第1章

科学计算

CHAPTER 1

1.1 引言

本书的主题通常称为数值分析. 数值分析主要是对求解许多领域(特别是科学和工程)中所提出的数学问题的算法进行设计和分析. 因此,近年来数值分析也被称为科学计算. 科学计算与计算机科学的大部分内容是有区别的,前者处理的是连续量,而后者处理的是离散量. 科学计算所涉及的是一些函数和方程,其基本变量在实际中都是连续的,像时间、距离、速度、温度、密度、压强、应力等.

原则上说大多数连续的数学问题(例如,几乎任何一个问题都要涉及导数、积分或非线性等)都不可能通过有限步的计算得到完全精确的结果,它们的求解往往都通过某种迭代过程(理论上是无限的) 最终收敛到解. 当然在具体操作时,人们不可能永远迭代下去,而只能得到一个"充分接近"所期望结果的近似解,所以科学计算的一个非常重要的方面就是找出一个能迅速收敛的迭代算法,并能估计近似解的精确度. 下面将会看到,如果收敛足够快,即使对于像线性代数方程组那样用有限步算法就能解决的问题,在某种条件下可能用迭代法求解也会更好一些.

因此,科学计算的第二个特点是它的近似效果. 许多求解技术都涉及各种类型的近似序列,甚至于计算本身也是近似的,因为数字计算机不可能精确地表示所有实数. 一个好的算法除了具有通常意义所指的性质,如效率以外,还要尽可能地可靠和精确.

1.1.1 计算问题

顾名思义,许多科学计算问题都来自科学和工程,它们的最终目的是了解某种自然现象或设计某种设备. 计算模拟是用计算机来实现或模拟物质系统或过程. 如果理论、观察或实验的方法本身很困难或不能很好地完成研究的话,计算模拟可以大大增强对所研究问题的科学了解. 例如,在天体物理学中,两个黑洞碰撞的详细情况是非常复杂的,它不能从理论上确定,也不能直接观察或在实验室中再现,但可以用计算来进行模拟. 所需要的只是适当的数学表达式(爱因斯坦的广义相对论方程)、数值求解这些方程的算法及

足以实现这个算法的计算机.

计算模拟不仅可以处理某些特殊情形或解决用其他手段无法解决的问题,还可以以合理的成本和时间解决大量的可以由实验来解决的“常规”问题.在工程设计中,计算模拟可以迅速、经济、安全地对大量的设计方案进行实验,这要比传统的用物理原型进行组合测试的方法优越得多.本书中,计算模拟也称为*虚拟模型*.例如,在汽车安全性的改进过程中,在计算机上进行碰撞实验要比实际碰撞更便宜、更安全,并且可以得到非常精确的设计参数,从而实现最佳设计.

计算模拟解决问题的过程通常包括以下几步:

1. 建立一个关于物理现象或所研究系统的数学模型,这个模型一般都用某种类型的方程表示;
2. 给出数值求解这个方程的算法;
3. 用计算机软件实现这个算法;
4. 在计算机上运行这个软件,对物理过程进行数值模拟;
5. 用容易理解的形式表示计算结果,如图解法;
6. 解释并确认计算结果,如果需要,重复前面的某些或全部步骤.

步骤1通常被称为*数学建模*.它不仅需要应用数学的知识,还要对问题所涉及的有关科学和工程学科的知识有一定的了解.步骤2和步骤3包括了设计、实现、分析、应用数值算法和软件等内容,是科学计算的主题,也是本书的主题.虽然我们主要关注步骤2和步骤3,但是也应对所有步骤有所了解,即从问题的表示到结果的表示及确定,这样可以使结果更有意义,更有应用价值.我们将会看到,虽然在原理和方法上可以对科学计算做出非常广泛、非常概括的研究,但也要时常考虑到给定问题的具体来源以及得到的结果要用来干什么,这两者之间往往是相互作用的.例如,原始问题的表示可能会对数值结果的精确度有很大影响,而数值结果又会对解的描述和确认产生影响.

如果一个数学问题的解存在、惟一,且连续依赖于问题的数据,则称这个问题是*适定的*.后一个条件说明问题数据的微小变动不会使解产生急剧、不匀称的变化.稍后会看到,这个性质对于数值计算是极其重要的,因为这类扰动通常是不可避免的.虽然物质系统的数学模型非常需要这种适定性,但事实上并非总可以实现这一点.例如,只靠外部的观察推断一个物理系统的内部结构,往往会使得到的数学问题本身是不适定的.用层析X射线或测震仪进行观测时就会产生这样的数学问题,因为明显不同的内部结构往往可能有相同的外部表现.

