Serious Math for Adults
给成年人的零基础数学进阶

吴金闪

数学是思维的语言，描述世界的语言

像数学家一样提出和思考问题

从现实到结构，从结构到现实

纯粹理性的愉悦

有系统，有联系，知识网络

教的更少，学得更多

List of changes

批注清单

1 在几何那一节设计一个相应的例题或者习题。 90
2 之前插入弧度和极坐标。 118
3 给出具体章节。 156
4 具体章节，在前面补上从集合构建自然数 162
6 这张图可能也得单独印一页，横版，或者折页。 339

目录

第一部分　初等数学篇 **25**

第一章　从数的提出看模式和抽象 **27**

1.1　集合以及自然数基于集合的定义 27
1.2　自然数的提出 38
1.3　等式的性质 42
1.4　记号、序数和基数 44
1.5　模式和抽象 47
1.6　自然数的公理化定义 47
1.7　推荐阅读材料 47
1.8　作业 47
1.9　本章小结 47

第二章　从四则运算的提出看模式和抽象 **49**

2.1　四则运算的提出和相互关系 49
2.2　模式和抽象 49
2.3　数的进制 49
2.4　数的扩张和封闭性 49
2.5　计算是标准化了的思考过程 49
2.6　运算律 50
2.7　从计算和计算的封闭性来定义数 50
2.8　推荐阅读材料 50
2.9　作业 50
2.10　本章小结 50

第三章　从整除看逻辑演绎证明 **51**

3.1 自然数的整除 51
3.2 几个数的整除性质及其证明 51
3.3 什么是逻辑演绎证明——初步 51
3.4 推荐阅读材料 51
3.5 作业 51
3.6 本章小结 51

第四章 从用符号表示对象看数学是语言 53
4.1 对象的层次 53
4.2 模式和抽象 53
4.3 表示论初步 53
4.4 推荐阅读材料 54
4.5 作业 54
4.6 本章小结 54

第五章 变量和常数——从数到量到代数式 55
5.1 完全已知的数 55
5.2 给定的（当作已知）的数 55
5.3 未知的或者重点关注的数 55
5.4 多项式的元、阶和解 55
5.5 推荐阅读材料 55
5.6 作业 55
5.7 本章小结 55

第六章 映射和函数 57
6.1 用于刻画集合内部结构的映射 57
6.2 表示是一种映射 57
6.3 函数 57
6.4 位置（速度）-时间函数 57
6.5 推荐阅读材料 57
6.6 作业 57
6.7 本章小结 57

第七章 从表示论看数学是语言 **59**
7.1 硬币翻转的数学描述 59
7.2 正三角形转动的数学描述 59
7.3 集合和群——对象和操作 59
7.4 操作的矩阵表式 59
7.5 对称性和相似变换 59
7.6 经典对象上的博弈的数学描述 60
7.7 当作对象属性或者过程的数 60
7.8 数当作加法和加法的表示 60
7.9 数学是对操作的研究对关系的研究 60
7.10 推荐阅读材料 60
7.11 作业 60
7.12 本章小结 60

第八章 存在性、唯一性以及求解 **61**
8.1 勾股定理和三角函数 61
8.2 等式的性质以及数的唯一性 61
8.3 三角形的稳定性和三边确定的三角形的唯一性 61
8.4 已知两边一角的三角形的存在性和唯一性 61
8.5 方程的解的存在性、唯一性和求解 61
8.6 推荐阅读材料 61
8.7 作业 61
8.8 本章小结 61

第九章 从存在性、唯一性和计算分开看分解和综合 **63**
9.1 方程的解的存在性、唯一性和计算 63
9.2 分解和综合 63
9.3 不动点的存在性、唯一性和计算 63
9.4 推荐阅读材料 63
9.5 作业 63
9.6 本章小结 63

第十章 从平面几何学看知识体系和逻辑演绎证明 65
10.1 《几何原本》的知识体系和逻辑演绎证明 65
10.2 现代数学的知识体系和逻辑演绎证明 65
10.3 现代物理学的知识体系和逻辑演绎证明加实验检验 65
10.4 从三角形到多边形看分解和综合 65
10.5 推荐阅读材料 65
10.6 作业 65
10.7 本章小结 65

第十一章 逻辑学和数学的关系 67
11.1 概念 67
11.2 命题 67
11.3 逻辑演绎证明 67
11.4 逻辑学的学科责任 67
11.5 数学归纳法 68
11.6 不完全归纳法 68
11.7 集合本身没有定义，怎么办? 68
11.8 推荐阅读材料 68
11.9 作业 68
11.10 本章小结 68

第十二章 从自然数的严格定义看数学的知识体系和逻辑演绎证明 69
12.1 自然数的严格定义 69
12.2 自然数的集合定义 69
12.3 为什么有了自然数的“自然”定义还需要严格定义 69
12.4 推荐阅读材料 69
12.5 作业 69
12.6 本章小结 69

第十三章 空间点和空间矢量的数学模型 71
13.1 空间的点和数——笛卡尔坐标 71
13.2 空间矢量 71

13.3 解析几何 71
13.4 来自物理学的启发和数学建模 71
13.5 推荐阅读材料 71
13.6 作业 71
13.7 本章小结 71

第十四章 概率的数学结构 **73**
14.1 简单状态集合和每个简单状态的概率 73
14.2 概率语言下的状态 73
14.3 概率映射 73
14.4 可列可加性 74
14.5 从概率的数学结构看到双向数学建模 74
14.6 从概率的数学结构看到代数是对关系的刻画 74
14.7 概率分布函数 74
14.8 推荐阅读材料 74
14.9 作业 74
14.10 本章小结 74

第十五章 数学解题四问 **75**
15.1 数学应用题的 WHWM 75
15.2 几何题的 WHWM 75
15.3 WHWM 背后的系统思维 75
15.4 一般性 WHWM 75
15.5 推荐阅读材料 75
15.6 作业 75
15.7 本章和本卷小结 75

第二部分 高等数学篇 **77**

第十六章 微积分 **79**
16.1 导数 80
16.1.1 导数受到的启发 80

16.1.2 导数计算的几个定理 93
16.1.3 不定积分计算的几个定理 97
16.1.4 基于导数的近似——Taylor 公式 99
16.2 定积分及其与导数的关系 102
16.2.1 定积分受到的启发 102
16.2.2 dx, dy, dF 独立有意义——微分 119
16.2.3 定积分计算的几个定理 121
16.3 微积分学习总结 124
16.4 积分和导数计算的计算机软件 125
16.5 微积分的应用 128
16.6 从微积分看双向数学建模 145
16.7 微积分体现的数学学科大图景 147
16.8 严格形式的微积分 149
16.8.1 极限 149
16.8.2 从极限到导数和积分 179
16.8.3 微积分基本定理 188
16.8.4 神奇的全微分：通过线性部分就可以得到原始函数 . . 196
16.9 微积分和人类思考深刻问题能力的爆发 200
16.9.1 基于导数的方程求解方法 201
16.9.2 力学——微分方程 203
16.9.3 自然进入问题的解的边界条件 207
16.9.4 勒让德变换 211
16.9.5 最优控制 215
16.9.6 傅里叶级数 217
16.10 推荐阅读材料 223
16.11 作业 224
16.12 本章小结 225

第十七章 矢量代数 229
17.1 矢量受到的启发 230
17.2 欧氏空间的抽象矢量表示 238
17.3 群环域和矢量空间 242

17.3.1 群的定义 243
17.3.2 群的定义受到的启发 248
17.3.3 域的定义 250
17.3.4 伽罗瓦理论 * 253
17.3.5 矢量空间的定义 253
17.3.6 矢量空间上的内积 257
17.3.7 矢量空间的基矢 * 261
17.4 矢量空间上的线性算符：矩阵乘法规则为什么这么奇怪 . . . 264
17.4.1 $V^* \to V^*$ 上的线性映射 * 269
17.4.2 矢量空间的基矢变换 272
17.5 表示论和广义表示论，数学建模 276
17.6 矩阵的谱分解 280
17.7 矩阵的 SVD 分解 290
17.8 算符当作矢量和算符内积 293
17.9 最大本征值对应的本征矢量和幂方法 296
17.10 线性方程的解 300
17.11 线性代数的典型应用场景 303
17.11.1 边值问题、函数和算符，解析解和数值解 303
17.11.2 矩阵的 SVD 分解的应用 310
17.11.3 矩阵的谱分解的应用 313
17.11.4 群、矢量和张量在物理学里面的应用 316
17.11.5 矢量和自然语言处理 324
17.11.6 矢量和概念空间 328
17.11.7 为什么仅仅是线性的线性代数这么通用 329
17.12 从矢量看双向数学建模 332
17.13 线性代数和人类思考抽象问题能力的爆发 335
17.14 推荐阅读材料 337
17.15 作业 337
17.16 本章小结 338

第十八章 概率论——提供有关概率这个数学对象的基本概念以及定理341
18.1 语言统一之后的好处 341

18.2 随机变量是什么——可测函数 342
18.3 概率密度分布函数 342
18.4 随机变量函数的概率密度分布函数 342
18.5 各阶矩、特征函数和中心极限定理 342
18.6 推荐阅读材料 342
18.7 作业 342
18.8 本章小结 342

第十九章 统计学——从样本数据到结论（到分布函数） 343
19.1 统计学的学科典型责任 343
19.2 最简单愚蠢的办法——重复生成样本 344
19.3 重复生成样本的低成本替代——Bootstrap 方法 344
19.4 已知分布函数的情况——参数估计 344
19.5 统计回归分析 344
19.6 Monte-Carlo 方法 344
19.7 统计学哪些地方需要概率论 344
19.8 推荐阅读材料 344
19.9 作业 344
19.10 本章小结 344

第二十章 变分法求极值及其应用 345
20.1 极值问题的欧拉-拉格朗日方程 345
20.2 用度规和微积分重构平面几何知识 346
20.3 最小作用量视角的力学和光学 347
20.4 变分法重写控制论 347
20.5 从函数微分到变分 347
20.6 推荐阅读材料 347
20.7 作业 347
20.8 本章小结 347

第二十一章 数学知识网络和学科大图景 349
21.1 数学一二层知识的系统性 349

21.2 数学学科大图景，及其和一二层知识的联系 349
21.3 从数学知识体系到科学知识体系 349
21.4 现代数学在干什么? 349
21.5 推荐阅读材料 350
21.6 作业 350
21.7 本章小结 350

第二十二章 双向数学建模 351
22.1 数学建模的神——对现实对象及其上操作的表示 352
22.2 典型模型及其建立过程举例 352
22.3 数学建模五步 352
22.4 已有典型模型 352
22.4.1 优化模型 352
22.4.2 矩阵模型 352
22.4.3 微分方程模型——变化率和累积量的问题 352
22.4.4 随机过程模型 352
22.4.5 博弈模型 352
22.5 推荐阅读材料 352
22.6 作业 352
22.7 本章和本卷小结 352

第二十三章 本书总结以及现代数学在干什么 353
23.1 像数学家一样提出和解决问题 353
23.2 数学建模五步 353
23.3 数学解题四问 353
23.4 从重复性使用到创造性使用到创造 353
23.5 现代数学在干什么? 353
23.6 推荐阅读材料 353

参考文献 358

名词索引 359

人名索引 **363**

插图目录 **366**

举例目录 **369**

作业目录 **371**

献给

我的父亲，吴迪平。

致谢

序

前言

本书是是“人类知识高速公路上以培养掌握高层知识生成器的创造者为目标的理解型学习”（简称“理解型学习”[1]）应用于具体学科的另一个例子。在这里我们把从小学到中学、大学、研究生院层次的数学知识做了梳理，用新的方式做了反映数学学科大图景的呈现。

一个学科的学科大图景指的这个学科的典型研究对象、典型研究问题、典型思维方式、典型分析方法。以及这个学科和世界以及其他学科的关系。在这个新的呈现里面，我们强调了学习是为了创造知识、创造性地使用知识，以及欣赏知识的创造和创造性使用，并且企图用理解型学习的方式来通过具体知识的学科之中体会好数学学科大图景，从而达成上面的学习的目的。

我们来补充一点点关于理解型学习的知识。在这个学习方法之中，最重要概念是知识的 $0+5$ 个层次：第零层前知识，经验和体验；第一层，事实性程序性知识；第二层，学科概念知识；第三层，学科大图景知识；第四层，一般性人类思维；第五层，教和习的方法。在学习三角形之前就看过摸过三角形，在学习加法之前就在生活中把两堆东西和起来数过，这成为经验和体验。乘法口诀、三角函数公式、题型和解题步骤口诀属于第一层知识。乘法是重复加法的简便记号、三角形是三条线段首尾相连构成的封闭图形、三角函数的定义以及各个三角函数之间的关系属于第二层知识。数学是思维的语言，加法来自于对“合起来数一数”的抽象，数来自于对“数 + 量 + 物”的模式的抽象，每个题型及其解题步骤的梳理总结的过程（提出或者被提出问题；找到这个问题中的对象，识别对象之间的关系，把关系转变成数学表达式，把所要解决的问题变成数学问题；求解问题；检验答案；进一步推广到一类问题。有的时候也被称为数学建模五步，或者简称数学

五步 [2]）属于第三层知识。例如从加法（和起来数一数，从经验和体验到学科概念）和乘法（重复的加法，为了更加简便的语言和记号）的概念的定义方式体会到抽象这个思维方式在数学中的重要性，体会到概念之间的联系，体会到不同层次知识之间的联系，进一步看到系联性思考（看到联系，通过联系从个体看到整体，从整体的角度来看个体），从证明中体会到批判性思维（在数学上，任何公理以外的命题必须经过我自己的严格证明，才能被我接受），进而把抽象、系联性思考、批判性思维迁移到其他学科的知识的理解和创造之中，属于第四层知识。通过知识之间的联系从而实现上下左右贯通来学习，通过具体知识的学习体会到学科大图景，以及从是否能够把学科大图景体现好的角度来看每一个具体知识的学习，是有关“理解型学习”这个教和习的方法的知识（甚至，从这样通过具体数学知识的学习体会到理解型学习，从理解型学习的角度来设计具体数学知识的学习，也是理解型学习的例子），属于第五层知识。

更进一步，我们把这样的通过不同层知识之间的关系来学习和思考，也就是从对相对低层知识的总结提炼中学习、体会甚至创造相对高层知识，从相对高层知识来指导相对低层知识的学习、理解和创造，称为上下贯通；我们把通过同一层知识之间的联系来理解好这一层知识，以及把来自于一个领域的高层知识用于学习着解决或者创造着解决另一个领域的问题称为左右贯通。我们可以借助概念地图（概念放在框内的和概念通过没有放入框内的连词相连，例如“加法–重复多次以后成为–乘法”、“减法–重复多次以后成为–除法”、乘除法关系和加减法关系被抽象为逆运算–体现了–抽象和具体化、抽象和具体化–指导了–乘除法关系和加减法关系被抽象为逆运算）来呈现和运用上下左右贯通，如图 1。各个学科的概念地图的总和被称为人类知识高速公路。正是由于第三层和第四层知识具有帮助解决本领域或者其他领域问题，创造知识的能力，我们把它们称为高层知识生成器。

因此，知识的层次中的第三、四、五层知识，也就是高层知识生成器；人类知识高速公路，也就是各个学科的概念地图的总和；上下左右贯通，也就是从低层知识中总结体会到高层知识，从高层知识来看低层知识，运用高层知识来学习着或者创造着解决本领域或者其他领域的问题从而学习或者创造知识，是理解型学习最重要的核心概念和理念。更多信息可参见吴金闪的《教的更少，学得更多》[1]。

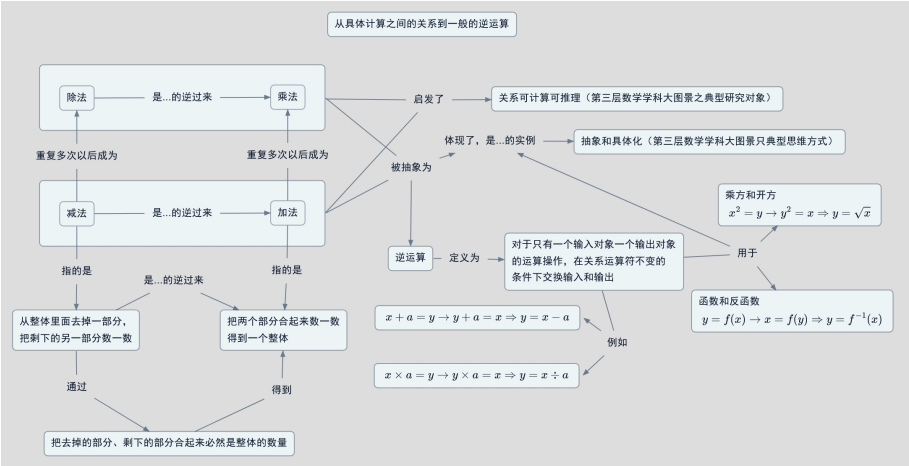


图 1: **用于描述数学概念间关系的概念地图举例**。这张图本身的数学内容暂时不用仔细学习，只需要看到概念地图可以通过概念和连词来描述概念之间的关系。

因此，本书目的就是用理解型学习来帮助已有大量经验体验的对理解数学和学会像数学家一样来提出和思考问题、对于体验到数学的美、体验到数学思维的美和苦恼有严肃的需求和要求的成年学习者学习数学——从零基础到象数学一样来提出和思考问题，同时也学会理解型学习，甚至未来进而用理解型学习来学习其他知识，拥有更大的机会成为知识的创造者和创造性使用者。或者，至少也成为数学知识的创造和创造型使用的欣赏者。

写给成年读者们的

对经验体验和意志品质的要求

兴趣要求很高，思维主动性和链长度要求很高，知识基础基本没要求。

如果你不想成为数学知识本身或者基于这里学会的数学知识来学会其他知识从而成为其他知识的创造者和创造性使用者，甚至都成为数学知识以及其他知识的创造和创造型使用的欣赏者，那你不是本书的读者。

写给成年读者的教师们的

你得能理解到位数学的学科大图景，能够深入浅出。同时，这事儿有意义——帮助成年人拥有再一次成为知识的创造者和创造性使用者，或者知识的创造和创造性使用的欣赏者的机会，尤其在企业中，以及在人工智能高度发展的时代——深入思考创造体验创造的痛苦和快乐在这个时代会更加重要。

本书所体现的数学学科大图景

前面我们提到了本书的核心思想通过“人类知识高速公路上以高层知识生成器为目标的理解型学习”来实现“教和学的更少，学得更多”。对于数学学习，这里的高层知识生成器指的是第三层知识数学学科大图景和第四层超越学科的一般性人类思维，以及教和学的方法。为了实现这个“人类知识高速公路上以高层知识生成器为目标的理解型学习”，我们大概需要完成以下步骤 [1]：

1. 梳理这个学科的学科大图景；
2. 选择体现这个学科大图景的核心概念、概念间关系来构建学科概念网络；
3. 选择合适的案例来体现每个概念的提出和运用，促进对概念的理解；
4. 标注书籍、论文、项目、案例、习题等到这个学科概念网络上；
5. 结合概念网络上的分析算法（例如学习顺序算法、自适应诊断性检测算法）来设计和开展教学实践。

这其中最关键的就是对学科大图景的梳理。那么，数学学科大图景，也就是数学的典型研究对象、典型研究问题、典型思维方式、典型分析方法、和世界还有其他学科的关系，到底是什么呢？我们在《小学数学这样学》[2] 中有专门的一章来展开讨论。在这里，我们把其中的主要意思复述一下。更完整的版本请参考吴金闪的《小学数学这样学》[2]。

- 数学观
  - 数学是思维的语言
  - 数学是描述世界的语言，“双向”（从现实提炼数学结构，把数学结构用于描述现实）数学建模
  - 数学是研究关系和模式的学科
- 数学思维
  - 抽象是数学的核心思维
  - 批判性思维在数学中的地位：证明和计算是数学最重要的论证方式
  - 表示数学结构和研究表示间的转换是的数学研究的目标和研究方式
  - 分解和综合是数学结构研究的典型方式
  - 数学知识有系统：公理化体系，层次性和跨越层次的联系，不同数学分支的融合
  - 化归的思想
- 分析方法
  - 数学建模五步
  - 数学解题 WHWM 四问，及其嵌套合并
  - 试错和分析错误是数学的重要分析方法之一
  - 近似是数学重要分析方法之一
  - 微积分——累积量和变化率的关系
  - 无穷小和极限的概念
  - 连续性及其分析方法
  - 变换和不变量的思想和相应的分析方法
  - 约束下求极值

- 最基础的概念（其他高级结构的基础）
  - 集合和映射
  - 范畴论
  - 数理逻辑：什么是演绎证明，为什么能用

本书会围绕以上内容来决定教什么怎么教，在可能的情况下把数学学科大图景体现在具体知识的学习之中，把理解型学习体现在数学的学习之中。

其实，这个具体知识和学科大图景以及超越学科的一般性人类思维还有教和学的方法的关系，就是系统思维 [3]：通过直接和间接联系，从个体看到整体，从整体的角度来看个体。我们把它叫做“系联性思考”。不仅仅是学科学习，在面对任何系统的任何问题的时候，系联性思考——也就是通过把这个系统在各个层次做分解和综合，看到在各个层次上各个部分之间的联系，进而从整个系统的整体功能的角度来看其每一个部分的作用和地位，从这个系统的每一个部分来看到这个系统的整体的功能——往往可以给我们提出和解决问题的启发。而且，更进一步，系统科学还在此思维方式的基础上，发展了一些分析方法，解决了一些典型问题 [3]。因此，我们也希望通过这本书，你不仅仅学会数学，体会到数学学科大图景，学会理解型学习，还能体会到系统思维的核心“系联性思考”，而且都是通过做中学、创造中学，而不是通过这些高层知识的被说教，或者被强迫记住，来学习。

第一部分

初等数学篇

第一章　从数的提出看模式和抽象

数的提出有两个来源，对现实世界的经验和体验以及日常语言的提炼，也就是对现实世界做抽象，以及通过对数学的最基本的研究对象先做抽象，得到一个叫做集合的数学结构，然后对集合做进一步的思考构建更加复杂的数学结构而来。

通常来说，我们小时候学习数学是先走的第一条路。今天，你已经是成年人（或者我要求你小学五年级以上），我们先来走第二条路。走第二条路的好处是，上来就能看到数学不是研究数和形的，而是研究任意的对象的，数学为描述任何对象和解决关于任何对象的问题提供语言。这一点是非常重要的，称为数学学科的典型责任。

当然，更加悲惨的情形是小时候也没有走第一条路：老师和教材根本就没有教你什么是 $0,1,2,3,\cdots$ 的含义，而仅仅教给你了 $0,1,2,3,\cdots$ 的记号。在我们这里，这是绝对不可接受的。你不仅仅要理解每一个数学概念、数学符号的含义，还要体会到这些概念是为什么会被提出来，会被定义成这样的，甚至往往还要明白这个数学概念为什么叫这个名字。

1.1　集合以及自然数基于集合的定义

学习任何一个学科，首先就要了解其研究对象。以后我们会知道数学的研究对象可以是任何东西，数学为描述任何东西和解决任何问题提供语言，为人类的思考提供含义明确的语言，可以用来帮助进行长步骤的深入的思考的语言。

既然这样，数学就要先发明一套描述这样的任何东西的语言。数学经

过了多年的发展之后，基本上形成了以集合为这样的最基础的描述对象的语言。

集合的朴素含义就是一堆东西。比如说，我们会说桌子上的盘子里面有苹果。那么，盘子里面的苹果就是一个集合。我们有可能会说，我的笔袋子里面有笔。那么，笔袋子里面的笔就是一个集合。加上我们前面说过，数学的基本要求是含义明确。因此，集合就是含义明确的一堆东西。对于集合来说，含义明确指的是给我任何一个对象，我知道这个对象到底属于还是不属于这个集合。属于集合的东西就称为这个集合的元素。

这不是一个集合的定义，而是关于集合的一个描述，一个关于其要满足的性质的描述。我们暂时凑合着把它当作定义。

定义 1.1 (集合). 集合是一个满足元素确定性的一堆东西的整体。集合记做 A，其元素记做 x。这里的元素确定性指的是，对于任何一个 x，必须而且仅仅满足以下两条之一：x 是 A 中的元素，记做 $x \in A$ 和 x 不是 A 中的元素，记做 $x \notin A$。

再次强调，这个不算集合的定义，只算描述，或者称为集合的朴素定义。原则上，定义的要求更高：所谓一个关于某事物（通常有一个这个事物的名称甚至符号）的定义就是给我们任何一个对象我们得能够判断出来这个对象是否满足这个定义算这个事物。但是，以后我们会注意到，集合的这个说法不满足这个要求[1]。以后如果有机会进一步学习，你会遇到集合的真正的定义。

比如说，“桌子上的盘子里面的苹果”这个集合确实满足这个要求。已经放在这个盘子里面的苹果都属于这个集合，没有放在这个盘子里面的苹果不属于这个集合。你说，我再往盘子里面放一个苹果，会怎样？这时候，我们说在 t 时刻，你约定好了一个“桌子上的盘子里面的苹果”集合 A_t，在 τ 时刻，你约定好了一个“桌子上的盘子里面的苹果”集合 A_τ。这两个集合可以不一样，这没问题。只需要你在告知其他人的时候明确地把这个时间信息也带上就行了。比如说，

$$A_t = \{a_1, a_2\}, A_\tau = \{a_1, a_2, a_3\}. \tag{1.1}$$

[1]如果你着急，我可以先透露一下你认为“所有集合的集合”是一个集合吗？未来我们会发现这个问题会造成逻辑上的问题。如果你着急现在想了解更多，可以先去看看“理发师悖论”。

这表示在 t 时刻盘子里面有一个编号为 a_1 和一个编号为 a_2 的苹果，在 τ 时刻盘子里面有编号为 a_1, a_2, a_3 的苹果。注意，这里的 a_1, a_2, a_3 不过就是对这些对象的记号。我们完全可以替换成“最红的苹果，有虫口的苹果，青色的苹果，最大的苹果”等等这样的记号，只要这些记号对应着哪一个苹果是明确的。习惯上，我们可以用这样的列出来一个集合中的每一个元素的方式，用大括号括起来，来表示一个集合。

比如说，“我家里的人”这个集合也确实满足这个要求，但是，这个集合的具体包含的元素可能和回答这个问题的人有关。例如“我的爷爷”可能会把我的奶奶、我的叔叔、我的姑姑、我的爸爸、我的妈妈、我和我的妹妹，如果叔叔和姑姑家有婶婶和姑夫还有孩子的话，都算上；但是，“我”认为的“我家里的人”只包含我的爷爷、我的奶奶、我的外公、我的外婆、我的爸爸、妈妈、我和我的妹妹。不过，只要把到底包含哪些人说清楚，则这仍然是一个集合。“我家里的人”并不要求对于不同的“我”其包含的元素相同，只需要对于给定的“我”，其对象是明确的就可以了。

我家门前的那颗树上的枣，我家里养的狗，我身上传的衣服的兜里的硬币，这本书里面所有的汉字，这些都是集合。除了看得见摸得着感受得到的集合，例如某个给定的时刻池塘里面所有的已经被孵化出来的鱼也是一个集合，某个时刻一个人身上所有的水分子也是一个集合，两条直线相交得到的所有角的度数也是一个集合，尽管这些集合里面的元素你可能都难以一个个找出来[2]。

集合的另一个表示方式是明确写下来这个集合所包含的所有元素的共性特征。例如，这里的“我家门前的那颗树上的枣”就是通过对树和枣的约定来表示一个集合的。当然，这里默认我家门前只有一颗枣树，或者“那颗”在上下文中明确有所指。我们把这样的集合的表示方式记做，

$$A = \{x | P(x) = 1\}. \tag{1.2}$$

还是用大括号括起来，然后中间用一条竖线 | 分割，左侧就是这些对象的通用符号，右侧是这样的对象所要满足的性质 $P(x) = 1$，其含义是“关于 x 的某性质 $P(x)$ 成立”。

[2] “难以”不是“不可以”。

习题 1.1 (我的家庭成员). 分别用列举全部对象和用特征概括的方式来表述“我的家庭成员”这个集合。这里面的“我”可以替换成你自己、你的孩子、你的父母等等。

习题 1.2 (我的家庭成员). 对于用特征概括的方式来表述的“我的家庭成员”，请你尝试构建一个具有一定一般性的特征描述。这里的一般性指的是，如果都是爷爷的角色，是不是其定义的家庭成员的特征都是类似的；如果都是爸爸的角色，是不是其定义的家庭成员的特征都是类似的；等等等等。也就是让这个特征的定义跨越具体对象的种类而适用。甚至，如果你能够找出来一个跨越所有这些种类，或者更广泛一点点的种类的特征的描述就更好了。例如，对于爷爷和奶奶这个特征是相同的，或者说对于所有人其实都相同。比如说，我就尝试写下来这个集合的这样的特征描述：$\{x \mid P(x,I)=1, P(I,x)=1, M(I,x)=1\}$。其中 $P(x,I)$ 表示 x 是 I 的父母，$M(I,x)$ 表示 x 和 I 具有婚姻关系，I 表示的是我自己。当然，这个定义不一定就是你认为的家庭成员。我只是提供一个这样的建立起来一个通用的家庭成员的特征描述的例子。

从这里，你就应该能够体会到，数学是一个可以用于描述任何对象的语言了。集合对域所要描述的对象基本上没有什么限制。那么，还有什么东西不是集合的呢？比如说，所有的秃头、所有的美女、所有的帅哥都不是集合。为什么？什么样的算秃头、算美女、算帅哥，不仅仅和评价的人有关（如果那样的话约束一下叫做“我认为的秃头”就可以）和评价的时间点有关，甚至对于给定评价的人给定时间点，其判断也很难有一个可重复的标准。比如说，到底有 100 根头发以下的算秃头，还是有 1000 根头发以下的算秃头，还是说，还得看头发在整个脑袋顶上覆盖的面积？更严重的问题，以后我们会更加严肃地来讨论，是下面的关于秃头的推理的问题。

例 1.1 (秃头悖论). 证明：所有人都是秃头。

证明 1.1. 假设一个只有一根头发的人我们叫做秃头。然后，我们说，下面的假设也很合理：对于一个秃头，不管其头上有几根头发，如果仅仅增加一根头发，我们仍然认为其是秃头。在这个两个假设下，则所有人都是秃头。如果某个有 N 根头发的人不是秃头，我们就会看有 $N-1$ 根头发的人是不是秃头，假设还不是秃头，我们就继续看有 $N-2$ 根头发的人是不是秃头，

不断地这样做下去，我们总会去看有 1 根头发的人是不是秃头。根据第一个假设是秃头。然后，按照第二个假设，有 $1+1=2$ 根头发的人是秃头，于是有 $2+1=3$ 根头发的人是秃头，一直返回来会得到有 N 根头发的人是秃头。因为，无论你有几根头发，只要前面两个假设成立，则所有人都是秃头。

这显然不对。一个显然不对的结论可以被从某些看起来非常合理的假设证明的时候，这个结论就被称为悖论。你说这两条假设不合理？好想还真挺合理的。未来我们会知道其根本原因是秃头不构成一个集合。上面的论证方式——以后会正式介绍其名称“数学归纳法”——适用的前提是其对象对应着一个可数的集合。“可数”这个概念我们也稍后会介绍，大概来说就是对应着自然数的意思。不过，目前我们还没有自然数呢。

集合是一个数学对其研究对象的最最基本最最没有结构的描述。数学概念的世界里面，再也没有比集合更加基本的概念了。小结一下：集合就是对象明确的一堆东西。很多东西可以算做集合，有的东西不能算。注意，数学不是研究数和形的，它是研究任何对象的——在这里约定是任何一个可以算做集合的对象。未来我们会知道，只不过是很多集合往往可以被表示成数和形而已，以及历史上数学也确实曾经主要研究数和形，所以，习惯上才会有人认为数学是研究数和形的。

我们现在就来从集合走到数。有了集合我们就可以来定义自然数了。有一个特别特殊的集合，叫做空集，记做 ϕ，它什么都不包含，也就是任何对象 x 都不属于这个集合，也就是 $x \notin \phi$。接着，以这个空集为基础，我们来定义下面这样的集合们，

$$\phi, \{\phi\}, \{\phi, \{\phi\}\}, \{\phi, \{\phi\}, \{\phi, \{\phi\}\}\}, \cdots \tag{1.3}$$

然后，我们定义这些集合相应的大小为自然数：$0, 1, 2, 3, \cdots$。引入集合大小

的记号，我们就有了

$$|\phi| = 0, \tag{1.4}$$

$$|\{\phi\}| = 0 + 1 = 1, \tag{1.5}$$

$$|\{\phi, \{\phi\}\}| = (0 + 1) + 1 = 2, \tag{1.6}$$

$$|\{\phi, \{\phi\}, \{\phi, \{\phi\}\}\}| = (0 + 1 + 1) + 1 = 3, \tag{1.7}$$

$$\cdots \tag{1.8}$$

这里的“+”仅仅表示比前面的数还多 1 的意思，而 1 就表示 $|\{\phi\}|$，而后面更大的数就是约定的符号，比如“比 1 还多 1”这个集合大小就被记做集合大小为 2。如果你愿意，你可以坚持用 $|\phi|, |\{\phi\}|, |\{\phi, \{\phi\}\}|, \cdots$ 来表示 $0, 1, 2, \cdots$ 而不用 $0, 1, 2, \cdots$ 这些符号。

这样我们就从集合开始定义了自然数。在本书中，我们把零也当作自然数。有了自然数，我们就可以来定义任何包含元素有限的集合的大小了。对于成年人的你来说，就是数一数每一个集合里面有几个元素，就把元素的多少称作这个集合的大小。

更严格和抽象地来说，就是凡是和 $\phi, \{\phi\}, \{\phi, \{\phi\}\}, \{\phi, \{\phi\}, \{\phi, \{\phi\}\}\}, \cdots$ 的其中之一建立起来同构映射的集合的大小都是相应的前面约定的自然数。这里的同构的意思就是，元素之间可以建立起来一对一映射——稍后会更加明确地定义其含义。一个任何元素都不包含的集合是空集 ϕ，具体到空集之间的映射，就是两边都没有元素，因此也就不需要建立起来映射，就是一一对应的。对于更一般的集合之间，我们还需要定义映射。

下面我们来定义映射，还是从集合开始。给定两个不同的集合 A, B，或者同一个集合 A, A，可以建立一个集合的卡氏积，定义为

定义 1.2 (集合的卡氏积). 给定集合 A, B，下面的集合称做集合 A, B 的卡氏积，记做 $A \otimes B$，$\{(a, b) \mid \forall a \in A, \forall b \in B\}$。

所谓两个集合卡氏积就是从这两个集合里面各自去出来一个元素 a, b 组合起来成为 (a, b) 这样子的组合而得到的所有的组合构成的集合。

比如说，我们对 $A = \{1, 2\}$ 做卡氏积 $A \otimes A$ 会得到 $\{(1, 1), (1, 2), (2, 1), (2, 2)\}$。

一个从 A 到 B 的映射指的是满足如下的特定性质的 $A \otimes B$ 的子集，

定义 1.3 (映射). 给定集合 A, B 和集合 A, B 的卡氏积 $A \otimes B$，$f \subset A \otimes B$

称作从集合 A 到集合 B 的映射，记做 $f\big|_{A\to B}$，如果满足：$\forall a\in A$，$\exists! b\in B$，使得 $(a,b)\in f$。

为了熟悉这套语言，我们来把 “$\forall a\in A$，$\exists! b\in B$，使得 $(a,b)\in f$” 翻译一下：对于 A 中任意的一个元素 a，存在着唯一的一个 B 中的元素 b，组合起来以后得到的 (a,b) 得属于 f 这个集合。注意，f 是卡氏积 $A\otimes B$ 的一个子集。

这里 “存在唯一” 的含义是，首先得存在，其次还只有一个。如果 $b_1\neq b_2\in B$，对于某一个 a，(a,b_1) 和 (a,b_2) 都属于 f 这个集合，那就不满足要求了。或者某个 a，找不到任何一个 $b\in B$ 使得 (a,b) 属于 f，那就也不满足要求了。

朴素地说，映射就是对于任何一个 A 中的元素 a，都可以找到一个 B 中的 b，使得 a 经过这个映射以后成为 b，也就是 $f(a)=b$。

但是，上面的定义更加抽象，好处是只需要集合这个数学概念当基础。直观地说，卡氏积的操作使得 A 中的任何一个 a 都可以对应着 B 中所有的 b。而映射就是对于每一个这样的 a，只能从所有的 b 中选择一个当作其映射的结果。这个时候 b 称作 a 在映射 f 作用下的像。

注意，映射的这个定义并不要求给定一个 $b\in B$，一定有一个 $a\in A$ 使得 $(a,b)\in f$，也没有要求给定一个 $b\in B$，如果有只有唯一的一个 $a\in A$ 使得 $(a,b)\in f$。也就是说，给定一个 $b\in B$，可以是没有，也可以是多个，$a\in A$ 使得 $(a,b)\in f$。如果满足一定有一个，把这个映射称为满射，就是 B 被射满了的意思。如果满足只要有就唯一有，则称这个映射为单射。如果一个映射既是满射也是单射，就被称为可逆映射。这个时候，对于任何一个 $b\in B$，存在唯一的一个 $a\in A$，使得 $(a,b)\in f$。也就是反过来也是一个映射。所以称为可逆映射。

有了映射，我们来给任意一个有限大小的集合定义一个大小，也就是建立从集合到自然数的映射。给定一个集合 A，如果可以从这个集合 A 到 $\phi,\{\phi\},\{\phi,\{\phi\}\},\{\phi,\{\phi\},\{\phi,\{\phi\}\}\},\cdots$ 中的任何一个集合之间建立起来一个一一到上的映射，也就是可逆映射，则称集合 A 的大小为相应的那个建立起来可逆映射的集合的大小。例如，如果 A 和 $\{\phi,\{\phi\}\}$ 建立起来可逆映射，则 A 的大小就是 2，记做 $|A|=2$。

例如两个苹果的集合 $\{a_1,a_2\}$ 就可以建立一个这样的到 $\{\phi,\{\phi\}\}$ 的映

射，给 a_1 配上 ϕ，给 a_2 配上 $\{\phi\}$，也就是 $f=\{(a_1,\phi),(a_2,\{\phi\})\}$，或者说，$\phi=f(a_1)$，$\{\phi\}=f(a_2)$。

通过这样抽象的方式，我们就建立了对任何一个大小有限的集合的大小的数数的过程和数数的结果，也就有了从集合 A 到集合大小 $|A|$ 的映射。

既然对任意的有限大小的集合我们都定义了其大小——可以用自然数表示，我们来定义自然数的加法。我们说，当我们把两个集合里面的元素合起来都包含到一个新的集合里面，并且这两个原来的集合之间没有共同的元素的时候，新的集合的大小等于原来的两个集合的大小相加。

定义 1.4 (并集). 我们把两个集合的元素合起来构成一个新集合的操作称为集合的并，得到的集合称为这两个集合的并集，记做 $A\cup B$。

定义 1.5 (交集). 我们把两个集合的共同元素找出来构成的新集合的操作称为集合的交，得到集合称为这两个集合的交集，记做 $A\cap B$。

用这个语言，加法运算的基础是集合的并和交：先把两个集合的交算出来，得保证其为空集，然后在这个前提下，把这两个集合的并算出来，得到的并集的大小，就对应着这两个集合的大小相加。用前面学过的数学符号来表示就是，

$$|A|+|B|=|A\cup B|, \tag{1.9}$$

当且仅当

$$A\cap B=\phi. \tag{1.10}$$

可以证明，这样定义的加法和我们稍后要介绍的也是你熟悉的加法的定义——两个数的加法是把这两个数各自对应的对象合起来数一数——是等价的。

为了未来使用集合的语言方便，我们再定义集合的补集和集合的包含关系。

定义 1.6 (集合的包含关系). 如果集合 A 中的元素都属于集合 B，则称 A 被包含于 B，或者等价地 B 包含 A，记做 $A\subseteq B$。这时候，A 也被称作是 B 的子集。

如果 B 中还存在 A 中没有的元素，则称称 A 被真包含于 B，或者等价地 B 真包含 A，记做 $A\subset B$。这时候，A 也被称作是 B 的真子集。我

们来引入两个记号，任意 $\forall$ 和存在 $\exists$。这样，$A \subseteq B$ 可以被表述为

$$\forall x \in A, x \in B. \tag{1.11}$$

而 $A \subset B$ 可以被表述为

$$\forall x \in A, x \in B; \exists y \in B, y \notin A. \tag{1.12}$$

注意，数学符号是数学非常重要的部分。只有尽量用数学符号来表达含义，才能保证数学含义的明确性以及未来的对数学符号的可操作性，数学论证的严格性。我们要尽量避免用日常语言来表达数学含义。

定义 1.7 (补集). 如果 $A \subseteq B$，则那些属于 B 但是又不属于 A 的元素，当我们把 B 当作全集的时候，构成的集合称为 A 在 B 中的补集，记做 $\bar{A}$。

更一般地，如果 B 没有被当作全集，而有另一个全集 I 可以用来定义 A 的补集，则 A 在 B 中的补集记做 $\bar{A} \cap B$。

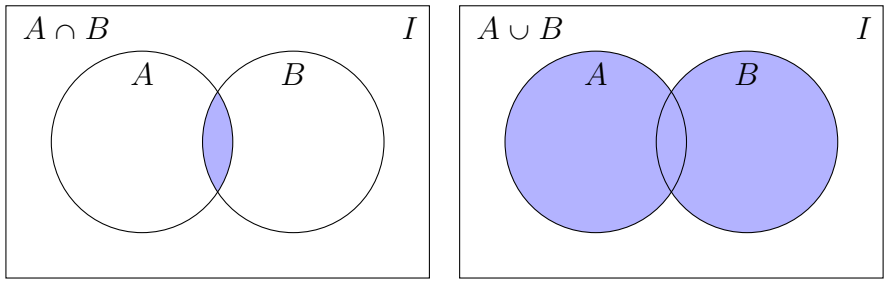
集合的交并补运算、集合的包含关系，如图 1.1 所示，都可以用下面的称为 Venn 图的示意图来表示。其中圆形（有的时候也用椭圆形）表示一个集合，全集通常用长方形来表示。两个圆形相交的共有的部分就是其交集，两个圆形合起来占据的部分就表示其并集。如果一个圆形被完全包含在另一个圆形之中就是集合包含关系。一个集合的外部，被包含在全集中的那些就代表了这个集合的补集。

纯粹当作为了更熟悉这些语言和符号，我们来用这些知识证明一个定理。

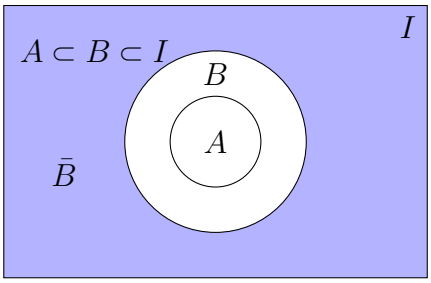
定理 1.1: 集合相等

当且仅当两个集合具有相互包含关系的时候，也就是 $A \subseteq B$ 并且 $B \subseteq A$ 的时候，集合 $A = B$。

两个集合相同 $A = B$ 的意思是集合这两个集合中的元素完全一一对应，也就是 $\forall x \in A, x \in B$，以及 $\forall x \in B, x \in A$。根据集合包含关系的定义，前者正好就是 $A \subseteq B$，而后者正好就是 $B \subseteq A$。因此，$A = B \Rightarrow (A \subseteq B \& B \subseteq A)$，这里我们引入了一个逻辑运算符号 & 表示这个符号前后的两个条件同时成立。



(a) (b)



(c)

图 1.1: **交集、并集、子集、全集和补集**。用 Venn 图展示了 (a) 交集 $(A \cap B)$, (b) 并集 $(A \cup B)$ 和 (c) 子集 $(A \subset B \subset I)$、全集 (I)、补集 $(\bar{B})$。

反过来，$A \subseteq B$ 就意味着 $\forall x \in A, x \in B, B \subseteq A$ 就意味着 $\forall x \in B, x \in A$，合起来正好就是 $A = B$ 的含义。因此，$(A \subseteq B \& B \subseteq A) \Rightarrow A = B$。

这就证明了当 $A \subseteq B \& B \subseteq A$ 的时候，有 $A = B$。也就是只需要有 $A \subseteq B \& B \subseteq A$ 的时候就有 $A = B$，这是“当”的意思，也就是前者是后者的充分条件。我们也证明了当 $A = B$ 的时候有 $A \subseteq B \& B \subseteq A$。也就是只要 $A = B$ 则必定有 $A \subseteq B \& B \subseteq A$，这就是“只有当，仅当”的意思，也就是后者是前者的必要条件。我们证明了“当且仅当”。以后我们进一步学习命题的时候会再一次来学习充分必要条件。

注意，本书的第一节主要是集合和映射的知识。这部分知识是相当抽象的。在这里，我们看到你以前所熟悉的数和加法运算竟然是可以这样通过非常抽象的方式来得到的，只要有集合这个概念就能得到。我们为什么第一节要选这么一个你以前就如此熟悉的学科概念知识但是采用如此抽象的讲法呢?

第一，从一个基本的对象来构建更加复杂的对象，例如这里的从空集构建自然数，并且补充上这个过程中需要用到的新的概念，证明所需要的命题，这是数学知识的系统性的表现，也是逻辑演绎证明在数学中的重要作用的体现。

第二，这样做可以帮助你体会到数学不是研究数和形的，是为了给描述任何对象——例如这里就是最没有结构的数学结构集合，也就是一堆东西，可以用于任何具体的东西——提供语言的，只不过数（和形，但是暂时这里仅仅展示了数）经常会是这些所要描述的对象的表示方式。

第三，通过这一节我希望你对数学的学科典型责任，数学的思维方式都有了一点点初步的体会。

当然，希望经过这一节之后你对于集合这个概念本身，数和加法本身，也有了一定的理解。

这一节和下一节走的是数学知识体系发展的两条路。在这里，除了一开始的集合的概念可以一定程度上认为是对现实世界的对象“一堆东西”的抽象之外，走的是先定义简单的结构，然后从简单的结构来构建复杂的结构的道路。例如，先定义了集合，再定义了自然数和映射，然后建立了集合的并的运算，再看到加法实际上对应的集合的并。

稍后就会看到，下一节走的是先从现实世界的对象中总结归纳出来模

式，对模式做抽象提炼，得到数学概念。例如，下面，我们会先从关于现实世界的自然语言提炼出来“数 + 量词 + 物”的模式，然后忽略掉“量词 + 物”而得到“数”中的“一、二”的概念，进而定义更大的数，进而定义加法和减法。

你可能对第二条路更习惯一些。但是，在数学发展上，这两条路是相辅相成的。我希望你从第一第二节就感受到这两条路。这也是本书和所有其他数学书的不同：用学科概念知识上最简单最基础的知识当素材，用最简单甚至都不需要进行的计算当素材，帮助你体会到高层知识，要么是学科大图景，要么是一般性人类思维，要么是教和习的方法。

理解和体会到高层知识确实是在智力上和思维习惯上是有挑战的，但是它不会无趣，也不会带来低层次的重复性活动所带来的负担。只要你跟上去了，你未来提出、思考和解决问题的境界就会不一样——会当凌绝顶，一览众山小。反过来，你一直在山脚下重复练习，则永远都不会太累，也永远都不会看透这个风景。

1.2 自然数的提出

和上一节从集合中得到自然数相比，从对现实世界的经验、体验和自然语言中提炼自然数的概念就非常直观了。在日常生活中，我们肯定已经遇到了“一头牛”、“一条鱼”、“一个人”、“一个西瓜”、“一个鸡蛋”、“两个苹果”、“三个鸡蛋”这样的说法。

把这些说法综合一下，我们发现，它们具有共同的结构：假装我们没有学过自然数，我们仍然可以知道，这个共同的结构是——“A+ 量词 + 物”，其中 A 就代表我们还不知道叫什么未来会被叫做自然数的那个东西。也就是这里是“一个玩意 + 一个量词 + 物（的种类）”这样的一个模式，这样的结构。而且，我们注意到，同一个的符号，例如“一”后面可以跟着不同的量词和不同的物，例如“一个人”、“一个西瓜”、“一个鸡蛋”。这三个表达方式之间只有物的差异。如果我们忽略这个差异，那么，我们只剩下“一个”。再联系到“一头牛”、“一条鱼”去掉物的部分剩下的“一头”、“一条”，也就是说，它们之间的不同仅仅是量词的不同。如果我们也去掉这个不同，就剩下“一”。

一旦我们得到“一”，我们就可以把别的量词和别的物加回来，来构成“一 + 量词 + 物”的模式。看到了模式，我们就可以迁移了。于是，我们也就很容易能够看懂甚至创造出来“一个词”、“一种模式”。

类似地，我们可以通过对日常生活中遇到的“二 + 量词 + 物”做抽象得到“二”，未来再结合量词和物来做看懂和创造新的表达方式。类似地，我们可以得到“三、四”等等。

这样地，先看到不同表达方式之间的差异和共性，去掉差异留下共性的过程就被称为抽象的过程，这个思维方式就被称为抽象。未来把已经抽象出来的东西加上那些差异，得到一个对更加具体的东西的描述，就叫做具体化。

因此，数“一、二、三、四”就是对关于现实世界的经验、体验、自然语言描述作抽象的结果。有了这几个数，我们就可以构建起来更加一般的数。

我们注意到，对于这些天生有一个量词单位——也就是有一个天生的单位——的对象，我们把一个新的对象和已经有的数量的对象合起来就能得到一个新的数量的对象。我们可以通过这个过程来得到这样的天生有一个量词单位的对象的所有下列的数量：

- 假设已知：“一”（+ 量词 + 物，后面的被抽象掉了）
- 加上一个单位的物之后得到“二”（+ 量词 + 物，后面的被抽象掉了）
- 再加上一个单位的物之后得到“三”（+ 量词 + 物，后面的被抽象掉了）
- 再加上一个单位的物之后得到“四”（+ 量词 + 物，后面的被抽象掉了）
- 再加上一个单位的物之后得到“五”（+ 量词 + 物，后面的被抽象掉了）
- .....

每两个相邻的数之间间隔都是一。然后，我们引入比汉字更加简单的符号，尤其是未来足够大——大于十和大于一百——之后可以更简单，相应为 $1, 2, 3, 4, 5, \cdots$。

实际上，这样我们不仅有了这些从 1 开始的自然数，还有了自然数之间的加上 1 的操作。这个时候，我们注意到，$1+2$ 可以看作是 $1+(1+1)$，因为 $2=1+1$ 显然成立。同样地，$1+3=1+(1+1+1)$。更一般地 $2+3=(1+1)+(1+1+1)$。也就是说任意的 $n+m$ 都可以通过还原到若干个 $1+1$ 来完成，

$$n+m=\left(\underbrace{1+\cdots+1}_{n}\right)+\left(\underbrace{1+\cdots+1}_{m}\right)=\underbrace{1+\cdots+1}_{n+m}. \tag{1.13}$$

我们把这个其操作过程称做加法：先把参与相加的两个数都还原成相应个数的 1 连着数的形式，然后把两个还原之后的相应个数的 1 再合起来连着数。

有了这个加法操作，我们就可以来回答把两堆东西合起来数一数有多少个的问题了。用集合的语言，一堆东西就是一个集合。因此，加法这个操作就可以用来回答两个集合里面的元素合起来数一数的问题了。这就是加法的含义。这样我们就从加法的操作走到了加法的含义。

注意，原则上，加法操作的计算过程总是要回到多少个 1 的问题，然后再合起来连着数一数的。不过，在实际计算中，你很容易可以记注个位数的加法口诀，然后基于十进制来把多位数的加法计算问题变成个位数的加法计算问题。不过，计算永远不是本书的重点。我们交代清楚原理就不管计算的问题了。我们看到，从原理来说，反而通过数手指头计算加法，也就是每次用连着数一数，是最合理的计算方式。

有了一般的加法我们可以问反过来的问题，3 是多少加上 1 得到的，n 是多少加上 m 得到的。这个时候，我们实际上就是在处理

$$?+m=n, \tag{1.14}$$

也就是

$$\left(\underbrace{1+\cdots+1}_{?}\right)+\left(\underbrace{1+\cdots+1}_{m}\right)=\left(\underbrace{1+\cdots+1}_{n}\right). \tag{1.15}$$

这个时候，我们只需要从 m（不包含从到 m 开始数的那一次）数到 n（包含数到 n 的那一次）看看要数多少次，就知道 ? 是多少了，或者反过来从 n（不包含从到 n 开始数的那一次）倒着数到 m（包含数到 m 的那一次）

看看要数多少次，也行。我们给这个问题和过程一个名称，称做减法，记做 $?=n-m$。

这个操作回到对象的语言上，那就是从整体 n 个对象中去掉其中的 m 个剩下多少个。这也就是得到了减法的含义——整体里面去掉一部分（之后剩下的数量）。

通过对比公式 (1.13) 和公式 (1.15) 的形状，我们发现，除了 ? 的位置不同——在加法公式 (1.13) 中待计算出来的 ? 出现在右侧，在减法公式 (1.15) 中待计算出来的 ? 出现在左侧，除此之外两者完全相同啊。于是，我们看到减法和加法是密切联系的：减法是加法的逆运算，逆操作。减法 $?=n-m$ 相当于是问 m 加上 ? 会成为 n。

一旦有了减法，很自然地我们就可以倒过来走，问

$$?+1=1 \Leftrightarrow 1-1=? \tag{1.16}$$

我们把这个数记为 0。其含义是加上 1 等于 1 的东西，或者说从只有 1 个对象的集合中去掉一个对象剩下的数量，于是，正好就是空集的数量。借用上一节的记号，就是其中包含 0 个对象。

那更进一步，

$$?+1=0 \Leftrightarrow 0-1=? \tag{1.17}$$

那个比零还少 1 的数存在吗？有什么含义？我们发现如果这个数存在，那么这个数的含义就是加上 1 以后等于零，等于不包含任何对象的集合。也就是说，这个对象就好像一个专门留给 1 的空洞——需要一个相应的对象来填补。比如说，欠一块钱，就有这个意思：如果得到了一块钱，补上这个洞，则刚好填平等于零块钱。我们把这个数记做 -1，表示一个空洞的意思，需要一个对象来填平的意思，也就是 $-1+1=0$。

或者直接用减法的含义，-1 就是比 0 小 1 的数。这样，我们就可以定义 $-n$ 为比 0 小 n 的数，需要 n 个对象来填平的意思 $-n=0-n$，$-n+n=0$。

这样我们就从对现实世界的经验和体验、自然语言得到了所有的自然数 $0,1,2,3,\cdots$，还有负整数 $-1,-2,-3,\cdots$。自然数包含 0 和正整数 $1,2,3,\cdots$。合起来，0、正整数和负整数称为整数。

并且，我们看到了，这样得到的自然数和从集合得到的自然数以及相应的加法完全一致。

从中，我们更加看到了抽象和具体化：从现实世界的经验和体验、自然语言中看到模式“数 + 量词 + 物”，去掉每次都不同的“物”，再去掉每次都不同的“量词”，剩下的东西就是“数”，未来加上新的具体的“量词”和“物”之后得到被抽象之前的模式，可以用来描述世界。去掉的过程是做抽象，加上的过程是具体化。

数学是研究模式的。做抽象和具体化是数学做这样的研究的典型思维方式之一。

纯粹依靠人类的想象力和推理，尽管想象力的部分可能已经受到了现实世界的启发，和通过对现实世界做抽象，做模式的提炼，是数学发展的两大动力。未来我们还会看到数学发展的第三个动力——数学知识本身的系统性和严格化。等我们遇到这样的例子再来展示这个第三个动力。

通过这一节我希望你对数学的学科典型责任，数学的思维方式都有了更加深刻的体会。对自然数、整数和加减法也有了比较深刻的理解。

习题 1.3 (做抽象的其他例子). 尝试描述一下一个其他的数学概念的从具体对象到数学概念的抽象过程。如果你更加熟悉其他学科，也可以是一个其他学科的概念的抽象过程。最好说说为什么做这个抽象。

1.3 等式的性质

其实，我们在前面已经悄悄地引入了等号，例如公式 (1.13) 和公式 (1.15) 中。在公式 (1.13) 交代了两个数加起来的含义和计算方式。在那里等号的意思是等号左右两侧的东西是（某种意义下，在这里是数的意义下）相同的东西，可以（在相应的意义下，在这里是数的意义下）相互替换。一般来说，习惯上，等号不会用在形式上也完全相同的两个东西之间。当然，那时候仍然可以使用等号把两者连接起来，但是，就相当于一句没有信息量的话，没有没有思考，没有意义，例如 $2 = 2, 1 + 2 = 1 + 2$。

由于在下面很长一段时间内，“某种意义下”都会代表是“在数的意义下”，我们忽略对每个等式的具体这样的意义的阐述，而把等号的含义简单说成“等号左右两侧的对象形式不同但是含义相同”，简称“等号两侧相等”。

有了这个等号的含义，等式就有了一些性质。例如，等式两边同时加上

一个数等式仍然成立，也就是

$$a = b \Rightarrow a + c = b + c. \tag{1.18}$$

这里的 a, b, c 可以是任何一个我们已经学习过的整数。等式之所以成立是因为 a, b 是同一个数，因此，$a + c$ 和 $b + c$ 也是同一个数。同样，对于减法这个等式的性质也成立，

$$a = b \Rightarrow a - c = b - c. \tag{1.19}$$

这里的符号 “⇒” 是 “推理出、得到、导致” 的意思。具体的推理过程就是下面那一句话的说明：“等式之所以成立是因为 a, b 是同一个数，因此，$a + c$ 和 $b + c$ 也是同一个数”。等这句话成为已经被证明的结论的时候，读者和作者都接受的时候，就可以被省略，以后直接用 ⇒ 替代。有的时候，我们也这样来写这个推理过程，

$$a = b \Rightarrow a + c = b + c(\text{同一个数加上同一个数}). \tag{1.20}$$

其中为了公示排版问题，我们把前面那句话 “等式之所以成立是因为 a, b 是同一个数，因此，$a + c$ 和 $b + c$ 也是同一个数” 简写为 “同一个数加上同一个数”，或者也可以写成 “等号和加法的含义”，

$$a = b \Rightarrow a + c = b + c(\text{等号和加法的含义}). \tag{1.21}$$

这样的明确地把每一步推理过程表达出来的方式，用推理符号建立两个等式（未来会拓展到两个命题）之间的关系在后面把这样推理的理由写在括号里，非常有助于作深入和正确的思考。这是数学学习的一个很大的好处——形成可以完成更长链条的正确的思考的习惯、技能和意愿。我们把完成某项任务的技能、习惯和意愿合起来称作完成这项任务的能力。这里就是长链条思考的能力。我们先从简单的一步推理开始就这样做，慢慢地来形成长链条思考的能力。

顺便，在学科概念知识上，这里得到的等式的性质—— “等式两边加上同一个数等式仍然成立” ——是未来等式的更加复杂的性质的基础。你也得掌握这个具体的学科概念知识，尽管这一节的主要目的是帮助你掌握推

理的表达方式，于是未来形成长链条思考的能力。这是一个超越数学学科的一般性人类思维层面的知识。具体来说是逻辑演绎证明，以及把逻辑演绎证明用于批判性思维。我们后面会更多地遇到这个高层知识，等积累了更多的例子了我们再来专门讨论逻辑演绎证明和批判性思维。

1.4 记号、序数和基数

我们已经学习了所有的整数。这一节我们来看看在整数的日常使用中，我们都把它们用来做哪些事情。首先，我们会用整数来给事物做记号。例如，这一节是本书的 1.4 节。这个编号的含义是“本章是本书的编号为 1 的章，本节是本书的编号为 4 的节”。例如，我们可以给一堆苹果中的每一个一个编号，给一个班里面的每一个人一个编号，给每一道习题一个编号。给了编号以后，我们就可以用编号来指代这个对象。注意，编号之间没有其他的关系。例如，一个编号为 3 的苹果和一个编号为 5 的苹果之间没有关系。甚至，原则上，本书的页码和章节号码也可以不遵循先后出现的顺序，只需要每一页都有一个页码，每一个章节都有一个编号，我们就可以实现找到相应的页和章节的功能了。

当然，如果页码和章节还是按照顺序编号的，那么，我们可能就能更顺利地找到相应的页和章节。但是，这里有一个额外的要求，这些被顺序编号的对象之间要具有没有用来编制顺序的属性。例如，在页面和章节这两个对象上，有一个其自身具有的这样的属性——页面和章节在本书中内在的出现顺序。因此，一个可以做顺序编号的对象的类别，意味着这个类别的对象的具体成员，具体对象，之间具有某种顺序。在数学上，人们把这个叫做集合上的序映射。我们就不去严格定义序映射和有序集合了。我们只需要明白给对象做有顺序的编号其实需要对象具有某个可以用来比顺序的属性。

比如说，本章的的习题，是顺序编号还是记号编号？大概来说，出现在章节内部的不同章节的习题之间是有一个自然的内在的顺序的，就算在同一章节内，我也尽量按照难度来先出更简单的习题。因此，一定程度上，习题的编号是一个顺序编号，而不仅仅是记号编号。但是，有的时候，同一章节内的两道题难度差不多，尤其是章节后面的各种知识点的习题放在一起的时候，就不一定能找到某种内在的顺序。这个时候，就只能承担记号编号

的作用了。

在比如说，本书的定义的编号，是顺序编号还是记号编号？大概整体上是顺序编号，前面的定义是后面的定义在知识结构上的基础。例如我们只有先定义集合再来定义卡氏积，然后定义映射。但是，并集和交集这两个概念尽管定义的时候并集在前，就没有内在顺序了，完全可以选择交集定义在前。

习题 1.4 (扑克牌是顺序编号吗？). 你拿到了一盒子扑克牌。每一张牌上都有一个数的编号（忽略花色，把 J, Q, K 看作是 $11, 12, 13$，忽略大小王）。什么时候这一盒子扑克牌是记号编号，什么时候是顺序编号？顺序编号的“顺序”是内在的吗，只有一种顺序吗？

更多的时候，其实我们是把数拿来做加减法的计算。这个时候我们说数起到的是基数的作用，其要求是：任何两个相邻的数之间的差是相等的。当我们问“本书总共有多少道习题”的时候，我们很自然地把每一章的习题的数量加起来，每一章的习题又来自于每一节的习题加起来再加上每一章最后部分的习题的数量。问题是，这样的加法给出来的答案真的有意义吗？

比如说，今天老师布置了 5 道习题当家庭作业，昨天老师布置了 8 道习题当作家庭作业。这个信息意味着什么？昨天的作业量更大，你需要更多的时间，而今天估计用的时间会更少？平均意义上也许有一点点时间和习题数量之间的相关性，但是具体到这个例子，完全有可能没有丝毫关系。比如说，今天的题尽管更少，但是有一道题非常难，或者计算非常繁琐呢？那么，这里不能加起来算习题的总数也不能把两个总数相减的根本原因是什么呢？是因为每道题的难度和相应的要求的时间不是一样长的。也就是破坏了“任何两个相邻的数之间的差是相等的”。

同样地，说一本书具有 15 章，另一本书具有 10 章，也不能做加减计算。甚至在本书内部，说阅读其中的 3 章和阅读其中的 5 章，除非这 5 章就包含了那 3 章，也不能做加法和减法计算。

那我们怎么天天做数的加减法计算呢？因为有的时候我们可以忽略对象之间的差别，当作大概满足“任何两个相邻的数之间的差是相等的”。比如说，我们经常计算篮子里面的苹果的加减法，盒子里面的鸡蛋的加减法。这个时候，我们其实假设了苹果之间的差异很小，至少这个差异远远小于一个苹果。所以，这时候我们才能说，5 个苹果的比 4 个苹果多一个，1 个

苹果和 1 个苹果一样多，合起来就有了 2 个苹果，和另外那个拿着 2 个苹果的人拥有的苹果一样多。但是，如果这一堆苹果之间的差异很大，超过了一个苹果，也就是有的时候两个苹果合起来也没有其中的一个苹果大，那这个加减计算就出问题了。拿到那两个小苹果的人就会说，“我的两个和你的两个根本不是一回事，不公平”。

那我们来思考一下：本书的页码之间是可以计算加减法的吗？看你的目的。如果你是在估算印刷成本或者翻书多少次，可以计算页码加减法。但是，如果你是在估算阅读时间，你不能计算加减法。对于前面的目的，大概来说，忽略插图彩印的问题（假设本书没有插图）每一页的印刷成本类似，从某一页到某一页的翻书的次数呢确实和书的内容没有联系。因此，基本上满足相邻数——这里是页码——之间等间隔的要求。对于后面的目的来说，每一页的内容是最主要的决定因素，因此，很有可能某一页的内容比另一页抽象很多，需要更多的思考，所以也就需要更多的阅读时间。因此，基本上不满足相邻数——这里是页码——之间等间隔的要求。

更进一步，我们还需要看到，用什么数学结构来描述一个对象，其实是由这个对象本身的属性还有你通过这个描述所要解决的问题你的目标来决定的。回到数当作记号、序数和基数的作用这个问题，记号对对象几乎就没有要求，只需要对象可以叫做“一个”就行（也就不同对象之间有明显的自然的区隔）；序数要求对象之间有某种内在的序，可以按照这个序来排列；基数要求可以按照一种序来排列，并且排列出来之后相邻的两个数之间的间隔，对应到所描述的对象属性上去之后，是相同的。

每一个对象加上你要实现的相应的目标，合起来决定了这个对象适合用什么数学结构来描述。这就是数学建模。不是任何一个数学结构都可以用于描述任何一个现实世界的对象的，也不是说把现实世界的对象用数学结构来描述以后，关于这个数学结构能够做的一切的操作，都可以在这个数学结构和相应的现实世界的对象上进行的。

1.5 模式和抽象

1.6 自然数的公理化定义

1.7 推荐阅读材料

1.8 作业

习题 1.5 (生活中的集合). 从生活中举出来几个集合的例子，分别用列举全部对象和用特征概括的方式来表述这些集合，然后证明它们确实是集合。

习题 1.6 (生活中的记号、序数和基数). 从生活中举出来几个把整数用来当作记号、序数和基数的例子。

1.9 本章小结

第二章　从四则运算的提出看模式和抽象

2.1　四则运算的提出和相互关系

2.2　模式和抽象

2.3　数的进制

$(ab)_M = a \times M + b$

2.4　数的扩张和封闭性

负数其实已经运用了封闭性。之所以要求封闭性是因为表示论。虚数的拓展也可以看作表示论和封闭性吗?

2.5　计算是标准化了的思考过程

鸡兔同笼、解析几何用于证明平面几何定理。如果这部分例子看不懂,可以在学习了方程、平面几何和解析几何之后再会来看

2.6 运算律

为什么需要，为什么成立

2.7 从计算和计算的封闭性来定义数

2.8 推荐阅读材料

2.9 作业

2.10 本章小结

第三章　从整除看逻辑演绎证明

3.1　自然数的整除

3.2　几个数的整除性质及其证明

3.3　什么是逻辑演绎证明——初步

3.4　推荐阅读材料

3.5　作业

3.6　本章小结

第四章　从用符号表示对象看数学是语言

4.1　对象的层次

数值、量、物、表达式、对表达式的操作

对象的某属性的值，对象的某属性，对象，对对象的操作，对对象的操作的操作

4.2　模式和抽象

4.3　表示论初步

用数学表示现实世界的对象、关系、操作、对操作的操作。用一种数学形式（例如数或者一群数）表示另一种数学形式（例如操作、关系、操作的操作）。

4.4　推荐阅读材料

4.5　作业

4.6　本章小结

第五章　变量和常数——从数到量到代数式

5.1　完全已知的数

5.2　给定的（当作已知）的数

5.3　未知的或者重点关注的数

5.4　多项式的元、阶和解

幂、多项式、方程求解

5.5　推荐阅读材料

5.6　作业

5.7　本章小结

第六章　映射和函数

6.1　用于刻画集合内部结构的映射

6.2　表示是一种映射

6.3　函数

变量之间的关系
函数以及几个典型具体函数

6.4　位置（速度）-时间函数

6.5　推荐阅读材料

6.6　作业

6.7　本章小结

第七章　从表示论看数学是语言

从符号到数学内涵——数学结构，表示论

7.1　硬币翻转的数学描述

对象，对对象的操作

7.2　正三角形转动的数学描述

7.3　集合和群——对象和操作

对对象的操作，操作的组合

7.4　操作的矩阵表式

7.5　对称性和相似变换

对操作的操作

7.6 经典对象上的博弈的数学描述

7.7 当作对象属性或者过程的数

7.8 数当作加法和加法的表示

7.9 数学是对操作的研究对关系的研究

任何一个操作的集合，再保持某个研究对象的某个特征不变，例如一个皮球还是皮球没有被剪开，的前提下所有的可能的操作的集合。研究这个操作集合的元素之间的关系，从而从已知的具体状态经过操作得到新的状态，或者得到所有新状态的可能，或者研究选择什么样的操作来调控这个对象会满足对这个对象的状态的某个特定要求。这样的操作构成群。

因此，从这个角度可以认为，所有的数学，建立在群论的基础上。不过就是某些群表现出来的结构适合用整数、分数、实数、矩阵等来表示。

7.10 推荐阅读材料

7.11 作业

7.12 本章小结

数学研究对象从此完全不受限制，数、代数、几何图形以及它们的结合，仅仅是这些对象的表示以及这些表示结合起来，背后是对象。什么样的对象决定了描述这个对象需要什么数学结构。

第八章　存在性、唯一性以及求解

8.1　勾股定理和三角函数

8.2　等式的性质以及数的唯一性

8.3　三角形的稳定性和三边确定的三角形的唯一性

8.4　已知两边一角的三角形的存在性和唯一性

8.5　方程的解的存在性、唯一性和求解

8.6　推荐阅读材料

8.7　作业

8.8　本章小结

第九章　从存在性、唯一性和计算分开看分解和综合

9.1　方程的解的存在性、唯一性和计算

9.2　分解和综合

9.3　不动点的存在性、唯一性和计算

9.4　推荐阅读材料

9.5　作业

9.6　本章小结

第十章　从平面几何学看知识体系和逻辑演绎证明

10.1　《几何原本》的知识体系和逻辑演绎证明

以距离公式为公理的平面几何知识体系。

10.2　现代数学的知识体系和逻辑演绎证明

10.3　现代物理学的知识体系和逻辑演绎证明加实验检验

10.4　从三角形到多边形看分解和综合

10.5　推荐阅读材料

10.6　作业

10.7　本章小结

第十一章　逻辑学和数学的关系

从逻辑演绎证明到逻辑学、逻辑的数学基础（集合）、数学的逻辑基础（所有论证过程成立的道理）、数学知识体系的问题（集合没有定义）

11.1　概念

集合

11.2　命题

集合关系

11.3　逻辑演绎证明

三段论（集合关系）

11.4　逻辑学的学科责任

把判断一个命题是否成立分成判断这个命题的基础是否成立和判断从基础开始来论证命题成立的过程分开。或者说，只要保证论证过程正确，则未来只需要判断其基础命题是否成立。

在科学上，往往还会去额外检验得到的命题是否成立，检验的方法往往是实验检验，也就是和现实世界对比。沿着逻辑链条回溯，当前命题的

基础命题是否成立的检验也通过：第一，检验其基础命题是否成立；第二、检验其论证过程是否成立；第三，如果其和现实世界中的对象相关，则实验检验其是否成立。

这也是分解和综合。

11.5 数学归纳法

11.6 不完全归纳法

条件概率和不完全归纳法

11.7 集合本身没有定义，怎么办？

理发师悖论、范畴论

11.8 推荐阅读材料

11.9 作业

11.10 本章小结

逻辑帮助你判断论证的对错，做更严格的论证，但是逻辑不启发论证的思路，不能代替实验检验。后两者，靠对知识结构的深刻的认知，尤其是长程联系的建立，靠直觉、想象力、现实的启发。

第十二章　从自然数的严格定义看数学的知识体系和逻辑演绎证明

12.1　自然数的严格定义

皮亚诺公理化定义

12.2　自然数的集合定义

12.3　为什么有了自然数的“自然”定义还需要严格定义

数是描述世界的典型结构。严格来说，每个数学概念都需要有严格的定义。可以用来构建其他定义，用来证明定理，用来严格地构建知识体系。

12.4　推荐阅读材料

12.5　作业

12.6　本章小结

第十三章　空间点和空间矢量的数学模型

13.1　空间的点和数——笛卡尔坐标

直角坐标系

13.2　空间矢量

一般的矢量应该是什么样子的?

13.3　解析几何

13.4　来自物理学的启发和数学建模

13.5　推荐阅读材料

13.6　作业

13.7　本章小结

第十四章　概率的数学结构

14.1　简单状态集合和每个简单状态的概率

基于等概率简单事件的古典概型

样本空间，简单事件集合

14.2　概率语言下的状态

集合的子集构成的集合，σ 代数

14.3　概率映射

从 σ 代数到 $[0,1]$ 的映射

14.4 可列可加性

14.5 从概率的数学结构看到双向数学建模

14.6 从概率的数学结构看到代数是对关系的刻画

14.7 概率分布函数

连续变量和离散变量

14.8 推荐阅读材料

14.9 作业

14.10 本章小结

第十五章　数学解题四问

15.1　数学应用题的 WHWM

从实际问题到数学问题，到表达式

15.2　几何题的 WHWM

对象、关系、概念、总结反思

15.3　WHWM 背后的系统思维

15.4　一般性 WHWM

15.5　推荐阅读材料

15.6　作业

15.7　本章和本卷小结

第二部分

高等数学篇

第十六章　微积分

Calculus works by making visible the infinitesimally small.
微积分通过使得无穷小可见而发挥作用。

– Keith Devlin

Most mathematics deals with static objects such as circles and triangles and numbers. But the great universe "out there," not made by us, is in a constant state of what Newton called flux. At every microsecond it changes magically into something different. Calculus is the mathematics of change.
大多数数学（结构）是关于圆、三角形和数等这样的静态的对象的。但是，我们伟大的不是我们造出来的就那在里的宇宙，处在一个被 *Newton*（牛顿）叫做“流数”的状态下：任何一个微秒，它都可以从一个状态变成另一个状态。微积分就是关于改变的数学。

– Martin Gardner

Calculus is the most powerful weapon of thought yet devised by the wit of man
微积分是人类智慧所创造的威力最大的思想武器。

– Wallace B. Smith

Calculus destroys self-esteem on contact.
微积分一接触就伤自尊。

– James Lileks

我们先从导数和积分的概念的历史发展过程中都到了哪些现实世界的和数学知识体系内部的启发，然后再来看这背后的双向数学建模的思想。这

样，等我们正式非常形式化地学习微积分的时候，可以形成对微积分的更好的理解。

16.1 导数

我们先来学习导数，再来学习积分，以及中间的微分。对于每一个概念的学习，我们都从现实世界或者数学本身的启发开始。得到启发以后提出的概念和命题的严格化，我们稍后再来完成。

16.1.1 导数受到的启发

对于微积分里面叫做导数的那个东西其实，我们是有现实世界的体验的。不知道什么原因，也许就是我们能够改变它操控它，我们是认为一个在行使的车每时每刻都有一个速度的，而且还可以通过油门和刹车来改变这个速度的大小，通过转方向盘于是转动轮胎来改变其方向。甚至，我们也可以想象，同一辆车如果开的很快撞到另一量静止不动的车或者一个人，其事故的严重程度要远远比开得很慢要高。

这么一个朴素的观察和体验，我们能够想明白吗？Zeno（芝诺）也思考了这个问题，他发现这个问题他想不明白，而且他认为这个想不明白不仅仅是他自己的原因。我来复述一下Zeno的思考。考虑一个飞着的箭。一方面，既然是飞着的就表示是其速度不为零。但是，我们来考虑一个时刻，一个时间点，而不是一个运动过程。那么，在这个时刻，箭还没有形成运动，也就是 $\Delta x = 0$。按照速度 $v = \frac{\Delta x}{\Delta t}$，当 $\Delta x = 0$ 的时候，“自然地” $v = 0$。因此，“飞矢不动”。

我们当然知道这个 $v = 0$ 的结论是错的。因此，Zeno “阿基利斯追不上乌龟” 佯谬实际上质疑的不是这个结论，而是我们如何才能从一个 $\frac{0}{\Delta t}$ 中得到一个不等于零的值。显然，如果 $\Delta t \neq 0$ 则 $\frac{0}{\Delta t} = 0$。幸好，这里确实 $\Delta t = 0$。那问题就成了，

$$\frac{0}{0} \tag{16.1}$$

怎么算出来一个不等于零的数。

从物理上说，我们永远可以通过先定一个时间间隔 t 到 $t + \Delta t$，记作

$[t, \Delta t]$，来测量其前后两个时间点的位置，从而得到这段时间间隔内的这个物体运动的平均速度，

$$\bar{v}_{t,\Delta t} = \frac{x\left(t+\Delta t\right) - x\left(t\right)}{\Delta t}. \tag{16.2}$$

并且让这个时间间隔的长度 Δt 不断地变小，都得到一个平均速度。问题是，物理上无论多么小，总是需要一个有线时间间隔来实际做这个测量的（都不考虑测量位置本身所需要的时间的问题）。因此，物理上没法解决真的当 $\Delta t = 0$ 的时候的情况。好了，我们可以做到 Δt 越来越小，但是不能做到 $\Delta t = 0$。

那反过来，数学上呢?

我们来看几个例子。从最简单的匀速运动开始。我们知道，匀速运动有 $x = x_0 + v_* t$。我们来完成上面的物理上的计算，

$$\begin{aligned}
\bar{v}_{t,\Delta t} = & \frac{x\left(t+\Delta t\right) - x\left(t\right)}{\Delta t}, \\
= & \frac{\left[x_0 + v_*\left(t+\Delta t\right)\right] - \left[x_0 + v_* t\right]}{\Delta t}, \\
= & \frac{v_* \Delta t}{\Delta t}, \\
= & v_*.
\end{aligned} \tag{16.3}$$

我们竟然发现，上下自动消去了 Δt。也就是说，对于我们特殊选择的匀速运动 $\left(x = x_0 + v_* t\right)$，实际上，我们可以发现，无论时间间隔取多少，$\bar{v} = v_*$ 一直都成立。那么，很自然，我们就可以说，对于我们不会算的 $\Delta t = 0$ 的情况，因为在之前其越来越接近于零的时候都有 $\bar{v} = v_*$，那就算在那个时候，也有

$$v_t = \bar{v}_{t,\Delta t=0} = v_*. \tag{16.4}$$

如果都是这样那就好了。那我们来试试自由落体运动 $x = x_0 + v_0 t + \frac{1}{2} g t^2$，

$$\begin{aligned}
\bar{v}_{t,\Delta t} = & \frac{x\left(t+\Delta t\right) - x\left(t\right)}{\Delta t}, \\
= & \frac{\left[x_0 + v_0\left(t+\Delta t\right) + \frac{1}{2} g\left(t+\Delta t\right)^2\right] - \left[x_0 + v_0 t + \frac{1}{2} t^2\right]}{\Delta t}, \\
= & \frac{v_0 \Delta t + g t \Delta t + \frac{1}{2} \Delta^2 t}{\Delta t}, \\
= & v_0 + g t + \frac{1}{2} \Delta t.
\end{aligned} \tag{16.5}$$

其中最后一项，会随着我们说的物理上测量的时间间隔 Δt 越来越小而可以和零无限接近。那是，是不是说，我们可以让这里的

$$\Delta t \to 0 \tag{16.6}$$

呢？如果可以，则对于前面的自由落体运动，我们也有了

$$v_t = \bar{v}_{t,\Delta t=0} = v_0 + gt. \tag{16.7}$$

而且这个公式给出的结果竟然也能通过物理实验的检验：只要让测量的时间间隔的 Δt 足够小，测出来的 t 时刻的速度就是可以和 $v_0 + gt$ 在误差范围内相符。也就是说，Zeno “阿基利斯追不上乌龟” 佯谬所关注的 $\frac{0}{0}$ 的问题，也就是瞬时速度的问题，可以这样算 [4, 5]：

1. 先把 Δt 当作正常代数量，计算分子 $x(t+\Delta t) - x(t)$，得到一个关于 Δt 的表示为其幂次函数（$\Delta^m t$）的加和形式的代数表达式，这时候 t 等其他量都看作是常数，

2. 然后，把分母 Δt 算到分子中每一项内部去，成为 $\Delta^{m-1} t$ 项的加和，

3. 对于，每一项，用 $\Delta^0 t = 1$，$\Delta^q t = 0$，$\Delta^{(-q)} t = \frac{1}{\Delta^q t} = \frac{1}{0} = \infty$，其中 $q > 0$，得到最终结果。

这个过程记作 $\frac{dx(t)}{dt}$。这些符号来自于Gottfried Leibniz（戈特弗里德 · 莱布尼茨）[6, 7]。未来我们会说明为什么这样的记号是合适的。其中 $\frac{1}{0}$ 的含义我们以后再进一步解释。暂时，可以看到，任何一个有限大小的数 a 乘上 0 都等于零不可能等于 1，也就是 $a \times 0 = 0 < 1$。因此，$\frac{1}{0}$ 必须比任何有限大小的数还要大，记作 ∞，读作 “无穷大”。至于这个比任何数都要大的数到底是什么，以后再说。

为什么可以这样算以后再说。不过，只要能够这样算，至少，我们可以从 $\frac{0}{0}$ 中得到一个数，这个数可能是有限大小，也可以是正负无穷大。这样瞬时速度的定义和Zeno “阿基利斯追不上乌龟” 佯谬就都解决了。

实际上，关于导数，也就是

$$\frac{dx(t)}{dt} \tag{16.8}$$

Isaac Newton（伊萨克 · 牛顿）和Leibniz也就创造性地提出了这么多 [8]。但是，正如George Berkeley（乔治 · 贝克莱）指出 [8]，这里是有逻辑问题的：一会儿 Δx 被当作了一个不是零的一般量（只要当作零了那么分子分母就都是零 $\frac{0}{0}$，于是我们就不知道怎么算了），一会儿又被当作一个零，而且是分子分开成每一项并且除掉分母之后再当作零之前当作不是零的一般量，凭什么？

因此，当年微积分的发展是受运动学研究的启发，以及曲线求切线、求曲线下面积等启发——后面这个稍后再说，先得到计算导数的方法，然后看到导数计算的逻辑问题，进而发现需要一个叫做 “极限” 的概念来补充成一个逻辑上严密的知识体系的。稍后我们来看极限这个概念。

现在，让我们再一次回到导数和积分。前面我们提到了，对每个时刻的位置 $x(t)$ 算导数可以得到速度 $v(t)$，也就是位置的变化率。那么，对速度算导数就可以得到速度的变化率。对于一个一般的函数 $f(x)$ 呢，导数 $\frac{df(x)}{dx}$ 有什么含义？

我们还是从 x 点的一段间隔 Δx 内的平均变化率——记为 $\bar{f}'(x)_{\Delta x}$——开始，

$$\bar{f}'(x)_{\Delta x} = \frac{f(x+\Delta x) - f(x)}{\Delta x}. \tag{16.9}$$

我们发现，$\frac{f(x+\Delta x)-f(x)}{\Delta x}$ 是曲线 f 在两个点的取值的差除以两个点的间距，也就是 $P = (x, f(x))$ 和 $Q = (x+\Delta x, f(x+\Delta x))$ 两个点的连线的斜率，如图 16.1 所示。

于是，很自然地，当 Δx 越来越小的时候，P 和 Q 两个点的连线就成了过 P 点的切线——和原来的曲线 $f(x)$ 在曲线上某个点附近有且只有一个交点的直线。于是，如果我们接受曲线的切线存在——具体什么条件下存在以后再说，那么，Δx 越来越小的时候，$\bar{f}'(x)_{\Delta x}$ 就成了曲线 $f(x)$ 在 $P = (x, f(x))$ 的切线的斜率。这个时候，就像我们用 $\bar{v}_{t,\Delta t}$ 在 Δt 越来越小的时候记作 $\bar{v}_{t,\Delta t=0} = v_t$ 一样，我们把 $\bar{f}'(x)_{\Delta x}$ 也就写成 $\bar{f}'(x)_{\Delta x=0} = f'(x)$。再类比 $v(t) = \frac{dx(t)}{dt}$ 的记号，我们记

$$f'(x) = \frac{df(x)}{dx}. \tag{16.10}$$

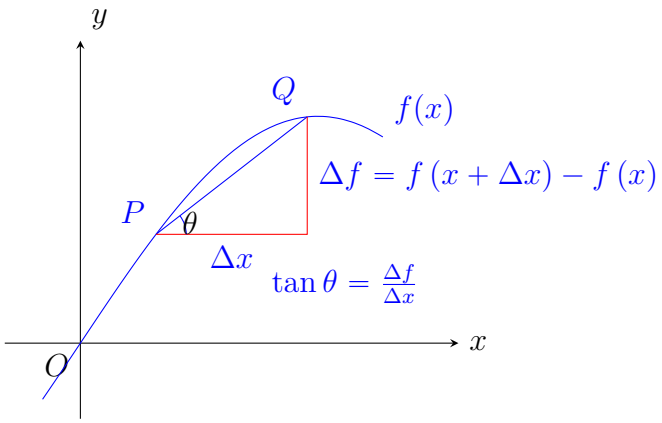


图 16.1: **函数的平均变化率和连线斜率**。函数 $f(x)$ 在某两个 x 点之间的平均变化率对应着这两个点对应的函数上的点连线的斜率。

有的时候，为了更明确地表示导数是对那个变量做的，我们也把上式写成，

$$f'(x)_x = \frac{df(x)}{dx}. \tag{16.11}$$

或者更偷懒地记作

$$f'_x = \frac{df(x)}{dx}. \tag{16.12}$$

注意很多曲线有切线。例如，过圆上一点 $P = (x_P, y_P)$ 的切线是和圆心到 P 的的半径垂直的那条直线。自然，斜率是存在的，也是用几何方法等其他非微积分的方法算得出来的。

例 16.1 (过圆上一点的切线). 用几何方法求出来过圆上一点 $P = (x_P, y_P)$ 的切线的斜率。

如图 16.2 所示，根据圆的含义（到圆心距离相等的点的集合，为简单计，这里我们取圆心为 $(x_0, y_0) = (0, 0)$）和方程，

$$(x - x_0)^2 + (y - y_0)^2 = r^2, \tag{16.13}$$

有 $y_P = \pm\sqrt{r^2 - x_P^2}$。为了成为函数——一个 x 值只能得到一个 y 值，我们取正号的分支：$y_P = \sqrt{r^2 - x_P^2}$。

于是，圆心到 P 的直线的方程为 $y = \frac{y_P}{x_P}x$——它同时经过圆心 $(0, 0)$ 和 P 点。这可以通过代入圆心 $(0, 0)$ 和 P 点的坐标值进入直线方程 $y = \frac{y_P}{x_P}x$

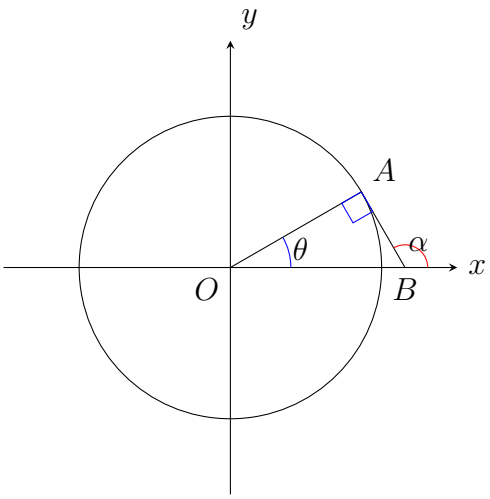


图 16.2: **圆上切线的斜率**。$k_2 = \tan\alpha = \tan 90^\circ + \theta = -\cot\theta = -\frac{1}{k_1}$。

验证这两个点的坐标值确实满足这个直线方程。用已经得到的符号来把直线方程表示为，

$$y = \frac{\sqrt{r^2 - x_P^2}}{x_P} x. \tag{16.14}$$

其斜率为 $k_1 = \frac{\sqrt{r^2-x_P^2}}{x_P}$。

再根据相互垂直的两条直线的斜率满足 $k_1 k_2 = -1$，得到过圆上一点 P 的切线的斜率为 $k_2 = -\frac{x_P}{\sqrt{r^2-x_P^2}}$。

那么，我们是不是真的可以得到同样的结果呢?

