现代机电技术系列教材

XIANDAI JIDIANJISHU XILIE JIAOCAI

现代设计方法及其应用

主编　芮延年

苏州大学出版社

现代机电技术系列教材

XIANDAI JIDIANJISHU XILIE JIAOCAI

机电一体化原理及应用
产品创新设计
工程制图
工程力学
现代工程材料基础
现代设计方法及其应用
计算机集成制造
现代制造技术
先进电子制造技术
模具设计
模具制造
液压与气压传动
传感器与检测技术
机电专业毕业设计指导

ISBN 7-81090-455-8

9 787810 904551 >

ISBN 7-81090-455-6/TM·3(课) 定价：29.00元

《现代机电技术系列教材》

现代设计方法及其应用

主编　芮延年

编者　陈　洁　时忠明　沈惠平　周国良

主审　葛友华

苏州大学出版社

图书在版编目(CIP)数据

现代设计方法及其应用/芮延年主编. 苏州:苏州大学出版社,2005.3
(现代机电技术系列教材)
ISBN 7-81090-455-8

Ⅰ.现… Ⅱ.芮… Ⅲ.机电设备-设计-教材
Ⅳ.TM

中国版本图书馆 CIP 数据核字(2005)第 015790 号

内容提要

本书介绍的主要内容包括:概述、创新设计、优化设计、可靠性设计、弹性力学与有限元、设计中的评价与决策等。本书是编者在多年从事现代设计方法教学科研积累经验的基础上编写的,内容新颖实用,结构体系完整,重点突出,理论联系实际,由浅入深,易于阅读和自学。

本书可作为高等学校工科机械类和近机类专业学生的教材,也可作为各类工科专业的选修课教学用书和工程技术人员继续教育的培训教材。

现代设计方法及其应用

芮延年 主编

责任编辑 陈兴昌

苏州大学出版社出版发行
(地址:苏州市干将东路 200 号 邮编:215021)
常熟高专印刷有限公司印装
(地址:常熟市元和路 98 号 邮编:215500)

开本 787mm×1 092mm 1/16 印张 13.75 字数 341 千
2005 年 3 月第 1 版 2005 年 3 月第 1 次印刷
ISBN 7-81090-455-8/TM·3(课) 定价:29.00 元

苏州大学版图书若有印装错误,本社负责调换
苏州大学出版社营销部 电话:0512-67258835

《现代机电技术系列教材》
编 委 会

总　序

本世纪头 20 年，对我国来说，是一个必须紧紧抓住并且可以大有作为的重要战略机遇期。在世界科技进步日新月异、经济全球化深入发展、国际间生产要素重组和产业转移加快的新形势下，苏州作为全国经济发达地区之一，应利用有利的时机和条件加快发展。要实现更快更好的发展，就必须抓住新科技革命带来的又一次历史机遇，正确驾驭其发展趋势，全面实施科教兴市战略，大力推动科技进步，加强科技创新，加速科技成果向生产力的转化。尤为重要的是，要大力培养一支高素质的人才队伍，在更高的平台上实现科技和经济发展的新跨越。

部分相关高校从 21 世纪对科技创新和人才培养的新要求出发，认真贯彻落实教育部关于面向 21 世纪教学内容和课程体系改革的指示精神，组织有关专家、学者、教授编写了《现代机电技术系列教材》，包括《机电一体化原理及应用》、《现代设计方法及其应用》、《产品创新设计》、《工程制图》、《工程力学》、《现代工程材料基础》、《计算机集成制造》、《现代制造技术》、《先进电子制造技术》、《模具设计》、《模具制造》、《液压与气压传动》、《传感器与检测技术》、《机电专业毕业设计指导》等，很有必要，颇有价值。我们相信，《现代机电技术系列教材》的出版，必将对苏州地区乃至全国机电人才的培养以及机电工业的发展产生积极的促进作用。

南京航空航天大学博士生导师　王 [illegible]

