些字段在编程语言中的含义.

floatmax() 之后更大的数由于机器无法表示，所以会返回 Inf.

-Inf 表示无穷小量，意思是非常接近 0，比其更大的下一个浮点数就是 ϵ_{mach}，所以会返回机器精度表示的最小数.

在 Python 中暂时没有类似的函数.

2 Problems and conditioning

Solution 2.1. 很显然，这几个函数都是存在一阶导数的，因此可以直接按照条件数的定义公式

$$\kappa_f(x) = \left|\frac{xf^{'}(x)}{f(x)}\right|$$

来计算获得结果.

Solution 2.2. 这几个函数也存在一阶导数，因此可以直接用条件数的定义来求解，也可以将原函数看作复合函数来求解条件数，既若 $h(x) = f(g(x))$，那么

$$\kappa_h(x) = \kappa_f(g(x)) \cdot \kappa_g(x).$$

(a) 观察可得 $h(x) = \sqrt{x+5} = f(g(x))$，其中 $f(x) = \sqrt{x}, g(x) = x+5$，这样就可以用上面的公式来求解出条件数. 首先

$$\kappa_f(g(x)) = \left|\frac{g(x)f^{'}(g(x))}{f(g(x))}\right| = \frac{1}{2}, \quad \kappa_g(x) = \left|\frac{x}{x+5}\right|,$$

因此

$$\kappa_h(x) = \kappa_f(g(x)) \cdot \kappa_g(x) = \left|\frac{x}{2x+10}\right|.$$

(b) 观察可得 $h(x) = \cos(2\pi x) = f(g(x))$, 其中 $f(x) = \cos(x), g(x) = 2\pi x$, 其他过程略;

(c) 观察可得 $h(x) = \mathrm{e}^{-x^2} = f(g(x))$, 其中 $f(x) = \mathrm{e}^{x}, g(x) = -x^2$, 其他过程略.

Solution 2.3. 我们可以先求解条件数，再进行求极限的操作，这里只给 (a) 的参考解答

(a) 首先

$$\kappa_f(x) = \left|\frac{xf^{'}(x)}{f(x)}\right| = \left|\frac{x\mathrm{sech}^2(x)}{\tanh(x)}\right| = \left|\frac{x(\frac{2\mathrm{e}^x}{\mathrm{e}^{2x}+1})^2}{\frac{\mathrm{e}^{2x}-1}{\mathrm{e}^{2x}+1}}\right| = \left|\frac{x4\mathrm{e}^{2x}}{\mathrm{e}^{4x}-1}\right| = \left|\frac{4x}{\mathrm{e}^{2x}-\mathrm{e}^{-2x}}\right|.$$

不难看出，对于任意的 $x \neq 0$，$\kappa_f(x) \neq 0$ 且不等于 $\pm\infty$，而当 $x = \pm\infty$ 时，由于指数函数增长快于 $4x$，因此 $\kappa_f(x) = 0$，因此只需要研究 $x \to 0$ 的情况，即

$$\lim_{x\to 0}\frac{4x}{\mathrm{e}^{2x}-\mathrm{e}^{-2x}} = \lim_{x\to 0}\frac{4}{2\mathrm{e}^{2x}+2\mathrm{e}^{-2x}} = 1.$$

因此，对于 $f(x) = \tanh(x)$，不存在 x 使得 $\kappa_f(x) = \pm\infty$.

Solution 2.4. 这题非常容易，借用链式法则 $h^{'}(x) = f^{'}(g(x))g^{'}(x)$ 即可推出.

Solution 2.5. 首先注意 $f^{-1}(x) \neq 1/f(x)$，而是反函数，即给定 $y = f(x)$，f^{-1} 被定义为

$$x = f^{-1}(y)$$

下面我们先计算 $(f^{-1}(y))^{'}$，显然 $y = f(f^{-1}(y))$，因此

$$1 = \frac{d}{dy}(y) = \frac{d}{dy}(f(f^{-1}(y))) \tag{5}$$

$$= \frac{d(f(f^{-1}(y)))}{d(f^{-1}(y))}\frac{d(f^{-1}(y))}{dy} \tag{6}$$

$$= f^{'}(f^{-1}(y))\frac{d(f^{-1}(y))}{dy} \tag{7}$$

$$= f^{'}(f^{-1}(y))(f^{-1})^{'}(y) \tag{8}$$

整理可得

$$(f^{-1})'(y) = \frac{1}{f'(f^{-1}(y))}$$

因此

$$\kappa_{f^{-1}}(y) = \left|\frac{y(f^{-1})'(y)}{f^{-1}(y)}\right| = \left|\frac{f(f^{-1}(y))}{f^{-1}(y)f'(f^{-1}(y))}\right| = \left|\frac{1}{\kappa_f(f^{-1}(y))}\right|.$$

再把 y 换成 x 就行, 注意, 这里 x 只是代表变量而已, 为的是与题目的表达达到形式的统一, 与前面的 x 不同, 前面代表 x 是属于 f 的定义域, 这里的 x 属于 f^{-1} 的定义域.