例 16.2 (过圆上一点的切线). 用微积分方法求出来过圆上一点 $P = (x_P, y_P)$ 的切线的斜率。

我们把函数 $y = f(x) = \sqrt{r^2 - x^2}$ 代入到公式 (16.9) 和公式 (16.10)

有，过这条曲线上的 $P=\left(x_P,\sqrt{r^2-x_P^2}\right)$ 的导数值为

$$\frac{df\left(x_P\right)}{dx} = \frac{\sqrt{r^2-\left(x_P+\Delta x\right)^2}-\sqrt{r^2-x_P^2}}{\Delta x}, \tag{16.15}$$

$$= \frac{r^2-\left(x_P+\Delta x\right)^2-r^2+x_P^2}{\Delta x\left[\sqrt{r^2-\left(x_P+\Delta x\right)^2}+\sqrt{r^2-x_P^2}\right]}, \tag{16.16}$$

$$= \frac{-2x_P\Delta x-\Delta^2 x}{\Delta x\left[\sqrt{r^2-\left(x_P+\Delta x\right)^2}+\sqrt{r^2-x_P^2}\right]}, \tag{16.17}$$

$$= \frac{-2x_P-\Delta x}{\left[\sqrt{r^2-\left(x_P+\Delta x\right)^2}+\sqrt{r^2-x_P^2}\right]}. \tag{16.18}$$

中间我们用了一次平方差公式 $(a+b)(a-b)=a^2-b^2$。注意，这里已经不是一个 $\frac{0}{0}$ 了，当 Δx 越来越小的时候分子为 $-2x_P$，分母为 $2\sqrt{r^2-x_P^2}$，也就是

$$\frac{df\left(x_P\right)}{dx} = \frac{-2x_P}{2\sqrt{r^2-x_P^2}}, \tag{16.19}$$

$$= -\frac{x_P}{\sqrt{r^2-x_P^2}}. \tag{16.20}$$

这个结果和之前通过集合算出来的结果 $k_2=-\frac{x_P}{\sqrt{r^2-x_P^2}}$ 完全吻合。

这个在圆这条具体曲线上的几何计算结果和“导数”——尽管还没有严格定义但是可操作了——的计算结果完全相同的结论表示，至少，把曲线的切线的斜率和导数联系起来，对于一般的函数一般的曲线，很可能也是对的。这是导数这个概念的提出得到的另一个启发以及需求。

合起来，我们有了：**无论是来自于物理还是来自于几何的启发，导函数 $f'(x)$ 就是原函数 $f(x)$ 在 $(x,f(x))$ 这个点的切线的斜率**。

除了物理和几何的启发，导数的提出同时还受到求函数极值的需求的推动。对一个函数求极值是一个无论在现实生活还是数学本身都是常见的问题。那极值有什么特征呢？以极大值 $f\left(x_*\right)$ 和极大值点 x_* 为例，大概来说，就是这个点的附近的点都比它小 $f\left(x\right)\leqslant f\left(x_*\right)$，或者说，当自变量 x 从左侧接近 x_* 这个点的时候，$f\left(x\right)$ 在增加，当自变量 x 离开 x_* 往右侧走的时候，$f\left(x\right)$ 在减小。

于是，大概来说，在 $x < x_*$ 又在 x_* 附近的时候，曲线是往上走的，或者说，曲线 $f(x)$ 上两点之间的连线的斜率大于零 $\frac{f(x+\Delta x)-f(x+\Delta x)}{\Delta x} > 0$；在 $x > x_*$ 又在 x_* 附近的时候，曲线是往下走的，或者说，曲线 $f(x)$ 上两点之间的连线的斜率小于零 $\frac{f(x+\Delta x)-f(x+\Delta x)}{\Delta x} < 0$。那么，如果 $\frac{f(x+\Delta x)-f(x+\Delta x)}{\Delta x}$ 可以在 $x = x_*$ 的点定义——这时候就相当于有了 $f'(x_*)$，并且这个函数 $f'(x)$ 还比较好——有连续性（从大于零到小于零中间必定经过等于零），我们就有了切线的斜率在极大值点 x_* 为零。也就是，

$$f'(x_*) = 0. \tag{16.21}$$

这被称为费马（Fermat）定理。

定理 16.1: 费马（Fermat）定理

函数 $f(x)$ 在 x_* 有导数并且 $f(x)$ 在 x_* 是极值（在 x_* 附近比其他 $f(x)$ 都大或者都小），则 $f'(x_*) = 0$.

有了费马（Fermat）定理，一个求函数极值的问题就变成了一个求其到函数零点的问题，而后者就是解方程。这样一个可能需要高度技巧性来解决的问题就成了一个有比较通用的方法来求解的问题了。这也是数学和科学的一个追求：找到问题的通用的解决方法。有了之后，人们只要照着这个方法做就可以解决问题了。我们用几个例子来体会一下一旦有了导数之后，求极值的难易程度和没有导数时候的差异。

我们看到，求极值的需求也使得我们考虑计算 Δx 越来越小的时候，计算 $\frac{f(x+\Delta x)-f(x+\Delta x)}{\Delta x}$ 的问题。

例 16.3 (没有约束的极值问题). 求二次函数 $y = x^2 + 2x + 3$ 的极值点和极值。

不用导数的概念，对于二次函数我们总是可以凑成完全平方。于是，$y = x^2 + 2x + 3 = (x+1)^2 + 2 \geqslant 2$。因此，$y = x^2 + 2x + 3$ 在 $x = -1$ 处有一个极小值，在 $x = -1$ 的左右两侧，其都会随着偏离 $x = -1$ 的增加而越来越大。

我们来用上导数。按照导数计算三步，我们有

$$\begin{aligned}
&\frac{\left[(x+\Delta x)^2+2(x+\Delta x)+3\right]-\left[x^2+2x+3\right]}{\Delta x},\\
=&\frac{\left[x^2+2x\Delta x+\Delta^2 x+2x+2\Delta x+3\right]-\left[x^2+2x+3\right]}{\Delta x},\\
=&\frac{2x\Delta x+\Delta^2 x+2\Delta x}{\Delta x},\\
=&2x+2+\Delta x. \qquad (16.22)
\end{aligned}$$

当 Δx 越来越小的时候，

$$y'=\frac{dy}{dx}=2x+2. \qquad (16.23)$$

我们来求出其零点 x_*，

$$2x_*+2=0. \qquad (16.24)$$

因此，$x_*=-1$，是 $y=x^2+2x+3$ 的极值点，相应的极值为 $y_*=2$。

两个方法得到的答案完全相同。

求导数的方法和依赖于对二次函数的凑完全平方的方法的好处是，我们可以什么关于这个具体函数的知识都没有，就能解决问题。

例 16.4 (约束下的极值问题). 用一根长度固定为 L 的线围成一个长方形，要求面积最大，长方形的长和宽分别为多少。

我们假设长方形的长和宽分别为 x,y，则其周长为 $x+y=\frac{L}{2}$，面积为 $S=xy$。我们希望找出来合适的 (x,y) 使得 S 最大。我们把 $x+y=\frac{L}{2}$ 这个关于 x,y 的必须满足的条件称为约束。这个约束也可以写成右侧等于零的形式，

$$C(x,y)\triangleq x+y-\frac{L}{2}=0. \qquad (16.25)$$

做一个变量替换，我们得到 $S=xy=x(\frac{L}{2}-x)=-x^2+\frac{L}{2}x$。我们回到了上一题的情况。用求导数或者凑完全平方可以解决。这里，我们用求导数。

$$\begin{aligned}
&\frac{-\left[(x+\Delta x)^2+\frac{L}{2}(x+\Delta x)\right]-\left[-x^2+\frac{L}{2}x\right]}{\Delta x},\\
=&\frac{-\left[x^2+2x\Delta x+\Delta^2 x+\frac{L}{2}x+\frac{L}{2}\Delta x\right]-\left[-x^2+\frac{L}{2}x\right]}{\Delta x},\\
=&\frac{-2x\Delta x-\Delta^2 x+\frac{L}{2}\Delta x}{\Delta x},\\
=&-2x+\frac{L}{2}-\Delta x. \qquad (16.26)
\end{aligned}$$

当 Δx 越来越小的时候，

$$S'=\frac{dS}{dx}=-2x+\frac{L}{2}. \qquad (16.27)$$

我们来求出其零点，

$$-2x_*+\frac{L}{2}=0. \qquad (16.28)$$

因此，$x_*=\frac{L}{4}$ 是极值点，相应的极值为 $S_*=\frac{L^2}{16}$。

还有一个更简单的不需要做变量替换的称为*Lagrangian* 乘子法的方式——变量替换要求解出来某个函数（这里是 $x+y=\frac{L}{2}$）的反函数而这个不一定能够求得出来。这里我们仅仅展示一下其核心思想和计算结果。

改造目标函数为

$$F(x,y;\lambda)=S(x,y)+\lambda C(x,y), \qquad (16.29)$$

然后把 $(x,y;\lambda)$ 都看作自变量，分别求导数，并且让这些导数等于零。

我们来求一下这些导数。其中 x,y 的我们仅仅给出结果，

$$F'_x=y+\lambda, \qquad (16.30)$$

$$F'_y=x+\lambda. \qquad (16.31)$$

而 λ 的导数我们做一个详细的计算。按照导数计算三步，我们有

$$\begin{aligned}
&\frac{[S(x,y)+(\lambda+\Delta\lambda)C(x,y)]-[S(x,y)+\lambda C(x,y)]}{\Delta\lambda},\\
=&\frac{\Delta\lambda C(x,y)}{\Delta\lambda},\\
=&C(x,y). \qquad (16.32)
\end{aligned}$$

当 Δx 越来越小的时候，

$$F'_\lambda = C(x,y). \tag{16.33}$$

我们必须使得这个导数等于零，也就是

$$C(x,y) = 0. \tag{16.34}$$

而这正好就是题目中要求满足的约束。

回过头来，既然这个约束总是要被满足，因此求极值的目标函数 $F(x,y;\lambda) = S(x,y) + \lambda C(x,y)$ 其实就回到了原来的目标函数 $F(x,y;\lambda) = S(x,y)$。其实是可以更加严格地证明这个结论的，但是，这个直觉的理解暂时就够用了。

把三个导数等于零的条件整合到一起，就成了我们最后要求解的方程，

$$0 = F'_x = y + \lambda, \tag{16.35}$$

$$0 = F'_y = x + \lambda, \tag{16.36}$$

$$0 = x + y - \frac{L}{2}. \tag{16.37}$$

这个方程的解为

$$x = \frac{L}{4} = y = -\lambda. \tag{16.38}$$

在这个极值点上，$S = xy = \frac{L^2}{16}$。两个方法得到的答案完全相同。

这里面我们用到了对于有两个和两个以上的自变量的函数 $F(x,y,\lambda)$ 的导数 F'_x, F'_y, F'_λ。这些导数就是假设其他自变量不变当作常数，仅仅改变所针对的自变量的导数。这样的仅仅考虑某个自变量的变化的导数被称为偏导数，记为

$$\frac{\partial F(x,y,\lambda)}{\partial x} = F'_x. \tag{16.39}$$

有了Lagrangian 乘子法，对于更加复杂的约束更加复杂的目标函数，我们就有了一般性的方法了。例如，我们可以问在圆上找一点使得目标函数 $F(x,y) = x + y$ 最大。

习题 16.1 (圆上的约束极值问题). 从单位圆 $x^2 + y^2 = 1$ 上找一个点 (x,y) 使得 $F(x,y) = x + y$ 最大。(提示：这个问题可以用几何方式求解，用本节的方法就需要改造目标函数为 $F(x,y;\lambda)$。在几何那一节设计一个相应的例题或者习题。)

更一般地，如果可以围成任何平面图形，要求面积最大，这个图形长什么样?

例 16.5 (最短距离问题). 已知在平面上的两点 $P=(x_P,y_P)$ 和 $Q=(x_Q,y_Q)$，求过这两点的距离最短的曲线。

根据平面几何知识，我们知道两点之间直线距离最短，而且两点之间必然存在一条而且是唯一的一条直线。这些知识在*Euclid* (欧几里得) 的平面几何知识体系里面是公理。我们可以基于其他的知识来得到这些知识吗？

如果可以的话，未来“其他的知识”一旦改变了，会怎样？比如说，我们也知道一个事实性知识：在地球表面这样的球形表面上两点之间大圆线——半径为球体半径的圆——的长度最短。也就是说，在球形表面大圆线就取代了平面集合中的直线的位置。这其实就是非欧几何。那么，大圆的什么造成了这个结果呢？一旦我们可以找到这个是什么，那未来一旦这个“什么”改变了，我们也就自然可以得到新的情况下的距离最短的曲线的问题了。

我们把这个结果写在这里——这是对上面的用文字描述的问题的数学化，但是，你暂时不需要搞懂它。给定两点 $P=(x_P,y_P)$ 和 $Q=(x_Q,y_Q)$ 的曲线可以被描述为

$$x=x(\tau),y=y(\tau). \tag{16.40}$$

其中，初始点和结束点分别为 P,Q，也就是

$$x(0)=x_P,y(0)=y_P; \tag{16.41}$$

$$x(t)=x_Q,y(t)=y_Q. \tag{16.42}$$

在这个描述下，曲线的长度——这个未来会解释——为

$$\mathscr{L}=\int_0^t d\tau\sqrt{(x'(\tau))^2+(y'(\tau))^2}. \tag{16.43}$$

其中 $x'(\tau)=\frac{dx(\tau)}{d\tau},y'(\tau)=\frac{dy(\tau)}{d\tau}$。

这样我们就有了一个自变量为起点和终点确定函数形式任意的 $(x(\tau),y(\tau))$ 函数，或者说轨迹，或者说曲线，因变量为由公式 (16.43) 给出的这个轨迹的长度的极值问题：问过固定两点的什么样的曲线可以使得曲线长度最小。

不过，可惜这个问题的解决，尽管思想我们已经掌握了，就是通过求到函数零点来求极值，但是具体求解的计算还需要进一步学习才能掌握。我们以后会回到这个问题。

通过这个问题，我们知道了有些很有学科和现实意义的问题可以被表述为以函数为自变量的极值问题。实际上，物理学中的光线的问题就可以从这个角度来理解：光选择的路径是所有路径中距离最短的那一条。

还有另一个自变量为函数的极值问题——最速下降线问题，本身就在物理现象里面有意义，同时历史上大大推动了微积分的发展。我们把这个问题也写在这里，但是留到以后来解决。

例 16.6 (最速下降线问题). 已知在平面上的两点 $P = (x_P, y_P)$ 和 $Q = (x_Q, y_Q < y_P)$，求如果从 P 点以初始速度为零（$\vec{v} = 0$ 表示 $v_x = 0 = v_y$）释放一个小球沿着曲线到 Q 点的用时最短的那条曲线。其中 Q 点在 P 点的下方（$y_Q < y_P$）。

在这里，我们暂时跳过把这个问题数学化的过程。也就是我们暂时就不给出明确的自变量、目标函数和约束了。这个问题的数学化和求解的过程，我们都会未来来完成。

这个问题为什么在物理学中是有意义的呢？比如说，我们要设计一条轨道，比一比，起点和终点正好是给定的 P, Q 两点，我们说哪个轨道上的小球先完成全程谁就赢，就正好就是这个最速下降线的问题。至于为什么会赢，我们可以在终点放一大笔奖金，或者放个触发把对方点击一下的惩罚按钮，或者就是为了在工程上节省时间。

通过这几个例子，我们想传达给你这样一个信息：回到微积分被创造出来之前，如果你也受到了来自于物理的（现实世界的）、几何的启发，受大了求极值问题的刺激，只要你足够努力，又不傻，稍微有点运气，你也可以创造出来微积分。当然，真的把微积分真的创造出来这个贡献绝对不会因为这个“回到当时，我也可以”的回顾性梳理导致的认知而被看得更低。我们所传达的这个信息的目的在于帮助你体会到：通过提出和解决问题，无论是现实世界的还是数学本身的，你也可以来创造数学知识。

16.1.2 导数计算的几个定理

有了导数计算三步和导数的含义，我们来证明几个关于导数的定理。注意，证明这几个定理主要是为了熟悉导数计算三步。

定理 16.2: 常数乘以函数的导数

证明 $\frac{d[af(x)]}{dx} = a\frac{df(x)}{dx}$。其中 a 是不依赖于 x 的常数。

证明. 首先，按照导数计算三步，$af(x)$ 的导数来自于

$$\frac{af(x+\Delta x)-af(x)}{\Delta x} = a\frac{f(x+\Delta x)-f(x)}{\Delta x}, \tag{16.44}$$

而当 Δx 越来越小的时候，已知

$$\frac{f(x+\Delta x)-f(x)}{\Delta x} \to \frac{df(x)}{dx}, \tag{16.45}$$

因此

$$\frac{d[af(x)]}{dx} = a\frac{df(x)}{dx}. \tag{16.46}$$

□

定理 16.3: 和函数的导数

证明 $\frac{d[f(x)+g(x)]}{dx}=\frac{df(x)}{dx}+\frac{dg(x)}{dx}$。

证明. 首先，按照导数计算三步，$f(x)+g(x)$ 的导数来自于

$$\begin{aligned}&\frac{[f(x+\Delta x)+g(x+\Delta x)]-[f(x)+g(x)]}{\Delta x}\\&=\frac{f(x+\Delta x)-f(x)}{\Delta x}+\frac{g(x+\Delta x)-g(x)}{\Delta x},\end{aligned}\tag{16.47}$$

而当 Δx 越来越小的时候，已知

$$\frac{f(x+\Delta x)-f(x)}{\Delta x}\to\frac{df(x)}{dx},\tag{16.48}$$

$$\frac{g(x+\Delta x)-g(x)}{\Delta x}\to\frac{dg(x)}{dx},\tag{16.49}$$

因此

$$\frac{d[f(x)+g(x)]}{dx}=\frac{df(x)}{dx}+\frac{dg(x)}{dx}.\tag{16.50}$$

□

以上两个结果合起来我们有导数计算的线性性，

$$\frac{d[af(x)+bg(x)]}{dx}=a\frac{df(x)}{dx}+b\frac{dg(x)}{dx}.\tag{16.51}$$

或者记为

$$[af(x)+bg(x)]'=af'(x)+bg'(x).\tag{16.52}$$

我们来用导数计算三步证明另一个重要的导数计算公式——莱布尼兹法则，或者称为积法则，乘积函数的导数。

定理 16.4: 乘积函数的导数——莱布尼茨法则

证明 $[f(x)g(x)]' = f(x)'g(x) + f(x)g'(x)$。

证明. 首先，按照导数计算三步，$[f(x)g(x)]'$ 来自于

$$\begin{aligned}
&\frac{[f(x+\Delta x)\,g(x+\Delta x)] - [f(x)\,g(x)]}{\Delta x} \\
&= \frac{f(x+\Delta x)\,g(x+\Delta x) - f(x)\,g(x+\Delta x)}{\Delta x} \\
&\quad + \frac{f(x)\,g(x+\Delta x) - f(x)\,g(x)}{\Delta x}, \\
&= \frac{f(x+\Delta x) - f(x)}{\Delta x} g(x+\Delta x) + f(x)\,\frac{g(x+\Delta x) - g(x)}{\Delta x}. \qquad (16.53)
\end{aligned}$$

而当 Δx 越来越小的时候，上式的左侧和右侧分别成为

$$[f(x)g(x)]' = f(x)'g(x) + f(x)g'(x). \qquad (16.54)$$

□

我们再来证明一个复杂点的复合函数的导数公式——链式法则。

定理 16.5: 复合函数的导数——链式法则

证明 $\frac{d[f(g(x))]}{dx} = \frac{df(g)}{dg}\frac{dg(x)}{dx}$。

证明. 首先，按照导数计算三步，$f(g(x))$ 的导数来自于[a]

$$\begin{aligned}&\frac{[f(g(x+\Delta x))]-[f(g(x))]}{\Delta x}\\ =&\frac{[f(g(x+\Delta x))]-[f(g(x))]}{g(x+\Delta x)-g(x)}\frac{g(x+\Delta x)-g(x)}{\Delta x}\end{aligned} \tag{16.55}$$

而当 Δx 越来越小的时候，已知

$$\frac{[f(g(x+\Delta x))]-[f(g(x))]}{g(x+\Delta x)-g(x)} \to \frac{df(g)}{dg}, \tag{16.56}$$

$$\frac{g(x+\Delta x)-g(x)}{\Delta x} \to \frac{dg(x)}{dx}, \tag{16.57}$$

因此

$$\frac{d[f(g(x))]}{dx} = \frac{df(g)}{dg}\frac{dg(x)}{dx}. \tag{16.58}$$

或者记为

$$f'(g(x))_x = f'(g)_g\, g'(x)_x = f'_g g'_x. \tag{16.59}$$

□

[a]严格证明比这个复杂：$g(x+\Delta x)-g(x)$ 并不独立，实际上是受 Δx 控制，通过 $g(x)$ 来发挥作用。而在 $\frac{df(g)}{dg}$ 中 g 是独立变化的。严格证明要处理这个情况。此外，当遇到某些 $\Delta x \neq 0$ 但是 $g(x+\Delta x)-g(x)=0$ 的时候，会出现分母等于零的情况。不过，我们暂时“满足”于这个不那么严格的证明。

特别地当 f 就是 g 的逆函数 $f = g^{-1}$ 的时候 $g^{-1}(g(x)) = x$，例如 $g(x) = x^3$，$g^{-1}(x) = x^{\frac{1}{3}}$，$g^{-1}(g(x)) = (x^3)^{\frac{1}{3}} = x$。因此，对 $g^{-1}(g(x)) = x$ 左右做导数计算得到

$$\left(g^{-1}\right)'_g g'_x = x'_x = 1, \tag{16.60}$$

因此，

$$\left(g^{-1}\right)'_g = \frac{1}{g'_x}. \tag{16.61}$$

或者记为

$$\frac{dx}{dy}=\frac{1}{\frac{dy}{dx}}. \tag{16.62}$$

逆函数的导数等于原来的函数的导数的倒数。其中我们用了 $y=g(x)$，$x=g^{-1}(y)$。这使得 $\frac{dy}{dx}$ 看起来更像一个分数了：没准 dy 和 dx 真的各自有意义?

再次强调，对于我们来说，记住这些具体公式是次要的，重要的是通过证明这些公式更好地理解导数的含义和导数计算三步。当然，下面的内容可能也会用到这几个证明了的定理。

16.1.3 不定积分计算的几个定理

为了给后面计算定积分服务，Leibniz为这个原函数和导函数之间的关系专门发明了一个逆向计算的记号，

$$F\left(x\right)=\int dxf\left(x\right). \tag{16.63}$$

原函数 $F\left(x\right)$ 被称为是导函数 $f(x)$——这里也称为被积分的函数，简称被积函数——的不定积分。本来这个不定积分的记号有意义的前提是 $\int$ 和 dx 分开有意义。但是，目前，我们仅仅把这句话的含义当作——$f(x)$ 是原函数 $F(x)$ 的导数，也就是

$$\frac{dF\left(x\right)}{dx}=f\left(x\right), \tag{16.64}$$

没有任何新的含义。稍后我们有了定积分的概念会更清楚为什么这叫不定积分。这里的符号你也暂时不用管。等学习了定积分我们会更理解这些符号。

注意，$F(x)$ 和 $F(x)+C$ 对应着相同的 $f(x)$，也就是同一个 $f(x)$ 可以对应着一群相差一个常数项的不定积分函数。

在下面的计算中，对于不定积分函数 $F\left(x\right)$，我们总是把相差一个常数的函数认为是同一个，也就是不再区分 $F\left(x\right)$ 和 $F\left(x\right)+C$。

根据导数计算的线性性公式 (16.51)，我们同样也有不定积分计算的线性性，

$$\int dx\left[af\left(x\right)+bg\left(x\right)\right]=a\int dxf\left(x\right)+b\int dxg\left(x\right). \tag{16.65}$$

这个只需要对公式 (16.51) 两边求不定积分就能得到上面的结果。证明留作练习。

习题 16.2 (不定积分线性性). 从公式 (16.51) 证明不定积分计算的线性性公式 (16.65)。

根据乘积函数的导数计算公式 (16.54)，我们有下面的称为 “分步积分” 的公式，

$$\int dx f(x)g'(x) = f(x)g(x) - \int dx f'(x)g(x). \tag{16.66}$$

这个只需要对公式 (16.54) 两边求不定积分就能得到上面的结果。证明留作练习。

习题 16.3 (分步积分). 从公式 (16.54) 证明不定积分计算的分步积分公式 (16.66)。

根据复合函数的导数计算公式 (16.58)，我们有下面的复合函数积分公式，

$$f\left(g(x)\right) = \int dx f'_g\left(g(x)\right) g'_x(x). \tag{16.67}$$

这个只需要对公式 (16.58) 两边求不定积分就能得到上面的结果。

我们来验证一下。对公式 (16.58) 的左侧做不定积分，我们得到左侧 (LHS，left-hand side) 的结果为

$$\text{LHS} = \int dx \frac{d\left[f\left(g\left(x\right)\right)\right]}{dx} = f\left(g(x)\right). \tag{16.68}$$

对右侧做不定积分，则得到右侧 (RHS，right-hand side) 的结果为

$$\text{RHS} = \int dx \frac{d\left[f\left(g\right)\right]}{dg}\frac{d\left[g\left(x\right)\right]}{dx} = \int dx f'\left(g\right) g'\left(x\right). \tag{16.69}$$

根据等式的性质，对等式的两边做相同的结果唯一 (我们已经把相差一个常数的不定积分函数认同) 的操作，等号仍然成立。于是，我们得到公式 (16.67)。未来公式 (16.67) 可以帮助我们对积分做变量替换。我们在定积分再来讨论变量替换。

为什么我们会定义不定积分呢？除了表示导数的积运算，实际上更重要的是因为受到定积分的启发。那我们为什么会定义定积分呢，定积分本身受到了什么的启发呢？

16.1.4 基于导数的近似——Taylor 公式

我们来证明一个未来可以进一步拓展的对函数做近似的基于导数的表达式——Taylor 公式。其一阶形式如下，

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + o(x - x_0). \tag{16.70}$$

其中，$o(x - x_0)$ 表示比 $(x - x_0)$ 更高阶的小量，也就是当 $(x - x_0) \to 0$ 的时候，

$$\frac{o(x - x_0)}{(x - x_0)} \to 0. \tag{16.71}$$

我们来用反证法，假设当 $(x - x_0) \to 0$ 的时候，

$$\frac{o(x - x_0)}{(x - x_0)} \to \alpha \neq 0, \tag{16.72}$$

则 $f(x)$ 在 x_0 这一点的导数为，

$$\begin{aligned} f'(x_0) = \frac{f(x) - f(x_0)}{x - x_0} &= \frac{f'(x_0)(x - x_0) + o(x - x_0)}{x - x_0}, \\ &\to f'(x_0) + \alpha \neq f'(x_0). \end{aligned} \tag{16.73}$$

矛盾。因此，原假设错误，$(x - x_0) \to 0$ 的时候 $\frac{o(x-x_0)}{(x-x_0)} \to 0$ 必然成立。

我们可以进一步假设

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + a(x - x_0)^2 + o(x - x_0)^2, \tag{16.74}$$

来找找这样的 a。如果找到了，那么我们就进一步使得这个近似更加准确了。为了解决这个问题，我们来算一下 $f(x)$ 的导数，

$$f'(x) = f'(x_0) + 2a(x - x_0) + o(x - x_0). \tag{16.75}$$

我们发现，这里的 $2a$ 相当于是 $f'(x)$ 的导数，也就是

$$a = \frac{1}{2}\frac{df'}{dx}(x_0). \tag{16.76}$$

其中 $\frac{df'}{dx}(x_0)$ 也称为 $f(x)$ 的二阶导数，记为

$$f''(x) = \frac{d^2 f}{dx^2}(x) = \frac{df'}{dx}(x) = \frac{d\frac{df}{dx}}{dx}(x). \tag{16.77}$$

因此，

$$f(x)=f(x_0)+f'(x_0)(x-x_0)+\frac{1}{2}f''(x_0)(x-x_0)^2+o(x-x_0)^2. \quad (16.78)$$

原则上，我们可以一直做下去，找到三、四、五等等阶近似的系数，这样就用多项式来近似了一个任意的函数，

$$f(x)=\sum_{j=0}^{n-1}\frac{1}{j!}f^{(j)}(x_0)(x-x_0)^j+o(x-x_0)^n. \quad (16.79)$$

其中

$$f^{(j)}(x)=f^{\overbrace{\prime\cdots\prime}^{j}}(x)=\frac{d^jf}{dx^j}(x) \quad (16.80)$$

是对函数 $f(x)$ 求了 (j) 次导数，称为 $f(x)$ 的 (j) 阶导数。证明留作练习。

习题 16.4 (Taylor 公式). 证明公式 (16.79) 成立。(提示：可以用数学归纳法)

有了导数就可以用多项式逼近任意函数是导数的另一个典型应用场景。例如，我们来计算一下 $\sqrt{1+x}$ 在 $x=0$ 附近的多项式表达形式。

例 16.7 (Taylor 公式应用). 计算 $f(x)=\sqrt{1+x}$ 在 $x=0$ 附近的多项式表达形式，近似到 x^2 截至。

按照 *Taylor* 公式公式 (16.79)，我们来计算 $f(x)=\sqrt{1+x}$ 的第一阶和第二阶导数。

我们先用导数计算三步，再用前面证明的导数计算的定理来得到第一阶导数。先计算差和比

$$\begin{aligned}\frac{f(x+\Delta x)-f(x)}{\Delta x}&=\frac{\sqrt{1+x+\Delta x}-\sqrt{1+x}}{\Delta x},\\&=\frac{1+x+\Delta x-1-x}{\left(\sqrt{1+x+\Delta x}+\sqrt{1+x}\right)\Delta x},\\&=\frac{1}{\sqrt{1+x+\Delta x}+\sqrt{1+x}}.\end{aligned} \quad (16.81)$$

再让 $\Delta x\to 0$，得到

$$f'(x)=\frac{1}{2\sqrt{1+x}}. \quad (16.82)$$

我们再用乘积函数导数公式——莱布尼茨法则来计算这个一阶导数。选 $g(x) = f(x)f(x) = 1 + x$，则

$$g'(x) = 1, \tag{16.83}$$

同时

$$g'(x) = 2f'(x)f(x), \tag{16.84}$$

得到，

$$f'(x) = \frac{g'(x)}{2f(x)} = \frac{1}{2\sqrt{1+x}}. \tag{16.85}$$

我们再用复合函数求导的链式法则来计算这个一阶导数。选 $g(x) = x^2$，则 $g(f(x)) = (\sqrt{1+x})^2 = 1 + x$，于是

$$\frac{dg(f(x))}{dx} = 1. \tag{16.86}$$

同时按照链式法则

$$\frac{dg(f(x))}{dx} = g'_f f'_x = 2f(x) \cdot f'_x. \tag{16.87}$$

因此，

$$f'(x) = \frac{1}{2f(x)} = \frac{1}{2\sqrt{1+x}}. \tag{16.88}$$

三种方法计算出来的结果完全相同。

我们再来计算二阶导数，也就是 $\frac{1}{2\sqrt{1+x}}$ 的一阶导数，

$$f''(x) = \frac{d\frac{1}{2\sqrt{1+x}}}{dx}. \tag{16.89}$$

我们这次正面来使用复合函数求导的链式法则：选 $u(x) = x^{-\frac{1}{2}}, v(x) = 1+x$，则 $\frac{1}{\sqrt{1+x}} = u(v(x))$，根据链式法则得到

$$\frac{du(v(x))}{dx} = u'_v(x)v'(x) = -\frac{1}{2}v^{-\frac{3}{2}}(x) \cdot 1 = -\frac{1}{2}(1+x)^{-\frac{3}{2}}. \tag{16.90}$$

于是，

$$f''(x) = -\frac{1}{4}(1+x)^{-\frac{3}{2}}. \tag{16.91}$$

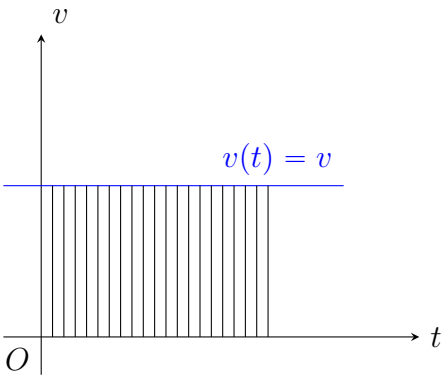


图 16.3: **速度曲线下的面积**。对于最简单的 $v(t)=v$ 的情况速度曲线下的面积是位移。

在 $x=0$ 点取值，这些导数项的值为，

$$f'(0)=-\frac{1}{2}, f''(0)=-\frac{1}{4}. \tag{16.92}$$

于是根据 *Taylor* 公式，

$$\sqrt{1+x}=1+\frac{1}{2}x-\frac{1}{8}x^2+o\left(x^2\right). \tag{16.93}$$

更一般地，还可以证明在 $x=0$ 附近，

$$(1+x)^n=1+nx+\frac{n(n-1)}{2}x^2+o\left(x^2\right). \tag{16.94}$$

具体证明留作练习。

16.2 定积分及其与导数的关系

同样我们也来创造性地体验一下当时创造出来定积分的概念和相关命题的人的思考过程，感受一下其所受到的启发，其所面对的挑战。

16.2.1 定积分受到的启发

讲完了导数——未来我们也称之为微分——得到的启发，现在，让我们再一次回到运动学里面来得到一个关于定积分的启发。如图 16.3 所示，

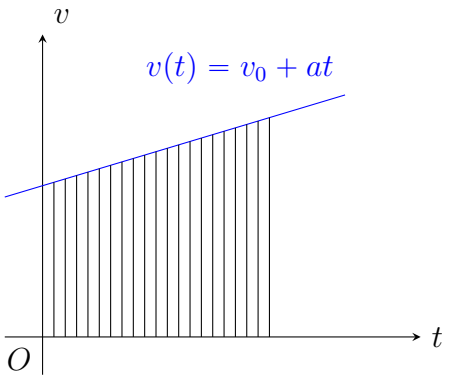


图 16.4: **速度曲线下的面积**。对于匀加速直线运动的 $v(t)=v_0+at$ 的情况速度曲线下的面积是位移。

先从最简单的开始，考虑匀速直线运动也就是当 $v(t)=v$ 是常数的时候，$x(t)-x_0=vt$，也就是位置的变化量——物理上称为位移——是曲线 $v(t)=v$ 到 t 轴在时间间隔 $0\sim t$ 之间所包含的面积：$v(t)=v$ 是一条水平直线，t 轴到 t 时刻截至构成一个长方形，其面积为 $S=vt$。

这启发我们猜测：是不是对于更加一般的 $v(t)$ 和相应的 $x(t)$，这个关系也对呢？我们再来试一个已知的运动——匀加速直线运动，如图 16.4 所示，满足 $v(t)=v_0+at$ 和 $x(t)=x_0+v_0t+\frac{1}{2}at^2$。我们发现，速度函数和 t 轴在时间间隔 $0\sim t$ 之间包围起来的是一个梯形，其面积为 $\frac{1}{2}(v_0+v_0+at)t=v_0t+\frac{1}{2}at^2$。完全和这个关系相符。

反过来，之前学习导数的时候，我们就已经知道 $x'(t)=v(t), v'(t)=a(t)$。也就是说，曲线下的面积在计算上似乎可以运用导函数 $f(x)$ 和原函数 $F(x)$ 的关系。

于是，我们得到下面两个启发：第一，对于给定曲线 $f(x)$，没准我们经常会需要计算这个曲线和 x 轴在间隔 $a\sim b$ 之间包围起来的面积，记作 $F(x)=F(f(x),a,b)$——其依赖于函数 $f(x)$ 以及间隔区间的边界 a,b; 第二，这个新定义的 $F(x)$ 和 $f(x)$ 的关系似乎和导函数和原函数之间的关系，或者前面已经定义的被积函数和不定积分有关系。

我们从第一个启发提出来一个叫做定积分的定义，记作

$$F(f(x),a,b)=\int_a^b dx f(x). \tag{16.95}$$

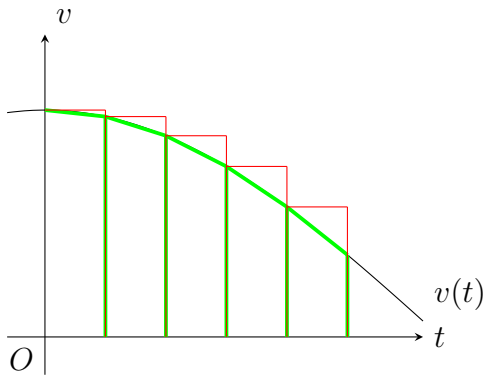


图 16.5: **曲线下的面积的计算方式**。计算曲线下面积的两种方式：梯形法则（绿色）和长方形法则（红色）。

我们从第二个启发就得到

$$F(f(x), a, b) = F(b) - F(a) = F(x)\big|_a^b . \tag{16.96}$$

其中右侧的 $F(b), F(a), F(x)$ 是公式 (16.63) 定义的不定积分，而求不定积分不过就是求导运算的逆运算。

把公式 (16.95) 和公式 (16.96) 合起来写在一起成为

$$F(b) - F(a) = \int_a^b dx f(x) . \tag{16.97}$$

它有一个响亮的名字——牛顿-莱布尼兹公式，又称为微积分基本定理。它把定积分的计算和导数计算联系了起来。未来我们会来证明这个定理。

下面是关于符号的一点点说明：如果我们把公式 (16.64)$\frac{dF}{dx} = f(x)$ 的左边真的看作一个分数（注意，这是基于导数计算三步来定义的，因此不一定能够这样看），则好像我们真的可以把其中的 dx 乘到左边的 $f(x)$ 上去。公式 (16.95) 中的右侧除了 $\int_a^b$ 的部分就这么来的。那是不是真的可以这样看呢，把 $\frac{dF}{dx}$ 真的看作一个分数，或者说把 dx 乘到 $f(x)$ 上去?

如图 16.5 所示，在上面计算曲线下面积的场景下，我们还可以这样来看：我们把整个间隔 $x \in [a, b]$ 分成了宽度为 $\Delta x = \frac{b-a}{N}$ 的小格子：$x_0 = a, x_1 = a + \Delta x, \cdots, x_N = a + N\Delta x$。每一个起点在 x_n 处的小格子其实是一个梯形：上底和下底的高度分别为 $f(x_n)$ 和 $f(x_n + \Delta x)$，高度为 Δx，于

是其面积为

$$\Delta^{(1)}S_n = \frac{1}{2}\left(f\left(x_n\right) + f\left(x_n + \Delta x\right)\right)\Delta x. \tag{16.98}$$

这里最右侧已经有一个 Δx 的一次方项，如果我们忽略更高阶的项，这个小梯形的面积可以近似为一个长方形的面积

$$\Delta^{(2)}S_n = f\left(x_n\right)\Delta x. \tag{16.99}$$

我们来把所有的这样的小梯形的面积加起来，每准就能得到整个区域的面积，

$$\begin{aligned} S^{(1)} = & \sum_{n=0}^{N-1}\Delta^{(1)}S_n, \\ = & \sum_{n=0}^{N-1}\left\{\frac{1}{2}\left(f\left(x_n\right) + f\left(x_n + \Delta x\right)\right)\Delta x\right\}, \\ = & \frac{1}{2}f\left(x_0\right)\Delta x + \frac{1}{2}f\left(x_N\right)\Delta x + \sum_{n=1}^{N-1}\left\{f\left(x_n\right)\Delta x\right\}, \\ = & -\frac{1}{2}f\left(x_0\right)\Delta x + \frac{1}{2}f\left(x_N\right)\Delta x + \sum_{n=0}^{N-1}\left\{f\left(x_n\right)\Delta x\right\}. \end{aligned} \tag{16.100}$$

而

$$S^{(2)} = \sum_{n=0}^{N-1}\Delta^{(2)}S_n = \sum_{n=0}^{N-1}f\left(x_n\right)\Delta x, \tag{16.101}$$

两者相差的项几乎就可以忽略，尤其是当 N 越来越大，Δx 越来越小的时候。从现在起我们忽略这两者的差别。这一点的在数学上的严格证明以后再说。

也就是说，计算函数曲线 $f(x)$ 下的面积的时候，我们看起来就是在计算这样一个取和，

$$S\left[f\left(x\right); a, b; N\right] = \sum_{n=0}^{N-1}f\left(x_n\right)\Delta x. \tag{16.102}$$

这个取和依赖于所给的函数 $f(x)$，依赖于所定的积分的区间 $[a, b]$，依赖于分成小格子的数量 N。但是，我们主要关注 N 越来越大的时候这个 S 的取值，也就是大概可以记作

$$S\left[f\left(x\right); a, b; N \to \infty\right]. \tag{16.103}$$

注意这个时候 $\Delta x=\frac{b-a}{N}$ 也相应地越来越小。

现在终于可以来解释积分的记号了。之前我们有导数计算的记号 $f'(x)=\frac{df(x)}{dx}$。之前，我们说，这个记号要整体认知的，$\frac{df(x)}{dx}$ 就是对 $f(x)$ 求导数的意思，不能拆成分子分母来分别看。具体的计算方法就对应着求导数三步。

现在，我们说，没准我们也可以试试分子分母拆开看。你看在我们导数计算三步中，我们确实先算 $\Delta f=f(x+\Delta x)-f(x)$，再算比值 $\frac{\Delta f}{\Delta x}$。只不过在算比值的时候，先把分子的每一项都除以分母得到 $\Delta^q x$ 的项，再按照 $q>0,q=0,q<0$ 来得到最后的比值。也就是说，在不太准确的含义下，我们确实在计算一个比值，

$$\frac{\Delta f}{\Delta x}\longrightarrow\frac{df(x)}{dx}. \tag{16.104}$$

当然，再一次强调，将来我们会严格地用“极限”的概念给出来上面这个过程的数学上的含义和操作的。但是，目前，有了这个理解之后，似乎我们就可以把 Δx 看作是 dx 的对应（在“越来越小”这个含义下），把 $\Delta f=f(x+\Delta x)-f(x)$ 看作是 $df(x)$ 的对应。

如果仅仅是在导数的时候这样看，也就算了。这仅仅是记号问题。但是，对于积分，这样看是有额外的好处的。未来，我们就可以把公式 (16.102) 写成

$$F[f(x);a,b]=\sum_{n=0}^{N-1}f(x_n)\Delta x=\int_a^b f(x)\,dx. \tag{16.105}$$

也就是把取和号（Sum）改成大大的“S”当作积分号，把 Δx 直接当作 dx。在实践中这个先分成小格子再取和，再转化为积分的过程非常有用，被称为微元法。我们用这个微元法就可以把对问题的很多直觉认知转化为积分表达式。

我们来对前面受到的所有的关于定积分的启发做一个总结。第一，通过曲线下面积的启发我们定义了定积分：先划分成小格子再取和，再转化成定积分。第二，通过速度函数和位置函数之间的关系、加速度函数和速度函数之间的关系，我们得到启发定积分的计算可以转化为不定积分，而求不定积分就是求导数的逆运算。未来我们会把这些受物理和几何启发的结论严格化。在那之前，我们先来体会一下这几样东西的威力：导数的定义公式 (16.12) 和相应的求导三步法，牛顿-莱布尼兹公式也就是公式 (16.97)，以及微元法。

两个特别简单的问题先留作作业。

习题 16.5 (匀加速直线运动从 $x(t)$ 到 $a(t)$). 已知匀加速直线运动的位置函数为 $x(t)=x_0+v_0t+\frac{1}{2}at^2$，通过前面的导数计算三步得到速度函数为 $v(t)=v_0+at$，加速度函数为 $a(t)=a$。其中加速度定义为速度的变化率 $a=\frac{dv}{dt}$。

习题 16.6 (匀加速直线运动从 $a(t)$ 到 $x(t)$). 已知匀加速直线运动的加速度函数为 $a(t)=a$，求其速度函数和位置函数。假设在 $t=0$ 的时刻，$v_0=0$，$x_0=0$。

然后，我们来尝试几个稍微复杂一点的问题。

例 16.8 (x^n 的导数和积分). 求 x^n 的导数 $\frac{dx^n}{dx}$ 和积分 $\int_0^x d\xi\xi^n$。其中 n 是非负整数。

首先，按照导数计算三步骤，有

$$\begin{aligned}\frac{(x+\Delta x)^n-x^n}{\Delta x} &= \frac{\sum_{m=0}^{n}C_n^m x^m\Delta^{n-m}x-x^n}{\Delta x},\\ &= \frac{\sum_{m=0}^{n-1}C_n^m x^m\Delta^{n-m}x}{\Delta x},\\ &= \sum_{m=0}^{n-2}C_n^m x^m\Delta^{n-m-1}x+C_n^{n-1}x^{n-1}\Delta^0 x,\\ &= nx^{n-1}. \qquad (16.106)\end{aligned}$$

最后一步我们用了凡是 $(n-m-1)>0$ 的时候 $\Delta^{n-m-1}x=0$。

于是，我们就得到了 $\frac{d(x^n)}{dx}=nx^{n-1}$，或者说 $(x^n)'=nx^{n-1}$。也就是 $\left(\frac{1}{n}x^n\right)'=x^{n-1}$，或者说

$$\left(\frac{1}{n+1}x^{n+1}\right)'=x^n. \qquad (16.107)$$

因此，按照公式 (16.97)，我们得到

$$\int_0^x d\xi\xi^n=\frac{1}{n+1}x^{n+1}. \qquad (16.108)$$

注意这里的积分区间被我们特意从 $[a,b]$ 换成了 $[0,x]$。

这里我们用到了导数的定义公式 (16.12) 和相应的求导三步法，牛顿-莱布尼兹公式也就是公式 (16.97)。我们从非负整数的指数变成分数指数试试。

例 16.9. *[$x^{\frac{q}{p}}$ 的导数和积分]* 求 $x^{\frac{q}{p}}$ 的导数 $\frac{d\left(x^{\frac{q}{p}}\right)}{dx}$ 和积分 $\int_0^x d\xi \xi^{\frac{q}{p}}$。其中 p,q 是正整数。

设 $u(x)=x^{\frac{q}{p}}$，则 $u^p(x)=x^q$，于是两边做对 x 的求导运算，运用复合函数导数的链式法则，得到

$$pu^{p-1}(x)\cdot u'(x)=qx^{q-1}, \tag{16.109}$$

于是

$$u'(x)=qx^{q-1}\cdot p^{-1}u^{1-p}(x)=\frac{q}{p}x^{q-1}\left(x^{\frac{q}{p}}\right)^{1-p}=\frac{q}{p}x^{\frac{q}{p}-1}. \tag{16.110}$$

也就是

$$\left(\frac{p}{q+p}x^{\frac{q+p}{p}}\right)'=x^{\frac{q}{p}}. \tag{16.111}$$

因此，按照公式 (16.97)，我们得到

$$\int_0^x d\xi x^{\frac{q}{p}}=\frac{p}{q+p}x^{\frac{q+p}{p}}=\frac{1}{\frac{p}{p}+1}x^{\frac{q}{p}+1}. \tag{16.112}$$

顺便，对比公式 (16.107) 和公式 (16.111) 就会发现，两公式相同，也就是其实是同一个公式对正整数幂次方和分数幂次方都成立。

这个公式对于负的整数和分数也成立吗？我们来试试负整数。

例 16.10. *[x^{-n} 的导数和积分]* 求 x^{-n} 的导数 $\frac{d\left(x^{-n}\right)}{dx}$ 和积分 $\int_0^x d\xi\xi^{-n}$。其中 n 是正整数。

设 $u(x)=x^{-n},v(x)=x^n$，则 $u(x)v(x)=x^0=1$，于是两边做对 x 的求导运算，得到

$$[u(x)v(x)]'=(1)'=0, \tag{16.113}$$

于是

$$u'(x)v(x)+u(x)v'(x)=0. \tag{16.114}$$

也就是

$$\left(x^{-n}\right)'x^n+x^{-n}\left(nx^{n-1}\right)=0, \tag{16.115}$$

得到

$$\left(x^{-n}\right)'=-x^{-n}x^{-n}\left(nx^{n-1}\right)=-nx^{-n-1}. \tag{16.116}$$

也就是 $\left(-\frac{1}{n}x^{-n}\right)' = x^{-n-1}$。

因此，按照公式 (16.97)，我们得到

$$\int_{0+\epsilon}^{x} d\xi x^{-n-1} = -\frac{1}{n}x^{-n} + \frac{1}{n}\epsilon^{-n}. \tag{16.117}$$

注意，未来我们看到，这里其实我们遇到了一个 $0^{-n} = \frac{1}{0^n}$ 的问题，所以我们悄悄地修改了积分的间隔为 $[\epsilon > 0, x]$。我们未来会回到这个问题。

我们看到公式 (16.116)、公式 (16.107) 和公式 (16.111) 实际上是相同的公式。你还可以证明这个公式对于负的分数指数也成立。

习题 16.7 ($x^{-\frac{q}{p}}$ 的导数和积分). 求 $x^{-\frac{q}{p}}$ 的导数和积分。其中 p, q 是正整数。(提示：参考例 16.9 或者例 16.10，把转化为已知——顺便这是解决数学问题和一般问题的通用思路之一。)

合起来，我们受关于速度的生活经验和物理学关于速度和位移的关系的启发，受计算曲线上某点的切线和曲线下面积的启发，提出了导数和积分的定义，形成了一套导数和积分的计算方法。尽管这个定义和相应的计算还有逻辑问题，但是，我们能用导数和积分这一对概念还能作计算了。具体计算方法是：导数按照导数计算三步做，积分需要按照牛顿-莱布尼兹公式反过来运用导数——找到什么样的原函数 $F(x)$ 的导数刚好就是被积分的函数 $f(x)$。

其中，我们发现从位置函数 $x(t)$ 算出来的速度函数 $v(t)$ 还真的可以和测量结果比较。当然，在本书中我们被没有去做这样的实验检验，而是通过和我们之前学习过的已经被实验检验过的物理学知识相比，发现完全相符，来做的间接检验。这部分其实是科学研究方法中的数学建模和试验检验，也是科学精神的核心。我们稍后会在双向数学建模一节更详细地展开讨论。

其中，我们还发现数形结合很多时候能够帮助我们思考，化归的思想——把未知转化为已知或者看到不同的对象的背后相同的结构（尤其是数学结构）于是成了同一个对象——对提出和解决问题也很有帮助。

我们再来受另一个需求的启发，从计算曲线长度和计算曲线包围的面积来得到积分。在这里，微元法将发挥作用。

根据平面几何的知识，在 x 轴上的 P, Q（假设 $x_Q > x_P$）两点之间的距离为 $d = x_Q - x_P$。如果我们这样来做一个计算：先把 P, Q 之间划分成

N 个小格子，每个小格子的长度为 $\Delta x = \frac{d}{N}$，则总的长度为

$$\sum_{n=0}^{N-1} \Delta x, \tag{16.118}$$

利用微元法转化成积分，并利用牛顿-莱布尼兹公式完成积分的计算，得到

$$\int_{x_P}^{x_Q} dx = x|_{x_P}^{x_Q} = x_Q - x_P. \tag{16.119}$$

这里我们已经识别出来 $f(x) = 1$，因此 $F(x) = x$。结果就是我们期待的。

那么，如果两个点不在 x 轴上呢？只需要把这两个点之间的直线找出来，算出来这个直线的长度。按照平面几何知识中的勾股定理，应该得到

$$PQ = \sqrt{(x_Q - x_P)^2 + (y_Q - y_P)^2}. \tag{16.120}$$

我们可以用积分把这个算出来吗？我们来试试。

PQ 两点之间的直线的方程为

$$\frac{y - y_P}{x - x_P} = \frac{y_Q - x_P}{x_Q - y_P}, \tag{16.121}$$

也就是

$$y = \frac{y_Q - x_P}{x_Q - y_P}(x - x_P) + y_P, \tag{16.122}$$

或者记为 $y = kx + b$，其中 $k = \frac{y_Q - x_P}{x_Q - y_P}, b = y_P - \frac{y_Q - x_P}{x_Q - y_P} x_P$。

根据这个方程，我们有当 x 增加了 Δx 的时候，

$$\Delta y = [k(x + \Delta x) + b] - [kx + b] = k\Delta x \tag{16.123}$$

这个和从 $y = kx + b$ 直接求导数得到的结论相同 $\frac{dy}{dx} = k$。

有了这个结果，我们就知道了，每一次一个小小的 Δx 带来的相应的小小的 Δy 的变化，也就得到了线长的小小的变化，

$$\Delta s = \sqrt{\Delta^2 x + \Delta^2 y}, \tag{16.124}$$

$$= \sqrt{1 + k^2}\Delta x \tag{16.125}$$

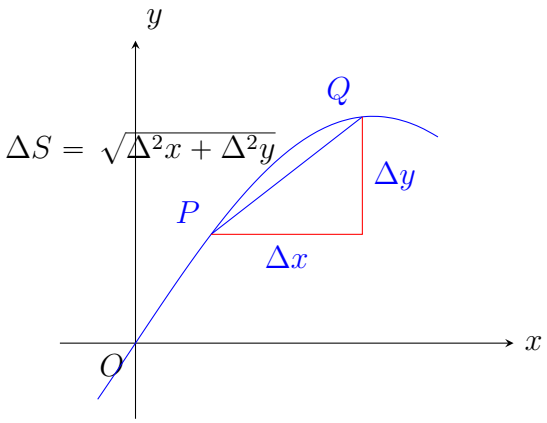


图 16.6: **从曲线上一小段的长度到曲线长度**。基于距离公式可以得到曲线上一小段的长度，对这个长度积分可以得到曲线的长度。

于是，我们相当于要计算这样一个积分，

$$PQ = \int_{x_P}^{x_Q} dx \sqrt{1+k^2}. \tag{16.126}$$

找到 $\sqrt{1+k^2}$ 相应的原函数 $\sqrt{1+k^2}x$，得到

$$\begin{aligned} PQ &= \sqrt{1+k^2}x\Big|_{x_P}^{x_Q}, \\ &= \sqrt{1+\frac{(y_Q-x_P)^2}{(x_Q-y_P)^2}}(x_Q-y_P), \\ &= \sqrt{(x_Q-y_P)^2+(y_Q-x_P)^2}. \end{aligned} \tag{16.127}$$

这正好和公式 (16.120) 的结果完全相同。

如图 16.6 所示，既然可以用一小段线长的积累来算直线的长度，是不是可以用这样的一小段线长的积累来算任何一条曲线的长度呢？比如

$$S = \int_P^Q ds, \tag{16.128}$$

而

$$ds = \sqrt{d^2x + d^2y}. \tag{16.129}$$

如果用 $(x = x(\tau), y = y(\tau))$ 来描写这个条曲线则，

$$ds = \sqrt{(x'(\tau))^2 + (y'(\tau))^2}d\tau. \tag{16.130}$$

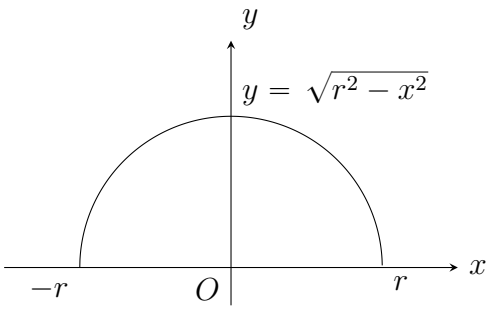


图 16.7: **半圆的长度的积分计算**。按照线长公式计算给定曲线——这里是半圆 $y=\sqrt{r^2-x^2}$ 的长度。

于是

$$S=\int_{\tau_P}^{\tau_Q}\sqrt{\left(x'\left(\tau\right)\right)^2+\left(y'\left(\tau\right)\right)^2}d\tau. \tag{16.131}$$

比如我们就取 $\tau=x$，则，

$$ds=\sqrt{1+\left(y'\left(x\right)\right)^2}dx. \tag{16.132}$$

而整个曲线的长度为

$$S=\int_{x_P}^{x_Q}\sqrt{1+\left(y'\left(x\right)\right)^2}dx. \tag{16.133}$$

我们来算几个例子试试。

例 16.11 (半圆的长度). 半圆的函数为 $y=\sqrt{r^2-x^2}$，$x\in[-r,r]$。求这个曲线的长度，检验是否就是 πr。

按照线长公式 (16.131) 和图 16.7，

$$S=\int_{x=-r}^{x=r}\sqrt{1+\left(y'\left(x\right)\right)^2}dx. \tag{16.134}$$

其中，我们会用到 $y=\sqrt{r^2-x^2}$ 的导数。我们这样算，

$$y'\left(x\right)=\sqrt{u\left(x\right)}, \tag{16.135}$$

而

$$u\left(x\right)=r^2-x^2. \tag{16.136}$$

根据复合函数求导的链式法则定理 16.1.2，或者我们之前已经计算出来的结果例 16.2，有

$$y'(x)=\frac{1}{2}u^{-\frac{1}{2}}(x)\,u'(x)=\frac{-x}{\sqrt{r^2-x^2}}. \tag{16.137}$$

于是，

$$\begin{aligned} S&=\int_{x=-r}^{x=r}\sqrt{1+\frac{x^2}{r^2-x^2}}dx,\\ &=\int_{x=-r}^{x=r}r\frac{1}{\sqrt{r^2-x^2}}dx,\\ &=r\int_{\xi=-1}^{\xi=1}\frac{1}{\sqrt{1-\xi^2}}d\xi. \end{aligned} \tag{16.138}$$

不过要计算出来

$$\int_{\xi=-1}^{\xi=1}\frac{1}{\sqrt{1-\xi^2}}d\xi \tag{16.139}$$

不是特别简单。我们很难看出来导函数 $f(x)=\frac{1}{\sqrt{1-x^2}}$ 的原函数。未来我们会做一个计算来得到这个值。不过，无论算出来这个值是多少，反正是一个常数。假设我们非常凑巧刚好把这个常数记作 π，就有了半圆的长度为 πr。

有了线长，我们来算面积。我们先来计算之前学习过的函数和 x 轴包围起来的面积，再来计算圆的面积。

例 16.12 ($f(x)$ 下到 x 轴的面积)**.** 已知函数 $f(x)$ 在 x 轴上方，计算其和 x 轴在间隔 $x\in[x_1,x_2]$ 之内的面积。

我们先来算在固定的 x 值的地方，从 x 轴到 $f(x)$ 的线长。按照线长公式 (16.131)，

$$l(x)=\int_{y=0}^{y=f(x)}dy=f(x)-0=f(x). \tag{16.140}$$

其中，我们选择了用 y 去描述这条曲线，因为正好当 x 固定的时候，每一条从 x 轴到 $f(x)$ 的直线就是 y 从 0 到 $y=f(x)$ 的直线。

接着我们把这样的直线的长度累计起来算面积，

$$S=\int_{x_1}^{x_2}f(x)dx. \tag{16.141}$$

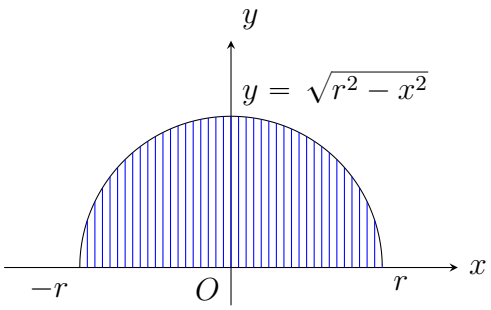


图 16.8: **半圆面积的积分计算**。先计算 y 方向的线长，再把这些线长通过积分累积起来成为面积。

这正好就是我们之前基于几何和物理的启发得到的曲线下面积的计算公式。进一步，如果找得到 $f(x)$ 的原函数 $F(x)$，则

$$S = F(x_2) - F(x_1). \tag{16.142}$$

更一般地，如果 $f(x)$ 有小于零在 x 轴下方的部分，只要我们约定这个时候算出来的面积也是负的，那就在这一段公式 (16.141) 仍然成立。于是，整体合起来，对于有正有负的，只要每一段的积分存在，则公式 (16.141) 仍然成立。

这样我们就从点的积分得到了线长，从线长的积分得到了面积。下面，我们用这个面积的含义来计算圆的面积。

例 16.13 (圆的面积). 我们来算半圆这个曲线 $y = \sqrt{r^2 - x^2}$ 和 x 轴在 $x \in [-r, r]$ 范围内包含的面积。

按照公式 (16.141) 和图 16.8，有

$$S = \int_{-r}^{r} \sqrt{r^2 - x^2} dx = r^2 \int_{-1}^{1} \sqrt{1 - \xi^2} d\xi \tag{16.143}$$

第二步我们用了变量替换 $x = r\xi$。稍后我们会证明一个关于定积分变量替换的定理。这个得到的被积函数的原函数也不容易看出来。未来我们会把这个积分的数值计算出来。这里我们先仅仅给出答案

$$\int_{-1}^{1} \sqrt{1 - \xi^2} d\xi = \frac{\pi}{2}. \tag{16.144}$$

注意，这里的 π 可不能随便给这个常数一个记号了。我们得证明这个积分和之前那个积分得到的是同一个 π，

$$\int_{-1}^{1} \frac{1}{\sqrt{1-\xi^2}} d\xi = \pi. \tag{16.145}$$

未来我们会通过计算来完成这个证明。按照这个结果，我们得到半圆的面积为 $S = \frac{\pi}{2} r^2$，因此圆的面积为 $S = \pi r^2$。

除了先从点用积分算出来线长，再通过线长的积分得到面积之外，我们还可以用微元法来得到面积。例如我们这样来计算圆的面积。

例 16.14 (圆的面积微元法). 基于圆的周长公式和微元法来计算圆的面积。

我们把圆分成多个半径为 $r \in [\xi, \xi + \Delta\xi]$ 的环形，如图 16.9 所示，其中的 $\xi \in [0, r]$。这样每一个环形的周长为 $2\pi\xi$。其宽度为 $\Delta\xi$。这样的近似为一个长方形图形的面积为 $2\pi\xi\Delta\xi$。按照前面的记号和理解，这就相当于求下面的积分

$$S = \int_0^r 2\pi\xi d\xi. \tag{16.146}$$

这个就很容易找到原函数了，$(\pi x^2)' = 2\pi x$。于是

$$S = \int_0^r 2\pi\xi d\xi = \pi\xi^2\big|_0^r = \pi r^2. \tag{16.147}$$

这样我们就基于圆的周长公式证明了圆的面积公式。从这里我们也看出来，结果上前面的那两个 π 确实是同一个常数。

有了这个理解，我们进一步来算一下长方形的面积。我们把这个长方形放到二维平面直角坐标系里面。把长方形的左侧下方的角落放到坐标原点。这样我们的长方形就是这样一个区域：$x \in [0, a]$，$y \in [0, b]$。

例 16.15 (长方形面积积分计算). 计算以下区域的面积：$x \in [0, a], y \in [0, b]$。

按照之前的思路，先从点通过积分得到线的长度，再从对线的积分得到面积，如图 16.10 所示。我们先来算给定 x 点的线段的长度，

$$l(x) = \int_0^b dy = b. \tag{16.148}$$

再把这个线的长度沿着 x 方向做一个积分来得面积，

$$S = \int_0^a l(x)dx = S = \int_0^a bdx = ba. \tag{16.149}$$

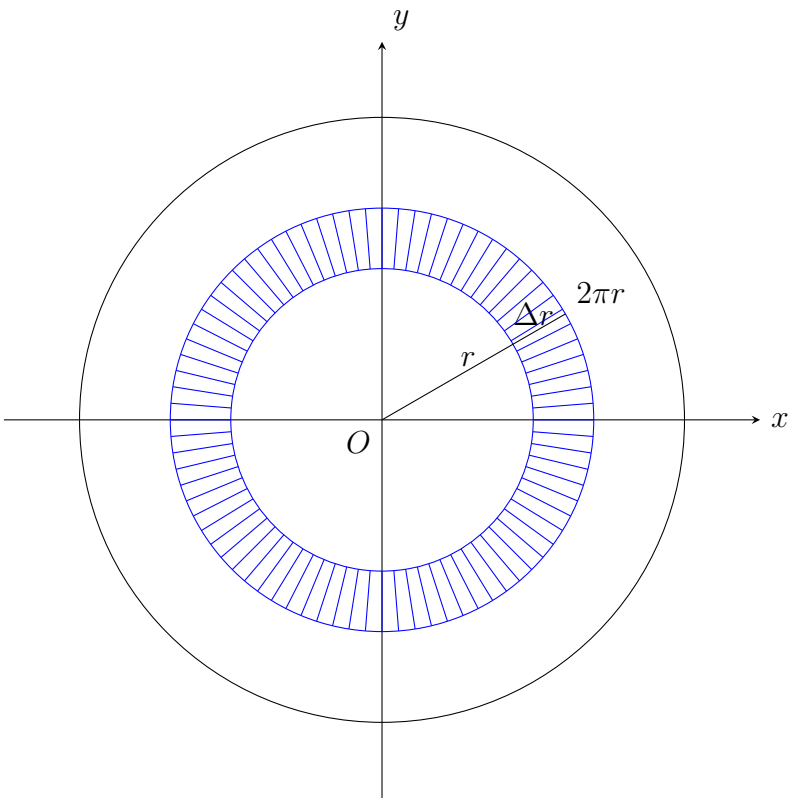


图 16.9: **圆切分成环形微元的面积**。阴影部分表示了周长为 $2\pi r$ 宽度为 Δr 的环形的面积。

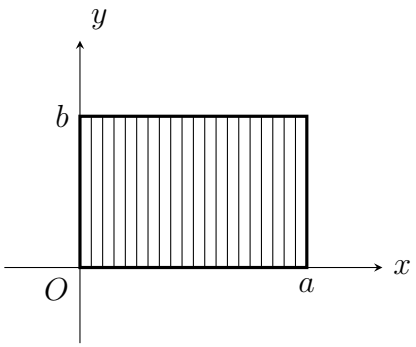


图 16.10: **直角坐标中长方形的面积的积分计算**。先积分得到 y 方向线段的长度然后对这些线段的长度做积分得到面积。

这样我们就从两点之间的距离公式得到了直线的长度（以及任意曲线的长度），从直线的长度得到了长方形的面积公式（以及任意平面图形的面积）。

我们来换一个思路用微元法再来算一遍长方形的面积。

例 16.16（长方形面积的面积元计算）. 计算以下区域的面积：$x \in [0, a]$，$y \in [0, b]$。

我们把这个 x 轴和 y 轴区域分别分成 N 份，每一个小格子都是一个边长为 Δx 和 Δy 的长方形。每一个小格子就如图 16.11 所示。为了视觉效果我们画图的时候放大了这个小格子。因此，整个长方形的面积就相当于求下面的取和，

$$S = \sum_{m,n} \Delta x \Delta y. \tag{16.150}$$

按照前面的从小格子取和和积分之间的对应关系，我们可以把上面的取和表示为

$$S = \iint dxdy. \tag{16.151}$$

其中两个积分号连起来的符号就表示这里有对两个变量的积分。具体算的时候可以先完成对一个变量的积分再完成对另一个变量的积分。因此，我们得到

$$S = \int_0^a dx \left(\int_0^b dy \right) = ab. \tag{16.152}$$

两种方法得到的答案完全相同。

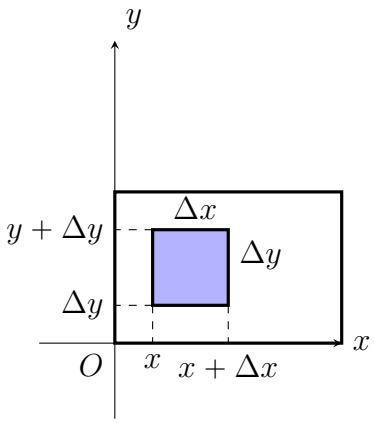


图 16.11: **直角坐标中长方形的微元的面积**。阴影部分表示了在直角坐标中给定 Δx 和 Δy 的一小块类似长方形的形状的面积。

注意微元法其实是基于其画出来的近似小长方形的面积的，因此，和从点积分成为线长，从线长的积分成为面积这个可以用来证明面积公式的求面积思路不同。微元法就是一个把直觉转化为积分表达式的好工具。我们现在来试试用这个微元法来计算圆的面积。

例 16.17 (圆面积的计算——极坐标). 用平面极坐标和微元法计算圆的面积。

在极坐标下，圆的方程为：$\rho \in [0, R]$，$\theta \in [0, 2\pi]$。之前插入弧度和极坐标。 如图 16.12，我们把 ρ 分成 N 份，每一份是 $\Delta\rho$，我们把 θ 分成 M 份，每一份是 $\Delta\theta$。然后可以选取一个小小的面积为下面的形状：$\rho \in [\rho_n, \rho_n + \Delta\rho]$，$\theta \in [\theta_m, \theta_m + \Delta\theta]$。这个小格子的面积近似为

$$\Delta S = (\rho_n \Delta\theta) \cdot \Delta\rho. \tag{16.153}$$

其中我们用到了半径为 ρ 夹角为 $\Delta\theta$ 的圆弧弧长为

$$l_{\Delta\theta} = \rho\Delta\theta. \tag{16.154}$$

按照微元法取和和积分的对应关系，我们有圆面积

$$S = \iint \rho d\rho d\theta = \int_0^r \rho d\rho \left(\int_0^{2\pi} d\theta \right). \tag{16.155}$$

其中后面的积分要先算出来，1 原函数为 θ，因此得到 2π。我们再来算前一个积分 $\int 2\pi\rho d\rho$。其原函数为 $\pi\rho^2$，代入积分边界得到

$$S = \pi\rho^2|_0^r = \pi r^2. \tag{16.156}$$

除了两个不太好算的积分没有算出来，

$$\int_{-1}^{1} \sqrt{1-\xi^2}d\xi = \frac{\pi}{2}, \tag{16.157}$$

$$\int_{-1}^{1} \frac{1}{\sqrt{1-\xi^2}}d\xi = \pi. \tag{16.158}$$

我们展示了导数的定义公式 (16.12) 和相应的求导三步法，牛顿-莱布尼兹公式也就是公式 (16.97)，以及微元法的威力，以及顺便从对点的积分得到了线长，从对线长的积分得到了面积，从而可以用来计算任意平面图形的面积，并且得到了长方形和圆的面积公式。

下面，我们来提供求这两个积分以及比较复杂的积分要用到的原理。

16.2.2 dx, dy, dF 独立有意义——微分

我们从逆函数的导数是原来的函数的导数的倒数公式 (16.62) 得到启发，没准 dx, dy, dF 自己就有意义，而不是一定得整体的 $\frac{dy}{dx}, \frac{dF}{dx}$ 才有意义。我们从定积分的表达式 $\int_a^b f(x)dx$，尤其是 $\int_a^b dx$ 受启发，dx, dy, dF 自己就有意义。我们还要注意到，其实，定积分的积分变量可以随便取，不影响结果，也就是 $\int_a^b f(x)dx = \int_a^b f(\xi)d\xi$。这样的变量还名字写出来的取和表达式是完全相同的，$\sum_{n=1}^{N-1} f(x_n)\frac{b-a}{N}$。

接着，我们可以这样来看积分，

$$F(b) - F(a) = \int_{F(a)}^{F(b)} dF. \tag{16.159}$$

联系到牛顿-莱布尼兹公式也就是公式 (16.97)，重写在这里

$$F(b) - F(a) = \int_a^b dx f(x), \tag{16.160}$$

再次启发我们，

$$dF = F'(x)\,dx = f(x)dx. \tag{16.161}$$

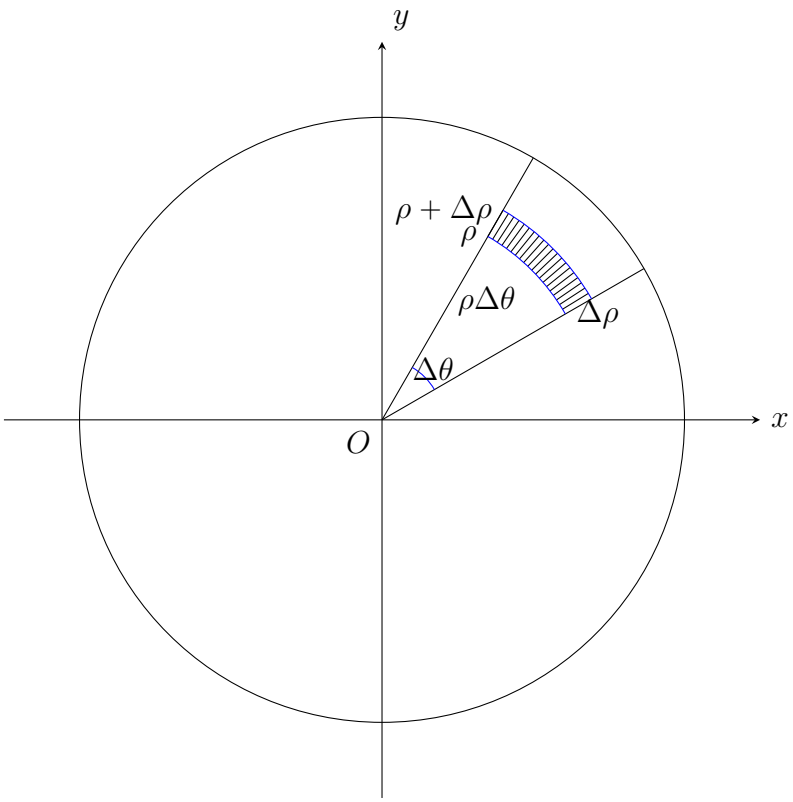


图 16.12: **极坐标中圆的微元的面积**。阴影部分表示了在极坐标中给定 $\Delta\theta$ 和 $\Delta\rho$ 的一小块类似长方形的形状的面积。

未来我们会依赖极限的概念把这些表达式严格化。目前，我们看到，把 dx, dy, dF 看作独立有意义，然后导数看作分数 $\frac{dy}{dx}, \frac{dx}{dy}, \frac{dF}{dx} = f(x)$，把积分看作是 $\int_a^b f(x)dx, \int_{F(a)}^{F(b)} dF$，把导数关系看作是 $dF = f(x)dx$，完全和导数、积分、分数计算不冲突。从现在开始，我们就把 dx, dy, dF 看作独立有意义，并且称为微分，分别叫做 x, y, F 的微分。

终于，我们有了导数、积分（分不定积分和定积分）和微分这些微积分的概念，也有了它们之间的关系。原则上，除了极限这个概念以及它所要解决的问题——避免导数计算三步法中的逻辑问题——之外，我们已经从物理、几何的启发出发构建了整个微积分的核心概念体系。未来我们会在极限这个概念的基础上严格地来构建这个体系。

16.2.3 定积分计算的几个定理

首先，按照不定积分的线性性，以及定积分和不定积分之间的牛顿-莱布尼兹公式，有定积分的线性性。证明留作练习。

习题 16.8 (不定积分的线性性). 证明定积分的线性性，

$$\int_a^b dx \left[\mu f(x) + \nu g(x)\right] = \mu \int_a^b dx f(x) + \nu \int_a^b dx g(x). \tag{16.162}$$

其次根据不定积分的分步积分有定积分的分步积分公式。证明留作练习。

习题 16.9 (分步积分). 证明定积分计算的分步积分公式，

$$\int_a^b dx f(x)g'(x) = f(x)g(x)\Big|_a^b - \int_a^b dx f'(x)g(x). \tag{16.163}$$

对于一个符合函数 $F(x(u))$，我们可以把积分写成几种不同的形式，

$$\int_{F_a}^{F_b} dF = \int_a^b \frac{dF}{du} du, \tag{16.164}$$

$$\int_{F_a}^{F_b} dF = \int_{x(a)}^{x(b)} \frac{dF}{dx} dx, \tag{16.165}$$

其中 $F(x(a)) = F_a, F(x(b)) = F_b$。再根据不定积分的复合函数积分公式或者说符合函数求导的链式法则，

$$\frac{dF}{du} = F'(x(u))_u = F'_x(x(u))\, x'_u(u). \tag{16.166}$$

于是，我们得到

$$\int_{x(a)}^{x(b)} f(x)dx = \int_{F_1}^{F_2} dF = \int_a^b duf\left(x(u)\right)x'_u(u). \tag{16.167}$$

这个变量替换的公式对于计算复杂的积分特别有用：当 $\int dxf\left(x\right)$ 比较复杂的时候，可以引入一个新的变量 $x = x(u)$ 来完成这个积分。

我们用这几个公式来完成之前没有算出来结果的几个积分。我们在一次强调，这些公式你都不需要真的记住，求导运算和积分运算你也不需要真的会多少个，只需要直到常用的几个，例如 $x^q, e^x, \sin x, \cos x$，然后剩下的就用计算机软件来完成就行。

我们先来证明

$$\int_{-1}^{1} \frac{1}{\sqrt{1-\xi^2}}d\xi = \pi. \tag{16.168}$$

令 $\xi = \sin\theta$，则 $-1 = \sin -\frac{\pi}{2}, 1 = \sin\frac{\pi}{2}$，并且在 $\theta \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$ 范围内 θ 和 ξ 一一对应，因此，

$$\int_{-1}^{1} \frac{1}{\sqrt{1-\xi^2}}d\xi = \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \frac{1}{\cos\theta}\cos\theta d\theta = \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} d\theta = \pi. \tag{16.169}$$

我们再来证明

$$\int_{-1}^{1} \sqrt{1-\xi^2}d\xi = \frac{\pi}{2}. \tag{16.170}$$

做上面相同的变换，得到

$$\begin{aligned}\int_{-1}^{1} \sqrt{1-\xi^2}d\xi = \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \cos^2\theta d\theta &= \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \frac{1+\cos 2\theta}{2}d\theta, \\ &= \left.\frac{\theta + \frac{1}{2}\sin 2\theta}{2}\right|_{-\frac{\pi}{2}}^{\frac{\pi}{2}} = \frac{\pi}{2}.\end{aligned} \tag{16.171}$$

这里我们用到了三角函数的导数和三角函数的几个公式：$\sin'\theta = \cos\theta$, $\cos\left(\alpha+\beta\right) = \cos\alpha\cos\beta - \sin\alpha\sin\beta$，尤其是当 $\alpha = \beta$ 的时候，$\cos 2\theta = \cos^2\theta - \sin^2\theta = 2\cos^2\theta - 1$。其中我们用到了 $\cos^2\theta + \sin^2\theta = 1$。我们一会儿还会用到 $\sin\left(\alpha \pm \beta\right) = \sin\alpha\cos\beta \pm \cos\alpha\sin\beta$ 以及从这里导出的和差化积公式 $\sin\alpha - \sin\beta = 2\cos\frac{\alpha+\beta}{2}\sin\frac{\alpha-\beta}{2}$。这些三角函数公式是可以基于三角函数的定义在单位圆内来证明的。但是，这里只有用来计算 $\sin'\theta = \cos\theta$

的作用。你完全可以不管。下面的证明只不过为了说服你，$\sin'\theta=\cos\theta$ 确实是可以基于定义算出来的，而不是需要单独记忆或者不得不接受的事实。

我们终于可以来补充证明 $\sin'\theta=\cos\theta$ 了。

例 16.18 ($\sin'\theta=\cos\theta$)**.** 证明 $\sin'\theta=\cos\theta$。

按照导数计算三步法，我们有第一步计算差的比

$$\frac{\sin(\theta+\Delta\theta)-\sin\theta}{\Delta\theta}=\frac{2\cos\frac{2\theta+\Delta\theta}{2}\sin\frac{\Delta\theta}{2}}{\Delta\theta},\tag{16.172}$$

然后，第二步把分母放到分子的每一项里面去：按照我们对 $\frac{2\theta+\Delta\theta}{2}$ 的认识，当 $\Delta\theta$ 越来越小的时候，$\frac{2\theta+\Delta\theta}{2}\to\theta$。因此，在第三步的时候 $\cos\frac{2\theta+\Delta\theta}{2}\to\cos\theta$。这部分简单。那剩下的就是 $\frac{2\sin\frac{\Delta\theta}{2}}{\Delta\theta}$ 了。我们来看看这一项当 $\Delta\theta$ 越来越小的时候会成为什么。

利用单位圆，如图 16.13，我们发现，以弧度来度量角度，OB 的长度就是 $\Delta\theta$，然后利用图中两个直角三角形 $\triangle ACO$ 和 $\triangle OAB$，有

$$\sin\Delta\theta<\Delta\theta<\tan\Delta\theta=\frac{\sin\Delta\theta}{\cos\Delta\theta}.\tag{16.173}$$

这个等式对于等于 $\Delta\theta$ 都成立。当 $\Delta\theta$ 越来越小的时候，这个等式接近

$$\sin\Delta\theta<\Delta\theta<\tan\Delta\theta=\sin\Delta\theta.\tag{16.174}$$

其中我们用到了当 $\Delta\theta$ 越来越小的时候

$$\cos\Delta\theta\to 1.\tag{16.175}$$

于是，当 $\Delta\theta$ 越来越小的时候

$$1\leqslant\frac{\sin\Delta\theta}{\Delta\theta}\leqslant 1.\tag{16.176}$$

因此，

$$\frac{\sin\Delta\theta}{\Delta\theta}=1.\tag{16.177}$$

现在我们来完成第三步，

$$\begin{aligned}\frac{2\cos\frac{2\theta+\Delta\theta}{2}\sin\frac{\Delta\theta}{2}}{\Delta\theta}&=\frac{\cos\frac{2\theta+\Delta\theta}{2}\sin\frac{\Delta\theta}{2}}{\frac{\Delta\theta}{2}},\\&=\cos\frac{2\theta+\Delta\theta}{2}\cdot\frac{\sin\frac{\Delta\theta}{2}}{\frac{\Delta\theta}{2}},\\&\to\cos\theta.\end{aligned}\tag{16.178}$$

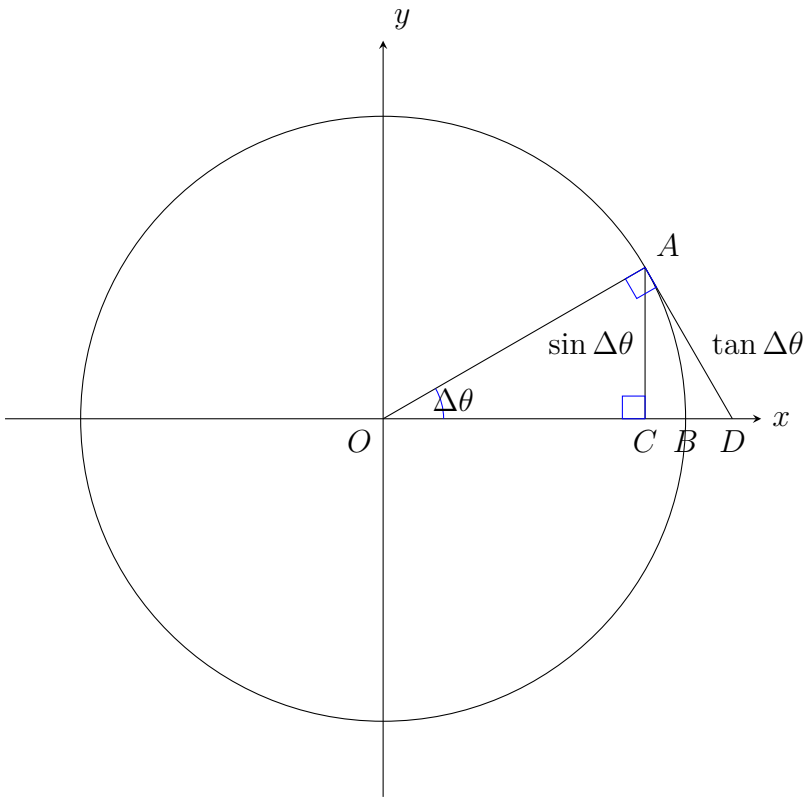


图 16.13: $\frac{\Delta\theta}{\sin\Delta\theta}$ **的极限**。利用单位圆上的圆弧长度和两个直角三角形可以得到 $\frac{\Delta\theta}{\sin\Delta\theta}$ 的在 $\Delta\theta\to 0$ 时候的极限。

因此，

$$\sin'\theta=\cos\theta. \tag{16.179}$$

习题 16.10 ($\cos'\theta=-\sin\theta$)**.** 证明 $\cos'\theta=-\sin\theta$。

16.3 微积分学习总结

以函数和四则运算这些初中数学知识为知识基础，在受到现实生活的经验体验、物理学和几何学上的启发和挑战下，在这一章中我们已经体验到了导数（不定积分）、积分和微分被创造和被创造性使用的过程。具体来说，我们有了导数的概念——曲线在某点的切线——和导数计算三步法（尽管由于没有极限的概念还不严格）、不定积分做为求导数的逆运算、定积分

的概念（曲线下面积、对微元求和）、定积分做为对微分的积分、定积分和不定积分的关系（牛顿-莱布尼兹公式）。

当然，在个别具体计算上，我们用到了三角函数这样的高中知识。但是，那仅仅是为了计算圆的面积这个具体问题的需要。原则上，我们展示了长方形的面积如何由积分得到也就能说明问题了——根据两点之间的距离公式（勾股定理），我们可以由点的积分得到线长，由线长的积分得到面积。

你根本就不需要太多的知识基础就可以学会微积分。并且，稍后我们就会看到，只有学会微积分才能体会到一些人类极具创造性的思想和相应的问题解决。如果你不理解微积分，不理解微积分在这些这些极具创造性的思想和问题解决过程中的作用，那你就没法站在巨人的肩膀上来继续提出和解决极具创造性的问题，发展极具创造性的思想，你甚至都没法欣赏到那些做这样的事情的人的痛苦和快乐。

16.4 积分和导数计算的计算机软件

前面我们解除了一些导数和积分的计算。有一部分是为了理解概念和证明命题。有一部分是遇到具体问题要解决的时候刚好出现的。在数学上，人类的作用是把问题转化成数学语言——具体到这一部分理解微积分的精神（极限的思想、变化率和累积量的思想）、概念和典型使用场景之后用微积分的语言来表述问题，然后，找到或者运用求解这些问题的方法。当然，也包含通过表达和解决现实世界的问题推动微积分这门数学的发展。或者，通过受到微积分本身的知识的系统性的启发或者逼迫来推动这门学科的发展。不过，这两个任务，已经很大程度上被完成了，除非我们考虑弯曲表面的线长、面积、导数、微分和积分等问题。因此，在实数集或者欧氏空间中的微积分的范围内，人类的主要作用基本上就剩下前一个了，也就是用微积分的语言来表述和解决问题。

当具体来达成这个作用的时候，我们还是要做一些关于微积分的具体计算的。这些具体计算，除非计算是为了理解概念和证明命题，都可以通过计算机来完成。因此，在这里，我们介绍几样计算微积分的数学软件。

计算微积分的数学软件分成数值计算和符号计算两类。符号计算就是完成类似本书之中的例题的那些问题。数值计算就是给定一个数值形式的函

数 $f(x)$——相当于给定两个数列一个对应着 x_n 一个对应着相应的 $f(x_n)$，然后计算出来相应的导数和积分。当 x_n 是均匀间隔的时候，也可以只给对应着 $f(x_n)$ 的一个数列。当然，当有些用符号给出的函数的积分不能解析求解出来的时候，人们也用数值方法来求这样的积分。

关于符号计算，一个最简单粗暴的方式就是查表——查能够算出来的基本的原函数-导函数表——和运用变量替换、线性性、乘积函数、分步积分等导数和积分计算的公式。当然，要构成一个计算机软件，则还需要发现一堆最基本的函数的基本公式，以及相应的可以尝试的变换。James Slagle（詹姆士 · 斯莱格）的符号积分系统 [9] 就是这样的一个例子。当然，现代符号计算软件的比Slagle的符号积分系统完善很多。

一个最简单粗暴的从给定数列求导数和积分的方法就是运用基本定义。例如，对于导数，为了简单计，以均匀分割的小格子为例

$$f'(x_n) = \frac{f(x_{n+1}) - f(x_n)}{\Delta x}. \tag{16.180}$$

当然，每准用下面的表达式也可以，

$$f'(x_n) = \frac{f(x_{n+1}) + f(x_{n-1}) - 2f(x_n)}{2\Delta x}. \tag{16.181}$$

数学家就得研究哪一种更准确的问题，以及是否有更加准确的导数的计算方法的问题。已知导函数计算原函数实际上就是在算积分，例如，

$$f(x_{n+1}) = f(x_n) + f'(x_n)\Delta x. \tag{16.182}$$

例如，对于更一般的积分问题，可以直接用定义来这样算

$$\int_a^b f(x)\,dx = \sum_{l=0}^{n-1} f(x_l)\Delta x, \tag{16.183}$$

或者这样，

$$\int_a^b f(x)\,dx = \sum_{l=0}^{n-1} \frac{f(x_l) + f(x_{l+1})}{2}\Delta x. \tag{16.184}$$

当然数学家也得研究哪一种更准确，是否有更加准确的积分的计算方法。在这些问题上，数学家都提出了比这几个基于定义的计算方式准确得多得数值计算方法，例如 Runge-Kutta 方法 [10]。

在符号计算上，我们可以用 Mathematica[1]、Maple[2]和 SageMath[3]这些数学计算软件。除了本地安装和使用这些关键之外，其中 Mathematica 还有一个网页版的近亲 Wolfram Alpha[4]，SageMath 也有合作伙伴提供网页版或者你自己可以运行一个服务器来支持网页版。这样你就可以在网页上直接运行计算代码。

在数值计算上，可以直接使用 BLAS[5]、LAPACK[6]、ATLAS[7]、GLS[8]等开源基础计算软件，也可以使用 Intel 数学核[9]和 AMD 数学核[10]这样的商业版本的专门针对 CPU 优化的软件，也可以使用 Mathematica、Maple、SageMath 和 MATLAB[11]。SageMath 可以同时用于符号计算和数值计算，并且还有一个额外的优势：免费。本书的一部分计算就是由 SageMath 完成的。