2004 年 7 月

前　言

现代设计方法学是一门正在形成和发展的新兴学科，它研究产品的设计规律、设计程序及设计各阶段的具体方法。本书试图用系统工程的观点，综合各门课程基础知识，使学生掌握机电产品设计的通用方法。其目的在于总结设计规律、启发创造性，在给定条件下实现高效、最优化设计，培养开发性、创造性机电产品设计人才。

基于现代设计方法种类繁多，内容又十分广泛的特点，编者结合多年来从事教学与科研工作的经验，将本书的重点放在目前应用十分广泛的几种现代设计方法：创新设计、优化设计、可靠性设计、弹性力学与有限元及设计中评价与决策方面。这几种方法，突出体现了现代设计与传统设计在思想方法上所发生的三个方面的深刻变化，即设计最优化、分析精确化与设计参数随机化的思想。本书着重介绍这些方法的基本原理及应用。希望读者在较短的时间内对现代设计方法有较全面的了解。

现代设计方法学是机械工程及自动化等相关专业的一门重要课程，课程计划 36～45 学时，本书是按照这个要求编写的。本书也可作为其他机械类专业的本科教材和产品设计人员的培训教材。对从事机电产品设计的工程技术人员和科研工作者、有关专业教师也是一本有用的参考书。

本书由芮延年主编。参加本书编写的有陈洁、时忠明、沈惠平、周国良老师等，盐城工学院葛友华教授主审。苏州职业大学姜左教授，苏州大学出版社陈兴昌老师对本书的编写和出版给予了大力支持；林杰、沈铭、夏威、马艳平、易敏捷、董桂岩、任芸丹等研究生为本书的插图、整理做了大量工作。在此表示衷心的感谢！同时，在本书编著过程中借鉴了不少同行专家和学者的宝贵材料，编者在此向他们表示真诚的谢意。

由于编者水平有限，错误和不当之处恳请广大读者批评、指正。

编　者

2005 年 1 月于苏大

目　录

第1章 绪 论

本章是本书的总论，概括地介绍了现代设计方法的基本概念、研究的主要内容与特点，并对创新设计、计算机辅助设计、优化设计、可靠性设计、有限元法、动态设计、智能设计等部分现代设计方法作了简介，使读者对现代设计方法有一个总体概括的了解。

1.1 现代设计方法的基本概念

1.1.1 设计科学概论

从古至今，人类生活在大自然和人类自身所“设计”的世界中。随着科学技术的发展，人类通过“设计”改变了大自然及人类社会的面貌，人们越来越生活在“人为”、“人技”设计的世界之中。

历史证明，人类文明的源泉就是创造；人类生活的本质就是创造；而设计，其本质上就是创造性的思维与活动。设计的历史也可以说就是人类的历史，但自觉的“设计”是开始于15世纪欧洲文艺复兴时期，直到20世纪中期，设计仍被限定在比较狭窄的专业范围内，单一的学科知识很难解决专业范围内的一些设计问题。

为了更好地满足人类的需求，设计方法必然要发展。随着创造性活动理论、现代决策理论、信息论、控制论、工业设计理论、系统工程等现代理论与方法的发展及传播，人们冲破了传统学科间的专业壁垒，在相邻甚至相远的学科领域内探索、研究，使现代设计科学走上日趋整体化的道路，促使单一的设计研究向广义的设计研究转变，从而形成了现代设计方法学。