但即使问题本身是适定的,解对问题数据的扰动也会具有高度敏感的响应(虽然连续).为了对这类扰动的影响作出估计,除了给出问题敏感性、连续性的概念外,还要定义一个关于敏感性的度量.另外,还要确保数值求解某个问题所使用的算法得到的结果不能比问题本身的结果更敏感(希波克拉底的誓言“不做有害的事”,除了给医生之外也可

以送给数值分析家). 这个要求引出了关于算法稳定的概念. 本章将引入这些概念和结果,并在后面的章节中针对不同类型的计算问题对它们加以详细的讨论.

1.1.2 一般策略

寻求某个计算问题解的一般策略是将复杂或困难的问题用同解或至少解相近的简单问题代替. 这种方案包括:

- 用有限维空间代替无限维空间;
- 用有限过程代替无限过程,如用有限和代替积分或无穷级数,用有限差分代替导数;
- 用代数方程代替微分方程;
- 用线性问题代替非线性问题;
- 用低阶方程组代替高阶方程组;
- 用简单函数,如多项式代替复杂函数;
- 用简单结构的矩阵代替一般矩阵.

例如,求解非线性微分方程组时,可以首先用一个非线性代数方程组来代替它,然后用线性代数方程组代替非线性代数方程组,再将线性方程组的矩阵用某种容易求解的具有特殊结构的矩阵代替. 整个过程中,每一步都要确保解没有改变,或解至少位于真解的某个限度内.

为了用这种一般策略解决问题,需要:

- 容易求解的另一个或另一类问题;
- 在某种意义下能将给定问题转化成另一类问题,并保持解不变的变换.

所以主要精力应放在如何将问题归结成简单的问题,即通过同解变换将其变为简单问题.

理论上,变换后问题的解与原来问题的解是相同的,但事实上并不如此. 这个解通常只能是原问题解的近似,但是通过附加的计算和存储,可使这个解达到任意的精度. 所以我们主要关心的是如何估计这个近似解的精确度及如何使其收敛到真解.

1.2 科学计算中的近似

1.2.1 近似的来源

科学计算中近似或不精确的来源很多,有些近似是在计算开始之前就出现的:

- 建模:这个过程可能简化或忽略问题或系统的某些物理特性(例如,摩擦、黏度、

空气阻力等).

- 经验测量：实验设备的精度是有限的. 它们的精确度有可能受到采样规模的限制，所得到的读数还有可能存在随机干扰或系统偏差. 例如，某些非常重要的物理常数，像牛顿万有引力常数或普朗克常数，即使非常精密的测量也只能精确到小数点后八九位，而大多数实验设备都达不到这个精确度.
- 前面的计算：输入数据可能是由前面的某些计算产生的，其结果只是近似值.

上面这些近似往往都超出我们的控制范围，但它们确实在确定计算的精确度时起重要作用. 我们要把注意力放在那些确实能够控制的近似上面，这类近似是在计算过程中产生的：

- 截断或离散化：数学模型的某些特性可能被忽略或简化，例如，用有限差分代替导数或在无穷级数中取有限项.
- 舍入：不论是手工、计算器，还是数字计算机，实数及实数运算都只能用有限精度表示，所以往往是不精确的.

计算的最后结果的精确程度将与上述任一因素或因素的综合作用有关，结果的扰动也可能由于所求解问题或所使用算法的特性被放大. 这种关于近似对数值算法精确度和稳定性影响的研究通常称为*误差分析*.

例 1.1　近似　地球的表面积可以通过半径为 r 的球体表面积公式

$$A = 4\pi r^2$$

计算，这个公式包括以下几种近似：

- 将地球看成球体，这是理想化的形状.
- 半径 $r \approx 6370\text{km}$ 可能由经验测量和前面计算得到.
- π 的值是由无穷极限过程得到的，只能截取到某一位.
- 在计算机和计算器上计算时，输入数据的值以及运算结果的值都可能有一定的舍入.