除了微积分，这些数学计算软件还可以用于线性代数、概率和统计等部分的符号计算和数值计算。因此，在其他部分，我们不再单独列举这些软件，除非有专门适用于那个任务的专门软件。

再一次强调，在双向数学建模这个数学科学的典型责任的实现的过程中，人类的主要工作是提出问题，把问题表述为数学问题，然后知道求解的思路（甚至这个很多时候也可以不了解），具体求解不是人类智慧应该做的事情，都应该交给计算工具去完成。当然，如果某个问题的解法没有或者还需要优化，那还是人类智慧该负责的部分。

---

[1]https://www.wolfram.com/mathematica/，2025 年 11 月 28 日访问。

[2]https://www.maplesoft.com/，2025 年 11 月 28 日访问。

[3]https://www.sagemath.org/，2025 年 11 月 28 日访问。

[4]https://www.wolframalpha.com/，2025 年 11 月 28 日访问。

[5]https://www.netlib.org/blas/，2025 年 11 月 28 日访问。

[6]https://www.netlib.org/lapack/，2025 年 11 月 28 日访问。

[7]https://math-atlas.sourceforge.net/，2025 年 11 月 28 日访问。

[8]https://www.gnu.org/software/gsl/，2025 年 11 月 28 日访问。

[9]https://www.intel.com/content/www/us/en/developer/tools/oneapi/onemkl.html，2025 年 11 月 28 日访问

[10]https://www.amd.com/en/developer/aocl/libm.html，2025 年 11 月 28 日访问。

[11]https://www.mathworks.com/，2025 年 11 月 28 日访问。

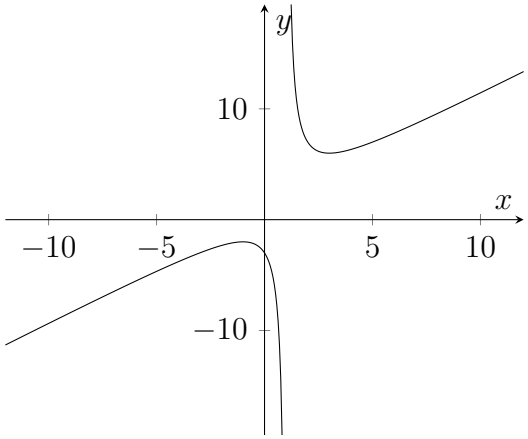


图 16.14: $\frac{x^2+3}{x-1}$ **的图像**。函数 $\frac{x^2+3}{x-1}$ 在 $x=1$ 不连续，在 $x_1=-1$ 和 $x_2=3$ 处有极值。再正负无穷远处分别趋向正负无穷大。通过计算和分析导函数，求原函数的极限，可以得到这个函数的所有的上面的性质。

16.5 微积分的应用

我们先来用导数求解几个很有技巧性的极值问题。

例 16.19 (函数极值问题). 求函数 $f(x)=\frac{x^2+3}{x-1}$ 的极值。

这个题的初等函数解法是这样的：

$$\frac{x^2+3}{x-1}=\frac{x^2-1+4}{x-1}=x+1+\frac{4}{x-1}=(x-1)+\frac{4}{x-1}+2. \tag{16.185}$$

当 $x-1>0$ 的时候，用基本不等式或者说完全平方公式 $a+b\geqslant 2\sqrt{ab}$ $(a,b\geqslant 0)$ 可以得到

$$f(x)=(x-1)+\frac{4}{x-1}+2\geqslant 2\sqrt{(x-1)\times\frac{4}{x-1}}+2=6. \tag{16.186}$$

当 $x-1<0$ 的时候，另 $y=1-x>0$，用基本不等式或者说完全平方公式 $a+b\geqslant 2\sqrt{ab}$ $(a,b\geqslant 0)$ 可以得到

$$f(x=-y)=-y-\frac{4}{y}+2\leqslant -2\sqrt{y\times\frac{4}{y}}+2=-2. \tag{16.187}$$

这里我们需要想到先把 $\frac{x^2+3}{x-1}$ 除法表达式展开，然后再观察到展开的表达式和基本不等式或者说完全平方公式之间的联系，才能解决问题。

下面，我们用求导函数然后让导函数等于零的方式来求极值试试。顺便，我们把整个函数的样子都基本可画出来。

首先，随着 x 绝对值越来越大，x^2 肯定会比 x 的绝对值更大，于是，这个函数到在 $x\to\pm\infty$ 极限肯定是相应的 $\pm\infty$。

其次，我们来看看这个函数的增减性。求 $f(x)$ 的导数，得到

$$\begin{aligned} f'(x) &= 2x(x-1)^{-1} + (-(x-1)^{-2}(x^2+3)) \\ &= \frac{2x(x-1)-(x^2+3)}{(x-1)^2} = \frac{x^2-2x-3}{(x-1)^2} \\ &= \frac{(x+1)(x-3)}{(x-1)^2}. \end{aligned} \tag{16.188}$$

分母大于等于零，分子当 $x<-1$ 的时候大于零，当 $x>3$ 的时候大于零，中间 $-1<x<3$ 的时候小于零。对应到 $f(x)$，那就是当 $x<-1$ 的时候 $f(x)$ 是一个增函数，当 $x>3$ 的时候也是增函数，中间部分是一个减函数。中间还有一个特殊点 $x=1$，函数值趋向无穷大，

$$\lim_{x\to 1^-} = \lim_{x\to 1^-}\frac{x^2+3}{x-1} = -\infty, \tag{16.189}$$

$$\lim_{x\to 1^+} = \lim_{x\to 1^+}\frac{x^2+3}{x-1} = \infty. \tag{16.190}$$

其中，$x\to 1^-$ 和 $x\to 1^+$ 分别表示从左侧和右侧接近 1。于是，我们就得到了 $f(x)$ 的图像，如图 16.14。自然地其中的极值点为 $f'(x)=0$ 的点，也就是

$$0 = f'(x) = \frac{(x+1)(x-3)}{(x-1)^2}. \tag{16.191}$$

也就是

$$x_1 = -1, x_2 = 3. \tag{16.192}$$

这个时候，$f(x)$ 的值为，

$$f(x_1) = -2, f(x_2) = 6. \tag{16.193}$$

我们得到了完全相同的结果，同时更清楚地掌握了整个函数的样子。

和初等数学的方法相比，我们不需要任何巧妙的聪明的技巧，我们只需要把导函数求出来，求解其等于零的点就可以。同时，一旦把导函数求出

来，我们就可以直到其再任何一个点的正负，也就知道了原函数的增减性。具体回到这道题里的函数，则还需要注意其发散点，求出来那个点上的函数的极限。这样我们就完全掌握了这个函数。

我们再来看一个例子。

例 16.20 (函数极值问题). $x+y=4,\ x>0, y>0$，求函数 $f(x,y)=\frac{x^2}{x+1}+\frac{y^2}{y+1}$ 的极值。

这个题的初等函数解法是这样的：

$$\frac{x^2}{x+1}+\frac{y^2}{y+1}=\frac{x+1+y+1}{x+1+y+1}\left(\frac{x^2}{x+1}+\frac{y^2}{y+1}\right) \tag{16.194}$$

$$=\frac{1}{6}\left(x^2+y^2+x^2\frac{(y+1)}{x+1}+y^2\frac{(x+1)}{y+1}\right) \tag{16.195}$$

$$\geqslant\frac{1}{6}\left(x^2+y^2+2\sqrt{x^2\frac{(y+1)}{x+1}\times y^2\frac{(x+1)}{y+1}}\right) \tag{16.196}$$

$$=\frac{1}{6}\left(x^2+y^2+2xy\right) \tag{16.197}$$

$$=\frac{1}{6}(x+y)^2 \tag{16.198}$$

$$=\frac{8}{3}. \tag{16.199}$$

其中第一步和第三步都很神奇，第四步需要保证最后的结果是一个 $(x+y)$ 的表达式，也很神奇。

下面，我们用求导函数然后让导函数等于零的方式来求极值试试，因为有约束，我们用*Lagrangian* 乘子法。

首先，令

$$F(x,y,\lambda)=\frac{x^2}{x+1}+\frac{y^2}{y+1}+\lambda(x+y-4). \tag{16.200}$$

求这个函数的所有偏导数并求解其等于零的方程，

$$0=\frac{\partial F}{\partial x}=\frac{2x(x+1)-x^2}{(x+1)^2}+\lambda, \tag{16.201}$$

$$0=\frac{\partial F}{\partial y}=\frac{2y(y+1)-y^2}{y+1}+\lambda, \tag{16.202}$$

$$0=\frac{\partial F}{\partial \lambda}=x+y-4. \tag{16.203}$$

得到，

$$x=2,y=2,\lambda=-\frac{2}{3}. \tag{16.204}$$

这个时候，$f(x,y)=\frac{8}{3}$。再一次，结果完全相同。

和初等数学的方法相比，我们不需要任何巧妙的聪明的技巧，我们只需要结合约束把问题转化为*Lagrangian* 乘子法能用的优化目标函数，把这个函数的导函数求出来，求解其等于零的点就可以。

以后我们再也不用担心找不找得到一个函数和基本不等式的联系，如何变化这个函数才能看到这样的联系，这种低层次的技巧了。

现在我们真正来求解前面已经数学化了的最短距离问题和最速下降线问题，以及拓展到在两种介质内的光的传播问题的数学化和求解。

例 16.21 (最短距离问题). 已知在平面上的两点 $P=(x_P,y_P)$ 和 $Q=(x_Q,y_Q)$，求过这两点的距离最短的曲线。

第一步，问题的数学化。给定两点 $P=(x_P,y_P)$ 和 $Q=(x_Q,y_Q)$ 的曲线可以被描述为

$$x=x(\tau),y=y(\tau). \tag{16.205}$$

其中，初始点和结束点分别为 P,Q，也就是

$$x(0)=x_P,y(0)=y_P; \tag{16.206}$$

$$x(t)=x_Q,y(t)=y_Q. \tag{16.207}$$

在这个描述下，曲线的长度为

$$\mathscr{L}=\int_0^t d\tau\sqrt{(x'(\tau))^2+(y'(\tau))^2}. \tag{16.208}$$

其中 $x'(\tau)=\frac{dx(\tau)}{d\tau},y'(\tau)=\frac{dy(\tau)}{d\tau}$。

这样我们就有了一个自变量为起点和终点确定函数形式任意的 $(x(\tau),y(\tau))$ 函数，或者说轨迹，或者说曲线，因变量为由公式 (16.208) 给出的这个轨迹的长度的极值问题：问过固定两点的什么样的曲线可以使得曲线长度最小。

第二步，求解数学化以后的问题。我们假设这样的长度最小的曲线存在，并且就是 $(x(\tau),y(\tau))$，那么，任何一条偏离这个曲线的曲线，记为 $(x(\tau)+$

$\Delta x(\tau), y(\tau)+\Delta y(\tau))$，其长度都大于 $(x(\tau), y(\tau))$ 的长度。按照导数和极值的关系——一个函数在其极值点导数为零，我们应该有，

$$\frac{\Delta \mathscr{L}}{\Delta x(\tau)} \to 0, \frac{\Delta \mathscr{L}}{\Delta y(\tau)} \to 0, \tag{16.209}$$

当 $\Delta x(\tau) \to 0, \Delta y(\tau) \to 0$ 的时候。

不过，这里的 $\mathscr{L}$ 是函数的函数——称为泛函，而我们没有学习过泛函的导数的定义和计算。我们来把之前学习过的普通函数的导数推广到泛函的导数。按照之前微积分常用的方式，我们把 $\tau \in [0,t]$ 分成 N 份，$\tau_n = \frac{n}{N}t$，这样我们就有了 $\{x(\tau_n), y(\tau_n)\}$ 个独立自变量。注意，在端点上，我们得保证 $x(0) = x_P, y(0) = y_P; x(t) = x_Q, y(t) = y_Q$。用这些变量我们来重新表述公式 (16.208)。

$$\begin{aligned} \mathscr{L} = & \sum_{n=0}^{N-1} t/N \sqrt{\left(\frac{x_{n+1}-x_n}{t/N}\right)^2 + \left(\frac{y_{n+1}-y_n}{t/N}\right)^2}, \\ = & \sum_{n=0}^{N-1} \sqrt{(x_{n+1}-x_n)^2 + (y_{n+1}-y_n)^2}. \end{aligned} \tag{16.210}$$

对这个函数求导数，并且让所有的导数等于零，得到

$$\begin{aligned} 0 = \mathscr{L}'_{x_n} = & \frac{x_n - x_{n-1}}{\sqrt{(x_n-x_{n-1})^2 + (y_n-y_{n-1})^2}} - \frac{x_{n+1}-x_n}{\sqrt{(x_{n+1}-x_n)^2 + (y_{n+1}-y_n)^2}}, \\ 0 = \mathscr{L}'_{y_n} = & \frac{y_n - y_{n-1}}{\sqrt{(x_n-x_{n-1})^2 + (y_n-y_{n-1})^2}} - \frac{y_{n+1}-y_n}{\sqrt{(x_{n+1}-x_n)^2 + (y_{n+1}-y_n)^2}}. \end{aligned}$$

(16.211a)

(16.211b)

进而得到

$$\frac{y_{n+1}-y_n}{x_{n+1}-x_n} = \frac{y_n - y_{n-1}}{x_n - x_{n-1}} = \cdots = \frac{y_1 - y_0}{x_1 - x_0} = k. \tag{16.212}$$

于是

$$y_n = k(x_n - x_{n-1}) + y_{n-1}. \tag{16.213}$$

其中 $x_0 = x_P, y_0 = y_P$。这是一个迭代公式，可以用来求出来所有的 y_n。例如，

$$y_1 = k\left(x_1 - x_0\right) + y_0, \tag{16.214}$$

$$y_2 = k\left(x_2 - x_1\right) + y_1 = k\left(x_2 - x_1\right) + k\left(x_1 - x_0\right) + y_0 = k\left(x_2 - x_0\right) + y_0, \tag{16.215}$$

直到一般形式

$$y_n = k\left(x_n - x_0\right) + y_0, \tag{16.216}$$

以及最后的 $n = N - 1$，

$$y_N = k\left(x_N - x_0\right) + y_0. \tag{16.217}$$

利用终点的条件 $x_N = x_Q, y_N = y_Q$，我们得到 $k = \frac{y_Q - y_P}{x_Q - x_P}$，以及一般形式

$$y_n = \frac{y_Q - y_P}{x_Q - x_P}\left(x_n - x_P\right) + y_P. \tag{16.218}$$

这刚好就是过 P, Q 的直线方程。

这样我们就证明了过给定两点距离最短的曲线是两点之间的直线段。

不过，这里我们先把 $\mathscr{L}$ 转化成了取和再来求导数 $\frac{d\mathscr{L}}{dx_n} = \mathscr{L}'_{x_n}$。既然我们把 x_n 也就是 $x(\tau_n)$ 当作了自变量，我们是不是可以试试直接求类似于 $\frac{d\mathscr{L}}{dx(\tau)}$ 这样的导数呢？

我们来试试。为了记号上区别对函数的导数和对普通自变量的导数，我们用另一套符号，$\frac{\delta\mathscr{L}}{\delta x(\tau)}$。跟普通函数的导数一样，我们先来用求导三步法。

第一步，引入 $\Delta x(\tau)$ 和 $\Delta y(\tau)$ 来得到 $\Delta\mathscr{L}$

$$\Delta\mathscr{L} = \int_0^t d\tau \sqrt{\left(x'(\tau) + (\Delta x)'(\tau)\right)^2 + \left(y'(\tau) + (\Delta y)'(\tau)\right)^2} - \int_0^t d\tau \sqrt{\left(x'(\tau)\right)^2 + \left(y'(\tau)\right)^2}. \tag{16.219}$$

接着我们把上式展开成 $\Delta x(\tau), \Delta y(\tau)$ 的多项式，为第二步计算 $\frac{\Delta\mathscr{L}}{\Delta x(\tau)}, \frac{\Delta\mathscr{L}}{\Delta y(\tau)}$ 做准备。我们来用 $\sqrt{1+x} = 1 + \frac{1}{2}x + o(x)$（其中 $o(x)$ 表示比 x 更高阶更快接近于零的量，当 x 趋近于零的时候）处理上式中的根式项。注意，为

了第二步，我们只需要保留到 $\Delta x(\tau),\Delta y(\tau)$ 的一阶项，

$$\begin{aligned}
&\sqrt{(x'(\tau))^2+(y'(\tau))^2+2x'(\tau)(\Delta x)'(\tau)+2y'(\tau)(\Delta y)'(\tau)+o(\Delta x)'(\tau)+o(\Delta y)'(\tau)},\\
=&\sqrt{(x'(\tau))^2+(y'(\tau))^2}\sqrt{1+2\frac{x'(\tau)(\Delta x)'(\tau)+y'(\tau)(\Delta y)'(\tau)}{(x'(\tau))^2+(y'(\tau))^2}+o(\Delta x)'(\tau)+o(\Delta y)'(\tau)},\\
=&\sqrt{(x'(\tau))^2+(y'(\tau))^2}\left(1+\frac{1}{2}2\frac{x'(\tau)(\Delta x)'(\tau)+y'(\tau)(\Delta y)'(\tau)}{(x'(\tau))^2+(y'(\tau))^2}+o(\Delta x)'(\tau)+o(\Delta y)'(\tau)\right),\\
=&\sqrt{(x'(\tau))^2+(y'(\tau))^2}+\frac{x'(\tau)(\Delta x)'(\tau)+y'(\tau)(\Delta y)'(\tau)}{\sqrt{(x'(\tau))^2+(y'(\tau))^2}}+o(\Delta x)'(\tau)+o(\Delta y)'(\tau).
\end{aligned}\tag{16.220}$$

其中第一项和后面相减刚好消去。第二步，把分母算到每一项里面去，

$$\frac{\Delta\mathscr{L}}{\Delta x(\tau)}\propto\frac{x'(\tau)}{\sqrt{(x'(\tau))^2+(y'(\tau))^2}}\frac{(\Delta' x)(\tau)}{\Delta x(\tau)}+\frac{o(\Delta' x)(\tau)}{\Delta x(\tau)}.\tag{16.221}$$

但是，我们发现，这里有个问题，本来我们期待出现 $\frac{(\Delta x)(\tau)}{\Delta x(\tau)}$ 这样的项而实际上出现的是 $\frac{(\Delta' x)(\tau)}{\Delta x(\tau)}$ 这样的项。我们不能直接这样去算第三步。幸好，我们还有一个对参数 τ 的积分没有用。我们来用上它。我们以前一项为例，后

一项的处理方式相同。

$$\Delta\mathscr{L} = \int_0^t d\tau \frac{x'(\tau)}{\sqrt{(x'(\tau))^2 + (y'(\tau))^2}} (\Delta' x)(\tau),$$

$$= \int_0^t \frac{x'(\tau)}{\sqrt{(x'(\tau))^2 + (y'(\tau))^2}} d\Delta x(\tau),$$

$$= \frac{x'(\tau)}{\sqrt{(x'(\tau))^2 + (y'(\tau))^2}} \Delta x(\tau)\Big|_0^t - \int_0^t \frac{d}{d\tau}\left[\frac{x'(\tau)}{\sqrt{(x'(\tau))^2 + (y'(\tau))^2}}\right] \Delta x(\tau)\, d\tau,$$

$$= -\int_0^t \frac{d}{d\tau}\left[\frac{x'(\tau)}{\sqrt{(x'(\tau))^2 + (y'(\tau))^2}}\right] \Delta x(\tau)\, d\tau.$$

其中我们使用了分步计分，还用到了在起点和终点 $\Delta x(\tau) = 0$。

$$\frac{\Delta\mathscr{L}}{\Delta x(\tau)} = \frac{-\int_0^t \frac{d}{d\tau}\left[\frac{x'(\tau)}{\sqrt{(x'(\tau))^2 + (y'(\tau))^2}}\right] \Delta x(\tau)\, d\tau}{\Delta x(\tau)}. \tag{16.222}$$

注意，这里我们把每一个时间点的 $\Delta x(\tau)$ 都当作独立自变量。因此，

$$\frac{\Delta\mathscr{L}}{\Delta x(\tau)} \propto \frac{d}{d\tau}\left[\frac{x'(\tau)}{\sqrt{(x'(\tau))^2 + (y'(\tau))^2}}\right]. \tag{16.223}$$

下面，我们终于可以来计算第三步了。第三步，让 $\Delta x(\tau) \to 0$ 来得到

$$\frac{\delta\mathscr{L}}{\delta x(\tau)} \propto \frac{d}{d\tau}\left[\frac{x'(\tau)}{\sqrt{(x'(\tau))^2 + (y'(\tau))^2}}\right]. \tag{16.224}$$

同理，

$$\frac{\delta\mathscr{L}}{\delta y(\tau)} \propto \frac{d}{d\tau}\left[\frac{y'(\tau)}{\sqrt{(x'(\tau))^2 + (y'(\tau))^2}}\right]. \tag{16.225}$$

取极值意味着，我们要求这些导数都等于零，也就是

$$\frac{d}{d\tau}\left[\frac{x'(\tau)}{\sqrt{(x'(\tau))^2 + (y'(\tau))^2}}\right] = 0, \tag{16.226a}$$

$$\frac{d}{d\tau}\left[\frac{y'(\tau)}{\sqrt{(x'(\tau))^2 + (y'(\tau))^2}}\right] = 0. \tag{16.226b}$$

也就是，求导内部的函数是一个不依赖于 τ 的常数，

$$\frac{x'(\tau)}{\sqrt{(x'(\tau))^2+(y'(\tau))^2}}=\alpha, \tag{16.227a}$$

$$\frac{y'(\tau)}{\sqrt{(x'(\tau))^2+(y'(\tau))^2}}=\beta. \tag{16.227b}$$

于是，把两式相除我们得到

$$\frac{y'(\tau)}{x'(\tau)}=\frac{\beta}{\alpha}=k. \tag{16.228}$$

也就是

$$y=kx+b. \tag{16.229}$$

这正好就是一条直线。接着我们要求这条直线过 P,Q 就可以确定 k,b 的值，从而得到公式 (16.218)。

在这个例子中，我们面对的极值问题是一个对函数积分的形式的泛函——给定一个函数当自变量映射成一个数为因变量——的问题。我们先用化积分为取和的方式解决了这个问题，再把导数计算三步法推广到了泛函再一次解决了这个问题。这个例子的主要目的是为了说明微积分的威力：从距离公式得到线长的微积分形式的表达式，然后运用微积分求出来这个表达式的极值，从而验证两点之间直线短最短。

在解决这个问题的过程中，我们注意到，对于一个更加一般的对函数积分的形式的泛函的极值问题，可能可以使用类似的方法来求解。我们把一般的对函数积分的形式的泛函表示为，

$$\mathscr{L}=\int_{t_1}^{t_2}d\tau L\left(\vec{x}(\tau),\vec{x}'(\tau),\tau\right). \tag{16.230}$$

我们想找出来使得这个 $\mathscr{L}$ 取极值的函数 $x(t)$ 满足

$$\vec{x}(t_1)=\vec{x}_1,\vec{x}(t_2)=\vec{x}_2. \tag{16.231}$$

我们来完成这个一般的泛函极值问题的求解。我们的思路还是相同的：我们需要 $\frac{\delta\mathscr{L}}{\delta x(\tau)}$，而我们计算这个泛函导数的方式是求导三步法，或者说简化了的求导三步法。

做差 计算 $\Delta f = f(x+\Delta x) - f(x)$，可以仅仅保留到 Δx 的一阶项；

相除 计算 $\frac{\Delta f}{\Delta x}$，如果上一步仅仅保留了一阶项则得到一阶项的系数；

求“极限” 让 $\Delta x \to 0$，得到结果就等于一阶项的系数。

这里我们把这个简化了的求导三步法用于 $\mathscr{L}$ 的极值问题。下面为了记号简单，我们把 $\vec{x}(\tau)$ 看作 $x(\tau)$。得到的结果可以直接推广到多维矢量的情况。第一步，做差：

$$\begin{aligned}
&\mathscr{L}\left(x(\tau)+\Delta x(\tau)\right)\\
=&\int_{t_1}^{t_2} d\tau L\left(x(\tau)+\Delta x(\tau), x'(\tau)+(\Delta x)'(\tau), \tau\right)\\
&-\int_{t_1}^{t_2} d\tau L\left(\vec{x}(\tau), \vec{x}'(\tau), \tau\right).
\end{aligned} \tag{16.232}$$

用Taylor 公式保留到 $\Delta x(\tau)$ 的一阶项得到

$$\begin{aligned}
&\int_{t_1}^{t_2} d\tau\left[L'_{x(\tau)}\Delta x(\tau)+L'_{x'(\tau)}(\Delta x)'(\tau)\right]\\
=&\int_{t_1}^{t_2} d\tau L'_{x(\tau)}\Delta x(\tau)+\int_{t_1}^{t_2} d\tau L'_{x'(\tau)}(\Delta x)'(\tau)\\
=&\int_{t_1}^{t_2} d\tau L'_{x(\tau)}\Delta x(\tau)+\int_{t_1}^{t_2} d\left[\Delta x(\tau)\right]L'_{x'(\tau)}\\
=&\int_{t_1}^{t_2} d\tau L'_{x(\tau)}\Delta x(\tau)+\Delta x(\tau)L'_{x'(\tau)}\Big|_{t_1}^{t_2}-\int_{t_1}^{t_2} d\tau\Delta x(\tau)\frac{d}{d\tau}L'_{x'(\tau)}\\
=&\int_{t_1}^{t_2} d\tau\left[L'_{x(\tau)}-\frac{d}{d\tau}L'_{x'(\tau)}\right]\Delta x(\tau).
\end{aligned} \tag{16.233}$$

再来算第二步相除和第三步“求极限”（由于我们仅仅保留了一阶项，第三步可以直接跳过），也就是得到一阶项的系数为零，

$$\frac{\delta\mathscr{L}}{\delta x(\tau)}=L'_{x(\tau)}-\frac{d}{d\tau}L'_{x'(\tau)}=0. \tag{16.234}$$

这个方程有一个自己的名字——Euler-Lagrangian 方程。用偏导数的记号，变量 τ 该回到 t，它可以写成，

$$\frac{d}{dt}\left(\frac{\partial L}{\partial x'(t)}\right)=\frac{\partial L}{\partial x(t)}. \tag{16.235}$$

也可以这样来理解为什么 $x(t)$ 一定要满足一阶项的系数为零，否则，我们永远可以选择

$$\Delta x(\tau) = \pm\left[L'_{x(\tau)} - \frac{d}{d\tau}L'_{x'(\tau)}\right], \tag{16.236}$$

来控制 $\Delta\mathscr{L}$ 的符号，使得出现比极小值还小，比极大值还大的情形，也就是不满足一阶项的系数为零的 $x(t)$ 就不可能是 $\mathscr{L}$ 的极值点。

从生活经验和体验的启发、物理的启发和几何的启发开始，基于函数和四则运算的初中数学知识，我们提炼了出来不严格的导数和积分的概念，导数的不严格的求法，到我们用导数和积分来解决了泛函的极值问题，顺便证明了两点之间直线段最短。这是非常了不起的一次成长。

现在，我们还发现，对于一般的表现为对函数的积分的泛函极值问题，我们只需要求解Euler-Lagrangian 方程公式 (16.234) 就可以。是不是有很多问题都可以被表达成对函数的积分的泛函极值问题？如果可以，那未来很多问题的解决都可以用同样的方法。

下面我们来用泛函极值问题的处理方法求解一个历史上非常有名的叫做“最速下降线问题（Brachistochrone Problem）”的问题。

小知识 16.1 (最速下降线问题的故事). 历史上,最速下降线问题是由*Johann Bernoulli*（约翰 · 伯努利）提出的对同时代数学家的挑战，给大家六个月的时间来给出这个问题的答案。*1696* 年六月，*Johann Bernoulli*在 *Acta Eruditorum* 期刊上发布了如下的问题[12]：

> 在竖直平面上有两个点 A, B。一个运动物体 M 需要从 A 点以初始速度为零出发到达 B。找出来用时最短的路径。忽略摩擦力。

既然这是一个挑战，那*Johann Bernoulli*自己肯定是已经完成的。他估计最厉害的数学家也需要六个月来解开这个问题。这个问题在六个月内没有得到其他人的回答。*Leibniz*也做了尝试。

*Newton*在 *1697* 年 *1* 月 *29* 日从信箱里看到这个问题之后，花了一个晚上解决了这个问题，然后匿名把答案寄给*Johann Bernoulli*。*Johann*

[12]来自于这里，https://www.cantorsparadise.org/the-famous-problem-of-the-brachistochrone-8b955d24bdf7/，2025 年 11 月 28 日访问。

*Bernoulli*说，我一看就知道是谁，就好像“从一只狮子的爪子就能看出来这只狮子”一样。

这个问题以及从这个问题发展起来的泛函求极值的方法影响了数学界（变分法、泛函分析）、物理学界（物理规律的表达方式之一——最小作用量原理）和很多工程领域（控制论）。后面大量的问题被表述为优化问题，然后用泛函求极值的方法来解决。这是人类智慧的伟大财富之一[13]。

例 16.22 (最速下降线问题). 已知在平面上的两点 $P=(x_P,y_P)$ 和 $Q=(x_Q,y_Q<y_P)$，求如果从 P 点以初始速度为零（$\vec{v}=0$ 表示 $v_x=0=v_y$）释放一个小球沿着曲线到 Q 点的用时最短的那条曲线。其中 Q 点在 P 点的下方（$y_Q<y_P$）。

我们选择用 $(x(\tau),y(\tau))$ 来描述一条任意的曲线，满足约束

$$x(0)=x_P=0,y(0)=y_P=0; \tag{16.237}$$

$$x(t)=x_Q,y(t)=y_Q. \tag{16.238}$$

为了下面记号简单，我们让 P 点定在原点 O。

在曲线的每一个点上，我们可以求出来其瞬时速度 $\vec{v}(\tau)=(x'(\tau),y'(\tau))$。这个瞬时速度的方向肯定是曲线的切线方向。其大小按照能量守恒定理为

$$v(\tau)=\sqrt{2g(-y(\tau))}. \tag{16.239}$$

根据 $ds=vdt$，我们有

$$dt=\frac{ds}{v(\tau)}=\frac{\sqrt{(x'(\tau))^2+(y'(\tau))^2}}{\sqrt{2g(-y(\tau))}}. \tag{16.240}$$

于是我们的目标函数——总时间最短就成了，

$$t=\int d\tau\frac{\sqrt{(x'(\tau))^2+(y'(\tau))^2}}{\sqrt{2g(-y(\tau))}}. \tag{16.241}$$

现在问题数学化已经完成，我们有了自变量、目标函数和约束。下面，我们来求解这个问题。

[13] 见https://engineerine.com/newton-vs-bernoulli-the-legendary-brachistochrone-challenge/，2025 年 11 月 28 日访问。

首先，识别出来这个目标函数对应的 $L(x,y,x',y')$，

$$L(x,y,x',y')=\left[(x')^2+(y')^2\right]^{\frac{1}{2}}(2g(-y))^{-\frac{1}{2}}, \tag{16.242}$$

再根据*Euler-Lagrangian* 方程，我们也可以得到 $x=x(t),y=y(t)$ 要满足的方程，

$$\frac{d}{dt}\frac{\partial L}{\partial x'}=\frac{\partial L}{\partial x}, \tag{16.243a}$$

$$\frac{d}{dt}\frac{\partial L}{\partial y'}=\frac{\partial L}{\partial y}. \tag{16.243b}$$

成为

$$\frac{d}{dt}\left\{x'\left[(x')^2+(y')^2\right]^{-\frac{1}{2}}\left[2g(-y)\right]^{-\frac{1}{2}}\right\}=0,$$

$$\frac{d}{dt}\left\{y'\left[(x')^2+(y')^2\right]^{-\frac{1}{2}}\left[2g(-y)\right]^{-\frac{1}{2}}\right\}=g\left[(x')^2+(y')^2\right]^{\frac{1}{2}}\left[2g(-y)\right]^{-\frac{3}{2}}. \tag{16.244a}$$

(16.244b)

这是一组微分方程——其中出现导数项——的方程。这里我们将不再讨论微分方程的求解。不过可以验证，以下的函数满足这个方程，

$$x=r(t-\sin t), \tag{16.245a}$$

$$y=r(\cos t-1). \tag{16.245b}$$

人们发现这是一条旋轮线——在直线上无滑滚动（也就是旋转）一个轮子的时候轮子上固定的一点的运动轨迹。

下面我们来用泛函极值问题的处理方法证明光的折射定律。

例 16.23 (反射率和折射律). 如图 16.15 所示，光经过两种物质的表面的时候，入射角等于反射角 $\theta_r=\theta_1$，折射角和反射角之间存在下面的关系 $n_1\sin\theta_1=n_2\sin\theta_2$。

光子走的路径满足时间最短，也就是把之前的距离最短的公式 (16.208) 改称如下的形式：

$$\mathscr{L}=\int_0^t d\tau\frac{1}{v(x(\tau),y(\tau))}\sqrt{(x'(\tau))^2+(y'(\tau))^2}. \tag{16.246}$$

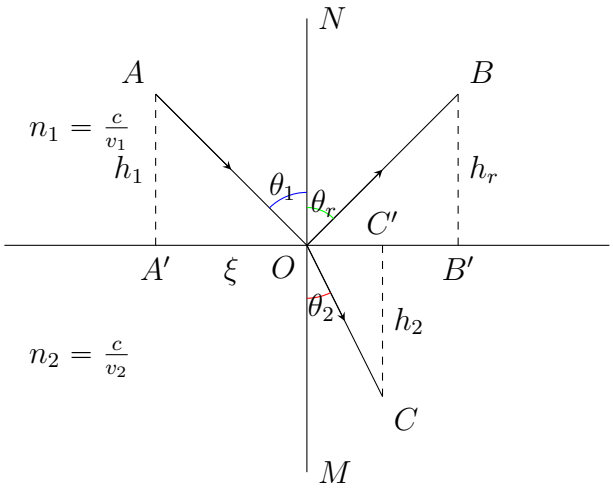


图 16.15: **光的折射和反射**。反射角等于入射角 $\theta_r = \theta_1$，折射角和入射角之间存在下面的关系 $n_1 \sin\theta_1 = n_2 \sin\theta_2$。

其中，我们用到了光子的速度依赖于介质，也就是依赖于光子所处的点 (x, y)。注意，这里的 τ 和 t 是用来给曲线 $(x(\tau), y(\tau))$ 参数化的参数，不是物理时间。其实为了避免这个误解我们应该使用另一套符号。不过，为了看到这个问题和之前的最短路径问题的联系，我们还是用同样的符号。

我们来用之前的结果简化现在的求解。如果 $v(x(\tau), y(\tau))$ 是一个常数，也就是光子在所有的路程中都在同一种介质里面，则问题直接就回到了之前的最短路径问题。于是，我们得到的答案就是光子必然走直线。

现在光子的路径有可能会在两种介质中，那就相当于我们要在界面上选一个点 O，对于反射光把已知的入射起点 A 和反射终点 B 和 O 连接起来，对于折射光把已知的入射起点 A 和折射终点 C 和 O 连接起来。因此，这里唯一的变量就成了 O 点的位置——$\xi = A'O$。

给定的约束是 A, B, C 的位置，也就是 $AA' = h_1, BB' = h_r, CC' = h_2$ 都给定，$A'B' = l_r, A'C' = l_2$ 也给定。光子所用的总时间如下。对于反射部分，

$$t_r = \frac{AO}{v_1} + \frac{OB}{v_1} = \frac{\sqrt{h_1^2 + \xi^2} + \sqrt{h_r^2 + (l_r - \xi)^2}}{v_1}. \tag{16.247}$$

对于折射部分，

$$t_2 = \frac{AO}{v_1} + \frac{OC}{v_2} = \frac{\sqrt{h_1^2 + \xi^2}}{v_1} + \frac{\sqrt{h_2^2 + (l_2 - \xi)^2}}{v_2}. \tag{16.248}$$

把这些需要求出极小值的目标函数乘以代表一个光速的常数 c，然后用 $n_1 = \frac{c}{v_1}, n_2 = \frac{c}{v_2}$ 替换其中的 v_1, v_2，就得到需要求极小值的目标函数为：对于反射部分，

$$t_r = n_1 \left(\sqrt{h_1^2 + \xi^2} + \sqrt{h_r^2 + (l_r - \xi)^2} \right). \tag{16.249}$$

对于折射部分，

$$t_2 = n_1 \sqrt{h_1^2 + \xi^2} + n_2 \sqrt{h_2^2 + (l_2 - \xi)^2}. \tag{16.250}$$

我们也可以写成带约束的极值问题：对于反射部分，

$$t_r = n_1 \left(\sqrt{h_1^2 + \xi^2} + \sqrt{h_r^2 + \eta^2} \right), \tag{16.251}$$

$$\xi + \eta = l_r. \tag{16.252}$$

对于折射部分，

$$t_2 = n_1 \sqrt{h_1^2 + \xi^2} + n_2 \sqrt{h_2^2 + \gamma^2}, \tag{16.253}$$

$$\xi + \gamma = l_2. \tag{16.254}$$

下面我们来求解这四个极值问题。我们用约束下的极值问题来当例子。对反射和折射分别构建带约束的目标函数

$$F_r(\xi, \eta, \lambda) = n_1 \left(\sqrt{h_1^2 + \xi^2} + \sqrt{h_r^2 + \eta^2} \right) + \lambda (\xi + \eta - l_r), \tag{16.255}$$

和

$$F_2(\xi, \gamma, \lambda) = n_1 \sqrt{h_1^2 + \xi^2} + n_1 \sqrt{h_2^2 + \gamma^2} + \lambda (\xi + \gamma - l_2), \tag{16.256}$$

对相应的自变量求导并让导数等于零，有

$$0 = \frac{\partial F_2}{\partial \xi} = n_1 \frac{\xi}{\sqrt{h_1^2 + \xi^2}} + \lambda, \tag{16.257a}$$

$$0 = \frac{\partial F_2}{\partial \eta} = n_1 \frac{\eta}{\sqrt{h_r^2 + \eta^2}} + \lambda, \tag{16.257b}$$

$$0 = \xi + \eta - l_r. \tag{16.257c}$$

和

$$0 = \frac{\partial F_2}{\partial \xi} = n_1 \frac{\xi}{\sqrt{h_1^2 + \xi^2}} + \lambda, \tag{16.258a}$$

$$0 = \frac{\partial F_2}{\partial \gamma} = n_2 \frac{\gamma}{\sqrt{h_2^2 + \gamma^2}} + \lambda, \tag{16.258b}$$

$$0 = \xi + \gamma - l_2. \tag{16.258c}$$

从这里我们可以求解出来所有的变量 ξ, η, γ。

不过，在引入发现 MON——经过入射点垂直于界面的线和入射角 θ_1、反射角 θ_r、折射角 θ_2 这些量之后，我们注意到

$$\sin\theta_1 = \frac{\xi}{\sqrt{h_1^2 + \xi^2}}, \sin\theta_r = \frac{\eta}{\sqrt{h_r^2 + \eta^2}}, \sin\theta_2 = \frac{\gamma}{\sqrt{h_2^2 + \gamma^2}}, \tag{16.259}$$

就可以用这些正弦函数值来简化得到的结论。对于反射，有

$$n_1 \frac{\xi}{\sqrt{h_1^2 + \xi^2}} = -\lambda = n_1 \frac{\eta}{\sqrt{h_r^2 + \eta^2}}, \tag{16.260}$$

也就是

$$\sin\theta_1 = \sin\theta_r, \tag{16.261}$$

在 $0^\circ - 90^\circ$ 的范围内（这个夹角是入射和出射光线和法线 MN 的夹角，自然在这个范围内），自然有

$$\theta_r = \theta_1. \tag{16.262}$$

对于折射，有

$$n_1 \frac{\xi}{\sqrt{h_1^2 + \xi^2}} = -\lambda = n_2 \frac{\gamma}{\sqrt{h_2^2 + \gamma^2}}, \tag{16.263}$$

也就是

$$n_1 \sin\theta_1 = n_2 \sin\theta_2. \tag{16.264}$$

至此，基于光子选择用时最短的路径的假设，你用微积分完全解决了光线的同种介质内传播、界面上的反射和折射问题。

当然，物理学是实验科学。实际上的逻辑是，物理学家已经观察到了上面的光线的同种介质内传播、界面上的反射和折射的规律，然后，上面所做的事情是物理学家们从另一个角度来给这样的规律一个解释。这样做的好处是，每准未来这个从“用时最短、路程最短”来考虑的问题的角度，可以

推广到其他更多的物理现象，甚至其他学科的现象中去。实际上，后来我们发现，确实如此：把问题转化为优化问题能够帮助很多个学科来表述和解决问题。

而把问题表述为优化问题和解决优化问题，微积分都可以发挥巨大的作用。你已经掌握了打开这个世界的秘密的一个非常管用的钥匙。

顺便，在这里我们还看到泛函极值问题和函数极值问题之间的自然的转化：本来这个问题是一个泛函极值问题，但是，在两种介质的情形下，问题退化为一个函数极值问题。因此，泛函极值问题并不比函数极值问题抽象太多。当然，当讨论光在一个处处都不均匀的介质（$n(x,y)$ 处处都不相同）中的传播的时候，那我们就要回到原始的泛函极值问题了。

到现在，你不仅掌握了导数、积分、微分的概念，还掌握了导数、积分的计算，还用导数和积分解决了一些典型问题，包含：无约束极值问题，有约束极值问题，任意函数的多项式近似问题，泛函极值问题，以及所有可以转化成这些问题的问题。除了极限的概念你还没有掌握所以知识体系上还有漏洞之外，你已经以函数和四则运算这些初中数学知识为基础相当完整地学会了微积分的学科概念知识、计算和应用。这是非常了不起的。

不过更加了不起的，是你在这些学科概念知识的学习的过程中，看到了启发这些学科概念知识的提出的点，包含：生活经验和体验、物理学知识和之前的数学知识，还看到了数学知识的系统性——从极少数的数学概念和假设建立起来整个数学知识的体系，如果有漏洞就补上。

数学不是纯粹逻辑的游戏，如果只有逻辑，没有现实的启发，则人们很难会去发展新的数学，新的来自于现实世界的需求是启发数学发展的动力。数学也不是说主要就受现实的启发而不需要逻辑，没有逻辑我们不能从已知的定义和假设来证明命题，不能建构起来数学学科知识之间的联系，看不到数学知识体系内部的漏洞所以来进一步发展数学。数学和科学一样，都是同时需要现实的启发和逻辑的。只不过，数学更加自由：科学（的概念和命题、模型的计算结果）要接受现实的检验，但是，数学（的概念和命题、计算结果）不需要接受现实的检验。只要定义和假设已经被提出来，不管受现实还是受数学本身的启发，建立在它们的基础之上的其他的数学定义和命题，只要建立的过程是逻辑上严密的，就不再需要现实世界来检验。因此，数学可以研究现实世界中完全不存在的对象。当然，神奇的是，往往

这样的本来仅仅从人类想象力中发展起来的数学会在其发展起来之后被科学家发现可以用于描述现实世界。例如，矩阵、非欧几何。这是另一个有意思的话题。

16.6 从微积分看双向数学建模

在我们严格地来学习微积分之前，基于现在已经学习过的微积分的学科概念知识和相应的得到的现实世界、物理学的启发，微积分的应用，我们来体会一项非常非常重要的数学思维——双向数学建模，也就是从现实世界或者其他学科中受到启发提炼出来数学概念和命题，以及反过来，把已经建立的数学概念和命题用于描述现实世界，提出和解决现实世界或者其他学科的问题。

首先，我们看到，导数的概念受到了瞬时速度这个物理概念的启发，受到了位置函数 $x(t)$ 到速度函数 $v(t)$ 之间的关系的启发，受到了曲线切线的启发。在导数概念形成之后，进一步解决了瞬时速度的定义的问题、加速度的定义的问题，以及从速度 $v(t)$ 到加速度 $a(t)$ 的关系的问题，求曲线上任意点切线的问题。

其次，我们看到，定积分的概念受到了从速度函数 $v(t)$ 到位置函数 $x(t)$ 之间的关系的启发，受到了求曲线下面积的启发，受到了求任意封闭平面图形面积的需求的启发。在积分的概念形成之后，进一步解决了如何从速度函数 $v(t)$ 得到位置函数 $x(t)$，从加速度 $a(t)$ 得到速度 $v(t)$，可以求出来任意曲线下的面积，平面上任意封闭图形的面积。

接着，我们看到导数还受到求函数极值和过程或者曲线极值需求的启发。并且，有了导数的概念之后，我们发现，极值确实对应着导数为零，于是也就一般性地解决了极值问题。当我们从没有约束的极值问题过渡到有约束的极值问题的时候，我们发现，很自然地需要考虑偏导数，也就是对于有多个自变量的函数来说仅仅考虑其中某一个自变量的变化带来的导数。有了偏导数的概念之后我们就可以处理具有多个变量的函数的极值问题，以及有约束的极值问题了。更进一步，把偏导数拓展到关于过程或者曲线的极值问题之后，我们得到了泛函的变分导数的概念，并且有了这个概念之后也统一地解决了积分型的泛函的极值问题。

基于泛函极值问题，从应用数学结构来描述世界的角度，我们用它来理解了光的传播行为，求解了最速下降线。反过来，从解决问题促进数学发展的角度，正是由于有了最速下降线问题等这样的问题的推动，才使得数学概念和定理的这个知识体系得到了更大和更具有方向性的发展。

你看偏导数的概念是自然地从把导数的概念用于解决问题中而产生的，泛函导数的概念也是从把导数的概念用于解决问题中而产生的。

然后，在我们把目光放到导数的另一个典型使用场景——用多项式近似任意函数的时候，我们自然地发现，需要把一阶导数推广到高阶导数。

最后，当我们把积分用于计算任意而为封闭图形的面积的时候，以及计算定积分的时候，我们还发现需要提出一个叫做“微元法”的帮助把直觉转化为积分表达式的方法。例如，在二维平面直角坐标系下，每一个格子的范围可以记作 $[x, x+\Delta x]\otimes[y, y+\Delta x]$，因此面积可以认为是 $\Delta x\Delta y$，最后变成积分的时候成为 $dxdy$。类似地，在二维极坐标下，积分表达式里面的小块面积——称为面积元——为 $\rho d\theta d\rho$。这微元法也是在把数学概念——这里是积分——用于解决问题的过程中被创造出来的，并且创造出来之后大大方便了人们写下来所要解决的问题的积分表达式。

这里所有的例子都展示了双向数学建模是什么，以及回答了为什么双向数学建模很重要：技能推动数学本身的发展，还能促进用数学来提出、表达和解决问题。

在这些所有的关于导数和积分的例子中，其实有一个共性：它们都是关于变化率和累积量的问题。因此，微积分其实就是关于变化率和累积量的数学。例如，导数的有限尺度版本 $\frac{\Delta f}{\Delta x}, \frac{\Delta x}{\Delta t}$ 要么关注的是空间变量 x 的小小变化带来的函数值 f 的变化，要么关注的是时间便量 t 的小小的变化带来的函数值 x 的变化。以后者及其在物理学中的作用为例，我们发现 $\frac{\Delta x}{\Delta t}$ 对应着速度 v，而 $\frac{\Delta v}{\Delta t}$ 对应着加速度，而加速度联系着力，其通常是 x 的函数，也就是 $F(x)=ma$。于是，我们就得到了一个方程，

$$m\frac{d}{dt}\left(\frac{dx}{dt}\right)=F(x). \tag{16.265}$$

或者说，反过来看：在给定的时间点 t，我们已知当时的速度 $v(t)$ 和位置 $x(t)$，那么我们就知道了当时的 $F(x)$，也就知道了当时的 $a(t)=\frac{F}{m}$。于是，

下一个时刻的速度也就能够得到了，

$$v(t+\Delta t)=v(t)+a(t)\Delta t. \tag{16.266}$$

并且这个 Δt 可以原则上要多小就取多小，直到精度满足要求。同样的道理，知道了下一个时刻的速度也就知道了下一个时刻的位置，

$$x(t+\Delta t)=x(t)+v(t)\Delta t. \tag{16.267}$$

实际上，这个方法称为求解微分方程的欧拉方法。当然，实际需要通过计算来解决问题的时候，可以直接求解上面的方程公式 (16.265)，或者用比这更准确更快收敛的方法。我们这里仅仅想展示一个道理：把一个问题转化为时间上的累积过程，就可用微积分。把一个问题转化成一个变量在时间上的变化率也是如此，就可以用微积分。空间上的累计过程和变化率也一样。

例如，当我们关注一根绳子的运动的时候，我们可以把绳子分成很多个小块，每一块受到周围的其他绳子小块的作用力就可以得到振动方程——一个描写振动的微分方程。当我们把一个带点物体形成的电场看作是这个带电物体的每一小块形成的电场的和，如果每一小块形成的电场是已知的，我们就可以用积分来计算整个带电物体形成的电场。

这是为什么微积分在如此众多的问题中被使用的原因，通过微元法，或者把一个量在时间或者空间上做累积，或者求变化率，我们就可以把这些问题转化成微积分的问题。

把这样的一个理念——微积分就是关于变化率和累积量的数学——放在心里，你就可以在所有的关于变化率和累积量的问题中自然地想起来用微积分，甚至把一个静态的问题转化为一个变化率和累积量的问题去解决。

16.7 微积分体现的数学学科大图景

除了下面在严格形式得微积分中体现的数学知识的系统性，以及逻辑演绎证明在构建严密的具有数学知识系统中的作用，通过前面的学习我们已经看到了数学的学科大图景。比如说，通过学习微积分，我们看到的数学的学科典型责任是什么？我们看到数学是受现实（这里有描述运动，尤其是从位置函数到速度函数，以及反过来）和已有的数学知识（这里有求极值，

求曲线下面积，求曲线切线）启发的，是用来帮助我们进行更加深入和准确的思考的（这里有Zeno “阿基利斯追不上乌龟”佯谬和Zeno “飞矢不动”佯谬，有对无穷小的纠结一会儿不等于零一会儿等于零，有对光子传播路径问题的重新表述）。

比如说，通过学习微积分，我们看到的数学的典型思维方式是什么？首先，我们看到了做抽象（导数的概念来自于对从位置函数到速度函数的抽象，来自于对曲线切线的抽象，积分的概念来自于对从速度函数到位置函数的抽象，来自于对曲线下面积和曲面面积的计算的抽象），看到了逻辑演绎证明（一个对象的定义是什么，如何操作，这个对象有哪些和其他对象的关系也就是命题，如果其中有逻辑问题就要看到问题然后想办法补全），看到了对数学知识的系统性的追求，还看到了极限的思想尽管不严格，看到了从累积量和变化率的角度来看函数的思想或者说从一个函数在某个点附近的行为来看这个点的行为[14]。

比如说，通过学习微积分，我们看到的数学的典型分析方式是什么？首先，我们看到了求函数和泛函极值的一般性的分析方法——求函数或者泛函的导数以及对于有约束的问题拓展到Lagrangian 乘子法。我们还看到了微元法可以用于把一个问题转化为积分和微分的问题。同时，极限和夹逼不仅仅是思想，也变成了可操作的方法，尽管严格的极限的操作还需要下一节再来体会。我们还看到了用熟悉的函数尤其是级数去逼近任意函数（Taylor公式）。当然，还有一些帮助我们计算的定理也提供了一些分析方法，例如利用分步积分、变量替换、复合函数求导来解决复杂的积分和求导的问题。

比如说，通过学习微积分，我们看到的数学的典型研究对象和典型研究问题是什么？我们发现，只要是能够帮助我们开展更加深入的思考的，同时又可以跨越对象跨越学科领域的，都是数学的研究对象和研究问题。数学不像物理学和化学、生物学等学科，是按照研究对象的种类来划分和决定是不是这个学科的研究对象的。这个和数学的学科典型责任也是有关系的。在这个意义上，也经常有人说数学是关于思维的学科。数学的研究对象其实是对象之间的关系或者说模式。例如，我们发现速度和位置之间的关系、加速度和速度之间的关系、曲线和切线之间的关系、曲线和曲线下面积的关系，未来还会发现势函数和力的关系、点电荷电场和带电体电场的关

[14]未来在学习函数极限和函数的连续性的时候，会把这一点体会的更好

系本质上都是微积分。而数学的典型研究问题就是这些从关系和模式中抽象出来的数学结构的定义、定义的基础以及所有的数学结构之间的关系也就是定理是什么，以及一些方便对这些数学结构做操作的辅助计算和操作的定理是什么。

未来在学习其他数学知识的时候，我们会再一次回到数学学科大图景。希望你体会到数学学科大图景之后，第一对这部分知识可以理解得更加深刻和到位，第二可以把这部分知识用得更好，第三可以迁移到其他数学知识得学习中去，第四帮助自己创造出来新的数学知识或者创造性地使用现有的数学知识。最后，第五，你还能够把这样的从学习一个学科的具体的学科概念和学科思维的知识的过程中体会到一个学科的大图景，然后再用这个学科的大图景来辅助进一步学习和创造的过程——称为人类知识高速公路上以培养掌握高层知识的创造者为目标的理解型学习，简称理解型学习——用到其他学科的知识的学习和创造、创造性使用中去。

16.8 严格形式的微积分

下面我们来补上微积分在概念上的逻辑漏洞——一会把 Δx 当作普通量一会当作零的问题，顺便证明一下微积分基本定理。为了完成这两件事情，我们需要补充很多关于数和函数的基本概念和相关的定理。不过，这些具体的学科概念知识都不是我们的重点。本节的首要目的是告诉你下面这个信息：你之前学会的不严格的微积分的每一个概念和定理都是可以被严格化的。本节的次要目的是帮助你体会到数学知识具有系统性，也就是从少数的概念和假设出发可以得到所有的其他定理，而构建这个系统的方法不过就是逻辑演绎证明。所以，请你在学习本节的时候处理好学会本节具体的数学学科概念知识和体会到这两点之间的关系。

16.8.1 极限

我们先来解决在导数计算中 Δx 一会儿是普通量一会是零的问题。我们说，Δx 是一个过程，一个无论和什么数比绝对值都比那个数要小的过程。写成数学语言构成的表达式如下：对于任意（$\forall$）的 $\epsilon > 0$，存在（$\exists$）一个 δ（实际这里只需要取 $\delta = \epsilon$），对于任意（$\forall$）的 $\Delta x \in (-\delta, \delta)$，满足 $|\Delta x| < \epsilon$。

这句话的意思是，Δx 是一个绝对值要多小就可以多小的变化的量变化的过程，或者说离 0 的距离要多小就有多小的数。从过程的角度这是可以理解的，那就是 $\Delta x \to 0$ 的意思呗。但是，过程会是或者说会对应着一个确定数吗？这确实是一个问题，我们来看几个过程就是或者说就对应着确定数的例子。

既是过程也是数的例子——数列的极限

我们说 $a = 0.\dot{9} = 0.99999\cdots$ 是一个过程，因为这数永远也写不完。这数看起来好想永远也比 1 小那么一点。比如说，当我们在第 n 位小数停下来的时候，

$$1 - a_n = 10^{-n}. \tag{16.268}$$

因此，对任意的 n，

$$a_n = 1 - 10^{-n} < 1. \tag{16.269}$$

那这里的 a 相当于是让这个 n 不断地变大的一个过程，我们还有 $a < 1$ 吗？

我们来做这样的一个之前就做过的计算，

$$10a = 9.\dot{9} \Rightarrow 10a - a = 9 \Rightarrow 9a = 9 \Rightarrow a = 1. \tag{16.270}$$

我们得到 $a = 1$，而不是 $a < 1$。这怎么理解：所有的有限位的 $a_n < 1$，但是 $a = 1$？

我们来套用一下之前的 Δx 当作过程的语言：$\forall \epsilon > 0$，$\exists N = \lfloor \log \frac{1}{\epsilon} \rfloor + 1$（这里 $\lfloor x \rfloor$ 表示对一个小数向下取其整数部分），对于任意的 $n > N$，有 $|1 - a_n| < \epsilon$。其证明如下，10^{-n} 是 n 的减函数，也就是 n 越大 10^{-n} 越小。因此当 $n > N = \lfloor \log \frac{1}{\epsilon} \rfloor + 1$ 时

$$0 < 1 - a_n = 10^{-n} < 10^{-N} = 10^{-\lfloor \log \frac{1}{\epsilon} \rfloor - 1} < 10^{\log \epsilon} = \epsilon. \tag{16.271}$$

因此

$$|1 - a_n| < \epsilon. \tag{16.272}$$

其中我们用到了

$$\lfloor x \rfloor \leqslant x < \lfloor x \rfloor + 1 \tag{16.273}$$

于是，对于减函数 10^{-n} 来说，

$$10^{-\lfloor x\rfloor-1}<10^{-x}. \tag{16.274}$$

这句话说的是，无论你希望 a_n 和 1 有多么接近，我们都可以找到满足这个条件的 N，并且从此之后所有的 a_n 都满足这个条件。于是，我们说，a_n 可以做到要多接近 1 就有多接近 1。或者说，a_n 和 1 之间的距离要多小就有多小，并且会随着 n 变大越来越小。

于是，我们发现，随着 n 变大 $a_n\to 1$ 就是很自然的一件事情了，而 $a_n\to a$ 啊，或者说 $a=0.\dot{9}$ 就是 n 无穷大的时候的 a_n 啊！因此，1 和 $a=0.\dot{9}$ 都就应该相同啊！一会儿我们就来把上面这个过程定义为数列的极限。用极限这个词，那 1 和 $a=0.\dot{9}$ 都是数列 a_n 在 n 趋向无穷大 $(n\to\infty)$ 的时候的极限啊。同一个动词的极限，只要极限存在，那么相同就是很自然的事情了。

这样，我们就从代数计算公式 (16.270) 和数列极限两个角度论证了 $0.\dot{9}=1$。注意：我们有了第一个同时是一个过程也是一个确定的数的例子。

另一个很有意思的是一个过程也是一个数的例子 $\sqrt{2+\sqrt{2+\sqrt{2+\cdot}}}$，并且这个例子从看起来是无理数——停在任何一个有限次的根号下面都是一个无理数——得到了一个有理数。

例 16.24 (从无理数构建出来有理数). 尝试求出来 $\sqrt{2+\sqrt{2+\sqrt{2+\cdot}}}$ 的值。

令 $x=\sqrt{2+\sqrt{2+\sqrt{2+\cdot}}}$，则 $x^2=2+\sqrt{2+\sqrt{2+\sqrt{2+\cdot}}}=x+2$。于是，$x$ 就是方程 $x^2-x-2=0$ 的解，并且是大于零的那个。

于是，求解方程 $x^2-x-2=(x-2)(x+1)=0$ 得到 $x=2$ ($x=-1<0$ 不符合本题的要求)。

历史上，求出来 π 的值——其定义是圆的周长和半径的比值的一半 $\pi=\frac{C}{2r}$——也是通过一些过程，例如祖冲之求 π 的过程就构建了这样的一个过程，而且这个过程包含了极限和夹逼的思想。祖冲之计算 π 的方法，用现代的语言来表达是这样的一个过程：用一个内接正多边形和外切正多边形，例如正四边形，来从内外把圆夹起来，可以直到圆的周长介于个正多边形的周长之间，只要计算出来两个正多边形的周长就得到了圆的周长的一个近似值；随着正多边形边长数的增加，例如下一个计算正八边形，就可以

得到更加接近圆的周长的两个多边形的周长的值。一直这样做下去，就可以得到越来越接近圆的周长的多边形的两个周长。如果这样的操作可以一直做下去，可以预期所得到的周长就会是圆的周长。因此，这样就可以得到任意精度的 π。一会儿我们会看到，这就是两个数列的极限。因此，历史上，在极限的概念被正式定义之前，极限的思想已经被使用。

下面，我们来定义数列的极限。

定义 16.1 (数列的极限). 给定一个实数数列 a_n，$n \in \mathbb{N}$ 为自然数。如果存在一个实数 a，满足下面的要求，我们就说 a 是数列 a_n 的极限：$\forall \epsilon > 0$，$\exists N$，$\forall n > N$，有 $|a_n - a| < \epsilon$。记作 $\lim_{n\to\infty} a_n = a$，或者 $a_n \to a$ 当 $n \to \infty$。

我们来看几个数列极限的例子。

例 16.25 (Zeno 序列). 证明数列 $a_n = 2^{-n}$ 的极限是 0。

也就是我们要证明 $\forall \epsilon > 0$，$\exists N$，$\forall n > N$，有 $|a_n - 0| < \epsilon$。我们取 $N = \lfloor \log_2 \frac{1}{\epsilon} \rfloor + 1$ 试试。$2^{-n} > 0$ 是一个减函数，再根据

$$\lfloor x \rfloor \leqslant x < \lfloor x \rfloor + 1 \tag{16.275}$$

我们有

$$2^{-N} = 2^{-\lfloor \log_2 \frac{1}{\epsilon} \rfloor - 1} < 2^{\log_2 \epsilon} = \epsilon. \tag{16.276}$$

因此，对于 $n > N$，$|2^{-n} - 0| = 2^{-n} < 2^{-N}$。合起来，我们得到，对于 $n > N$，

$$2^{-n} < \epsilon. \tag{16.277}$$

也就是说，数列 $a_n = 2^{-n}$ 的极限是 0，$\lim_{n\to\infty} 2^{-n} = 0$。 □

我们再来看数列 $a_n = 2^{-n}$ 的数列和的极限。有了这两个极限，我们就可以理解Zeno的另一个佯谬——Zeno “阿基利斯追不上乌龟”佯谬——了。

例 16.26 (Zeno 序列和). 证明数列 $a_n = 2^{-n}$ 的数列和 $S_n = \sum_{l=0}^{n} a_l$——注意这也是一个数列——的极限是 2。

首先，我们尝试把 S_n 的具体表达式算出来。对于等比序列 $a_n = aq^n$ $(n = 0, 1, 2, \cdots)$，我们可以直接运用之前学习过的等比序列求和公式，

$$S_n = \frac{a\left(1 - q^{n+1}\right)}{1 - q}. \tag{16.278}$$

我们也可以来推导出来这个公式。

$$S_n = a + aq + \cdots + aq^n, \tag{16.279}$$

则两边乘以 q 得到

$$qS_n = aq + aq^2 + \cdots + aq^n + aq^{n+1}, \tag{16.280}$$

把两个等式相减得到

$$S_n - qS_n = a - aq^{n+1}, \tag{16.281}$$

于是，我们就得到了公式 (16.278)。

回到我们这里，$a = 1$，$q = \frac{1}{2}$，因此，$S_n = 2\left(1 - 2^{-(n+1)}\right)$。我们前面已经证明了 $2^{-(n+1)}$ 的极限是 0，因此，很自然地（未来我们会补充这个关于极限的四则运算的定理），这里的 S_n 的极限是 $2(1-0) = 2$。

更一般地，如果 $|q| < 1$，我们可以证明 q^n 的极限是零。因此，$S_n = \frac{a\left(1-q^{n+1}\right)}{1-q}$ 的极限是 $\frac{a}{1-q}$。

或者，我们换一个办法来求这个极限。我们假设等比序列求和的极限 $S = \sum_{l=0}^{\infty} aq^l$ 真的存在并且可以写下来，例如这样

$$S = a + aq + aq^2 + \cdots \tag{16.282}$$

那么，两边同时乘以 q 得到

$$Sq = aq + aq^2 + aq^3 + \cdots \tag{16.283}$$

于是两者做差得到

$$S - Sq = a. \tag{16.284}$$

因此

$$S = \frac{a}{1-q}. \tag{16.285}$$

当然，对于 $|q| > 1$ 这个极限不一定存在。这个我们稍后再说。

为了熟悉数列极限的语言，我们再来直接用 $\epsilon - N$ 语言来求这个极限。也就是我们要证明 $\forall \epsilon > 0$，$\exists N$，$\forall n > N$，有

$$\left| \frac{a\left(1-q^{n+1}\right)}{1-q} - \frac{a}{1-q} \right| < \epsilon. \tag{16.286}$$

我们来求解这个方程，从而找出来合适的 N。为了简单。我们设定 $0<q<1$。对于我们的具体问题 $q=\frac{1}{2}$。于是，公式 (16.286) 成为

$$q^{n+1}<\frac{(1-q)\,\epsilon}{a}. \tag{16.287}$$

q^n 是关于 n 的减函数，因此，只要

$$n+1>\log_q\frac{(1-q)\,\epsilon}{a}, \tag{16.288}$$

就满足要求。于是，我们取 $N=\lfloor\log_q\frac{(1-q)\epsilon}{a}\rfloor$，当 $n>N$ 的时候，公式 (16.288) 成立，于是，公式 (16.286) 成立。也就是说，S_n 的极限是 $S=\frac{a}{1-q}$，$\lim_{n\to\infty}S_n=\frac{a}{1-q}$。 □

S_n 是一个过程——这里是把无穷多个 a_n 加起来的过程，同时也是一个数的另一个例子。从物理上，我们可以把 S_n 看作是下面这个被称作“阿基利斯追乌龟”过程的时间取和或者路程取和。

小知识 16.2 (阿基利斯追不上乌龟). 前面提出促进了导数的定义的*Zeno*“飞矢不动”佯谬的*Zeno*还提出了另一个促进极限的提出的*Zeno*“阿基利斯追不上乌龟”佯谬。

*Zeno*设计了下面这个场景：乌龟奔跑的速度是每秒零点五米（$v_T=0.5m/s$），半神阿基利斯的速度是乌龟的两倍（$v_A=2\times0.5m/s=1m/s$），乌龟在阿基利斯前面一米处（$L_0=1m$），两者同时开跑，请问过多长时间（$t=?$）阿基利斯追上乌龟。

我们小学就会求解这个问题：当阿基利斯追上乌龟的时候，两者的位置相同，也就是 $v_At=v_Tt+L_0$，于是 $t=\frac{L_0}{v_A-v_T}=\frac{1m}{0.5m/s}=2s$。我们的生活经验也可以告诉我们这样的答案是对的，阿基利斯是可以追上乌龟的。我们甚至可以来做实验检验这个答案。

但是，*Zeno*给出了这样的回答：当阿基利斯跑到乌龟之前的位置的时候，乌龟用这段时间往前走了一段。尽管是更短的一段（是阿基利斯同时间的这一段的一半，也是乌龟自己前一次走的一半），但是毕竟是在阿基利斯的前面。这样，乌龟就能保证永远出现在阿基利斯的前面，也就是阿基利斯追不上乌龟。

我们来用数学语言翻译一下这个回答。把一开始的时刻记为 $t=0$。第一次阿基利斯跑到乌龟之前的位置——也就是 L_0——的时间记为 $t_1=\frac{L_0}{v_A}$，

这段时间乌龟跑的路程为 $L_1 = \frac{L_0 v_T}{v_A}$。之后重复相同的模式，也就是，

$$t_{n+1} = \frac{L_n}{v_A}, \tag{16.289a}$$

$$L_{n+1} = \frac{v_T}{v_A} L_n. \tag{16.289b}$$

根据给定的 L_0, v_T, v_A 的值求解出来一般表达式，有

$$t_n = 2^{-n+1} s, \tag{16.290a}$$

$$L_n = 2^{-n} L_0 = 2^{-n} m. \tag{16.290b}$$

其中 $n = 1, 2, 3, \cdots$。

于是，乌龟的到第 n 次这样的过程为止的总路程和总时间可以表示为，

$$\mathscr{T}_n = \sum_{l=1}^{n} 2^{-l+1} s, \tag{16.291a}$$

$$\mathscr{L}_n = \sum_{l=1}^{n} 2^{-l} m. \tag{16.291b}$$

如果我们让这个取和不断地做下去，也就是 $n \to \infty$（n 趋向无穷大），按道理应该得到乌龟的总路程 $\mathscr{L}$ 和总时间 $\mathscr{T}$，用我们刚学会的数列的极限的语言也就是 $\mathscr{T}$（$\mathscr{L}$）是数列 $\mathscr{T}_n$（$\mathscr{L}_n$）在 $n \to \infty$ 下的极限，

$$\mathscr{T} = \mathscr{T}_{n\to\infty} = \sum_{l=1}^{\infty} 2^{-l+1} s, \tag{16.292a}$$

$$\mathscr{L} = \mathscr{L}_{n\to\infty} = \sum_{l=1}^{\infty} 2^{-l} m. \tag{16.292b}$$

这刚好就是之前我们计算出来的等比序列的和的极限，我们发现，

$$\mathscr{T} = \frac{1}{1-\frac{1}{2}} s = 2s, \tag{16.293a}$$

$$\mathscr{L} = \frac{\frac{1}{2}}{1-\frac{1}{2}} m = 1m. \tag{16.293b}$$

这说明，尽管按照*Zeno*的思考，阿基利斯需要无穷多次（按照*Zeno*定义的 n）才能追得上乌龟，但是，实际上把累计的时间和路程按照数列的极

限算出来，并不是无穷大，而是有限值；同时，我们也看到，阿基利斯和乌龟之间的距离差 $L_n = 2^{-n}m$ 不断地在缩小，要多小就可以多小只要 n 足够大，于是按照极限的定理，这个距离差的极限就是零。

因此，我们发现，阿基利斯能够追上乌龟，并且所用的时间和经过的距离有限。

其中的逻辑时间，也就是按照*Zeno*定义的 n，确实得到无穷大才行，而无穷大是一个过程又似乎永远到不了，于是就产生了*Zeno*“阿基利斯追不上乌龟”佯谬。

但是，在这里，我们看到，只要用上数列极限的概念再加上“有极限的过程也是一个数”的理解，我们就发现总时间总路程有限所以能追上，并且这里的两者之间的距离这个过程的极限就是零，所以能追上。

通过这个例子我们再一次看到了数列极限这个概念的威力：数列极限可以帮助我们理解“有极限的过程也是一个数”——那些有极限的数列可以看作过程其实同时就是数，就如同 $0.\dot{9} = 1$；有了数列极限和这个“有极限的过程也是一个数”我们就完美地解决了Zeno“阿基利斯追不上乌龟”佯谬。

无限不循环小数是过程也是数？——柯西收敛

我们看到了循环小数是过程也是数，那么，无线不循环小数呢？从这个讨论我们将引入一个叫做数列的柯西收敛的概念。

之前我们也学习过所有的循环小数都是分数，而分数可以在数轴上表示。于是，我们得到了所有循环小数都是数轴上的点。给出具体章节。 我们也构造出来了数轴上的无限不循环小数，也能够对给定的任意一个无限不循环小数来构造一个找到数轴上的点的过程。那，无限不循环小数是一个过程还是一个数？数轴上除了循环小数和无限不循环小数还有别的数吗？我们先来复习一下这部分，然后再来用极限的概念来回答这个问题，或者如果还需要其他概念才能回答这个问题的话，我们看看需要提出什么新的概念。

整数 n 可以用尺规（圆规和不带刻度的直尺）作图这样来表示在约定了原点和正方向的数轴上：约定一个单位长度，用来表示 1，然后用圆规来复制这个单位一的长度 n 次每一次都把新的长度向着正方向连到已有的长

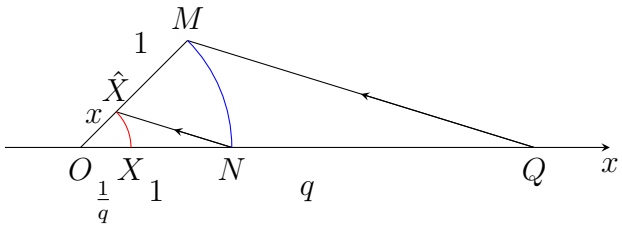


图 16.16: **分数在数轴上的表示**。$ON = 1, OQ = q, OM = 1$。红色和蓝色是圆弧。基于平行基本定理或者相似三角形的判定定理和性质，可以得到 OX 就是分数 $\frac{1}{q}$ 在数轴上的表示。

度上，最后到达的位置就是整数 n 在数轴上的位置。

分数 $\frac{q}{p}$ 可以用尺规作图这样来表示在数轴上：先在数轴上表示出来 $\frac{1}{p}$ 的点 P，再用上面的方式把从原点 O 到 P 的这一段复制 q 份，接起来。

如图 16.16，在数轴上先找到 1 和 q 这两个整数对应的点，分别记作 N 和 Q。过原点 O 做一条不和数轴重合的直线段，取其长度为 1，记其另一个终点为 M。连接 QM，过 P 做平行于 QM 的直线交 OM 于 X——这个也可以用尺规作图来完成。则，OX 的长度为 $x = \frac{1}{q}$。最后我们通过做一个以原点为圆心，OX 为半径的圆，交 x 轴于一点。那个点就是 $\frac{1}{q}$ 所对应的点。

证明 16.1. 根据平行基本定理——平行线切分线段成比例，或者根据相似三角形判定定理和性质定理，有 $\frac{x}{1} = \frac{OX}{OM} = \frac{ON}{OQ} = \frac{1}{q}$，因此，$x = \frac{1}{q}$。

换句话说，在数轴上表示分数不过就是改变一下单位一，用更小的尺度——原来的单位一的 $\frac{1}{q}$——来当新的单位一，则在新的单位一下，分数 $\frac{p}{q}$ 就相当于以前的整数 p。这也是为什么有理数会被称为“比例数”：只要缩放一定比例就可以变成整数。那么，无理数呢，怎么表示？运用勾股定理可以表示任何一个 $\sqrt{n}$ 以及更一般的 $\sqrt{\frac{p}{q}}$。但是更一般的无理数就不好表示了。

我们也来从有限小数和无线循环小数（注意后者包含前者，相当于循环体是零）在数轴上的小数的表示开始。小数的整数部分我们已经解决。现在我们只关心小数部分的表示。未来把小数部分的表示和整数部分的表示用首尾相连的方式“加起来”就可以。

在数轴上表示纯小数的办法是，对于每一个小数位的上的整数（必然

是个位数，在 $0 \sim 9$ 之间，包含 $0, 9$），相应地把上一个小数位（对于十分位其上一个数位是个位）的一个单位均匀划分成十份，找到下一个小数位上的整数对应的位置，就行了。这里面唯一需要的技术就是把一份分成十份。刚好，我们之前已经学会了从 q 得到 $\frac{1}{q}$ 的表示方法。这里只需要取 $q = 10$ 就行了。对于有限小数，这样做原则上是准确的。如果遇到无限小数，这样做就会没完没了。

我们来考察这样得到的两个数——无限循环小数在 n 分位和 $n+1$ 分位的两个表示 a_n, a_{n+1}——之间的差，

$$\Delta a_n \triangleq a_{n+1} - a_n < 10^{-n}. \tag{16.294}$$

而这个差会随着 n 趋向无穷而不断趋于零，也就是当 $n \to \infty$ 的时候，

$$\Delta a_n \to 0. \tag{16.295}$$

也就是 a_n 和 a_{n+1} 之间的差会随着 $n \to \infty$ 越来越小。那么，如果 a_n 和 a_{n+1} 各自的极限都存在，它们的极限就应该是同一个数。对于无限循环小数，那个极限必然存在，因为所有的无限循环小数都是分数，而分数可以在数轴上找到表示。也就是说，数轴上有这样一个数。

那么，是不是所有满足 $\Delta a_n \to 0$ 的数列，极限都存在呢？我们来构造几个这样的数。在那之前，我们给这个 $\Delta a_n \to 0$ 稍微做一个拓展并取一个专门的名字——柯西收敛。

定义 16.2 (柯西收敛). 一个数列称为柯西收敛的，如果满足下面的条件，对任意给定的 $\epsilon > 0$，总存在正整数 N，使得当 $m, n > N$ 时，

$$|a_m - a_n| < \epsilon. \tag{16.296}$$

例 16.27 (趋向 $\sqrt{2}$ 的数列). 定义 $a_n = \lfloor \sqrt{2} \times 10^n \rfloor \times 10^{-n}$。我们检验一下这个序列是否柯西收敛。注意，这里的 a_n 是有理数，并且是有限长度的有理数。

我们来计算几个例子：$a_0 = \lfloor \sqrt{2} \times 10^0 \rfloor \times 10^0 \approx \lfloor 1.414 \rfloor = 1$，$a_1 = \lfloor \sqrt{2} \times 10^1 \rfloor \times 10^{-1} \approx \lfloor 14.14 \rfloor \times 10^{-1} = 1.4$，$a_2 = \lfloor \sqrt{2} \times 10^2 \rfloor \times 10^{-2} \approx \lfloor 141.4 \rfloor \times 10^{-3} = 1.41$。因此，这个函数的作用就是逐位近似 $\sqrt{2}$。这个函数其实正好就是上面描述的每次找到下一位的在数轴上找数的方法。

于是，很自然地，我们有

$$\sqrt{2}\times 10^n \leqslant \lfloor \sqrt{2}\times 10^n \rfloor < \sqrt{2}\times 10^n + 1, \tag{16.297}$$

因此

$$\sqrt{2} \leqslant a_n = \lfloor \sqrt{2}\times 10^n \rfloor \times 10^{-n} < \sqrt{2} + 10^{-n}, \tag{16.298}$$

也就是

$$0 \leqslant a_n - \sqrt{2} < 10^{-n}. \tag{16.299}$$

于是

$$\begin{aligned} |a_m - a_n| = & \left|\left(a_m - \sqrt{2}\right) - \left(a_n - \sqrt{2}\right)\right| \\ \leqslant & \left|\left(a_m - \sqrt{2}\right)\right| + \left|\left(a_n - \sqrt{2}\right)\right| \\ < & 10^{-m} + 10^{-n}. \end{aligned} \tag{16.300}$$

因此，只要我们取 $N = \lfloor \log \frac{\epsilon}{2} \rfloor + 1$，我们就有 $\forall m, n > N$，$10^{-m} < \frac{\epsilon}{2}$，$10^{-n} < \frac{\epsilon}{2}$，因此，

$$|a_m - a_n| < 10^{-m} + 10^{-n} < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon. \tag{16.301}$$

也就是说，$a_n = \lfloor \sqrt{2}\times 10^n \rfloor \times 10^{-n}$ 是柯西收敛的有理数数列。

这个方法可以给所有的无限不循环小数找到一个柯西收敛的有理数数列。这个数列中的每一个数都是有理数，都可以表示为分数，都可以在数轴上表示出来。于是，我们产生了一个很自然的问题——这个有理数列的极限是不是存在，并且也可以在数轴上表示出来呢？

但是，我们已经知道 $\sqrt{2}$ 不是有理数。也就是说，如果这个 a_n 的极限存在，那么，这就是一个有理数的数列，但是其极限是一个无理数。也就是有理数集合对极限的运算不封闭。那问题来了，是不是还可以造出来一个无理数的数列，其极限不是一个无理数呢？换句话说，是不是加入无理数之后，任何一个柯西收敛的数列的极限就封闭了，仍然是无理数了呢？

由于“柯西收敛的数列极限存在、数轴上表示的点的极限仍然在数轴上”这个性质太好了，并且万一不满足这个性质，那就按照极限来定义新的数，补充到已有的数的集合之中，我们就希望给那个满足“柯西收敛的数列极限存在”的数的集合一个名称——实数。

反过来，任何一个数轴上的点 P，我们也可以找到一个这样的有理数数列。先从整数开始，看 x_P 落在哪两个相邻的整数之间，这样我们就确定了其整数部分。然后，再把这个长度为 1 的间隔分成十个等份，分的方法还是之前的 $\frac{1}{q}$ 的尺规作图。对了，之所以用尺规作图而不是依靠事先画好的刻度，是因为尺规作图第一是完全准确的第二是理论上可以不断地做下去的，而不是说没有标出来的刻度就没法继续下去。这样，我们可以问在十分位这个层面上，x_P 介于 $0.0 \sim 1.0$ 或者说写成整数 $0 \sim 9$ 的哪两个数之间，这样就确定了十分位上的数。我们可以不断地继续做下去。至于实际做的时候太小了眼睛看不清，第一我们可以用显微镜来放大或者通过之前的 $\frac{1}{a}$ 反过来来放大，第二物理上能做到多么精细和数学上原则上可以做到多精细是两件事情。这样忽略整数位，我们就得到了对应着数轴上任意一个点 P 的数列 $a_1, a_2, \cdots, a_n, \cdots$。

根据前面的过程，相邻的两个数 a_{n+1} 和 a_n 相差最大就是 $9 \times 10^{-(n+1)}$，于是我们得到

$$|a_{n+1} - a_n| = a_{n+1} - a_n \leqslant 9 \times 10^{-(n+1)} < 10^{-n}. \tag{16.302}$$

更一般地对于 $m > n$，

$$\begin{aligned} a_{n+1} - a_n &\leqslant 9 \times 10^{-(n+1)}, \\ a_{n+2} - a_{n+1} &\leqslant 9 \times 10^{-(n+2)}, \\ &\cdots, \\ a_m - a_{m-1} &\leqslant 9 \times 10^{-(m-1)}. \end{aligned} \tag{16.303}$$

所有的等式加起来，得到

$$\begin{aligned} a_m - a_n &\leqslant 9 \times 10^{-(n+1)} + 9 \times 10^{-(n+2)} + \cdots + 9 \times 10^{-(m-1)} \\ &= 9 \times 10^{-(n+1)} \left(1 + 10^{-1} + \cdots + (10^{-1})^{m-n-1}\right) \\ &= 9 \times 10^{-(n+1)} \frac{1 - (10^{-1})^{m-n}}{1 - 10^{-1}} \\ &< 10^{-n}. \end{aligned} \tag{16.304}$$

因此，$\forall \epsilon > 0$，只要我们选择 $N = \lfloor \log \epsilon \rfloor + 1$，我们就能够使得 $\forall m, n > N$，

$$|a_m - a_n| < \epsilon. \tag{16.305}$$

也就是这样选出来的对应着数轴上任意一个点 P 的数列 $a_1, a_2, \cdots, a_n, \cdots$ 是柯西收敛的。因此，如果我们的实数却是满足任意柯西收敛的数列必然有极限，则任何数轴上的点必然对应着一个实数。这样我们就证明了，如果实数这个数的集合满足“柯西收敛则极限存在”，那么数轴上的点和实数完全一一对应。而在我们这里，实数就是通过不断地扩张来补齐数从而满足“柯西收敛则极限存在”的。因此，这一条天生就满足。当然，我们可能很难真的穷尽所有的柯西收敛的数列，因此，依靠这条路来真的补齐是很难的。只要这样的补齐了的数集存在，那么，原则上，我们就证明了实数和数轴上的点一一对应，并且在实数范围内“柯西收敛则极限存在”成立。我们在一次强调，在有理数范围内，这两条都不成立。

在这里，我们只需要知道原则上能够通过补齐来得到一个更大的数的集合就行了。下面这一节，我们介绍了补齐的方式——戴德金分割，以及我们得证明为什么那样就相当于是补齐了，使得“柯西收敛则极限存在”成立了。不过，这部分相当抽象和烧脑，不做学习要求。可以等你更加熟悉了这些语言、符号和内容之后，有兴趣的时候回过头来再看。不过，这部分对于构建严格的微积分的知识体系是非常重要的。它们代表了数学发展的另一个驱动力——使得数学知识体系本身更加严密。微积分的不严格的概念和计算、应用发展起来之后，正是这个使得微积分严格化去掉所有逻辑漏洞的努力，才使得这部分知识发展起来。这也是人类智慧的瑰宝。如果你能够欣赏得了，那也是一种幸福。

顺便，从数列有极限转向数列柯西收敛的讨论也是一个人类智慧的瑰宝：对于一个一般的数列 $\{a_n\}$，你可能很难直接猜出来其极限 a，甚至对于有理数数列其极限可能不是一个有理数，也就数列的极限都可能超越这个数列本身所定义的集合，因此，直接去判断是否能够找到足够大的 N 使得 $n > N$ 时候 $|a_n - a| < \epsilon$ 对于任意的 $\epsilon > 0$。但是，柯西收敛的判断不需要事先知道这个极限，而且前面我们已经看到从数轴上找对应的数，从数找数轴上对应的点，自然就是一个从 a_n 到 a_{n+1} 的过程。因此，数列柯西收敛是一个远远比数列有极限好很多的概念。下面唯一要完成的就是，找到合适的条件，或者说合适的集合，使得这两者成为一回事，也就是“柯西收敛则极限存在”，就行了。你想，数学家们当年得多么殚精竭虑才构建起来这样一条孤立拐弯的路才使得微积分的概念严谨起来啊。如果你不能欣赏

这样的精巧深刻的思考，那绝对是遗憾。

习题 16.11 ($\frac{1}{n}$ 的极限). 证明 $\frac{1}{n}$ 的极限是 0，同时证明其满足柯西收敛的要求。注意，这里的没有任何一个 a_n 真的满足 $a_n = 0$。

习题 16.12 ($\frac{n}{n+1}$ 的极限). 证明 $\frac{n}{n+1}$ 的极限是 1，同时证明其满足柯西收敛的要求。注意，这里的没有任何一个 a_n 真的满足 $a_n = 1$。

有的时候我们也要讨论一个命题的否命题和逆否命题。我们来介绍一下，如何来把出现了 $\forall, \exists$ 的命题变成否命题。我们说一般公式是：$\forall \to \exists$，$\exists \to \forall$，然后把最终结论中的（其他地方的不变）关系运算符改变符号，例如 $> \to \leqslant$。

例如，我们来把下面这段之前出现过的话改成否命题：

> 对于任意（$\forall$）的 $\epsilon > 0$，存在（$\exists$）一个 δ（实际这里只需要取 $\delta = \epsilon$），对于任意（$\forall$）的 $\Delta x \in (-\delta, \delta)$，满足 $|\Delta x| < \epsilon$。

反过来，就成了：存在 $(\exists)$ $\epsilon > 0$，对于任意的 $(\forall)$ δ，存在 $(\exists)$ $\Delta x \in (-\delta, \delta)$，满足 $|\Delta x| \geqslant \epsilon$。

最终的结论中，我们实际上先把结论修改为其否定形式“不满足 $|\Delta x| < \epsilon$”，然后把这个“不满足”明确地写成等价的数学表达式“满足 $|\Delta x| \geqslant \epsilon$”。

初学这套语言的时候，这个转换不太容易。你尝试着多做几次，就会发现这个转换很有道理。

戴德金分割 *

实际上，我们还需要一个构造性的道路来给出来一个这样的数的集合，来使得“柯西收敛则极限存在”成立。比如说，我们试试从有理数集合[15]构建出来一个这样的一个数集。这就是戴德金分割完成的事情。Julius Wilhelm Richard Dedekind（尤利乌斯 · 威廉 · 理查德 · 戴德金）这样来定义有理数的戴德金分割：定义两个非空的有理数集合 $\mathbb{Q}$ 的子集 L, U，满足第一不相交 $L \cap U = \phi$，第二合起来等于有理数集合 $L \cup U = Q$，第三有序

[15]有理数集合本身我们是从自然数集合加上除法运算就可以建立起来的。至于自然数，我们暂时认为是从生活经验中直接提炼出来的，只要有了 0 和 1，然后通过加法就可以构建出来所有的自然数。或者，我们前面也展示了从集合如何构建自然数：$\{\phi, \{\phi\}, \{\phi, \{\phi\}\}, \cdots\}$。具体章节，在前面补上从集合构建自然数 未来，我们也会学习到自然数的公理化定义。

$x < y, \forall x \in L, y \in U$。这个分割记作 (L, U)。这个分割可能在有理数集合中存在分界点，也就是 L 中存在一个最大值有理数但是 U 中没有最小值有理数，或者 U 中存在一个最小值有理数但是 L 中没有最大值有理数，也可能在有理数集合中不存在分界点，也就是在 L 中不存在属于最大值有理数 U 中也不存在最小值有理数。

顺便，同时在 U 中存在一个最大值有理数 U 中存在最小值有理数的情况本来就不可能存在。如果存在，前者记为 a，后者记为 $b \neq a$（$L \cap U = \phi$，不能有共同元素），则 $c = \frac{a+b}{2}$ 还是一个有理数（有理数对加减乘除运算封闭），而 $a < c < b$，因此，$c \notin L, c \notin U$，这个和 $L \cup U = Q$ 矛盾。

我们给其他三种情况举几个例子。如图 16.17 所示，(a) 取 $L = \{x | x \leqslant 1, x \in \mathbb{Q}\}$，$U = \{x | x > 1, x \in \mathbb{Q}\}$，可以验证这确实是有理数的一个戴德金分割，同时 L 中存在一个最大值有理数 1 但是 U 中没有最小值有理数。(b) 取 $L = \{x | x < 1, x \in \mathbb{Q}\}$，$U = \{x | x \geqslant 1, x \in \mathbb{Q}\}$，可以验证这确实是有理数的一个戴德金分割，同时 U 中存在一个最小值有理数 1 但是 L 中没有最大值有理数。实际上，把其中的分界点换成 1 任何一个有理数，都可以得到一个戴德金分割。(c) 取 $L = \{x | x^2 \geqslant 2, x \in \mathbb{Q}\} \cup \{x | x < 0, x \in \mathbb{Q}\}$，$U = \{x | x^2 > 2, x > 0, x \in \mathbb{Q}\}$，可以验证这确实是有理数的一个戴德金分割，但是 L 中不存在一个最大值有理数而且 U 中也没有最小值有理数（因为最小值实际上是无理数 $\sqrt{2}$）。(d) 取 $L = \{x | x^2 < 2, x \in \mathbb{Q}\} \cup \{x | x < 0, x \in \mathbb{Q}\}$，$U = \{x | x^2 \geqslant 2, x > 0, x \in \mathbb{Q}\}$。得到的情况和 (c) 相同。