从广义角度出发，设计有许多定义，如：

(1) 设计是“一种针对目标的问题求解活动”。

(2) 设计是“将人为环境符合人类社会心理、生理需求的过程”。

(3) 设计是“从现存事实转向未来可能的一种想像跃迁”。

(4) 设计是“一种创造性活动——创造前所未有的、新颖而有益的东西”。

(5) 设计是“一种构思与计划，以及把这种构思与计划通过一定的手段符号化的活动过程”。

(6) 设计是“建立在一定生产方式上的造型计划”。

(7) 设计是“使人造物产生变化的活动”。

(8) 设计是“一种社会文化活动。一方面，设计是创造性的、类似于艺术的活动；另一方面，它又是理性的、类似于条理性的科学活动”。

(9) 设计是“对一批特殊的实际需要的总和，得出最恰当的答案”。

(10) 设计是“实现信念的一种非常复杂的行动”。

(11) 设计是“一种约定俗成的活动，是在规定和创造将来”。

(12) 设计是“完成委托人的要求、目标，获得使设计师与用户均能满意的结果”。

(13) 设计是“一种研讨生活的途径”。

(14) 设计是“综合社会的、经济的、技术的、心理的、生理的、人类学的、艺术的各种形态的特殊的美学活动及其产品”。

(15) 设计是“通过分析、创造与综合，达到满足某些特定功能系统的一种活动过程”。

由此可见，设计的含义并不受学科或专业本身的限制，这些含义具有普遍性与广义性。

1.1.2 传统设计与现代设计

20 世纪以来，由于科学和技术的进步与发展，对设计的基础理论研究得到加强，随着设计经验的积累，以及设计和工艺的结合，已形成了一套半经验、半理论的设计方法。依据这套方法进行的机电产品设计，称为传统设计。所谓“传统”是指这套设计方法已沿用了很长时间，直到现在仍被广泛地采用着。传统设计又称常规设计。

传统设计是以经验总结为基础，运用力学和数学而形成的经验公式、图表、设计手册等作为设计的依据，通过经验公式、近似系数或类比等方法进行设计。传统设计在长期运用中得到不断地完善和提高，是符合当代技术水平的有效设计方法。但由于所用的计算方法和参考数据偏重于经验的概括和总结，往往忽略了一些难解的问题或非主要的因素，因而造成设计结果的近似性较大，也难免有不确切和失误。此外，在信息处理、参量统计和选取、经验或状态的存储和调用等方面还没有一个理想的有效方法，解算和绘图也多用手工完成，所以不仅影响设计速度和设计质量的提高，而且也难以做到精确和优化的效果。

图 1-1 所示为一般传统机械设计过程。由图可见，这一过程的特点是：第一，它的每一个环节都依靠设计者用手工方式来完成。从本质上来说，这些都是凭藉设计者直接的或间接的经验，通过类比分析或经验公式来确定方案，对于特别重要的设计或计算工作量不太大的设计，有时可对拟订的几个方案作计算对比。方案选定后按机械零件的设计方法或按标准选用，最后绘出整机及部件装配图和零件图，编写技术文件，从而完成整机设计。第二，按传统机械设计方法，设计人员的大部分精力耗费在零部件的常规设计(特别是繁重而费时的绘图工作)中，而对整机全局问题难以进行深入的研究，对于一些困难而费时的分析计算，常常不得不采用作图法或类比定值等粗糙的方法，因此具有很大的局限性。这些局限主要表现在：

(1) 方案的拟订很大程度上取决于设计者的个人经验，即使同时拟订了少数几个方案，也难以获得最优方案。

(2) 在分析计算工作中，由于受人工计算条件的限制，只能采用静态或近似的方法而难以按动态精确的方法计算，计算结果未能完全反映零部件的真正工作状态，影响了设计质量。

(3) 设计工作周期长，效率低，成本高。

所以，传统设计方法是一种以静态分析、近似计算、经验设计、手工劳动为特征的设计方法。显然，随着现代科学技术的飞速发展，生产技术的需要和市场的激烈竞争，以及先进设计手段的出现，这种传统设计方法已难以满足当今时代的要求，从而迫使设计领域不断研究和发展新的设计方法和技术。