计算结果的精确度与所有这些近似有关. ■

1.2.2　绝对误差和相对误差

误差的意义与其度量或计算的量的大小有关. 例如，计算地球人口时误差 1 的意义要远远小于计算电话亭中人数时误差 1 的意义. 因此引出了*绝对误差*和*相对误差*的概念：

$$\text{绝对误差} = \text{近似值} - \text{真值},$$
$$\text{相对误差} = \text{绝对误差} / \text{真值}.$$

有些作者将绝对误差定义为上述值的绝对值，但这里只在需要误差大小的时候才取

绝对值(或向量及矩阵的范数). 如果真值为零,相对误差没有定义.

也可以用百分比表示相对误差,即将相对误差乘上 100. 例如,相对于真值 10 为 0.1 的绝对误差,其相对误差为 0.01,或 1%. 当相对误差至少是 1 或至少是百分之百时,所对应的近似将是完全错误的,这时意味着绝对误差与真解一样大.

相对误差的另一种解释为,如果某个近似值的相对误差为 10^{-p},则这个近似值的十进制小数表示有 p 个有效数字(*有效数字*包括首位非零数字和后面的所有数字). 这样,有必要区别精度和精确度:*精度*是与所表示的数的数字个数有关的,而*精确度*则与近似某个量时的有效数字的个数有关(即相对误差). 例如,3.252603764690804 是一个非常精确的数,但用它来近似 π 就不是精确的. 后面会看到,用给定的精度去计算一个量并不一定说明结果可以达到那个精确度.

绝对误差和相对误差之间关系也可以表示为

$$\text{近似值} = \text{真值} \times (1 + \text{相对误差}).$$

当然,一般并不知道真值,否则,就不用去近似它了. 由于不知道真值,所以一般只能估计或限定误差的范围,而不能准确地计算出其值. 因此,相对误差也只能是相对于近似值,而不是像前面定义的那样,相对于真值.

1.2.3 数据误差和计算误差

我们已经看到,有些误差是由输入数据产生的,有些则来自于计算过程. 虽然这种区别并不是很明确(例如,四舍五入既影响输入数据,也影响其后的计算),但它对了解数值计算中近似的总体影响还是有帮助的.

许多实际问题都是多维的,为了简化,本章只考虑一维问题. 有关定义及结果可以直接推广到多维空间,一般只需用适当的范数(见 2.3.1 节)代替绝对值即可. 一维空间中的典型问题是函数值的计算,即 $f: \mathbb{R} \to \mathbb{R}$,由输入值到输出值的映射. x 表示输入的真值,所求的真实结果为 $f(x)$. 假定所用的输入值是不精确的,记为 $\hat{x}$,这样就只能算出函数的近似值,记为 $\hat{f}$. 然后,利用加减同一个量总量不变这一标准的数学手段,有

$$\begin{aligned}\text{总误差} &= \hat{f}(\hat{x}) - f(x)\\ &= (\hat{f}(\hat{x}) - f(\hat{x})) + (f(\hat{x}) - f(x))\\ &= \text{计算误差} + \text{数据传播误差}.\end{aligned}$$

和式中第一项是同一个输入时精确函数值和近似函数值间的差,因而可以将其看成是纯*计算误差*. 第二项是由输入的误差所产生的精确函数值间的差,可以看成是纯的*数据传播误差*. 需要注意的是算法的选择对数据传播误差是没有影响的.

例 1.2　数据误差和计算误差　假定不用计算机或计算器,求 $\sin(\pi/8)$ 的近似值. 首

先需要 π 的值以确定输入. 在中学里学过，π 的古典近似为 22/7，但这个值无法转换成小数格式，所以用简单的“圣经”近似，$\pi \approx 3$，这样实际的输入是 3/8. 为了计算函数的值，利用微积分知识，自变量很小时可以用其泰勒级数展开的第一项近似函数，这样，$\sin x$ 的近似就是 x. 最后的结果

$$\sin(\pi/8) \approx \sin(3/8) \approx 3/8 = 0.3750.$$

第一个近似——用扰动输入 $\hat{x}=3/8$ 代替了真值 $x=\pi/8$——产生了数据传播误差. 由于使用了不正确的输入，所以即使精确地计算出正弦函数的值，得到的结果也是不正确的. 第二个近似是计算误差，尽管这里的“计算”只是复制了输入（计算误差往往是这类“省略的误差”，当然一般不会像本例题一样如此极端.）！利用前面的记号，用截断的数学表达式 $\hat{f}(x)=x$ 代替了真正的函数 $f(x)=\sin x$，这说明即使使用正确的输入也不会得到正确的结果. 总的精确度由这两种近似的组合决定.