于是，我们从有理数定义出来的所有的戴德金分割的集合 $\mathbb{R} \triangleq \{(L, U)\}$，实际上是超越了有理数集合 $\mathbb{Q}$ 的：任何一个有理数 q 可以对应于一个分割 $L = \{x | x \leqslant q, x \in \mathbb{Q}\}, U = \{x | x > q, x \in \mathbb{Q}\}$ 或者等价地 $L = \{x | x < q, x \in \mathbb{Q}\}, U = \{x | x \geqslant q, x \in \mathbb{Q}\}$[16]；但是，我们已经看到了分割的集合 $\mathbb{R}$ 里面有不是 $\mathbb{Q}$ 的元素。因此 $\mathbb{Q} \subset \mathbb{R}$。

既然在有理数范围内，任何一个戴德金分割相当于定义了一个分界点，那么，我们推而广之认为戴德金分割的集合的所有元素都对应着一个分界点。我们问这些戴德金分割的集合 $\mathbb{R}$ 是不是就是我们已经听说过的实数呢？比如说，$\mathbb{R}$ 对戴德金分割封闭吗？也就是如果我们对戴德金分割的集合 $\mathbb{R}$ 再做戴德金分割，是不是可以保证每一个戴德金分割对应的分界点都在 $\mathbb{R}$

[16]在这里等价的意思是，这两个分割对应着同一个分界点 q。

里面呢？也就是我们要证明，对于戴德金分割的集合 $\mathbb{R}$，其上的戴德金分割$(\mathscr{L},\mathscr{U})$，要么 $\mathscr{L}$ 在 $\mathbb{R}$ 内有最大值，要么 $\mathscr{U}$ 在 $\mathbb{R}$ 内有最小值。

回到我们熟悉的但是不严格的实数——一一对应着数轴上的点，在实数范围内，戴德金分割就是选定了一个数轴上的点 x_P，把 L 取为其左侧的所有的点的集合 $\{x|x < x_P\}$，U 取为其右侧的所有点的集合 $\{x|x > x_P\}$，至于这个点本身，要么归于左侧 L，要么归于右侧 U，但是不能同时属于两者（$L \cap U = \phi$，不能有共同元素），则这个分割就对应着点 x_P。如果 $x \in \mathbb{Q}$ 限制为有理数，则当 x_P 为无理数的时候，就算这个点归入到左侧或者有侧，实际上在有理数范围内，也取不到这个实数，因此就出现了“L 中不存在属于最大值 U 中也不存在最小值”的情况。实数和有理数相比，就是使得第三种情况，“L 中不存在属于最大值 U 中也不存在最小值”不再出现。也就是对于图 16.17 的 (c) 和 (d)，在实数范围内，必然出现要么在 L 中有最大值，要么在 U 中有最小值。也就是前面的命题完全成立。不过，现在我们不能先通过这样的基于几何直觉的方式来理解实数和戴德金分割，而是通过戴德金分割来构建的集合 $\mathbb{R}$，我们要重新来证明这个命题成立。

一旦这个命题成立，则，我们确实从有理数集合构建了一个把有理数集合当作真子集的更大的数集，并且这个数集对于戴德金分割封闭。有了这个封闭性，就使得任何一个有上界的实数集合的子集有上确界，任何一个有下界的实数集合的子集有下确界。而这一点使得“柯西收敛则极限存在”成立。

顺便，学习和理解任何一个数学命题的时候，要从外部和内部两个角度来学习：第一，外部的角度，这个命题的提出受到了什么东西的启发（数学知识本身的系统性，或者现实世界，或者两者同时）；第二，外部的角度，有了这个命题，对数学知识体系本身或者现实世界——可能得通过物理学等科学学科——会起到什么作用；第三，内部的角度，这个命题的具体证明过程是怎样做的；第四，有了内部之后的外部，这个命题的证明过程受到了什么启发，这个证明过程是否还可以迁移到其他命题的证明中。其中，前两个是非常重要的无论如何要先看到的，后两个可以慢慢看到慢慢体会到。其实，理解和掌握其他的对象的时候，也是如此，要内外角度相结合。

下面我们来证明这个命题，并且从这个命题建立起来“柯西收敛则极限存在”。

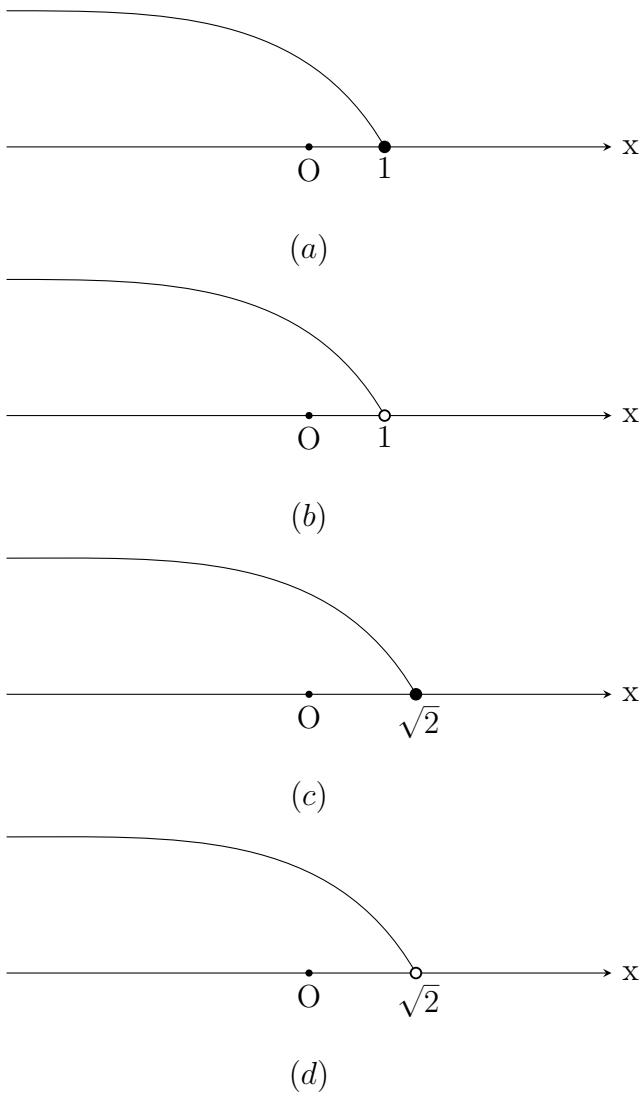


图 16.17: **戴德金分割举例**。戴德金分割的四种情况：分界点为有理数和无理数，L 包含和不包含分界点。

从戴德金分割到柯西收敛 *

这部分更加抽象，但是同样非常非常地精妙。

我们先来证明戴德金分割的集合 $\mathbb{R}$ 对戴德金分割封闭。为了给戴德金分割的集合 $\mathbb{R}$ 构建戴德金分割，我们来定义戴德金分割之间的大小关系。

定义 16.3 (戴德金分割比大小). 对于两个戴德金分割，(L_1, U_1) 和 (L_2, U_2)，如果 $L_1 \subset L_2$（这是是真包含关系，不允许出现 $L_1 = L_2$），我们说 $(L_1, U_1) < (L_2, U_2)$。

补充上两个集合相等的时候对应着等于号。这样我们就有了戴德金分割之间的 $\leqslant$ 关系。

我们先来证明任意两个不同的有理数集合上的戴德金分割(L_1, U_1) 和 (L_2, U_2) 必然存在一个大于另一个。注意到以有理数集合 $\mathbb{Q}$ 为全集 $U_1 = \bar{L}_1$ 和 $U_2 = \bar{L}_2$。未来以实数为全集这一点仍然成立。这个结论稍后会用到。

定理 16.6: 戴德金分割有序

证明 (L_1, U_1) 和 (L_2, U_2) 必然存在一个大于另一个。也就是要么 $L_1 \subset L_2$，要么 $L_2 \subset L_1$。

证明 16.2. 我们用反证法，假设结论不成立，则必定同时存在 $x_1 \in L_1$ 但是 $x_1 \notin L_2$ 和 $x_2 \in L_2$ 但是 $x_2 \notin L_1$。根据 $x_1 \notin L_2$，我们知道 $x_1 \in U_2$。根据 $x_2 \notin L_1$，我们知道 $x_2 \in U_1$。根据 $x_1 \in L_1$ 和 $x_2 \in U_1$，我们有 $x_1 < x_2$。根据 $x_2 \in L_2$ 和 $x_1 \in U_2$，我们有 $x_2 < x_1$。矛盾。因此假设不成立，也就是 $L_1 \subset L_2$，要么 $L_2 \subset L_1$。

这样我们就把所有的有理数集合上的戴德金分割按照顺序排列起来了。注意，对于有理数 q，$L_q = \{x | x < q\}$ 和 $\tilde{L}_q = \{x | x \leqslant q\}$ 的分界点相同。因为对于前者 L_q 每有极大值但是 $U_q = \bar{L}_q = \{x | x \geqslant q\}$ 有极小值 q，所以分界点是 q；对于后者 $\tilde{L}_q$ 有极大值 q 而 $\tilde{U}_q = \bar{\tilde{L}}_q = \{x | x > q\}$ 没有极小值，所以分界点也是 q。也就是说，我们把有理数集合上的戴德金分割按照顺序排列起来以后，实际上是这样的，

$$\cdots \subset L_p \subset \tilde{L}_p \subset \cdots \subset L_q \subset \tilde{L}_q \subset \cdots \subset L_r \subset \tilde{L}_r \subset \cdots \tag{16.306}$$

其中 $p < q < r$，中间还有对应着其他有理数和无理数的戴德金分割。

这样我们就可以构建出来戴德金分割的集合 $\mathbb{R}$ 的戴德金分割，$(\mathscr{L},\mathscr{U})$，满足：

- 第一、$\mathscr{L},\mathscr{U}$ 是 $\mathbb{R}$ 的非空子集，
- 第二、不相交 $\mathscr{L}\cap\mathscr{U}=\phi$，
- 第三、合起来等于实数集 $\mathscr{L}\cup\mathscr{U}=\mathbb{R}$，
- 第四、有序 $L<U,\forall L\in\mathscr{L},U\in\mathscr{U}$。

我们来证明 $(\mathscr{L},\mathscr{U})\in\mathbb{R}$。

既然我们已经在公式 (16.306) 中把 $\mathbb{R}$ 做了一个排序，则任何一个 $(\mathscr{L},\mathscr{U})$ 必然是对这个排序在某个点上的截断，例如 $\mathscr{L}=\{L|L<L_\xi\}$ 或 $\mathscr{L}=\left\{L|L<\tilde{L}_\xi\right\}$，或者 $\mathscr{L}=\{L|L\leqslant L_\xi\}$ 或 $\mathscr{L}=\left\{L|L\leqslant\tilde{L}_\xi\right\}$。对于前者，则 $\mathscr{U}$ 中有最小值。对于后者，则 $\mathscr{L}$ 中有最大值。

我们还看到这个最大值或者最小值必然对应着一个对有理数集合的戴德金分割L_ξ，也就是集合 $\mathbb{R}$ 的元素。也就是说，对集合 $\mathbb{R}$ 的戴德金分割仍然是 $\mathbb{R}$ 的元素。这样我们就用有理数集合上的戴德金分割构建了一个对进一步的戴德金分割封闭的集合，记作实数集 $\mathbb{R}$。

其中最关键的性质是：实数集 $\mathbb{R}$ 的戴德金分割(L,U)，必然会出现 L 内有最大值或者 U 内有最小值的情况之一（也不能同时出现），因此这个分界值是唯一的。所以，一个分割就对应着一个分界值。我们把实数上的戴德金分割的这个性质写成一条定理——称为戴德金定理。上面我们已经证明了这条定理。

定理 16.7: 戴德金原理

实数集 $\mathbb{R}$ 的戴德金分割(L,U)，必然会出现和仅仅出现以下两种情况之一：L 内有最大值 U 中没有最小值，或者 U 内有最小值 L 中没有最大值。

我们还可以定义分割——注意这是集合——的加减乘除，来保证其得到的结果对应着分界值的加减乘除运算。比如说，两个 L 的加法被定义为：$L_{12}=\{x_1+x_2|x_1\in L_1,x_2\in L_2\}$。如果这 L_1,L_2 都内有最大值 ξ_1,ξ_2，自然

地，我们得到的 L_{12} 的最大值为 ξ_1, ξ_2。你可以自己来求解下面的习题来验证这个结论。其他的运算我们就不展开了。

习题 16.13 (两个 L 的加法). 设 $L_1 = \{x|x \leqslant \xi_1, x \in \mathbb{R}\}, L_2 = \{x|x \leqslant \xi_2, x \in \mathbb{R}\}$，证明 $L_{12} = \{x_1 + x_2|x_1 \in L_1, x_2 \in L_2\}$ 等价于 $\{x|x \leqslant \xi_1 + \xi_2, x \in \mathbb{R}\}$。

这样就把有理数集合上的戴德金分割和你之前就听说过的实数完全对应起来了。

有了实数，我们怎么得到在实数集合上，“柯西收敛则极限存在”呢?

我们先来证明柯西收敛的数列有界。按照柯西收敛的定义，我们有，$\forall \epsilon > 0, \exists N \in Z^+$，使得对于 $\forall m, n > N$，有 $|a_m - a_n| < \epsilon$。我们取 $m = N+1$，于是

$$|a_n - a_{N+1}| < \epsilon \Rightarrow a_{N+1} - \epsilon < a_n < a_{N+1} + \epsilon. \tag{16.307}$$

把前面的有限个元素 $a_1, a_2, \cdots, a_{N+1}$ 放在一起构成一个集合。因为是有限个，肯定有界。后面的所有的 a_n $(n \geqslant N+2)$ 也有界。因此，合起来有界。我们把这个边界记为 $a_n \in [\mu, \nu]$。

我们来根据这个 a_n 数列构造一个实数集上的戴德金分割。我们把集合 L 的元素定为这样的 x：$(-\infty, x)$ 中最多包含了 a_n 的有限项，(x, ∞) 包含了 a_n 的无限项。比如说，比 μ 更小的实数都在 A 中。具体哪些实数在 A 中得看 a_n 的具体取值。

直观地说，A 里面的元素就是用来逼近 a_n 的极限的：从比 a_n 所有的数都更小的数开始，一点点增加，靠近 a_n 的极限。可以发现，无论 x 靠的多么近，只要不等于这个极限值，就可以满足上面的要求：$(-\infty, x)$ 中最多包含了 a_n 的有限项，(x, ∞) 包含了 a_n 的无限项。但是，a_n 的极限，以及比这个极限还大的数，例如 ν 不能在 L 里面。否则，$(-\infty, \nu)$ 中就得包含 a_n 的无穷多项了。

我们取 $U = \bar{L}$，全集为实数。我们已经看到，ν 和大于 ν 的实数都在 U 里面。因此，这样的 (L, U) 肯定是实数集的一个戴德金分割。两者非空，互斥和全集为 $\mathbb{R}$ 都是显然的。我们来证明两者有下面的关系：任意 $x \in L, y \in U$，有 $x < y$。

按照 U 的定义，其元素 y 必须满足：$(-\infty, y)$ 中含有 a_n 的无穷多项。因此，必然有 $x < y$。否则，如果 $x = y$，则 $(-\infty, y = x)$ 中最多含有 a_n 的有限项；如果 $x > y$，则 $(-\infty, y) \subset (-\infty, x)$，而后者最多含有 a_n 的有限项。

矛盾。

既然我们构建的 (L,U) 肯定是实数集的一个戴德金分割，要么 L 中有最大值，要么 U 中有最小值。假设这个分界点是 a，则 a 要么是 L 中的最大值，要么是 U 中的最小值。如果是前者，则对于任意的 $\epsilon_1 > 0$，有 $a-\epsilon_1 \in L$。如果是后者，也有 $a-\epsilon_1 \in L$。否则，根据戴德金分割的定义，就有 $a-\epsilon_1 \in U$，于是 a 不再是 U 中的最小值，至少 $a-\epsilon_1$ 更小，也在 U 中。矛盾。因此，我们得到 $a-\epsilon_1 \in L$。

同理，我们得到对于任意的 ϵ_2，$a+\epsilon_2 \in U$。如果 a 是 U 中的最小值这显然成立。如果 a 是 L 中最大值，假设 $a+\epsilon_2 \in U$ 不成立，那么 $a+\epsilon_2 \in L$，于是 $a+\epsilon_2$ 更大而且在 L 中，这个和 a 是 L 中最大值矛盾。因此，我们得到 $a-\epsilon_2 \in U$。

既然如此，我们就发现，$(-\infty, a-\epsilon_1)$ 只能含有 a_n 中的有限项，也就是在某个 $m > N_L \in Z^+$ 之后，所有的 $a_m > a-\epsilon_1$。同理，$(-\infty, a+\epsilon_2)$ 必须含有 a_n 中的无限多项，也就是在某个 $m_k > N_U \in Z^+$ 之后，有 $a_{m_k} \leqslant a+\epsilon_2$。这里的 a_{m_k} 表示至少有无穷多个这样的 a_n 中的元素，但是不能保证其中每一个都满足上面的条件。

合起来，取 $N_{LR} = \max\{N_L, N_R\}$，对于任意的 $\epsilon > 0$，我们都能得到存在一个子数列 a_{m_k}

$$-\frac{\epsilon}{2} < a_{m_k} - a \leqslant \frac{\epsilon}{2}. \tag{16.308}$$

根据柯西收敛，我们还有对于任意的 $\epsilon > 0$，存在 $N_C \in Z^+$，对于任意的 $m_k, n > N_C$，有

$$-\frac{\epsilon}{2} < a_n - a_{m_k} < \frac{\epsilon}{2}. \tag{16.309}$$

合起来，我们有当 $n > N = \max\{N_{LR}, N_C\}$ 的时候，

$$-\epsilon < a_n - a < \epsilon. \tag{16.310}$$

也就是 a_n 有极限。

这样我们就证明了柯西收敛则有极限。

定理 16.8: 柯西收敛则有极限

数上的柯西收敛数列有极限。

函数版本的这个定理的证明就不重复了。参照着这个数列版本的定理的证明就可以完成。

回顾这个证明，我们先证明柯西收敛的数列有界，再构造了一个戴德金分割来得到有界数列必然存在有极限的子数列，然后证明柯西收敛的数列整体都收敛到这个子数列的极限。

顺便，有极限则柯西收敛这个结论是很容易证明的。

习题 16.14 (数列有极限则柯西收敛). 证明一个数列如果 $n \to \infty$ 的极限存在 $\lim_{n\to\infty} a_n = a$，则这个数列柯西收敛，也就是 $\forall \epsilon > 0$，$\exists N \in Z^+$，$\forall m, n > N$，$|a_m - a_n| < \epsilon$。

习题 16.15 (函数有极限则柯西收敛). 证明一个函数如果在 $x = x_0$ 极限存在 $\lim_{x\to x_0} f(x) = a$，则这个函数在 $x = x_0$ 处柯西收敛，也就是 $\forall \epsilon > 0$，$\exists \delta > 0$，$\forall x_1, x_2 \in O(x_0, \delta)$，$|f(x_1) - f(x_2)| < \epsilon$。

这样合起来，我们就把有极限和柯西收敛完全等同了。

上面的证明过程中得到的中间结论：有界序列，必然存在一个有极限的子序列，被称为聚点定理（Bolzano–Weierstrass），也非常重要。我们把它专门写在这里。

定理 16.9: 聚点定理（Bolzano–Weierstrass）

实数集上的任意有界闭区间中的无限集合都有一个极限点（即聚点），或一个有界闭区间中的任意数列必有一子数列在其集合中收敛。

这里闭区间（例如 $[a, b]$，跟开区间 (a, b) 相比，包含边界点 a, b）就是为了保证极限点都在这个集合内部。之前我们已经证明了，有界的无穷数列保证有一个子数列收敛，也就是有极限点。合起来，就得到了上面的定理。

习题 16.16 (极限是有理数的无理数序列). 通过柯西收敛证明 $\sqrt{2+\sqrt{2+\sqrt{2+\cdots}}}$ 有极限。（提示：可以先算 $a_{n+1} - a_n$，然后再计算 $a_m - a_n$。注意运用平方差公式，例如 $a_2 - a_1 = \sqrt{2+\sqrt{2}} - \sqrt{2} = \frac{\left(\sqrt{2+\sqrt{2}}-\sqrt{2}\right)\left(\sqrt{2+\sqrt{2}}\sqrt{2}\right)}{\left(\sqrt{2+\sqrt{2}}+\sqrt{2}\right)} = \frac{\sqrt{2}}{\left(\sqrt{2+\sqrt{2}}+\sqrt{2}\right)}$。）

极限和实数学习阶段性小结

到此为止，我们已经定义了数列的极限，并且把数列的柯西收敛等同于数列有极限。按照数列的柯西收敛和有极限，我们把有极限的过程和做为其极限的数做了认同。从今往后，我们不再区分 $0.\dot{9}$ 和 1 了，尽管前者是过程后者是其极限。更一般地，我们说，一个数列和这个数列的极限，在极限的意义下是一个东西，也就是

$$\lim_{n\to\infty} a_n = a. \tag{16.311}$$

这不是说任意的 a_n 都是 a，甚至可能其中没有任何一个 a_n 真的等于 a，我们仅仅说 $|a_n - a|$ 会随着 n 越大而越来越小，或者说 $|a_m - a_n|$ 会随着 m, n 越大而越来越小。

请你给自己祝贺一下，吃点好吃的，或者干点好玩的。能从函数和四则运算这么一点初中知识走到极限和实数，已经是一件特别不容易的事情，能够体会到其中的人类智慧，体会到创造者们当年创造它们的时候痛苦和快乐，其中的思考和思想，已经是一个巨大的快乐（和痛苦，不可避免的，第一遍看不懂本书的部分内容没关系，先从外部来看，直到没个部分再做什么，起到什么作用，然后内部的多看几遍）。

有了数列的极限的概念，我们就可以来讨论自变量和函数的极限了。如果你跳过了关于戴德金分割的这一小节，你只需要知道下面的信息：数学家确实构建起来了一个叫做实数的数集，使得“柯西收敛则极限存在”成立，就行了。同时我们还证明了“极限存在则柯西收敛”。因此，以后我们把极限存在和柯西收敛完全等同。

极限不存在的数列

之前我们只展示了极限存在的数列，千万不要误以为数列都存在极限。为此，我们来看几个极限不存在的数列。

例 16.28 ($(-1)^n$ 的极限). 证明 $a_n = (-1)^n$ 的极限不存在。

按照定义，我们需要证明对于任何的 a，任何的 $\epsilon > 0$，不存在 N，使得 $n > N$ 之后，$|a_n - a| < \epsilon$。或者说，总有某一个 n 使得 $|a_n - a| > \epsilon$。

比如说，我们取 $a = -1, 0, 1$ 都可以得到这个结论。我们以 $a = 0$ 为例。$|a_n| = 1$，因此，只要 $\epsilon < 1$，我们就能找到（实际上所有的）n 使得

$|a_n - a| > \epsilon$。可是关键问题是，我们得针对所有的 a 来完成这个证明，否则完全有可能是我们没找到合适的 a。这个证明每发完成。幸好我们有柯西收敛和有极限等价。

于是，我们转而证明这个数列不满足柯西收敛性。我们发现，无论 n 取多少，相邻的两个元素 a_{n+1} 和 a_n 的差的绝对值永远是固定的

$$|a_{n+1} - a_n| = \left|(-1)^{n+1} - (-1)^n\right| = |(-1)^n|\,|(-1) - 1| = 2. \tag{16.312}$$

因此，只要 $\epsilon < 2$，我们就能找到（实际上所有的）n 使得 $|a_{n+1} - a_n| > \epsilon$，也就是存在 m, n 使得柯西收敛条件不满足。

例 16.29 (2^n 的极限)**.** 证明 $a_n = 2^n$ 的极限不存在。

我们直接证明这个数列不满足柯西收敛性。我们发现，无论 n 取多少，相邻的两个元素 a_{n+1} 和 a_n 的差的绝对值满足

$$\left|2^{n+1} - 2^n\right| = 2^n. \tag{16.313}$$

因此，只要 $\epsilon < 1$，我们就能找到（实际上所有的）n 使得 $|a_{n+1} - a_n| > \epsilon$，也就是存在 m, n 使得柯西收敛条件不满足。

例 16.30 (-2^n 的极限)**.** 证明 $a_n = -2^n$ 的极限不存在。

我们直接证明这个数列不满足柯西收敛性。我们发现，无论 n 取多少，相邻的两个元素 a_{n+1} 和 a_n 的差的绝对值满足

$$\left|-2^{n+1} + 2^n\right| = 2^n. \tag{16.314}$$

因此，只要 $\epsilon < 1$，我们就能找到（实际上所有的）n 使得 $|a_{n+1} - a_n| > \epsilon$，也就是存在 m, n 使得柯西收敛条件不满足。

这里出现的三种极限不存在的数列具体情形还是不太一样的。第一个属于在两个值之间跳来跳去，第二个属于要多大就有多大，第三个属于要多小就有多小。为了区分这样的数列，我们定义一个不同一般的极限——无穷大。

定义 16.4 (以无穷大为极限)**.** 如果一个数列 a_n 满足下面的条件，就说这个数列的极限是无穷大：对于任意的 $M > 0$，存在一个 N，使得对于任意的 $n > N$，有 $a_n > M$。记作 $\lim_{n\to\infty} a_n = \infty$。

定义 16.5 (以负无穷大为极限). 如果一个数列 a_n 满足下面的条件，就说这个数列的极限是负无穷大：对于任意的 $M>0$，存在一个 N，使得对于任意的 $n>N$，有 $a_n<-M$。记作 $\lim_{n\to\infty}a_n=-\infty$。

有了无穷大的正式的定义，我们终于可以名正言顺地使用 $n\to\infty$ 的记号了。可以证明前面的三个数列中第二个的极限是正无穷大，也称正无穷，第三个的极限是负无穷大，也称负无穷。证明留做作业。

习题 16.17 (2^n 的极限). 证明 $a_n=2^n$ 的极限是正无穷大。

习题 16.18 (-2^n 的极限). 证明 $a_n=-2^n$ 的极限是负无穷大。

注意，无穷大不是一个一般的数，是一个过程——一个要多大就有多大，比任何一个具体数都大的过程[17]。

函数和自变量的极限

我们说，在数列上，我们把有极限的过程和做为其极限的数——在极限的意义下——做了认同，也就是 $\lim_{n\to\infty}a_n=a$。现在，我们来把 Δx 和 0 做一个极限意义下的认同，

$$\lim_{\Delta x\to 0}\Delta x=0. \tag{16.315}$$

其中，完全可以没有任何一个具体的 $\Delta x=0$。我们参照数列极限的 $\epsilon-N$ 语言来建立上面的自变量趋于零的极限：对于任意的 $\epsilon>0$，存在着一个 $\delta=\epsilon>0$，使得 $-\delta<\Delta x<\delta$ 范围内的任何 Δx 满足 $|\Delta x-0|<\epsilon$。我们把这样的极限等于零的自变量的一个变化过程，称为无穷小。

这样的自变量的极限也可以看作是函数的极限，只不过这里的函数就是 $y=f(x)=x$，并且在 $x_0=0$ 处来讨论极限。更一般地，函数的极限可以这样来定义。

定义 16.6 (函数在某点的极限). 一个函数 $y=f(x)$ 在 x_0 点的极限 y_0 存在，记作 $\lim_{x\to x_0}f(x)=y_0$，如果满足下面的条件：对于任意的 $\epsilon>0$，存在着一个 $\delta>0$，使得任何 $x\in(x_0-\delta,x_0)\cup x\in(x_0,x_0+\delta)$ 满足 $|f(x)-y_0|<\epsilon$。

[17]这个时候，你就可以追问：前面我们有过程也就是数的例子，尤其是那个过程的极限存在的时候。那么，这里的这个极限为无穷大的过程，对应着一个数吗?

为了以后符号简单，我们记 $O(x_0,\delta)\triangleq(x_0-\delta,x_0)\cup(x_0,x_0+\delta)$，称为以 x_0 为中心的宽度为 δ 的去心邻域。如果把中心点 x_0 也包含进来，则记为 $\bar{O}(x_0,\delta)\triangleq(x_0-\delta,x_0+\delta)$，称为以 x_0 为中心的宽度为 δ 的邻域。

讨论函数极限的时候要去心，是因为决定是否有极限和极限的值的不是中心点这个点上的函数行为，而是周围的非中心的其他点的行为。稍后我们讨论连续的时候会把中心点的行为加回来。

我们把数列的柯西收敛也拓展到函数的柯西收敛。

定义 16.7 (函数的柯西收敛). 一个函数 $y=f(x)$ 在 x_0 点的柯西收敛，如果满足下面的条件：对于任意的 $\epsilon>0$，存在着一个 $\delta>0$，使得任何 $x_1,x_2\in O(x_0,\delta)$ 满足 $|f(x_1)-f(x_2)|<\epsilon$。

和数列的情形一样，我们也把函数的柯西收敛和函数的极限认同。

如果这里的 $y_0=f(x_0)$，我们称这个函数在 x_0 连续。

定义 16.8 (函数连续). 一个函数 $y=f(x)$ 在 x_0 点的极限 y_0 存在，并且这个值 $y_0=f(x_0)$，合起来也就是 $\lim_{x\to x_0}f(x)=f(x_0)$。

如果一个函数处处连续，我们称这个函数为一个连续函数。

我们来看几个例子。

例 16.31 (常值函数的极限). 证明常值函数 $f(x)=C$ 处处连续。

我们来证明对于任意的 x_0，$y=f(x)=C$ 的极限都存在，并且这个极限刚好就是 $f(x_0)=C$。对于任意的 $\epsilon>0$，我们只要取 $\delta=\epsilon>0$，就可以使得任何 $x\in O(x_0,\delta)$ 满足 $|f(x)-f(x_0)|=0<\epsilon$。

例 16.32 (x^n 的极限). 证明 $y=f(x)=x^n$ $(n>0,n\in\mathbb{N})$ 在 $x=0$ 点的极限是零。

对于任意的 $\epsilon>0$，我们只要取 $\delta=\epsilon^{\frac{1}{n}}>0$，就可以使得任何 $x\in O(0,\delta)$ 满足

$$|x^n-0^n|=|x|^n<\delta^n=\epsilon. \tag{16.316}$$

x 的正整数次方在 $x=0$ 点的极限是零是一个很重要的结论。

例 16.33 (x 处处连续). 证明 $y=f(x)=x$ 处处连续。

我们来证明对于任意的 x_0，$y=f(x)=x$ 的极限都存在，并且这个极限刚好就是 $f(x_0)=x_0$。对于任意的 $\epsilon>0$，我们只要取 $\delta=\epsilon>0$，就可以使得任何 $x\in O(x_0,\delta)$ 满足 $|x-x_0|<\epsilon$。

例 16.34 (x^2 处处连续)**.** 证明 $y=f(x)=x^2$ 处处连续。

我们来证明对于任意的 x_0，$y=f(x)=x^2$ 的极限都存在，并且这个极限刚好就是 $f(x_0)=x_0^2$。对于任意的 $\epsilon>0$，我们只要取 $\delta=\sqrt{x_0^2+\epsilon}-|x_0|>0$，就可以使得任何 $x\in O(x_0,\delta)$ 满足 $|x^2-x_0^2|<\epsilon$。

我们来检验一下最后的不等式是否成立，

$$\left|x^2-x_0^2\right|=|x+x_0|\,|x-x_0|\leqslant(2|x_0|+\delta)\,\delta, \tag{16.317}$$

而

$$\begin{aligned}(2|x_0|+\delta)\,\delta&=\left(2|x_0|+\sqrt{x_0^2+\epsilon}-|x_0|\right)\left(\sqrt{x_0^2+\epsilon}-|x_0|\right)\\&=\left(\sqrt{x_0^2+\epsilon}+|x_0|\right)\left(\sqrt{x_0^2+\epsilon}-|x_0|\right)\\&=\epsilon.\end{aligned} \tag{16.318}$$

合起来我们得到

$$\left|x^2-x_0^2\right|\leqslant\epsilon. \tag{16.319}$$

不过，这里有可能取到了等于号，和我们的定义不完全相符。未来我们会知道这个等于号无所谓。不过，在这里为了更严格，我们取 $\delta=\frac{\sqrt{x_0^2+\epsilon}-|x_0|}{2}>0$，来得到

$$\left|x^2-x_0^2\right|<\epsilon. \tag{16.320}$$

首先，我们发现 $g(\delta)=(2|x_0|+\delta)\,\delta-\epsilon=(\delta+|x_0|)^2-x_0^2-\epsilon$ 是一个二次函数，开口向上。因此，$g(\delta)$ 在其两个根 g_1,g_2 之间小于零。我们把两个根求出来，得到

$$\delta_{1,2}=-|x_0|\pm\sqrt{x_0^2+\epsilon}. \tag{16.321}$$

于是

$$g(\delta_2<\delta<\delta_2)<0. \tag{16.322}$$

而我们选的 $\delta=\frac{\sqrt{x_0^2+\epsilon}-|x_0|}{2}$ 确实处于两根之间，因此，

$$g(\delta)<0, \tag{16.323}$$

也就是

$$(2|x_0|+\delta)\,\delta<\epsilon. \tag{16.324}$$

举个具体的 $x_0=0$ 的例子，我们取 $\delta=\frac{\sqrt{x_0^2+\epsilon}-|x_0|}{2}=\frac{\sqrt{\epsilon}}{2}$，则 $|x^2-0^2|<\frac{\epsilon}{4}<\epsilon$ 确实成立。

例 16.35 ($|x|$ 处处连续)**.** 证明 $y=f(x)=|x|$ 处处连续。

对于 $x>0$，我们得到 $y=f(x)=x$，之前已经证明其连续。对于 $x<0$，我们得到 $y=f(x)=-x$。我们来证明对于任意的 x_0，对于任意的 $\epsilon>0$，我们只要取 $\delta=\epsilon>0$，就可以使得任何 $x\in O(x_0,\delta)$ 满足 $|(-x)-(-x_0)|=|x-x_0|<\epsilon$。

于是，我们只剩下 $x_0=0$ 的特殊情况需要证明。我们需要求解不等式 $||x|-|0||<\epsilon$，也就是 $|x|<\epsilon$。因此，我们只需要选 $\delta=\epsilon>0$ 就满足这个条件。

例 16.36 ($sign(x)$ 在 $x=0$ 处不连续)**.** 证明符号函数 $sign(x)$ 在 $x=0$ 处不连续。

符号函数的定义为

$$sign(x)=\begin{cases}1 & x>0\\ 0 & x=0\\ -1 & x<0\end{cases}. \tag{16.325}$$

我们来证明这个函数在在 $x=0$ 处不连续。

我们先来看看 $x=0$ 处，符号函数是否有极限，再来看看这个极限是否等于这个函数在这个点的值 $sign(0)=0$。

从 $x>0$ 的这一侧来看，对于 $x=0$，对于任意的 $\epsilon>0$，我们只要取 $\delta=\epsilon>0$，就可以使得任何 $x\in(0,\delta)$ 满足 $|sign(x)-1|=0<\epsilon$。但是，对于 $x<0$，我们有 $|sign(x)-1|=2$，只要 $\epsilon<2$，永远不满足极限的定义。同理，从 $x<0$ 的一侧来看，对于 $x=0$，对于任意的 $\epsilon>0$，我们只要取 $\delta=\epsilon>0$，就可以使得任何 $x\in(-\delta,0)$ 满足 $|sign(x)-(-1)|=0<\epsilon$。但是，对于 $x>0$，我们有 $|sign(x)-(-1)|=2$，只要 $\epsilon<2$，永远不满足极限的定义。也就是说，符号函数在 $x=0$ 处的极限不存在。更加就不用谈这个极限是不是正好就是 $sign(0)=0$ 的问题了。

不过这个例子很有意思，其实符号函数的单侧极限，如果我们分别定义 $x\in(x_0,x_0+\delta)$ 和 $x\in(x_0-\delta,x_0)$ 范围内满足不等式 $|f(x)-y_0|<\epsilon$ 为

单侧极限的话，是存在的。实际上，在数学分析课程中，确实是有单侧极限的概念的。不过，我们不用管它。

函数在某点不连续除了在那个点没有极限之外，还有虽然有极限，但是极限值不等于函数在那个点的取值的情况。例如，我们可以定义一个函数 $y=x^2$ 在 $x=0$ 点的值为 $y=4$。这不影响其仍然是一个函数，但是这个函数就不再是连续函数了。

习题 16.19 (a^x 连续). 证明函数 a^x $(a>1)$ 处处连续。

习题 16.20 (a^x 连续). 证明函数 a^x $(0<a<1)$ 处处连续。

几个数列和函数极限的性质定理

我们来证明几个未来会用到的数列和函数极限的性质定理。注意，这里的用到主要是在证明其他结论的时候用到。具体计算的时候，确实也会用到，但是，未来主要的极限计算都会由计算机完成。因此，这些定理主要是概念上的作用而不是计算上的作用。

定理 16.10: 数列极限的线性性

如果 a_n, b_n 有极限 a, b，则 $\lim_{n\to\infty}(\mu a_n+\nu b_n)=\mu a+\nu b$。

证明 16.3. 我们先假设 μ,ν 都不等于零。对于任意的 $\epsilon>0$，我们选择 $N=\max\{N_a,N_b\}$，也就是两者之中取大的，其中 $n>N_a$，则 $|a_n-a|<\frac{\epsilon}{2|\mu|}$，$n>N_b$，则 $|b_n-b|<\frac{\epsilon}{2|\nu|}$。于是，对于 $n>N$，

$$\begin{aligned}|(\mu a_n+\nu b_n)-(\mu a+\nu b)| &= |\mu(a_n-a)+\nu(b_n-b)|,\\ &\leqslant |\mu(a_n-a)|+|\nu(b_n-b)|,\\ &< |\mu|\frac{\epsilon}{2|\mu|}+|\nu|\frac{\epsilon}{2|\nu|}=\epsilon. \end{aligned} \tag{16.326}$$

当 μ,ν 其中有一个等于零的时候，证明很简单。我们就省略这个证明了。

我们还可以证明函数的极限的相应的线性性定理：

定理 16.11: 函数极限的线性性

如果 $f(x), g(x)$ 在 $x = x_0$ 有极限 a, b, 则 $\lim_{x\to x_0}(\mu f(x) + \nu g(x)) = \mu a + \nu b$。

在此我们也省略这个证明。

定理 16.12: 数列乘积的极限

如果 a_n, b_n 有极限 a, b, 则 $\lim_{n\to\infty} a_n b_n = ab$。

证明 16.4. 我们先假设 a, b 都不等于零。对于任意的 $\epsilon > 0$, 我们选择 $N = \max\{N_a, N_b\}$, 也就是两者之中取大的, 其中 $n > N_a$, 则 $|a_n - a| < \frac{\frac{\epsilon}{2}}{\frac{\epsilon}{2}|a|+|b|}$, $n > N_b$, 则 $|b_n - b| < \frac{\epsilon}{2|a|}$。于是, 对于 $n > N$,

$$\begin{aligned} |a_n b_n - ab| &= |a_n b_n - ab_n + ab_n - ab|, \\ &\leqslant |a_n - a|(|b| + \epsilon_b) + |a||b_n - b|, \\ &< \epsilon_a(|b| + \epsilon_b) + |a|\epsilon_b, \\ &= \epsilon. \end{aligned} \tag{16.327}$$

当 a, b 其中有一个等于零的时候, 例如 $a = 0, b \neq 0$, 我们取 $\epsilon_a = \frac{\epsilon}{\frac{\epsilon}{2}+|b|}$, $\epsilon_b = \frac{\epsilon}{2}$, 有

$$\begin{aligned} |a_n b_n| = |a_n||b_n| &\leqslant |a_n|(|b| + \epsilon_b), \\ &< \epsilon_a(|b| + \epsilon_b), \\ &= \epsilon. \end{aligned} \tag{16.328}$$

定理 16.13: 函数乘积的极限

如果 $f(x), g(x)$ 在 $x = x_0$ 有极限 a, b, 则 $\lim_{x\to x_0} f(x)g(x) = ab$。

在此我们也省略这个证明。

定理 16.14: 复合函数的极限

如果 $g(x)$ 在 $x=x_0$ 有极限 a,b，$f(x)$ 在 $x=a$ 有极限 b，则 $\lim_{x\to x_0} f(g(x))=b$。

证明 16.5. 对于任意的 $\epsilon>0$，我们选择合适的 δ_f（根据已知 $\lim_{x\to a} f(x)=b$，这个 δ_f 必然存在），使得对于任意的 $g\in O(a,\delta_f)$，$|f(g)-b|<\epsilon$，

根据 δ_f，我们选择合适的 δ_g（根据已知 $\lim_{x\to x_0} g(x)=a$，这个 δ_g 必然存在），使得对于任意的 $x\in O(x_0,\delta_g)$，$|g(x)-a|<\epsilon_f$。

于是，合起来，就有对于任意的 $\epsilon>0$，存在 δ_g，使得对于任意的 $x\in O(x_0,\delta_g)$，$|f(g(x))-b|<\epsilon$，也就是

$$\lim_{x\to x_0} f(g(x))=b. \tag{16.329}$$

16.8.2 从极限到导数和积分

有了数列和函数的极限的概念，我们终于可以来定义导数和积分了。

定义 16.9 (导数). 对于一个函数 $f(x)$，如果在 x_0 点，下面的极限存在，

$$\lim_{\Delta x\to 0}\frac{f(x_0+\Delta x)-f(x_0)}{\Delta x} \tag{16.330}$$

则就称这个函数在 x_0 点可导，这个极限就称为函数 $f(x)$ 在 x_0 点的导数，记作 $f'(x_0)$，也记作 $\left.\frac{df(x)}{dx}\right|_{x_0}$，$\left.\frac{df}{dx}\right|_{x_0}$。如果一个函数（在某个区间内）处处都存在导数，就称这个函数（在这个区间内）处处可导。这个时候每个点的导数就称为原函数 $f(x)$ 的导函数，记作 $f'(x)$，也记作 f'_x，$\frac{df(x)}{dx}$，$\frac{df}{dx}$，$\left.\frac{df}{dx}\right|_x$。

按照Newton的记号，在物理学中，当自变量是时间的时候，也记为 $\dot{f}(t)$，例如速度函数是位置函数对时间的导数，$v(t)=\dot{x}(t)$。

注意，这个极限是对 Δx 做的，而不是对 x_0 做的。我们来看几个例子。

例 16.37 (x^n 的导数). 证明 x^n（n 为正整数）在 $x=0$ 导数存在。我们要检查下面的极限是否存在，

$$\lim_{\Delta x\to 0}\frac{(0+\Delta x)^n-(0)^n}{\Delta x}=\lim_{\Delta x\to 0}(\Delta x)^{n-1}. \tag{16.331}$$

当 $n=1$ 的时候，这是一个常值函数 $(\Delta x)^{n-1}=1$。我们之前已经证明了常值函数极限处处存在。因此，满足要求。当 n 是大于 1 的正整数的时候，$(n-1)$ 是大于零的整数，之前我们也已经证明，其在 $x=0$ 点的极限存在，并且其值为零。因此，合起来，$\lim_{\Delta x\to 0}(\Delta x)^{n-1}$ 存在。也就是 x^n（n 为正整数）在 $x=0$ 导数存在。

顺便这里得到的两个结论会在下面用到。

$$\lim_{\Delta x\to 0}(\Delta x)^n=\delta_{n,0}. \tag{16.332}$$

其中 $\delta_{n,0}$ 称为Kronecker δ 记号，其定义为

$$\delta_{n,0}=\begin{cases}1 & n=0,\\ 0 & \text{others.}\end{cases} \tag{16.333}$$

我们再来看几个之前用不严格的导数三步法得到过的结果。

例 16.38 (x^n 的导数)**.** 证明 x^n（n 为正整数）处处可导。我们要检查下面的极限是否存在，

$$\begin{aligned}\lim_{\Delta x\to 0}\frac{(x+\Delta x)^n-(x)^n}{\Delta x}&=\lim_{\Delta x\to 0}\left[\sum_{l=0}^{n}C_n^l x^{n-l}(\Delta x)^{l-1}\right],\\ &=\sum_{l=0}^{n}\left[C_n^l x^{n-l}\lim_{\Delta x\to 0}(\Delta x)^{l-1}\right],\\ &=\sum_{l=0}^{n}C_n^l x^{n-l}\delta_{l-1,0},\\ &=nx^{n-1}.\end{aligned} \tag{16.334}$$

这里我们用到了极限的线性性。注意，线性性对于有限项求和显然成立，但是对于无限项求和不一定成立。

有了导数的定义，结合极限的线性性，乘积函数的极限，我们就很容易严格证明之前得到导数计算的定理，而不再是依靠不严格的导数计算三步法。下面，我们来证明其中的一个，例如乘积函数的导数的莱布尼茨法则。

定理 16.15: 乘积函数的导数——莱布尼茨法则

证明 $[f(x)g(x)]' = f(x)'g(x) + f(x)g'(x)$。

证明.

$$\begin{aligned}
[f(x)g(x)]' &= \lim_{\Delta x\to 0}\frac{[f(x+\Delta x)\,g(x+\Delta x)]-[f(x)\,g(x)]}{\Delta x}\\
&= \lim_{\Delta x\to 0}\frac{f(x+\Delta x)\,g(x+\Delta x)-f(x)\,g(x+\Delta x)}{\Delta x}\\
&\qquad + \lim_{\Delta x\to 0}\frac{f(x)\,g(x+\Delta x)-f(x)\,g(x)}{\Delta x},\\
&= \lim_{\Delta x\to 0}\frac{f(x+\Delta x)-f(x)}{\Delta x}g(x+\Delta x)\\
&\qquad + \lim_{\Delta x\to 0}f(x)\,\frac{g(x+\Delta x)-g(x)}{\Delta x},\\
&= f(x)'g(x)+f(x)g'(x). \qquad (16.335)
\end{aligned}$$

其中我们用到了乘积函数的极限以及极限的线性性。 □

并不是所有的函数都是处处可导的。下面我们来看一个在某个点不可导的函数。

例 16.39 ($|x|$ 的导数). 证明 $|x|$ 在 $x=0$ 导数不存在。我们要检查下面的极限是否存在，

$$\lim_{\Delta x\to 0}\frac{|0+\Delta x|-|0|}{\Delta x}=\lim_{\Delta x\to 0}\frac{|\Delta x|}{\Delta x}. \qquad (16.336)$$

当 $\Delta x<0$ 的时候，$\frac{|\Delta x|}{\Delta x}=-1$，而当 $\Delta x>0$ 的时候，$\frac{|\Delta x|}{\Delta x}=1$。也就是说，只要我们选择 $\Delta x_1>0, \Delta x_2<0$，我们可是使得

$$\left|\frac{|\Delta x_1|}{\Delta x_1}-\frac{|\Delta x_2|}{\Delta x_2}\right|=2. \qquad (16.337)$$

因此，函数 $\frac{|\Delta x|}{\Delta x}$ 在 $\Delta x=0$ 点不满足柯西收敛的条件，也就是说函数 $\frac{|\Delta x|}{\Delta x}$ 在 $\Delta x=0$ 点没有极限。

于是，我们得到函数 $|x|$ 在 $x=0$ 点导数不存在，不可导。

很容易证明，对于 $x>0$ 和 $x<0$，$|x|$ 是可导的。

我们之前已经证明函数 $|x|$ 处处连续。现在，我们看到这个函数并不处处可导。至于更细节的连续和可导之间的关系是什么，可导的充分必要条

件是什么，我们就不讨论了。不过这里有一个有趣的知识——竟然存在处处连续处处不可导的函数。有兴趣的可以自己去了解一下这样的函数。

可以证明，一个在 x_0 可导的函数肯定在 x_0 连续，但是连续函数不一定可导。直观上，连续可以是折线，可导必须是比较光滑地连起来的，折线不可导。实际上，我们还可以基于可导来定义光滑程度，例如一个函数是一阶可导但是二阶不可导的，表示函数本身是连续的而且处处都不是折线，但是这个函数的导函数就会出现折线。

定义 16.10 (光滑函数). 一个连续函数被称为 C^0 阶光滑函数，一个 n 阶可导的函数被称为 C^n 阶光滑函数，一个无穷多阶可导的函数被称为 C^∞ 阶光滑函数。后者也被简称为光滑函数。

我们来证明一下一个可导的函数必定连续。

定理 16.16: 可导函数连续

证明在 x 点可导的函数 $f(x)$ 必然在 x 点连续。

证明 16.6. 函数 $f(x)$ 在 x 点可导的含义是极限 $f'(x) = \lim_{\Delta x\to 0}\frac{f(x+\Delta x)-f(x)}{\Delta x}$ 存在，是一个有限大小的值，也就是存在实数 M，使得 $|f'(x)| < M$。于是，对于任意的 $\epsilon > 0$，存在 $\delta > 0$，使得任意的 $\Delta x \in \bar{O}(0,\delta)$，满足

$$\begin{aligned}
&\left|\frac{f(x+\Delta x)-f(x)}{\Delta x} - f'(x)\right| < \epsilon \\
\Rightarrow &\left|\frac{f(x+\Delta x)-f(x)}{\Delta x}\right| < |f'(x)| + \epsilon \\
\Rightarrow &\left|\frac{f(x+\Delta x)-f(x)}{\Delta x}\right| < M + \epsilon \\
\Rightarrow &|f(x+\Delta x)-f(x)| < (M+\epsilon)|\Delta x| \qquad (16.338)
\end{aligned}$$

因此，只要选择一个新的 $\delta' = \min\left(\delta, \frac{\delta}{M+\epsilon}\right)$，就能使得对于任意的 $\Delta x \in \bar{O}(0,\delta')$，$|f(x+\Delta x)-f(x)| < \epsilon$。因此，函数 $f(x)$ 在 x 点连续。

我们再来通过极限定义定积分。

定义 16.11 (定积分). 对于一个函数 $f(x)$ 和一个区间 $[a,b]$，我们定义下面的分割[18]和取和的过程：

- 把区间 $[a,b]$ 分割成 n 份，每一份的大小为 $\Delta x = \frac{b-a}{n}$。
- 在每一个分割好的小格子内按照下面的长方形来计算这个格子的面积，$S_l = f(x_l)\,\Delta x$。
- 把所有小格子的面积加起来，得到

$$S = \sum_{l=0}^{n-1} f(x_l)\,\Delta x. \tag{16.339}$$

如果这个过程的极限存在，我们就说函数 $f(x)$ 在一个区间 $[a,b]$ 可积，记为

$$\int_a^b f(\xi)\,d\xi = \lim_{n\to\infty} \sum_{l=0}^{n-1} f(x_l)\,\Delta x. \tag{16.340}$$

我们来看几个例子。

例 16.40 (常数函数的积分). 证明 $f(x)=1$ 在区间 $[a,b]$ 上的积分存在，并且求出来。

对于常数函数，

$$S_n = \sum_{l=0}^{n-1} f(x_l)\,\Delta x = \sum_{l=0}^{n-1} \Delta x = \sum_{l=0}^{n-1} \frac{b-a}{n} = (b-a). \tag{16.341}$$

于是

$$\int_a^b d\xi = (b-a). \tag{16.342}$$

这个例子还有一个额外的好处，看起来 dx 的含义有了一个说法：$\sum_{l=0}^{n-1} \Delta x$ 在极限下变成了 $\int_a^b dx$。如果这个不仅仅是一个记号，而是 $\int$ 和 dx 的意义的话，这意味着 Δx 的极限下成了 dx，$\sum$ 在极限下成了 $\int$。会是这样吗？我们能给 dx 一个含义吗？

[18] 其实可以允许满足一定条件的更一般的分割过程。原则上还得证明不同的分割过程不影响极限存在和极限的值。不过，那部分内容留给数学分析去完成。

之前，我们提到过 dx 叫做 x 的微分，提到过微元法——把所要研究的对象先分成一个个的 Δx，然后，得到关于 Δx 的表达式之后，把 Δx 看作 dx 来计算。

之前，我们还证明了逆函数的导数是原来的函数的导数的倒数，也就是对于 $y=f(x)$，$y'_x=\frac{dy}{dx}$，而 $x=f^{-1}(y)$ 的导数为 $x'_y=(y'_x)^{-1}=\frac{dx}{dy}$。

更进一步，如果 dx 和 dy 可以分开来写，有自己独立的含义，那么，对于 $y=(x)$，我们就有了

$$df=dy=f'(x)\,dx. \tag{16.343}$$

而根据 Taylor 公式 $\Delta y=f(x+\Delta x)-f(x)=f'(x)\,\Delta x+\frac{1}{2}f''(x)\,\Delta^2 x+\cdots$，这个完全可以看作是保留到 Δx 的一阶的

$$\Delta f=\Delta y=f'(x)\,\Delta x \tag{16.344}$$

在极限下的版本。公式 (16.344) 被称为对一个函数 $f(x)$ 来说其差 $f(x+\Delta x)-f(x)$ 的线性主部。从这个角度来看，dx 就对应着 $f(x)=x$ 这个函数的差 $f(x+\Delta x)-f(x)$ 的线性主部。这样，通过线性主部的概念，加上极限，似乎 df 和 dx 有了统一的含义[19]

这些都使得 dx 和 dy 看起来很像有自己的含义，而不仅仅是符号。没准到了我们给它们自己的含义的时候了？不过，在那之前，无论如何，在上面的求导数和求积分的场合你可以像通常的表达式一样来独立地使用 dx 和 dy 了。

例 16.41 ($f(x)=x$ 的积分)**.** 证明 $f(x)=x$ 在区间 $[a,b]$ 上的积分存在，并且求出来。

[19]但是，其实，我们并没有解释清楚，为什么保留到线性主部就够了，而且用了极限未来也不影响算积分。我们稍后会再一次回到这个问题。

对于 $f(x)=x$，

$$\begin{aligned} S_n &= \sum_{l=0}^{n-1} f(x_l)\,\Delta x = \frac{b-a}{n}\sum_{l=0}^{n-1} x_l, \\ &= \frac{b-a}{n}\sum_{l=0}^{n-1}(l\Delta x + a) = \frac{b-a}{n}\sum_{l=0}^{n-1} l\Delta x + \frac{b-a}{n}\sum_{l=0}^{n-1} a, \\ &= \left(\frac{b-a}{n}\right)^2 \sum_{l=0}^{n-1} l + (b-a)\,a, \\ &= \left(\frac{b-a}{n}\right)^2 \frac{n(n-1)}{2} + (b-a)\,a, \\ &= \frac{(b-a)^2}{2}\frac{n(n-1)}{n^2} + (b-a)\,a. \end{aligned} \tag{16.345}$$

其中待检验的极限部分为

$$\lim_{n\to\infty}\frac{n(n-1)}{n^2} = \lim_{n\to\infty}\left[1-\frac{1}{n^2}\right] = 1. \tag{16.346}$$

因此，S_n 的 $n\to\infty$ 的极限存在，按照积分的定义

$$\int_a^b \xi d\xi = \frac{(b-a)^2}{2} + (b-a)\,a = \frac{b^2-a^2}{2}. \tag{16.347}$$

$f(x)=1$ 的原函数为 $F(x)=x$，$f(x)=x$ 的原函数为 $F(x)=\frac{1}{2}x^2$。通过这两个例子，我们已经看到了似乎 $f(x)$ 定积分和其原函数 $F(x)$ 之间存在联系，

$$\int_a^b f(\xi)\,d\xi = F(b) - F(a). \tag{16.348}$$

注意，如果 dF, dx 有了自己的含义，并且对于原函数 $F(x)$ 有 $dF = f(\xi)\,d\xi$，上面的公式可以写成，

$$\int_a^b f(\xi)\,d\xi = \int_{F(a)}^{F(b)} dF. \tag{16.349}$$

于是这就是一件非常自然的事情，是一个积分变量的替换，而不是一个有待证明的公式、定理。反过来说，当我们证明了这个公式之后，我们确实就可以把 dF, dx 当作有自己独立的含义的东西来看了，而不再仅仅是用来记导数（$\frac{dF}{dx}$）和积分（$\int_a^b f(x)dx$）的符号了。

就像我们在下一节即将展示的，这个公式是可以被证明的。因此，我们也提前给了 dF, dx 它们自己的名称——微分，函数的微分和自变量的微分。未来，我们证明这个命题之后，我们就可以真的像通常的表达式一样来独立地使用 dF 和 dx 了[20]。

我们再来看几个例子。

例 16.42 ($f(x)=x^2$ 的积分). 证明 $f(x)=x^2$ 在区间 $[a,b]$ 上的积分存在，并且求出来。

对于 $f(x)=x^2$，

$$\begin{aligned} S_n &= \sum_{l=0}^{n-1} f(x_l)\,\Delta x = \frac{b-a}{n}\sum_{l=0}^{n-1} x_l^2, \\ &= \frac{b-a}{n}\sum_{l=0}^{n-1}(l\Delta x + a)^2, \\ &= \frac{b-a}{n}\sum_{l=0}^{n-1}\left[l^2\Delta^2 x + 2al\Delta x + a^2\right], \\ &= \left(\frac{b-a}{n}\right)^3\sum_{l=0}^{n-1} l^2 + 2a\left(\frac{b-a}{n}\right)^2\sum_{l=0}^{n-1} l + \left(\frac{b-a}{n}\right)a^2\sum_{l=0}^{n-1} 1. \end{aligned} \tag{16.350}$$

其中三个取和分别计算出来，得到

$$\sum_{l=0}^{n-1} 1 = n, \tag{16.351}$$

$$\sum_{l=0}^{n-1} l = \frac{n(n-1)}{2}, \tag{16.352}$$

$$\sum_{l=0}^{n-1} l^2 = \frac{n(n+1)(2n+1)}{6}. \tag{16.353}$$

其中新的待检验的极限部分为

$$\lim_{n\to\infty}\frac{2n^3+3n^2+n}{6n^3} = \lim_{n\to\infty}\left[\frac{1}{3}+\frac{1}{2n}+\frac{1}{6n^2}\right] = \frac{1}{3}. \tag{16.354}$$

[20]不过，尽管这里我们把 dF, dx 叫做微分，又通过极限和线性主部联系了起来，又可以联系到积分，但是其真正含义只有未来在一门叫做微分几何的课程中才能看明白：它们的名字叫做微分形式。

其他两个极限已经在之前的计算中得到，

$$\lim_{n\to\infty}\frac{n^2-n}{2n^2}=\frac{1}{2}, \tag{16.355}$$

$$\lim_{n\to\infty}\frac{n}{n}=1. \tag{16.356}$$

因此，S_n 的 $n\to\infty$ 的极限存在，

$$\lim_{n\to\infty}S_n=(b-a)^3\frac{1}{3}+2a(b-a)^2\frac{1}{2}+(b-a)a^2=\frac{1}{3}\left(b^3-a^3\right). \tag{16.357}$$

按照积分的定义

$$\int_a^b\xi^2d\xi=\frac{b^3-a^3}{3}. \tag{16.358}$$

$f(x)=x^2$ 的原函数为 $F(x)=\frac{1}{3}x^3$。看起来积分和原函数之间的联系真的存在。

例 16.43 ($f(x)=\cos x$ 的积分)**.** 证明 $f(x)=\cos x$ 在区间 $[a,b]$ 上的积分存在，并且求出来。

对于 $f(x)=\cos x$，

$$\begin{aligned}S_n=\sum_{l=0}^{n-1}f(x_l)\Delta x&=\frac{b-a}{n}\sum_{l=0}^{n-1}\cos x_l,\\&=\frac{b-a}{n}\sum_{l=0}^{n-1}\cos(l\Delta x+a),\\&=\frac{\frac{b-a}{n}}{\sin\frac{b-a}{n}}\sum_{l=0}^{n-1}\cos(l\Delta x+a)\sin\Delta x,\\&=\frac{1}{2}\frac{\frac{b-a}{n}}{\sin\frac{b-a}{n}}\sum_{l=0}^{n-1}\{\sin[(l+1)\Delta x+a]-\sin[(l-1)\Delta x+a]\}.\end{aligned} \tag{16.359}$$

其中大括号内的大部分项都会相互削去，就剩下

$$\begin{aligned}&\sin[N\Delta x+a]+\sin[(N-1)\Delta x+a]-\sin[-\Delta x+a]-\sin[a],\\&=\sin b+\sin(b-\Delta x)-\sin(a-\Delta x)-\sin a.\end{aligned} \tag{16.360}$$

用上

$$\lim_{x\to0}\frac{x}{\sin x}=1, \tag{16.361}$$

$$\lim_{\Delta x\to0}\sin(x+\Delta x)=\sin x. \tag{16.362}$$

之后，我们得到 S_n 的 $n \to \infty$ 的极限存在，

$$\lim_{n\to\infty} S_n = \sin b - \sin a. \tag{16.363}$$

按照积分的定义

$$\int_a^b \cos\xi d\xi = \sin b - \sin a. \tag{16.364}$$

$f(x) = x^2$ 的原函数为 $F(x) = \frac{1}{3}x^3$，$\cos x$ 的原函数是 $\sin x$。看起来积分和原函数之间的联系真的存在。

顺便，历史上，少数的这些能够完全用初等数学的方式计算出来的极限和积分对于微积分的发展的推动作用也是很强的。希望你也从中看到数列（整个极限和积分都建立在数列的概念之上）、函数增减性（讨论极限的时候常用）的作用。很大程度上，当前的初等数学的教学就是为了给学习微积分做铺垫的。

那我们能不能证明积分和原函数之间的这样的关系呢？下面我们来完成这个证明。

16.8.3 微积分基本定理

这即将是你的另一场努力和对人类智慧的瑰宝的欣赏。这是数学家们的艰苦卓绝的努力。

我们从戴德金定理、费马（Fermat）定理和聚点定理（Bolzano–Weierstrass）开始。戴德金定理告诉我们任何实数上的戴德金分割(L, U) 必然会出现和仅仅出现以下两种情况之一：L 内有最大值 U 中没有最小值，或者 U 内有最小值 L 中没有最大值。费马（Fermat）定理告诉我们对于可导函数 $f(x)$ 其极值点 x_* 的导数为零 $f'(x_*) = 0$。聚点定理（Bolzano–Weierstrass）说的是有界闭区间中的无限集合都有一个极限点，或者说有界闭区间中的数列必然存在有极限的子序列。我们现在来用这两个定理以及连续函数的定义证明确界原理和罗尔（Rolle）引理。

一个集合 A 的上界 M 的定义是 $x \leqslant M, \forall x \in A$。一个集合 A 的下界 m 的定义是 $x \geqslant m, \forall x \in A$。一个集合 A 的上确界 M_{min} 的定义是 M_{min} 是 A 的最小的上界，也就是说 $x \leqslant M_{min}, \forall x \in A$，同时对于任意的 $\epsilon > 0$，存在 $x \in A$ 使得 $x > M_{min} - \epsilon$。一个集合 A 的下确界 m_{Max} 的定义是 m_{Max}

是 A 的最大的下界，也就是说 $x \geqslant m_{Max}, \forall x \in A$，同时对于任意的 $\epsilon > 0$，存在 $x \in A$ 使得 $x < m_{Max} + \epsilon$。

定理 16.17: 确界原理

非空有界集合必定有确界。

证明 16.7. 我们仅仅证明上确界，下确界同理可得。

A 为非空有界集合，表示存在 $a \in A$，并且存在 M，使得对于任意的 $x \in A$，$x \leqslant M$，存在 m，使得对于任意的 $x \in A$，$x \geqslant m$。我们把所有的 A 的上界合起来构成集合 U。定义 $L = \bar{U}$，全集为实数集 $\mathbb{R}$。我们来证明 (L,U) 是一个戴德金分割。

首先，$L \cap U = \phi, L \cup U = \mathbb{R}$。其次，我们来证明 U, L 非空。任意取 $\epsilon > 0$，$m - \epsilon \in L$（否则 $m - \epsilon \in U$，也就是 $m - \epsilon$ 是 A 的上界,$x \leqslant m - \epsilon$，也就是 $x + \epsilon < m$，这个和 $x \geqslant m$ 矛盾）。当然，$M \in U$。接着，我们来证明任意 $l \in L, u \in U, l < u$。用数学语言表示 U 为

$$U = \{u | \forall x \in A, x \leqslant u\}, \tag{16.365}$$

于是其补集的定义为不满足上面的条件的所有实数，

$$L = \{l | \exists x \in A, x > l\}. \tag{16.366}$$

因此，对于 $l \in L, u \in U$ 肯定存在一个 $x \in A$，使得

$$l < x < u, \tag{16.367}$$

也就是 $l < u$。于是 (L,U) 是一个戴德金分割。其中要么 L 有最大值，要么 U 有最小值。

我们假设 L 中有最大值 η。也就是对于任意的 $\epsilon > 0$，$\eta + \epsilon \in U$。根据 U 的定义，这意味着对于任意的 $x \in A$，$x \leqslant \eta + \epsilon$。同时，根据 L 的定义，存在 $x \in A$ 使得 $x > \eta$。对于任意的 $\epsilon > 0$，$x \leqslant \eta + \epsilon$ 和 $x > \eta$ 不可能同时成立，例如取 $\epsilon = \frac{x-\eta}{2} > 0$，就可以得到 $x \leqslant \eta + \epsilon = \eta + \frac{x-\eta}{2}$，也就是

$$2\epsilon = x - \eta \leqslant \frac{x-\eta}{2} = \epsilon \Rightarrow \epsilon \leqslant 0. \tag{16.368}$$

这显然不成立。

因此，只能出现另一种情况，U 中有最小值。U 中任何元素都是 A 的上界，其最小值就是上确界。因此，A 存在上确界。

有了确界原理，我们来证明罗尔（Rolle）引理。

定理 16.18: 罗尔（Rolle）引理

对于闭区间 $[a,b]$ 上连续，开区间 (a,b) 上可导的函数 $f(x)$，如果有 $f(a)=f(b)$，则存在 $\xi\in(a,b)$ 使得 $f'(\xi)=0$。

证明 16.8. 根据连续函数的定义，对于任意的 $x\in[a,b]$，我们有 $\lim_{\xi\to x}=f(x)$，也就是 $\forall\epsilon>0$，$\exists\delta>0$，使得

$$|f(\xi)-f(x)|<\epsilon\Rightarrow|f(\xi)|<|f(x)|+\epsilon \tag{16.369}$$

让我们比如说从取 $x=a$ 开始，则我们就得到了在 $\xi\in\bar{O}(a,\delta_a)$（这里仅仅在 a 的右侧管用，因为 a 的左侧这个函数是否连续甚至有没有定义都没有交代）内的所有 $f(\xi)$ 有界。接着我们选这个邻域内的另一个点当作 x：第一、这个点的 $f(x)$ 存在并且有界，第二，函数在这个点连续。重复上面的证明就可以找到一个 $\bar{O}(x,\delta_x)$，使得这个邻域内的所有 $\xi\in\bar{O}(a,\delta_a)$ 满足 $f(\xi)$ 有界。因为邻域总是两侧延展的，除了在端点 a,b（只能而且只有必要在一侧延展），因此，可以把这个证明过程拓展到整个闭区间 $[a,b]$。

既然闭区间上的连续函数是有界的，根据确界原理，其必定有上下确界也就是最大值和最小值。我们不妨记这个最大值和最小值为 M,m。我们来证明这个最大值 M 肯定至少能够找到一个 $\eta\in[a,b]$ 使得 $f(\eta)=M$。

首先，根据 M 为最大值，得到 $\forall x\in[a,b]$，$f(x)\leqslant M$。其次，根据 M 为上确界得到，$\forall\epsilon>0$，存在 $f(\eta)>M-\epsilon$。合起来，$\forall\epsilon>0$，$M-\epsilon<f(\eta)<M+\epsilon$。因此，$\lim_{x\to\eta}f(x)=M$，根据连续性，$f(\eta)=\lim_{x\to\eta}f(x)=M$。

同理，这个结论对最小值也同样成立：肯定至少能够找到一个 $\eta\in[a,b]$ 使得 $f(\eta)=m$。

根据最大值和最小值的含义，我们肯定有 $m\leqslant f(a)=f(b)\leqslant M$。这时候我们来看是否取等号。这里有几种情况：第一，第一个不等式取等号但是第二个不等式只取不等号；第二，第二个不等式取等号但是第一个不等式只取不等号；第三，两个不等式都取不等号。

对于第三种情况，则 $f(x)=f(b)$，是一个常值函数。常值函数的导数处处为零。因此，命题成立。

对于第一（第二）种情况，肯定存在一个 $x_*\neq a,x_*\neq b$，但是仍然 $x_*\in[a,b]$，也就是 $x_*\in(a,b)$，使得 $f(x)=M$（$f(x)=m$）。根据费马（*Fermat*）定理，必然有 $f'(x_*)=0$。因此，命题成立。

我们用罗尔（Rolle）引理来证明拉格朗日（Lagrange）微分中值定理。

定理 16.19: 拉格朗日（Lagrange）微分中值定理

对于闭区间 $[a,b]$ 上连续，开区间 (a,b) 上可导的函数 $f(x)$，$\xi \in (a,b)$ 使得 $f'(\xi) = \frac{f(b)-f(a)}{b-a}$。

证明 16.9. 取 $g(x) = f(x) - \frac{f(b)-f(a)}{b-a}x$，则 $g(a) = \frac{bf(a)-af(b)}{b-a} = g(b)$，同时 $g(x)$ 继承了 $f(x)$ 对于闭区间 $[a,b]$ 上连续，开区间 (a,b) 上可导的性质。根据罗尔（Rolle）引理，存在 $\xi \in (a,b)$ 使得 $g'(\xi) = 0$，也就是 $f'(\xi) = \frac{f(b)-f(a)}{b-a}$。

有了拉格朗日（Lagrange）微分中值定理，我们就可以建立起来微分和积分之间的联系了。

考虑在取极限之前的积分表达式 $\sum_{l=0}^{n-1} f(x_l)\Delta x$ 的每一项，$f(x_l)\Delta x = f(x_l)(x_l - x_{l-1})$。$f(x)$ 的原函数为 $F(x) = \int f(x)dx$，也就是 $F'(x) = f(x)$，根据拉格朗日（Lagrange）微分中值定理，有 $\xi_l \in (x_{l-1}, x_l)$ 使得

$$F(x_l) - F(x_{l-1}) = f(\xi_l)\Delta x. \tag{16.370}$$

于是

$$F(b) - F(a) = \sum_{l=0}^{n-1}\left(F(x_{l+1}) - F(x_l)\right) = \sum_{l=0}^{n-1} f(\xi_l)\Delta x. \tag{16.371}$$

这里我们用了 $a = x_0, x_n = b$。如果我们能证明在 $\Delta x \to 0$ 也就是 $n \to \infty$ 的极限下，

$$\lim_{n\to\infty}\sum_{l=0}^{n-1} f(\xi_l)\Delta x = \lim_{n\to\infty}\sum_{l=0}^{n-1} f(x_l)\Delta x. \tag{16.372}$$

我们就有牛顿-莱布尼兹公式，这个沟通定积分和导数（也就是不定积分）之间的微积分基本定理了，

$$\int_a^b f(\xi)\,d\xi = F(b) - F(a). \tag{16.373}$$

下面我们来计算 $\sum_{l=0}^{n-1}\left[f(\xi_l) - f(x_l)\right]\Delta x$。因为 $f(x)$ 是连续函数，对于任意的 $\epsilon > 0$，我们总是可以找到 $\delta_l > 0$，使得 $\xi_l \in \bar{O}(x_l, \delta_l)$，$|f(\xi_l) - f(x_l)| <$

$\frac{\epsilon}{b-a}$。对于给定的 n 下的每一组分割，我们都找出来这样的 δ_l 中最小的当作 δ。这样，我们就有了

$$\left|\sum_{l=0}^{n-1}\left[f\left(\xi_l\right)-f\left(x_l\right)\right]\Delta x\right| \leqslant \sum_{l=0}^{n-1}\left|f\left(\xi_l\right)-f\left(x_l\right)\right|\Delta x < \frac{\epsilon}{b-a}\sum_{l=0}^{n-1}\Delta x = \epsilon. \tag{16.374}$$

唯一剩下的问题就是这样的 δ 是不是会没有已经通过拉格朗日（Lagrange）微分中值定理确定的 ξ_l 包含在 $\bar{O}\left(x_l,\delta_l\right)$ 里面。这个问题怎么解决呢?

如果函数是一致连续——能够保证对于任意的 ϵ 所有的 $\bar{O}\left(x,\delta\right)$ 的邻域内的任意两个点 ξ,x 都满足 $\left|f\left(\xi\right)-f\left(x\right)\right|<\epsilon$——就好办了。对于任意的 ϵ，先找到这个 δ 划分好这样大小为 δ 的邻域，在每一个邻域内在运用拉格朗日（Lagrange）微分中值定理来确定 ξ_l。

我们来补充一个定理：闭区间上的连续函数一致连续。

定理 16.20: 闭区间上一致连续

闭区间 $[a,b]$ 上的连续函数 $f\left(x\right)$ 在这个闭区间上一致连续。

证明 16.10. 假设 $[a,b]$ 上的连续函数 $f\left(x\right)$ 在 $[a,b]$ 上非一致连续，则 $\exists\varepsilon>0,\forall\delta>0,\exists x_1,x_2\in[a,b],\left|x_1-x_2\right|<\delta$ 但是 $\left|f\left(x_1\right)-f\left(x_2\right)\right|\geqslant\varepsilon$。我们取一个 δ_n 的数列满足 $\lim_{n\to\infty}\delta_n=0$，对于每一个这样的 δ_n 的取值，我们把找到的 x_1,x_2 记下来，x_1 构成数列 $\{a_n\}$，x_2 构成数列 $\{b_n\}$。则必然有 $\lim_{n\to+\infty}\left(a_n-b_n\right)=0$。但是，每一个 n，都有 $\lim_{n\to\infty}\left|f\left(a_n\right)-f\left(b_n\right)\right|\geqslant\varepsilon$。

根据聚点定理（*Bolzano–Weierstrass*），有界闭区间上的数列必然存在有极限的子数列，所以 $\{a_n\},\{b_n\}$ 存在收敛子列 $\{a_{n_k}\},\{b_{n_k}\}$，设其分别收敛到 p,q。$\lim_{k\to\infty}\left(a_{n_k}-b_{n_k}\right)=0$，所以 $p=q$。对于闭区间，极限点必然属于闭区间本身，因此 $p,q\in[a,b]$。

根据 $f\left(x\right)$ 的连续性，$\lim_{k\to\infty}\left[f\left(a_{n_k}\right)-f\left(b_{n_k}\right)\right]=f\left(p\right)-f\left(q\right)=0$。这个和每一组 $\left(x_1,x_2\right)$ 都有 $\left|f\left(x_1\right)-f\left(x_2\right)\right|\geqslant\varepsilon$ 矛盾——每一对 $\left(a_{n_k},b_{n_k}\right)$ 不过也就是前面的 $\left(x_1,x_2\right)$ 中的一个。因此，原假设错误，一致收敛必然成立。

注意，在开区间上这个命题就不成立。例如，我们可以考虑一个 $(0,1)$ 上的函数 $\frac{1}{x}$，是连续的但是不是一致连续的。

其实还有一个更加直观的方式来理解上面的结果。我们假设 $f(x)$ 处处可导，那么，$|f(\xi_l)-f(x_l)|=f'(\eta_l)(\xi_l-x_l)\leqslant|f'(\eta_l)|\Delta x$，于是

$$\begin{aligned}\left|\sum_{l=0}^{n-1}\left[f(\xi_l)-f(x_l)\right]\Delta x\right| &\leqslant \sum_{l=0}^{n-1}|f'(\eta_l)|\Delta^2 x\\ &\leqslant Mn\left(\frac{b-a}{n}\right)^2=M(b-a)^2/n.\end{aligned}\tag{16.375}$$

其中我们用到了 $|f'(\eta_l)|$ 极大值为 M。也就是说，闭区间上的连续函数，如何可导，其导数也是有界的。闭区间上的连续函数导数有界，意味着在闭区间上的连续函数不可能出现无限增加到正无穷和无限减少到负无穷的情况。这个就很容易建立起来直观的理解了。

当然，我们不用这个更强的条件来证明这个命题，而是仅仅给出一个直观的解释。证明之前已经完成了。

到此，我们的微积分的严格的理论也就完成了。我们先重新建立了实数（通过戴德金分割），从而定义了柯西收敛和极限，证明了在实数上柯西收敛和极限完全等价，接着用极限定义了导数和定积分，最后证明了定积分和导数（不定积分）之间的联系，也就是牛顿-莱布尼兹公式。

用今天微积分的知识体系已经完善的角度，我们其实不过就是引入了五六个概念按顺序证明了七八个命题。但是，实际上在微积分的发展过程中可不是这么简单的。有的概念和命题在体系形成之前就有了。有的概念和命题是在构建体系的过程中发现需要有的。然后，数学家们选择一个整体上没有逻辑漏洞的思路来把这些已有的和新的概念和命题联系起来。你发现，这正好也是《几何原本》[11] 的构建知识体系的方式。

此外，做者我是一个物理学家，本书呢是一本尽量追求有启发又严格的数学入门书，所以在证明每一个命题的时候，我们都会联系到这个命题在知识体系中的作用，这个命题的证明实际上可以如何从直觉来理解。

顺便，有了牛顿-莱布尼兹公式之后，我们这样来看不定积分，

$$F(x)=\int_a^x f(\xi)\,d\xi.\tag{16.376}$$

其中下边界 a 可以任意取定一个，只要 $f(x)$ 在那一点可积。很容易证明 $F'(x)=f(x)$。留作练习。

习题 16.21 (变上限定积分). 证明变上限定积分 $F(x)=\int_a^x f(\xi)\,d\xi$ 的导数是 $f(x)$。

当然，在到此为止的课程中，我们没有仔细展开讨论可导和可积的条件。这些就留给进一步的微积分课程、实分析课程去完成了。

16.8.4 神奇的全微分：通过线性部分就可以得到原始函数

有了牛顿-莱布尼兹公式，我们就可以把 $\int_a^b f(x)\,dx$ 当作数学对象而不仅仅是记号来理解了。$\int_a^b dx$ 是对 dx 的积分，$\int_a^b f(x)\,dx$ 是对 dx 做了加权积分。再结合 $dF=f(x)\,dx$，以及牛顿-莱布尼兹公式，$\int_a^b f(x)\,dx$ 可以看作是对 dF 的积分 $\int_{F(a)}^{F(b)} dF$。这里我们还用到了定积分的变量替换——需要把函数形式（函数值不变）、微分和积分上下限同时做替换，见公式(16.167)。这个时候 F 是自变量，当然，我们就会得到原函数为 F，也就是 $\int_{F(a)}^{F(b)} dF=F|_{F(a)}^{F(b)}$。

之前，我们也提到了对 dF 和 dx 的另一个理解：根据Taylor 公式，它们可以看作是极限下的函数的线性主部。$\Delta F=f(x)\,\Delta x+o(\Delta x)$，极限下就得到了 dF，而当 $F(x)=x$ 的时候，自然 $f(x)=1$，极限下就得到了 dx。

显然这两个理解是有差别的。再后者看来，dF 仅仅是一个近似；但是，从前者看来，好像 dF 包含了所有的信息。由于有了导函数 $f(x)$ 就可以得到原函数 $F(x)$，而这个过程可以看作是对 dF 的变上限定积分，我们把这个包含了 $F(x)$ 的所有信息的 dF 称作全微分。

如果推广到多元函数 $F(x,y)$，按照Taylor 公式，自然我们会有由于 x 的变化到 $x+\Delta x$ 带来的 F 的变化，以及由于 y 的变化到 $y+\Delta y$ 带来的 F 的变化，也就是

$$\Delta F=\left(\frac{\partial F}{\partial x}\right)_y(x,y)\,\Delta x+\left(\frac{\partial F}{\partial y}\right)_x(x,y)\,\Delta y+o(\Delta x)+o(\Delta y). \quad (16.377)$$

因此，极限下自然就得到

$$dF=\left(\frac{\partial F}{\partial x}\right)_y dx+\left(\frac{\partial F}{\partial y}\right)_x dy. \quad (16.378)$$

这样我们就把全微分推广到了多变量函数的情形。同样的道理，多变量函数的全微分也包含了原函数的所有的信息，从 dF 可以得到原函数 F。

当我们写下来

$$dF = u(x,y)\,dx + v(x,y)\,dy \tag{16.379}$$

并且宣称这是一个全微分的时候，我们实际上就等于说，

$$\left(\frac{\partial F}{\partial x}\right)_y = u(x,y) \quad \left(\frac{\partial F}{\partial y}\right)_x = v(x,y). \tag{16.380}$$

于是，我们就可以通过对 $u(x,y)$ 做对 x 的积分，对 $v(x,y)$ 做对 y 的积分来得到 $F(x,y)$。不过，我们注意到对同一个 $F(x,y)$ 先做一个对 x 的导数再做一个对 y 的导数，应该和先做一个对 y 的导数再做一个对 x 的导数得到相同的结果，也就是

$$\left(\frac{\partial u}{\partial y}\right)_x = \left(\frac{\partial v}{\partial x}\right)_y, \tag{16.381}$$

或者记为

$$\frac{\partial^2 F}{\partial y \partial x} = \frac{\partial^2 F}{\partial x \partial y}. \tag{16.382}$$

我们来做几个例子验证一下这个结果，然后求一下多元函数的原函数。

例 16.44 (计算偏导数). 求 $F(x,y) = x^2 + 2xy$ 的偏导数，并且检验公式 (16.382) 是否成立。

先把 y 看作常数，求函数对 x 的导数，

$$\left(\frac{\partial F}{\partial x}\right)_y = 2x + 2y. \tag{16.383}$$

再把 x 看作常数，求函数对 y 的导数，

$$\left(\frac{\partial F}{\partial y}\right)_x = 2x. \tag{16.384}$$

接着求对 $\frac{\partial F}{\partial x}$ 的对 y 的偏导数，

$$\frac{\partial^2 F}{\partial y \partial x} = \frac{\partial (2x+2y)}{\partial y} = 2. \tag{16.385}$$

再求对 $\frac{\partial F}{\partial y}$ 的对 x 的偏导数，

$$\frac{\partial^2 F}{\partial x \partial y} = \frac{\partial (2x)}{\partial x} = 2. \tag{16.386}$$

公式 (16.382) 成立。

例 16.45 (计算偏导数). 求 $F(x,y)=\sin(2x+y)$ 的偏导数，并且检验公式 (16.382) 是否成立。

先把 y 看作常数，求函数对 x 的导数，

$$\left(\frac{\partial F}{\partial x}\right)_y = 2\cos(2x+y). \tag{16.387}$$

再把 x 看作常数，求函数对 y 的导数，

$$\left(\frac{\partial F}{\partial y}\right)_x = \cos(2x+y). \tag{16.388}$$

接着求对 $\frac{\partial F}{\partial x}$ 的对 y 的偏导数，

$$\frac{\partial^2 F}{\partial y \partial x} = \frac{\partial(2\cos(2x+y))}{\partial y} = -2\sin(2x+y). \tag{16.389}$$

再求对 $\frac{\partial F}{\partial y}$ 的对 x 的偏导数，

$$\frac{\partial^2 F}{\partial x \partial y} = \frac{\partial(\cos(2x+y))}{\partial x} = -2\sin(2x+y). \tag{16.390}$$

公式 (16.382) 成立。

现在，我们反过来，已知全微分来求原函数。

例 16.46 (已知全微分求原函数). 已知 $dF=(2x+2y)\,dx+(2x)\,dy$，求 F。

根据全微分公式公式 (16.378)，有

$$\left(\frac{\partial F}{\partial x}\right)_y = 2x+2y, \tag{16.391}$$

$$\left(\frac{\partial F}{\partial y}\right)_x = 2x. \tag{16.392}$$

对上述第一式中的函数做 x 的积分，

$$\int_a^x (2x+2y)\,dx = \left(x^2+2yx\right)\Big|_a^x = x^2+2yx+C(y). \tag{16.393}$$

注意，其中的“常数”项 $C(y)$ 可以包含 y，因为我们把 y 当作了常数。这个常数也可以看作是积分的下界的选择带来的。对上述第二式中的函数做 y 的积分，

$$\int_a^y (2x)\,dy = (2yx)\Big|_a^y = 2yx+D(x). \tag{16.394}$$

其中的“常数”项 $D(x)$ 可以包含 x。要保证 $x^2+2yx+C(y)$ 和 $2yx+D(x)$ 相同，则 $D(x)=x^2+C, C(x)=C$。因此，

$$F(x,y)=x^2+2yx+C. \tag{16.395}$$

例 16.47 (已知全微分求原函数). 已知 $dF=2\cos(2x+y)dx+\cos(2x+y)dy$，求 F。

根据全微分公式公式 (16.378)，有

$$\left(\frac{\partial F}{\partial x}\right)_y=2\cos(2x+y), \tag{16.396}$$

$$\left(\frac{\partial F}{\partial y}\right)_x=\cos(2x+y). \tag{16.397}$$

对上述第一式中的函数做 x 的积分，

$$\int_a^x 2\cos(2x+y)dx=\sin(2x+y)|_a^x=\sin(2x+y)+C(y). \tag{16.398}$$

注意，其中的“常数”项 $C(y)$ 可以包含 y。对上述第二式中的函数做 y 的积分，

$$\int_a^y \cos(2x+y)dy=\sin(2x+y)|_a^y=\sin(2x+y)+D(x). \tag{16.399}$$

其中的“常数”项 $D(x)$ 可以包含 x。要保证 $\sin(2x+y)+C(y)$ 和 $\sin(2x+y)+D(x)$ 相同，则 $D=C, C(x)=C$。因此，

$$F(x,y)=\sin(2x+y)+C. \tag{16.400}$$

从这几个例子我们看到，全微分对于多变量函数，也包含了全部信息，可以从全微分里面的偏导数函数——通过计算积分——得到多变量的原函数。

这个看起来仅仅是对于线性主部的近似竟然包含了全部的信息。这里面肯定包含了更加深刻的数学上的含义。如果我们搞清楚了这个含义，很有可能，牛顿-莱布尼兹公式真的就是这个含义的直接结果而不是一个需要证明的命题了。

实际上，后来发展起来的微分几何就是这样的一个分支学科。在那里，数学家找到了导数的含义——切空间（所有可能的曲线的切线的集合）的元素，而全微分成了微分形式——切空间的对偶空间的元素。并且，我们已经看到对于任何一条曲线，已知其每个点上的切线的斜率的值都可以得到这条曲线本身。因此，对于曲线的研究完全可以变成对于其上的左右的切线的研究。实际上，这就是李群（曲线）和李代数（曲线上的切线）的关系。把对一个量的研究转变为对这个量的变化——并且这个变化合起来的空间是一个线性空间（关于线性空间见第十七章）而线性空间总是比原对象所在的集合更容易处理——的研究，这个微积分的核心思想再一次在对更加复杂的对象的研究上发挥了作用。

16.9 微积分和人类思考深刻问题能力的爆发

尽管我们前面已经提到了促进微积分提出和发展的各种来自于现实世界和数学本身的需求已经存在，极限、夹逼、积分的思想也已经存在，但是在微积分被提出和发展之前（先不算严格化的那部分），基本上所有的代数方面的数学研究都在关心解方程[21]、作计算以及某些具体的函数的研究，几何方面不能处理一般的曲线，也就能把所有的几何对象都化成三角形和圆来处理。有了微积分，一切都不一样了，数学的研究内容和研究对象就不是以前的样子了。

几何上可以处理任何曲线了，只需要把曲线和函数，或者曲线联系到其参数化表示。曲面甚至更高维的几何对象也不过就是曲线的累积，如果要算其面积体积之类的，那不过就是一个积分问题。代数上，任何一个性质好一点的（例如可导的）函数在任何一点附近都可以用一个多项式去逼近[22]，函数甚至泛函的极值问题也不再是高度技巧性的问题而是一个有一般性的方法的问题了，任何一条曲线下的面积都可以通过积分得到了，任何一条曲线在任何一点（只要函数在这一点可导）的切线也能求出来了。最关键的是，不是这些问题我们会求解了，而是求解这些问题的方法竟然是同一套