Solution 2.6. 完全按照书本正文描述方法就可以求出来, 这里不赘述, 参考答案为

$$\kappa_f(b) = \left|\frac{r_1 + r_2}{r_1 - r_2}\right|.$$

Solution 2.7. (a) 对于方程

$$x^2 - (2+\epsilon)x + 1 = 0$$

可以使用求根公式

$$r_1(\epsilon) = \frac{(2+\epsilon) + \sqrt{4\epsilon + \epsilon^2}}{2}, \quad r_2(\epsilon) = \frac{(2+\epsilon) - \sqrt{4\epsilon + \epsilon^2}}{2}$$

实现遍历不同的 ϵ 得到系列值的参考代码参考 Gitee 仓库.

(b) 由求根公式可得

$$|r_1(\epsilon) - 1| = \left|\frac{\epsilon + \sqrt{4\epsilon + \epsilon^2}}{2}\right| \tag{9}$$

$$|r_2(\epsilon) - 1| = \left|\frac{\epsilon - \sqrt{4\epsilon + \epsilon^2}}{2}\right| \tag{10}$$

显然 $\max\{|r_1(\epsilon) - 1|, |r_2(\epsilon) - 1|\} = |r_1(\epsilon) - 1|$, 且 $\epsilon^2 = o(\epsilon), \epsilon = o(\epsilon^{1/2})$, o 代表高阶无穷小, 代入可得

$$\max\{|r_1(\epsilon) - 1|, |r_2(\epsilon) - 1|\} = |r_1(\epsilon) - 1| \tag{11}$$

$$= \frac{o(\epsilon^{1/2}) + \epsilon^{1/2}\sqrt{4 + o(1)}}{2} \tag{12}$$

$$= \frac{\left(o(1) + \sqrt{4 + o(1)}\right)\epsilon^{1/2}}{2} \tag{13}$$

令 $C = 1, q = 1/2$, 则有, 当 $\epsilon \to 0$ 时,

$$\max\{|r_1(\epsilon) - 1|, |r_2(\epsilon) - 1|\} \approx C\epsilon^q$$

成立.

Solution 2.8. 定义函数 $r(a_k) = r$, 那么在多项式 $p(a_k)$ 的左右两端对 a_k 求导得到

$$\frac{dr}{da_k} = \frac{-r^k}{p'(r)},$$

再代入条件数定义即可得证.

3 Algorithms

Solution 3.1. 代码参考 Gitee 仓库.

Solution 3.2. 代码参考 Gitee 仓库.

Solution 3.3. 代码参考 Gitee 仓库.

这里我们解释一下题中的公式. 这个公式是基于格林定理 (Green's theorem) 推导出来用于计算多边形面积的公式。

格林定理建立了平面区域上的二重积分与围绕该区域边界的曲线积分之间的联系，其形式为:

$$\iint_D \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\right) dA = \oint_{\partial D} P dx + Q dy$$

其中 D 是平面区域，∂D 是 D 的边界曲线，P 和 Q 是关于 x 和 y 的函数.

对于计算多边形面积的情况，我们可以选择 $P = -\frac{1}{2}y$ 和 $Q = \frac{1}{2}x$，此时 $\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} = 1$. 那么格林定理就能导出

$$\iint_D dA = \frac{1}{2}\oint_{\partial D} x dy - y dx$$

等式左边的 $\iint_D dA$ 就是区域 D（即多边形）的面积.

对于一个多边形，我们可以将其边界曲线 ∂D 看作是由 n 条线段组成. 对于每一条线段从顶点 (x_k, y_k) 到 (x_{k+1}, y_{k+1}) (注意，这里 $x_{n+1} = x_1$ ，$y_{n+1} = y_1$)，曲线积分

$$\int_{(x_k,y_k)}^{(x_{k+1},y_{k+1})} x dy - y dx$$

可以通过参数化线段并计算积分得到. 简化后，多边形的面积 A 就可以表示为

$$A = \frac{1}{2}\left|\sum_{k=1}^{n} x_k y_{k+1} - x_{k+1} y_k\right|.$$

4 Stability

Solution 4.1. 参考计算条件数的公式以及书中 Example.

(a) 根据相对条件数的公式，有

$$\kappa_f(x) = \left|\frac{x f'(x)}{f(x)}\right| = \left|\frac{x}{\sin x}\right|$$

显然 $\kappa_f(x) \to 1$，当 $x \to 0$.

(b) 参考书中 Table 1.2.1，一步一步地算，有表 1. 我们发现，在 $1 - \cos(x)$ 时出现了 subtractive cancellation，不好.

(c) 类似 (b) 容易得到. 此时避开了 subtractive cancellation，是更优的算法.

表 1: $f(10^{-6})$ 的计算结果和条件数

Calculation	Result	κ
$u_1 = \cos(10^{-6})$	0.9999999999995	1×10^{-12}
$u_2 = 1 - \cos(10^{-6})$	$5.000444502911705 \times 10^{-13}$	1999822214639.54
$u_3 = \sin(10^{-6})$	$9.999999999998333 \times 10^{-07}$	10^{-6}
$u_4 = u_2/u_3$	$5.000444502912538 \times 10^{-07}$	1

(d) 根据 (b) 和 (c) 容易看出，$g(10^{-6})$ 更精确.

Solution 4.2. 代码参考 Gitee 仓库.

根据条件数公式计算 $\kappa_f(x)$，与上题类似比较两种算法. 其中计算多项式可以调用前面作业实现的函数.

Solution 4.3. 代码参考 Gitee 仓库.

根据条件数公式可以计算出给定点的 $\kappa_f(x_i)$，虽然这个问题是良态的，但是算法并不稳定，因为出现了 subtractive cancellation.