接下来，用计算器算出正确答案，取 4 位小数，有

$$\sin(\pi/8) \approx 0.3827,$$

总误差为

$$\hat{f}(\hat{x}) - f(x) \approx 0.3750 - 0.3827 = -0.0077.$$

注意到对于扰动输入，正确答案为

$$f(\hat{x}) = \sin(3/8) \approx 0.3663,$$

所以，由不精确输入产生的数据传播误差为

$$f(\hat{x}) - f(x) = \sin(3/8) - \sin(\pi/8) \approx 0.3663 - 0.3827 = -0.0164.$$

由无穷级数截断产生的计算误差为

$$\hat{f}(\hat{x}) - f(\hat{x}) = 3/8 - \sin(3/8) \approx 0.3750 - 0.3663 = 0.0087.$$

这两个误差之和就是我们看到的误差. 对这个特殊的例子，两个误差的符号相反，所以它们有部分补偿；在其他情形，它们的符号可能相同，这时则相互增强. 对于这个特殊的输入数据，数据传播误差和计算误差的值大致相同，相差大约两倍，但对于其他的输入数据，任何一种误差都可能占优势. 对 π 及 $\sin x$ 做同样的近似，如果输入数据很小，数据传播误差可能占优势，而当输入数据较大时，计算误差又将占优势（为什么?）. 为降低整体误差，可以使用更精确的 π 值以降低数据传播误差，也可以使用更精确的 $\sin x$ 表达式（如在无穷级数中多取几项）以降低计算误差. ■

1.2.4　截断误差和舍入误差

计算误差（计算过程产生的误差）可以分为截断（或离散化）误差和舍入误差：

- *截断误差*是真实结果（对于实际输入）与用给定算法经精确运算得到的结果之间

的差，一般由无穷级数的截断、有限差分代替导数或在收敛之前中止迭代之类的近似产生.

- *舍入误差*是用给定的算法经精确计算得到的结果与用同样的算法经有限精度(即舍入运算)得到的结果之差.它是由实数的不精确表示以及对它们进行的操作引起的，在 1.3 节还要详细讨论.

由定义，计算误差是截断误差和舍入误差的和.在例 1.2 中，输入数据经过了舍入，但在计算过程中并没有进行舍入，所以计算误差只包含了用无穷级数的一项作为近似所带来的截断误差.如果在无穷级数中多取一些项，则可以降低截断误差，但在计算级数值时所做的运算还可能产生舍入误差.这种截断误差和舍入误差间的相互制约在数值计算中是经常的.

例 1.3　有限差分近似　对于可微函数 $f:\mathbb{R}\to\mathbb{R}$，考虑一阶导数的有限差分近似

$$f'(x)\approx\frac{f(x+h)-f(x)}{h}.$$

由泰勒定理得

$$f(x+h)=f(x)+f'(x)h+f''(\theta)h^2/2,$$

其中 $\theta\in[x,x+h]$.所以有限差分近似的截断误差估计为 $Mh/2$，其中 M 是 t 接近 x 时 $|f''(t)|$ 的上界.假定函数值的误差为 ε，计算有限差分公式的舍入误差为 $2\varepsilon/h$，则总的计算误差可由它们的和

$$\frac{Mh}{2}+\frac{2\varepsilon}{h}$$

估计，其中第一项随 h 的减少而减少，第二项随 h 的减少而增加.所以，在选择步长 h 时要权衡截断误差和舍入误差.将这个函数对 h 求导，并设导数为零，可以看到当

$$h=2\sqrt{\varepsilon/M}$$

时总的计算误差取最小值.图 1.1 所表示的是用有限差分近似计算函数 $f(x)=\sin x$ 在 $x=1$ 点的导数值时所产生的总的计算误差与步长 h 之间的函数关系，同时还分别画出了截断误差和舍入误差的界与步长 h 之间的函数曲线，其中取 $M=1$ 且所用计算机的精度为 $\varepsilon\approx10^{-16}$.可以看到，总误差在 $h\approx10^{-8}\approx\sqrt{\varepsilon}$ 处取最小值.在 h 值较大时，由于截断误差增加，总误差增加，而在 h 较小时，由于舍入误差增加，总误差也增加.