[21]解析求解方程，对于大多数方程都是不可能的，例如高于五次方的多项式方程。我们一回会看到，方程数值求解其实也得依赖微积分。

[22]未来我们还会学习到用其他熟悉的函数，例如三角函数，去逼近任意函数。

概念——微积分。

现在，我们来看看，有了这个一般概念和相应的思维方式和分析方法，以及上面提到的其典型应用场景之后，人类基于微积分在数学上、科学上、工程上都取得了哪些没有微积分不敢想像的进步，人类可以思考哪些没有微积分就不能思考甚至就不可能会想到去思考的问题了。这部分的目的是让你更加深刻地感受到，如果你没有学会微积分的概念、思维方式和分析方法、典型应用场景，那将是你人生的一大憾事。微积分使得你能够站在巨人的肩上，使得能够欣赏巨人的创造，使得你能够提出和解决更加深刻的问题。

分享一段人们对Newton的物理学和数学（从牛顿的角度，微积分这个他作为创造者之一的数学是为了解决物理问题）的工作的一段评价：

> Nature and Nature's laws lay hid in night: God said, Let Newton be, and all was light.
>
> – Alexander Pope
>
> 自然和自然的规律藏在黑夜之中：上帝说，让牛顿去吧！于是，世界一片光明。

我希望你，亲爱的读者，也能看到这个光，未来创造自己的光。

16.9.1 基于导数的方程求解方法

第一个例子我们来看看微积分是如何帮助更好地解决微积分之前的数学研究的主要问题之一——求解方程的。未来，在群和矢量空间的部分我们会帮助大家理解一个叫做“伽罗瓦（Galois）理论”的东西——五次以上的方程不存在解析解。因此，微积分也只能在数值解上做出贡献。我们今天来见识一下两个基于微积分的方程数值求解方法。

基于Taylor 公式或者说导数的定义，我们有

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + o(x - x_0). \tag{16.401}$$

在这里我们忽略高阶小量。如果刚好有 $f(x) = 0$，则

$$x = -\frac{f(x_0)}{f'(x_0)} + x_0. \tag{16.402}$$

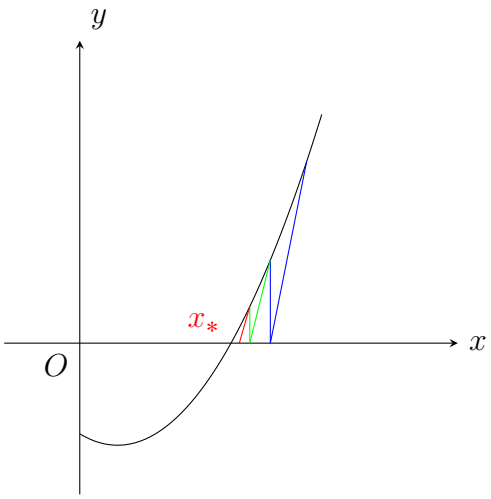


图 16.18: **求方程数值解的牛顿方法**。根据牛顿法可以通过基于导数和迭代方法找到方程的解。从图中可以看到，随着迭代，切线在 x 轴上的交点越来越接近这个函数的零点。

如果不是正好有 $f(x)=0$，是不是我们可以这样来算，

$$x_{n+1}=-\frac{f(x_n)}{f'(x_n)}+x_n. \tag{16.403}$$

当 x_n 足够接近函数 $f(x)$ 的零点 x_* 的时候，$f(x_n)$ 非常接近于零，只要函数 $f(x)$ 在这个点的导数不同时也趋近于零（或者说以比 $f(x)$ 更慢的速度趋近于零），我们就有了

$$x_{n+1}-x_n=-\frac{f(x_n)}{f'(x_n)}\to 0. \tag{16.404}$$

也就是说，x_* 确实是迭代方程公式 (16.403) 的不动点。

几何上，我们可以通过下面的图获得一个直观的理解。从数轴上任意一个 x_0 开始，我们先找到 $f(x_0)$，计算导数 $f'(x_0)$，然后，我们寻找直线 $y=f'(x_0)(x-x_0)+f(x_0)$ 和 x 轴也就是 $y=0$ 的交点，得到 x_1。然后，重复这个过程。只要 x_* 存在，每一次，这样的迭代都会使得 x_{n+1} 比 x_n 更加接近 x_*，见图 16.18。

有了直观理解和前面从Taylor 公式受到的启发之后，我们其实还需要找到一个牛顿法收敛（$\lim_{n\to\infty}x_n=x_*$，这个时候 $x_{n+1}-x_n$ 随着 n 变大越

来越小）的条件然后证明在这个条件下牛顿法确实收敛。当然，只要收敛就可以得到

$$\begin{aligned} & f\left(x_{n+1}\right)=f\left(x_n\right)+f'\left(x_n\right)\left(x_{n+1}-x_n\right)+o\left(x_{n+1}-x_n\right) \\ =& f\left(x_n\right)+f'\left(x_n\right)\left(-\frac{f\left(x_n\right)}{f'\left(x_n\right)}+x_n-x_n\right)+o\left(x_{n+1}-x_n\right) \\ & =o\left(x_{n+1}-x_n\right). \end{aligned} \tag{16.405}$$

也就是，$\lim_{x_n\to x} f\left(x_n\right)=0$。因此，我们需要证明 $\lim_{n\to\infty} x_n=x_*$。可以看到，这个证明必然也是通过微积分这部分的极限的知识来证明的。

具体证明超出了本书的范围。但是，我们要看到的是初等数学的问题，其解法和解法成立的条件和证明，也都可以由微积分这个高等数学的概念来解决。

我们还可以把牛顿法推广到牛顿弦截法，用割线代替切线，也就是把切线的斜率替换为割线的斜率——这样就不用要求函数可导了，得到

$$x_{n+1}=x_n-\frac{f\left(x_n\right)}{\frac{f(x_n)-f(x_{n-1})}{x_n-x_{n-1}}}. \tag{16.406}$$

这样只要从两个初始点 x_0,x_1 开始我们就可以通过迭代得到 $x_2,x_3,\cdots$。直观上，我们可见图 16.19。同样我们需要用微积分来找到这个迭代收敛的条件和证明相应的定理。

这样的微积分这个新概念新工具在解决旧问题上的应用仅仅是微积分开创的新的数学时代的序幕。

16.9.2 力学——微分方程

用代数方程来表述和求解问题是在微积分之前就被研究和使用的方法。其思想就是对于一个位未知的对象 x，例如某个东西的量，先假设已知，对这个对象做一番操作，记作 $f(x)$，然后得到的东西，或者得到的东西和这个未知的对象世界的关系，记作 $g(x)$，是已知的。这样就得到了一个 $f(x)=g(x)$ 的方程。

这个方程的思想，当研究的是在外场力下质点的运动的时候，就成了一个微分方程。例如，一个受固定大小和方向的力的物体，根据 $F=ma$

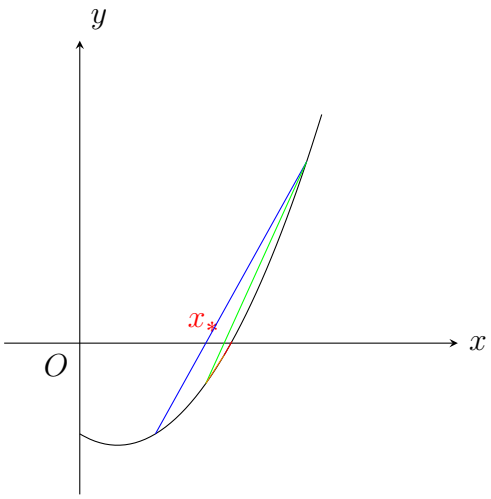


图 16.19: **求方程数值解的牛顿弦截法**。根据牛顿弦截法可以通过基于割线的斜率和迭代方法找到方程的解。从图中可以看到牛顿弦截法所定义的割线在 x 轴上的交点随着迭代越来越接近这个函数的零点。

（下面用一维运动为例，多维只需要改用矢量形式的 $\vec{F}=m\vec{a}$，也就是在每个方向上都有 $F=ma$ 就行）得到，

$$m\frac{d^2x}{dt^2}=F. \tag{16.407}$$

这是一个二阶微分方程——出现了二阶导数项。通过积分可以得到

$$\frac{dx}{dt}=v, v=\frac{F}{m}t+v_0. \tag{16.408}$$

继续积分得到

$$x=\frac{1}{2}\frac{F}{m}t^2+v_0t+x_0. \tag{16.409}$$

这就是我们熟悉的匀加速直线运动。如果 F 依赖于 t，那就需要求解

$$m\frac{d^2x}{dt^2}=F\left(t\right). \tag{16.410}$$

这个仍然可以通过积分来求解，得到

$$v=\int_0^t F\left(t\right)d\tau+v_0, \tag{16.411}$$

$$x=\int_0^t\int_0^\xi F\left(t\right)d\tau d\xi+v_0t+x_0. \tag{16.412}$$

更加复杂的情形出现在 $F = F(x)$ 的时候（当然，$F = F(x,t)$ 就更更复杂了）。这个时候，

$$m\frac{d^2x}{dt^2} = F\left(x\left(t\right)\right), \tag{16.413}$$

或者更一般的

$$m\frac{d^2x}{dt^2} = F\left(x\left(t\right), t\right), \tag{16.414}$$

就不能简单通过积分来求解了。

这样的同时出现了自变量、函数、函数的导数的表达式的方程被称为微分方程。当然，当 F 是常数，或者仅仅依赖于 t 的时候，也是微分方程。微分方程的求解是一个专门的数学分支学科。这部分内容的学习留给专门的关于微分方程的课程和书。

比如说，在路上行驶的汽车，被发送到太空的卫星，背后都是这样的运动方程。以卫星的发射过程为例，我们需要考虑的外场力有：地球的引力（其中可以分成把地球当作完美的球形而简化以后的引力和地球由于偏离球形而造成的引力）、大气阻力、太阳辐射压、太阳和月亮以及其他星体的引力、地球反照辐射和热辐射形成的力。需要考虑的卫星的运动除了平动，还要考虑转动和振动，甚至形变。不过，我们在这里仅仅以平动为例。相当于这里的 F 里面有很多项，并且其中很多项和 x 甚至 v 有关系。

更进一步，这里的运动也不是一个仅仅控制初始速度和初始位置，就随着卫星在外场力作用下走到哪里算哪里的问题，而是一个有预定的点和到那个点以后的预定的轨道的问题。因此，实际上，卫星上天是一个以微分方程的形式表现出来的运动方程的控制问题。所谓控制问题就是需要把卫星最终送到预定轨道上去。

由于不管控制的那部分都已经是用微分方程描述的，自然，加了控制以后，仍然是需要用微分方程来描述的问题。并且，由于外场力很多，还依赖于时间、位置和速度，因此这些方程往往只能采用数值方式求解。这些背后都需要微积分。

我们再来看另一个力学问题是微分方程的根源。当我们再考虑一个对象的状态和状态变化——称为运动——的时候，我们往往需要把这个对象打开，走到这个对象内部的子对象上去，看到这些子对象之间的相互作用。

比如说，我们不再把汽车当作一个整体，而是要考虑汽车的轮子有主动有被动——开车的人知道四轮驱动和两轮驱动的差别于是汽车的设计师

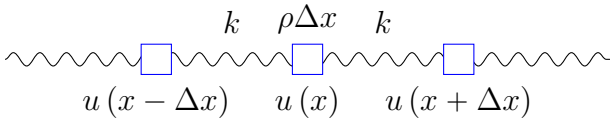


图 16.20: **纵波的波动方程的推导**。使用微元法可以推导出来弹簧上的纵波的波动方程。

在设计的时候肯定要考虑这个差别，甚至汽车的轮子也要进一步分开成轮胎和地面的接触部分和轮轴等其他部分——当汽车的抓地力出了问题的时候就要做这样的分析。

为了更具体地看到这个由于分解对像得到的微分方程，我们来举一个弹簧的振动形成的纵波的例子。

先定义变量 $u(x,t)$ 表示平衡位置在 x 长度为 Δx 的弹簧的微元在 t 时刻的时候的位置。然后，我们考虑这个微元受到的力：在其左侧的微元给它的力为 $k(u(x-\Delta x,t)-u(x,t))$（以右侧为力和位移的正方向，当 $u(x-\Delta x,t)-u(x,t)>0$ 的时候，弹簧处于压缩态，在 x 处的微元受到的力向右，也就是大于零），在其右侧的微元给它的力为 $k(u(x+\Delta x,t)-u(x,t))$（当 $u(x+\Delta x,t)-u(x,t)>0$ 的时候，弹簧处于拉伸态，在 x 处的微元受到的力向右，也就是大于零），合起来其受到的力为，

$$F=k(u(x-\Delta x,t)-u(x,t))+k(u(x+\Delta x,t)-u(x,t)). \tag{16.415}$$

这个微元的质量为

$$m=\rho\Delta x \tag{16.416}$$

因此，根据 $F=ma$，有

$$\begin{aligned}\rho\Delta x u''_{tt}(x,t)=F&=k(u(x-\Delta x,t)-u(x,t))\\&\quad+k(u(x+\Delta x,t)-u(x,t)),\\&=ku'_x(x,t)\Delta x-ku'_x(x-\Delta x,t)\Delta x,\\&=ku''_{xx}(x,t)\Delta^2 x\end{aligned} \tag{16.417}$$

其中每一小段的弹簧的 k 和 $\frac{L}{\Delta x}$ 个这样的弹簧微元合起来的弹性系数之间

有如下关系[23]:

$$K = \frac{k}{N} = \frac{k\Delta x}{L}. \tag{16.418}$$

因此

$$\rho\Delta x u''_{tt}(x,t) = \frac{KL}{\Delta x} u''_{xx}(x,t)\,\Delta^2 x. \tag{16.419}$$

两边除以 Δx 也就是

$$u''_{tt}(x,t) = \frac{KL}{\rho} u''_{xx}(x,t). \tag{16.420}$$

现在我们看到当包含了对象内部的子对象的时候，把子对象之间的相互作用考虑进来的时候，物理系统的方程往往是微分方程。其背后的原因是：物理系统很自然地能够使用微元法，而微元法往往导致微分方程。

16.9.3 自然进入问题的解的边界条件

尽管我们说了微分方程的求解不是本书所要包含的内容，但是为了突出微积分在求解特定形状特定边界条件问题上的非凡的作用，以及联系到后面的用三角函数和更加专门的函数来把任意函数表示为矢量，我们还是来求解几个微分方程。

前面我们得到了一维的波动方程公式 (16.421), 二维波动方程我们记作，

$$u''_{tt} = c^2\left(u''_{xx} + u''_{yy}\right). \tag{16.421}$$

其中我们用常数 c 替代了之前的那些常数的组合。记 $\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}$，称为拉普拉斯算符拉普拉斯算符，上式可以重新写成

$$u''_{tt} = c^2\nabla^2 u. \tag{16.422}$$

[23]这个可以这样证明。一根弹性系数为 k 的弹簧在力 F 作用下伸长量为 $\Delta l = \frac{F}{k}$。当 N 个这样的弹簧首位相连的时候每一根弹簧受力都相同，贡献的伸长量都是 Δl。因此，整个弹簧的伸长量为 $N\Delta l = \frac{NF}{k}$。因此整个弹簧的弹性系数为 $\frac{F}{\frac{NF}{k}} = \frac{k}{N}$。

我们先来在一维固定长度固定端点，也就是

$$u\left(0,t\right)=0=u\left(L,t\right), \tag{16.423}$$

$$u\left(x,0\right)=0, \tag{16.424}$$

$$u\left(x,0\right)_x=v_0\delta\left(x-\frac{L}{2}\right). \tag{16.425}$$

的情形下求解这个方程。这个情形表示：在端点 $x=0,L$ 处，$u(x,t)=0$ 固定不动；同时，在 $t=0$ 的时刻，$u\left(x,0\right)=0$ 也就是弹簧完全静止，$u\left(x,0\right)_x=v_0\delta\left(x-\frac{L}{2}\right)$ 表示在 $x=\frac{L}{2}$ 的地方被给了一个初始速度，其他地方的初始速度为零。

任何一个两阶微分方程可以通过给出两个初始条件或者两个边界条件来确定解。这里，u 对 x 和 t 都是两阶，因此，我们对于空间变量，给出来两个边界条件，对于时间变量给出来两个初始条件。不过，之前我们在最速下降线问题、光子路径问题中也看到了，对于时间也可以给出两个边界条件，也就是起点和终点。这里选择这样的边界条件和初始条件是为了后面展示适合问题（特定形状）的特定函数的作用。

对于一维问题，在这个边解条件下，我们可以证明最一般的函数是

$$u\left(x,t\right)=\sum_{n=0}^{\infty}\left[a_n\sin\left(\frac{n\pi}{L}x\right)\cos\left(\omega_n t\right)+b_n\sin\left(\frac{n\pi}{L}x\right)\sin\left(\omega_n t\right)\right], \tag{16.426}$$

其中

$$\omega_n=\frac{n\pi c}{L}. \tag{16.427}$$

注意，这里的 b_0 实际上没有意义，因为对应着的 $\sin\left(\frac{n\pi}{L}x\right)=\sin\left(0x\right)=0$。保持 b_0 项仅仅是为了记号简单。

我们可以这样来检验这个结论。先求出来 u''_{tt}，

$$u_{tt}\left(x,t\right)=-\sum_{n=0}^{\infty}\omega_n^2\left[a_n\sin\left(\frac{n\pi}{L}x\right)\cos\left(\omega_n t\right)+b_n\sin\left(\frac{n\pi}{L}x\right)\sin\left(\omega_n t\right)\right], \tag{16.428}$$

再求出来 u''_{xx}，

$$u_{xx}(x,t) = -\sum_{n=0}^{\infty}\left(\frac{n\pi}{L}\right)^2\left[a_n \sin\left(\frac{n\pi}{L}x\right)\cos(\omega_n t) + b_n \sin\left(\frac{n\pi}{L}x\right)\sin(\omega_n t)\right], \tag{16.429}$$

最后检验是否满足 $u''_{tt} = c^2 u''_{xx}$。我们发现，当 $\omega_n = \frac{n\pi c}{L}$ 的时候，完全满足。

下面，我们来根据初始条件求出来 a_n 和 b_n。根据 $u(x,0) = 0$ 得到 $a_n = 0$，这是因为 $u(x,0) = 0$ 会表现为

$$0 = u(x,0) = \sum_{n=0}^{\infty} a_n \sin\left(\frac{n\pi}{L}x\right). \tag{16.430}$$

未来在傅里叶级数那一节（16.9.6 节）我们会证明，只有当 $a_n = 0$ 的时候，这个方程才能成立。

$$v_0\delta\left(x - \frac{L}{2}\right) = u_t(x,0) = \sum_{n=0}^{\infty} b_n\omega_n \sin\left(\frac{n\pi}{L}x\right). \tag{16.431}$$

未来在傅里叶级数那一节（16.9.6 节）我们会证明，这个方程成立的条件是

$$b_n = \frac{2}{\omega_n L}\int_0^L v_0\delta\left(x - \frac{L}{2}\right)\sin\left(\frac{n\pi}{L}x\right)dx = \frac{2v_0}{L\omega_n}\sin\left(\frac{n\pi}{2}\right). \tag{16.432}$$

如果边界条件改为二维平面上的 $L_1 \times L_2$ 的长方形，长方形的四周都固定，则相应的最一般的满足波动方程的解肯定是

$$\begin{aligned} u(x,t) = \sum_{m,n}\Bigg[&a_{m,n}\sin\left(\frac{m\pi}{L_1}x\right)\sin\left(\frac{n\pi}{L_2}y\right)\cos(\omega_{m,n}t) \\ &+ b_{m,n}\sin\left(\frac{m\pi}{L_1}x\right)\sin\left(\frac{n\pi}{L_2}y\right)\sin(\omega_{m,n}t)\Bigg]. \end{aligned} \tag{16.433}$$

其中

$$\omega_{m,n} = \pi c\sqrt{\left(\frac{m}{L_1}\right)^2 + \left(\frac{n}{L_2}\right)^2}. \tag{16.434}$$

检验这个结论的方法相同，计算出来 u''_{tt} 和 $u''_{xx} + u''_{yy}$，然后验证确实公式 (16.421) 成立。接着结合初始条件，我们可以求出来 a_n, b_n。例如，如果

初始时刻还是 $u(x,y,0)=0$ 处处为零，则仍然可以得到 $a_{m,n}=0$。结合 $u_t'(x,y,0)=v_0\delta\left(x-\frac{L_1}{2}\right)\delta\left(y-\frac{L_2}{2}\right)$，可以得到

$$b_{m,n}=\frac{4}{\omega_{m,n}L_1L_2}\int_0^L v_0\delta\left(x-\frac{L_1}{2}\right)\delta\left(y-\frac{L_2}{2}\right)\sin\left(\frac{m\pi}{L_1}x\right)\sin\left(\frac{n\pi}{L_2}x\right)dxdy$$
$$=\frac{4v_0}{\omega_{mn}L_1L_2}\sin\left(\frac{m\pi}{2}\right)\sin\left(\frac{n\pi}{2}\right). \tag{16.435}$$

那么，如果是三维长方体的上的波动呢？只需要把这里的函数推广到三维的情形。如果是三维球体的表面或者球体呢？或者更加单一点，二维的圆或者圆环呢？方法完全一样，只需要在例如二维圆或者圆环上把满足波动方程公式 (16.421) 的一般的解的形式求出来。这个一般解的形式往往是在那个形状上满足

$$\nabla^2u=\lambda u, \tag{16.436}$$

称为拉普拉斯算符的本征函数。然后，按照边解条件把这个一般形式的解写成本征函数的系数组合的形式，最后通过初始条件来确定系数。

我们就不再具体求解这些在二维圆上满足波动方程公式 (16.421) 的一般的解的形式了。我们要注意的是：不同的承载波动的物体的形状和边界条件，会先确定拉普拉斯算符的本征函数，然后通过初始条件来确定出现在本征函数之前的组合系数。前面的例子中，我们仅仅对平面直角坐标系写出来了拉普拉斯算符的本征函数。对于更一般的承载波动的物体的形状和边界条件，这个本征函数要复杂很多。

但是，我们看到了，物理现象发生的载体的信息和边界条件的信息，自然地进入了用来表示一般解的本征函数。或者说，对于由波动方程这样的[24]微分方程描述的物理现象，我们甚至可以提前求解出来空间部分的算符的本征函数，然后，未来在结合边界条件和初始条件来确定最后的解所对应的本征函数前面的组合系数。这样的一个如此自然地把物理现象的载体、边界条件、初始条件分离，并且先把载体、边界条件用上的高效的解决问题的方式，只有在把问题表述为微分方程的形式之后才能得到。而微分方程

[24]这里的“这样的”含义其实是时间和空间部分的导数算符可以分离。

的基础是微积分。没有微积分，我们是不可能把问题分解和解决得如此漂亮的。于是，在物理学中，大量的问题是通过这样得本征函数的方式来解决的。

16.9.4 勒让德变换

这一节我们来体会一个非常神奇的变换，物理学中经常要使用到的变换。我们来看看这个变换为什么要做，怎么做，以及怎么做和微积分的联系。没有微积分，我们都不知道于原来还可以这样做，但是明明来自于物理学的启发已经要去我们做这样的变换啊。

我们从通常的函数的变换开始。函数 $f(x)$ 的变换有两种。一种是变量替换——维持函数形式不变，把 $f(x)$ 中的 x 换成 $x=x(u)$。这相当于复合函数 $\hat{f}(x)=f(u(x))$。一种是维持函数值不变，也就是使得 $\tilde{f}(u(x))=f(x)$，然后找到这个新的函数形式 $\tilde{f}(x)=f(u^{-1}(x))$。我们来看一个例子。

例 16.48 (函数形式不变和函数值不变的变换). 我们把 $f(x)=x^2+1$ 变成关于新的自变量 $u=x+1$ 的函数。第一，保持函数形式不变，也就是

$$\hat{f}(x)=f(u(x))=f(x+1)=(x+1)^2+1. \tag{16.437}$$

第二，保持函数值不变，也就是“新函数在新自变量的取值等于老函数在老自变量的取值——同一个点的自变量变了但是这个点上的值不变”，

$$\begin{aligned} &\tilde{f}(u(x))=f(x), \\ \Rightarrow\ &\tilde{f}(x+1)=f(x), \\ \Rightarrow\ &\tilde{f}(x)=f(x-1). \end{aligned} \tag{16.438}$$

得到

$$\tilde{f}(x)=(x-1)^2+1. \tag{16.439}$$

“新函数在新自变量的取值等于老函数在老自变量的取值——同一个点的自变量变了但是这个点上的值不变”其实意味着在同一个点的同一个物理量上的测量结果相同。

但是，在物理学的研究中，还有另一种情况：所测量的物理量都可以不一样，但是是对同一个物理过程的测量。比如说，在运动学问题中，我们可

以测量位置和速度。运动学方程为 $F = ma$，也就是

$$x''_{tt} = \frac{F(x)}{m}. \tag{16.440}$$

或者用Newton的符号，

$$\ddot{x} = \frac{F(x)}{m}. \tag{16.441}$$

这里面出现的描述物理系统的变量是位置，然后时间是这个位置的参数。为了简单计，我们忽略了 F 对 t 和 v 的直接依赖关系。$F = ma$ 还可以写成下面的形式，

$$\begin{cases} \dot{x} = \frac{p}{m}, \\ \dot{p} = F(x). \end{cases} \tag{16.442}$$

这里面出现的描述物理系统的变量是位置和动量（$p = mv$ 和 v 等价），然后时间是这个位置和动量的参数。这两种描述运动的方式的区别是：前者的自变量是位置，后者的自变量是位置和动量。

物理学家引入一个势函数

$$V(x) = -\int_a^x F(\xi)\, d\xi, \tag{16.443}$$

于是

$$\frac{d}{dx}V(x) = -F(x), \tag{16.444}$$

来把上面的运动方程重新写成下面的形式，

$$L(x, \dot{x}) = \frac{1}{2}m\dot{x}^2 - V(x), \tag{16.445}$$

然后物体的运动就是满足下面这个泛函极值的解，

$$S = \int_{t_0}^{t_f} L(x(\tau), \dot{x}(\tau))\, d\tau. \tag{16.446}$$

下面，我们来用Euler-Lagrangian 方程公式 (16.234)，重写在这里

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{x}}\right) = \frac{\partial L}{\partial x}, \tag{16.447}$$

来展示一下得到的运动方程就是 $F=ma$,

$$\frac{\partial L}{\partial \dot{x}}=m\dot{x}, \tag{16.448}$$

$$\frac{\partial L}{\partial x}=-\frac{dV}{dx}=F\left(x\right), \tag{16.449}$$

$$m\ddot{x}=\frac{d}{dt}\frac{\partial L}{\partial \dot{x}}=\frac{\partial L}{\partial x}=F\left(x\right). \tag{16.450}$$

顺便，这个结论本身也很有意思：通过引入正确的 $L\left(x,\dot{x}\right)$ 可以把质点运动问题转化为泛函极值问题。这个量 $L\left(x,\dot{x}\right)$ 和 S 还有一个名称，分别称为Lagrangian和作用量。因此，相应地这个泛函极值问题也被称为最小作用量原理。未来，任何质点运动问题只需要给出来这个 $L\left(x,\dot{x}\right)$ 就知道了这个质点的运动方程，而且这个 $L\left(x,\dot{x}\right)$ 非常容易被推广到一组质点，甚至连续介质上的质点们。比如说，我们之前的绳子上的波，也可以写下来这个系统的 $L\left(u\left(x,t\right),\dot{u}\left(x,t\right)\right)$。

注意，从质点运动来说，从变量 x 变成 (x,p) 非常简单。但是，对于 $L\left(u\left(x,t\right),\dot{u}\left(x,t\right)\right)$ 的，要找到对应的“动量”就很不简单了。因此，我们就希望得到一个从以 $x,\dot{x}$ 为自变量的 L 到以 x,p 为自变量的另一个函数的另一组方程的一般方法。

我们还是从质点力学的问题里面先得到一点点启发。我们发现，如果我们定义这样一个函数，称为Hamiltonian就可以得到公式 (16.442) 了，

$$H\left(x,p\right)=\frac{p^2}{2m}+V\left(x\right), \tag{16.451}$$

然后把方程改成

$$\dot{x}=\frac{\partial H}{\partial p}, \tag{16.452}$$

$$\dot{p}=-\frac{\partial H}{\partial x}. \tag{16.453}$$

我们简单验证一下：

$$\dot{x}=\frac{\partial H}{\partial p}=\frac{p}{m}, \tag{16.454}$$

$$\dot{p}=-\frac{\partial H}{\partial x}=-\frac{d}{dx}V=F. \tag{16.455}$$

完全一致。

现在我们就有了目标了：我们需要把 $L\left(x\left(t\right),\dot{x}\left(t\right)\right)$ 变成一个新的函数 $L\left(x\left(t\right),p\left(t\right)\right)$，注意这里函数形式和自变量都发生了改变，但是，要求保持其中的物理过程，也就是描述物理过程的微分方程等价。这相当于要求所测量的物理量发生了变化，但是要求测量到的物理过程没变。这是一个非常非常高的要求。但是，实际上，不仅仅在力学问题中，在热力学问题中更常见：在热力学问题中，我们可以对同一个物理过程选择测量不同的能量函数，例如内能、熵、焓、自由能，我们希望看到通过测量这些不同的能量函数，得到的信息量——关于这个物理系统的状态和过程的信息——一样多。

这样的变换，既改变了自变量也改变了函数值和函数形式，但是要维持描述物理过程的方程不变[25]，却对超过了我们之前提到的两种函数变换。

那这样的变换怎么做呢？神奇的Legendre 变换来帮我们了。让我们从 L 的全微分 dL 开始。注意，我们之前提到过全微分包含了所有的信息。之前，我们求解泛函极值问题的时候也是从 dL 开始的。

$$dL=\frac{\partial L}{\partial\dot{x}}d\dot{x}+\frac{\partial L}{\partial x}dx=\frac{\partial L}{\partial\dot{x}}d\dot{x}+\frac{d}{dt}\left(\frac{\partial L}{\partial\dot{x}}\right)dx. \tag{16.456}$$

这里我们把运动方程，也就是 $\frac{d}{dt}\left(\frac{\partial L}{\partial\dot{x}}\right)=\frac{\partial L}{\partial x}$ 也写到了全微分里面去了。

注意，我们的目标是不想要 $\dot{x}$ 出现在表达式里面当作自变量。我们试试这样，

$$d\left(\frac{\partial L}{\partial\dot{x}}\dot{x}-L\right)=\dot{x}d\left(\frac{\partial L}{\partial\dot{x}}\right)-\frac{d}{dt}\left(\frac{\partial L}{\partial\dot{x}}\right)dx. \tag{16.457}$$

如果我们就定义新的自变量——注意这里的这个自变量的变换其实是得到了一个 $\dot{x}$ 和 p 之间的关系 $\dot{x}\left(p\right)$——为

$$p=\frac{\partial L}{\partial\dot{x}}, \tag{16.458}$$

而新的函数为

$$H\left(x,p\right)=p\dot{x}\left(p\right)-L\left(x,\dot{x}\left(p\right)\right), \tag{16.459}$$

那么，公式 (16.457) 就成了

$$dH=\dot{x}dp-\dot{p}dx. \tag{16.460}$$

[25]未来还会超越这个不变，而要求描述物理过程的方程协变——就是协调一致地变化，同样对应着物理过程不变。相对论就是如此。

根据全微分的一般表达式 $dU = \frac{\partial U}{\partial x}dx + \frac{\partial U}{\partial y}dy$，我们有

$$\begin{cases} \dot{x} = \frac{\partial H}{\partial p}, \\ \dot{p} = -\frac{\partial H}{\partial x}. \end{cases} \tag{16.461}$$

这个方程被称作Hamiltonian 方程，而其中的函数 $H(x,p)$ 被称为Hamiltonian——哈密顿量。

回到我们质点运动的例子，$L = \frac{1}{2}m(\dot{x})^2 - V(x)$。按照定义公式 (16.458)，

$$p = \frac{\partial L}{\partial \dot{x}} = m\dot{x}, \tag{16.462}$$

也就是说 $\dot{x} = \frac{p}{m}$。于是，按照定义公式 (16.459)，H 成为

$$\begin{aligned} H(x,p) = p\dot{x}(p) - L(x,\dot{x}(p)) &= \frac{p^2}{m} - \frac{1}{2}m\left(\frac{p}{m}\right)^2 + V(x) \\ &= \frac{p^2}{2m} + V(x). \end{aligned} \tag{16.463}$$

再运用Hamiltonian 方程公式 (16.461)，我们正好得到公式 (16.442)。

$$\begin{cases} \dot{x} = \frac{\partial H}{\partial p} = \frac{p}{m}, \\ \dot{p} = -\frac{\partial H}{\partial x} = F(x). \end{cases} \tag{16.464}$$

不过，具体的物理知识在这里不重要。我们通过这个例子展示的是：微积分，在这里尤其是全微分，才使得我们能够找到这样神奇的变换，它变换自变量，也变换函数值和函数形式，但是它维持物理过程也就是描述物理过程的运动方程——注意运动方程的形式也完全变了仅仅保持其含义，在这里就是公式 (16.447) 和公式 (16.461)——不变。物理上这个要求是自然的，但是没有微积分，我们不可能找到这样的变换。再一次，我希望你体会到微积分对于人类深层次思考的推动作用。

16.9.5 最优控制

我们已经看到了把运动问题转化成泛函极值问题的价值以及微积分在其中的作用。现在我们来看把控制问题表述为泛函极值问题以及微积分在其中的作用。控制问题在现实世界中广泛存在。例如，卫星发射场景下，我

们希望其最终达到的位置和速度接近理想值，中间我们可以通过火箭的动力来施加控制从而来实现这个目标。例如，在退休理财的场景下，我们希望初始有一笔固定的钱以及结合每个月领到的退休金之后，每一个月花的钱第一足够生活，第二还能尽可能花的比较多，第三留下一部分钱在理财项目里面继续生钱。当然，也可以加上末期约束[26]，也就是去世的时候还留下一笔遗产给家人或者其他人。这个问题中我们可以通过每个月花的钱来控制，控制的对象变量是当前的可以花的总的钱。

一般地，我们可以把问题表述为下面的称为“最优控制”的形式。在 t 时刻，系统有一个状态变量 $x(t)$，我们可以通过一个控制变量 $u(t)$ 去控制这个系统的演化，也就是微分方程 $\dot{x} = f(x,u,t)$ 或者差分方程 $x_{n+1} = f(x_n,u_n)$。我们控制的目标是使得下面的泛函取极值，

$$J = \int_0^T F(x(t),u(t),t)\,dt + S(x(T),T). \tag{16.465}$$

控制方式是选择合适的 $u(t)$。用经济学场景的语言，其中 $F(x(t),u(t),t)$ 表示过程中每时每刻时刻的收益，$S(x(T),T)$ 表示最终时刻剩下的钱的收益。

当我们的控制目标只考虑最终点的状态的时候，$F=0$。当我们不考虑最终点，或者最终点完全固定不可变的时候，$S=0$。

我们之前的最速下降线问题也可以表示为这样的控制问题 [12, 13]。实际上，我们要寻找的最速下降线相当于一个外力 $(u_x(t),u_y(t))$，系统的演化方程为 $\dot{x}=v_x,\dot{y}=v_y,\dot{v}_x=\frac{u_x}{m},\dot{v}_y=\frac{u_y}{m}$。接着我们就可以写下来相应的边界条件和优化目标 [12, 13]。我们也可以这样看最速下降线问题：我们要寻找的是充当控制方向作用的一个函数 $\left(u(x),v(x)=-\sqrt{1-u^2}\right)$，使得每时每刻的速度刚好沿着这个方向，也就是 $\frac{dx}{dt}=\sqrt{2g(-y)}u,\frac{dy}{dt}=\sqrt{2g(-y)}v$，于是总时间可以表述为

$$t = \int_{x_P}^{x_Q} \frac{dt}{dx}dx = \frac{1}{\sqrt{2g(-y)}u}dx. \tag{16.466}$$

把问题表述为最优控制的形式之后如何求解的问题我们就交给更加专门的关于最优控制的课程了。本质上仍然是Lagrangian 乘子法，而其背后

[26]其实还有一个活多久的问题，假设这个时间我们已知，按照某种平均值（例如类似的健康条件和生活方式的人）来估计。

仍然是微积分。当前阶段，看到通过微积分的语言可以把大量的问题转化为最优控制问题对我们来说就够了。毕竟用语言来表述问题，也就是数学建模的过程，比解决问题更加重要。除非遇到还没有解决方法的问题，那时候就要开展数学研究来找到这一类问题的解决方法，否则解决问题仅仅是一个迁移使用已有解决方法的问题。

16.9.6 傅里叶级数

前面我们已经在公式 (16.426) 看到了用三角函数，以及适合特定问题的某个算符（之前是拉普拉斯算符）的本征函数，来表示一般函数的例子。现在，我们来把为什么可以用三角函数逼近这些例子的原理讲清楚。我们把一个 $x \in [0, L]$ 上的周期函数 $\phi(x)$ 表示为

$$\phi(x) = \sum_{n=0}^{\infty} \left[a_n \cos \frac{n2\pi x}{L} + b_n \sin \frac{n2\pi x}{L} \right] \tag{16.467}$$

的方式称作 $x \in [0, L]$ 上的周期函数 $\phi(x)$ 的三角函数级数展开，也称为 $\phi(x)$Fourier 级数。

$[0, L]$ 上的周期函数最简单的形式就是 $\cos \frac{n2\pi x}{L}$ 和 $\sin \frac{n2\pi x}{L}$。我们也可以把这里的 $[0, L]$ 上的周期函数看作是：对于一个非周期的一般函数仅仅考虑某一段区间内的情况，去逼近这一段区间内的函数值，然后，我们的方法是把其他区间的看作是这一段区间的周期函数。这样，我们就可以考虑一个一般函数在任意一段区间内的逼近的问题了。

更进一步，我们还得证明这群三角函数的完备性，也就是任何一个 $[0, L]$ 上的周期函数都可以写成这些三角函数的叠加。这个任务我们放到后面来完成。

首先，我们来计算一下对于给定的 $\phi(x)$ 如何得到展开系数。为此，我

们先来算几个积分，

$$\begin{aligned}
&\int_0^L \cos\frac{m2\pi x}{L}\sin\frac{n2\pi x}{L}dx,\\
=&\frac{1}{2}\int_0^L\left[\sin\left(\frac{m2\pi x}{L}+\frac{n2\pi x}{L}\right)-\sin\left(\frac{m2\pi x}{L}-\frac{n2\pi x}{L}\right)\right]dx\\
=&\frac{1}{2}\int_0^L\left[\sin\frac{m2\pi+n2\pi}{L}x-\sin\frac{m2\pi-n2\pi}{L}x\right]dx\\
=&-\frac{1}{2}\left[\frac{L}{m2\pi+n2\pi}\cos\frac{m2\pi+n2\pi}{L}x\Big|_0^L\right.\\
&\left.-\frac{L}{m2\pi-n2\pi}\cos\frac{m2\pi-n2\pi}{L}x\Big|_0^L\right]\\
=&0.
\end{aligned}\tag{16.468}$$

再来看看

$$\begin{aligned}
&\int_0^L \cos\frac{m2\pi x}{L}\cos\frac{n2\pi x}{L}dx,\\
=&\frac{1}{2}\int_0^L\left[\cos\left(\frac{m2\pi x}{L}+\frac{n2\pi x}{L}\right)+\cos\left(\frac{m2\pi x}{L}-\frac{n2\pi x}{L}\right)\right]dx\\
=&\frac{1}{2}\int_0^L\left[\cos\frac{m2\pi+n2\pi}{L}x+\cos\frac{m2\pi-n2\pi}{L}x\right]dx\\
=&\frac{1}{2}\left[\frac{L}{m2\pi+n2\pi}\sin\frac{m2\pi+n2\pi}{L}x\Big|_0^L\right.\\
&\left.+\frac{L}{m2\pi-n2\pi}\sin\frac{m2\pi-n2\pi}{L}x\Big|_0^L\right]\\
=&\frac{L}{2}\delta_{m,n}.
\end{aligned}\tag{16.469}$$

不过这里有一个特例，当 $m=n=0$ 的时候，上述积分等于 L，

$$\int_0^L \cos(0\times x)\cos(0\times x)dx=L.\tag{16.470}$$

形式上不能被包含在上式之中。因此，对于 a_0 必须单独处理。正好，b_0 项永远等于零，没必要出现。因此，我们把公式 (16.467) 调整为下面的表达式

$$\phi(x)=\frac{a_0}{2}+\sum_{n=1}^{\infty}\left[a_n\cos\frac{n2\pi x}{L}+b_n\sin\frac{n2\pi x}{L}\right]\tag{16.471}$$

这样形式上，a_0 的表达式也就符合了公式 (16.469) 的结果。

最后，我们来看看

$$\begin{aligned}
&\int_0^L \sin\frac{m2\pi x}{L}\sin\frac{n2\pi x}{L}dx,\\
=&-\frac{1}{2}\int_0^L\left[\cos\left(\frac{m2\pi x}{L}\frac{n2\pi x}{L}\right)-\cos\left(\frac{m2\pi x}{L}-\frac{n2\pi x}{L}\right)\right]dx\\
=&-\frac{1}{2}\int_0^L\left[\cos\frac{m2\pi+n2\pi}{L}x-\cos\frac{m2\pi-n2\pi}{L}x\right]dx\\
=&\frac{1}{2}\left[\frac{L}{m2\pi+n2\pi}\sin\frac{m2\pi+n2\pi}{L}x\Big|_0^L\right.\\
&\left.-\frac{L}{m2\pi-n2\pi}\sin\frac{m2\pi-n2\pi}{L}x\Big|_0^L\right]\\
=&\frac{L}{2}\delta_{m,n}. \qquad (16.472)
\end{aligned}$$

然后，对公式 (16.467) 的左右两边乘以 $\cos\frac{m2\pi x}{L}$ 然后做积分，就可以得到

$$\int_0^L\cos\frac{n2\pi x}{L}\phi(x)\,dx=\sum_{n=0}^{\infty}a_n\frac{L}{2}\delta_{m,n}=a_m\frac{L}{2}, \qquad (16.473)$$

于是

$$a_m=\frac{2}{L}\int_0^L\cos\frac{m2\pi x}{L}\phi(x)\,dx. \qquad (16.474)$$

对公式 (16.467) 的左右两边乘以 $\sin\frac{m2\pi x}{L}$ 然后做积分，就可以得到

$$\int_0^L\phi(x)\cos\frac{n2\pi x}{L}dx=\sum_{n=0}^{\infty}b_n\frac{L}{2}\delta_{m,n}=b_m\frac{L}{2}, \qquad (16.475)$$

于是

$$b_m=\frac{2}{L}\int_0^L\phi(x)\sin\frac{m2\pi x}{L}dx. \qquad (16.476)$$

这样我们就把一个一般的 $[0,L]$ 上的函数，用一组数 $\{a_n,b_n\}$ 来表示。

我们通过几个例子来看看这个逼近的效果。

例 16.49 (傅里叶级数逼近 x 的效果). 取区间 $[0,1]$，我们来看看用傅里叶级数逼近函数 $f(x)=x$ 的效果。

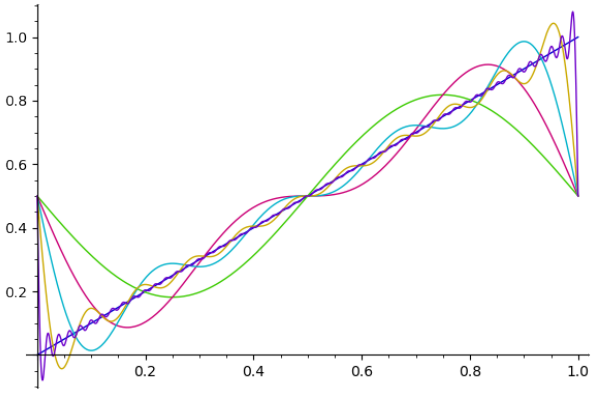


图 16.21: x **的傅里叶级数**。取了 $L = 1, 2, 4, 10, 50$ 项，分别对应着绿色、红色、蓝色、橘色和紫色的曲线。

$$a_n = 2\int_0^1 \cos(n2\pi x)xdx = 2\int_0^1 \cos(n2\pi x)xdx = \delta_{0,n}. \tag{16.477}$$

$$b_n = 2\int_0^1 \sin(n2\pi x)xdx = -\frac{1}{\pi n}. \tag{16.478}$$

于是

$$x_L = \frac{1}{2} - \sum_{n=1}^{L} \frac{1}{\pi n}\sin(n2\pi x). \tag{16.479}$$

而 $L \to \infty$ 的极限就是被逼近的函数 $f(x) = x$。

为了展示有限项的效果，在图 16.21 中我们分别展示了 $L = 1, 2, 4, 10, 50$ 项加起来的逼近效果。我们发现，随着 L 的增加逼近的效果越来越好。每新增加的一项提升的精度正比于 $\frac{1}{n\pi}$。这个随着 n 的增加而减小。

$\cos\frac{n2\pi x}{L}\sin\frac{n2\pi x}{L}$ 作为 $[0, L]$ 上的周期函数的完备性，可以从满足周期边条件的拉普拉斯算符的本征函数的角度来解决。在一维坐标上，拉普拉斯算符的本征函数表示为

$$u''_{xx} = \nabla^2 u = \lambda u. \tag{16.480}$$

满足这个方程的函数只有 $\sin(kx)$ 和 $\cos(kx)$，并且 $\lambda = -k^2$；或者 e^{kx}, e^{-kx}，并且 $\lambda = k^2$，这两组。但是，后者不可能满足周期边界条件。对于前者，满

足周期边界条件 $u(0)=u(L)$ 就得到

$$\sin(k0)=\sin(kL),\cos(k0)=\cos(kL). \tag{16.481}$$

也就是要满足 $KL=n2\pi$，也就是 $K=\frac{n2\pi}{L}$。于是，我们就得到最一般的解的形式是

$$u(x)=\sum_n\left[a_n\cos\left(\frac{2n\pi x}{L}\right)+b_n\sin\left(\frac{2n\pi x}{L}\right)\right]. \tag{16.482}$$

而这刚好就是Fourier 级数的一般形式。实际上这也是Joseph Fourier（约瑟夫 · 傅里叶）发现Fourier 级数的思路 [14]：通过求解热转到方程——其中的空间算符部分正好就是拉普拉斯算符——发现的拉普拉斯算符的本征函数为三角函数[27]，进而用三角函数表示一般解。

未来我们就可以用任意算符的任意边界条件下的本征函数来逼近一般的函数，从而把函数表示为用矢量。并且我们看到，往往这样的矢量的系数是越来越小的，取决于计算需要的精度前面若干项就够用来表达这个函数。这是高等数学对描述一般的函数提供的很好用的工具——级数展开，并且系数往往随着级数的阶数越大而越小。这样我们就再也不用给每个问题对应的函数取发明新的函数形式函数定义了，直接用级数，或者说一个理论上无穷个分量的矢量，但是实际上往往只需要有限个分量的矢量来表示就可以了。

后来人们还把 $[0,L]$ 上的傅里叶级数推广到 $(-\infty,\infty)$ 上的傅里叶变换。实际上，傅里叶级数和傅里叶变换已经成了解决大量的问题的基础，无论是计算上还是理论上，例如，量子信息、量子计算、求解微分方程和现代信号处理等问题。

小知识 16.3 (泛函分析). 用矢量的方式来研究函数的数学分支学科叫做泛函分析。其核心思想就是所有的函数或者说所有的满足某种性质的函数可以看作构成一个矢量空间，用某一组基函数——就像上面的三角函数和更一般的某个算符的本征函数——来把这些函数表示为一组数，就跟通常的二维或者更高维欧氏空间 $\mathbb{R}^n$ 的矢量一样。

[27]热传导问题的边界条件可以是固定边界条件，也就是两个端点的温度不同但是固定。这个和边界上取值为零等价。

未来在矢量空间的部分（见第??章）我们会证明函数确实构成一个矢量空间。

当然，除了形式上这样做，如果我们真的把 $\mathbb{R}^n$ 迁移到函数空间，也就是一个函数就是矢量空间中的一个点。我们还得定义某种函数之间的距离的概念，才能重新构建一套函数空间上的极限和微积分。

这里，甚至有一个很好的关于距离的启发。两点之间的距离的定义可以基于两点之间的矢量的长度来定义，而矢量完全可以把其中一个点移动到原点。也就是说，在矢量空间 $\mathbb{R}^n$ 中，长度可以这样定义，

$$r = \sqrt{\sum_{j=1}^{n} a_j^2}. \tag{16.483}$$

于是 P, Q 两点之间的距离可以这样定义，

$$r = \sqrt{\sum_{j=1}^{n} (P_j - Q_j)^2}. \tag{16.484}$$

那么，是不是在把函数用三角函数表示为一系列数 (a_n, b_n) 之后，长度仍然可以像下面这样定义呢？

$$r = \sqrt{\sum_{n} (a_n^2 + b_n^2)}. \tag{16.485}$$

如果是，回到函数的表达式，相当于定义了什么样的长度呢？很容易验证[28]实际上是这样的定义（a_0 项需要调整表达式，暂时先忽略这个问题），

$$|\phi| = \sqrt{\int_0^L \phi^2(x)\, dx}. \tag{16.486}$$

于是两个函数之间的距离就成了

$$(\phi, \psi) = \sqrt{\int_0^L [\phi(x) - \psi(x)]^2\, dx}. \tag{16.487}$$

下一步我们就要回答，这样的函数之间的距离的定义在函数本身是有意义的吗？可以用到什么场景？或者反过来，关于函数本身如果我们可以从应

[28]把公式 (16.471) 代入到公式 (16.486) 就可以验证。

用场景受到一些启发，那个时候函数之间的距离有该怎么定义？比如说，我们是不是可以用函数之间的距离来讨论一个函数和用来毕竟这个函数的多项式（见*Taylor* 公式公式 (16.79)）或者三角函数之和公式 (16.471) 之间的距离呢？

无论如何，一旦有了函数当作空间的点以及有了这些点之间的距离，我们就可以来定义极限了。有了极限，就可以定义这个空间的函数——给定一个这个空间的点也就是一个函数当作输入输出一个数，这正好就是我们提到的类似 $S\left[f\left(t\right)\right]$ 这样的泛函啊——的连续性和导数、积分等概念。

你看，除了从有限维到了无穷维这个关键差别我们还完全没有讨论之外，我们已经看到了，通过把函数看作矢量，泛函就成了通常的函数，就可以定义这个函数的微积分。**在这个过程中，我们也看到了为什么在我们学习微积分的时候，在数学家研究微积分的时候，需要构建微积分的严密的概念体系：未来构建新的对象的微积分的时候，只需要把之前的严密的体系迁移过去针对这个新的对象重新构建一次。这就是创造的形式之一——把一个学科知识结构迁移到全新的对象上去。**

顺便，我们还看到了，为了领会某些高层次的思想，例如上面的黑体字部分——构建和理解严密知识和创造的关系，你其实不一定需要真的把所有的细节都搞懂。你只需要明白整体思路和关键当然都上下左右贯通地彻底搞懂会更好。这个背后是系统思维和理解型学习 *[1]*。

16.10 推荐阅读材料

如果你想更细节地感受一下微积分受到的启发，微积分发展过程中的挑战和克服这些挑战的过程，以及背后的思想，推荐你看看Carl Boyer（卡尔 · 波耶）的《The History of the Calculus and Its Conceptual Development》[8]。

如果你想看一本严肃的关于微积分的科普，则推荐看看Steven Strogatz（史蒂文 · 斯托加茨）的《Infinite Powers: How Calculus Reveals the Secrets of the Universe》[15]。更一般的关于数学的严肃科普可以看Timothy Gowers（蒂莫西 · 高尔斯）的《牛津通识读本：数学》[16]。

如果你想更加严肃认真地学习微积分，你可以参考张筑生的《数学分

析新讲》第一册 [17]，Morris Kline（莫里斯 · 克莱因）的《Mathematical thought from ancient to modern times》[5] 和《Calculus: An Intuitive and Physical Approach》[18]，Michael Spivak（麦克 · 斯皮瓦克）的《Calculus》[19], Tom Apostol（汤姆 · 阿波斯托）的《Calculus, Vol. 1: One-Variable Calculus, with an Introduction to Linear Algebra》[20]，Terence Tao（陶哲轩）的《Analysis I》[21] 和《Analysis II》[22]，以及Walter Rudin（沃尔特 · 鲁丁）的《Principles of Mathematical Analysis》[23]。

如果你对本书背后的写作和教学的理念和方法感兴趣，可以阅读Alfred Whitehead（阿尔弗雷德 · 怀特海）的《Aims of Education》[24] 和吴金闪的《教的更少，学得更多》[1]。

16.11 作业

习题 16.22 (夜宵的数学建模). 我有 $M = 300$元 钱，打算去买啤酒和小龙虾吃。啤酒的价格是 $P_b = 10$元/瓶，小龙虾的价格是 $P_l = 80$元/斤。我习惯上吃两只小龙虾要和一口啤酒，大概也就相当于一斤小龙虾需要两瓶啤酒。如果两者不配合而是单独享用，我觉得比不上它们配合起来。我要怎么花这个钱，才能吃的最开心？你要获取哪些数据以及如何获取才能建立起来这个问题的数学描述？这个问题的数学描述大概怎么样？把“我”换成你自己或者你身边熟悉的人，尝试去建立这个问题的数学描述并解决它。

习题 16.23 (球体体积). 用积分算出来三维球体的体积。(提示：直角坐标和球坐标下的计算难度不同)

习题 16.24 (圆柱体体积). 用积分算出来圆柱体的体积。(提示：直角坐标和柱坐标下的计算难度不同)

习题 16.25 (圆锥体体积). 用积分算出来圆锥体的体积。(提示：直角坐标和柱坐标下的计算难度不同)

习题 16.26 (球体表面积). 用积分算出来三维球体的表面积。提示：直角坐标和球坐标下的计算难度不同)

习题 16.27 (圆柱体表面积). 用积分算出来圆柱体的表面积。(提示：直角坐标和柱坐标下的计算难度不同)

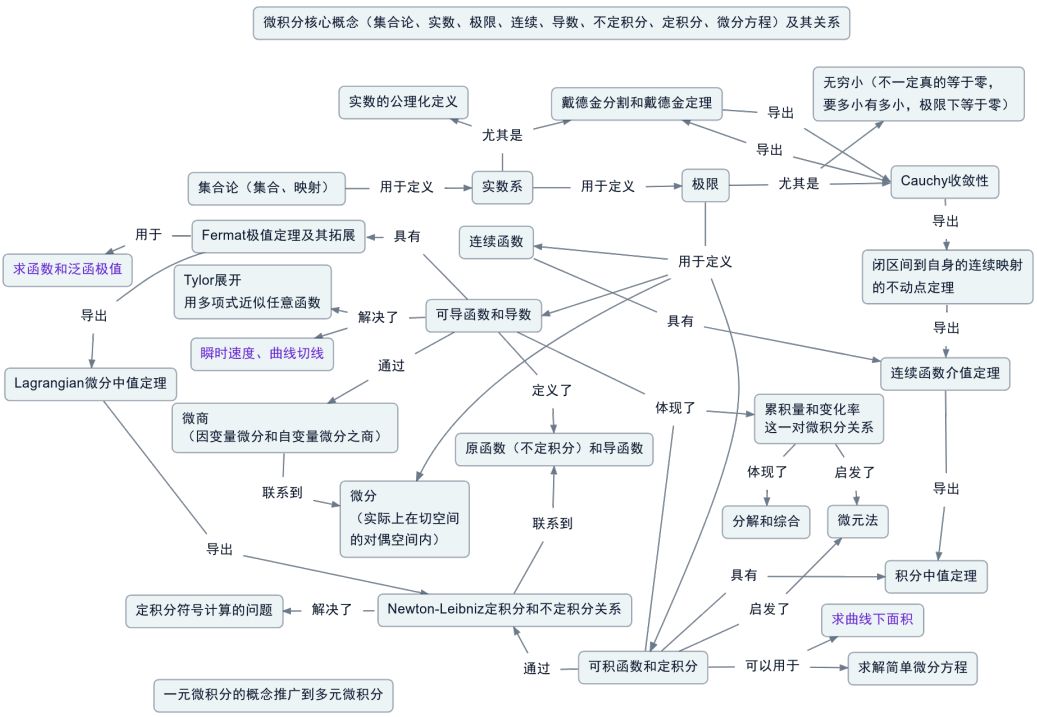


图 16.22: **微积分部分的知识体系**。其中彩色的概念节点代表了微积分受到的启发，也是微积分所解决的典型问题。

习题 16.28 (圆锥体表面积). 用积分算出来圆锥体的表面积。(提示：直角坐标和柱坐标下的计算难度不同)

习题 16.29 (横波波动方程的推导). 参考纵波的波动方程的推导，基于牛顿运动定理和微元法，推导一下横波的波动方程的推导。除了导数和极限，推导过程中你可能会用到三角函数。

16.12 本章小结

如图 16.22 所示，在本章中，我们先和创造微积分的前辈们一样，接受了当年提出导数和积分受到的启发，提出了不严格的但是可操作的导数和积分的定义，解决了一些启发导数和积分的问题。然后，我们对这个双向数学建模的过程做了总结梳理，看到数学是从解决现实问题中受启发，发展

数学的概念和命题也就是数学的学科概念知识体系以及数学学科的大图景(典型思维方式、典型分析方法、学科典型责任、典型研究问题、典型研究对象),以及把发展了的数学学科概念知识和学科大图景用来解决现实世界的问题。

接着,我们看到这样的不严格的微积分的问题,进而沿着前人的路径来解决这个问题使得微积分的知识体系严格化。从这里,我们还体会到这是数学发展另一个推动力——发现和解决数学知识体系里面的逻辑问题。

最后,我们展示了有了微积分之后人类的思考能力非凡提升的几个例子。

在具体学科概念知识上,我们仅仅保留了极限、导数(原则上就是极限)、积分(原则上还是极限)、微积分基本定理——牛顿-莱布尼兹公式。任何一个学科最终学生需要掌握的学科概念知识都应该是极少的,并且一旦学生掌握了学科大图景、一般性人类思维这些更高层次的知识,又受到了提出和解决问题的启发,学生完全就应该可以重新把这些学科概念知识"创造"出来。

甚至,从学科概念知识的学习的角度,你能够掌握不严格的能算的微积分就够了。从欣赏数学发展的角度,我希望读者你同时体会到从被启发被需求到真正建立微积分的知识体系的整体思路,看到其中的数学家们的艰苦卓绝的创造,看到微积分在解决问题上巨大的作用。但是,更加重要的目标是你可以通过微积分看到数学的学科大图景。尤其是其中的双向数学建模,也就是数学的学科典型责任。再看到数学的典型思维方式:这里有做抽象、逻辑演绎证明、分解和综合、极限、数学知识的系统性。再看到数学的典型分析方法:极限、微元法、函数极值问题和泛函极值问题。从运用数学知识的角度,我希望你能看到微积分的典型使用场景,在把微积分的学科概念知识和学科大图景理解好的基础上,未来可以至少参考这些典型使用场景来或远或近地迁移使用微积分的学科概念知识和学科大图景知识。

更高一层,我还希望你体会到学习和思考的方法。比如说,在思维方式的层面,不仅仅数学中有知识系统性、逻辑演绎证明、做抽象、从现实受启发再发展概念性知识再回到现实,其他学科甚至日常生活和工作中也需要运用这些思维方式。比如说,在学习方法的层面,你可能已经体会到:第一、学科的发展往往来自于现实世界或者学科自身知识体系的启发,所以,

经验体验这些知识以及学科已有的知识对于创造新知识很重要；第二、没有建立起来学科概念知识的过程，我们没法体会到学科大图景、一般性人类思维、教和学的方法这些更高层次的知识；第三，学习的目标不是记住所有的学科概念知识，甚至都不是形成学科概念知识的体系，而是掌握学科大图景、一般性人类思维、教和学的方法这些更高层次的知识，从而万一需要的时候，我们也可以通过提出和解决现实世界的问题或者学科已有知识中的问题来创造知识和创造性地使用知识，或者至少可以欣赏到其他人对知识的创造和创造性使用。

更更高一层，我还希望未来教材的写作者，能够借鉴本书的写作方式：交代一个学科概念知识受到的来自于现实世界或者学科知识内部的启发，看到学科概念知识之间的联系从而建构起来整个学科概念知识的体系，把学科概念的发展过程和应用联系到学科大图景，让读者体会到这些学科概念的典型应用场景，甚至从其中的某些场景再一次推动学科发展。这样，学习者在学习之后就很可能会成为一个问题解决者，甚至问题提出者，就是一个创造知识的人和创造性使用知识的人，也就是一个创造者。不能只有学科知识，但是没体系；不能有学科概念知识的体系，但是没有学科概念提出和发展所受到的启发，所体现的学科大图景和这些学科概念知识和学科大图景的典型应用。每一个点都不需要全，不需要多，而是需要帮助学习者体会深刻，用最典型的例子，体现最深刻的思想，用尽可能简单的学科概念知识。这就是“教得更少，学得更多”的思想，也是人类知识高速公路上以培养掌握高层知识生成器的理解型学习的核心思想 [1]。

最后，回头看一下微积分的整个学习过程，你从分数的四则运算律和函数（函数本身的概念，像三角函数、幂函数等一些具体的函数）这些初中的数学知识开始，已经学会了微积分的基本概念，领会了其中的思维方式，体验到了其中的分析方法，以及这些思维方式和分析方法在典型场景中是如何发挥作用的，了解了微积分的典型应用场景，体会到了微积分帮助人类思维实现的巨大的飞跃。你已经非常了不起了。低起点，启发和严格化要同时照顾到，高终点，是我在写作本书的时候的要求。其中有一些细节你可能很难第一轮就学会。这个没关系。先学会从外部去看它，看到这部分起到什么作用，传达什么信息。如果外部就能够看清了，不跳进去从内部看都无所谓。当然，如果这个内部本身很有意思，或者某些信息不跳进去其实体会

不到，那么，你最好还是跳一跳试试。

祝贺你，给你自己的思维的巨大的飞跃做好了准备。去享受一下生活的其他方面，然后我们稍后回来做另一个巨大的甚至我个人认为更大的飞跃——矢量代数。

第十七章　矢量代数

The beauty of mathematics only shows itself to more patient followers. ... You have to spend some energy and effort to see the beauty in math.

数学的美只会展示给耐心的追寻者。..... 你得花一些时间和努力才能看到。

– Maryam Mirzakhani

Mathematics compares the most diverse phenomena and discovers the secret analogies that unite them.

数学比较差异最大的现象们然后发现把它们统一起来的秘密。

– Joseph Fourier

A mathematician who is not also something of a poet will never be a complete mathematician.

如果一个数学家不是某种意义上的诗人，那么他/她就不可能是一个真正的数学家。

– Karl Weierstrass

Mathematics is the art of reducing any problem to linear algebra.

数学就是把一切问题都转化为线性代数的艺术。

– William Stein

The purpose of computation is insight, not number.

计算是为了获得洞见，而不是得数。

– Richard Hamming

I personally believe that many more people need linear algebra than calculus.

我个人相信对许多人来说，相比微积分，他们更需要线性代数。

– Gilbert Strang

17.1 矢量受到的启发

矢量的概念首先受到空间矢量的启发。空间矢量是我们从之前的数学知识的学习和日常生活中就已经看到的数学结构。例如，我们先往东边走 3 米，再往北边走 4 米，我们就会到达一个相对于起点来说的东北方向距离为 5 米的地方。这里，我们用到了矢量的概念——（地球“平面”上[1]）从起点到终点的有向线段、勾股定理和空间矢量的加法——首尾相连。$\vec{a}+\vec{b}$ 也就是矢量 b 的起点（首）连接到矢量 a 的终点（尾）上合起来从 a 的起点到 b 的终点的有向线段。

因此，我们就可以把空间矢量通过其起点和终点来定义。也就是说起点 P 和终点 Q 定义了矢量 $P \to Q$，记作 $\overrightarrow{PQ}$。以平面上的矢量为例，结合平面直角坐标系，我们就有 $\overrightarrow{PQ} = P \to Q = (x_P, y_P) \to (x_Q, y_Q)$。

在做矢量加法的时候我们需要移动矢量的起点，于是终点也跟着移动了，因为移动的时候要保持第二个矢量不变。因此我们得经常去平行移动矢量——也就是对这个矢量的起点和终点的坐标加上同一个值[2]于是，我们把所有的起点不同但是一旦平行移动到相同的起点以后都相同的矢量认同。也就是说，$(x_P, y_P) \to (x_Q, y_Q)$ 和 $(x_P + \Delta x, y_P + \Delta y) \to (x_Q + \Delta x, y_Q + \Delta y)$

[1]未来，不一定只考虑在地球上的矢量，地球也不一定是平面。但是，按照我们日常生活经验，我们先这么认为。

[2]在球表面等弯曲的空间中如何平行移动是一个很有意思的问题。但是，我们这里默认在平直空间中讨论移动。至于在平直空间中矢量的平行移动就是“矢量的起点和终点的坐标加上同一个值”我们目前基于生活经验和过去的数学知识来理解。未来如果你有幸学习到微分几何，那里会有定义更一般的平行移动的定义以及明白为什么平直空间中矢量的平行移动是这样的。

是同一个矢量。于是，一个自然的办法就是把矢量定义为 “其终点坐标和起点坐标之差”，也就是

$$\overrightarrow{PQ} = P \to Q = (x_Q, y_Q) - (x_P, y_P) = (x_Q - x_P, y_Q - y_P)\,. \tag{17.1}$$

这里我们还补充了一个关于两个点的坐标之差的定义 $(x_Q, y_Q) - (x_P, y_P) = (x_Q - x_P, y_Q - y_P)$。这实际上也是两点坐标之和以及矢量乘以一个数的定义，

$$(x_Q, y_Q) + (x_P, y_P) = (x_Q + x_P, y_Q + y_P)\,, \tag{17.2}$$

$$a \times (x_P, y_P) = (a \times x_P, a \times y_P)\,. \tag{17.3}$$

来得到

$$\begin{aligned}(x_Q, y_Q) - (x_P, y_P) &= (x_Q, y_Q) + (-1) \times (x_P, y_P) \\ &= (x_Q, y_Q) + (-x_P, -y_P) \\ &= (x_Q - x_P, y_Q - y_P)\,. \end{aligned} \tag{17.4}$$

我们来用这样的语言重新表述一下矢量加法 $\vec{a} + \vec{b}$。假设 $\vec{a}$ 和 $\vec{b}$ 的终点分别是 O, E, N，我们有

$$\begin{aligned}\vec{a} + \vec{b} = \overrightarrow{OE} + \overrightarrow{EN} &= (x_E - x_O, y_E - y_O) + (x_N - x_E, y_N - y_E) \\ &= (x_E - x_O + x_N - x_E, y_E - y_O + y_N - y_E) \\ &= (x_N - x_E, y_N - y_E) = \overrightarrow{ON}. \end{aligned} \tag{17.5}$$

刚好就是我们所期待的。合起来，从平面上的矢量或者日常生活中的矢量，我们得到的启发是，一个矢量得有

- 矢量的含义:起点指向终点的有向线段,用坐标可以表示为公式 (17.1)。
- 两个矢量的加法 $\vec{a} + \vec{b}$: 用坐标可以表示为公式 (17.2)，
- 矢量的数乘 $\mu\vec{a}$，也就是矢量乘以一个数，用坐标可以表示为公式 (17.3)，

进一步依赖于生活经验和之前学习过的平面上的坐标和矢量，我们还可以得到，矢量加法满足交换律 $\vec{a} + \vec{b} = \vec{b} + \vec{a}$。从生活经验上，我们先往北

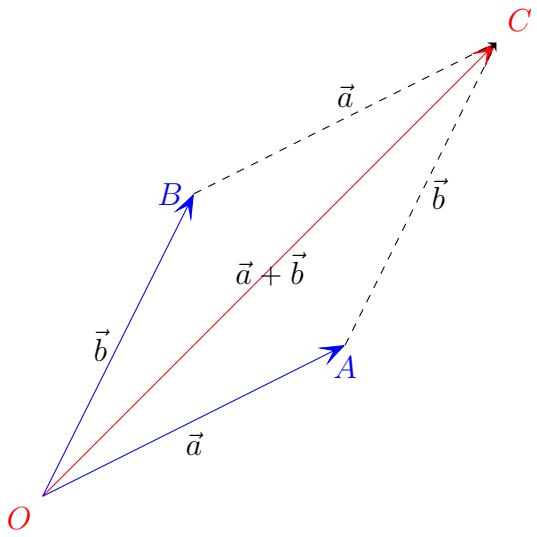


图 17.1: **矢量叠加的平行四边形表示**。矢量叠加的平行四边形法则保证了 $\vec{a}+\vec{b}=\vec{b}+\vec{a}$。注意：平行的而且长度相等的矢量可以平移以后重叠，因此都是同一个矢量。

边走 4 米，再往东边走 3 米，我们仍然会到达一个相对于起点来说的东北方向距离为 5 米的地方。更一般地，我们可以画一个平行四边形来验证这个可交换性，如图 17.1：根据平行四边形的性质以及我们把平移以后的矢量认同，把 $\vec{b}$ 平移以后接到 $\vec{a}$ 上还是把 $\vec{a}$ 平移以后接到 $\vec{b}$ 上，得到的结果相同。如果你学习过物理学，你也可以从力、速度和加速度受到启发，它们也满足前面的三条，在加上矢量加法的交换律。

或者从依赖于坐标的计算上，我们可以这样来验证：

$$\vec{a}+\vec{b}=(a_x,a_y)+(b_x,b_y)=(a_x+b_x,a_y+b_y)=\vec{b}+\vec{a}. \tag{17.6}$$

于是，我们得到了矢量加法满足交换律。所有的这些通过生活经验、物理学、平面上的矢量得到的关于矢量的启发，都需要被保留到一般的矢量的概念之中吗？比如说，其实从平面矢量都到力和速度、加速度，第一条“矢量是有向线段”的含义就不怎么适用了。未来我们会进一步来思考这个问题。

另一个问题是，这些个条件够了吗，还需要补充其他条件吗？这个问题我们也稍后再讨论。

再剩下的问题是，那位移、速度和加速度、力都是矢量，空间位置看作从零点到这个位置的矢量则也是矢量，数看起来那也是一维矢量，那是不是其实所有的对象都是矢量啊，于是矢量的概念也就没有了意义——既然可以描述所有对象，也就不再需要这个概念来区分对象了。我们先来回答这个问题。

幸好不是的，很多对象不是矢量。钱的多少不是矢量。原则上，如果我们把欠钱记作负数，则钱也有正负，可以做加减，可以有数乘，但是，没有方向。如果你足够仔细，你问：是不是钱很多维的时候就是矢量了呢？如果这个很多维指的是很多种货币，那么，实际上也不是矢量：尽管每一种货币都有正负，可以做加减，可以有数乘，于是合起来多种货币很像矢量，但是由于货币之间还存在着可以相互转化的汇率，因此，多种汇率也要实际上转换以后看作一种。因此，还是没有方向性。比较麻烦的情景是，在优化问题中，当你考虑购买物品的时候，钱（m）和物品的数量（q）是你要考虑的两个维度，这样构成的 (m, q) 算不算矢量。这个时候，我们就需要具体问题具体考虑了。有的时候可以看作是矢量。其次，质量（或者用非物理的语言说，物体的重量）也不是一个矢量：只有正，可以做加减，可以有数乘，但是，没有方向。体积、能量、电压、人数等等，都不是矢量。我们把这些没有方向的量称为标量[3]。

未来我们还会接触到比矢量更加复杂的东西，例如把平面物体或者三维物体转化为大小不同的二维图象的相机，把地球画成地图的地理画图法，把你的意思表达成字词的线性序列的你的写作功能以及相应的大脑，把已经写成字词的线性序列转化为你所理解到的意思的你的阅读功能以及相应的大脑，在不同的能量形式之间相互转换的像电动机和发电机这样的物理设备。

因此，矢量不过就是比标量稍微复杂一点的数学对象，而且很多现实世界的对象是矢量，也有更多不是矢量。

现在，我们回来继续讨论什么是矢量的问题。前面那几条都要保留吗，够了吗？这取决于我们像用矢量来描述什么。也就是说，我们需要继续从已有的数学、物理等学科知识来受启发，继续从现实世界获得启发。

方向还意味着有夹角的概念。可以是从一个固定的参考线到这个矢量

[3]如果你有幸学习到微分几何，会看到标量的更严格的定义。

的夹角，或者任意两个矢量之间的夹角。对于二维平面上的矢量，其夹角是这样计算的

$$\cos\theta = \frac{\vec{a}\cdot\vec{b}}{|\vec{a}|\left|\vec{b}\right|}. \tag{17.7}$$

其中 $\vec{a}\cdot\vec{b}$ 是两个矢量的内积，$|\vec{a}|$ 是矢量的长度。

考虑一个平面上的矢量，例如钟表表面的指针，我们发现如果我们把钟表的指针也算做矢量的话，这样的矢量还可以被旋转。也就是一个矢量被旋转以后还是一个矢量。当我们考虑相似图形的时候，则我们还得允许矢量可以被缩放。

那么，如果我们还是用平面直角坐标系的两个数 (a_x, a_y) 来描述一个矢量 $\vec{a}$——其实对应着从原点到 $A=(a_x,a_y)$ 的有向线段的话，我们如何来描述转动和缩放呢?单纯的缩放似乎就是其面已经提到的数乘 $\mu\vec{a}=(\mu a_x, \mu a_y)$——方向不变长度发生变化。旋转就比较复杂了。但是，通过之前学习过的相差九十度的角的三角函数，就可以知道，如果是转动九十度，应该就成了 $(-a_y, a_x)$，记作

$$(-a_y, a_x) = \vec{a}' = R_{90^\circ}\vec{a} = R_{90^\circ}(a_x, a_y). \tag{17.8}$$

我们发现，这里的缩放和旋转都具有线性性，也就是

$$\mu\left(\eta\vec{a}+\xi\vec{b}\right) = \eta\mu(\vec{a}) + \xi\mu\left(\vec{b}\right), \tag{17.9}$$

以及

$$R_{90^\circ}\left(\eta\vec{a}+\xi\vec{b}\right) = \eta R_{90^\circ}\vec{a} + \xi R_{90^\circ}\vec{b}. \tag{17.10}$$

第一条我们就不验证了。我们来基于公式 (17.8) 验证一下第二条。

$$\begin{aligned} R_{90^\circ}\left(\eta\vec{a}+\xi\vec{b}\right) &= R_{90^\circ}(\eta a_x + \xi b_x, \eta a_y + \xi b_y), \\ &= (-\eta a_y - \xi b_y, \eta a_x + \xi b_x), \\ &= (-\eta a_y, \eta a_x) + (-\xi b_y, \xi b_x), \\ &= \eta(-a_y, a_x) + \xi(-b_y, b_x), \\ &= \eta R_{90^\circ}\vec{a} + \xi R_{90^\circ}\vec{b}. \end{aligned} \tag{17.11}$$

是不是一般的对矢量的转动 R_θ 也具有线性性呢?

反过来，如果一个操作 f 具有线性性

$$f(\mu a+\nu b)=\mu f(a)+\nu f(b), \tag{17.12}$$

那么，自然也就说明同时有了 f 和 a 的数乘和加法。那么，是不是其实线性性是矢量真正的内涵呢?

比如说，之前我们学习过的函数的求导运算和求积分运算都是具有线性性的，见公式 (16.51) 和公式 (16.162)，是不是意味着：函数和函数的求导和积分运算都是矢量呢? 它们的加法和数乘都是有自然的定义的。如果它们是矢量，它们的长度、角度、旋转、缩放（这个好办，就是数乘）这些东西怎么定义? 之前我们使用的矢量的基矢和坐标分量的形式呢，在函数和函数的求导和积分运算里面都表现为什么?

回忆一下 16.9.6 节中的用三角函数来表示一般的函数。那时候我们就说，用了三角函数之后，一个一般的函数被表示成了一堆系数 $\{a_n,b_n\}$。是不是这些数就很像是矢量的分量形式? 顺便，被表示为这样的形式以后，

$$\phi(x)=\frac{a_0}{2}+\sum_{n=1}^{\infty}\left[a_n\cos\frac{n2\pi x}{L}+b_n\sin\frac{n2\pi x}{L}\right] \tag{17.13}$$

求导数和积分的运算也会特别简单。例如，对于求导数，

$$\frac{d}{dx}\phi(x)=\frac{02\pi}{L}\frac{a_0}{2}+\sum_{n=1}^{\infty}\frac{n2\pi}{L}\left[-a_n\sin\frac{n2\pi x}{L}+b_n\cos\frac{n2\pi x}{L}\right] \tag{17.14}$$

或者表示为

$$a_n\to\frac{n2\pi}{L}b_n,b_n\to-\frac{n2\pi}{L}a_n,a_0\to 0. \tag{17.15}$$

也就是说，求导就成了这些系数的代数运算——乘上一个相应的 $\frac{n2\pi}{L}$。积分也类似。

习题 17.1 (傅里叶级数的积分)**.** 证明对于傅里叶级数，除了 a_0 项，不定积分成为 $a_n\to-\frac{L}{n2\pi}b_n,b_n\to\frac{L}{n2\pi}a_n$。

因此，函数、对函数求导数、对函数做积分，都可以被表示为矢量。注意，这里把函数称作当作矢量，不是说函数都是线性函数 $f(\mu a+\nu b)=\mu f(a)+\nu f(b)$（可以证明只有 $f(x)=kx+b$ 才能满足这个要求），而是

函数可以被线性地操作，于是函数之间的加法（$f+g$）和函数乘上一个数（μf）是可以自然定义的，并且这个加法和数乘看起来很像欧氏空间的矢量。操作其他对象时候的线性性、自身的线性性、被其他对象操作的时候的线性性，并不完全是一个东西。

此外，从一个对象可以被如何操作——在数学上这些操作会成为运算会成为关系，这些操作具有什么性质的角度来定义一个矢量（以及未来会看到那些更加一般的数学对象），被称为代数。关于这一点，我们即将有更加深刻的体会。顺便，这就是为什么关于矢量的数学被称为线性代数的原因——矢量的核心是线性性。用了这样的定义方式，而不是之前小学学习的时候先搞清楚参与运算的对象和运算的含义再来总结这些运算满足的性质的方式，就可以把定义出来的对象用于描述任何一个参与运算的对象。这就是代数——适用于代表任何一个对象甚至都不仅仅是数的数学结构，也就是抽象性。合起来，线性代数指的是一种关于满足线性性的对象的数学结构，并且这个数学结构是通过这些对象参与的运算的性质来定义的，而不是先通过搞清楚对象是什么来定义的，具有抽象性、这些对象也不一定是数。

我们还可以把什么东西看作矢量，其中会用到矢量的相加、数乘、旋转、缩放、分量表示，会具有被其他对象来操作的时候的线性性，或者自身的线性性，或者去操作其他对象的时候的线性性？把我们这部分的目标记在心里：我们希望把什么是矢量尽可能一般地但是又不能太松地提炼出来。

我们来尝试看更多的长得像矢量的对象。比如说，数学概念之间的关系是不是也很像矢量相加？例如，“直角”和“三角形”这两个概念合起来是“直角三角形”。“正方形”和“周长”这个概念合起来是“正方形的周长”。甚至，一般的语言也很像矢量相加。例如，“中国的首都”是“中国”和“首都”这两个概念合起来，并且还等于“北京”这个城市。类似地，“日本”和“首都”合起来，“美国”和“首都”合起来，都有其含义，这个表达方式都成立。甚至，我们，没准可以发现，“中国——北京”的关系，“日本——东京”的关系，“美国——华盛顿”之间的关系，在某种表示下是一种。

比如说，我出这样一道题，下列那一组关系和其他组的不同：“中国——北京”、“日本——东京”、“美国——华盛顿”、“法国——马赛”，你肯定可以选出来“法国——马赛”。在这里，实际上，你就做了每组的两个对

象之间的关系的计算，然后，看哪些关系比较接近，哪些关系离得比较远。

如果我们可以用矢量来表达，相当于我们在计算下面这些矢量：$C_{cb} = C_{\text{中国——北京}}$、$C_{jt} = C_{\text{日本——东京}}$、$C_{aw} = C_{\text{美国——华盛顿}}$、$C_{fm} = C_{\text{法国——马赛}}$，然后发现，$C_{cb}, C_{jt}, C_{aw}$ 很像——具体可能是方向相似或者不仅方向相似并且长度也差不多相等，但是 C_{fm} 跟其他的差别很大，尽管我们还不知道这里的方向和长度是什么意思，怎么表示出来，如何做对比。

如果词可以用矢量表示，那么是不是一句话的意思，一篇文章的意思，都可以用一个矢量来表示。这个矢量构成的空间应该就是所有的含义构成的空间。我们暂时就把它称作概念含义空间。如果确实可以这样做，那不同语言之间的翻译，是不是就可以看成是概念空间的同一个含义在不同语言的词汇和语法先的表示，就跟同一个欧氏空间的矢量在不同的基矢的选择下的分量表示一样？甚至，人类的写作过程是不是就是把这个意思的网络翻译成一个特定语言的时间序列，也可以看作是某个线性变换。反过来的阅读过程也是如此。当然，在语言现象中，可能线性性不是跟欧氏空间的矢量那样这么好。不过，至少这是一个可以尝试的来看待语言甚至处理语言的思路。实际上，这个思路——用概念含义空间的矢量来表示篇章、句子、词汇等语言单位——就是当前兴起的自然语言大模型的基础 [25, 26]。

有了矢量的长度、方向、加法、数乘（缩放）、内积、旋转、被操作的线性性、操作其他对象的线性性这几个概念，我们终于可以在下面开始严肃地定义矢量了。

顺便，为什么要把那些能够看作矢量的东西看作是矢量呢，为什么要让矢量这个概念在保留其核心的条件下尽量宽泛呢？一旦我们能够把一个对象看作是矢量，我们就可以用一堆数，也就是一个矢量在所选择的基矢下的分量，来表示这个对象了。这样，我们就把关于这个对象的研究变成了数的加减乘除等数学操作的研究。也就是说，在这里，我们追求的是对对象的矢量表示，达到这个对象对其他对象的操作、其他对象对这个对象的操作、一切关于这个对象的问题，都可以用这个矢量表示来表达。当然，更一般的问题和更一般的对象没准会需要其他的数学结构来表示。但是，很自然地我们追求把尽可能多的对象都表示为矢量。

17.2 欧氏空间的抽象矢量表示

为了给理解后面一般的矢量的定义和记号做准备，我们来把我们熟悉的二维欧氏空间的矢量用新的记号表示一下。根据笛卡尔平面直角坐标系，一个二维平面上的点 P 可以表示为 (x_P, y_P)。这也可以表示以坐标原点 O 到 P 的矢量，

$$\overrightarrow{OP} = \begin{bmatrix} x_P \\ y_P \end{bmatrix} = [x_P, y_P]^T . \tag{17.16}$$

其中的方括号也可以替换为圆括号，这里的符号 a^T 表示对 a 的转置，就是把行的形式变成列的形式，或者把列的形式变成行的形式。以后会更一般地来定义这个转置操作。如果是一般的从 P 指向 Q 的矢量，可以写做

$$\overrightarrow{PQ} = \begin{bmatrix} x_Q - x_P \\ y_Q - y_P \end{bmatrix} = [x_Q - x_P, y_Q - y_P]^T . \tag{17.17}$$

在二维矢量空间也就是二维欧氏空间中，由于我们把点 P 和矢量 OP 认同了，我们就把这样的排成一行（一列）的两个数称为行矢量（列矢量）。

现在，我们这样来写一个二维矢量，

$$\overrightarrow{OP} = x_P \hat{e}_1 + y_P \hat{e}_2. \tag{17.18}$$

其中，$\hat{e}_1$ 和 $\hat{e}_2$ 分别被称作 x 和 y 方向的单位矢量，而 x_P 和 y_P 看作是乘上这些单位矢量的系数。

我们约定，未来会看到这样约定的理由，行矢量乘以列矢量的规则如下，

$$[a_1, a_2] \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = a_1 b_1 + a_2 b_2. \tag{17.19}$$

于是，在这个规则下

$$\overrightarrow{OP} = [\hat{e}_1, \hat{e}_2] \begin{bmatrix} x_P \\ y_P \end{bmatrix} = x_P \hat{e}_1 + y_P \hat{e}_2. \tag{17.20}$$

在这个记号下，$\begin{bmatrix} x_P \\ y_P \end{bmatrix}$ 成了矢量 $\overrightarrow{OP}$ 在给定的一组基矢 $\hat{e}_1, \hat{e}_2$ 下的系数列。

至于 $\hat{e}_1, \hat{e}_2$ 到底是什么这是写不出来的，所以它们被称为抽象矢量记号。如果在它们自己的基矢下写那就是，它们的系数列可以这样

$$\hat{e}_1 \to \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \hat{e}_2 \to \begin{bmatrix} 0 \\ 1 \end{bmatrix}. \tag{17.21}$$

你没准会说，把矢量 $\overrightarrow{OP}$ 的列矢量形式，公式 (17.16) 的第一个等号项，看作

$$\overrightarrow{OP} = \begin{bmatrix} x_P \\ y_P \end{bmatrix} = x_P \begin{bmatrix} 1 \\ 0 \end{bmatrix} + y_P \begin{bmatrix} 0 \\ 1 \end{bmatrix}. \tag{17.22}$$

然后，和公式 (17.18) 去对比，就发现，

$$\hat{e}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \hat{e}_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}. \tag{17.23}$$

我们说，不是这样的，在公式 (17.16) 中的 $\overrightarrow{OP}$ 是分量形式，已经取了一组基矢的条件下，而公式 (17.18) 中的 $\overrightarrow{OP}$ 是没有取定任何一组基矢的情况下的表达式，也就是矢量的抽象表达式。公式 (17.21) 仅仅表示基矢在选定自己当作基矢的条件下的列矢量形式，而不是说基矢本来就是那些列矢量。这一步是矢量的抽象记号的核心。

这个时候，如果我们用另一组基矢，例如

$$\hat{e}'_1 = \frac{\hat{e}_1 + \hat{e}_2}{\sqrt{2}}, \hat{e}'_2 = \frac{\hat{e}_1 - \hat{e}_2}{\sqrt{2}} \tag{17.24}$$

则，在这一组新的基矢下，原来的基矢被表达为

$$\hat{e}_1 \to \begin{bmatrix} \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} \end{bmatrix}, \hat{e}_2 \to \begin{bmatrix} \frac{\sqrt{2}}{2} \\ -\frac{\sqrt{2}}{2} \end{bmatrix}. \tag{17.25}$$

我们看到同一个矢量的列矢量形式是会变化的！但是，同一个量就应该是同一个量！因此，我们预期只需要永远把抽象形式写完，就能发现，其实同一个量仍然是同一个量，例如

$$\hat{e}_1 = \frac{\sqrt{2}}{2}\hat{e}'_1 + \frac{\sqrt{2}}{2}\hat{e}'_2 = \frac{\sqrt{2}}{2}\frac{\hat{e}_1 + \hat{e}_2}{\sqrt{2}} + \frac{\sqrt{2}}{2}\frac{\hat{e}_1 - \hat{e}_2}{\sqrt{2}} = \hat{e}_1. \tag{17.26}$$

把一个矢量的分量形式——不管是列矢量形式还是行矢量形式——和其相应的抽象表示的基矢写在一起是一个非常好的记号。在抽象记号下，一个矢量是什么它就是什么，不会随着采用不同的基矢而变化。变化的仅仅是这个矢量的分量形式。未来我们还会发现，这就能够自然地导出一个矢量在不同基矢情况下的分量表示。这通常被成为矢量的坐标变换。在我们这里是抽象矢量记号的自然结果。

我们再来引入矢量的另一套抽象记号——矢量的Dirac 符号表示。二维矢量空间中是有长度和方向的定义的，而长度和方向是基于内积定义的。用矢量分量形式来表述，二维矢量 $\vec{a},\vec{b}$ 的内积定义为

$$\left(\vec{a},\vec{b}\right)=a_xb_x+a_yb_y. \tag{17.27}$$

当 $\vec{a}=\vec{b}$ 的时候，$(\vec{a},\vec{a})\geqslant 0$，因此 $\sqrt{(\vec{a},\vec{a})}$ 可以当作矢量的长度，记作 $|\vec{a}|$ 或者 a，并且刚好和勾股定理距离公式相符。有了长度，就可以定义两个矢量的夹角，

$$\theta=\arccos\frac{\left(\vec{a},\vec{b}\right)}{ab}. \tag{17.28}$$

基于内积，我们来定义一个映射 $(\vec{a},\cdot)|_{\mathbb{R}^2\to\mathbb{R}}$。这个映射是完全由 $\vec{a}$ 决定的。对于任何一个输入的矢量 $\vec{b}$，这个映射的输出是一个实数。可以验证这个映射是线性的，

$$\left(\vec{a},\mu\vec{b}+\nu\vec{c}\right)=\mu\left(\vec{a},\vec{b}\right)+\nu\left(\vec{a},\vec{c}\right). \tag{17.29}$$