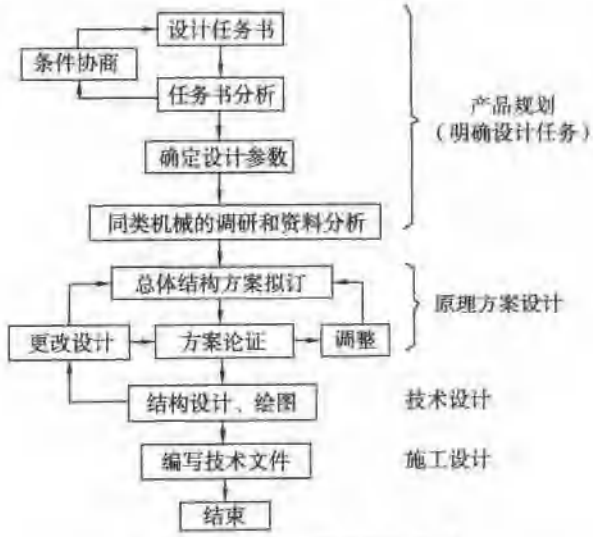


图 1-1 传统机械设计过程

20 世纪 70 年代以来，由于科学技术的飞速发展和计算机技术的应用与普及，给设计工作包括机电产品的设计带来了新的变化。随着科技发展，新工艺、新材料的出现，微电子技术、信息处理技术及控制技术等新技术对产品的渗透和有机结合，与设计相关的基础理论的深化和设计新方法的涌现，都给产品设计开辟了新途径。在这一时期，国际上在设计领域相继出现了一系列有关设计学的新兴理论与方法。为了强调它们对设计领域的革新，以区别于传统设计理论和方法，把这些新兴理论与方法统称为现代设计。当然，现代设计不仅指设计方法的更新，也包含了新技术的引入和产品的创新。目前现代设计所指的新兴理论与方法主要包括：优化设计、可靠性设计、设计方法学、计算机辅助设计、动态设计、有限元法、工业艺术造型设计、人机工程、并行工程、价值工程、协同设计、反求工程设计、模块化设计、相似性设计、虚拟设计、疲劳设计、三次设计、摩擦学设计、模糊设计、人工神经网络、遗传算法等设计分析方法。

1.1.3 现代设计方法的特点

由传统设计与现代设计方法的比较可以看出，现代设计方法的基本特点如下：

(1) 程式性：研究设计的全过程。要求设计者从产品规划、方案设计、技术设计、施工设计到试验、试制进行全面考虑，按步骤有计划地进行设计。

(2) 创造性：突出人的创造性，发挥集体智慧，力求探寻更多突破性方案，开发创新产品。

(3) 系统性：强调用系统工程处理技术系统问题。设计时应分析各部分的有机关系，力求系统整体最优。同时考虑技术系统与外界的联系，即人—机—环境的大系统关系。

(4) 最优化：设计的目的是得到功能全、性能好、成本低的价值最优的产品。设计中不仅考虑零部件参数、性能的最优，更重要的是争取产品的技术系统整体最优。

(5) 综合性：现代设计方法是建立在系统工程、创造工程基础上，综合运用信息论、优化论、相似论、模糊论、可靠性理论等自然科学理论和价值工程、决策论、预测论等社会科学

理论，同时采用集合、矩阵、图论等数学工具和电子计算机技术，总结设计规律，提供多种解决设计问题的科学途径。

(6) 计算机化：将计算机全面地引入设计。通过设计者和计算机的密切配合，采用先进的设计方法，提高设计质量和速度。计算机不仅用于设计计算和绘图，同时在信息储存、评价决策、动态模拟、人工智能等方面将发挥更大作用。

与人们对设计的要求相比，我国现阶段的设计工作相对而言是比较落后的。面对这种形势，唯一的出路就是：设计必须科学化、现代化。也就要求设计人员不仅要有丰富的专业知识，而且还需掌握先进的设计理论、设计方法和设计手段及工具，科学地进行设计工作，这样才能设计出符合时代要求的新产品。

产品是设计结果的物质表现。现代产品的设计不仅依赖于自然科学技术，而且还要受到社会科学和社会因素的支配与影响。这就是说，现代产品设计，除了要求考虑技术方面的因素外，它还要求设计者应将"产品－人－环境－社会"视为一个完整的系统。设计时，必须从系统角度来全面考虑各方面的问题。既要考虑产品本身，还要考虑对系统和环境的影响；不仅要考虑技术领域，还要考虑经济、社会效益；不但要考虑当前问题，还需考虑长远发展。

例如，汽车设计，不仅要考虑汽车本身的有关技术问题，还要考虑使用者的安全、舒适、操作方便等；另外，还需考虑汽车的燃料供应、车辆存放、环境污染、道路发展以及国家能源政策、资源条件、道路建设、城市规划等政策及社会条件限制等问题。因此，现代产品设计已要求设计者把自然科学、社会科学、人类工程学，以及各种艺术、实际经验和聪明才智融合在一起，用于设计中。