截断误差可以通过更精确的有限差分公式降低，如中心差分公式(见 8.6.1 节)

$$f'(x)\approx\frac{f(x+h)-f(x-h)}{2h}.$$

而舍入误差也可以通过使用精度更高的运算而降低. ■

对于给定的计算，虽然截断误差和舍入误差都起重要作用，但在整个计算误差中总是某一种误差成为主要因素.简单地说，有限次求解的纯代数问题中，舍入误差往往占主

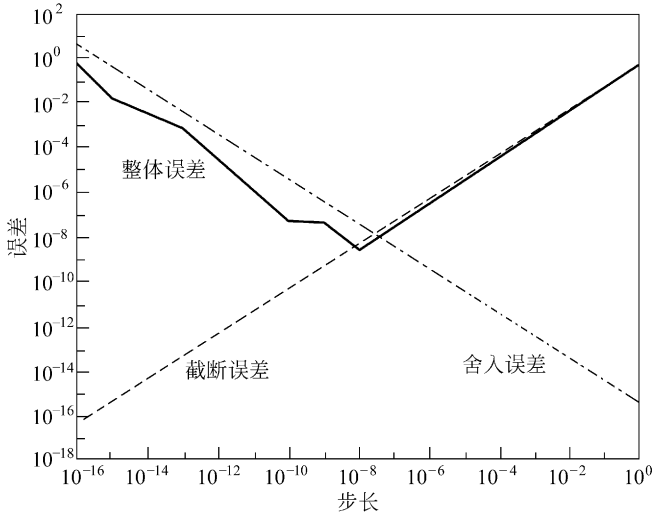

图 1.1　有限差分近似关于步长的计算误差

要地位，而对于涉及积分、导数、非线性化这类理论上是无限求解过程的问题，截断误差则占主要地位.

以上关于误差类型的说明对于了解数值算法的性态及影响它们精确度的因素是很重要的. 但是精确地给出每种类型的误差往往是没有必要甚至是不可能的. 后面将看到，通常的做法是将所有误差归并到一起，视它们为输入数据所引起的.

1.2.5　向前误差和向后误差

计算结果的质量与输入数据的质量及其他因素有关. 如果输入数据只精确到 4 位有效数字，那么无论计算多么精确，结果的有效数字也不会超过 4 位. 俗话说“废料进、废料出”就是这个逻辑. 所以在评价计算结果的质量时，不能忽略输入数据在不同程度上的扰动作用.

例 1.4　数据误差的影响　假如要预报某个国家 10 年后的人口数量，首先建立一个描述人口随时间变化的数学模型. 假定任一时刻出生人数和死亡人数与人口总数成正比，可建立一个简单模型

$$P(t+\Delta t) = P(t) + (B-D)P(t)\Delta t,$$

其中 $P(t)$ 表示 t 时刻的人口，Δt 是某个时间间隔(假定为 1 年)，B 和 D 是出生率和死亡率(例如，净增长率为 $B-D=0.04$，或每年 4%). 固定时间间隔 Δt，可得到离散模型，或者令 $\Delta t \to 0$ 取极限，得到微分方程

$$\mathrm{d}P(t)/\mathrm{d}t = (B-D)P(t).$$

方程的解为著名的指数增长定律

$$P(t) = P(0)\exp((B-D)t).$$

对任何情况，模型只是对实际的一种近似．例如，这里我们忽略了人口移居及容量限制的影响．为了阐述问题方便，这里并没有考虑此类模型误差，但这类误差在任何实际科学问题中都是存在的，所以就不必苛求接下来的计算精确度．不管用离散模型还是连续模型，都要给出当前的人口数量及出生率和死亡率．任何一次人口普查都不可能逐个去数每个人，且不发生遗漏，所以人口的初值只能在一定的精确度范围内．相应地初始人口通常用“舍入数字”（即只有几位有效数字）表示，这样可以表现出其不确定程度．但不能将这个过程看成舍入误差，因为去掉的位数并不一定是可信的，也许根本就是毫无意义的．类似地，可以用若干离散事件的平均值代表出生率和死亡率，所以它们的精确度也是在一定范围内的．

模型中输入数据的这些不确定性，在某种程度上意味着所得到的人口方案也是不确定的．通常，我们可以把模型看作在输入空间的模糊区域上与输出空间的模糊区域上建立的关系（换句话说，在不确定区域上，输入数据的所有可能的结合得到的所有可能的结果）．在实现这个模型时，可能会产生某些计算误差（截断误差或舍入误差），但只要产生的结果仍能落在相对于不确定输入数据的输出空间的模糊区域中，结果就是可信的．换一种说法，任何结果，无论它怎样得到，都可以看成是由某种输入而得到的精确结果，而这个输入实际上就是我们能够期望得到的最好结果的那一个．■