习题 17.2 (左矢量映射的线性性)**.** 根据公式 (17.27) 证明映射 $(\vec{a},\cdot)$ 的线性性，也就是公式 (17.29)。

同样地，我们也可以建立一个这样的映射 $(\cdot,\vec{a})|_{\mathbb{R}^2\to\mathbb{R}}$。这个映射是完全由 $\vec{a}$ 决定的。对于任何一个输入的矢量 $\vec{b}$，这个映射的输出是一个实数。这个映射也是线性的。

$(\vec{a},\cdot)$ 完全对应着矢量 $\vec{a}$，$(\cdot,\vec{a})$ 也是。是不是它们就是一个东西呢？比如说，它们的加法和数乘的结果都一一对应吗？我们来检验一下。首先，

$$\mu\left(\cdot,\vec{a}\right)+\nu\left(\cdot,\vec{b}\right)=\left(\cdot,\mu\vec{a}+\nu\vec{b}\right), \tag{17.30}$$

这个可以这样来验证，对于任何的 $\vec{c}$，我们有

$$\mu\left(\vec{c},\vec{a}\right)+\nu\left(\vec{c},\vec{b}\right)=\left(\vec{c},\mu\vec{a}+\nu\vec{b}\right). \tag{17.31}$$

同样地，对于任何的 $\vec{c}$，我们有，对于任何的 $\vec{c}$，

$$\mu\left(\vec{a},\vec{c}\right)+\nu\left(\vec{b},\vec{c}\right)=\left(\mu\vec{a}+\nu\vec{b},\vec{c}\right), \tag{17.32}$$

因此，也就有

$$\mu\left(\vec{a},\cdot\right)+\nu\left(\vec{b},\cdot\right)=\left(\mu\vec{a}+\nu\vec{b},\cdot\right). \tag{17.33}$$

于是对矢量 $\vec{a}$ 能做的加法和数乘，就成了对 $(\vec{a},\cdot)$ 和 $(\cdot,\vec{a})$ 能做的加法和数乘。也就是说，$\vec{a}$、$(\vec{a},\cdot)$ 和 $(\cdot,\vec{a})$ 完全是一个东西。

Paul Dirac（保罗 · 狄拉克）注意到了这个事实，把 $(\vec{a},\cdot)$ 和 $(\cdot,\vec{a})$ 分别记作

$$(\vec{a},\cdot)\triangleq\langle a|, \tag{17.34}$$

$$(\cdot,\vec{a})\triangleq|a\rangle, \tag{17.35}$$

相应地称为一个矢量的做矢量形式和右矢量形式，也称为一个矢量的对偶矢量和矢量形式。在这个记号下，内积自然地成为

$$\langle a|\ b\rangle=\left(\vec{a},\vec{b}\right). \tag{17.36}$$

一个矢量写成基矢和分量的形式就成了

$$\langle a|=a_x\langle e_1|+a_y\langle e_2|, \tag{17.37}$$

$$|a\rangle=a_x|e_1\rangle+a_y|e_2\rangle. \tag{17.38}$$

我们发现，这个记号也使得一个矢量明确地把分量和基矢量用相加的方式写在了一起，而不是用一行或者一列数来表示一个矢量。这正和我们之前建议使用的矢量的抽象记号公式 (17.20) 一模一样。

做一下矢量的抽象记号的小结，注意，这个抽象矢量记号——公式 (17.20) 或者公式 (17.38)——真的非常非常重要：**把一个矢量的分量形式和相应的抽象形式的基矢写在一起写成相加的形式，可以使得一个矢量的**

整体形式永远不变，不随着基矢的选择而变化（而其分量形式自然地随着基矢的选择而变化[4]）。

在这个记号下，矢量不是一列或者一行数，这些数仅仅是其分量形式，一切的矢量性全都来源于抽象的基矢。相比于通常的把一列或者一行数看作矢量，这是非常非常大的差别。

下面，我们开始严肃地来定义矢量。

17.3 群环域和矢量空间

为了定义矢量，我们先来定义群环域。我们从集合和映射开始来定义群环域。

我们来复习一下集合和映射。用朴素的语言来说，集合就是一堆东西，满足元素的确定性，以及无序性、互异性。不过，后面两个性质其实不重要。如果出现了相同的，那我们就每一个相同的都算一个。如果有序，那我们就忽略顺序。因此，集合的关键性质就是对于给定的任何一个元素，这个元素要么属于这个集合，要么不属于这个集合，两者不能同时成立，两者必然有一个成立，就行。映射就是两个集合之间存在着这样一种关系，以 $f|_{X\to Y}$ 为例：任意一个 $x\in X$，存在着唯一的一个 $y\in Y$，使得 $f(x)=y$，也就是通过 f 使得 x 找到了对应的 y。其中 y 称为 x 在 f 作用下的像，x 称为 y 的原像。

映射我们也可以用集合来定义。为此，我们定义了集合的卡氏积 $X\otimes Y=\{(x,y)\,|\forall x\in X,y\in Y\}$。也就是从两个集合 X,Y 中各自取一个元素构成一个有序元素对 (x,y)，然后所有的这样的有序对的集合就是 X,Y 的卡氏积 $X\otimes Y$。然后映射是这个卡氏积的满足下面的条件的子集 $F\subseteq X\otimes Y$：任意给定的 x 有且只有一个 y 组成元素对 (x,y)。这样下面所有的定义就都可以从集合来定义了。

[4]这个后半句不作要求，以后会进一步体会到。

17.3.1 群的定义

我们从一个集合开始来定义更加复杂的数序结构。如果只有一个集合 G，没有其他集合，没有映射，那么我们没什么其他的数学结构好定义的。我们来试试增加一个映射。你也可以试试增加一个不同于集合 G 的集合 L。我们先保持只有一个基础集合 G。如果我们引入一个映射从 G 到 G 的映射 $f|_{G\to G}$，那么，我们还是没有任何新的数学结构，因为相当于给 G 的元素做了重命名。自然地，当我们约定只有一个基础集合的时候，我们试试集合的卡氏积，也就是 G 上的一个从 $G\times G$ 到 G 的映射 "$\cdot|_{G\times G\to G}$"。下面我们把这个映射称作乘法，因为看起来其映射的形式为对于任意的 $g_1, g_2 \in G$，有 $g_1 \cdot g_2 \in G$，像一个乘法算式。

定义 17.1 (群). 如果乘法满足下面的性质，则这个集合 G 配上乘法映射一起称作一个群：

- 封闭性：$g_1 \cdot g_2 \in G$，$\forall g_1, g_2 \in G$,
- 结合律：$(g_1 \cdot g_2) \cdot g_3 = g_1 \cdot (g_2 \cdot g_3)$，$\forall g_1, g_2, g_3 \in G$,
- 存在单位元，记作 e：$g \cdot e = g = e \cdot g$，$\forall g \in G$,
- 存在逆元：$\forall g \in G$，$\exists g^{-1} \in G$ 使得 $g^{-1} \cdot g = e = g \cdot g^{-1}$。

可以验证整数配上加法当作群乘是一个群，但是自然数配上加法当作群乘不是一个群。

例 17.1 (整数配上加法构成群). 对照群的定义，验证当 G 取为整数集合 $\mathbb{Z}$，群乘取为加法的时候，是一个群。

- 封闭性：一个整数加上一个整数还是一个整数,
- 结合律：$(z_1 + z_2) + z_3 = z_1 + (z_2 + z_3)$，对于任意整数 z_1, z_2, z_3,
- 存在单位元，$e = 0$：$z + 0 = z = 0 + z$，对于任意整数 z,
- 存在逆元：$\forall z \in \mathbb{Z}$，$\exists g^{-1} = -z$ 使得 $-z + z = 0 = z + (-z)$。

但是，如果取 $G=\mathbb{N}$，则所有的大于零的整数在 G 中都没有逆元。因此，不满足定义要求，不是群。

可以验证整数配上乘法不是一个群（因为整数的乘法逆元如果存在应该是 $\frac{1}{z}$ 不是整数），分数配上乘法不构成群（因为 0 的乘法逆元如果存在应该是 $\frac{1}{0}$ 没有意义），分数扣除 0 以后配上乘法构成群，实数配上乘法不构成群（因为 0 的乘法逆元如果存在应该是 $\frac{1}{0}$ 没有意义），实数扣除 0 以后配上乘法构成群。我们选择验证其中的一个。

例 17.2 (分数扣除 0 以后配上乘法构成群). 对照群的定义，验证当 G 取为整数集合 $\mathbb{Z}$，群乘取为加法的时候，是一个群。

- 封闭性：一个分数乘上一个分数还是一个分数 $\frac{p_1}{q_1}\times\frac{p_2}{q_2}=\frac{p_1p_2}{q_1q_2}$，
- 结合律：$(x_1\times x_2)\times x_3=x_1\times(x_2\times x_3)$，对于任意分数 x_1,x_2,x_3，
- 存在单位元，$e=1$：$x\times 1=x=1\times x$，对于任意分数 x，
- 存在逆元：$\forall x=\frac{p}{q}\in\mathbb{Q}$，$x^{-1}=\frac{q}{p}$ 使得 $x\times x^{-1}=1$。

但是，如果取 $G=\mathbb{Q}$，则 $0\in G$ 但是它在 G 中没有逆元，$\frac{1}{0}\notin G$。因此，不满足定义要求，不是群。

以后我们会看到群是我们在集合之上构建的最基本的数学结构[5]，为什么这个最基本的数学结构会要求这四条呢? 甚至我们为什么会定义群呢? 在我们这里我们是因为定义矢量空间的需要。但是，实际上群的定义是有受到其他的启发的。

封闭性的要求很好理解。任何一个集合上的任何一个运算，我们都需要而且可以追求封闭性。可以这样去追求是因为，我们可以通过不断地把新出现的运算结果包含到集合里面去的方式来使得集合对运算封闭。需要追求封闭性是因为只要运算有现实世界中的意义或者数学中的意义，或者相应的其他学科中的意义，那么，运算出现的新结果就是有意义的，于是就应该被包含在我们所考虑的集合之中。就好像我们要在整数集合之中增加由于除法才出现的分数，而除法是作为乘法的逆运算而出现的。所谓逆运算就是当我们来求解一个乘法问题的逆问题，例如求解乘法的方程 $3x=4$，

[5]其实还有半群，不过我们在此忽略半群。

就自然会出现的。这也顺便解释了群的定义中对逆元存在的要求。至于单位元的存在，那是因为定义逆元所需要的。

那就剩下结合律这个要求了。有了结合律，我们就可以写下来类似于 $a\cdot b\cdot c$、a^m 这样的表达式，否则连这样的表达式都是有多个可能的含义的。此外，结合律也是对对象操作的普遍性质。例如，当我们对一个硬币做翻转（记作 X）和不翻转（记作 I）的操作的时候，IXX（从右侧开始，也就是将来把硬币的状态写到这个操作的右边，例如 $IXXu$）可以理解为“先翻转，再翻转，然后再不翻转”也就是逐次作用到硬币上 $I\left[X\left(Xu\right)\right]$，这也就是 $\left[I\left(XX\right)\right]u$。顺便，这里先把操作的部分计算出来，最后作用到硬币上，就比逐次作用到硬币上更方便。当然，必须保证两者的结果相同。IXX 也可以理解为先把 IX 的联合操作算出来，再去和剩下的操作结合，最后，再把这个算出来的操作作用到硬币上，也就是 $\left[\left(IX\right)X\right]u$。

我们可以具体来算一下这个操作的结合的结果。我们说，$XX=I$，$IX=X$，$XI=X$，$II=I$。为什么？硬币翻转两次，无论其处于什么状态都回到原状态，也就是相当于没有被翻转。因此，$\left(XX\right)u=Iu$ 对于任意的硬币状态 u，也就是说 $XX=I$。其他的你可以自己来验证一下。于是

$$I\left(XX\right)=II=I,\left(IX\right)X=XX=I. \tag{17.39}$$

确实两者得到结果相同。

可以验证，确实这个 I,X 配上这个翻转结合的操作合起来构成群。群就是对对象的操作的抽象，要求满足这一类的操作都得在集合 G 里面，要求操作之间得可以先做操作结合的运算，在作用到被操作对象上，而且这些结合要满足结合律。

习题 17.3 (Z_2 群)**.** 检验 I,X 配上这个翻转结合是否构成群。

好了，我们解释了这四个要求的每一个。当然，不是说不满足这四个要求的运算就不能存在。但是，我们发现，对于绝大多数运算，尤其是前面提到的根据受到的来自于现实世界的启发，这样的要求是很低的。

满足群乘交换律 $g_1\cdot g_2=g_2\cdot g_1$ 的群被称作阿贝尔群，有的时候也被称作加群。对于加法群，单位元也被称为零元。

这样看起来很基本的结构就有一些很有意思的性质。比如说，

定理 17.1: 群乘的一一性和到上性

对于任意的 $g_2 \neq g_3 \in G$，$g_1 g_2 \neq g_1 g_3$，以及对于任意的 $g \in G$，取任意的 $g_1 \in G$，存在着 $g_2 \in G$，使得 $g g_2 = g_1$。

证明 17.1. 我们先来证明第一条。假设有 $g_1 g_2 = g_1 g_3$，根据逆元存在性和结合律，我们有

$$g_1^{-1}(g_1 g_2) = g_1^{-1}(g_1 g_3) \Rightarrow \left(g_1^{-1} g_1\right) g_2 = \left(g_1^{-1} g_1\right) g_3 \Rightarrow g_2 = g_2. \quad (17.40)$$

对于第二条，我们取 $g_2 = g^{-1} g_1$ 就行。

因此群乘 $g\cdot$ 是一个一一到上映射，是一个可逆映射。实际上，逆映射就是 $g^{-1}\cdot$。按照这个定理，对于只有有限个元素的群，我们把群乘的结果列成一张表格（表头横向纵向的每一格为群的一个元素，可以约定纵向为 $g_1 \cdot g_2$ 的 g_1），则这张表格的每一行每一列都会每一个元素都出现一遍，而且任意两行或者两列的对应位置不会出现同一个元素。

根据这个结果，我们来定义子群和证明一个关于子群的定理。

定义 17.2 (子群). 如果群 G 有一个子集 $H \subseteq G$，在同一个群乘下也是一个群，则把 H 称为 G 的子群。

$\{e\}$ 和 G 肯定是群 G 的子群。同时，任意一个 $h \in H$，集合 hH 肯定仍然是 H。其中集合 hH 的定义是 $\{hh_1 | \forall h_1 \in H\}$。子群内部的群乘运算封闭。对于一个不在 H 中但是在 G 中的 g，gH 会怎样？顺便，这个集合 gH 有一个专门的名称——群 G 的子群 H 在元素 g 左乘下的陪集，简称陪集。陪集有左陪集和右陪集。

首先，$gH \cap H = \phi$。如果两者交集非空，那么，就存在一个 $h_1, h_2 \in H$ 使得 $g h_1 = h_2$，则 $g = h_2 h_1^{-1} \in H$。矛盾。接着，我们让所有的 $g \notin H, g \in G$ 都来乘一下得到一个集合 gH。我们发现这个陪集的大小都和 H 的大小相等，因为之前证明了群乘得到的结果不重复，也就是一一性。然后，我们证明，通过这样的操作肯定能够穷尽 G 的所有的元素，也就是对于任意的 $g \in G$，存在 $g_1 \in G$ 和 $h \in H$，使得 $g_1 h = g$。实际上，我们取 $h = e, g_1 = g$ 就可以。于是，我们得到了一系列的陪集——子群 H 自己可以看作是陪集 eH，合起来的元素的数量正好就是群 G 的元素的数量。因此，如果 G 是一

个有限大小的集合，则 H 的大小（记作 $|H|$）必然是 G 的大小（记作 $|G|$）的整数倍。

定义 17.3 (正规子群). 如果群 G 有一个子群 N，左陪集永远等于右陪集，也就是 $\forall g \in G$，$gN = Ng$，则称这个子群为正规子群。

对于正规子群，我们就可以不区分左陪集和右陪集了。我们来证明，正规子群的所有陪集的集合构成一个群。我们来这样定义陪集之间的乘法：

$$g_1 N \circ g_2 N = (g_1 g_2) N. \tag{17.41}$$

根据我们的定义，包含了所有的陪集，封闭性自然成立。首先，我们来证明 $eN = N$ 是单位元：$g_1 N \circ N = (g_1 e) N = g_1 N$。接着，我们来证明 $g^{-1}N$ 是 gN 的逆元：$g^{-1}N \circ gN = (g^{-1}g) N = N$。最后，我们来证明满足结合律：

$$g_1 N \circ (g_2 N \circ g_3 N) = g_1 N \circ (g_2 g_3) N = [g_1 (g_2 g_3)] N, \tag{17.42}$$

$$(g_1 N \circ g_2 N) \circ g_3 N = (g_1 g_2) N \circ g_3 N = [(g_1 g_2) g_3] N, \tag{17.43}$$

而根据群乘的结合律两者相等。

这个正规子群的所有的陪集构成的群也有一个专门的名称，称为群 G 在正规子群 N 下的商群，简称商群。有了商群，我们相当于把原来的群简化了，得到了一个正规子群 N，以及 $\frac{|G|}{|N|}$ 个和正规子群同样大小的集合，这些集合之间具有群乘。于是，只需要研究这些数量更少的集合之间的关系和正规子群，就可以得到整个群的性质。

你看一个满足看起来这么简单的封闭性和结合律的群，就可以有这么多非平凡的性质。未来，任何一个构成群的对象，都会具有这些性质。这就是代数的好处——从操作满足的条件的角度来定义对像和操作，而不在是从被操作的对象的属性来定义对象然后得到操作的性质。这样做确实更抽象了，但是适用范围已经比原来高了很多。

对于有限群（或者说元素可数的群，我们暂时只关心有限群，未来甚至可以推广到无限群），我们还可以这样来定义一个生成元：如果群 G 的一个子群 H 可以包含的元素是 $\left\{h, h^2, h^3, \cdots, h^L\right\}$（要求这些元素各不相同），则这个子群 H 称为元素 h 生成的群，而 h 称为 H 的生成元。有时候也把 h 称为群 G 的生成元。

这个由 h 生成的子群 H 的结构特别简单。我们来证明 $h^L = e$。如果之前就有元素 $h^M = e, M < L$，则 $h^{M+1} = h^m h = h$ 就和第一个元素 h 重复了。其中又必须有单位元，因此，只能是 $h^L = e$。也就是说，这个群 H 就是一个由 h 的各阶乘法构成的循环群。任意两个元素的相乘计算也特别简单，例如 $h^m h^n = h^{m+n} = h^{(m+n) \pmod L}$。其中 $k \pmod L$ 表示整数 k 除以整数 L 得到的余数。

如果一个群能够被表示为由若干个生成元生成的，那这个群上的计算就简单了。当然，如果有多个生成元，还得搞清楚这些生成元之间的乘法得到的结果，类似于 $h_1 h_2$ 这样的两个生成元的乘法项，但是，至少在每一个生成元自己构成的循环子群内就简单了。实际上，未来我们就是这样来研究李群的。

17.3.2 群的定义受到的启发

在 17.11.4 节我们会更加详细地展开论述物理学个一个很大的主题是研究状态的描述和状态的演化。而对状态的演化的研究可以被转化到对状态的操作的算符的研究。也就是本来我们要看每时每刻的状态，现在我们固定一个初始时刻的状态把每时每刻的状态看作是一个那个时刻的算符作用到物理对象上从而改变其状态的结果。这样，物理学就整个从对状态的研究变成了对对状态的操作的研究。对状态的所有操作合起来必然构成群。

因此，群的概念是受到物理学研究的启发的，物理学也是群的典型应用领域之一。为什么说，对状态的所有的操作合起来必然构成群呢？首先，这些操作必然可以联合起来，也就是一个操作先作用到状态上，然后另一个操作再作用上去。这个联合操作就定义为操作之间的乘法，大概可以表示为这样，

$$(BA)\, u \triangleq B\,(Au)\,. \tag{17.44}$$

首先，这样的联合操作具有封闭性，因为 BA 合起来必然仍然代表一个对这个状态的一个操作。

其次，这个操作显然具有不变操作，也就是完全不改变状态的操作，

$$Iu = u. \tag{17.45}$$

这样一旦这个操作和其他操作联合起来，自然就有

$$IA = A = AI, \tag{17.46}$$

因为

$$(IA)\,u = I\,(Au) = Au, (AI)\,u = A\,(Iu) = Au. \tag{17.47}$$

逆元也很好理解，尤其对于有限个状态的对象，只要把某一个操作不断地进行下去，由于状态的总数有限，其必然会回到这个状态，这样就得到了 $A^m = I$，也就是 A 的逆元是 A^{m-1}。

至于结合律未来有了演化算符之后会更好理解，我们放到 17.11.4 节去展示。比如说，我们旋转一个刚体，旋转一个自旋，都是前面的状态和作用在状态上的算符的例子，我们推动一个质点也是。

但是，其实，更一般地，我们对一个硬币的翻转或者不翻转的操作，我们对于一个三角形的旋转一定角度的操作，也可以当作物理学的研究对象，甚至是博弈论的研究对象。比如说，我们约定一开始硬币的状态固定为正面向上。然后，两个博弈的参与者可以选择各自的策略，翻转或者不翻转，然后需要把两个人选择的策略按照先第一个人作用再第二个人作用的方式作用到这个硬币上面，最后根据硬币的状态来判断谁输输赢——例如，如果最后是正面向上则第一个人从第二个人那里赢一块钱，否则第二个人从第一个人那里赢走三块钱。

这个时候，博弈者的行动选择的集合就是一个群 I, X。其中 I 表示不翻转，X 表示翻转。很显然，$XX = I$，$IX = X = XI$，$II = I$。这样我们就给出了所有的联合操作的结果。如果要继续操作下去，比如说 XXX，就可以按照顺序得到 $XXX = X\,(XX) = XI = X$。所以，群论也是自然地从对任何一个对象的操作上受启发的。

类似地我们可以来讨论对一个三角形的操作。为了简单，我们先来讨论对一个正三角形的保证操作前后还都是正三角形的所有的操作。我们记原始的三角形为 $\triangle ABC$，其中 A 在三角形的做下方，B 在右下方，C 在 AB 的上方，ABC 按照逆时针的方向编号，如图 17.2 的左下方的三角形。

然后，下一排的三个三角形（其中一个是原始状态），从左到右，依次是逆时针转动 0°（不动），120° 和 240°。而上下两排之间可以看作是上一排把下一排对着的三角形的底边两个顶点交换了一下，或者说，沿着底边

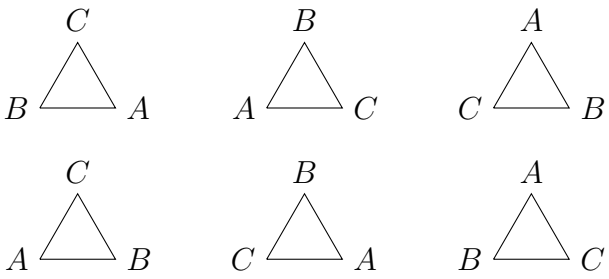


图 17.2: **对正三角形的所有翻和转的操作。对一个正三角形的所有的转动和翻转的操作构成群。**

上的中线做了一次翻转。或者，都从原始状态开始看，则上面的三个可以分别看作是对原始状态做沿 AB 上高的翻转、沿 BC 上高的翻转和沿 AC 上高的翻转。在保持这个三角形还是正三角形，又仅仅考虑其仍然停留在二维平面内的操作就只有这六个了。

于是，显然，这六个操作有乘法（联合操作），乘法满足封闭性、有单位元、有逆元。

其实还有一个有意思的问题，我们为什么不包含把硬币融化了、粉碎了的操作，把三角形拉伸了的操作呢？其实，只要我们允许这样的操作，那么，相当于在另一个层次上保持了构成这些对象的元素的不变，例如，拉伸不断开是一种不变性，硬币融化了但是分子层面没有变也是一种不变性。正是这种另一个层次上还会存在的不变性，使得这些对对象的操作仍然满足封闭性，会有乘法的定义，会有单位元和逆元。因此，群其实就是在保持被操作对象某种意义上不变的条件下的所有可能的操作的集合。

有了这个启发，结合未来 17.11.4 节中来自物理学的启发，群的定义就很好理解了。

顺便，未来把群用于博弈论也就是一件很自然的事情了。这就是看到一个概念受到的启发和这个概念的内涵之间的联系的好处。

17.3.3 域的定义

定义 17.4 (域). 一个集合 F 上定义了加法（记作 $a+b$）和乘法（记作 $a\times b$）两种从 $F\otimes F$ 到 F 的运算，如果它满足

- F 在加法运算下构成一个交换群，记加法零元为 0，
- F 在扣除 0 之后在乘法运算下构成一个交换群，记乘法单位元为 1，
- 乘法对加法的左侧分配律：$a \times (b+c) = a \times b + a \times c$，

则称 F 在这个加法和乘法下是一个域。

有的时候也简称 F 是个域。

可以验证有理数、实数、复数配上数的加法和乘法是一个域。整数不是一个域，因为整数不满足乘法下构成群（整数的乘法逆元不是整数）。可以证明 $a+b\sqrt{2}$ 其中 $a,b \in \mathbb{Q}$ 是有理数是域。

例 17.3 ($a+b\sqrt{2}$ 是个域)**.** 证明 $a+b\sqrt{2}$ 其中 $a,b \in \mathbb{Q}$ 是有理数是域。

我们先来证明第一条，$a+b\sqrt{2}$ 在数的加法下构成交换群。交换律和结合律必然满足我们不再检验。我们来看看封闭性、零元和逆元是否成立。

$$a_1 + b_1\sqrt{2} + a_2 + b_2\sqrt{2} = (a_1+a_2) + (b_1+b_2)\sqrt{2}. \tag{17.48}$$

因此封闭性成立。零元是 0 对应着 $a=0=b$ 存在。$a+b\sqrt{2}$ 的逆元是 $-a-b\sqrt{2}$，因此零元和逆元都存在。

我们先来证明第二条，$a+b\sqrt{2}$ 在数的乘法下构成交换群。交换律和结合律必然满足我们不再检验。我们来看看封闭性、零元和逆元是否成立。

$$\left(a_1 + b_1\sqrt{2}\right) \times \left(a_2 + b_2\sqrt{2}\right) = (a_1a_2 + 2b_1b_2) + (a_1b_2 + a_2b_1)\sqrt{2}. \tag{17.49}$$

因此封闭性成立。零元是 1 对应着 $a=1, b=0$ 存在。$a+b\sqrt{2}$ 的逆元是 $\frac{1}{a+b\sqrt{2}} = \frac{a-b\sqrt{2}}{a^2-2b^2} = \frac{a}{a^2-2b^2} - \frac{b}{a^2-2b^2}\sqrt{2}$，因此零元和逆元都存在。

第三条乘法对加法的左侧分配律必然成立，因为这是实数的一部分，所有的实数都满足乘法对加法的左侧分配律。

因此，$a+b\sqrt{2}$ 是个域。

类似地你可以证明 $a+b2^{\frac{1}{3}}+c2^{\frac{2}{3}}$ 是一个域，$a+b2^{\frac{1}{3}}$ 不是。

习题 17.4 ($a+b2^{\frac{1}{3}}+c2^{\frac{2}{3}}$ 是个域)**.** 证明 $a+b2^{\frac{1}{3}}+c2^{\frac{2}{3}}$ 是一个域，$a+b2^{\frac{1}{3}}$ 不是一个域。

从定义矢量空间的角度，之所以要先定义域是为了给矢量空间的数乘准备好“数”。这个数子集得能做加法和乘法，因为未来会出现两个矢量相加

和矢量数乘，例如 $ax, ax+bx, a(bx)$，这时候很自然就需要有 $(a+b)x, (ab)x$。也就是说，矢量加法和数乘最好和其数乘用到的数的加法和乘法是配套的相容的。为什么还要要求分配律呢？类似地，将来矢量运算上会出现 $abx+acx, a(b+c)x$，也就是把一个矢量放大 ab 倍和 ac 倍然后加起来，这自然就很像先放大 $(b+c)$ 倍然后再放大 a 倍。

但是实际上，域的定义也是有其自身受到的启发的。下面要介绍的伽罗瓦（Galois）理论就是其中之一。

回顾群和域的定义，实际上它们是把小学的时候学过的四则运算律（加法结合律，加法交换律，乘法结合律，乘法交换律，乘法对加法的左侧分配律）做了一个层次上的区分，增加了封闭性的明确要求（小学的时候其实也有，从整数到分数就是为了除法的封闭性，但是那个时候不强调封闭性这个词），然后强调从这些运算或者说操作要满足的条件来定义对象，而不是像小学的时候一样从对象和运算的含义上来确认这些对象有这些规律。这两件事情都是数学走向更加一般化更加抽象的表现。这样做的好处是，这些新的数学结构不再受任何具体对象的制约了。只要对这些对象的操作和运算具有这些性质，那么，未来关于这些数学结构的研究的结果就可以自动套用到这样的对象上。

我们先来说，为什么把四则运算律和封闭性分层次，怎么分的层次。首先，封闭性是集合和计算的最低要求，并且逆元和单位元实际上也是封闭性带来的。因此，放到要求低的群里面。其次，前面从硬币翻转的例子中我们也看到了结合律对于操作来说特别重要，因此也放到要求低的群里面。至于更高的要求的分配律，专门定义一个群的子类——阿贝尔群。

然后，域实际上不过就是定义了两种群运算的同一个集合。因为同时有两种群乘运算——为了区分分别称作加法和乘法，自然就要去协调这两种运算之间的关系，于是才有了 $a(b+c), (b+c)a, ab+c, c+ab$ 这些运算。其中，自然就会要求这些运算——回到群的语言这相当于对背后的对象的状态的操作——满足可以先结合起来的要求，并且最好不同的结合方式得到的结果相同。比如说，最好有 $a(b+c)=ab+ac$ 和 $ab+a=ab+ae=a(b+e)$ 等等。

比如说，我们来把一个实数和另一个实数的加法运算 $a+b$ 解释为把一个在初始位置的物体，沿着坐标轴所在的方向先推动距离 b，再推动距离 a，

也就是 $P_a + P_b$。我们发现，这个操作确实有如下的关系，

$$P_a + P_b = P_{a+b}. \tag{17.50}$$

从推动操作来说，这其实就是 $a+b$ 的真正含义。同时，我们再定义 $P_{nb} = \underbrace{P_b + P_b + \cdots + P_b}_{n} = nP_b$，这其实就是 b 的整数乘法的含义。接着我们可以推广到任何实数的乘法 ab，然后对应到 P_{ab}，就得到了 ab 对应的操作的定义。

这个时候，我们就自然要处理 $ab + a$ 这一类的问题了。并且实际上前面的 P_{nb} 的定义中，就藏了一个这样的问题：$P_{2b} = P_b + P_b$，回到没有 P 的语言，这就是 $b + b = 2b$，其实就相当于 $b + b = b(1 + 1) = b2 = 2b$。也就是说，在这里分配律和交换律的成立特别的自然，是相应的背后的推动操作的自然性质。

也就是说，当只有一个运算的时候，也就是定义群的时候，我们其实不需要考虑分配律。只有当考虑多个运算的时候，也就是需要定义域的时候，才需要考虑分配律。

17.3.4 伽罗瓦理论 *

17.3.5 矢量空间的定义

有了域，我们来定义域上的矢量空间。

定义 17.5 (矢量空间). 如果一个集合 V 上定义上加法运算 $V \otimes V \to V$，记作 $a + b, \forall a, b \in V$，和数乘 $V \otimes F \to V$ 运算，记作 $\mu a, \forall a, b \in V, \mu, \nu \in F$，满足下面的性质：

- F 是一个域；
- V 在加法下构成一个交换群；
- 域上运算和矢量运算的相容性：
  - $1a = a$，其中 1 是 F 中的乘法单位元，
  - 双向分配律：$(\mu + \nu)(a + b) = \mu a + \mu b + \nu a + \nu b$，

– 域结合律：$\mu\left(\nu a\right)=\left(\mu\nu\right)a$.

注意，这样定义的 V 到底每个元素 a 是什么含义，这里我们完全是没有给出的。其实，甚至域 F 的每个元素 μ 是什么意思也没有给出。只不过，从小学学习到的数学知识，我们已经有了一些熟悉的域：有理数域、实数域、复数域等等。

另外，我们前面得到的启发中的线性性体现在哪里？矢量的长度、方向、内积、缩放、旋转体现在哪里？线性性其实通过定义矢量之间的加法和数乘已经包含进去了，因为有了加法和数乘我们就可以计算 $\mu a+\nu b$ 了。未来如果这个矢量其实是有一个操作对象的——把那些被操作的对象记作 x,y，就会出现类似于

$$v\left(\mu x+\nu y\right)=\mu v\left(x\right)+\nu v\left(y\right) \tag{17.51}$$

这样的结果。数乘也就是缩放，也有了。但是，矢量的长度、方向、内积以及矢量的分量形式，还有通过分量形式表达的旋转还没有包含进去。稍后，我们再来讨论这些更加复杂的数学结构。

我们先来证明几个有意思的结果。

例 17.4 $(0\times a=0)$. 我们来证明 $0\times a=\underline{0}$，其中 0 是 F 的加法零元，$\underline{0}$ 是 V 中的加法零元。

证明 17.2. 我们来计算下式：

$$a+0\times a=1\times a+0\times a=\left(1+0\right)\times a=1\times a=a, \tag{17.52}$$

两边加上 $-a$，得到

$$0\times a=a+\left(-a\right)=\underline{0}. \tag{17.53}$$

顺便，从现在开始我们就不再区分 0 和 $\underline{0}$，都记作 0。同理可以得到 $(-1)a=-a$。

例 17.5 $((-1)a=-a)$. 我们来证明 $(-1)a=-a$，其中 1 是 F 的乘法单位，a 是 V 中元素。

证明 17.3. 我们来计算下式：

$$a+(-1)a=1\times a+(-1)\times a=(1+(-1))\times a=0\times a=0, \tag{17.54}$$

两边加上 $-a$，得到

$$\begin{aligned} -a+a+(-1)a &= 0+(-a) \\ \Rightarrow (-a+a)+(-1)a &= 0+(-a)=-a \\ \Rightarrow 0+(-1)a &= -a \\ \Rightarrow (-1)a &= -a. \end{aligned} \tag{17.55}$$

我们来看看我们熟悉的矢量按照这个定义还是不是矢量。

例 17.6 (二维平面上的矢量). 我们来检验一下二维平面上的矢量的集合 $\mathbb{R}^2$ 是不是满足矢量空间的定义。

首先，对于二维平面矢量，其用来数乘的域是实数 $\mathbb{R}$，而实数确实是一个域。

其次，用二维平面上的矢量的分量表示 $\vec{a}=\begin{bmatrix} a_x \\ a_y \end{bmatrix}$ 来表示就可以证明 $\mathbb{R}^2$ 确实在加法下构成一个交换群。

接着，相容性也非常容易得到验证。

再来看看我们不熟悉的。

例 17.7 (所有的可导函数). 我们来检验一下在一个区间，例如 $(-\infty,\infty)$ 上的所有的可导函数（这个条件可以换成别的，例如连续函数、可积函数、平方可积函数、无穷次可导函数或者称为光滑函数），是不是矢量空间。

首先，对于函数，其数乘可以定义为 $f\to\mu f$，也就是把函数值做一个数乘 $f(x)\to\mu f(x)$，其用来数乘 μ 的域是实数 $\mathbb{R}$，而实数确实是一个域。

其次，函数之间的加法 $f+g$ 也由自然的定义：$(f+g)(x)=f(x)+g(x)$。例如，加法封闭性，加法结合律，加法交换律，加法零元和加法逆元都容易得到。封闭性是因为两个可导函数相加仍然得到可导函数。零元就是函数值取值处处为零的函数。$f(x)$ 的逆元就是 $-f(x)$。

最后，域运算和矢量运算的相容性也很容易得到。例如，我们来证明分配律：$(\mu+\nu)(f+g)=\mu f+\mu g+\nu f+\nu g$，对于任意的 x，有

$$\begin{aligned} (\mu+\nu)(f+g)(x) &= (\mu+\nu)[f(x)+g(x)] \\ &= \mu f(x)+\mu g(x)+\nu f(x)+\nu g(x). \end{aligned} \tag{17.56}$$

例 17.8 (导数算符). 我们来检验一下二维平面上的导数的集合 $\left\{v_x\frac{\partial}{\partial x}+v_y\frac{\partial}{\partial y}\right\}$ 是不是满足矢量空间的定义。其中 v_x, v_y 是任意一个 $\mathbb{R}^2$ 中的矢量。对于这样的导数算符，其有一个天然的作用对象：两变量可导函数 $f(x,y)$。

首先，对于二维平面矢量，其用来数乘的域是实数 $\mathbb{R}$，而实数确实是一个域。

其次，用二维平面上的矢量的分量表示 $v_x\frac{\partial}{\partial x}+v_y\frac{\partial}{\partial y}$ 来表示就可以证明这个矢量确实在加法下构成一个交换群。例如，加法封闭性，加法结合律，加法交换律，加法零元和加法逆元都容易得到。我们先来搞清楚这个导数算符的加法应该怎么定义，怎么表示。

对于一个任意的两变量可导函数 $f(x,y)$，

$$\begin{aligned}&\left[\left(a_x\frac{\partial}{\partial x}+a_y\frac{\partial}{\partial y}\right)+\left(b_x\frac{\partial}{\partial x}+b_y\frac{\partial}{\partial y}\right)\right]f(x,y)\\&=\left(a_x\frac{\partial f}{\partial x}+a_y\frac{\partial f}{\partial y}\right)+\left(b_x\frac{\partial f}{\partial x}+b_y\frac{\partial f}{\partial y}\right)\\&=(a_x+b_x)\frac{\partial f}{\partial x}+(a_y+b_y)\frac{\partial f}{\partial y}\\&=\left[(a_x+b_x)\frac{\partial}{\partial x}+(a_y+b_y)\frac{\partial}{\partial y}\right]f(x,y),\end{aligned}\tag{17.57}$$

因此我们得到，

$$\left(a_x\frac{\partial}{\partial x}+a_y\frac{\partial}{\partial y}\right)+\left(b_x\frac{\partial}{\partial x}+b_y\frac{\partial}{\partial y}\right)=(a_x+b_x)\frac{\partial}{\partial x}+(a_y+b_y)\frac{\partial}{\partial y}\tag{17.58}$$

实际上，这个加法和 $\mathbb{R}^2$ 中的矢量加法完全同构：不过就是把 $\hat{e}_1,\hat{e}_2$ 换成了 $\frac{\partial}{\partial x},\frac{\partial}{\partial y}$。可以证明，数乘也完全同构。于是，只要 $\mathbb{R}^2$ 中的矢量加法满足的性质，这个导数的加法都满足。

对于这样的导数算符，我们还可以这样看：它们是作用在函数这个矢量空间上的线型算符，也就是

$$\frac{\partial}{\partial x_i}\left(\mu f(x,y)+\nu g(x,y)\right)=\mu\frac{\partial f}{\partial x_i}+\nu\frac{\partial g}{\partial x_i}.\tag{17.59}$$

这里的 $x_1=x, x_2=y$。稍后我们就来证明矢量空间上的线性算符的集合肯定构成矢量空间。

17.3.6 矢量空间上的内积

在那之前，我们来补充矢量的大小和方向的定义，也就是矢量的内积的定义。我们先抽象地定义内积，再来讨论内积的分量表述形式。下面，我们回到实数域或者复数域上的有限维矢量空间 $\mathbb{R}^n$ 和 $\mathbb{C}^n$，并且采用带有基矢的语言来讨论矢量空间。下面，我们将以 $\mathbb{C}^n$ 为例，因为其中的定义和定理都基本适用于 $\mathbb{R}^n$。

定义 17.6 (内积). 一个矢量空间上的内积的是指从 $V\otimes V$ 到 F 的映射，通常记为 $(\cdot,\cdot)$，其中两个 "$\cdot$" 就是留给两个矢量的位置，满足以下要求：

1. 线性性：$(a,\mu b+\nu c)=\mu\,(a,b)+\nu\,(a,c)\,,\forall a,b,c\in V,\forall\mu,\nu\in F$，
2. 长度大于等于零：$(a,a)\geqslant 0,\forall a\in V$，取等号当且仅当 $a=O$，
3. 复共轭对称性：$(a,b)=(b,a)^*\,,\forall a,b\in V$。

稍后我们就会发现，用分量形式表示的 $\mathbb{C}^n$ 中的内积满足这三条要求。当我们把线性性与复共轭结合之后，我们看到，内积对于左边的矢量的线性性与右边的矢量的线性性不太一样：

$$(\mu a+\nu b,c)=(c,\mu a+\nu b)^*=\mu^*\,(a,c)+\nu^*\,(b,c)\,. \tag{17.60}$$

我们把这个性质称作反线性性，其实更合适的名词应该是共轭线性性。当然，如果数域 F 中，$a^*=a,\forall a\in F$，那么实际上内积对左边和右边的矢量的线性性就是相同的。当 F 中一个元素的复共轭不再回到这个元素，这就不一样了，例如当 $F=\mathbb{C}$ 的时候。

我们来验证，通常的 $\mathbb{C}^n$ 中通过分量表示的内积满足这个不依赖分量的内积的定义。用分量形式定义的 $\mathbb{C}^n$ 中的内积是这样的，

$$(a,b)=\sum_{l=1}^{n}a_l^*b_l. \tag{17.61}$$

首先，这个定义满足右侧线性性，

$$\begin{aligned}(a,\mu b+\nu c)&=\sum_{l=1}^{n}a_l^*\,(\mu b_l+\nu c_l)\\&=\mu\sum_{l=1}^{n}a_l^*b_l+\nu\sum_{l=1}^{n}a_l^*c_l\\&=\mu\,(a,b)+\nu\,(a,c)\,.\end{aligned} \tag{17.62}$$

其次，长度大于等于零，

$$(a,a)=\sum_{l=1}^{n}a_l^*a_l\geqslant 0, \tag{17.63}$$

并且如果等于零，则所有的 $a_l=0$，也就是 $a=0$，反之亦然。最后，复共轭对称性也自然满足，

$$\begin{aligned}(b,a)^*&=\left(\sum_{l=1}^{n}b_l^*a_l\right)^*\\&=\sum_{l=1}^{n}a_l^*b_l\\&=(a,b)\,.\end{aligned} \tag{17.64}$$

内积的抽象定义的好处是第一可以不依赖于分量形式来定义，第二可以拓展到其他的内积的定义而不仅仅是 $\mathbb{C}^n$ 中的常用的分量形式。

定义 17.7 (内积空间)**.** 定义了内积的矢量空间被称为内积空间。

有了内积这个映射之后，我们来看 $(\cdot,a)$ 和 $(a,\cdot)$ 是什么。我们发现，这两个东西都可以看成是一个从 V 到 F 的映射，它需要一个 V 的元素 b 的输入，产生一个 $(b,a)\in F$ 或者 $(a,b)\in F$ 的输出。而且，这个映射与 V 的元素是一一对应的：有一个 $a\in V$ 就有一个 $(\cdot,a)$ 和一个 $(a,\cdot)$。

记它们的集合分别为 $V^*=\{(a,\cdot)\,|\forall a\in V\}$ 和 $V^{**}=\{(\cdot,a)\,|\forall a\in V\}$。我们发现 V 与 V^* 以及 V^{**} 之间存在着一一到上的可逆映射。

不过这两个映射也不太一样。$(\cdot,a)$ 满足和 a 一样的加法和数乘，也就是

$$(\cdot,\mu a+\nu b)=\mu\,(\cdot,a)+\nu\,(\cdot,b)\,. \tag{17.65}$$

这可以通过内积的定义看出来：对于任何一个 $c\in V$，

$$(c,\mu a+\nu b)=\mu\,(c,a)+\nu\,(c,b)\,. \tag{17.66}$$

也就是说 $(\cdot,a)$ 和 a 在加法和数乘上一模一样。由此也很容易进一步验证 V^{**} 中的元素满足矢量空间的定义。也就是说 V^{**} 就是 V。但是 $(a,\cdot)$ 和 a 一样的加法和数乘具有对应关系但是稍有不同——是共轭线性性而不是普通的线性性，

$$(\mu a+\nu b,\cdot)=\mu^*\,(a,\cdot)+\nu^*\,(b,\cdot)\,. \tag{17.67}$$

但是，我们也可以定义一个 V^* 中的 $(a,\cdot)$ 的加法和数乘分别为：

$$(a,\cdot)+(b,\cdot)=(a+b,\cdot)\,, \tag{17.68}$$

$$\mu\,(a,\cdot)=(\mu^* a,\cdot)\,. \tag{17.69}$$

可以验证，这个数乘和加法满足矢量空间的定义。

例 17.9 (V^* 是矢量空间)**.** 证明 V^* 满足矢量空间的定义。

首先，参与其数乘的数的集合 $F=\mathbb{C}$ 是一个域。

其次，V^* 在加法下构成交换群。封闭性和结合律根据 a 的封闭性和结合律自然满足。零元是 $(0,\cdot)$，$(a,\cdot)$ 的逆元是 $(-a,\cdot)$。

接着，域上运算和矢量运算的相容性：第一条 $1\,(a,\cdot)=(1^* a,\cdot)=(a,\cdot)$。第二条双向分配律，

$$\begin{aligned}(\mu+\nu)\left[(a,\cdot)+(b,\cdot)\right]&=(\mu+\nu)\left[(a+b,\cdot)\right]\\&=\left((\mu+\nu)^*\,(a+b)\,,\cdot\right)\\&=(\mu^* a+\mu^* b+\nu^* a+\nu^* b,\cdot)\\&=\mu\,(a,\cdot)+\mu\,(b,\cdot)+\nu\,(a,\cdot)+\nu\,(b,\cdot)\end{aligned} \tag{17.70}$$

第三条域结合律，

$$\mu\left[\nu\,(a,\cdot)\right]=\mu\,(\nu^* a,\cdot)=(\mu^*\nu^* a,\cdot)=(\mu\nu)\,(a,\cdot)\,. \tag{17.71}$$

都满足。

这样我们就证明了 V^* 其实是一个矢量空间。顺便，我们也看到了共轭线性性是一种线性性。我们还看到了一般性的矢量空间定义的好处——你看完全看不出来分量形式该怎么写而且看起来不像满足线性性的 V^* 实际上是一个矢量空间。

至于 V^{**} 是一个矢量空间就不再证明了，因为 V^{**} 中的元素和 V 中一一对应，加法和数乘也相同。借助内积，我们来把 $(\cdot,a)$ 看作是 V^* 上的线性映射，也就是任意一个 $(b,\cdot)\in V^*$，$(\cdot,a)$ 作用上去之后得到，

$$(\cdot,a)\left[(b,\cdot)\right]=(b,a)\,. \tag{17.72}$$

我们来证明这个映射是线性的，也就是

$$(\cdot,a)\left[\mu(b,\cdot)+\nu(c,\cdot)\right]=\mu(\cdot,a)\left[(b,\cdot)\right]+\nu(\cdot,a)\left[(c,\cdot)\right]. \tag{17.73}$$

我们从左边开始，

$$\begin{aligned}(\cdot,a)\left[\mu(b,\cdot)+\nu(c,\cdot)\right]&=(\cdot,a)\left[(\mu^{*}b+\nu^{*}c,\cdot)\right]\\&=(\mu^{*}b+\nu^{*}c,a)\\&=\mu(b,a)+\nu(c,a).\end{aligned} \tag{17.74}$$

然后计算右边，

$$\mu(\cdot,a)\left[(b,\cdot)\right]+\nu(\cdot,a)\left[(c,\cdot)\right]=\mu(b,a)+\nu(c,a). \tag{17.75}$$

左边右边得到的结果完全相同。因此，映射的线性性成立。

注意 V^{*} 中的元素 $(a,\cdot)$ 是从矢量空间 V 到域 F 的线性映射，并且自己构成矢量空间，而 V^{**} 中的元素 $(\cdot,a)$ 从矢量空间 V^{*} 到域 F 的线性映射，并且自己也构成矢量空间，而且这个矢量空间的元素和 V 中的元素完全一一对应。因此，我们把 V^{*} 称作对偶矢量空间，对偶的含义是把原来的矢量空间映射为数，矢量空间的含义是其自身是矢量空间。同理，V^{**} 是 V^{*} 的对偶矢量空间。对于有了内积定义的矢量空间，我们看到，对偶矢量空间的对偶矢量空间是原矢量空间。

对于有限维矢量空间 $V=\mathbb{C}^{n}$，由于存在着 V,V^{*},V^{**} 之间的一一对应关系，V^{*} 也必然是 n 维矢量空间。

为了明确地把这样的对应关系表示出来，联系到 $(a,\cdot)$ 和 $(\cdot,a)$ 的记号，Dirac把这些记号改写了一下，成为

$$(a,\cdot)\rightarrow\langle a|\,, \tag{17.76}$$

$$(\cdot,a)\rightarrow|a\rangle. \tag{17.77}$$

这个记号被称为Dirac 符号，分别称作左矢量和右矢量。用这个记号，内积就表示为，

$$\langle a|\,|b\rangle=\langle a|\,b\rangle=(a,b)\,, \tag{17.78}$$

也就是说内积就是左右矢量凑起来，尖角向外，凑成原来内积的括号的样子。其中，凑起来的时候我们为了简化，扔掉了中间重复的竖线。

我们称两个内积为零的矢量具有正交关系。这是来自于平面直角坐标系中的 x 和 y 坐标方向的矢量的正交关系，也就是 $(\hat{e}_1, \hat{e}_2) = 0$，在一般的矢量空间的推广。

有了内积，在 $\mathbb{R}^n$ 中还可以定义矢量之间的夹角，如公式 (17.28)。但是，在 $\mathbb{C}^n$ 由于矢量的内积可以是一个复数，我们不能定义夹角。只能定义矢量的长度，

$$|a| = \sqrt{(a,a)} = \sqrt{\langle a\,|a\rangle}. \tag{17.79}$$

有了内积，我们就可以来定义正交归一的基矢了。严格地说，我们得证明，任何一个 n 维线性空间内可以找到一组基矢，然后对这些基矢量做正交归一化，得到一组正交归一的基矢。不过，我们借助熟悉的 $\mathbb{R}^2$ 来避免完全从头来构建这些基矢。

17.3.7 矢量空间的基矢 *

对于任意两个不完全成线性关系（$b \neq \gamma a$）的矢量 a, b，我们总是可以通过下面的过程来得到两个正交归一的矢量 a', b'，

$$|a'\rangle = \frac{|a\rangle}{\sqrt{\langle a\,|a\rangle}}, \tag{17.80}$$

$$|b'\rangle = \frac{|b\rangle - |a'\rangle\langle a'\,|b\rangle}{\sqrt{\left(\langle b| - \langle a'|\langle b|\,a'\rangle\right)\left(|b\rangle - |a'\rangle\langle a'\,|b\rangle\right)}}. \tag{17.81}$$

习题 17.5 (a', b' 正交归一). 证明上面定义的 a', b' 正交归一，当 $b \neq \gamma a$ 的时候。

如果有第三个矢量 c，则我们进一步先从 c 中去掉 a' 和 b'，再来归一化，就可以，

$$|\tilde{c}\rangle = |c\rangle - |a'\rangle\langle a'\,|c\rangle - |b'\rangle\langle b'\,|c\rangle, \tag{17.82}$$

$$|c'\rangle = \frac{|\tilde{c}\rangle}{\sqrt{\langle\tilde{c}\,|\tilde{c}\rangle}}. \tag{17.83}$$

继续这个操作，我们就可以从一组不完全成线性关系的矢量中得到一组正交归一的矢量。这个正交归一化的过程被称为施密特 (Schmidt) 正交化。同

时，两个矢量之间的不完全线性相关也需要被推广到一组矢量之间的不完全线性相关。

定义 17.8 (线性相关和线性无关). 如果存在着 F 中的不完全为零的数 μ_l $(\sum_{l=1}^{m}|\mu_l|\neq 0)$ 使得

$$\sum_{l=1}^{m}\mu_l|a_l\rangle=0, \tag{17.84}$$

我们就说这 m 个自身不为零的矢量 $\{|a_l\rangle\}$ $(\langle a_l|a_l\rangle\neq 0)$ 之间线性相关。否则，称这 m 个矢量线性无关。

习题 17.6 (从原命题得到否命题). 把上面定义中的“否则”明确写下来。(提示：这里牵涉到从“存在”到“任意”的转换。)

在下面的例子中，我们解答了这个习题。请你先自己尝试来解答这个习题。

例 17.10 (从原命题得到否命题). 把上面定义中的“否则”明确写下来。

如果对于任意的 F 中的不完全为零的数 μ_l $(\sum_{l=1}^{m}|\mu_l|\neq 0)$ 都有

$$\sum_{l=1}^{m}\mu_l|a_l\rangle\neq 0, \tag{17.85}$$

我们就说这 m 个自身不为零的矢量 $\{|a_l\rangle\}$ $(\langle a_l|a_l\rangle\neq 0)$ 之间线性无关。

注意，由于每一个 a_l 都不为零，因此如果 μ_{l_0} 只有其中一个不为零的时候，肯定满足

$$\sum_{l=1}^{m}\mu_l|a_l\rangle=\mu_{l_0}|a_{l_0}\rangle\neq 0 \tag{17.86}$$

因此，线性相关必定是至少出现两个不为零的 μ_l 使得 $\sum_{l=1}^{m}\mu_l|a_l\rangle=0$。比如说，当仅有两项的时候，就意味着

$$\begin{aligned}&\mu_{l_1}|a_{l_1}\rangle+\mu_{l_2}|a_{l_2}\rangle=0\\ \Rightarrow\ &|a_{l_1}\rangle=-\frac{\mu_{l_2}}{\mu_{l_1}}|a_{l_2}\rangle.\end{aligned} \tag{17.87}$$

这正好就是两个矢量的方向完全相同的含义。类似地，如果有三项系数不为零，但是合起来满足线性相关性，就表示，

$$\Rightarrow|a_{l_1}\rangle=-\frac{\mu_{l_2}}{\mu_{l_1}}|a_{l_2}\rangle-\frac{\mu_{l_3}}{\mu_{l_1}}|a_{l_3}\rangle. \tag{17.88}$$

也就是其中一个矢量可以表示为其他矢量的线性叠加。

如果我们找到了这样的把一个矢量表达成其他矢量的线性叠加的形式，然后对用来做线性叠加的右侧的矢量做一个施密特（Schmidt）正交化，我们不就可以得到这个矢量空间的一组正交归一基矢了吗？因此，找到一个最大线性无关的矢量组，简称最大线性无关组，是找到矢量空间的正交归一基矢的非常重要的一步。最大线性无关组的意思是说，在整个矢量空间中把尽可能多的矢量纳入进来，使得这些矢量之间线性无关，直到加入任何一个额外的矢量，这些矢量必然线性相关。

有数学定理（Zorn 引理）证明线性空间必然存在着这样的最大线性无关组，也就是存在着一组正交归一基矢。不过，这个定理比较抽象，我们不再去证明了。留给更加专门的矢量空间的课程和书去证明。在这里，我们暂时接受，任何矢量空间必然存在基矢的结论。当然，另一方面，由于我们总是以 $\mathbb{R}^n$ 和 $\mathbb{C}^n$ 为我们的矢量空间的原型，因此，这个有基矢的结论也很容易接受。

比如说，在 $\mathbb{R}^n$ 中，最大线性无关组的矢量个数就是 n。其中一个取法就是取各个基矢，$\{\hat{e}_j\}$。可以证明，它们线性无关。否则，

$$\sum_{j=1}^{n} \mu_j \left|\hat{e}_j\right\rangle = 0, \tag{17.89}$$

两边做对任意的 $\hat{e}_k$ 的内积，根据

$$\left\langle \hat{e}_k \middle| \hat{e}_j \right\rangle = \delta_{kj}, \tag{17.90}$$

得到

$$\sum_{j=1}^{n} \mu_j \delta_{kj} = \mu_k = 0. \tag{17.91}$$

这个和要求 μ_l 不全为零矛盾。因此，它们肯定线性无关。其实，任何一个基矢不能由其他基矢来线性叠加，这一点我们也早就从关于 $\mathbb{R}^n$ 的已有知识中知道了。

反过来，任何一个其他的矢量，都可以表示成这些基矢的线性叠加，也就是存在不完全为零的系数，使得对于任何一个 $\mathbb{R}^n$ 中的不为零的矢量，有

$$\left|a\right\rangle = \sum_{j=1}^{n} a_j \left|\hat{e}_j\right\rangle, \tag{17.92}$$

而这个正好就提供了一组满足线性相关性的要求的 $\{\mu_l\}$。任何多于 n 个的矢量放在一起必然线性相关。同时，存在着一组线性无关的 n 个矢量。因此，$\mathbb{R}^n$ 的最大线性无关组的矢量个数就是 n。

更一般地，对于一个矢量空间，我们只要找到了一个最大线性无关组[6]，我们也就知道了这个空间的维数，得到了这个空间的一组正交归一基矢。

有了基矢，一个矢量 a 在每一个基矢下的投影 a_j 可以通过内积算出来，对公式 (17.92) 的左右两边乘上 $\langle\hat{e}_i|$ 得到，

$$\langle\hat{e}_i|\,a\rangle = \langle i|\left[\sum_{j=1}^{n} a_j\,|\hat{e}_j\rangle\right] = \sum_{j=1}^{n} a_j\delta_{ij} = a_i, \tag{17.93}$$

因此，

$$a_i = \langle\hat{e}_i|\,a\rangle. \tag{17.94}$$

进一步与公式 (17.92) 合起来，得到

$$|a\rangle = \sum_{j=1}^{n} |\hat{e}_j\rangle\langle\hat{e}_j|\,a\rangle = \sum_{j=1}^{n} a_j\,|\hat{e}_j\rangle. \tag{17.95}$$

我们终于在一个 V 内的抽象矢量 a 和一群 F 上的数的行或者列 $\{a_j\}$ 之间建立起来了联系。

17.4 矢量空间上的线性算符：矩阵乘法规则为什么这么奇怪

下面我们来用基矢看看矢量空间上的线性算符是什么。一个附带的目标是帮你看清楚为什么矩阵乘法的规则会约定成这么特殊的样子，

$$(AB)_{ij} = \sum_{k} A_{ik}B_{kj}. \tag{17.96}$$

从矢量空间 V 到 V 的线性算符的定义如下。

[6]对于无限维矢量空间，我们还得另论。不过，在本书中，我们基本上不关注无限维矢量空间。

定义 17.9. 一个 $V \to V$ 上的映射，记作 $\hat{A}$，如果满足条件 $\hat{A}(\mu a + \nu b) = \mu\hat{A}(a) + \nu\hat{A}(b)$，则被称为 V 上的线性算符。

我们前面已经得到了从 V 到 F 的线性映射 $V^* = \{\langle a|\}$。那么，我们来看看 $|b\rangle\langle a|$ 是什么，是不是正好就是从 V 到 V 的线性映射?

首先，它作用在任何一个 V 中的元素 $|c\rangle$ 上，

$$|b\rangle\langle a|\, c\rangle = (\langle a|\, c\rangle)\,|b\rangle. \tag{17.97}$$

确实得到一个 V 中的元素。

其次，这个 $|b\rangle\langle a|$ 也满足线性性，

$$\begin{aligned} |b\rangle\langle a|\,[\mu\,|c\rangle + \nu\,|d\rangle] &= (\mu\langle a|\, c\rangle + \nu\langle a|\, d\rangle)\,|b\rangle \\ &= \mu\,(\langle a|\, c\rangle)\,|b\rangle + \nu\,(\langle a|\, d\rangle)\,|b\rangle \\ &= \mu\,|b\rangle\langle a|\, c\rangle + \nu\,|b\rangle\langle a|\, d\rangle. \end{aligned} \tag{17.98}$$

上一节，我们说了矢量可以表达为正交归一基矢的线性组合，

$$a = \sum_j a_j\,|\hat{e}_j\rangle = \sum_j a_j\,|j\rangle, \tag{17.99}$$

其中为了偷懒我们简化了基矢的记号为 $|j\rangle$。所谓正交归一指的是，

$$\delta_{ij} = \langle i\,|j\rangle. \tag{17.100}$$

其中 δ_{ij} 是Kronecker δ 记号，i 和 j 相等的时候等于 1，否则等于零。

于是，每一个 $|b\rangle\langle a|$ 可以写成如下的线性组合，

$$\begin{aligned} |b\rangle\langle a| &= \sum_i b_i^*\,|i\rangle \sum_j a_j\langle j| \\ &= \sum_{i,j} b_i^* a_j\,|i\rangle\langle j|\,. \end{aligned} \tag{17.101}$$

也就是说，更基本的单位是那些基矢形式的算符，$|i\rangle\langle j|$。以它们为基础，则其他像 $|b\rangle\langle a|$ 这样的算符都可以表达为这些算符的线性叠加的形式。

于是，我们猜测，更一般的 $V \to V$ 上的线性算符 $\hat{A}$，也可以表示为

$$\hat{A} = \sum_{i,j} A_{ij}\,|i\rangle\langle j|\,. \tag{17.102}$$

沿着这个思路，我们试试。根据 $\hat{A}$ 的线性性，$\hat{A}$ 作用到任何一个矢量上，可以表示为

$$\begin{aligned}\hat{A}a &= \hat{A}\left(\sum_j a_j \left|j\right\rangle\right) \\ &= \sum_j a_j \left(\hat{A}\left|j\right\rangle\right).\end{aligned} \tag{17.103}$$

而这个 Aa 也是一个矢量，用同一组基矢展开就得到，

$$\sum_l \left(\hat{A}a\right)_l \left|l\right\rangle = \sum_j a_j \left(\hat{A}\left|j\right\rangle\right). \tag{17.104}$$

等式的左右两边都做一个和左矢量 $\langle i|$ 的内积来计算 $\left(\hat{A}a\right)$ 在 $\langle i|$ 基矢上的投影，我们得到

$$\begin{aligned}\left\langle i\right| \left(\sum_k \left(\hat{A}a\right)_k \left|k\right\rangle\right) &= \left\langle i\right| \left[\sum_j a_j \left(\hat{A}\left|j\right\rangle\right)\right], \\ \Rightarrow \sum_k \left(\hat{A}a\right)_k \delta_{ik} &= \sum_j a_j \left(\left\langle i\right| \hat{A}\left|j\right\rangle\right), \\ \Rightarrow \left(\hat{A}a\right)_i &= \sum_j a_j \left(\left\langle i\right| \hat{A}\left|j\right\rangle\right).\end{aligned} \tag{17.105}$$

如果我们定义

$$A_{ij} = \left\langle i\right| \hat{A}\left|j\right\rangle, \tag{17.106}$$

我们真的就可以用这些数来计算算符 $\hat{A}$ 如何作用在一个矢量 a 上，

$$\left(\hat{A}a\right)_i = \sum_j A_{ij} a_j, \tag{17.107}$$

或者写成完整的矢量的形式而不是上面的投影分量的形式，有

$$\begin{aligned}\hat{A}a &= \sum_i \left(\hat{A}a\right)_i \left|i\right\rangle = \sum_{ij} \left|i\right\rangle A_{ij} \left\langle j\right| a\rangle \\ &= \sum_i \left|i\right\rangle\left\langle i\right| \hat{A} \left(\sum_j \left|j\right\rangle\left\langle j\right|\right) \left|a\right\rangle.\end{aligned} \tag{17.108}$$

也就是说，对于一个给定的矢量 a，$\hat{A}a$ 只需要按照公式 (17.107) 基于 A_{ij} 算出来。甚至，在公式 (17.108) 中，我们只需要在 $\hat{A}a$ 的两个地方分别插入

$\left(\sum_j |j\rangle\langle j|\right)$ 就行。这样的插入不改变矢量的唯一的可能性是插入的部分是一个恒等变换，或者说保持矢量不变的线性算符——记作 I，于是，

$$I = \sum_j |j\rangle\langle j| . \tag{17.109}$$

习题 17.7 (验证 $I = \sum_j |j\rangle\langle j|$). 通过把 $I = \sum_j |j\rangle\langle j|$ 作用在一个矢量上的方式，或者找到对应的矩阵的元素的方式，证明 I 确实是一个很等变换：对于任意的 $a \in V$，$Ia = a$。

我们终于在一个 $V \to V$ 上的线性算符 $\hat{A}$ 和一个 F 上的数的方阵 (A_{ij}) ——一个由 F 上的数构成的方阵称为矩阵——之间建立起来了联系，并且如果把这个算符作用到一个矢量上，其投影分量可以通过公式 (17.107) 来计算。

如果我们有连续两个线性算符 $\hat{A}$ 和 $\hat{B}$ 作用到一个矢量上，我们来看看其结果会怎样。我们记 $b = \hat{A}a$，$\hat{B}b$，则

$$c_i = \sum_j B_{ij} b_j, b_j = \sum_k A_{jk} a_k, \tag{17.110}$$

合起来，

$$c_i = \sum_{j,k} B_{ij} A_{jk} a_k = \sum_k \left(\sum_j B_{ij} A_{jk} \right) a_k = \sum_j \left(\sum_k B_{ik} A_{kj} \right) a_j. \tag{17.111}$$

最后一步我们交换了一下被取和的角标——被取和的部分的具体的变量名称是无所谓的只需要保证每一个取和的内部关系不变，例如 $\sum_{k=1}^{2} a_k = a_1 + a_2 = \sum_{j=1}^{2} a_j$。从这个结果我们得到，如果我们把 $(\sum_k B_{ik} A_{kj})$ 看作一个整体，记作算符 $\hat{BA}$，则

$$(BA)_{ij} = \sum_k B_{ik} A_{kj}. \tag{17.112}$$

这就是矩阵乘法的规则：如果我们把矩阵写成长方形，则两个矩阵相乘先得保证第一个矩阵的列数等于第二矩阵的行数——也就是这里的 k 的个数，然后，把第一个矩阵的每一行乘上第二个矩阵的每一列，得到一个数，行乘以列的时候每一个元素都对应相乘，然后加起来。我们看到，这么一个看起来非常特殊的规则，并不是天上掉下来的，而是由于线性空间有正交归一基矢以及线性算符的线性性的自然结果。

我们来做一个从 V 到 V 的抽象的线性算符 $\hat{A}$ 和表示成 F 上的数的矩阵 A 的小结：**矢量不是一列或者一行数，算符不是一个数的方阵。后者仅仅是前者在 V 上的给定的一组正交归一基矢下的投影分量。所有的矢量性和算符性，完全由那个不知道具体对象是什么的抽象矢量部分承担，数仅仅是这些抽象矢量前面的参与相加的系数。**在给定正交归一基矢下，$\hat{A}$ 和矩阵 A 的联系是 $\hat{A} = \sum_{ij} |i\rangle\langle i| \hat{A} |j\rangle\langle j| = \sum_{ij} A_{ij} |i\rangle\langle j|$。这个关系也可以这样看：在合适的地方插入 $I = \sum_j |j\rangle\langle j|$。更进一步，无论是两个算符前后连续作用——对应到矩阵相乘，还是算符作用到一个矢量上，还是两个矢量自己的内积，都可以看作是在在合适的地方插入 $I = \sum_j |j\rangle\langle j|$。在这个过程中，我们主要运用的公式就是基矢之间的正交归一关系，也就是 $\langle i| k\rangle = \delta_{ij}$。

关于最后这个结论，我们重写写一下之前分散写的结果：两个算符前后作用，

$$\hat{B}\hat{A} = \sum_i |i\rangle\langle i| \hat{B} \sum_k |k\rangle\langle k| \hat{A} \sum_j |j\rangle\langle j| = \sum_{i,j} |i\rangle \left(\sum_k B_{ik} A_{kj} \right) \langle j| . \quad (17.113)$$

算符作用到矢量上，

$$\hat{A} |a\rangle = \sum_i |i\rangle\langle i| \hat{A} \sum_j |j\rangle\langle j| a\rangle = \sum_i |i\rangle \left(\sum_j A_{ij} a_j \right) . \quad (17.114)$$

两个矢量内积，

$$\langle b |a\rangle = \langle b| \left(\sum_i |i\rangle\langle i| \right) |a\rangle = \sum_i b_i^* a_i . \quad (17.115)$$

这样我们就从抽象矢量记号完全走到了熟悉的分量的记号。

Dirac 符号还有一个好处：$|a\rangle$ 是 V 中的元素，是 $V^* \to F$ 的线性映射；$\langle a|$ 是 V^* 中的元素，是 $V \to F$ 的线性映射；$|a\rangle\langle b|$ 是 $V \to V$ 的线性映射，$\langle a| b\rangle$ 是内积；$|a\rangle$ 和 $\langle a|$ 之间，存在着一一对应关系。一个符号就表示了这一切。

写成给定正交归一基矢下的投影分量的形式，则：列矢量是 $|a\rangle$ 是 V 中的元素（在给定基矢下的投影分量），是从行矢量到数的映射；行矢量是 V^* 中的元素，是 $V \to F$ 的线性映射；列矢量乘以行矢量得到的是矩阵，是 $V \to V$ 的线性映射，行矢量乘以列矢量是内积；行矢量和列矢量之间存

在着一一对应关系——把每一个元素做一个复共轭操作，然后把行排成列或者把列排成行。顺便，这个复共轭和转置联合起来的操作成为共轭转置。

对比这两段，抽象矢量概念和记号这部分，其实就相当于我们什么也没学，不过就是用新的更加抽象的符号把关于 $\mathbb{R}^2$ 的中学知识换了一个说法。未来你会看到这个说法意义巨大作用巨大。其实，你现在就看到了，在这个概念和记号下，矩阵乘法不再是一个规定，而是自然的要求。否则，把一堆数排成一个特定的模样，用特定的方式来计算乘法，简直就是没头没脑。

17.4.1 $V^* \to V^*$ 上的线性映射 *

下面我们来说明，基于 $V \to V$ 上的线性映射 $\hat{A}$，加上在 V 到 V^* 之间的左右矢映射——记作 Φ 和 Φ^{-1} 分别表示把右矢量映射为左矢量和反过来的映射，可以得到 $V^* \to V^*$ 上的线性映射 $\hat{A}^\dagger = \Phi\hat{A}\Phi^{-1}$。

如图 17.3 所示，我们已经有了一个从 $V \to V$ 上的线性映射 A，可以把一个右矢量 $|a\rangle$ 映射为一个右矢量 $\hat{A}\,|a\rangle$。我们希望可以把一个左矢量 $\langle a|$ 自然地映射为另一个做矢量，在已有的 $\hat{A}$ 的帮助下。如果能够找到这个自然的映射，我们就把这个映射记作 $\hat{A}^\dagger$，表示这是 $\hat{A}$ 诱导出来的 $V^* \to V^*$ 上的线性映射。一会我们还会证明，这个 $\hat{A}^\dagger$ 的符号，对应到这个诱导出来的算符的分量表示，正好就是共轭转置。

要让 $\hat{A}$ 发挥作用，我们需要一个右矢量，我们的出发点是一个左矢量 $\langle a|$。怎么办呢？如图 17.3 所示，我们可以运用在 V 到 V^* 之间的一个左右矢之间的自然映射。也就是，$\langle a| \xrightarrow{\Phi^{-1}} |a\rangle$ 和 $|a\rangle \xrightarrow{\Phi} \langle a|$。当得到 $|a\rangle$ 之后，再让 $\hat{A}$ 作用上去，得到 $A\,|a\rangle$。接着，把这个右矢量通过 Φ 再变成一个左矢量，就完全实现了我们的要求。

定义 17.10 (伴算符，共轭算符). 由作用在右矢量上的算符 $\hat{A}$ 通过 $\hat{A}^\dagger = \Phi\hat{A}\Phi^{-1}$ 定义的作用在左矢量上的算符 $\hat{A}^\dagger$ 被称为 $\hat{A}$ 的伴算符，也叫做共轭算符。

我们来运用 $\hat{A}$ 的投影分量形式的矩阵 A 来看看这个 $\hat{A}^\dagger$ 的投影分量形

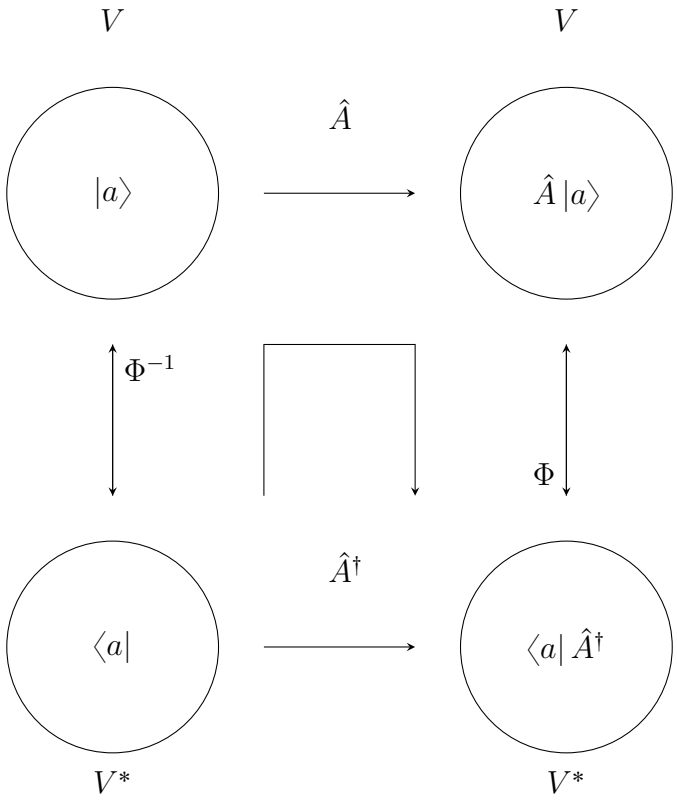


图 17.3: $V^* \to V^*$ **上的线性映射** $\hat{A}^\dagger$。基于 $V \to V$ 上的线性映射 $\hat{A}$ 可以得到 $V^* \to V^*$ 上的线性映射 $\hat{A}^\dagger = \Phi\hat{A}\Phi^{-1}$。

式——下面记作 $A^\dagger$。如图 17.3 所示，我们定义 $\hat{A}^\dagger$ 为，

$$\begin{aligned}\langle i|\,\hat{A}^\dagger = \Phi\left(\hat{A}\,|i\rangle\right) &= \Phi\left(\sum_k |k\rangle\langle k|\,\hat{A}\,|i\rangle\right)\\ &= \Phi\left(\sum_k |k\rangle\, A_{ki}\right) = \sum_k A^*_{ki}\langle k|\,.\end{aligned} \tag{17.116}$$

对这个等式两边做对 $|j\rangle$ 的内积，得到

$$A^\dagger_{ij} = \langle i|\,\hat{A}^\dagger\,|j\rangle = \sum_k A^*_{ki}\langle k|\,j\rangle = \sum_k A^*_{ki}\delta_{jk} = A^*_{ji}. \tag{17.117}$$

这刚好说明 $A^\dagger$ 矩阵就是 A 矩阵的共轭转置。

注意，按照定义 $\hat{A}$ 是作用在右矢量 $|a\rangle$ 上的，而 $\hat{A}^\dagger$ 是作用在左矢量 $\langle a|$ 上的。可是明明我们已经做了类似 $\langle i|\,\hat{A}\,|j\rangle$ 和 $\langle i|\,\hat{A}^\dagger\,|j\rangle$ 这样的计算啊。当然，我们总是可以这样来理解：例如对于 $\langle i|\,\hat{A}\,|j\rangle$，我们先让 $\hat{A}$ 作用到 $|j\rangle$ 上——这个符合定义，得到一个右矢量 $|b\rangle = \hat{A}\,|j\rangle$，然后我们求这个右矢量 $|b\rangle$ 的在给定基矢下的分量，于是，需要计算 $\langle i|\,\hat{A}\,|j\rangle$。我们是不是索性从概念上就允许 $\hat{A}$ 作用到左矢量 $\langle j|$ 上得了呢？当然，这样允许以后还得和之前的计算结果保持一致。

我们来思考，如果能够作用了，我们期待的 $\langle a| = \langle j|\,\hat{A}$ 是什么，是一个左矢量。左矢量是什么，是一个从 V 到 F 的映射。也就是说，如果我们能够给出来对于任何一个右矢量 $|b\rangle$ 上的作用结果 $\langle a|\,b\rangle$，并且证明这个映射是线性的，那么，我们就得到了一个左矢量。于是，我们定义

$$\left(\langle j|\,\hat{A}\right)|b\rangle \triangleq \langle j|\left(\hat{A}\,|b\rangle\right), \tag{17.118}$$

就得到了 $\left(\langle j|\,\hat{A}\right)$。甚至，我们可以通过运用 $I = \sum_j |j\rangle\langle j|$ 把这个 $\left(\langle j|\,\hat{A}\right)$ 的投影分量形式求出来，

$$\langle j|\,\hat{A} = \left(\langle j|\,\hat{A}\right)\sum_k |k\rangle\langle k| = \langle j|\left(\hat{A}\sum_k |k\rangle\right)\langle k| = \sum_k A_{jk}\langle k| \tag{17.119}$$

同样的道理，我们定义

$$\langle a|\left(\hat{A}^\dagger\,|j\rangle\right) \triangleq \left(\langle a|\,\hat{A}^\dagger\right)|j\rangle, \tag{17.120}$$

就得到了 $\left(\hat{A}^\dagger |j\rangle\right)$，以及投影分量形式，

$$\hat{A}^\dagger |j\rangle = \sum_k A^*_{jk} |k\rangle = \sum_k \left(A^\dagger\right)_{kj} |k\rangle . \tag{17.121}$$

于是，从今以后，我们再也不区分 $\hat{A}$ 和 $\hat{A}^\dagger$ 到底作用在右矢量还是左矢量上了。我们看到能够这样做的秘诀还是 $I = \sum_j |j\rangle\langle j|$。

既然 $\hat{A}$ 和 $\hat{A}^\dagger$ 都可以作用在左矢量和右矢量上了，我们来定义一个厄米算符。

定义 17.11 (厄米算符). 如果一个算符 $\hat{A}$ 和它的诱导算符 $\hat{A}^\dagger$ 相等，$\hat{A} = \hat{A}^\dagger$，则称 $\hat{A}$ 为厄米算符，或者自伴算符。

17.4.2 矢量空间的基矢变换

我们说，一个矢量空间的正交归一基矢的选择可能不唯一。例如在 $\mathbb{R}^2$ 中，我们就可以选择水平和垂直方向当作基矢，也可以选择 45° 和 135° 的方向当作基矢，甚至任意的两个相互垂直的方向当作基矢都可以。我们来看看，如果同一个空间我们有两组或者两组以上的正交归一基矢，我们会看到什么？

我们会看到，同一个矢量可以表示在同一个矢量空间的不同的基矢下，它们的分量表示必定要满足下面的关系。假设我们有两组正交归一基矢，记作 $\{|u_l\rangle\}$ 和 $\{|v_l\rangle\}$，并且我们有两者之间的关系

$$|v_i\rangle = \sum_{j=1}^{n} S_{ji} |u_j\rangle . \tag{17.122}$$

这里我们定义了一堆已经给定的数排成的方阵 S。对公式 (17.122) 两边乘以左矢量 $|u_j\rangle$，可以得到

$$S_{ij} = \langle u_i | v_j\rangle . \tag{17.123}$$

因此，把 S 写成一个方阵的形式，有

$$S = \begin{bmatrix} \langle u_1 | v_1\rangle & \langle u_1 | v_2\rangle & \langle u_1 | v_3\rangle & \cdots \\ \langle u_2 | v_1\rangle & \langle u_2 | v_2\rangle & \langle u_2 | v_3\rangle & \cdots \\ \langle u_3 | v_1\rangle & \langle u_3 | v_2\rangle & \langle u_3 | v_3\rangle & \cdots \\ \cdot & \cdot & \cdot & \cdots \end{bmatrix} \tag{17.124}$$

这样看起来，S 的每一列，例如第 j 列，都是矢量 v_j 在基矢 $\{|u_l\rangle\}$ 下的投影分量。下面，我们还会用到 $S^\dagger$，

$$\left(S^\dagger\right)_{ij} = S_{ji}. \tag{17.125}$$

它是 S 的共轭转置——也就是把每个元素做复共轭然后把 i 行 j 列的元素放到 j 行 i 列上，

$$S^\dagger = \begin{bmatrix} \langle v_1 | u_1 \rangle & \langle v_1 | u_2 \rangle & \langle v_1 | u_3 \rangle & \cdots \\ \langle v_2 | u_1 \rangle & \langle v_2 | u_2 \rangle & \langle v_2 | u_3 \rangle & \cdots \\ \langle v_3 | u_1 \rangle & \langle v_3 | u_2 \rangle & \langle v_3 | u_3 \rangle & \cdots \\ \cdot & \cdot & \cdot & \cdots \end{bmatrix} \tag{17.126}$$

也就是，$S^\dagger$ 的每一列，例如第 j 列，都是矢量 u_j 在基矢 $\{|v_l\rangle\}$ 下的投影分量。

根据基矢 $\{|v_l\rangle\}$ 的正交归一性，我们得到，

$$\begin{aligned} \delta_{ij} = \langle v_i | \, v_j \rangle &= \sum_{l,m=1}^{n} (S_{li})^* \, S_{mj} \langle u_l | \, u_m \rangle \\ &= \sum_{l,m=1}^{n} (S_{li})^* \, S_{mj} \delta_{lm} = \sum_{l}^{n} (S_{li})^* \, S_{lj} \\ &= \sum_{l}^{n} \left(S^\dagger\right)_{il} S_{lj} \end{aligned} \tag{17.127}$$

因此，我们得到这一个系数方阵必须满足的条件是

$$S^\dagger S = I. \tag{17.128}$$

有了这个条件，我们可以得到，

$$|u_i\rangle = \sum_{j=1}^{n} S^\dagger_{ji} \, |v_j\rangle . \tag{17.129}$$

因为公式 (17.122) 其实是 n 个 $|v_i\rangle$ 和 n 个 $|u_i\rangle$ 之间的线性方程，解必然唯一，而只有上面的那一组解才能得到

$$|v_i\rangle = \sum_{j=1}^{n} S_{ki} S^\dagger_{jk} \, |v_j\rangle = \sum_{j=1}^{n} \delta_{ij} \, |v_j\rangle = |v_i\rangle , \tag{17.130}$$

于是自然满足方程解的要求。

有了公式 (17.129)，我们可以参照前面证明 $S^\dagger S = I$ 的过程来证明

$$SS^\dagger = I. \tag{17.131}$$

习题 17.8 ($SS^\dagger = I$)**.** 证明 $SS^\dagger = I$。

现在我们来看任意一个矢量在两套基矢下的投影分量之间的联系。任意一个矢量 a 可以表达为两种形式，

$$|a\rangle = \sum_{j=1}^{n} a_j |u_j\rangle, |a\rangle = \sum_{j=1}^{n} a'_j |v_j\rangle. \tag{17.132}$$

于是，

$$a_j = \langle u_j | a\rangle = \sum_{k=1}^{n} \left(S^\dagger_{kj}\right)^* \langle v_k | a\rangle = \sum_{k=1}^{n} S_{jk} a'_k. \tag{17.133}$$

把两套投影分量都写成列矢量的形式 $a = \begin{bmatrix} a_1 \\ a_2 \\ \cdot \end{bmatrix}$，则凑巧了有

$$a = Sa'. \tag{17.134}$$

于是，两边乘上 $S^\dagger$ 得到，

$$a' = S^\dagger a. \tag{17.135}$$

这个关系被称为列矢量的坐标变换——同一个列矢量在两套不同的基矢下的投影分量之间的联系。

我们看到，这个本来是一堆确定的 F 上的数排成的方阵 S，看起来竟然形式上和一个算符作用到矢量上很像，那里我们有 $b = \hat{A}a$。在选定基矢之后，算符 $\hat{A}$ 也可以表示成一个 F 上的数的方阵。是不是我们真的能够给

这个 S 一个算符的含义呢？我们试试这个定义，

$$\begin{aligned} S = & \sum_{ij} S_{ij} |u_i\rangle\langle u_j| \\ = & \sum_{ij} \langle u_i |v_j\rangle |u_i\rangle\langle u_j| \\ = & \sum_j \left(\sum_i |u_i\rangle\langle u_i| \right) |v_j\rangle\langle u_j| \\ = & \sum_j |v_j\rangle\langle u_j| . \end{aligned} \tag{17.136}$$

其中我们再一次用到了 $I = \sum_i |u_i\rangle\langle u_i|$。这个时候，我们发现，$S$ 的含义非常自然：吃掉一个 $|u_j\rangle$ 得到一个 $|v_j\rangle$。这正好就是坐标变换的含义。于是，

$$S^\dagger = \sum_j |u_j\rangle\langle v_j| , \tag{17.137}$$

也就是吃掉一个 $|v_j\rangle$ 得到一个 $|u_j\rangle$。这样我们就得到了 S 和 $S^\dagger$ 的不依赖于基矢选择的算符的含义。从现在起，我们就把 S 和 $S^\dagger$ 也当作算符了。

最后，对于基矢选择和相应的坐标变换，实际上，我们可以在保持两套基矢整体不变的条件下对基矢重新排序，但是，仍然应该看作是同一个坐标变换。但是，对基矢重新排序会影响 S 和 $S^\dagger$。例如，交换第一号和第二号的基矢，我们会得到，

$$S_1 = |v_1\rangle\langle u_1| + |v_2\rangle\langle u_2| + \sum_{j>2} |v_j\rangle\langle u_j| , \tag{17.138}$$

$$S_2 = |v_2\rangle\langle u_1| + |v_1\rangle\langle u_2| + \sum_{j>2} |v_j\rangle\langle u_j| . \tag{17.139}$$

写成公式 (17.124) 那样的矩阵的形式，两者之间其实就相当于交换了两列。因此，我们把这样的相当于一次或者多次交换了两列的坐标变换看成是同一个坐标变换——称为相似变换。

我们再来看看，在两套不同的基矢下，一个算符 $\hat{A}$ 的投影分量，矩阵 A，将会如何变化。根据公式 (17.106)，

$$A_{ij} = \langle u_i| \hat{A} |u_j\rangle , \tag{17.140}$$

$$A'_{ij} = \langle v_i| \hat{A} |v_j\rangle , \tag{17.141}$$

则

$$\begin{aligned} A'_{ij} &= \langle v_i| \left(\sum_k |u_k\rangle\langle u_k| \right) \hat{A} \left(\sum_l |u_l\rangle\langle u_l| \right) |v_j\rangle, \\ &= \sum_{k,l} \left(\langle v_i|\, u_k\rangle\right) \langle u_k| \hat{A} |u_l\rangle \left(\langle u_l|\, v_j\rangle\right), \\ &= \sum_{k,l} S^\dagger_{ik} A_{kl} S_{lj} \end{aligned} \tag{17.142}$$

回到通常的分量形式的矩阵的语言，这表示

$$A' = S^\dagger A S. \tag{17.143}$$

这样我们就得到了在坐标变换下，或者说基矢变换下，矢量和算符的投影分量的变化形式（分别为公式 (17.135) 和公式 (17.143)）。注意矢量和算符它们自己是不变的，改变的仅仅是其在不同基矢的投影分量。

17.5 表示论和广义表示论，数学建模

在上一节中，我们看到一个算符可以用一堆数来表示，具体来说就是一个由一堆数排列而成的方阵，也就是矩阵，其含义是这个算符在选定基矢下的投影分量。尽管这样做以后，这些矩阵的乘法运算必须采取一种看起来相当奇怪的特殊的方式才能反映两个算符相乘以及一个算符作用到一个矢量上。从现在开始，我们把矢量也叫做矩阵，只不过列矢量是 $n \times 1$ 型的矩阵，而行矢量是 $1 \times n$ 型的矩阵。

更进一步，通过内积矢量可以把另一个矢量变成数：右矢量 $|a\rangle$ 把左矢量变成数，左矢量 $\langle a|$ 把右矢量变成数，矩阵把一个 A 把一个左矢量和一个右矢量合起来变成数（矩阵把一个 A 把一个右矢量变成一个右矢量，而右矢量需要一个左矢量来变成数，因此，矩阵需要一个左矢量和一个右矢量合起来，才能变成数）。我们用内积的符号，说，右矢量是一个 $(1,0)$ 型的张量，右矢量是一个 $(0,1)$ 型的张量，而矩阵是一个 $(1,1)$ 型的张量。这里的 (m,n) 型的张量这个名称的含义是这个数学对象把 m 个左矢量和 n 个右矢量变成数，并且对右矢量满足线性性，对左矢量满足共轭线性性。

定义 17.12 (张量). 从 $\left(\underbrace{V^* \otimes V^* \otimes \cdots \otimes V^*}_{m}\right) \otimes \left(\underbrace{V \otimes V \otimes \cdots \otimes V}_{n}\right)$ 到 F