总结一下，为了简化还是考虑一维问题．假定要计算函数 $y=f(x)$ 的值，其中 $f:\mathbb{R}\to\mathbb{R}$，但得到的只是一个近似值 $\hat{y}$．称计算值与真值之间的偏差 $\Delta y=\hat{y}-y$ 为*向前误差*．评估计算结果质量的一种方式是估计向前误差的相对值，这种做法简单与否依赖于具体的情况．但一般来说，分析计算过程中误差的向前传播是非常困难的，其原因后面将作介绍．而且，在每一步所做的最坏假定往往得到的是关于整体误差的最差估计．

另一种方案是将得到的近似解看成是某个修改问题的精确解，然后讨论为了得到这个实际解，修改问题与原始问题要相差多少．换句话说，初始输入的误差多大时才能解释最终计算结果的全部误差？具体地说，若 $f(\hat{x})=\hat{y}$，称 $\Delta x=\hat{x}-x$ 为*向后误差*，对这个相对量所做的估计称为*向后误差分析*．用这种观点，如果一个问题的近似解是其“邻近”问题的精确解，则认为这个近似是好的（即相对向后误差比较小）．事实上，如果邻近问题是在输入数据的不确定范围内，则计算解 $\hat{y}$ 可能就是我们所知道（或能知道的）的“真解”，因而是可信的．

上述关系的简图可由图 1.2 所表示，其中 x 和 f 分别代表精确的输入值和函数值，$\hat{f}$ 表示实际计算的函数值，$\hat{x}$ 表示能够精确地得出计算值的输入．由 $\hat{x}$ 的取法有等式 $\hat{f}(x)=f(\hat{x})$，实际上，正是这个条件确定了 $\hat{x}$．1.2.6 节将具体讨论向前误差和向后误差之间的关系．

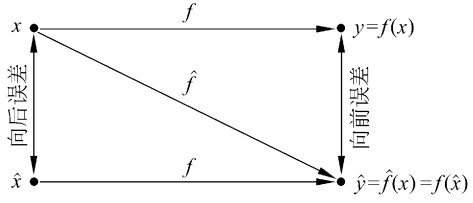

图 1.2 向前误差和向后误差的图示

例 1.5 向前误差和向后误差 $\hat{y}=1.4$ 是 $y=\sqrt{2}$ 的近似，它的绝对向前误差

$$|\Delta y|=|\hat{y}-y|=|1.4-1.41421\cdots|\approx 0.0142,$$

或者说相对误差大约为 1%. 为确定向后误差，利用 $\sqrt{1.96}=1.4$，其绝对向后误差为

$$|\Delta x|=|\hat{x}-x|=|1.96-2|=0.04,$$

或者说相对向后误差为 2%. ■

例 1.6 向后误差分析 求余弦函数 $y=f(x)=\cos x$ 在 $x=1$ 处的近似值. 余弦函数的无穷级数表示为

$$\cos x=1-\frac{x^2}{2!}+\frac{x^4}{4!}-\frac{x^6}{6!}+\cdots,$$

截去级数中第二项之后的项，得到近似值

$$\hat{y}=\hat{f}(x)=1-\frac{x^2}{2}.$$

这个近似值的向前误差为

$$\Delta y=\hat{y}-y=\hat{f}(x)-f(x)=1-\frac{x^2}{2}-\cos x.$$

为确定向后误差，需要找到一个使输出值为 $\hat{y}$ 的输入值 $\hat{x}$，即 $\hat{f}(x)=f(\hat{x})$. 对这个函数，该值可由

$$\hat{x}=\arccos\hat{f}(x)=\arccos\hat{y}$$

给出. 对 $x=1$，有

$$y=f(1)=\cos 1\approx 0.5403,$$

$$\hat{y}=\hat{f}(1)=1-\frac{1^2}{2}=0.5,$$

$$\hat{x}=\arccos\hat{y}=\arccos 0.5\approx 1.0472,$$

$$\text{向前误差}=\Delta y=\hat{y}-y\approx 0.5-0.5403=-0.0403,$$

$$\text{向后误差}=\Delta x=\hat{x}-x\approx 1.0472-1=0.0472.$$

由于输出值非常接近预期的结果，所以向前误差表明近似值的精确度还算可以，同时，由于对带有微小扰动的输入，仍得到了正确的结果，所以向后误差也表明这个近似值的精确度还算可以. 接下来，将定量地考虑向前误差和向后误差间的关系. ■