的对 V 中元素满足线性性，对 V^* 中元素满足共轭线性性的映射被称为张量，更具体地被称为 (m,n) 型张量。

狭义表示就是把数学对象，包含群元素、矢量、算符、张量都写成矩阵的形式，使得这些能够对数学对象做的操作可以用矩阵运算来得到，并且保证通过矩阵运算计算得到的结果和对原来的数学对象的操作结果相对应。例如，对于群元素，这就要求 $g_1 \to A_1$，$g_2 \to A_2$，使得 $g_1 g_2 \to A_1 A_2$ 成立。例如，对于算符，这就要求 $hatA \to A$，$\hat{B} \to B$，使得 $\hat{A}\hat{B} \to AB$ 和 $\mu\hat{A} + \nu\hat{B} \to \mu A + \nu B$ 成立。因此，实际上，一个表示不仅仅是把被表示对象和用来表示的对象之间的映射，更重要的是保持结构——也就是保持对被表示对象的操作，具体这里指的是被表示对象之间的乘法。

注意，对于群元素我们总是存在一个平凡表示：把所有的群元素都表示为单位矩阵 I。这样，无论群元素之间怎么算乘法，表示之后的乘法都会得到 I，而这个也正好是群元素乘法得到的结果的表示，也就是

$$a \cdot b = c \to II = I. \tag{17.144}$$

满足表示的定义。如果表示是一一的，也就是任何两个不同的被表示的对象的矩阵是不同的，则称这个表示为忠实表示。显然上面的平凡表示是不忠实的，表示映射把所有的被表示元素都映射成了同一个矩阵。显然，忠实表示是我们期待的，这样我们就可以用矩阵运算来表示一个原来不知道是什么的抽象对象。

例 17.11 (硬币翻转群 $\mathbb{Z}_2$ 的表示). 把包含 I, X 两个元素的群用矩阵来表示。

首先，我们说 $I \to 1, X \to 1$ 是一个表示，因为 $II = I \to 1 \times 1 = 1$，$IX = X \to 1 \times 1 = 1$，$XI = X \to 1 \times 1 = 1$，$XX = I \to 1 \times 1 = 1$。

其次，我们说 $I \to 1, X \to -1$ 是一个表示，因为 $II = I \to 1 \times 1 = 1$，$IX = X \to 1 \times (-1) = -1$，$XI = X \to (-1) \times 1 = -1$，$XX = I \to (-1) \times (-1) = 1$。

最后，我们说 $I \to \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, X \to \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ 是一个表示，因为 $II = I \to \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \times \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$。其他几个乘法的验算当作练习。

习题 17.9 (硬币翻转群 $\mathbb{Z}_2$ 的表示). 验算 $I \to \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, X \to \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ 是硬币翻转群 $\mathbb{Z}_2$ 的一个表示。

你看，本来我们只知道翻转和不翻转的操作的含义，它们不是数，但是，有了表示，它们就成了一堆数。我们就可以用对数的操作来操作这些对象了。这就是表示的为例。

更一般的表示指的是建立起来两种数学对象之间的映射，使得其中那种可以对被表示对象做的操作，也可以在用来对表示的数学对象做相应的操作，保证操作得到的结果，刚好就是被表示对象的操作结果对应的表示对像。也就是说，用来表示的对象不一定是矩阵。

最最一般的表示指的是，现实世界的对象（也可以是一种数学对象）等任何一个对象的某一个集合中的所有的元素，都被映射到了一个（另一种，如果本来就是数学对象的话）数学对象，并且保证能够对原来的被表示对象的操作得到的结果，也正好被映射为对用来表示的数学对象的操作的结果。

例如，我们用三维空间的矢量表示物理上的位置、速度、加速度，甚至我们用位置对时间的函数表示物理上的运动过程，甚至更进一步我们用对位置对时间的函数的导数来得到速度对时间的函数，以及反过来用对速度对时间的函数的积分来得到位置对时间的函数。这些都是最最广义的表示。

实际上，科学研究很大程度上，就是再找现实世界的对象的数学表示，然后检验经过这样的表示之后，对这些表示做得到的计算结果，是否正好就是被表示的现实世界的对象在相应的操作下的结果。前者称为数学建模，后者称为实验检验。

顺便，只有学习了群和群的表示，学习了代数，你才能看到实际上数学建模甚至整个科学研究就是找给现实世界足够简单同时尽量忠实的表示（是否确实是表示需要通过实验来检验）。

最后，我们给出来一个用一种数学对象——矩阵表示另一种数学对象——四元数的例子。这是一个第二种意义下的表示的很有意思的例子。

小知识 17.1 (四元素的矩阵表示). 四元数是由 *William Hamilton*（威廉 · 哈密顿）提出来的对复数的拓展

$$\xi = \xi_0 + \xi_1 i + \xi_2 j + \xi_3 k. \tag{17.145}$$

其中 ξ_0,ξ_1,ξ_2,ξ_3 都是实数。

四元素的数乘和加法很简单，就是通常数或者说矢量、复数的加法和数乘的拓展，

$$a\xi = a\xi_0 + a\xi_1 i + a\xi_2 j + a\xi_3 k, \tag{17.146}$$

$$\xi + \eta = (\xi_0 + \eta_0) + (\xi_1 + \eta_1)\, i + (\xi_2 + \eta_2)\, j + (\xi_3 + \eta_3)\, k. \tag{17.147}$$

四元素之间的乘法需要补充额外的规则，

$$ii = -1, jj = -1, kk = -1, \tag{17.148}$$

$$ij = k, ji = -k, jk = i, kj = -i, ki = j, ik = -j. \tag{17.149}$$

这样就能够通过分配律来计算任何两个四元素之间的乘法了。

$$\begin{aligned}\xi\eta = (\xi_0 + \xi_1 i + \xi_2 j + \xi_3 k)\,(\eta_0 + \eta_1 i + \eta_2 j + \eta_3 k)& \\ = (\xi_0\eta_0 - \xi_1\eta_1 - \xi_2\eta_2 - \xi_3\eta_3)& \\ + (\xi_0\eta_1 + \xi_1\eta_0 + \xi_2\eta_3 - \xi_3\eta_2)\, i& \\ + (\xi_0\eta_2 + \xi_2\eta_0 + \xi_3\eta_1 - \xi_1\eta_3)\, j& \\ + (\xi_0\eta_3 + \xi_0\eta_3 + \xi_1\eta_2 - \xi_2\eta_1)\, k.&\end{aligned} \tag{17.150}$$

注意，这个乘法规则不满足交换律，$ij = k, ji = -k$ 两者不同。注意复数乘法是满足交换律的。因此，这里的四元数尽管来自于对复数的拓展，但是可能已经远远突破了复数的概念了。

回顾已有的数学结构，矩阵也是不满足交换律的数学结构。那是不是，其实四元数可以看作是矩阵呢？或者说可以用矩阵来表示四元数呢？

数学家们发现，确实可以这样表示，而且这样表示以后，四元数就和旋转联系了起来，用来表示四元数的矩阵构成的这个群还有一个自己的名字——$SU(2)$ 群。这是在物理学上非常重要的一个群。这个表示就是，

$$\xi = \xi_0 I + i\xi_1 X + i\xi_2 Y + i\xi_3 Z, \tag{17.151}$$

其中 i 是虚数单位，

$$I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, X = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, Y = \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix}, Z = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}. \tag{17.152}$$

可以验证，

$$iXiX = -I, iYiY = -I, iZiZ = -I, \tag{17.153}$$

$$iXiY = iZ, iYiX = -iZ, iYiZ = iX,$$

$$iZiY = -iX, iZiX = iY, iXiZ = -iY. \tag{17.154}$$

完全和公式 (17.148) 和公式 (17.149) 对应起来了。因此，矩阵形式的 ξ 就是四元数形式的 ξ 的表示，并且是忠实表示——具有一一对应关系。

也就是说，其实，不需要再拓展复数到四元数，只需要在复数的基础上引入矩阵，就足以表示四元数。于是，任何一个可以用四元数来做的事情，就可以用复数域上的这几个矩阵来做。

至于用四元数来做什么呢？当年*Hamilton*提出来的时候，纯粹是好奇心驱动。不过，有了矩阵表式，我们就发现了，其实对应着旋转[7]，而且这个旋转在物理学里面对应着自旋在磁场中的旋转。至于为什么自旋在磁场中的旋转可以用前面这四个矩阵的加法构成的群来表示，那就是另一个故事了。等到学习量子物理学的时候，你就有机会体会到这一点了。

17.6 矩阵的谱分解

走过了前面的艰苦卓绝的把矢量的概念抽象化一般化的曲折的路，我们终于准备好开始学习矩阵或者说算符的两个最核心的定理了，一个是关于矩阵的谱分解的，一个是关于矩阵的 SVD 分解的。它们分别是这一节和写一节的内容。

之前，我们已经定义了厄米算符，$\hat{A} = \hat{A}^\dagger$，现在我们来讨论厄米算符的一个重要性质——其本征矢量的正交性。

定义 17.13 (本征矢量). 给定一个算符 $\hat{A}$，满足 $\hat{A}|\lambda\rangle = \lambda|\lambda\rangle$ 的不为零的矢量 $|\lambda\rangle$ ($|\lambda\rangle \neq 0$) 被称为算符 $\hat{A}$ 的本征矢量，λ 称为算符 $\hat{A}$ 本征值。

例如，硬币翻转算符 $X = |-1\rangle\langle 1| + |1\rangle\langle -1|$，其中 $|\pm 1\rangle$ 分别表示硬币的正面或反面，的本征矢量是 $|\pm\rangle = |-1\rangle \pm |1\rangle$。我们验算一下。

[7]这里有一个很不错的解释，不过需要的数学基础有点多，可以先看思路和结论：https://blog.csdn.net/weixin_45590473/article/details/122884112。2025 年 11 月 28 日访问。

例 17.12 (硬币翻转算符的本征矢量). 验证 $|\pm\rangle$ 是算符 X 的本征矢量。

$$\begin{aligned} X|+\rangle &= (|-1\rangle\langle 1| + |1\rangle\langle -1|)(|-1\rangle + |1\rangle) \\ &= |-1\rangle\langle 1|(|-1\rangle + |1\rangle) + |1\rangle\langle -1|(|-1\rangle + |1\rangle) \\ &= |-1\rangle + |1\rangle . \end{aligned} \tag{17.155}$$

其中我们用到了算符的线性性和基矢量的正交归一性 $\langle i|j\rangle = \delta_{ij}$。所以对应的本征值为 $\lambda = 1$。同理可得

$$\begin{aligned} X|-\rangle &= (|-1\rangle\langle 1| + |1\rangle\langle -1|)(|-1\rangle - |1\rangle) \\ &= |-1\rangle\langle 1|(|-1\rangle - |1\rangle) + |1\rangle\langle -1|(|-1\rangle - |1\rangle) \\ &= -|-1\rangle + |1\rangle \\ &= -1(|-1\rangle + |1\rangle) . \end{aligned} \tag{17.156}$$

所以对应的本征值为 $\lambda = -1$。

顺便，可以验证 $\langle +|-\rangle = 0$，

$$\begin{aligned} \langle +|-\rangle &= (\langle -1| + \langle 1|)(|-1\rangle - |1\rangle) \\ &= \langle -1|-1\rangle - \langle 1|-1\rangle + \langle -1|1\rangle - \langle 1|1\rangle \\ &= 1 - 0 + 0 - 1 = 0. \end{aligned} \tag{17.157}$$

下面我们来把这个例子变成一个一般性的定理。

本征值和本征矢量的方程 $\hat{A}|\lambda\rangle = \lambda|\lambda\rangle$ 可以用算符和矢量的投影分量的形式写成下面的线性方程，

$$\begin{aligned} \langle i|\hat{A}\left(\sum_j |j\rangle\langle j|\right)|\lambda\rangle &= \lambda\langle i|\lambda\rangle \\ \Rightarrow \sum_j A_{ij}\lambda_j &= \lambda\lambda_i \\ \Rightarrow \sum_j (A_{ij} - \lambda\delta_{ij})\lambda_j &= 0. \end{aligned} \tag{17.158}$$

这是一个线性方程组，写成矢量和矩阵的形式，而不是分量的形式就成了

$$\left(\hat{A} - \lambda\hat{I}\right)|\lambda\rangle = 0. \tag{17.159}$$

我们要求出来的解 λ_j 要满足不完全等于零。这个线性方程只要是每个方程都是独立的，不能通过加减乘除变成相同的，则必然要求所有的 $\lambda_j = 0$。例如，我们可以看到下面的两个方程，第一个只有零解（$x = 0, y = 0$），第二个有非零解（$x = 1, y = -1$），

$$\begin{cases} 2x + 3y = 0 \\ x + y = 0. \end{cases} \tag{17.160}$$

和

$$\begin{cases} 2x + 2y = 0 \\ x + y = 0. \end{cases} \tag{17.161}$$

两者的差别是：第一个方程的两个方程不能相互削去，而第二个方程的两个方程其实是同一个可以相互削去（把其中的第二个方程左右两边乘以 2，然后和第一个方程做差，就得到 $0 = 0$），

$$2x + 2y = 0 \Longleftrightarrow x + y = 0. \tag{17.162}$$

因此，这个方程有非零解的条件就是不同的 i 对应的行的方程通过加减乘除操作可以完全消掉其中的一行或者多行。我们把这个“左侧的系数矩阵可以通过不同的行之间的乘以一个常数再加到其他行上的操作得到一行或者多行为零”性质记做

$$Det\left(\hat{A} - \lambda\hat{I}\right) = 0, \tag{17.163}$$

称作矩阵或者说算符 $\left(\hat{A} - \lambda\hat{I}\right)$ 的行列式（determinant）为零。具体的行列式怎么算，完全可以通过这个要求来推导出来。不过，我们不管这个具体怎么算的问题。我们只需要直到，这是一个关于 λ 的 n 次多项式方程。根据代数基本定理，n 阶多项式方程在复数范围内有 n 个根，允许重根。于是，我们就得到算符 $\hat{A}$ 原则上有 n 个本征值 λ，不过允许有的 λ 相同。

定理 17.2: 厄米算符本征矢量

如果厄米算符具有本征矢量，则本征值为实数，并且不同本征值对应的本征矢量相互正交。

证明 17.4. 记厄米算符 $\hat{A}$ 的两个不同本征值对应的本征矢量为 $|\mu\rangle$ 和 $|\nu\rangle$ 对应的本征值为 μ 和 ν，而且 $\mu \neq \nu$。

我们先来证明本征值为实数。我们来计算下式，

$$\langle\mu|\hat{A}|\mu\rangle = \langle\mu|\left(\hat{A}|\mu\rangle\right) = \mu\langle\mu|\mu\rangle. \tag{17.164}$$

用上 $\hat{A} = \hat{A}^\dagger$ 的方式来计算，有

$$\begin{aligned}\langle\mu|\hat{A}|\mu\rangle = \langle\mu|\hat{A}^\dagger|\mu\rangle &= \left(\langle\mu|\hat{A}^\dagger\right)|\mu\rangle \\ &= \left[\langle\mu|\left(\hat{A}|\mu\rangle\right)\right]^* \\ &= \mu^*\langle\mu|\mu\rangle. \end{aligned}\tag{17.165}$$

于是

$$\begin{aligned}&\mu\langle\mu|\mu\rangle = \mu^*\langle\mu|\mu\rangle, \\ \Rightarrow\ &(\mu-\mu^*)\langle\mu|\mu\rangle = 0.\end{aligned}\tag{17.166}$$

只有 $\mu = \mu^*$ 才能满足条件，因为 $|\mu\rangle \neq 0$ 从而 $\langle\mu|\mu\rangle \neq 0$。

我们再来证明不同本征值对应的本征矢量正交。

$$\langle\mu|\hat{A}|\nu\rangle = \langle\mu|\left(\hat{A}|\nu\rangle\right) = \nu\langle\mu|\nu\rangle. \tag{17.167}$$

用上 $\hat{A} = \hat{A}^\dagger$ 的方式来计算，有

$$\begin{aligned}\langle\mu|\hat{A}|\nu\rangle = \langle\mu|\hat{A}^\dagger|\nu\rangle &= \left(\langle\mu|\hat{A}^\dagger\right)|\nu\rangle \\ &= \left[\langle\nu|\left(\hat{A}|\mu\rangle\right)\right]^* \\ &= \mu^*\langle\mu|\nu\rangle = \mu\langle\mu|\nu\rangle. \end{aligned}\tag{17.168}$$

因此

$$\begin{aligned}&\nu\langle\mu|\nu\rangle = \mu\langle\mu|\nu\rangle, \\ \Rightarrow\ &(\nu-\mu)\langle\mu|\nu\rangle = 0.\end{aligned}\tag{17.169}$$

根据 $\nu \neq \mu$，则有 $\langle\mu|\nu\rangle = 0$。

对于 $\mathbb{C}^n$，如果 $\hat{A}$ 刚好有 n 个不相等的本征值 α_l，那我们就可以把相应的本征矢量拿过来构成一个完备的正交的基矢集合，然后再归一化，就能得到一组正交归一基矢。在这一组基矢下，算符 A 可以表示成这样，

$$\hat{A} = \hat{A}\left(\sum_{l=1}^{n} |\alpha_l\rangle\langle\alpha_l|\right) = \sum_{l=1}^{n} \alpha_l |\alpha_l\rangle\langle\alpha_l| . \tag{17.170}$$

算符 $\hat{A}$，在其自身的本征矢量当作基矢的条件下，成了一堆数 α_l。这样所有的关于 $\hat{A}$ 的计算分析就只需要在这一堆数 α_l 上来做了。

但是，有的时候，会出现不同的本征值的数量小于空间维数 n 的情况。这个时候，我们可能就需要把算符 $\hat{A}$ 和别的算符合在一起来找到一组完备的正交的基矢集合了。这就需要这些算符之间满足一个特定的关系。

定义 17.14 (对易算符). 如果 $\hat{A}\hat{B} = \hat{B}\hat{A}$，则我们称 $\hat{A}, \hat{B}$ 对易。

对易算符的乘法可以交换顺序。一般情况下算符的作用顺序交换以后得到的结果不同。例如，我们可以通过计算发现之前定义的 X 和这里新定义的 $Z = |1\rangle\langle 1| - |-1\rangle\langle -1|$ 之间不对易，

$$\begin{aligned} XZ &= (|-1\rangle\langle 1| + |1\rangle\langle -1|)(|1\rangle\langle 1| - |-1\rangle\langle -1|) = |-1\rangle\langle 1| - |1\rangle\langle -1| , \\ ZX &= (|1\rangle\langle 1| - |-1\rangle\langle -1|)(|-1\rangle\langle 1| + |1\rangle\langle -1|) = |1\rangle\langle -1| - |-1\rangle\langle 1| . \end{aligned}$$

对于一组相互对易的算符，我们可以构建一组共同本征矢量。

定理 17.3: 对易算符的共同本征矢量

对易算符存在共同本征矢量。

证明 17.5. 设一对对易算符为 $\hat{A},\hat{B}$，已知 $\hat{A}\hat{B}=\hat{B}\hat{A}$，我们来证明 $\hat{A}$ 的本征矢量也是 $\hat{B}$ 的本征矢量。

$$\hat{A}\left(\hat{B}\left|\alpha\right\rangle\right)=\hat{B}\left(\hat{A}\left|\alpha\right\rangle\right)=\alpha\left(\hat{B}\left|\alpha\right\rangle\right). \tag{17.171}$$

如果只有一个 $|\alpha\rangle$ 对应着 $\hat{A}$ 的本征值 α，则

$$\hat{B}\left|\alpha\right\rangle=\beta\left|\alpha\right\rangle. \tag{17.172}$$

因此，在这个条件下，$|\alpha\rangle$ 必然也是 $\hat{B}$ 的本征矢量。

如果存在 $|\alpha_1\rangle,|\alpha_2\rangle$ 之类 L 个本征矢量——的对应着同一个 α。我们只需要在这个 L 维子空间内找到 B 的本征矢量 $|\beta\rangle$，它们必然就也是 $\hat{A}$ 的（本征值为 α 的）本征矢量。之所以有这个结果是因为，在这个子空间内，B 的本征矢量 $|\beta\rangle$ 必然是这些 α 对应的本征矢量的线性组合，

$$\left|\beta\right\rangle=\sum_{l=1}^{L}\mu_l\left|\alpha_l\right\rangle, \tag{17.173}$$

因此

$$\hat{A}\left|\beta\right\rangle=\sum_{l=1}^{L}\alpha\mu_l\left|\alpha_l\right\rangle=\alpha\left|\beta\right\rangle. \tag{17.174}$$

如果这个子空间内 $\hat{B}$ 的本征值都不同，那我们就得到了一个完备的正交的基矢集合。注意，就算这样做到了，仍然会出现算符 $\hat{B}$ 其实有相同本征值对应着的多个本征矢量，只要这些本征矢量对应着例如算符 A 的不同本征值，就没问题。如果在这个子空间内，$\hat{B}$ 仍然有一些本征值，我们就可以考虑引入算符 $\hat{C}$。只要它和之前的算符对易，就可以在 $\hat{B}$ 相同本征值对应的本征矢量的子空间内来求出来算符 $\hat{C}$ 的本征值和本征矢量。我们可以一直做下去，直到再也没有相同的本征值。

把这一节在前面的所有结论——算符一般有 n 个允许有相同取值的本征值，厄米算符本征值为实数，不同本征值对应的本征矢量正交，对易算

符具有共同本征矢量于是总能够通过包含足够多的对易算符来区分开相同取值的本征值——合起来，我们就得到了，给定一组相互对易的厄米算符 $\left\{\hat{A},\hat{B},\cdots\right\}$，我们就有一组完备的正交的基矢。这个按照一群对易算符的本征值和本征矢量来对矩阵做分解，当作基矢，来表示其他算符，尤其是这个算符 $\hat{A}$ 本身成了公式 (17.170) 的形式（重写在这里），

$$\hat{A}=\sum_{l=1}^{n}\alpha_l\left|\alpha_l\right\rangle\left\langle\alpha_l\right|, \tag{17.175}$$

称为算符 $\hat{A}$ 的谱分解。这里的谱就是本征值的意思。任何一个厄米算符 $\hat{A}$，只要在它自己也就是 $\hat{A}$ 的本征矢量当作基矢——称为算符和矢量的 $\hat{A}$ 表示——的条件下，其投影分量都只有对角元，也就是这个算符成了一个对角的矩阵。

我们记，

$$S=\sum_l\left|\alpha_l\right\rangle\left\langle l\right|, S_{ij}=\left\langle i\right|S\left|j\right\rangle=\left\langle i\right|\left.\alpha_j\right\rangle, \tag{17.176}$$

而

$$\Lambda=\sum_l\alpha_l\left|l\right\rangle\left\langle l\right|, \Lambda_{ij}=\left\langle i\right|\Lambda\left|j\right\rangle=\alpha_i\delta_{ij}, \tag{17.177}$$

则

$$\begin{aligned}S\Lambda S^{\dagger}&=\sum_l\left|\alpha_l\right\rangle\left\langle l\right|\sum_k\alpha_k\left|k\right\rangle\left\langle k\right|\sum_m\left|m\right\rangle\left\langle\alpha_m\right|\\&=\sum_{l=1}^{n}\alpha_l\left|\alpha_l\right\rangle\left\langle\alpha_l\right|=\hat{A}.\end{aligned} \tag{17.178}$$

我们看到原始矩阵 A 被表示为

$$A=S\Lambda S^{\dagger}, \tag{17.179}$$

其中 Λ 是一个对角矩阵，S 是一个本征矢量排成列得到的矩阵，并且 $SS^{\dagger}=I=S^{\dagger}S$。

那为什么找到一组算符对应着的完备的正交的基矢集合如此地重要呢？因为对于给定的问题场景，这些算符 $\hat{A},\hat{B},\hat{C}$ 都是适合这个问题场景的，而不是随便选取的或者通用的。我们前面已经提到，在这样找出来的完备的正交的基矢集合中，$\hat{A},\hat{B},\hat{C}$ 这些对易算符中的每一个都简单地成了数。很

好。其他的算符呢，也会在这个完备的正交的基矢集合中表示出来。适合问题场景的基矢可以使得对问题的表述和解决变得简单很多很多。这就是找到一组算符用来构建完备的正交的基矢集合的意义。

回顾上一个提到了“本征”函数、本征状态的地方是拉普拉斯算符等导数算符的本征函数。是不是其实那里的本征函数和这里的本征矢量有关系呢？我们也说了，在那里，拉普拉斯算符等导数算符的本征函数的意义也是所得到的本征函数是适合特定的方程和边界条件的。例如，固定边界条件就对应着正弦函数，周期边界条件就对应着正弦和余弦函数的组合，长方形就用拉普拉斯算符等导数算符直角坐标表示，圆形就用拉普拉斯算符等导数算符的极坐标表示。这也就是找到一组本征函数用来构建完备的正交的基矢集合的意义。未来我们会再一次回到这个问题，真的把拉普拉斯算符等导数算符的本征函数和这里的矢量空间上的线性算符的本征矢量联系起来。

下面，在本小节的最后，我们来通过一个例子展示如何来做一个算符的谱分解。

例 17.13. *[X 表象] 按照之前定义的 $X = |1\rangle\langle -1| + |-1\rangle\langle 1|$，$Z = |1\rangle\langle 1| - |-1\rangle\langle -1|$ 和 $I = |1\rangle\langle 1| + |-1\rangle\langle -1|$，以及下面补充的 $Y = -i|1\rangle\langle -1| + i|-1\rangle\langle 1|$，计算出来 X 的本征值和本征矢量，求出来其他几个算符在这个表象下的投影分量。*

用这里定义中的 $|1\rangle, |-1\rangle$（其实可以检验它们是 Z 表象下的基矢）当作基矢，X, Y, Z, I 的投影分量形式为，

$$X = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \tag{17.180}$$

$$Y = \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix}, \tag{17.181}$$

$$Z = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}, \tag{17.182}$$

$$I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}. \tag{17.183}$$

首先，我们来求 X 的本征值。

$$\left(X-\lambda\hat{I}\right)|\lambda\rangle=0, \tag{17.184}$$

也就是

$$\begin{bmatrix} 0-\lambda & 1 \\ 1 & -\lambda \end{bmatrix}\begin{bmatrix} u_1 \\ u_2 \end{bmatrix}=0, \tag{17.185}$$

也就是

$$-\lambda u_1+u_2=0, \tag{17.186}$$

$$u_1-\lambda u_2=0. \tag{17.187}$$

这两个方程必须能够相互削去才有非零解，也就是

$$u_2=\lambda u_1, \tag{17.188}$$

$$u_2=\frac{1}{\lambda}u_1. \tag{17.189}$$

必须完全相同，于是

$$\lambda=\frac{1}{\lambda}. \tag{17.190}$$

这就要求

$$\lambda_1=1,\lambda_2=-1. \tag{17.191}$$

这就得到了 X 的本征值。

接着，我们再来求 X 的本征矢量。当 $\lambda=1$ 的时候，

$$u_2=u_1, \tag{17.192}$$

因此，本征矢量可以写成

$$|+\rangle=\begin{bmatrix} 1 \\ 1 \end{bmatrix}. \tag{17.193}$$

当 $\lambda=1$ 的时候，

$$u_2=-u_1, \tag{17.194}$$

因此，本征矢量可以写成

$$|-\rangle=\begin{bmatrix} 1 \\ -1 \end{bmatrix}. \tag{17.195}$$

可以检验这两个本征矢量确实正交，但是不归一。为了得到正交归一基矢，我们做一下归一化，得到

$$|+\rangle = \frac{\sqrt{2}}{2}\begin{bmatrix} 1 \\ 1 \end{bmatrix}, \tag{17.196}$$

$$|-\rangle = \frac{\sqrt{2}}{2}\begin{bmatrix} 1 \\ -1 \end{bmatrix}. \tag{17.197}$$

下面我们来做基矢变换。第一个方式是用抽象矢量记号：用新的基矢 $|\pm\rangle$ 来表达之前的基矢 $|\pm 1\rangle$。也就是

$$|+\rangle = \frac{\sqrt{2}}{2}\left[|1\rangle + |-1\rangle\right], \tag{17.198}$$

$$|-\rangle = \frac{\sqrt{2}}{2}\left[|1\rangle - |-1\rangle\right]. \tag{17.199}$$

反解开这个方程，得到

$$|1\rangle = \frac{\sqrt{2}}{2}\left[|+\rangle + |-\rangle\right], \tag{17.200}$$

$$|-1\rangle = \frac{\sqrt{2}}{2}\left[|+\rangle - |-\rangle\right]. \tag{17.201}$$

然后把这个表达式带入到要变换的算符中。例如

$$\begin{aligned} Z &= |1\rangle\langle 1| - |-1\rangle\langle -1| \\ &= \frac{1}{2}\left[|+\rangle + |-\rangle\right]\left[\langle +| + \langle -|\right] - \frac{1}{2}\left[|+\rangle - |-\rangle\right]\left[\langle +| - \langle -|\right] \\ &= \frac{1}{2}\left[|+\rangle\langle +| + |+\rangle\langle -| + |-\rangle\langle +| + |-\rangle\langle -|\right] \\ &\quad -\frac{1}{2}\left[|+\rangle\langle +| - |+\rangle\langle -| - |-\rangle\langle +| + |-\rangle\langle -|\right] \\ &= |+\rangle\langle -| + |-\rangle\langle +|. \end{aligned} \tag{17.202}$$

第二种方法是用之前学习的基矢变换，构建 S 矩阵。按照公式 (17.124)，我们只需要把新的本征矢量排成一列一列，也就是

$$S = \begin{bmatrix} \langle 1|+\rangle & \langle 1|-\rangle \\ \langle -1|+\rangle & \langle -1|-\rangle \end{bmatrix} = \frac{\sqrt{2}}{2}\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \tag{17.203}$$

然后，算符 A 的投影分量会被变换为 $A' = S^\dagger AS$，于是

$$Z' = S^\dagger ZS = \frac{\sqrt{2}}{2}\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}\frac{\sqrt{2}}{2}\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}. \tag{17.204}$$

坐标变换的方法和抽象矢量记号的方法得到的算符 Z 完全相符。

一般地，我们可以求解出来本征值，再得到本征矢量，然后把本征矢量的分量排成一个列矢量得到坐标变换矩阵 S，再用 S 去变换其他矢量和矩阵。如果不用坐标变换而用抽象矢量的记号，则需要把用新的本征矢量把之前所用的基矢量表达出来，然后，把这个表达式代入到矢量和算符中就能得到这些矢量和算符如何表示为新的本征矢量的形式。

习题 17.10 (Y 表象)**.** 按照之前定义的 X, Y, Z, I，计算出来 Y 的本征值和本征矢量，求出来其他几个算符在这个表象下的投影分量。

17.7 矩阵的 SVD 分解

对于厄米算符或者说厄米矩阵（我们不再区分矩阵和算符），上面的谱分解可以把它变成一个对角的矩阵，成了 n 个数，或者更准确地说是公式 (17.170) 的形式，而不再是 n^2 个数的矩阵。这大大简化了对矩阵的讨论。对于一般的矩阵，我们能做到类似的分解吗？

矩阵的奇异值分解（Singular Value Decomposition，SVD）解决了这个问题。对于一个任意的矩阵 A，定义 $H = A^\dagger A$，则 $H^\dagger = A^\dagger A = H$，是一个厄米矩阵。并且，我们有下面的性质——$H$ 的半正定性，也就是：对于任意的矢量 a，

$$\langle a|\, H\, |a\rangle = \langle a|\, A^\dagger A\, |a\rangle = \left(\langle a|\, A^\dagger\right)\left(A\, |a\rangle\right) = \left(\langle aA|\right)\left(|Aa\rangle\right) \geqslant 0. \tag{17.205}$$

最后一步我们用到了内积的性质，$(v, v) \geqslant 0$，倒数第二步我们用到了伴算符的定义 $\langle a|\, A^\dagger = \langle aA|$。厄米矩阵 H 可以做谱分解，可以得到一个

$$H = V\Sigma^2 V^\dagger = \sum_j \sigma_j^2\, |v_j\rangle\langle v_j|\,. \tag{17.206}$$

其中的本征值被记作了 σ_j^2，这是因为对于一个半正定的厄米矩阵，其对角化以后的矩阵的对角元素必定大于等于零。否则，只需要选 a 为对角元素为负的本征值对应的本征矢量，则肯定就会得到 $\langle a| H |a\rangle < 0$，矛盾。既然所有的 σ_j^2 都大于等于零，是实数，我们把它们按照降序排列，同时按照这个顺序我们就得到了相应的 $|v_j\rangle$。

同理，我们可以定义 $G = AA^\dagger$，则 $G^\dagger = AA^\dagger = G$，是一个厄米矩阵。我们对 G 做谱分解，可以得到一个

$$G = U\tilde{\Sigma}^2 U^\dagger \tag{17.207}$$

只要我们证明得到的本征值刚好和之前矩阵 H 的相同 $\tilde{\Sigma}^2 = \Sigma^2$，也按照降序排列，我们就得到了一系列的 $|u_j\rangle$。这个时候，我们可以构建一个矩阵，未来会证明正好就是原始的 A 矩阵，

$$A = U\Sigma V^\dagger = \sum_j \sigma_j |u_j\rangle\langle v_j| . \tag{17.208}$$

其中矩阵 $\Sigma = \sum_j \sigma_j |j\rangle\langle j|$，而 $\sigma = \sqrt{\sigma^2}$，也就是仅仅取其中正的平方根。这个分解被称为矩阵的 SVD 分解。很容易验证，$A = U\Sigma V^\dagger$ 代入到 H, G 其实可以得到 $H = V\Sigma^2 V^\dagger$ 以及 $G = U\Sigma^2 U^\dagger$。因此，这看起来真的非常可能是对的。

于是，我们从厄米矩阵的谱分解受到了启发可以去做一般矩阵的 SVD 分解。下面，我们来正式证明这个分解成立。首先，根据构建出来的 H 矩阵的本征值和本征矢量，我们定义

$$|u_j\rangle = \frac{1}{\sigma_j} A |v_j\rangle \Longleftrightarrow A |v_j\rangle = \sigma_j |u_j\rangle . \tag{17.209}$$

于是，

$$A^\dagger |u_j\rangle = \frac{1}{\sigma_j}\sigma_j^2 |v_j\rangle = \sigma_j |v_j\rangle . \tag{17.210}$$

我们接着来证明 $|v_j\rangle, |u_j\rangle$ 都是一组正交归一基矢。

$$\langle v_i |v_j\rangle = \delta_{ij}, I = \sum_j |v_j\rangle\langle v_i| , \tag{17.211}$$

和

$$\langle u_i \left| u_j \right\rangle = \delta_{ij}, I = \sum_j \left| u_j \right\rangle \langle u_i \left| \right. . \tag{17.212}$$

前者可以依据谱分解定理，根据 $|v_j\rangle$ 是厄米矩阵 H 的本征矢量而得到。后者，我们这样来证明，

$$\begin{aligned} \langle u_i \left| u_j \right\rangle &= \frac{1}{\sigma_i} \langle v_i A \left| \frac{1}{\sigma_j} A \right| v_j \rangle \\ = \frac{1}{\sigma_i \sigma_j} \langle v_i \left| A^\dagger A \right| v_j \rangle &= \frac{1}{\sigma_i \sigma_j} \langle v_i \left| H \right| v_j \rangle \\ &= \frac{1}{\sigma_i \sigma_j} \langle v_i \left| \sigma_j^2 \right| v_j \rangle \\ &= \delta_{ij}. \end{aligned} \tag{17.213}$$

这样 $|u_j\rangle$ 相互正交，数量和已知的一组正交归一基矢 $|v_j\rangle$ 一样多，因此，$|u_j\rangle$ 也是一组正交归一基矢。

既然 $|v_j\rangle$ 和 $|u_j\rangle$ 都是一组正交归一基矢就好办了，我们对矩阵 A 做如下的计算，

$$A \sum_j \left| v_j \right\rangle \langle v_j \left| \right. = \sum_j \sigma_j \left| u_j \right\rangle \langle v_j \left| \right. , \tag{17.214}$$

$$A^\dagger \sum_j \left| u_j \right\rangle \langle u_j \left| \right. = \sum_j \sigma_j \left| v_j \right\rangle \langle u_j \left| \right. . \tag{17.215}$$

于是，我们证明了矩阵 A 的 SVD 分解成立。

算符的 SVD 分解和算符的谱分解相比，最主要的差别，这里的左矢量 $\langle v_j|$ 和右矢量 $|u_j\rangle$ 不完全就是把右矢量用内积变成左矢量的关系，而是两个不同的矢量，而谱分解中它们必然是 $\langle \alpha_j|$ 和右矢量 $|\alpha_j\rangle$。当然，这里的类似于本征值的部分成了奇异值，约定都取大于等于零，而谱分解中的本征值可以取所有的实数。两者都实现了把算符或者说矩阵变成数，或者说变成对角项的叠加的形式的要求。

算符或者说矩阵的 SVD 分解有一个很好的性质：任意矩阵的前若干项提供了这个矩阵的一个近似，也就是

$$\hat{A}_L = \sum_{j=1}^{L} \sigma_j \left| u_j \right\rangle \langle v_j \left| \right. \tag{17.216}$$

在 $L=n$ 的时候，就是原始矩阵 A，但是，在 $L\to n$ 的时候，L 越大，$\hat{A}_L$ 和 A 之间的某种距离越小。具体的距离我们稍后定义。但是，直觉上，这很好理解："$|u_j\rangle\langle v_j|$" 包含的都是正交归一的矢量，所以大小什么的肯定都相同，所有的具有大小含义的地方都被包含在 σ_j 里面，因为我们已经对 σ_j 按照降序排列，因此，随着 L 越大，剩下的没有包含进来的 σ_j 肯定越小。SVD 分解的这个性质未来会被用于很多问题的解决。

矩阵的谱分解和 SVD 分解可以用来解决各种各样的问题。下面，我们来看看它们的应用场景，以及其他的线性代数的概念的典型应用场景。任何对象，只要被证明符合矢量和算符的定义，我们这一章所学习的概念和定理就都能用在这些对象身上。

17.8 算符当作矢量和算符内积

前面我们提到了包含有限个奇异值和奇异值矢量的 $\hat{A}_L$ 和原始算符 A 之间的距离，但是，我们对两个算符之间的距离并没有定义。这一节，我们来试试把算符看作是矢量，然后试试把矢量的大小，也就是矢量之间的距离，推广到算符之间的距离。

首先，我们来证明，矢量空间上的线性算符构成矢量空间。我们定义两个算符的加法得到这样的算符：算符 $A+B$ 的定义是，对于任意的 $a\in V$，$(A+B)\,a=Aa+Ba$。算符的数乘定义为：对于任意的 $\mu\in F$，$(\mu A)\,a=\mu\,(Aa)$。我们接着来证明这样定义的加法和数乘满足矢量空间的要求。我们选择来证明其中的双向分配律：$(\mu+\nu)\,(A+B)=\mu A+\mu B+\nu A+\nu B$。我们证明的方式是证明对于任何一个矢量 a，左边的算符和右边的算符的作用结果相同。左侧作用于 a 得到，

$$\begin{aligned}&[(\mu+\nu)\,(A+B)]\,a=(\mu+\nu)\,[(A+B)\,a]\\&=(\mu+\nu)\,(Aa+Ba)=\mu Aa+\mu Ba+\nu Aa+\nu Ba,\end{aligned}\tag{17.217}$$

最后一步我们用了矢量的双向分配律。右侧作用于 a 得到

$$(\mu A+\mu B+\nu A+\nu B)\,a=\mu Aa+\mu Ba+\nu Aa+\nu Ba.\tag{17.218}$$

因此，两者完全相同。

然后，我们来定义算符的大小，或者说算符的内积，

$$(A,B)=tr\left(A^{\dagger}B\right), \tag{17.219}$$

其中 $tr(\cdot)$ 称作求一个矩阵的迹，定义如下：

$$tr(A)=\sum_j\langle v_j|A|v_j\rangle. \tag{17.220}$$

这里的 $|v_j\rangle$ 可以任意选一组正交归一基矢。结果不依赖于这个选择。例如，如果我们选了另一组，则

$$\sum_j\langle u_j|A|u_j\rangle=\sum_{jk}\langle u_j|v_k\rangle\langle v_k|A|u_j\rangle$$

$$=\sum_k\langle v_k|A\left(\sum_j|u_j\rangle\langle u_j|\right)|v_k\rangle=\sum_k\langle v_k|A|v_k\rangle. \tag{17.221}$$

我们可以证明这个定义满足内积的要求。例如 $(A,A)\geqslant 0$ 可以这样来证明：

$$(A,A)=tr\left(A^{\dagger}A\right)=\sum_j\langle v_j|A^{\dagger}A|v_j\rangle=\sum_j\langle v_jA|Av_j\rangle\geqslant 0. \tag{17.222}$$

最后一步我们用到了矢量内积的性质 $(a,a)\geqslant 0$。内积定义的其他要求也很容易得到。线性性我们就不证明了。我们来看看共轭对称性，也就是

$$(B,A)=(A,B)^*. \tag{17.223}$$

也就是要求

$$\sum_j\langle v_j|A^{\dagger}B|v_j\rangle=\left(\sum_j\langle v_j|B^{\dagger}A|v_j\rangle\right)^*. \tag{17.224}$$

而这个确实成立，

$$\sum_j\langle v_j|A^{\dagger}B|v_j\rangle=\sum_j\langle v_jA|Bv_j\rangle \tag{17.225}$$

$$=\left(\sum_j\langle v_jB|Av_j\rangle\right)^*=\left(\sum_j\langle v_j|B^{\dagger}A|v_j\rangle\right)^*. \tag{17.226}$$

最后一步我们用到了伴算符的定义，倒数第二步用到了矢量内积的共轭对称性 $(b,a)=(a,b)^*$。

这样我们就可以把矢量空间上的线性算符看成矢量空间了。如果我们要讨论两个算符 A,B 之间的距离，我们只需要计算 $\sqrt{tr\left[(A-B)^\dagger(A-B)\right]}$。例如，

$$\begin{aligned}
&tr\left[\left(A-\hat{A}_L\right)^\dagger\left(A-\hat{A}_L\right)\right]\\
=&tr\left[\sum_{j=L+1}^{n}\sigma_j|v_j\rangle\langle u_j|\sum_{k=L+1}^{n}\sigma_k|u_k\rangle\langle v_k|\right]\\
=&tr\left[\sum_{j,k=L+1}^{n}\sigma_j|v_j\rangle\sigma_k\delta_{jk}\langle v_k|\right]\\
=&tr\left[\sum_{j=L+1}^{n}\sigma_j^2|v_j\rangle\langle v_j|\right]\\
=&\sum_l\langle v_l|\left[\sum_{j=L+1}^{n}\sigma_j^2|v_j\rangle\langle v_j|\right]|v_l\rangle\\
=&\sum_{j=L+1}^{n}\sigma_j^2.
\end{aligned}\tag{17.227}$$

因此，

$$\left|A-\hat{A}_L\right|=\sqrt{\sum_{j=L+1}^{n}\sigma_j^2}\leqslant\sqrt{(n-L)}\sigma_{L+1}.\tag{17.228}$$

用 $\epsilon-N$ 语言，只要 $L>N=\max\left(n-\frac{\epsilon}{\sigma_{L+1}},0\right)$，我们就可以满足上面的条件。$\sigma_j$ 是一堆已经给定的数，我们总是可以把 $L=1,2,\cdots,n$ 的每一个代入到这个条件中取试试，直到得到满足条件的 L。并且，显然 $L=n$ 或者说 $N=n-1$ 满足这个条件——这个时候 $(n-L)=0$。

我们看到有了距离的概念，我们就可以把数列的极限用到这里来。而距离的概念，在这里，是借助于矢量的内积而来的。更进一步，矢量的内积在这里其实是算符的内积，通过把算法看作矢量而来的。这就是矢量空间和内积的抽象定义的好处：任何对象只要符合这些定义，就可以运用相应的定理，而不需要局限于之前的欧氏空间的矢量或者矢量的分量表示。

在这里，我们也初步看到了代数（矢量、内积）和分析（距离、极限）

是相互融合的。实际上，我们整个矢量空间就可以看作在给一些抽象的对象找一堆数来表示，然后定义这些对象之间的加法和数乘，以及距离。有了这些数的表示，还有了加法数乘和距离，我们就可以讨论这些对象的极限和微小变换，而这正好就是分析学发挥为例的地方。

在这里你也看到了，线性性是最根本的性质：有了线性性就有了自然的加法和数乘的定义，就可以检验是否满足矢量空间的定义。这也启发我们，未来任何一个具有线性性的对象，都可以尝试用矢量来描述。不过，还剩下一个概念上特别神奇的问题：线性性看起来似乎是一个包含内容非常具体也就是外延非常小的集合，为什么一个这么特殊的具体的小的集合，会发挥这么大的作用呢？大量的不是线性的东西多了去了啊！

我们稍后等到见识了矢量空间发挥作用的例子之后再来回答这个问题。不过，可以先透露答案：这个线性性的定义足够具有一般性，很多从一个角度来看非线性的东西，其实在更高层次上可以看作是线性的。也就是说，不是 $y=kx+b$ 这样的叫做线性性而一旦出现 $y=x^2$ 就不满足线性性，例如，$\frac{d^n}{dx^n}$ 和 $\frac{d^n}{dx^n}+f(x)$ 也可以看作是具有线性性的，就算这里的 $f(x)=x^2$ 或者更一般的函数。

17.9 最大本征值对应的本征矢量和幂方法

有一个非常简单的求出来厄米矩阵的最大本征值（绝对值意义上）对应的本征矢量的方法——幂方法。由于 SVD 分解中矩阵 A 和矩阵 $H=A^\dagger A$ 的关系，这个方法也是求出来 SVD 分解的最大奇异值对应着的奇异值矢量的方法。

对于给定的厄米矩阵 A 和随机选取的一个矢量 $\left|v^{(0)}\right\rangle$，我们做如下的计算

$$\left|v^{(n)}\right\rangle=\frac{A\left|v^{(n-1)}\right\rangle}{\sqrt{\left\langle v^{(n-1)}\right|A^\dagger A\left|v^{(n-1)}\right\rangle}},\tag{17.229}$$

在 $n\to\infty$ 的极限下，在运气不那么差（一会儿会看到这是什么意思）的情况下，这个 $\left|v^{(n)}\right\rangle$ 等于 A 的最大本征值对应的本征矢量。

我们用厄米矩阵 A 的本征矢量来展开表示上面的迭代过程。我们分开

这样两个步骤来看：

$$\left|v^{(0)}\right\rangle \to A\left|v^{(0)}\right\rangle, \tag{17.230}$$

$$A\left|v^{(0)}\right\rangle \to \frac{A\left|v^{(0)}\right\rangle}{\sqrt{\left\langle v^{(0)}\right| A^{\dagger}A\left|v^{(0)}\right\rangle}} = \left|v^{(1)}\right\rangle. \tag{17.231}$$

对于第一步，

$$A\left|v^{(0)}\right\rangle = A\sum_j |\alpha_j\rangle\langle\alpha_j\left|v^{(0)}\right\rangle = \sum_j \alpha_j |\alpha_j\rangle\langle\alpha_j\left|v^{(0)}\right\rangle. \tag{17.232}$$

也就是说，第 j 个本征矢量的系数会被放大 α_j 倍，

$$\langle\alpha_j\left|v^{(0)}\right\rangle \to \alpha_j\langle\alpha_j\left|v^{(0)}\right\rangle. \tag{17.233}$$

到了第二步归一化的时候，自然第一个矢量所占的比例会增加，因为第一个矢量是本征值的绝对值最大的那个矢量。每一次迭代都是这个效果，因此，只要满足下面两个条件：第一，$\langle\alpha_1\left|v^{(0)}\right\rangle \neq 0$ 和第二 $\alpha_1 \neq \alpha_2$，这个矢量中 $|\alpha_1\rangle$ 所占的比例会越来越高。我们迭代到满足再一次迭代得到的矢量和前一次迭代得到的矢量相差足够小，就可以停止迭代，把得到的矢量当作 $|\alpha_1\rangle$。

原则上，我们可以从整个空间中去掉这个 $|\alpha_1\rangle$，来得到第二大本征值所对应的本征矢量。例如，随机取一个初始矢量，选从中把这个随机矢量的 $|\alpha_1\rangle$ 成份去掉，也就是计算，

$$\left|v^{(0)}\right\rangle \to \left|v^{(0)}\right\rangle - \langle\alpha_1\left|v^{(0)}\right\rangle|\alpha_1\rangle. \tag{17.234}$$

然后，做之前的迭代相乘，为了保险还可以每一次迭代得到的新矢量都做一个上面这个去掉 $|\alpha_1\rangle$ 成份的过程。这样我们就按照顺序可以得到各个矢量了。当然，实际上，问题比这复杂很多：只要我们没有真正把 $|\alpha_1\rangle$ 算准确，每次都会留下一个 $|\alpha_1\rangle$ 的小小的分量，这个分量每次迭代都会被放大最大的倍数，于是，每次这样的迭代实际上很可能算出来的都是接近 $|\alpha_1\rangle$ 的矢量。因此，真正的矩阵本征矢量的计算需要找更好的方法。不过，幂方法至少给了一个很好的启发，很多更复杂的矩阵谱分解和 SVD 分解的数值计算算法，例如 Krylov 子空间方法 [27]，都是这个幂方法的变形。

数值线性代数是一个非常有意思的领域。实际上，当前的大量的数值计算都依赖于数值线性代数。数值线性代数领域的计算方法的进步可以大大推动一般的数值计算任务。例如，最近矩阵乘法的计算效率有了一个看起来很小的提高，但是未来将会带来很大的效率的提升。

顺便，矩阵乘法的数值计算也是一个很有意思的问题。对于一个两个各自有 $n \times n$ 个元素的矩阵的乘法，我们简单地想一想也就知道要算 n^3 个乘法以及更多的加法才能得到矩阵乘法的结果。加法相对于乘法来说在计算机里面的计算成本低很多，我们忽略不计。那么，我们就问是不是有一个利用更多的加法的步骤，但是更少的乘法的步骤来完成矩阵乘法的计算方式呢？初看起来这是一个天方夜谭，不可能啊！但是Volker Strassen（沃尔克 · 施特拉森）发现了一个不需要计算 n^3 次乘法的方法，而是只需要 $n^{\log_2 7} \approx n^{2.807}$ 次乘法的方法 [28]。从那以后，数值线性代数有了属于自己的领域和时代。最近依靠人工智能大模型，人们找到了计算效率更高的矩阵乘法的计算方式 [29]。

小知识 17.2 (Strassen 算法). 矩阵乘法的 *Strassen* 算法是人类巧妙创造的典范，但是也是无脑纯技巧创造的典范，特别适合人工智能来做，其背后本质上就是不断地试错和对更高效率的计算方法的追求。但是，现代人在受到了 *Strassen* 算法的启发之后，也找到了一些来通过有目的性和有方向性的思考来找到 *Strassen* 算法的方式[8]在这里，我们仅仅当作手工艺品展示一下 *Strassen* 算法。

对于两个 2×2 矩阵 A, B，

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}, B = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix}, \tag{17.235}$$

其乘积可以这样计算，

$$C = \begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix} = \begin{bmatrix} A_{11}B_{11} + A_{12}B_{21} & A_{11}B_{12} + A_{12}B_{21} \\ A_{21}B_{11} + A_{22}B_{21} & A_{21}B_{12} + A_{22}B_{22} \end{bmatrix}. \tag{17.236}$$

这样就需要 8 次乘法计算和 4 次加法计算。

[8] 见https://ilyi.net/page/deriving-stressen-matrix-multiplication-formulas/，和https://softwareengineering.stackexchange.com/questions/199627/how-did-strassen-come-up-with-his-matrix-multiplication-method。2025 年 11 月 28 日访问。

*Strassen*说你可以这样计算，

$$C = \begin{bmatrix} P_5 + P_4 - P_2 + P_6 & P_1 + P_2 \\ P_3 + P_4 & P_1 + P_5 - P_3 - P_7 \end{bmatrix}, \tag{17.237}$$

其中

$$\begin{aligned} P_1 = A_{11}\left(B_{12} - B_{22}\right), P_2 = \left(A_{11} + A_{12}\right) B_{22}, \\ P_3 = \left(A_{21} + A_{22}\right) B_{11}, P_4 = A_{22}\left(B_{21} - B_{11}\right), \\ P_5 = \left(A_{11} + A_{22}\right)\left(B_{11} + B_{22}\right), P_6 = \left(A_{12} - A_{22}\right)\left(B_{21} + B_{22}\right), \\ P_7 = \left(A_{11} - A_{21}\right)\left(B_{11} + B_{12}\right). \end{aligned} \tag{17.238}$$

这样，我们只需要计算 7 次乘法和 18 次加法。由于乘法的计算成本高很多，我们这里只关注乘法。为什么这样组合起来算会有这个好处呢？实际上是因为有分配律。例如 $P_6 = \left(A_{12} - A_{22}\right)\left(B_{21} + B_{22}\right) = A_{12}B_{21} + A_{12}B_{22} - A_{22}B_{21} - A_{22}B_{22}$ 实际上计算了将来会出现在最终计算结果中的四项。只不过，这里这四项之间直接给做了加法，混在一起了。因此，实际上，*Strassen*算法就是想办法再把这些混在一起的项区分出来，或者说直接用这些混在一起的项组合出来所有的 $C_{11}, C_{12}, C_{21}, C_{22}$。

更进一步，我们还可以把这个过程迭代进行——利用分治的思想（*Divide and Conquer*）。我们可以把 A_{11} 这些看作是子矩阵，也就是把任何一个矩阵先分割为四个子矩阵，然后相比于通常计算方式的时候要用到 8 次这些子矩阵的乘法，我们只需要用到 7 次这些子矩阵的乘法。接着，对于每一个子矩阵乘法，例如 $A_{11}\left(B_{12} - B_{22}\right)$，我们把 A_{11} 和 $\left(B_{12} - B_{22}\right)$ 进一步各自划分成四个子矩阵，来完成子矩阵乘法的计算。接着，再次划分成子矩阵，直到子矩阵只剩下一个元素。这样，我们只需要算一算 $7^{\log_2 n} = n^{\log_2 7}$ 次乘法就够了。

当然，实际上，由于 *CPU* 的处理能力的问题，实际操作中不见得需要一直划分到单个元素。实际上，现在已有不少数值线性代数计算包，例如

BLAS[9]、*ATLAS*[10]、*Intel* 数学核[11]和 *AMD* 数学核[12]做了基于 *Strassen* 和具体 *CPU* 计算能力的优化。

还会不会有更巧妙的组合得到更高的计算效率呢？到底这个巧妙组合的方式是不是可以还原成一种思考模式，甚至迁移到其他问题之中呢？这就是现代计算数学家在思考的问题了。

17.10 线性方程的解

一个包含 m 个等式的 n 元一次方程组，

$$\sum_{j=1}^{n} A_{ij}x_j = b_i, \tag{17.239}$$

可以写成矩阵乘上矢量的形式，

$$Ax = b. \tag{17.240}$$

其中 $A = (A_{ij})_{m\times n}$ 是一个 m 行 n 列矩阵，$x = [x_1, x_2, \cdots, x_n]^T$，$b = [b_1, b_2, \cdots, b_n]^T$。注意，当 m 个等式中去掉那些可以通过其他的等式组合而来的之后剩下的等式——也就是经过对行的数乘和加法操作使得其中某一行为零——的数量才是真正独立的方程的数量。我们这里假设已经去掉了那样的等式。我们进一步假设没有遇到某一行的整个左边等于零但是右边不等于零的情况。如果出现这个情况则这个方程没有解。在此基础上，如果 $m < n$ 则方程称为未定方程，有好多解，因为方程数量小于未知数的数量，总是有些未知数可以任意取值；如果 $m > n$ 则方程无解，方程的数量比未知数要多，肯定有某些方程不能被满足。因此，最终线性方程都可以化成方阵的形式，也就是 $m = n$。

这个时候，如果矩阵 A 的逆存在，记作 A^{-1}，也就是

$$AA^{-1} = I = A^{-1}A, \tag{17.241}$$

[9]https://www.netlib.org/blas/，2025 年 11 月 28 日访问。

[10]https://math-atlas.sourceforge.net/，2025 年 11 月 28 日访问。

[11]https://www.intel.com/content/www/us/en/developer/tools/oneapi/onemkl.html，2025 年 11 月 28 日访问

[12]https://www.amd.com/en/developer/aocl/libm.html，2025 年 11 月 28 日访问。

那么

$$x = A^{-1}b. \tag{17.242}$$

也就是我们只需要求出来矩阵 A 的逆就可以。但是，求一个矩阵的逆是一个比较复杂的过程，例如高斯消去法 [27]。我们来介绍另一种求解线性方程的方法 [30]。

比如说，我们把原始的线性方程公式 (17.240) 变形为

$$x = (I - A)\, x + b. \tag{17.243}$$

然后，引入迭代求解方式，

$$x_{n+1} = (I - A)\, x_n + b. \tag{17.244}$$

如果我们证明

$$\lim_{n\to\infty} x_n = x_*, \tag{17.245}$$

或者

$$\lim_{n\to\infty} (r_n = b - Ax_n) = 0, \tag{17.246}$$

或者

$$\lim_{n\to\infty} (x_{n+L} - x_n) = 0, \forall L \in \mathbb{Z}^+, \tag{17.247}$$

我们就相当于求解出来了原始的方程。当然，这样的迭代求解必须在有限步终止，只要满足 r_n 的最大值或者 $|r_n|$ 小于某一个满足要求的值 ϵ。第一个很难证明，因为我们不知道 x_*。后面两个的证明可能简单很多。我们来试试。

根据公式 (17.244)，我们有

$$r_n = b - Ax_n = x_{n+1} - x_n. \tag{17.248}$$

这表示 r_n 不仅仅代表了解离真实解的距离 $A\,(A^{-1}b - x_n)$，还代表了两次迭代得到的解前后的差 $x_{n+1} - x_n$。那我们来看看这个 r_n 的迭代行为，如果这个越来越小，就好办了。

$$\begin{aligned} Ar_n = Ax_{n+1} - Ax_n = Ax_{n+1} + x_{n+1} - x_n - b = -r_{n+1} + r_n \\ \Rightarrow r_{n+1} = (1 - A)\, r_n. \end{aligned} \tag{17.249}$$

但是，这个迭代公式不一定能够保证 r_n 越来越小，除非 $(1-A)$ 的所有本征值的绝对值小于 1。

实际上，对于一个更一般的 $B=(1-D^{-1}A)$，其中 D 是一个对角矩阵，原始的线性方程公式 (17.240) 变形为

$$x = Bx + D^{-1}b. \tag{17.250}$$

进而得到新的迭代求解方式，

$$x_{n+1} = Bx_n + D^{-1}b. \tag{17.251}$$

于是，我们完全可以得到

$$r_{n+1} = Br_n. \tag{17.252}$$

这个时候，只要 B 的所有本征值的绝对值都小于 1，就能保证 r_n 在不断地减小。于是，这提醒我们甚至就可以选为 $D=dI$，其中 d 为一个比 A 的绝对值最大的本征值还要大一点的数，我们就能保证所有 $B=(1-D^{-1}A)$ 本征值为

$$1-\frac{\alpha_j}{\max\{\alpha_j\}+\delta}<1. \tag{17.253}$$

当然，实际选的时候，我们也不知道 A 的所有的本征值。一方面，我们可以先用幂方法得到 A 的绝对值最大本征值，另一方面，没准选 D 为 A 的对角矩阵就行，如果 A 的对角元比非对角元要大的话。此外，前面直接选 $B=I-A$ 的方式，后来发展成了一个叫做 Krylov 子空间方法的求解线性方程组的方法 [27]。

这个把矩阵和矢量的乘法用于求解线性方程的例子的目的，除了提到线性方程组的几个解法之外，主要是说明线性方程组的求解实际上靠的还是矩阵、矩阵和矢量的乘法、矩阵本征值这些概念，而不是行列式——那是大多数线性代数教材的重要内容甚至起点 [31–33]。本书除了这里从负面的角度提到行列式之外，只在一个地方——17.11.3 节——提到了行列式。矩阵、矢量、算符，这些才应该是线性代数的学习内容，而不是行列式——那只是一个从矩阵得到数的映射的计算问题！和本书在这一点上理念相同的书还有Sheldon Axler（谢尔顿 · 阿克斯勒）的《Linear Algebra Done Right（线性代数这样学）》[34]。

17.11 线性代数的典型应用场景

在这一节中，我们来呈现线性代数的典型应用场景。其中一部分是我们在启发矢量和群的定义的时候就已经遇到的场景，一部分是前面所学习到的新的知识——例如矩阵的谱分解和 SVD 分解——的应用场景。在学习本节的时候，一方面我们是为了看到这些典型应用场景，未来可以迁移使用，另一方面是为了更好地理解前面学习到的学科概念知识本身。学以致用和用中学也会使得学更加有动力和方向感。当然，学以致用，学以致创造也是目标之一。

17.11.1 边值问题、函数和算符，解析解和数值解

在物理学和用物理学解决工程等实际问题中，我们会遇到大量的边值问题——给定一个事物发生的空间 Ω，例如长方形和圆形，甚至一般的空间，这个空间的边界上对于这个事物有明确的要求（顺便，这被称为这个问题的边界条件，这个要求被称为对边界值的要求，所以称为边值问题——约定了一个现象的边界值的问题）。之前在在微积分的部分的 16.9.3 节中，其实我们已经学习了一点点边值问题的处理方法，现在我们从矢量空间的角度来更好地理解和更一般地处理边值问题。

之前在 16.9.2 节，我们已经基于微积分和牛顿第二定律推导了波动方程，

$$\frac{\partial^2}{\partial t^2}u = c^2\nabla^2 u. \tag{17.254}$$

关于热扩散现象的热传导方程——我们不再推导了，可以基于关于热的物理知识和微积分来推导——为

$$\frac{\partial^2}{\partial t^2}u = k\nabla^2 u. \tag{17.255}$$

量子现象的Schrödinger 方程的形式如下，

$$i\hbar\frac{\partial}{\partial t}\psi = \left(-\frac{\hbar^2}{2m}\nabla^2 + V\right)\psi. \tag{17.256}$$

这些方程的特点是：它们都是线性方程（左右两侧的算符都满足线性性），右侧都出现了 ∇^2 这样的导数项。在物理学的研究和工程实践中，我们还会

遇到非线性的算符。很多时候，非线性的方程也可以通过线性方程来得到近似的研究，例如通过线性稳定性分析 [35, 36]。因此，线性方程的处理是后面的所有的方程的处理的基础。

我们把右侧的线性算符记作 L。我们来证明它是线性的，

$$\begin{aligned} L\left(a\psi+b\phi\right) &\propto \left(\nabla^2+V\right)\left(a\psi+b\phi\right) \\ &= \nabla^2\left(a\psi+b\phi\right)+V\left(a\psi+b\phi\right) \\ &= a\nabla^2\psi+b\nabla^2\phi+aV\psi+bV\phi \\ &= a\left(\nabla^2+V\right)\psi+b\left(\nabla^2+V\right)\phi. \end{aligned} \tag{17.257}$$

可以类似地证明左侧的算符也是线性的。于是，我们就得到了一个结论：对于上面的每一个方程，如果 ψ_1,ψ_2 是方程的解，则 $a\psi_1+b\psi_2$ 也是方程的解。

联系到矢量空间，这启发我们，是不是我们能够找到一组基本的满足方程的解来当作这个解空间的基矢，然后，其他的解都表达成这些基矢的线性组合呢?

我们已经知道，函数可以看作是矢量空间（因为函数的加法和数乘满足矢量空间的定义），那么，线性算符 L 就是这个矢量空间上的线性算符——把一个函数映射为另一个函数。比如说，如果我们找到了 L 的本征矢量，

$$L\psi=\lambda\psi, \tag{17.258}$$

上面的几个方程是不是就很容易求解了呢? 我们试试。

这时候，波动方程，热传导方程（扩散方程）和Schrödinger 方程，分别成了

$$\frac{\partial^2}{\partial t^2}\psi=\lambda\psi, \tag{17.259}$$

$$\frac{\partial^2}{\partial t^2}u=\lambda\psi, \tag{17.260}$$

$$i\hbar\frac{\partial}{\partial t}\psi=\lambda\psi. \tag{17.261}$$

在这个情况下，这些方程都非常容易求解，

$$\psi = e^{\pm\sqrt{\lambda}t}\psi\left(0\right), \tag{17.262}$$

$$\psi = e^{\pm\sqrt{\lambda}t}\psi\left(0\right), \tag{17.263}$$

$$\psi = e^{-i\frac{\lambda}{\hbar}t}\psi\left(0\right). \tag{17.264}$$

前两个实际上也往往是复数的形式，因为 $\sqrt{\lambda}$ 下的 λ 往往是一个负数。

也就是说，如果我们找到了 L 的一堆本征矢量，记作 ψ_n，满足

$$L\psi_n = L_n\psi_n, \tag{17.265}$$

我们的原始方程的一般解就搞定了，以Schrödinger 方程，其他方程也类似，有

$$\phi\left(x,t\right) = \sum_n a_n e^{-i\frac{L_n}{\hbar}t}\psi_n\left(x\right). \tag{17.266}$$

其中的 a_n 只需要通过初始条件来确定，

$$\phi\left(x,0\right) = \sum_n a_n\psi_n\left(x\right). \tag{17.267}$$

这是一个矢量的基矢展开形式，只要基矢之间是正交归一的，

$$a_n = \left(\phi\left(x,0\right),\psi_n\left(x\right)\right). \tag{17.268}$$

于是，我们唯一剩下的任务就是找到算符 L 的一堆本征矢量函数，然后证明其正交归一。

本来我们需要对具体的算符在具体的边界条件上来证明得到的本征矢量函数的正交归一性——之前在 16.9.6 节中对于三角函数 $\left\{\sin\left(\frac{n\pi x}{L}\right),\cos\left(\frac{n\pi x}{L}\right)\right\}$ 的正交归一性我们就是那样来证明的，但是，矢量空间的学习告诉了我们矩阵的谱分解定理——任意一个厄米算符，其本征矢量肯定正交归一。于是，我们只需要证明 L 算符是一个厄米算符就够了。这就是学习了矢量空间的好处。

其实，我们还不知道函数的内积是怎么做的。我们还是借助矢量空间的概念来试试找到这个函数的内积的定义。假设我们已经通过三角函数

$\left\{\sin\left(\frac{n\pi x}{L}\right), \cos\left(\frac{n\pi x}{L}\right)\right\}$ 这种 L 的本征矢量函数来表示了一个函数，一个函数成了一个列向量，

$$\phi(x) \rightarrow \begin{bmatrix} \phi_1 \\ \phi_2 \\ \cdot \\ \cdot \\ \cdot \end{bmatrix}, \psi(x) \rightarrow \begin{bmatrix} \psi_1 \\ \psi_2 \\ \cdot \\ \cdot \\ \cdot \end{bmatrix} \tag{17.269}$$

那根据通常的矢量空间的内积的定义，两个矢量的内积可以表示为

$$(\phi(x), \psi(x)) \rightarrow \begin{bmatrix} \phi_1^* & \phi_2^* & \cdots \end{bmatrix} \begin{bmatrix} \psi_1 \\ \psi_2 \\ \cdot \\ \cdot \\ \cdot \end{bmatrix}. \tag{17.270}$$

我们来试试从这个等式的右边去猜测这个等式的左边。根据三角函数这个 ∇^2 的本征函数的特例的启发，比如说，两个不同的三角函数的正交性是这样得到的，

$$\int_0^L \sqrt{\frac{2}{L}} \cos\frac{m2\pi x}{L} \sqrt{\frac{2}{L}} \cos\frac{n2\pi x}{L} dx = \delta_{m,n}. \tag{17.271}$$

这也就是相当于在求 $\begin{bmatrix} 0 & \cdots & 1_m & 0 & \cdots \end{bmatrix} \begin{bmatrix} 0 \\ \cdot \\ 1_n \\ \cdot \end{bmatrix}$ 的内积，前（后）者其他元素都是 0，只有在 m（n）分量上投影是 1。因此，两个一般函数的内积很自然地就是，

$$(\phi(x), \psi(x)) = \int_\Omega \phi^*(x) \psi(x) dx. \tag{17.272}$$

其中加一个负共轭是因为我们的取值从实数到了复数，这样一个函数和自己的内积就能保证得到大于等于零的数，Ω 表示在现象所发生的空间。有了这个猜测之后，你可以去用内积的定义来证明这个猜测确实满足内积的要求。

现在我们来证明 L 是一个厄米算符。为了简单以单变量函数为例。先来看其中的 V 项，

$$\begin{aligned}(\psi, V\phi) &= \int_\Omega dx\psi^*(x)\left(V(x)\phi(x)\right)\\ &= \int_\Omega dx\left(V^*(x)\psi(x)\right)^*\phi(x) = (V^*\psi, \phi),\end{aligned} \tag{17.273}$$

于是，只要

$$V(x) = V^*(x), \tag{17.274}$$

那 V 项就是厄米的。先来看其中的 $\nabla^2 = \frac{d}{dx}\left(\frac{d}{dx}\right)$ 项。为了简单，先看其中的 $\frac{d}{dx}$ 项，

$$\begin{aligned}\left(\psi, \frac{d}{dx}\phi\right) &= \int_\Omega dx\psi^*(x)\phi'(x) = \int_\Omega \psi^*(x)\,d\left(\phi(x)\right)\\ &= \psi^*(x)\phi(x)|_{\partial\Omega} - \int_\Omega \phi(x)\,d\left(\psi^*(x)\right)\\ &= -\int_\Omega \phi(x)\psi^{*,\prime}(x)\,dx = \left(-\frac{d}{dx}\psi, \phi\right).\end{aligned} \tag{17.275}$$

其中我们假设了 $\psi^*(x)\phi(x)|_{\partial\Omega} = 0$，也就是在空间 Ω 的边界 $\partial\Omega$ 上本征函数等于零。这是非常通用的边界条件，尤其是在正负无穷远处的时候，我们都假设物理量在这样的边界上为零。其他更一般的，我们暂时忽略不计。于是，我们看到

$$\left(\frac{d}{dx}\right)^\dagger = -\frac{d}{dx} \tag{17.276}$$

于是，只要

$$\left(\nabla^2\right)^\dagger = \nabla. \tag{17.277}$$

因此之前的算符 L 都是厄米算符。

顺便，这里我们也再一次看到了微积分的重要性（具体知识这里是导数、积分和分步积分公式），要是没有微积分我们都讨论不了这些算符的厄米性，而矢量空间上的算符谱分解定理需要算符的厄米性。我们也再一次看到了分析和代数是合在一起的。

有了函数看作矢量空间这个认知，我们就从概念上把边值问题变成了求解相应算符的本征矢量函数问题和这个本征矢量下的代数问题。

有了函数看作矢量空间这个认知，我们发现，泛函 $F[\cdot]$ 是从 V 到 $F=\mathbb{R}$ 的线性映射，也就是属于 V^* 中的元素，也就是函数这个右矢量对应的左矢量。确实，我们发现，在一组本征矢量函数当作基矢的条件下，这个映射也就成了左矢量，具体来说，就是

$$\begin{aligned} F[\phi(x)\cdot] &= F\left[\sum_n \phi_n\psi_n(x)\cdot\right] \\ &= \sum_n \phi_n F[\psi_n(x)\cdot] = \sum_n \phi_n F_n \\ &= \begin{bmatrix} F_1 & F_2 & \cdots \end{bmatrix} \begin{bmatrix} \phi_1 \\ \phi_2 \\ \cdot \\ \cdot \\ \cdot \end{bmatrix} \end{aligned} \tag{17.278}$$

其中

$$F_n = F[\psi_n(x)], \tag{17.279}$$

是泛函 $F[\cdot]$ 作用在基矢函数 $\psi_n(x)$ 上的结果。

我希望看到函数空间的左矢量其实是泛函，能够帮你更好地感受到矢量空间的威力。这是震撼心灵的洞见的威力！把函数看作矢量空间竟然能够带来这么多好处：函数可以用基矢函数来逼近了，而且往往保留不多的几项就可以就像我们在 16.9.6 节中用三角函数逼近一般函数所展示的；对函数的线性算符给基矢函数提供了很好的方案，并且这样的契合现象的发生空间和方程还有边界条件的基矢函数大大地简化了我们对问题的表述和解决；函数、算符、泛函，一切都成了投影形式的右矢量、矩阵、左矢量，于是可以做数值计算了。你以前熟悉的 $\mathbb{R}^n$ 和 $\mathbb{C}^n$ 中的计算可以迁移到函数空间里面来了。这是多么震撼心灵的威力啊！

这是数学的美，一个简单的洞见，帮助你理解世界，表达和解决问题，看到不容易看到联系的对象之间的联系。

这是矢量空间在概念上和解析求解上对于微分方程边值问题的意义。现在，我们来看一个在数值计算上矢量空间对于微分方程边值问题的意义。假设我们需要在一个一维区间 $[0,L]$ 上求解 ∇^2 的本征函数方程 $\nabla^2 u=\lambda u$，我们来把这个方程变成一个矩阵形式的方程。除了通过前面的解析计算来

得到 u 的解析表达式，我们还可以求数值解。这个时候，我们就要回到导数的定义式了，

$$\nabla^2 = \frac{d}{dx}\left(\frac{d}{dx}\right),$$
$$\frac{d}{dx}u \to \frac{u(x+\Delta x) - u(x+\Delta x)}{\Delta x}. \tag{17.280}$$

我们这样来表示函数 u

$$u \to \begin{bmatrix} u(0) \\ u(\Delta x) \\ \cdot \\ u(n\Delta x) \\ \cdot \\ u(L) \end{bmatrix}. \tag{17.281}$$

如果边界条件约定 $u(0) = 0 = u(L)$，则第一个分量和最后一个分量可以去掉。于是，

$$\frac{d}{dx} \to \frac{1}{\Delta x}\begin{bmatrix} -1 & 1 & 0 & \cdots & & \\ 0 & -1 & 1 & 0 & \cdots & \\ \cdots & 0 & -1 & 1 & 0 & \cdots \\ \cdots & & & & & \end{bmatrix}. \tag{17.282}$$

这样

$$\nabla^2 = \frac{d}{dx}\left(\frac{d}{dx}\right) \to$$
$$\frac{1}{\Delta^2 x}\begin{bmatrix} 1 & -2 & 1 & 0 & \cdots & & \\ 0 & 1 & -2 & 1 & 0 & \cdots & \\ \cdots & 0 & 1 & -2 & 1 & 0 & \cdots \\ \cdots & & & & & & \end{bmatrix} = \mathscr{D}. \tag{17.283}$$

于是，本征函数方程就成了矩阵方程，

$$\nabla^2 u = \lambda u \to \mathscr{D}u = \lambda u. \tag{17.284}$$

这样我们只需要求解矩阵 $\mathscr{D}$ 的本正值和本正矢量，我们就解决了算符 ∇^2 的本征矢量问题。于是，我们发现，原来线性微分方程的求解问题，如果不

能解析求解而是要数值求解，还是要转化为矩阵方程。这也是为什么线性代数计算软件是所有的通用数学软件的基础，为什么提升了矩阵乘法的效率就可以大大提升数值计算的效率。

17.11.2 矩阵的 SVD 分解的应用

矩阵的 SVD 分解，如公式 (17.214) 所示以及重写在这里，就是把一个矩阵展开成如下的形式，

$$A = \sum_j \sigma_j \left|u_j\right\rangle\left\langle v_j\right|, \tag{17.285}$$

以及用奇异值 σ 最大的有限多个来近似原始矩阵，

$$\hat{A}_L = \sum_{j=1}^{L} \sigma_j \left|u_j\right\rangle\left\langle v_j\right|. \tag{17.286}$$

下面，我们将看到这么一个简单的分解和近似将帮助我们解决多么丰富的问题。

我们先来看 SVD 如何帮助实现主成分分析 (Principal component analysis, PCA)。给定一组数据 $X = (x_{nl})_{NL}$，其中 $n = 1, 2\cdots, N$ 表示第 n 个体，$l = 1, 2\cdots, L$ 表示第 l 维度的指标，因此，x_{nl} 表示个体 n 在维度 l 上的取值，我们想看看这些指标的内部是不是有某些关联性，于是组合一下这些指标得到新的指标会更能够说明问题，而且没准组合出来的指标的重要性不一样，可能仅仅考察组合出来的新指标的少数几个就可以基本上替代考察所有原始指标。如果可以我们再用这些新指标中的满足要求（指的是某种意义上就可以取代所有的原指标）的少数，我们还可以把不同个体用这几个少数指标来表示，看看这些个体会不会有一些聚类现象，也就是一部分集中在这些新指标的某些范围内。这里有三个问题：指标变换、指标降维但是仍然具有某种替代性和数据可视化与聚类。

我们来看看，如何解决这三个问题。我们把 $X = (x_{nq})_{NQ}$ 看作一个矩阵，N 行 Q 列矩阵。于是，

$$C_{pq} = \left\langle p\right| X^\dagger X \left|q\right\rangle = \sum_{n=1}^{N} x_{pn} x_{nq}. \tag{17.287}$$

注意，这里的数据通常是实数，因此共轭转置 † 实际上就是转置。我们希望找一个新的指标，由不同的 q 指标组合而成。这样，如果由于某一个 x_{nq} 的均值比较大而导致 A_{qq} 比较大，就不是我们希望得到的了——因为我们可以整体移动数据来改变均值而不改变数据点之间的相对关系。我们想到了做归一化。如果我们先对这些指标做一个标准化，也就是用

$$z_{nq} = \frac{x_{nq} - \bar{x}_q}{\sigma_q} \tag{17.288}$$

代替 x_{nq}，就去掉了每一个维度的指标的均值和方差带来的影响。其中，

$$\bar{x}_q = \frac{1}{N}\sum_{k=1}^{N} x_{kq}, \sigma_q = \sqrt{\frac{1}{N}\sum_{k}(x_{kq} - \bar{x}_q)^2}. \tag{17.289}$$

从现在开始，我们假设 $X = (x_{nq})_{NQ}$ 就是已经做完上面的归一化的。在这个条件下，C 其实就成了这个数据的协方差矩阵，

$$C = \frac{1}{N}\left[C_{pq}\right]_{Q\times Q}. \tag{17.290}$$

至于前面的常数 N 不重要，我们忽略。

前面的三个问题——指标变换、指标降维、数据可视化与聚类中的第一个要求新指标的协方差非对角项最好等于零，表示它们是独立的维度。这个时候，我们发现，正好，可以对 C 做谱分解，得到本征矢量和对角矩阵。第二个指标降维是说，我们希望找到一组新指标，在这个新指标下，原来的信息很大程度上被保留了下来。这个时候，我们只需要保留做完谱分解之后的前几个本征值和相应的本征向量就可以了。其次，这还要求这些样本之间的差异尽可能地被保留下来了，否则不同个体对应的数据点之间的差异变小了，就不是对数据的信息的很好的保留。正好，这也表示应该使得新的指标下的协方差尽可能大。这两条合起来，协方差尽可能大，非对角元最好等于零，那就是要求得到一个对角矩阵，并且保留对角元最大的那些新矢量。这正好就是矩阵的谱分解所解决的问题。于是，我们把

$$\hat{C}_L = \sum_{p=1}^{L} c_p \left|p\right\rangle\left\langle p\right|. \tag{17.291}$$

至于到底保留几个，那是另外一个问题：如果本征值 $c_1, c_2, \cdots$ 有一个内在的大小的突变，比如说前两个差不多大，第三个忽然变得非常小，那就可以

尝试取在大小的突变之前；否则，完全可以多尝试几次，不同的数量的新指标的取法，看看对于后面所要解决的问题的效果。通过协方差矩阵的谱分解来解决—指标变换、指标降维、数据可视化与聚类这些问题的工具就叫做主成分分析，相当于是看到协方差矩阵里面的主要成分。

一旦得到新指标，而且确定好了 L，那就可以把原始的数据点用新指标来表示，也就是把每一个个体 n 的旧指标 $(x_{n1}, x_{n2}, \cdots, x_{nQ})$ 换成新指标下的 $(\tilde{x}_{n1}, \tilde{x}_{n2}, \cdots, \tilde{x}_{nL})$。这样就可以来做原始的数据点的低维可视化，甚至通过可视化来启发分类了。

但是，我们看到，为了做这个新旧指标的变换，我们最好回到对 X 矩阵，而不是 C 矩阵的分解。正好 $C = X^{\dagger}X$，因此，对 C 矩阵的谱分解正好就对应着对 X 矩阵的 SVD 分解。也就是说，我们上面做的事情的更自然的语言是 X 矩阵的 SVD 分解，

$$X = \sum_{np} |n\rangle\langle n| X |p\rangle\langle p| = \sum_{j} \sigma_j \left(\sum_{n} |n\rangle\langle n | u_j\rangle \right) \left(\sum_{p} \langle v_j| p\rangle\langle p| \right) \quad (17.292)$$

其中 $(\sum_n |n\rangle\langle n | u_j\rangle)$ 是一个左矢量,对所有的个体按照一定系数取和,$\left(\sum_p \langle v_j| p\rangle\langle p|\right)$ 是一个右矢量，对所有的指标按照一定系数取和。也就是说，u_j 其实是在个体空间的，而 v_j 是在指标空间的。也就是说，在 SVD 分解的语言下，我们同时得到了需要对个体和对指标做的变换。未来我们主要关注哪一个就对哪一个做相应的变换操作就可以。反过来，有了这两个变换，我们要知道 X_{nj} 就特别简单了，

$$\langle n| X |v_j\rangle = \sum_{pj} \langle n| X |p\rangle\langle p| v_j\rangle, \quad (17.293)$$

$$\langle n| X |v_j\rangle = \sigma_j \langle n| u_j\rangle. \quad (17.294)$$

用这两者的其中之一都可以得到 X_{nj}，然后就可以在新的指标下来表示每一个个体了。

当然，因为这里的协方差是对指标做的，归一化中的取和项也是为了保留指标从而对个体做的。如果想做到完全对个体和指标可以做同样的操作，则还得相应地在归一化的时候也做调整。或者，回到本来就不需要区分个体和指标的问题场景，比如说，在处理的是照片上的信息——把照片上的每一个像素点的颜色值看作是一个矩阵——的时候。

到现在为止，我们说的都是尽可能地保留原始信息的角度，是不是有可能有一些数据其实是噪音，反而我们需要仅仅保留那些占主要成份的部分而去掉噪音呢? 比如说，一张照片模糊了。这个时候，我们发现，仅仅保留这张照片的每个点上的颜色值的矩阵的前几个奇异值分解项可能可以锐化这个照片，使得其主要信息反而更清晰。在这个思路下，这也是一个压缩算法：我们把照片的所有信息用一个仅仅包含其前几个奇异值分解项来代表。本来我们需要 $N \times N$ 个数据点，现在只需要 L 个奇异值和相应的矢量，大概也就需要 $L \times N$ 个数。

在人脸和声音识别中，人们就存下来了这样的综合和大量样本（各个人的脸或者声音）少数的典型模式，把每一个已知的人在这些典型模式空间的分量算出来，当作一个矢量存下来。未来需要识别人脸和声音的时候，把遇到的人脸和声音做这些典型模式上的分解，然后对比得到的新矢量和已经存下来的各个人的矢量就可以 [37][13]。

这个任意一个方阵都可以做分解而尝试仅仅保留绝对值最大的几个奇异值项的 SVD 分解可以运用的场合实在是太多了，原则上所有的方阵都可以尝试用分解出来之后的前几项代替一下试试，看看在后续其他任务解决上的效果。实际上，直到今天，矩阵的 SVD 分解仍然被创造性地使用于物理学和机器学习 [38, 39]。例如矩阵乘积态（Matrix-Product State, MPS），也被称为张量网络（Tensor Network），在量子信息和多体量子物理学、机器学习中应用广泛，而矩阵乘积态的基础是矩阵的 SVD 分解 [38]。

17.11.3 矩阵的谱分解的应用

在物理学中，我们要求所有的可测量物理量都是厄米算符 $O^\dagger = O$。这是因为根据物理知识，物理量的测量平均值的定义为 $\langle s| O |s\rangle$，而原则上我们可以制备出来状态空间所允许的任何一个状态 $|s\rangle$，并且测量值必须是一个实数。也就是说，物理量对应的算符必须满足

$$\langle s| O |s\rangle = (\langle s| O |s\rangle)^* = \langle s| O^\dagger |s\rangle . \tag{17.295}$$

[13]转引自 Wikipedia “Eigenvalues and eigenvectors” 词条，https://en.wikipedia.org/wiki/Eigenvalues_and_eigenvectors，2025 年 11 月 28 日访问。

因此，看起来就得满足 $O^\dagger = O$。不过具体的证明可以用这样两个 $|s\rangle$ 来完成：第一个，$|s\rangle$ 只有两个非零元素分别是 $s_i = 1, s_j = 1$；第二个，$|s\rangle$ 只有两个非零元素分别是 $s_i = 1, s_j = i$。代入到上面的任意状态下的测量均值的一般表达式就能得到，

$$O_{ij} + O_{ji} = O_{ji}^* + O_{ij}^*, \tag{17.296}$$

$$iO_{ij} + (-i)\, O_{ji} = iO_{ji}^* + (-i)\, O_{ij}^*. \tag{17.297}$$

这个方程的解只有

$$O_{ij} = O_{ji}^*, \tag{17.298}$$

也就是

$$O^\dagger = O. \tag{17.299}$$

因此，物理问题所对应的微分方程的求解，往往就成了厄米算符的谱分解问题。例如之前我们已经见过的波动方程、热传导方程和Schrödinger 方程。更进一步，就像我们在边值问题的讨论中所看到的，只要我们完成了适合现象发生的空间和边界条件的谱分解，就相当于给这个问题找到了一组适合这个具体问题的本征矢量。

网络或者说图——最简单的情况下其核心就是把有联系的两个对象之间连上一条线，记作 $A_{ij} = 1 = A_{ji}$，否则记为 $A_{ij} = 0$——是一个描述能力非常强的数学工具。例如，我们可以这样来描述一群人中哪些人经常有接触，那就是选择一个最小的接触次数或者接触时间的阈值，大于这个阈值就算两个人有联系。这样我们可能可以发现这群人中的集团结构，就是哪些人组成了小团体。例如，我们可以把两个词一起出现在一句话中当作有联系，构建一个词共现网络。然后，我们就可以来研究这些经常在一起出现的词在意义上是不是也有联系。这个后来成了自然语言处理的基础。例如，我们可以把两篇论文联系起来如果它们之间有引用关系，构成一个引文网络。这样我们就可以来研究知识在文献中在学科中的扩散。

由于其应用的广泛性，基于网络的研究已经形成了一个应用数学或者也可以算是系统科学 [3] 的分支学科，叫做网络科学 [40]，或者网络分析。其主要任务就是把网络这个工具用于描述现实世界解决现实世界的问题，以及从描述和解决问题中发展出来新的网络分析方法。

很显然，前面提到的最简单的网络就对应着一个对称矩阵 A，而对称矩阵按照谱分解定理，有本征值和本征矢量。那么，是不是把 A 分解成本征值和本征矢量就可以来解决很多的网络科学的问题呢？是的，这个网络科学的子领域被称为网络的谱理论 [41]。发现网络的集团结构的问题就正好可以用这个谱分解来完成 [41]。

一句话，当我们遇到任何一个厄米矩阵的时候，我们都可以尝试来做一下这个矩阵的谱分解，然后把需要原来的矩阵来完成的事情变成在这个矩阵的本征值和本征矢量上来完成。或者，某些特殊的本征值对应着的矢量直接就对于某些问题具有特殊的意义。例如，对于微分方程，$\frac{d}{dt}u = Au$，其零本征值对应的本征矢量 $|0\rangle$ 具有特殊意义——是这个微分方程的稳态。对于迭代方程 $u_{n+1} = Au_n$，则其最大本征值所对应的本征矢量具有特殊意义——是这个迭代方程的最终状态。

前面我们谈到了两个关于矩阵的具体技术的应用——只要遇到矩阵就可以把它们用上去试试，然而，线性代数更大的应用在于矢量和算符的基本概念的应用，也就是通过把什么对象看作矢量和算符来描述这个对象，来和解决关于这个对象的问题。下一小节，我们来看这样的例子。当然，这样描述之后在解决问题的过程中还是会需要具体的技术性的概念和定理的，有些问题的解决还很有技巧性，甚至偶然性（例如 Strassen 算法）。但是，真正的关键反而是把数学结构的基本概念用于描述对象。

来自于现实世界的经验和体验，以及对于整体学科知识体系的深刻理解和反思，有助于提出问题，学科概念有助于表述问题把问题转化成有待解决的学科语言表述的问题甚至猜测出来答案或者回答的方向，定理和技巧有助于解决问题。它们在创造知识和创造性使用知识上的意义不同。未来，我们越来越需要的是提出和表述问题、提出猜想的能力，其次才是求解问题的能力。

因此，学习的时候主要注意力要放在数学结构的基本概念上，然后才是关于这个数学结构的最重要的定理，接着是支撑起来基本概念和最重要的定理的知识体系（围绕基本概念和最重要的定理来选择其他概念和定理，加上逻辑演绎证明，加上提出这些概念和定理、证明这些定理的过程中的数学思维），最后是这些概念和定理的典型应用，最后的最后才是其他定理。在这个过程中你可能得做一些习题，但是做习题的首要目标是促进对概念

的理解，促进对概念的使用场景和提炼概念的过程的理解（也就是双向数学建模），而不是为了掌握某些技巧除非这些技巧的背后是某种典型数学思维方式或者分析方法。

我还希望你体会到，每次都跳出一个层次来思考和总结提炼。例如，从 SVD 和谱分解的应用和矢量和算符的基本概念的应用的对比看到数学建模和数学问题求解之间的差别，更进一步看到数学学习应该关注基本概念和知识体系，而不是技巧，甚至看到一般性的学习要怎么做，以及看到这句话本身在做的事情——时时刻刻跳出一个层次来思考和总结提炼。这属于思维的层次性，也是系统思维 [3] 的体现。

17.11.4 群、矢量和张量在物理学里面的应用

物理学是研究物理对象的状态，包含状态的描述、状态的演化的描述（包含演化发生的原因，也就是建立方程，通常是微分方程，因为需要考虑时间上小小间隔之间的状态之间的变化）的学科。顺便，这样从状态的描述、状态演化的描述和状态演化的原因的描述来考虑问题称为力学的世界观。通常我们用矢量或者某个数学结构 $x(t)$ 来描述物理对象的状态，然后看 $x(t+\Delta t)$ 和 $x(t)$ 之间的联系。

但是，我们还可以换一个角度来这样看：我们把状态的前后变化看作是一个算符 $A(t)$，系统的状态只需要考虑其初始的时刻，也就是 $x(0)$，在 t 时刻系统的状态可以写成 $A(t)\cdot x(0)$。具体这样的数学对象是什么，中间的“作用 $\cdot$”如何计算，我们不管。如果这样，那么，我们 $x(t)\to x(t+\Delta t)$ 就成了 $A(t)\to A(t+\Delta t)$ 就成了。也就是说，我们可以通过建立 $A(t)$ 的方程来描述系统的演化，而不是 $x(t)$ 的演化。

当然，坏处是没准这个方程更复杂了，尽管内涵上应该和 $x(t)$ 的演化方程等价。实践上的好处是，对于任何一个初始条件 $x(0)$ 都适用，但是更大的好处是概念上的，通过研究算符的演化，我们可以更清楚地看到问题内在的结构。

这其实就是群在物理学的应用的根源：物理学的研究对象的数学描述就是群。因为 $A(t)\cdot x(0)$ 显然会要求有连续的两个作用 $A(t_2)\cdot[A(t_1)\cdot x(0)]$，

并且合起来肯定得可以看作是一个整体的作用，也就是

$$A(t_2)\cdot[A(t_1)\cdot x(0)]=A(t_2+t_1)\cdot x(0). \tag{17.300}$$

既然这个运算应该对任何初始值 $x(0)$ 都成立，那么，自然地，就要求有

$$A(t_2+t_1)=A(t_2)\cdot A(t_1) \tag{17.301}$$

的自然定义。于是，也就有了

$$A(t_3+t_2+t_1)=A(t_3)\cdot[A(t_2)\cdot A(t_1)]. \tag{17.302}$$

如果这个演化，在同一个场景下（对应于在物理学中相当于外界的力不变，或者说Hamiltonian不变），最好能够满足下面的条件：先看做系统被演化到 t_1 时候，然后，系统再经过了 t_2+t_3 这个时间间隔的演化，也就是，最好有

$$A(t_3+t_2+t_1)=[A(t_3)\cdot A(t_2)]\cdot A(t_1). \tag{17.303}$$

这个正好就是操作或者说演化算符的结合律。因此，我们发现，群是物理系统的状态变化的最自然的语言。

用这个语言，就算对于一个质点的平动，我们也可以重新表述如下。定义推动算符 $P_{\Delta x(t)}$，其作用如下，

$$P_{\Delta x(t)}x=x+\Delta x(t). \tag{17.304}$$

于是，对于给定的质点的初始位置，我们完全建立起来了其位置状态 $x+\Delta x(t)$ 和算符 $P_{\Delta x(t)}$ 之间的关系。如果需要我们可以把关于 $x+\Delta x(t)$ 的方程变成关于 $P_{\Delta x(t)}$ 的方程。类似地我们还可以定义作用在速度上的 $P_{\Delta v(t)}$。

对刚体的旋转也是，我们不再通过刚体的欧拉角来描述，而是通过转动来描述。这个时候，我们甚至会发现，我们需要某个数学结构来表示这些转动，这样才能完成具体的计算。于是，我们自然就需要群的表示。

实际上，我们还能够用导数算符把 $P_{x(t)}$ 明确写出来，以单变量函数和一维运动为例（多维的时候用得偏导数）

$$P_{\Delta x(t)}=e^{\Delta x(t)\frac{d}{dx}}=\sum_{n=0}^{\infty}\frac{\left[\Delta x(t)\frac{d}{dx}\right]^n}{n!}=\sum_{n=0}^{\infty}\frac{[\Delta x(t)]^n}{n!}\frac{d^n}{dx^n}. \tag{17.305}$$

可以验证，

$$P_{\Delta x(t)}x=\sum_{n=0}^{\infty}\frac{\left[\Delta x\left(t\right)\right]^{n}}{n!}\frac{d^{n}}{dx^{n}}x=x+\Delta x\left(t\right). \tag{17.306}$$

对于更加一般的函数，

$$\begin{aligned}P_{\Delta x(t)}f\left(x\right)&=\sum_{n=0}^{\infty}\frac{\left[\Delta x\left(t\right)\right]^{n}}{n!}\frac{d^{n}}{dx^{n}}f\left(x\right)\\&=f\left(x\right)+f'\left(x\right)\Delta x\left(t\right)+\frac{1}{2}f''\left(x\right)\Delta^{2}x\left(t\right)+\cdots\\&=f\left(x+\Delta x\left(t\right)\right).\end{aligned} \tag{17.307}$$

最后一步我们用了Taylor 公式，公式 (16.79)。

顺便，我们可以证明 $\left\{P_{\Delta x(t)}\right\}$ 构成群。

例 17.14 ($\left\{P_{\Delta x(t)}\right\}$ 构成群). 我们先来证明封闭性：对于任意的可导函数 $f\left(x\right)$，有

$$P_{\Delta_2 x}\left(P_{\Delta_1 x}f\left(x\right)\right)=f\left(x+\Delta_1 x+\Delta_2 x\right)=P_{\Delta_2 x+\Delta_1 x}f\left(x\right). \tag{17.308}$$

我们再来证明结合律：

$$\begin{aligned}P_{\Delta_1 x}\left[P_{\Delta_2 x}\left(P_{\Delta_3 x}f\left(x\right)\right)\right]&=f\left(x+\Delta_3 x+\Delta_2 x+\Delta_1 x\right)\\&=\left(P_{\Delta_1 x}P_{\Delta_2 x}\right)\left(P_{\Delta_3 x}f\left(x\right)\right).\end{aligned} \tag{17.309}$$

单位元和逆元分别是 $e=P_0$ 和 $P_{-\Delta x(t)}$。

实际上，我们这里不仅证明了 $\left\{P_{\Delta x(t)}\right\}$ 构成群，还证明了这个群是交换群。这里的证明我们全部都运用了基于Taylor 公式的公式 (17.307)。我们还可以直接从指数函数来证明上面的所有结果。例如，我们用指数函数的性质来证明交换律。

$$\begin{aligned}P_{\Delta_1 x}P_{\Delta_2 x}&=\sum_{m=0}^{\infty}\frac{\left[\Delta_1 x\right]^{m}}{m!}\frac{d^{m}}{dx^{m}}\sum_{n=0}^{\infty}\frac{\left[\Delta_2 x\right]^{n}}{n!}\frac{d^{n}}{dx^{n}}\\&=\sum_{m,n}\frac{\left[\Delta_1 x\right]^{m}\left[\Delta_2 x\right]^{n}}{m!n!}\frac{d^{m+n}}{dx^{m+n}}\\&=P_{\Delta_2 x}P_{\Delta_1 x}.\end{aligned} \tag{17.310}$$

实际上，更一般地，给定任何的算符 A，由于 $A^mA^n = A^{m+n} = A^nA^m$，也就是同一个算符永远可以交换顺序，因此 $e^{\tau_1 A}e^{\tau_2 A} = e^{(\tau_1+\tau_2)A}$。但是，一般地，对于 $AB - BA \neq 0$，$e^{\tau_1 A}e^{\tau_2 B} \neq e^{\tau_1 A+\tau_2 B}$。

现在，基于微积分（导数、Taylor 公式）和群我们竟然有了平动群的明确表示 $P_{\Delta x(t)} = e^{\Delta x(t)\frac{d}{dx}}$，甚至我们竟然理解到了所有的平动合起来竟然是群，我们竟然把运动状态的演化的描述变成了对平动群的元素的演化的描述。这些在学习微积分和群论之前，都是不可想象的。不过，这样做，好像也没啥具体好处，除了思考更深刻了，理解世界更深刻了——物理学是研究对对象的操作的。

下面，我们来体会一下好处。我们来看看如果我们对平动群的元素求导数，我们得到什么。通过这个求导数，没准我们可以把平动群和矢量也联系起来。

$$\lim_{\Delta x\to 0}\left[\frac{P_{x+\Delta x}-P_x}{\Delta x}\right]f(x_0) = \lim_{\Delta x\to 0}\frac{f(x_0+x+\Delta x)-f(x_0+x)}{\Delta x} = f'(x_0+x), \quad (17.311)$$

于是，我们得到

$$\left(\frac{d}{dx}\right)_P (P_x)_f = (P_x)_f \left(\frac{d}{dx}\right)_f. \quad (17.312)$$

为了明确指出第一个导数算符作用在算符 P_x 上，最后那个导数算符作用到一般函数上，我们给它们各自引入了下标。

这是一个非常神奇有非常自然的结论。神奇的地方在于 $P_x = e^{x(t)\frac{d}{dx}}$ 是一个函数空间上的算符，不是一个一般的函数，但是，我们对其参数 x 求导数，竟然得到了这个导数就是 $P_x\left(\frac{d}{dx}\right)_f$。自然的地方在于，对于指数函数，确实其导数为

$$\frac{d}{dx}(e^{x\alpha}) = e^{x\alpha}\alpha. \quad (17.313)$$

如果我们把这里的 $\frac{d}{dx}$ 看作 α，则这个结果正好就符合指数函数的导数的计算结果。反过来，我们也可以通过求解如下的微分方程来理解指数函数的上面的求导数的结果。

例 17.15 (指数函数满足的微分方程和求解). 求解微分方程 $\frac{d}{dx}f(x) = f(x)\alpha$。假设初始条件是 $f(0) = f_0$。

$$\begin{aligned}
&\frac{d}{dx}f\left(x\right)=f\left(x\right)\alpha\\
&\Rightarrow\frac{df}{f}=\alpha dx\\
&\Rightarrow\int_{f_0}^{f}\frac{df}{f}=\int_{0}^{x}\alpha dx\\
&\Rightarrow\ln f-\ln f_0=\alpha x\\
&\Rightarrow f\left(x\right)=f_0e^{\alpha x}.
\end{aligned}\tag{17.314}$$

这里例子说明 $\frac{d}{dx}f\left(x\right)=f\left(x\right)\alpha$ 也可以当作是指数函数的定义，也就是说，公式 (17.312) 完全可以看作是对平动群的元素 P_x 的定义。就好像 $\frac{d}{dx}f\left(x\right)=f\left(x\right)\alpha$ 定义了指数函数，公式 (17.312) 定义了算符的指数函数，为了强调是把算符放到指数里面，一般把这样 $P_x=e^{x\frac{d}{dx}}$ 称作算符 $\frac{d}{dx}$ 的指数映射，或者简单称为算符的指数映射。

我们看到，算符的指数 $\left\{P_x=e^{x\frac{d}{dx}}\mid\forall x\in\mathbb{R}\right\}$ 映射构成群，这些群元 P_x 的导数是 $P_x\frac{d}{dx}$，等价地我们可以说 P_x 是 $\frac{d}{dx}$ 的参数为 x 的指数映射。

这时候，我们再来看一维平动群：其所有的元素 P_x 都是 $\frac{d}{dx}$ 的参数为 x 的指数映射，其导数扣除其自身以后或者在 $x=0$ 的地方取导数就是 $\frac{d}{dx}$。前面我们学习矢量空间的时候就学习到了导数算符 $\frac{d}{dx}$ 是矢量，因此，一维平动群就是这个矢量的参数为 x 的指数映射。

类比于指数函数 $e^{\alpha x}$，就是一个 α 的参数为 x 的指数映射，其导数扣除其自身以后或者在 $x=0$ 的地方取导数就是 α。更一般地，如果我们是一个三个变量的函数 $e^{\alpha_1x+\alpha_2y+\alpha_3z}$，我们就可以得到其在 $x=y=z=0$ 的地方的三个偏导数分别为 $\alpha_1,\alpha_2,\alpha_3$。记为

$$\left(\frac{\partial}{\partial x}\hat{i}+\frac{\partial}{\partial y}\hat{j}+\frac{\partial}{\partial z}\hat{k}\right)\left(e^{\alpha_1x+\alpha_2y+\alpha_3z}\right)\Big|_{x=y=z=0}=\alpha_1\hat{i}+\alpha_2\hat{j}+\alpha_3\hat{k}.\tag{17.315}$$

这个时候，我们说 $e^{\alpha_1x+\alpha_2y+\alpha_3z}$ 是矢量 $\alpha_1\hat{i}+\alpha_2\hat{j}+\alpha_3\hat{k}$ 的指数映射。

当我们回到三维空间的平动的时候，注意下面的表达式中的导数算符相当于上面的 α 们，

$$P_{\hat{r}=x\hat{i}+y\hat{j}+z\hat{k}}=e^{x\frac{\partial}{\partial x}+y\frac{\partial}{\partial y}+z\frac{\partial}{\partial z}}.\tag{17.316}$$

其在 $x=y=z=0$ 的地方的导数的三个分量为

$$\left(\frac{\partial}{\partial x}\right)_P P_{x\hat{i}+y\hat{j}+z\hat{k}}=\left(\frac{\partial}{\partial x}\right)_f, \tag{17.317}$$

$$\left(\frac{\partial}{\partial y}\right)_P P_{x\hat{i}+y\hat{j}+z\hat{k}}=\left(\frac{\partial}{\partial y}\right)_f, \tag{17.318}$$

$$\left(\frac{\partial}{\partial z}\right)_P P_{x\hat{i}+y\hat{j}+z\hat{k}}=\left(\frac{\partial}{\partial z}\right)_f. \tag{17.319}$$

类比于前面的三变量指数函数,我们说三维空间的平动群是矢量 $\left(\frac{\partial}{\partial x},\frac{\partial}{\partial y},\frac{\partial}{\partial z}\right)$ 的指数映射。

我们来回顾一下：从数的指数函数开始，我们已经得到了导数算符的（参数为矢量 $\vec{r}$ 的）指数映射，并且这个指数映射的导数（扣除自身或者在参数零点取值）就是导数算符，而导数算符是矢量。更进一步，似乎物理系统的状态的研究都可以走这条状态变化算符群的道路，如果状态变化算符群未来都可以表示为矢量的指数映射，那也就是说，物理系统的研究都可以走这条矢量的指数映射的道路。

在量子力学里面，我们就既可以通过研究状态矢量的演化——称为 Schödinger 绘景，也可以通过研究算符的演化——称为 Heisenberg 绘景，来研究量子系统的演化。在场论中，往往我们选择研究算符的演化，而不是状态的演化。

联系到导数可以看作是一元函数的切线斜率，那么，导数算符就可以看作是物理系统状态变化算符群这个指数映射的切线，就是矢量。因此，实际上，**微积分是线性代数在函数空间的运用**。函数是矢量，对函数的求导和求不定积分都是从函数到函数的线性映射，也就是线性算符，而线性算符本身也可以看作是矢量。仅此而已。导数是生活在函数的切空间——在某个点上一切可导函数的切线方向所构成的矢量——中的对象。函数和这个切空间中的矢量存在着几乎一对一的关系（差一个常数，因为 $f(x)+c$ 和 $f(x)$ 的导数相等）。

实际上，这也回答了一个很有意思的问题：为什么可以本来仅仅是一个函数 $f(x+\Delta x)$ 在 x 点附近保留 Δx 的线性项的主要部分的 $f'(x)$（$f(x+\Delta x)=f(x)+f'(x)\Delta x$）竟然可以通过不定积分得到原函数的所有的信息：$f'(x)$ 给出了 $f(x)$ 在每个点的切线的斜率（或者说切线方向的矢

量），为了使得原函数给出的切线处处和 $f(x)$ 相同，这个原函数除了零点之外没有其他选择的自由度。

回到我们之前问的这样做的好处在哪里的问题。其实，如果仅仅考虑可交换的对物理状态的操作算符构成群，除了更好地理解世界之外，也没什么其他的好处。但是，一旦遇到不满足交换律的群，而只要是两个不对易的算符的指数映射就不会满足交换律，则研究这样的指数映射要比直接研究状态好多了。大量的在状态的层面看不清楚的问题，就可以在操作构成的群的层面看得更加清楚。而且研究这些操作的时候，可以通过研究这些操作的导数来进行。实际上，这就是李群和李代数的研究内容。李群就是李代数的指数映射，李代数就是李群的导数 [42]。

例如，在量子状态的世界中我们需要考虑状态叠加，也就是一个被叫做自旋的微观状态可以处于 z 方向的向下态和向上态的叠加，而且这个叠加起来的状态是 x 方向的向上态。用算符的语言可以很简单地看到，x（z）方向的自旋算符是之前定义的 X（Z）[14]。我们只要求一下这两个矩阵的本征矢量就可以看到它们的本征矢量之间确实有前面所说的关系。因此，矢量层面的难以理解的东西需要到矩阵的层面才能看清楚。

习题 17.11 (X 和 Z 的本征矢量). 按照之前在例 17.13 中定义的 X, Y, Z 算符，求解 X 和 Z 的本征矢量，并且验证 $|1_z\rangle = \frac{\sqrt{2}}{2}(|1_x\rangle + |-1_x\rangle)$。

更进一步，对于一个处在 $\hat{r}$ 方向的磁场里面的自旋，基于物理知识——具体指的是Schrödinger 方程这里我们不管它，其演化算符就可以写成下面的指数映射形式的算符，

$$U(t) = e^{-\frac{i}{\hbar}Ht} = e^{\mu B \frac{i}{\hbar}(\hat{r}_x X + \hat{r}_y Y + \hat{r}_z Z)t}. \tag{17.320}$$

于是，我们自然地看到了算符的指数映射的出现。既然出现了这样的指数映射，我们自然就想到如果可以简化为各自算符的指数映射就好了，也就是这样，

$$e^{(A+B)t} \stackrel{?}{=} e^{At} e^{Bt}. \tag{17.321}$$

或者为了记号简单我们把 t 也吸收到 A, B 里面去。我们知道，只有当 $[A, B] = AB - BA = 0$ 的时候，这个等式才能成立。对于一般情况，我们

[14]物理学的历史发展上，是先看到了现象，导致人们猜测其实状态之间有这样的关系，然后再构建出来这些算符。我们这里倒过来，从建立起来的理论来理解世界。

有 Baker-Campbell-Hausdorff (BCH) 公式，

$$e^A e^B = e^{A+B+\frac{1}{2}[A,B]+\frac{1}{12}[A,[A,B]]+\frac{1}{12}[B,[B,A]]+\cdots}. \tag{17.322}$$

这说明演化算符假设包含了一些由于 $[A,B] \neq 0$ 导致的 $e^{(A+B)t}$ 作用到矢量上的奇怪的信息，我们通过看算符的指数映射就可以看到这样的信息。

更更进一步，如果我们对于这些算符也有一个完成的基矢的集合 $\{A_j\}$，那么我们就可以把每一对这样的对易关系用这样的基矢来展开，

$$[A_i, A_j] = \sum_k \Gamma_{ij}^k A_k. \tag{17.323}$$

于是，

$$[A_i, [A_j, A_k]] = \sum_l \Gamma_{jk}^l [A_i, A_l] = \sum_l \Gamma_{jk}^l \Gamma_{il}^m A_m. \tag{17.324}$$

那 Baker-Campbell-Hausdorff (BCH) 公式的右侧就成了一堆带着系数 Γ 的 n 次方的但是关于 A 的一次方项，并且更高项的系数更小，

$$e^{A_i} e^{A_j} = e^{A_i + A_j + \frac{1}{2}\sum_k \Gamma_{ij}^k A_k + \frac{1}{12}\sum_l \Gamma_{jk}^l \Gamma_{il}^m A_m + \cdots}. \tag{17.325}$$

这样一切尽管还不是 $e^{A_i+A_j} = e^{A_i} e^{A_j}$，但是也不远了，大量的信息都藏在了系数 Γ 中。

真正理解到这样做的好处，需要在具体物理系统的这样处理能够更好地解决的问题上来体会。这一点，我们只能留给物理学去完成了。不过，至少，我希望你甚至在完全没看懂上面的细节计算的情况下，大概看懂了其中的逻辑，体会到了：

- 研究算符的演化至少和研究状态的演化是等价的，很多复杂的问题通过研究算符的演化可以更简单，更能建立起来深刻的认知；
- 顺便，在这个过程中，算符往往可以看作是指数映射，而对指数映射的求导得到的是（指数映射的切空间的）矢量；
- 更一般地，可以把函数的导数看作（函数的切空间的）矢量，因此微积分可以看作是矢量代数在函数这个矢量空间上的应用。

从概念上看到群和矢量、算符在物理学中的应用，在物理学中的核心地位之后，下面，我们来看更加具体的矢量、张量在物理学中的应用。

哪些物理对象是用矢量这个数学对象来描述的呢？这就非常多了，三维空间中的位移、速度、动量、力、电场强度、磁场强度、力矩、角动量这些和力以及运动有关的量的数学结构都是矢量。其背后的原因是因为它们都满足矢量相加的首尾相加的法则，或者说平行四边形法则。如果把时空看作一个合起来的舞台，也就是在相对论的层面，则动量和能量合起来的四维动量，电磁场的矢量势和标量势和起来的四维势，也都是矢量。甚至，量子系统的状态也是矢量——称作波函数，并且不是时间和位置空间的矢量，而是时间和位置的函数空间的矢量。量子力学中的物理量是算符，也是张量。刚体的转动惯量则是张量，应力、应变和弹性是张量，电导率、磁化率是张量，电动力学中的电磁张量是张量，广义相对论中的能量动量张量和曲率张量是张量。在广义相对论中我们还会遇到更高阶的张量。物理学需要用到这些矢量和张量来表示物理对象是因为物理对象的行为，通过实验测量看出来，展现了满足这些数学对象的要求或者是能够对这些数学对象做做的操作上的性质。

下面，我们来看一个初看起来不会是矢量的对象被研究者变成矢量从而进行更好的研究的例子。

17.11.5 矢量和自然语言处理

我们知道对于一个矢量来说，能对它做的事情是：数乘（缩放）、加法（首尾相加）和反过来的减法、旋转、相似性（内积、夹角）、求长度（内积）、基矢展开（一套给定的独立因素下的分量）和坐标变换（那组独立因素可以选）。那么，如果一个还不是矢量的对象，但是，能够对其做上面的操作中的一部分，往往我们就可以尝试来把它做矢量，然后看看对它的操作是不是可以看作算符或者其他的可以用数学结构表示的对象，甚至最后用数表示出来，这样，我们就有了可以对这个对象和相应的操作的作存储和计算的方式。这样，不仅人类可以操作它，计算机也可以操作它了。

在基于人工智能的自然语言处理技术兴起之前，计算机也是要处理语言的 [43]。例如，统计一下一篇文章里面有几个字，有几个不同的字，几个词几个不同的词（如果是中文还牵涉到分词，也就是也哪些字合起来看作

词），比较一下不同的作者的文章的字频和词频（同一个作者往往可能具有比较固定的字词使用频率，你看我就用很多括号和“于是”、“也就是说”），给文章按照主题来分一下类然后做类别的研究（同一主题类的文章可能词的频率分布函数比较相似），在不同的语言之间做翻译（最简单的翻译，如果每一种语言的每一个意思只有一个词来表示，一个词也仅仅表示一个意思，那就只需要查词表对应关系，再加上语法上的调整），甚至更加复杂的篇章的阅读理解的任务（最简单的阅读理解任务就是找信息，问文章中关于某个对象的某个特性说了什么，从文章中找到这个对象的这个特性的词就行）。

那个时候，主要依靠的是统计学。计算机把每一个词（对于汉语可能是字）当作一个独立的对象进行编码，例如说“汉语”就是编号为 000001 的词。然后，在计算机中就有了一篇文章的编码。接着取对这些编码做统计分析就行了。统计分析第一包含描述性统计，也就是字词的频率，某个主题下的字词频率（这个可能需要先由人类做一些分好的类来找到特定的频率分布特征，当然后来人们也发明了不需要人类先分好类的技术，例如一种叫做 Latent Dirichlet allocation 的主题分类技术），某个作者下的字词频率等等。如果我们建立了多语言之间的这个编码的对应关系，那我们也大概可以来做机器翻译了。统计分析还包含上下文关系的建模。例如，“当...”的后面经常跟着“的时候”，“我玩得很”后面经常跟着“开心”或者“累了”。这样的条件概率函数对于从当前的文本来预测下一个出现的词会很有用。再后来，人们就不再满足于仅仅对字词做编码和统计分析了，而是提出了一个叫做 $n-gram$ 的概念 [43]，就是把任意的出现在一起的 n 个字词放在一起，不管在语法上和语义上是不是一个通常认为的单位，当作一个语言单位，来做编码和统计分析。这个时候就会出现比把字词当作单位更丰富的字词表，因而也就有可能把某些问题解决的更好。

但是，这样的基于统计学的自然语言处理有一个非常大的缺陷——每一个字词或者 n-gram 都被当作独立单位进行的编码，而实际上，词之间是有相似程度和差的。我们如何找到一个语言单位的比较忠实的表示呢? 对于相似程度，比如说，“北京”和“北平”是不同时期对同一个城市的称呼，“汉语”和“中文”基本上是同义词，“开心”和“舒畅”可能很接近。对于两个词的差，比如说，“漂亮”和“丑陋”属于反义词，不同对的国家和首都

关系，例如“中国-北京”、“日本-东京”、“美国-华盛顿”都属于“国家-首都”这个关系。比如说，“曹操-曹植”、“苏洵-苏轼”、“祖冲之-祖🄵之”都是“父子”关系，而“父-子”、“母-女”、“父-女”、“母-子”都属于一代的直接亲属关系。显然，当我需要用到“父子”关系来举例子的时候——我在写这一段的时候就是这样想的——我就需要找找大家比较熟悉关于“父子”关系的例子。也就是说，字词或者更一般的 n-gram 这样的语言单位的后面其实藏了这些语言单位的含义上的联系。这个联系典型的有相似性和差。

联系到矢量，我们发现，矢量正好可以用来计算相似性和差啊。我们是不是可以用矢量来表示字、词、n-gram 等语言单位呢？如果可以没准篇章层面也可以表示成矢量，比如说通过直接把篇章内的所有更小的单位的矢量相加，或者其他更高级的方式。而且这样做还有一个额外的好处，没准我们可以用更加少的维度的矢量就来更准确地表示这些语言单位，而不像原来的编码方式，每一个字一个编码，所有的字的编码所在的维度是字的数量。如果这个时候联系到矩阵的 SVD 分解，没准这个降维的背后正好是某个矩阵的 SVD 呢。

于是，Tomas Mikolov（托马斯 · 米考洛夫）等人提出了一种把词转化成向量的技术，称作word2vec[25]。

word2vec的目标是把每一个词转化成为一个 V 维矢量，并且尽可能地保持词之间的意义联系。但是，计算机不会真正意义上的联系啊，计算机知识到一个词经常和哪些其他词出现在一起。是不是一个词的含义和其周围的词的含义有一定的相似性呢？如果有，没准可以通过调整这个词的矢量表示，使得经常一起出现的词的矢量表示更相似。于是，如图 17.4 所示，word2vec构造了一个从目标词开始来计算周围词（例如前后 c 个，于是总共 $2c$ 个周围词）的概率的一个学习机。这个学习机希望通过调整每一个词所对应的矢量来保证：给定这个词（所对应的向量 u）的输入学习机给出来的输出结果和真实语料中的周围的词（所对应的向量 $v_{-c},\cdots,v_{-1},v_1,\cdots,v_c$）的概率最大，输出结果和周围词不一样的概率最小，这个这个概率是通过如下的词矢量的内积来计算的，

$$P\left(v\right)=\frac{e^{v^{T}u}}{\sum_{w}e^{w^{T}u}}. \tag{17.326}$$

至于具体算法上如何找到满足这个目标的词矢量的表示，我们就不展开了。

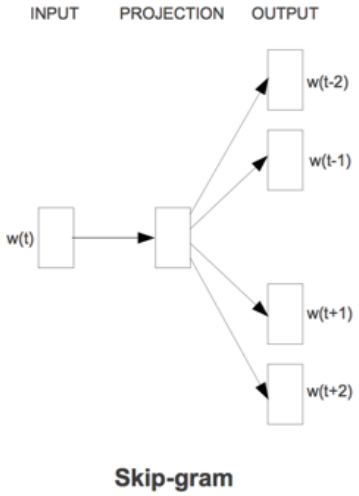


图 17.4: word2vec算法的 SkipGram 实现的示意图 [25]，从一个词的周围出现了其他什么词来得到这个词的矢量表示。

这里的关键，就是看到了要用矢量去表示字词等语法单位。后来，人们还拓展到用矢量去表示句、段、篇章。此外，由于词汇和语义之间没有一对一的关系，不同语言之间的翻译并不是一件简单的事情，但是，有了一定程度上反映语义的矢量表示，是不是机器翻译就会更简单和准确一点呢？有了矢量，统计语言处理中的条件概率建模是不是也会更简单更准确呢？因此，后续的自然语言处理就在矢量表示的基础上得到了很大的发展。

不过，最近的基于人工智能大语言模型（Large Language Model, LLM）的自然语言处理，接过了这个矢量化的思想但是走了完全不同的道路。大语言模型不再依赖于统计一个词附近的词来保证经常出现在一起的词具有相似性来得到词的矢量表示了，而是通过构建一个叫做 Transformer 的神经网络来完成词预测的功能——从以后的上下文来预测其中故意被掩盖住的字词——的方式来得到字词等语言单位的矢量表示 [26, 43]。

更进一步，实际上，某一句话（或者一个段落、一篇文章、一本书）的翻译可以看作是这句话的意思在不同的语种里面的表示。也就是说，所要翻译的语言单位的含义其实是一个抽象矢量，但是这个抽象矢量任何时候都需要一种具体的语种才能被表示出来，除了在大脑内部那时候可能是用

神经信号的方式表示的，然后人类也通过这个人所掌握的语言的形式来操作这个抽象的语言矢量。更进一步，人类的思考所用的语言，甚至所用的神经信号的方式，也是这个抽象矢量的一种表示而已。

那这个语言单位的语义的抽象矢量生活在哪个矢量空间之中呢？它们生活在概念空间之中。下面，我们来看看概念空间中的矢量。

17.11.6 矢量和概念空间

人类语言的内核是概念空间。人类知识也是生活在概念空间中的对象。

以数学知识为例，概念空间的基本元素是：数学概念、数学定理和公理，以及它们之间的关系。我们说过，数学概念其实是双向的命题。因此，数学的概念空间的基本单位就是命题、命题之间的联系。命题之间的联系包含哪些呢？第一种，证明关系，也就是从一个或者多个定理，一个或者多个公理，通过逻辑演绎证明，来得到另一个定理这样的关系。例如，从厄米算符的本征值和本征矢量分解定理我们得到了矩阵的 SVD 分解定理，从 Lagrangian 微分中值定理得到了 Newton-Leibniz 定积分和不定积分关系。第二种，集合的细化后者说集合的包含关系。例如，从三角形这个集合分裂出来直角三角形、锐角三角形和钝角三角形，例如从一般的线性算符走到厄米算符的定义。

当然，人类生活语言所对应的概念空间没有这么严格，概念有的时候不会明确定义出来，联系有可能还有别的类型。例如，自然语言中的“桌子”，比较狭窄的含义是“有四条腿和一个面，面相对来说比较平，可以用来放东西”的一个东西。但是，实际上，在野外我们也经常找一个没有分开的腿的，面也不太平的东西来当作桌子。关系上呢，还有对象之间的类比，比如说，“那个人就像花一样好看”，在好看程度的维度上把一个人和（也不知道是一朵特定的花，还是所有的花的平均）花做了类比。

之前，我们已经见过很多概念地图了。概念地图其实就是对概念空间的对象和关系的另一种表达方式。如果我们把自己先局限在学科概念知识单层网络里面，或者说和跟学科相关的对这个世界的经验和体验、学科概念知识、学科大图景、一般性人类思维、教和习的方法这个多层知识网络里面，我们就可以要求：从给定的单层知识网络，或者多层知识网络出发，构建出来表示其中的每一个概念和每一个关系的矢量。如果得到了这个矢量，

相比于从关于这些对象的自然语言素材而不是知识网络得到的关于这些对象的表示，可能这里从概念网络得到的表示会更加准确。当然，没准还能相互比较甚至相互结合，来得到更准确的这些对象的矢量表示。

那个时候，基于这些矢量表示再回到自然语言处理，类似于做机器翻译、内容总结、机器写作、篇章阅读理解题的回答这些任务就会更好了。甚至，反过来，那个时候，基于关于这些对象的自然语言素材来构建这些对象的单层和多层知识网络的任务都能完成的更好了。

而这一切的基础，都是概念、命题等这些对象的矢量表示。而这个表示的背后，不过就是搞清楚一个数学结构的内涵，然后，找到和这个内涵相匹配的现实世界的对象，或者反过来，基于所要描述的现实世界的对象而建立起来可以描述这些对象的数学结构。这也就是双向数学建模或者说广义的表示论。回到矢量，我们需要掌握的核心就是：数乘、加法和反过来的减法、旋转、相似性、求长度、基矢展开和坐标变换，这些核心的背后是线性性。希望未来你能够创造性地发挥矢量以及其他数学结构的描述作用，甚至根据你描述和解决问题的需要提出新的数学结构，创造新的数学知识。

顺便，这也是深刻理解一个数学概念（以及不仅仅数学概念）的好处：抓住尽可能少的核心，才能促进创造性使用，并且未来有必要和有机会提出自己的数学概念的时候，也可以参考这个抓住核心来定义概念的方式。

17.11.7 为什么仅仅是线性的线性代数这么通用

从来自于仅仅是原函数的线性主部的导函数可以得到完整的原函数，从曲线在每个点的切线就可以得到曲线，从泛函其实是函数空间的左矢量，从导数运算和积分运算都是函数空间的线性算符，从函数在不同算符的本征函数矢量下的展开表示，从矢量和张量在物理学、机器学习、自然语言处理等领域的应用，我们都可以体会到矢量和张量这些仅仅是线性的数学对象竟然如此通用。线性不是一个非常狭窄的概念吗？怎么这么狭窄的概念会这么通用呢？

比如说，当我们说一个函数是线性的时候，我们指的是一个函数只有可以表示为

$$y = f(x) = kx + b \tag{17.327}$$

因为只有这样才会有 $f(\mu x+\nu y)=\mu f(x)+\nu f(y)$。注意，这里的线性是针对 $f(\cdot)$ 作用在 x 上说的。

但是，比如说，当我们说一个微分方程是线性的时候，我们指的是，这个微分方程中的各阶导数项，包含零阶导数也就是这个函数它自己，都是一次的，也就是[15]，

$$\sum_{l=1}^{n} a_l(x)\frac{d^l}{dx^l}y+a_0(x)y=0. \tag{17.328}$$

因为只有这样才满足：如果 $y_1(x)$ 和 $y_2(x)$ 都是这个方程的解，则 $\mu y_1(x)+\nu y_1(x)$ 也是这个方程的解。因此，线性性具体指的对象不同。对于这个线性微分方程，显然其解就允许之前函数意义上非线性函数。例如，

$$\frac{d}{dx}y+ay=0. \tag{17.329}$$

满足线性微分方程的定义，而其解为

$$\frac{1}{y}dy=-adx\Rightarrow y=y_0e^{-ax}, \tag{17.330}$$

这是一个非线性函数。

以物理学的质点运动的牛顿方程为例，给定Lagrangian$L=\frac{1}{2}m\dot{x}^2-V(x)$（根据物理学知识，这个时候等价地Hamiltonian为 $H=\frac{p^2}{2m}+V(x)$）的条件下的运动方程为

$$\frac{d}{dt}\left(\frac{\partial L}{\partial\dot{x}}\right)=\frac{\partial L}{\partial x} \tag{17.331}$$

也就是

$$m\ddot{x}=-V_x'(x). \tag{17.332}$$

只要 $V(x)$ 中包含非 x 的常数项、一次项和二次项的其他项，则公式 (17.332) 必然出现非线性项。例如当 $V(x)=\frac{1}{2}kx^2$ 的时候，牛顿运动方程为 $m\ddot{x}=-kx$ 是线性方程，但是当当 $V(x)=\frac{1}{2}\phi x^4$ 的时候，牛顿运动方程为 $m\ddot{x}=-3\phi x^3$ 是一个非线性方程。

[15]下面的方程的右侧可以出现一个 $f(x)$，都算线性方程。不过，那个时候需要区分齐次方程的通解和非齐次方程的特解。为了不进入这个更复杂的问题，我们让右侧等于零。

但是，如果我们换一个角度，从 (x,p) 空间上的分布函数 $\rho(x,p;t)$，或者说量子态的波函数 $\psi(x,t)$ 的角度来考虑这个问题，这个方程就分别成了

$$\frac{\partial}{\partial t}\rho(x,p;t)=\frac{\partial H}{\partial x}\frac{\partial \rho}{\partial p}-\frac{\partial H}{\partial p}\frac{\partial \rho}{\partial x}=V'(x)\frac{\partial \rho}{\partial p}-\frac{p}{m}\frac{\partial \rho}{\partial x}. \tag{17.333}$$

和

$$i\hbar\frac{\partial}{\partial t}\psi(x,t)=H\psi(x,t)=\left[-\frac{\hbar^2}{2m}\nabla^2+V(x)\right]\psi(x,t). \tag{17.334}$$

可以验证这两个方程都是线性方程。你只需要验证如果两个函数各自是方程的解，则它们的线性叠加也满足方程就行。

我们看到了什么：从函数形式的线性性，到函数的微分方程的线性性，到函数当作自变量的分布函数的线性性，尽管每次还是线性方程，但是其研究对象的角度在不断拓展的同时使得原来的非线性的对象变成了新的角度下的具有线性性的对象。甚至，我们还可以进一步把研究对象的视角转化到分布函数所在的函数空间的函数，没准那个时候的线性方程也能够解决一部分在分布函数的视角下的非线性方程。

其次，线性函数是非线性函数的一个近似，线性系统的解是非线性系统的解的一个近似。对于前者，我们已经通过Taylor 公式有体会，重写在这里，

$$f(x+\Delta x)=f(x)+f'(x)\Delta x+o(\Delta x). \tag{17.335}$$

对于后者，你只有通过学习微分方程定性理论或者物理学中的场论的微扰计算来体会了。

合起来，我们说，为什么看起来那么狭窄的线性性这么通用呢？因为第一，其给更复杂的非线性情形提供了近似提供了基础，更重要的是第二，我们可以通过拓展我们的数学对象来使得拓展以后的数学对象上的线性性其实包含了拓展之前的数学对象上的非线性的情形。上一小节甚至我们看到了另一个非常深刻的拓展：算符的指数映射和算符的关系，实际上就是导数和积分的关系，而导数其实是矢量，是一种线性性，也就是用线性的切空间的矢量来描述原始的可能是非线性的对象。

17.12 从矢量看双向数学建模

现在终于到了我们通过回顾学习矢量代数部分的历程来感受到数学的学科大图景以及更高层知识的时候了。我们先从熟悉的矢量的具体例子感受到了什么是矢量，启发了我们应该把矢量定义成什么东西。例子有人们在地球表面行走的经验体验，有数学中已经学习过的二维矢量空间的矢量，有物理学中的位移、速度、加速度和力，有时钟的指针，有傅里叶级数，有概念之间的差和相似性。我们体会到的这些矢量具有的共性大概有矢量的加法、缩放、长度、方向、旋转、给定基矢下的分量表示、不同基矢下的坐标变换、被操作对象的线性性和操作的线性性。

然后，我们在这些启发下定义了矢量，发现加法和数乘以及它们要满足的性质是这个矢量的关键。为了说清楚它们要满足的性质，我们回去先从集合定义了群和域。顺便，群的定义也受到了实际问题、数学本身的发展和物理学的发展的启发。有了群和域，我们定义了抽象矢量——这个矢量可以完全就不是数。例如，后来我们看到函数可以是矢量，而那个时候对偶矢量是泛函，求导和积分是函数这个矢量空间上的线性算符。

有了基本的矢量空间的定义之后，我们定义了抽象的内积，当然是在二维空间矢量的内积的启发下。有了内积，依靠映射的语言，我们定义了矢量和对偶矢量，以及线性算符，还有张量，引入了 Dirac 符号，引入了空间的正交归一基矢，证明了对于厄米算符的谱分解定理和针对一般算符的 SVD 分解。在给定的一组基矢下，理解了矩阵乘法为什么遵循这么奇怪的规则，明白了矩阵不过就是抽象算符在给定的一组基矢下的投影。

在这中间，为了讲清楚厄米算符的概念，我们不得不引入了由矢量空间到矢量空间的线性算符 A 诱导的，基于内积建立的矢量和对偶矢量之间的一一对应关系的，从对偶矢量空间到对偶矢量空间的线性算符 $A^\dagger$。以及把这个 $A^\dagger$ 又作用回矢量上来——这个把定义域拉回来到矢量空间的过程还是通过内积来完成的。从矢量代数本身的知识上，这就是为了后面学习谱分解和 SVD 分解，但是，从双向数学建模的角度来说，这其实是给物理学做准备的，因为在物理学中，所有的可观测量都得是厄米算符。

最后，我们举了矢量代数的多个典型应用场景，包含把边值问题转化成函数空间上的谱分解，SVD 用在主成分分析、数据降维、数据去噪、以及广泛用于现代物理学和机器学习的张量网络，把谱分解用于物理学问题

和网络科学，群在物理学中的基础地位（这部分实际上属于李群和李代数这个更高阶的数学和物理学的主题，但是你看到我们以现有的矢量代数和微积分的知识可以掌握其整体思路），矢量和张量在物理学中的应用于描述物理量，矢量用于描述语言单位和概念——这个思维和知识的单位。

做完这个回顾，我们发现，本书中的矢量、算符、张量的概念的提出完全是在现实世界的经验体验的启发、物理学等其他学科的要求和启发、现有数学知识的启发之下来进行，这些概念明确定义之后就进入了数学学科概念知识之间的联系的环节，然后准备好了一些关键概念和定理之后，我们再通过把这些知识用于解决世界的问题，解决当时启发这些概念的提出和发展的场景的问题之后，来更好地理解这些概念。甚至，关于这些概念的定理的发展，也是受到这些应用场景的启发的。例如，厄米算符的普分解定理的发展可以看作是受到物理学发展的要求启发的。

这就是双向数学建模，从现实世界的经验体验、其他学科的要求，加上数学学科本身的知识系统性的要求，来提出和解决问题，来提炼概念来猜测命题，然后经过数学知识体系本身的完善之后得到明确的概念和已经被证明的命题，把这些概念和命题再用于解决问题，以及进一步提出和解决新问题。

你可能会说，所谓的双向数学建模现在主要也就是把数学结构用于描述现实这个单向的了，我为什么要理解到双向数学建模呢？这确实是事实，提出新的数学结构确实比以前更不常见了。但是，也不是没有啊，万一轮到你来受到现实世界或者其他学科的启发从而得到这样的机会呢？比如说，近些年实际上还是由新的数学结构的提出，例如非对易几何——代表空间上的不同点的数之间就非对易的，这样的几何上的代数和分析怎么做，有什么用。比如说，更早的时候发展起来的微分几何——弯曲空间上的微积分和代数是什么，有什么用。因此，也不是没有机会来提出和发展新的数学结构。另一方面，这世界的很多问题的表示和解决方案，目前阶段，尤其在更接近实践的领域和层面，还是靠经验甚至感觉的，没有经过严密的思考和数学建模、实验检验的。如果你学会了数学建模，就算是单向的，就有可能在这些领域和层面发挥巨大的作用。

当然，实在不行，我们说了本身的目标是，培养数学知识的创造和创造性使用的欣赏者。如果你能够欣赏到别人做的双向数学建模，别人在发

展数学知识的系统性本身上的艰苦卓绝，背后的学科思维，背后的苦和乐，那也将丰富你自己的人生。

顺便，我们也把本章中体现的其他的数学学科大图景，以及更高层的知识做一个总结。这本来是本章小结的任务。我们先在这里，和本章具体内容的学习的回顾过程结合，这样在本章小结中，我们就简单提一提这些名词就行，不再展开了。

中间，以用矩阵表示群为例，我们介绍了狭义的表示论——用矩阵表示群，和广义的表示论——用一种数学结构表示另一种数学结构，和最广义的表示论——用数学结构表示世界中的对象，也就是双向数学建模。

整个矢量代数的这一章，还体现了代数的思想——通过对对象的操作满足的性质来定义对象，更加具有一般性。对比初等数学，在那里我们先定义对象和操作，再来看对这些对象的这些操作满足什么性质。这样每一种对象都是具体对象具体讨论的。这也没问题，但是相比于代数的思想一般性不够。注意，前面我们还学习过一个初等数学范畴内的代数的思想——用符号表示对像，这个对象可以是现实世界的对象、某个单位下的现实世界的对象的量（的数学结构）、这个量的值（数）、这些量所参与的表达式，简称物、量、数、表达式。那个代数的思想和这里的代数的思想不是一个东西。一个是中文“用符号代替数”的意思，一个是“运算和要满足的性质”的意思。

通过把函数看作是矢量，我们还顺便看到了，微积分是线性代数在函数这个矢量空间的应用。看到了分析和代数天然的结合：空间中的点和点之间的距离需要有代数来提供，有了距离极限、导数和积分需要依靠分析来建立。更一般地说，这也体现了不同的数学分枝学科之间的天然的融合关系。

关于矢量和矢量的投影之间的关系，也被我们类比到了语义和不同语言对语义的表达，类比到了概念空间的中的思想在不同的语言甚至包含神经信号的方式的表达。

在思维方式和学习方法的层面，我们看到了学习更应该关注基本概念而不是技巧，甚至都不应该是定理，除非这个定理可以促进对概念的理解，或者这个定理是很多应用或者后续学科知识的基础。在创造知识和创造性使用知识上，来自于现实世界的经验和体验以及对先有学科知识体系的深

刻理解有助于提出问题和提出猜想，学科概念有助于表述问题，定理和技巧有助于解决问题。而我们更加需要的是提出和表述问题、提出猜想的人。从这一小节，以及 17.11.3 节，我们还体会到了，任何一个具体任务完成了，都要跳出这个具体任务来做高一个层次的总结反思。这属于思维的层次性，也是系统思维 [3] 的体现。

甚至，我还希望你们做到，在不太理解前面的所有细节的条件下能够看到整体思路从而大概理解这些高层的知识，关于数学学科大图景的，关于一般性人类思维方式的，或者关于教和习的方法的。当然，如果能够彻底地上下贯通，把这些高层知识理解得更好还是更好的。那个时候，甚至，我会更进一步，希望你在彻底上下贯通以后，忘掉大量的那些学科概念知识层面的细节，然后用学科大图景和一般性人类思维，配合教和习的方法的指导，来重新通过提出和解决问题来把学科概念知识重新“创造”一遍。

顺便，我想告诉你，这就是我在本书中完成的事情。我在写本书的时候，我尽量忘掉所有的已经学会的学科知识，也不迁移学习其他教材中的教什么和怎么教，完全以提出和解决问题，培养提出和解决问题的创造者和创造欣赏者为目标，重新来选择教什么，决定怎么教。我就是这样来重新“创造”这些知识的。当然，不得不承认，有的时候，这个“创造”的结果和前人的教材其实相差没有这么大，尽管大多数时候这个差别都不小。我不会故意去追求不同，但是，如果我觉得为了培养创造者和欣赏者这个教什么和怎么教自然就和绝大多数教材不同了，我一点也不遗憾，甚至会有一点点小小的开心——显摆自己对相应的知识和思维的深刻认知的开心。

17.13 线性代数和人类思考抽象问题能力的爆发

在线性代数自己来说，人们开始把各种各样的对象，只要能够做旋转、缩放、加法、长度、方向、线性作用到其他对象和被其他对象线性作用，都用矢量和算符来描述，同时矩阵的矩阵的谱分解和 SVD 分解也激发了大量的线性代数的应用案例。人们把矩阵的谱分解和 SVD 分解尝试用于任何一个矩阵。从前面的例子我们已经看到，不仅仅传统的物理学的物理量用矢量和算符来描述，物理量算符需要用到算符的谱分解和 SVD 分解，一般的数据压缩、数据去噪音、数据降维，自然语言处理和机器学习，都开始大量

地使用矢量和算符的语言。机器学习也开始使用张量网络。

线性代数对人类抽象问题思考能力的激发也不仅仅在线性代数自己的应用上。欧氏空间 $\mathbb{R}^n$ 和相应的复空间 $\mathbb{C}^n$ 可以从微积分和线性代数两个角度被推广到一般的空间。在微积分那里，我们主要是对欧氏空间 $\mathbb{R}^n$ 的函数做了深入的研究，包含用多项式级数或者三角函数级数来逼近或者说表示任意一个函数，函数的切线或者说导数，函数的积分，以及支撑这些概念的极限的概念，以及极限的概念背后的点之间的距离的概念、实数的构造和性质的概念。但是，实际上，在微积分中我们是把距离当作一个平凡的概念的。

在矢量空间中，我们是定义了矢量也就是之前微积分中的点，以及点之间的距离的——通过内积来定义的距离。而且，这里的矢量的定义更加的抽象。于是，任何定义了距离的矢量空间，我们都可以把微积分里面的基本概念，包含极限、导数、积分等等都迁移过来。并且，我们也已经看到，把函数看作矢量，定义函数的内积之后，求导数和积分就成了这个空间上的线性映射。因此，如果说我们要把欧氏空间 $\mathbb{R}^n$ 和相应的复空间 $\mathbb{C}^n$ 可以从微积分和线性代数都推广到其他更加一般的空间的话，我们发现，其实矢量空间是更加基本的。先定义空间的点，点之间的距离，我们才能谈极限和微积分。

当然，我们这是抠细节，实际上矢量空间和微积分本来就是结合在一起的。两者结合在一起之后，才能处理更加一般的问题。于是，人们开始在那些能够看作是矢量的对象上，尝试定义距离，然后尝试建立起来那样的对象的空间上的微积分。比如说，泛函分析就是把线性代数和微积分拓展到函数空间，微分几何就是企图建立起来一般的流形——局部看起来非常像 $\mathbb{R}^n$ 或者 $\mathbb{C}^n$ 但是整体上是弯曲的空间——上的微积分。前者在量子力学和量子场论中具有基础性的地位，后者在相对论中具有非常基础的地位——那里有引力而引力表现为时空弯曲。目前，物理学在研究的量子力学和相对论的统一——这是Albert Einstein（爱因斯坦）以及很多（大概应该算“所有”）物理学家的梦想。显而易见，泛函分析和微分几何的某种结合对于这个梦想也是重要的。当然，目前，物理学家和数学家还没有确定到底这个最合适这个统一理论的数学结构是什么。没准，这个问题上，将来会有你的贡献。

当然，遗憾的是，随着这些研究的发展，所用的数学语言也越来越抽象了。因此，尽管线性代数大大促进了人类思考抽象问题能力的爆发，但是，要体会到这个爆发对于非专业数学家和物理学家，或者非相应应用领域的专家，都成了一件不容易的事情。在这里，我已经给你描绘了一个大概的样子。等到梳理出来了更合适的更丰富的例子，再来带着你在你已经学会的现在这个知识基础上，跟你分享。

17.14 推荐阅读材料

和本书在学科概念知识上的选择和选择的理念比较接近的是Axler的《Linear Algebra Done Right（线性代数这样学）》[34]。矢量和算符远远比行列式重要，矢量和算符的抽象形式远远比分量形式重要。如果你想看比较传统的线性代数的教材，可以参考Gilbert Strang（吉尔伯特 · 斯特朗）的《Linear Algebra and Its Applications》[32] 和Serge Lang（塞尔日 · 兰）的《Linear Algebra》[31]，或者丘维声的高等代数 [33]。

在群和域这部分，你可以参考David Dummit（大卫 · 杜米特）和Richard Foote（理查德 · 福特）的《Abstract Algebra》[44]，John Fraleigh（约翰 · 弗瑞）和Neal Brand（尼尔 · 布兰达）的《A First Course in Abstract Algebra》[45] 和丘维声的《近世代数》[46]

关于矢量和映射这部分，强烈推荐梁灿彬和周彬的《微分几何入门与广义相对论（上册. 第二版）》[47]。

如果你对本书的教和习的理念和方法还不熟悉，你可以去看吴金闪的《教的更少，学得更多》[1]。本书也会渗透一些系统思维。关于系统思维你可以参考吴金闪的《系统科学导引》[3]。

17.15 作业

习题 17.12 (指数映射). 一个算符 A 的指数映射定义为 $e^A = \sum_{n=0}^{\infty} \frac{1}{n!} A^n$。请你计算以下算符的指数映射：$e^I, e^X, e^Y, e^Z$。其中 I, X, YZ 之前在例 17.13 中已经定义。（提示：采用直接套定义计算的方式，中间可能得需要三角函数的知识和欧拉公式 $e^{i\theta} = \cos\theta + i\sin\theta$。）

习题 17.13 (指数映射). 一个算符 A 的指数映射定义为 $a^A = \sum_{n=0}^{\infty} \frac{1}{n!} A^n$。请你计算以下算符的指数映射：$e^I, e^X, e^Y, e^Z$。其中 I, X, YZ 之前在例 17.13 中已经定义。(提示：采用各个算符自己的本征矢量来当作基矢)

习题 17.14 (SVD 分解). 把矩阵 $A = \begin{bmatrix} 1 & 2 \\ 0 & -1 \end{bmatrix}$ 做 SVD 分解。(提示：构建 $A^\dagger A$ 和 $AA^\dagger$)

习题 17.15 (谱分解). 把矩阵 $A = \begin{bmatrix} 1 & 2 \\ 2 & -1 \end{bmatrix}$ 做谱分解。

习题 17.16 (矢量或者算符的例子). 找一找，在自己的生活和工作中，或者听说的其他人的生活和工作中，或者你所阅读的其他的资料中，有没有什么东西可以看作是矢量或者算符的。

17.16 本章小结

在学科概念知识上，如图 17.5 所示，本章主要学习了群、域、矢量（加法、数乘、内积）、算符（张量）、厄米算符的谱分解、一般算符的 SVD 分解、群表示论。其中集合和映射的语言在这些概念的定义上起到了非常大的作用，Dirac 符号对于更清楚地理解和表达这些概念也起到了非常大的作用。

这些概念非常抽象，但是这个抽象性可以依靠来自于现实世界和已有数学知识的例子来变得更容易理解。因此，我们花了不少的篇幅来讨论启发这些概念和命题的提出的现实世界的经验和体验、物理学等学科的学科概念知识和待解决的问题，以及可以启发到这些概念和命题的已有的数学知识。当然，就算这样，这个抽象性对学习者的深入思考的意愿还是有很高的要求的。在学习了这些概念和命题以及它们构成的知识体系之后，我们还花了不少的篇幅来介绍这些概念和命题的引用。一方面，这可以帮助你更好地理解这些概念和命题，另一方面，可以帮助你未来把这些概念和命题使用的案例迁移到你自己提出和解决问题的场景中。

在学知识上，学习群、矢量、算符这些概念其实不需要多少已有数学知识当基础，大概来说得有整数、有理数、实数和复数的四则运算以及四则运算律、二维平面上的矢量的坐标表示和加法数乘，就够了。这些大概应该算

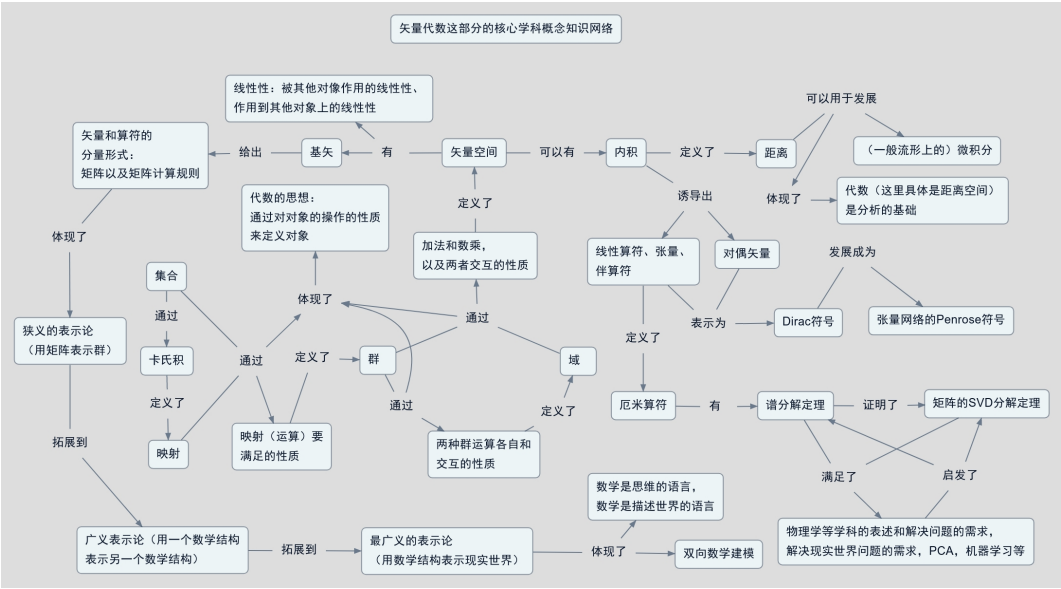


图 17.5: **矢量代数一章的知识体系**。矢量代数这一章主体内容其实特别少（只有矢量、算符、算符的谱分解和算符的 SVD 分解），相互联系紧密，但是比较抽象。从中我们还可以体会到数学的学科大图景。这张图可能也得单独印一页，横版，或者折页。

初中层次的知识。原则上，复数的概念在这里用到的也就是负共轭和绝对值。当然，为了学习群、矢量、算符，我们还需要集合和映射的语言。如果之前没有学习过，这些可以算是学习群、矢量、算符的当时来补的。从这里你也再一次体会到了数学知识的系统性，尤其是集合和映射、逻辑演绎证明在构建这个系统中所起到的作用。

在学科大图景的层面，我们体会到了，对于矢量来说，任何一个被其他对象作用的时候具有线性性或者作用到其他对象上具有线性性的对象，都可以尝试用矢量这个数学结构来描述。这属于矢量这部分的学科典型研究对象。这一章体现的比较深入的典型思维方式有代数的思想（通过对对象可以进行的操作的性质来定义对象，高度抽象和通用）、分析和代数相结合的思想（尽管我们强调了多次，代数比分析更加基本，更加普适）、双向数学建模。当然，做抽象和逻辑演绎证明、知识的系统性在这一章也有非常好的体现。这一章体现的比较好的分析方法有厄米算符的谱分解和一般算符的 SVD 分解。矢量的坐标变换是一个分析方法也是一个思维方式。例如，在矢量用于自然语言处理和用于概念空间的例子中，我们看到，不同的语言形式可以看作是语义的一种表象，人类用语言来表达思考甚至来用神经信号的方式表达也是人类思维的一种表象。

最后，希望本章的写作方式能够供教材编撰专家参考：对于学科概念，尽量要精简仅仅保留最能够体现学科大图景的最少量的学科概念知识，要交代启发这些概念的提出的问题场景和背后的思维方式，要呈现这些概念和已经学过的其他概念之间的联系帮助学习者建立起来知识体系，看到学科概念知识体系，看到发展学科概念知识体系的背后的思维方式，还要展示这些学科概念知识的典型应用场景，这些对于提出和解决问题，创造知识和创造性使用知识都是非常有意义的。

以这么一点点的知识基础，我们学会了群、矢量、算符这么抽象、深刻和应用广泛的概念和相关的定理，理解到了数学的学科大图景，理解到了一般性人类思维还有教和习的方法，这是非常了不起的成就。祝贺你，我再次为你而自豪，去休息休息，去吃顿好的，去玩个痛快，然后回来开启下一个旅程。在本书的范围内，你已经爬到山巅，后面的旅程就应该是从山巅滑沙、滑草、滑道而下，看看风景就行了。当然，偶尔还是会有几个小颠簸的。

第十八章　概率论——提供有关概率这个数学对象的基本概念以及定理

概率和测度，数学结构，公理化定义的提出，看模式、抽象、代数

18.1　语言统一之后的好处

迁移定理，例如通过测度论迁移不动点定理到概率论

对某些概念的理解更深刻

18.2　随机变量是什么——可测函数

18.3　概率密度分布函数

18.4　随机变量函数的概率密度分布函数

18.5　各阶矩、特征函数和中心极限定理

18.6　推荐阅读材料

18.7　作业

18.8　本章小结

第十九章　统计学——从样本数据到结论（到分布函数）

从观测到对对象的描述，到对结论是否成立的判断

19.1　统计学的学科典型责任

从观测到对对象的描述——到分布函数,到对结论是否成立的判断——到给出结论成立的概率或者说可靠性。对有没有差异的判断，对某个函数关系是否成立的判断

19.2 最简单愚蠢的办法——重复生成样本

19.3 重复生成样本的低成本替代——Bootstrap 方法

19.4 已知分布函数的情况——参数估计

19.5 统计回归分析

19.6 Monte-Carlo 方法

Monte Carlo——从分布函数到样本，统计学的逆问题。

19.7 统计学哪些地方需要概率论

19.8 推荐阅读材料

19.9 作业

19.10 本章小结

第二十章　变分法求极值及其应用

20.1　极值问题的欧拉-拉格朗日方程

$$S=\int_{t_i}^{t_f} L\left(x\left(\tau\right),\dot{x}\left(\tau\right),\tau\right)d\tau. \tag{20.1}$$

$$x\left(t_i\right)=x_i, x\left(t_f\right)=x_f. \tag{20.2}$$

$$x\rightarrow x\left(\tau\right)+\delta x\left(\tau\right). \tag{20.3}$$

$$\begin{aligned}\delta S=&\int_{t_i}^{t_f} L\left(x\left(\tau\right)+\delta x\left(\tau\right),\dot{x}\left(\tau\right)+\dot{\delta x}\left(\tau\right),\tau\right)d\tau\\&-\int_{t_i}^{t_f} L\left(x\left(\tau\right),\dot{x}\left(\tau\right),\tau\right)d\tau\\=&\int_{t_i}^{t_f}\left[\frac{\partial L}{\partial x}\delta x\left(\tau\right)+\frac{\partial L}{\partial\dot{x}}\dot{x}\left(\tau\right)\right]d\tau \qquad (20.4)\\=&\int_{t_i}^{t_f}\frac{\partial L}{\partial x}\delta x\left(\tau\right)d\tau+\int_{t_i}^{t_f}\frac{\partial L}{\partial\dot{x}}d\left(\delta x\left(\tau\right)\right) \qquad (20.5)\\=&\int_{t_i}^{t_f}\frac{\partial L}{\partial x}\delta x\left(\tau\right)d\tau+\left.\frac{\partial L}{\partial\dot{x}}\left(\delta x\left(\tau\right)\right)\right|_{t_i}^{t_f}-\int_{t_i}^{t_f}\delta x\left(\tau\right)\frac{d}{d\tau}\left(\frac{\partial L}{\partial\dot{x}}\right)d\tau \qquad (20.6)\\=&\int_{t_i}^{t_f}\left[\frac{\partial L}{\partial x}-\frac{d}{d\tau}\left(\frac{\partial L}{\partial\dot{x}}\right)\right]\delta x\left(\tau\right)d\tau \qquad (20.7)\end{aligned}$$

所有线性项的导数必须为零，或者针对性地取 $\delta x(\tau)=\left[\frac{\partial L}{\partial x}-\frac{d}{d\tau}\left(\frac{\partial L}{\partial \dot{x}}\right)\right]$，因此

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{x}(t)}\right)=\frac{\partial L}{\partial x(t)}. \tag{20.8}$$

20.2 用度规和微积分重构平面几何知识

对于任意曲线 $\vec{x}(t)=\left(x^1(t),x^2(t),\cdots,x^N(t)\right)$，有

$$S=\int_{t_i}^{t_f}\sqrt{g_{\mu\nu}\left(x^1,x^2,\cdots,x^N\right)\dot{x}^\mu(\tau)\dot{x}^\nu(\tau)}d\tau. \tag{20.9}$$

$$\frac{d}{dt}\left(\frac{\partial\sqrt{g_{\mu\nu}\left(x^1,x^2,\cdots,x^N\right)\dot{x}^\mu(\tau)\dot{x}^\nu(\tau)}}{\partial\dot{x}^\eta(\tau)}\right)=\frac{\partial\sqrt{g_{\mu\nu}\left(x^1,x^2,\cdots,x^N\right)\dot{x}^\mu(\tau)\dot{x}^\nu(\tau)}}{\partial x^\eta(\tau)}. \tag{20.10}$$

当 $g_{\mu\nu}$ 为欧氏几何的度规（距离公式的时候），$g_{11}=1=g_{22}$，则

$$\frac{d}{dt}\left(\frac{\partial\sqrt{\dot{x}(\tau)\dot{x}(\tau)+\dot{y}(\tau)\dot{y}(\tau)}}{\partial\dot{x}(\tau)}\right)=0, \tag{20.11}$$

$$\frac{d}{dt}\left(\frac{\partial\sqrt{\dot{x}(\tau)\dot{x}(\tau)+\dot{y}(\tau)\dot{y}(\tau)}}{\partial\dot{y}(\tau)}\right)=0. \tag{20.12}$$

也就是

$$\frac{\dot{x}(\tau)}{\sqrt{\dot{x}(\tau)\dot{x}(\tau)+\dot{y}(\tau)\dot{y}(\tau)}}=a, \tag{20.13}$$

$$\frac{\dot{y}(\tau)}{\sqrt{\dot{x}(\tau)\dot{x}(\tau)+\dot{y}(\tau)\dot{y}(\tau)}}=b. \tag{20.14}$$

得到满足这个要求的曲线为

$$\dot{y}(\tau)=k\dot{x}(\tau). \tag{20.15}$$

也就是

$$y=kx+b. \tag{20.16}$$

我们得到结论：两点之间直线距离最短，有且只有一条直线（满足这个给定两个边界条件的微分方程有且只有唯一解）。

接着，我们可以定义两直线平行然后证明平行的传递性。没有交点的两条直线称为平行线，记为 $l_1 \parallel l_2$。必然有 $k_1 = k_2, b_1 \neq b_2$。这个时候二元一次方程组

$$\begin{cases} y = k_1 x + b_1, \\ y = k_2 x + b_2 \end{cases} \tag{20.17}$$

没有解。

实际上，我们可以把平面几何的公理都证明一遍。这样后面的其他定理就可以用这些证明出来的“公理”来证明了。这样，我们就可以把平面几何的知识体系重新建立起来。在我们这里，起点是

1. 平面上的点和数对 (x, y) 一一对应，
2. 任意两点之间的距离，或者写成无穷小的形式为 $ds = \sqrt{d^2x + d^2y}$。

就这么点基础。后面都是可以计算和证明出来的。

20.3 最小作用量视角的力学和光学

20.4 变分法重写控制论

20.5 从函数微分到变分

20.6 推荐阅读材料

20.7 作业

20.8 本章小结

第二十一章　数学知识网络和学科大图景

Mathematics is the tool specially suited for dealing with abstract concepts of any kind and there is no limit to its power in this field.

数学是一个尤其适合描述各种抽象概念的工具。在这方面它的能力无限。

– Paul Dirac

21.1　数学一二层知识的系统性

21.2　数学学科大图景，及其和一二层知识的联系

21.3　从数学知识体系到科学知识体系

公理化方案

21.4　现代数学在干什么？

主要数学分支子学科。

（基本）不再出现数而是出现关系的数学。

21.5 推荐阅读材料

21.6 作业

21.7 本章小结

第二十二章　双向数学建模

从双向数学建模的思维到已有模型

22.1 数学建模的神——对现实对象及其上操作的表示

22.2 典型模型及其建立过程举例

22.3 数学建模五步

22.4 已有典型模型

22.4.1 优化模型

22.4.2 矩阵模型

22.4.3 微分方程模型——变化率和累积量的问题

微分方程数值解

微分方程定性理论

22.4.4 随机过程模型

22.4.5 博弈模型

22.5 推荐阅读材料

22.6 作业

22.7 本章和本卷小结

数学的发展来源于现实世界的启发和数学研究者对数学知识结构本身的逻辑严密性的追求。其中，在现实启发方面，主要通过双向的数学建模来完成，通过把数学结构用于描述现实来解决现实问题，以及通过把数学结构用来描述现实启发提出新的数学结构。

第二十三章　本书总结以及现代数学在干什么

23.1　像数学家一样提出和解决问题

23.2　数学建模五步

23.3　数学解题四问

23.4　从重复性使用到创造性使用到创造

23.5　现代数学在干什么？

拓扑学、数论微分几何——弯曲时空上的微积分范畴论——模式和关系

23.6　推荐阅读材料

参考文献

[1] 吴金闪. 教的更少, 学得更多 [M]. 科学出版社, 2021.

[2] 吴金闪. 小学数学这样学 [M]. 浙江人民出版社, 2022.

[3] 吴金闪. 《系统科学导引》第一卷 [M]. 科学出版社, 2018.

[4] Newton I. Philosophiæ Naturalis Principia Mathematica（中文版为《自然哲学之数学原理》，王克迪翻译）[M]. England: Apud G. & J. Innys, 1726.

[5] Kline M. Mathematical Thought from Ancient to Modern Times（《古今数学思想》）[M]. Oxford University Press, 1990.

[6] Leibniz G W. Nova methodus pro maximis et minimis, itemque tangentibus, quae nec fractas, nec irrationales quantitates moratur, et singulare pro illis calculi genus [J/OL]. Acta Eruditorum, 1684: 467–473. http://17centurymaths.com/contents/Leibniz/nova1.pdf.

[7] Cajori F. A history of mathematical notations [M/OL]. Dover Publications, 1993. http://www.maths.ed.ac.uk/~aar/papers/cajorinot.pdf.

[8] Boyer C B. The history of the calculus and its conceptual development [M]. Dover, 1949.

[9] Slagle J R. A Heuristic Program That Solves Symbolic Integration Problems in Freshman Calculus [J/OL]. J. ACM, 1963, 10 (4): 507–520. http://doi.acm.org/10.1145/321186.321193.

[10] Burden R L, Faires J D, Burden A M. Numerical Analysis [M]. Cengage, 2016.

[11] Euclid, edited by Heath T L. The Thirteen Books of Euclid's Elements, Books 1 and 2 （《几何原本》） [M]. New York, NY, USA: Dover Publications, Inc., 1956.

[12] Sussmann H, Willems J. 300 years of optimal control: from the brachystochrone to the maximum principle [J]. IEEE Control Systems Magazine, 1997, 17 (3): 32–44.

[13] SUSSMANN H J, WILLEMS J C. THE BRACHISTOCHRONE PROBLEM AND MODERN CONTROL THEORY [M]. In Contemporary Trends in Nonlinear Geometric Control Theory and Its Applications. : 113–166.

[14] 约瑟夫 · 傅立叶. 热的解析理论 [M]. 北京大学出版社, 2008.

[15] Strogatz S. Infinite Powers: How Calculus Reveals the Secrets of the Universe [M]. Mariner Books, 2019.

[16] Gowers T. Mathematics: A Very Short Introduction（《牛津通识读本——数学》） [M]. Oxford: Oxford University Press, 2002.

[17] 张筑生. 数学分析新讲（第一册） [M]. 北京大学出版社, 1990.

[18] Kline M. Calculus: An Intuitive and Physical Approach [M]. Dover Publications, 1998.

[19] Spivak M. Calculus [M]. Publish or Perish, 2008.

[20] Apostol T M. Calculus, Vol. 1: One-Variable Calculus, with an Introduction to Linear Algebra [M]. Publish or Perish, 1991.

[21] Tao T. Analysis I [M]. Hindustan Book Agency, 2016.

[22] Tao T. Analysis II [M]. Hindustan Book Agency, 2016.

[23] Rudin W. Principles of Mathematical Analysis [M]. McGraw-Hill Education, 1976.

[24] Whitehead A N. The Aims of Education and Other Essays（《教育的目的》） [M]. Free Press; Reissue edition, 1967.

[25] Mikolov T, Chen K, Corrado G, et al. Efficient Estimation of Word Representations in Vector Space [J/OL]. arXiv, 2013, abs/1301.3781. http://arxiv.org/abs/1301.3781.

[26] Vaswani A, Shazeer N, Parmar N, et al. Attention Is All You Need. 2023. https://arxiv.org/abs/1706.03762.

[27] Golub G H, van Loan C F. Matrix Computations [M]. JHU Press, 2013.

[28] Strassen V. Gaussian elimination is not optimal [J/OL]. Numerische Mathematik, 1969, 13 (4): 354–356. https://doi.org/10.1007/BF02165411.

[29] Fawzi A, Balog M, Huang A, et al. Discovering faster matrix multiplication algorithms with reinforcement learning [J/OL]. Nature, 2022, 610 (7930): 47–53. https://doi.org/10.1038/s41586-022-05172-4.

[30] Gutknecht M H. A Brief Introduction to Krylov Space Methods for Solving Linear Systems [C] // Kaneda Y, Kawamura H, Sasai M. In Frontiers of Computational Science, Berlin, Heidelberg, 2007: 53–62.

[31] Lang S. Linear Algebra [M]. 3rd ed. New York: Springer Verlag, 1987.

[32] Strang G. Linear algebra and its applications [M]. Belmont, CA: Thomson, Brooks/Cole, 2006.

[33] 丘维声. 高等代数 [M]. 北京: 清华大学出版社, 2019.

[34] Axler S J. Linear Algebra Done Right [M/OL]. New York: Springer, 1997. http://linear.axler.net/.

[35] Scheinerman E R. Invitation to Dynamical Systems [M]. Prentice Hall, 1995.

[36] Strogatz S H. Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry and Engineering [M]. Westview Press, 2000.

[37] Xirouhakis Y, Votsis G, Delopoulos A. Estimation of 3D Motion and Structure of Human Faces [M] // Tzafestas S G. In Advances in Intelligent Systems: Concepts, Tools and Applications. Dordrecht: Springer Netherlands, 1999: 333–344.

[38] Sengupta R, Adhikary S, Oseledets I, et al. Tensor networks in machine learning [J]. Eur. Math. Soc. Mag., 2022, 126: 4–12.

[39] Schollwöck U. The density-matrix renormalization group in the age of matrix product states [J]. Annals of Physics, 2011, 326 (1): 96–192. January 2011 Special Issue.

[40] Bollobás B. Modern Graph Theory [M]. Springer-Verlag New York, 1998.

[41] Chung F R K. Spectral Graph Theory [M]. American Mathematical Society, 1997.

[42] 梁灿彬, 周彬. 微分几何入门与广义相对论 (中册. 第二版) [M]. 科学出版社, 2009.

[43] Widdows D, Cohen T. Large Language Models: How They Work And Why They Matter [M]. SemanticVectors Publishing, 2025.

[44] Dummit D S, Foote R M. Abstract Algebra [M]. 3rd ed. Wiley, 2004.

[45] Fraleigh J B, Brand N E. A First Course in Abstract Algebra [M]. 8th ed. Pearson, 2021.

[46] 丘维声. 近世代数 [M]. 北京: 北京大学出版社, 2015.

[47] 梁灿彬, 周彬. 微分几何入门与广义相对论 (上册. 第二版) [M]. 科学出版社, 2006.

名词索引

B | D | E | F | G | J | K | L | M | N | Q | R | S | T | W | Z

B

聚点定理（Bolzano–Weierstrass） Bolzano–Weierstrass Theorem. 170, 188, 194

D

戴德金定理 Dedekind Cut. 167, 188

戴德金分割 Dedekind Cut. 161–164, 166–171, 188, 189, 195

Dirac 符号 Dirac 发明的用于表示矢量、对偶矢量和算符的符号. 240, 260, 268

E

Euler-Lagrangian 方程 Euler-Lagrangian Equation. 137, 138, 140, 212

F

费马（Fermat）定理 Fermat Theorem. 87, 188, 191

Fourier 级数 Fourier Series. 217, 221

G

伽罗瓦（Galois）理论 Galois Theory. 201, 252

J

经典力学 Classical Mechanics.
- **Hamiltonian** 哈密顿量. 213, 215, 317, 330
- **Hamiltonian 方程** 分析力学 Hamiltonian 形式的方程. 215
- **Lagrangian** 拉格朗日量. 213, 330

K

Kronecker δ 记号 Kronecker δ notation. 180, 265

L

拉格朗日（Lagrange）微分中值定理 Lagrange's Mean Value Theorem. 193, 194

Lagrangian 乘子法 Lagrangian Multiplyer. 89, 90, 130, 131, 148, 216

拉普拉斯算符 Laplacian. 207, 210, 217, 220, 221, 287

Legendre 变换 自变量、函数值、函数形式都变化，维持运动方程不变. 214

M

理解型学习 Meaningful Learning，人类知识高速公路上以高层知识生成器为目标的理解型学习. 19

N

牛顿-莱布尼兹公式 把定积分的计算和导数联系了起来，又称微积分基本定理. 104, 106, 107, 109, 110, 119, 121, 124, 193, 195, 196, 199, 226

Q

确界原理 Least Upper Bound Property. 188, 191

R

罗尔（Rolle）引理 Rolle's lemma. 188, 191, 193

S

施密特（Schmidt）正交化 Schmidt Orthogonalization. 261, 262
Schroedinger 方程 Schrödinger Equation. 303–305, 314, 322

T

Taylor 公式 Taylor's Formula. 98, 100, 101, 137, 148, 196, 201, 202, 222, 318, 331

W

word2vec 用矢量来表示词语的一种算法. 326

Z

最小作用量原理 Least Action Principle. 213
作用量 Action. 213

人名索引

A | B | D | E | F | G | H | K | L | M | N | Q | R | S | T | W | Z

A

Apostol Tom Apostol（汤姆 · 阿波斯托）. 223
Axler Sheldon Axler（谢尔顿 · 阿克斯勒）. 302, 337

B

Berkeley George Berkeley（乔治 · 贝克莱）. 82
Johann Bernoulli Johann Bernoulli（约翰 · 伯努利）. 138
Boyer Carl Boyer（卡尔 · 波耶）. 223
Brand Neal Brand（尼尔 · 布兰达）. 337

D

Dedekind Julius Wilhelm Richard Dedekind（尤利乌斯 · 威廉 · 理查德 · 戴德金）. 162
Dirac Paul Dirac（保罗 · 狄拉克）. 241, 260
Dummit David Dummit（大卫 · 杜米特）. 337

E

Einstein Albert Einstein（阿尔伯特 · 爱因斯坦）. 336
Euclid Euclid of Alexandria（欧几里得）. 91

F

Foote Richard Foote（理查德 · 福特）. 337
Fourier Joseph Fourier（约瑟夫 · 傅里叶）. 221
Fraleigh John Fraleigh（约翰 · 弗瑞）. 337

G

Gowers Timothy Gowers（蒂莫西 · 高尔斯）. 223

H

Hamilton William Hamilton（威廉 · 哈密顿）. 278, 280

K

Kline Morris Kline（莫里斯 · 克莱因）. 223

L

Lang Serge Lang（塞尔日 · 兰）. 337
Leibniz Gottfried Leibniz（戈特弗里德 · 莱布尼茨）. 82, 97, 138
梁灿彬 北京师范大学物理系教师. 337

M

Mikolov Tomas Mikolov（托马斯 · 米考洛夫）. 326

N

Newton Isaac Newton（伊萨克 · 牛顿）. 82, 138, 179, 201, 212

Q

丘维声 北京大学数学科学学院教师. 337

R

Rudin Walter Rudin（沃尔特 · 鲁丁）. 223

S

Bascones James Slagle（詹姆士 · 斯莱格）. 126
Spivak Michael Spivak（麦克 · 斯皮瓦克）. 223
Strang Gilbert Strang（吉尔伯特 · 斯特朗）. 337
Strassen Volker Strassen（沃尔克 · 施特拉森）. 298, 299
Strogatz Steven Strogatz（史蒂文 · 斯托加茨）. 223

T

Tao Terence Tao（陶哲轩）. 223

W

Whitehead Alfred Whitehead（阿尔弗雷德 · 怀特海）. 224
吴金闪 北京师范大学系统科学学院教师. 20, 22, 224, 337

Z

Zeno Zeno of Elea（芝诺）. 80, 152, 154–156
Zeno “飞矢不动”佯谬 古希腊哲学家 Zeno 提出的有关瞬时速度概念的和现实矛盾的结论. 147, 154
Zeno “阿基利斯追不上乌龟”佯谬 古希腊哲学家 Zeno 提出的有关极限概念的和现实矛盾的结论. 80, 82, 147, 152, 154, 156
张筑生 北京大学数学系教授. 223
周彬 北京师范大学物理系教师. 337
祖冲之 中国古代数学家. 151

插图

1 用于描述数学概念间关系的概念地图举例 21

1.1 交集、并集、子集、全集和补集 36

16.1 函数的平均变化率和连线斜率 84
16.2 圆上切线的斜率 85
16.3 速度曲线下的面积 102
16.4 速度曲线下的面积 103
16.5 曲线下的面积的计算方式 104
16.6 从曲线上一小段的长度到曲线长度 111
16.7 半圆的长度的积分计算 112
16.8 半圆面积的积分计算 114
16.9 圆切分成环形微元的面积 116
16.10 直角坐标中长方形的面积的积分计算 117
16.11 直角坐标中长方形的微元的面积 118
16.12 极坐标中圆的微元的面积 120
16.13 $\frac{\Delta\theta}{\sin\Delta\theta}$ 的极限 124
16.14 $\frac{x^2+3}{x-1}$ 的图像 128
16.15 光的折射和反射 141
16.16 分数在数轴上的表示 157
16.17 戴德金分割举例 165
16.18 求方程数值解的牛顿方法 202
16.19 求方程数值解的牛顿弦截法 204
16.20 纵波的波动方程的推导 206
16.21 x 的傅里叶级数 220
16.22 微积分部分的知识体系 225

17.1 矢量叠加的平行四边形表示 232
17.2 对正三角形的所有翻和转的操作 250
17.3 $V^* \to V^*$ 上的线性映射 $\hat{A}^\dagger$ 270
17.4 word2vec 算法示意图 327
17.5 矢量代数一章的知识体系 339

举例目录

1.1 例 (秃头悖论) 30

16.1 例 (过圆上一点的切线) 84
16.2 例 (过圆上一点的切线) 85
16.3 例 (没有约束的极值问题) 87
16.4 例 (约束下的极值问题) 88
16.5 例 (最短距离问题) 91
16.6 例 (最速下降线问题) 92
16.7 例 (Taylor 公式应用) 100
16.8 例 (x^n 的导数和积分) 107
16.9 例 108
16.10 例 108
16.11 例 (半圆的长度) 112
16.12 例 ($f(x)$ 下到 x 轴的面积) 113
16.13 例 (圆的面积) 114
16.14 例 (圆的面积微元法) 115
16.15 例 (长方形面积积分计算) 115
16.16 例 (长方形面积的面积元计算) 117
16.17 例 (圆面积的计算——极坐标) 118
16.18 例 ($\sin'\theta = \cos\theta$) 123
16.19 例 (函数极值问题) 128
16.20 例 (函数极值问题) 130
16.21 例 (最短距离问题) 131
16.22 例 (最速下降线问题) 139
16.23 例 (反射率和折射律) 140

16.24 例 (从无理数构建出来有理数) 151
16.25 例 (Zeno 序列) 152
16.26 例 (Zeno 序列和) 152
16.27 例 (趋向 $\sqrt{2}$ 的数列) 158
16.28 例 ($(-1)^n$ 的极限) 171
16.29 例 (2^n 的极限) 172
16.30 例 (-2^n 的极限) 172
16.31 例 (常值函数的极限) 174
16.32 例 (x^n 的极限) 174
16.33 例 (x 处处连续) 174
16.34 例 (x^2 处处连续) 175
16.35 例 ($|x|$ 处处连续) 176
16.36 例 ($sign\,(x)$ 在 $x = 0$ 处不连续) 176
16.37 例 (x^n 的导数) 179
16.38 例 (x^n 的导数) 180
16.39 例 ($|x|$ 的导数) 181
16.40 例 (常数函数的积分) 183
16.41 例 ($f(x) = x$ 的积分) 184
16.42 例 ($f(x) = x^2$ 的积分) 186
16.43 例 ($f(x) = \cos x$ 的积分) 187
16.44 例 (计算偏导数) 197
16.45 例 (计算偏导数) 198
16.46 例 (已知全微分求原函数) 198
16.47 例 (已知全微分求原函数) 199
16.48 例 (函数形式不变和函数值不变的变换) 211
16.49 例 (傅里叶级数逼近 x 的效果) 219

17.1 例 (整数配上加法构成群) 243
17.2 例 (分数扣除 0 以后配上乘法构成群) 244
17.3 例 ($a + b\sqrt{2}$ 是个域) 251
17.4 例 ($0 \times a = 0$) 254
17.5 例 ($(-1)a = -a$) 254

17.6 例 (二维平面上的矢量) 255
17.7 例 (所有的可导函数) 255
17.8 例 (导数算符) 256
17.9 例 (V^* 是矢量空间) 259
17.10 例 (从原命题得到否命题) 262
17.11 例 (硬币翻转群 $\mathbb{Z}_2$ 的表示) 277
17.12 例 (硬币翻转算符的本征矢量) 281
17.13 例 287
17.14 例 ($\left\{P_{\Delta x(t)}\right\}$ 构成群) 318
17.15 例 (指数函数满足的微分方程和求解) 319

作业目录

1.1 习题 (我的家庭成员) 29
1.2 习题 (我的家庭成员) 30
1.3 习题 (做抽象的其他例子) 42
1.4 习题 (扑克牌是顺序编号吗？) 45
1.5 习题 (生活中的集合) 47
1.6 习题 (生活中的记号、序数和基数) 47

16.1 习题 (圆上的约束极值问题) 90
16.2 习题 (不定积分线性性) 98
16.3 习题 (分步积分) 98
16.4 习题 (Taylor 公式) 100
16.5 习题 (匀加速直线运动从 $x(t)$ 到 $a(t)$) 107
16.6 习题 (匀加速直线运动从 $a(t)$ 到 $x(t)$) 107
16.7 习题 ($x^{-\frac{q}{p}}$ 的导数和积分) 109
16.8 习题 (不定积分的线性性) 121
16.9 习题 (分步积分) 121

16.10 习题 ($\cos'\theta = -\sin\theta$) 124
16.11 习题 ($\frac{1}{n}$ 的极限) 162
16.12 习题 ($\frac{n}{n+1}$ 的极限) 162
16.13 习题 (两个 L 的加法) 168
16.14 习题 (数列有极限则柯西收敛) 170
16.15 习题 (函数有极限则柯西收敛) 170
16.16 习题 (极限是有理数的无理数序列) 170
16.17 习题 (2^n 的极限) 173
16.18 习题 (-2^n 的极限) 173
16.19 习题 (a^x 连续) 177
16.20 习题 (a^x 连续) 177
16.21 习题 (变上限定积分) 196
16.22 习题 (夜宵的数学建模) 224
16.23 习题 (球体体积) 224
16.24 习题 (圆柱体体积) 224
16.25 习题 (圆锥体体积) 224
16.26 习题 (球体表面积) 224
16.27 习题 (圆柱体表面积) 224
16.28 习题 (圆锥体表面积) 225
16.29 习题 (横波波动方程的推导) 225

17.1 习题 (傅里叶级数的积分) 235
17.2 习题 (左矢量映射的线性性) 240
17.3 习题 (Z_2 群) 245
17.4 习题 ($a + b2^{\frac{1}{3}} + c2^{\frac{2}{3}}$ 是个域) 251
17.5 习题 (a', b' 正交归一) 261
17.6 习题 (从原命题得到否命题) 262
17.7 习题 (验证 $I = \sum_j |j\rangle\langle j|$) 267
17.8 习题 ($SS^\dagger = I$) 274
17.9 习题 (硬币翻转群 $\mathbb{Z}_2$ 的表示) 278
17.10 习题 (Y 表象) 290
17.11 习题 (X 和 Z 的本征矢量) 322

17.12 习题 (指数映射) 337
17.13 习题 (指数映射) 338
17.14 习题 (SVD 分解) 338
17.15 习题 (谱分解) 338
17.16 习题 (矢量或者算符的例子) 338