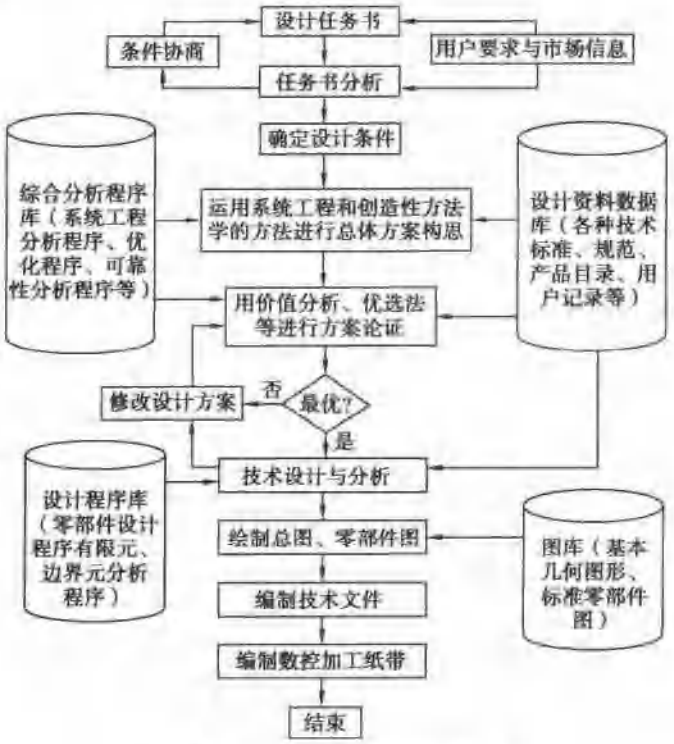


图 1-2 现代设计基本作业过程

最后，应该指出，设计是一项涉及多种学科、多种技术的交叉工程。它既需要方法论的指导，也依赖于各种专业理论和专业技术，更离不开技术人员的经验和实践。现代设计方法是在继承和发展传统设计方法的基础上融会新的科学理论和新的科学技术成果而形成的。因此，学习使用现代设计方法，并不是要完全抛弃传统的方法和经验，而是要让广大设计人员在传统方法和实践经验的基础上掌握一把新的思想钥匙。由于设计方法具有时序性和继承性，之所以冠以“现代”二字是为了强调其科学性和前沿性以引起重视，其实有些方法也并非是现代的，当前传统设计与现代设计正处在共存性阶段。如图 1-2 所示为现代设计的基本作业过程。与传统设计方法相比，它则是一种以动态分析、精确计算、优化设计为特征的设计方法。所以，不能把现代设计与传统设计截然分开，传统设计方法在一些适合的工业产品设计中还在应用。当然，现代设计方法也并非万能良药，现代设计中各种方法都有其特定的作用和应用场合，如优化设计，目前只能在指定方案下进行参数优化，不可能自行创造最优设计方案。而计算机辅助设计也只能在“寻找”方面帮助人的脑和手工作，面决不能代替人脑进行“创造性思维”。这就是现代设计与传统设计方法上的继承与改革的辩证关系。

现代设计方法是一门种类繁多，知识面广的学科群，它所涉及的内容十分广泛，而且随着科学技术的飞速发展，必将还会有许多新的设计方法不断涌现，因此它的内容还会有不断发展。

1.2 设计过程与设计方法

1.2.1 机电产品设计的一般过程

从产品设计角度出发，机电产品设计过程有产品规划阶段、原理方案设计阶段、技术设计阶段和施工设计阶段四个主要步骤。现代设计要求设计者要以系统的、整体的思想来考虑设计过程中许多综合性技术问题。为了避免不必要的经济损失，开发机电产品时应该遵循一定的科学开发生产原则。下面就对开发机电产品设计的一般步骤加以论述。

1. 产品设计规划阶段

产品设计规划，就是决策开发新产品的设计任务，为新技术系统设定技术过程和边界，是一项创造性的工作。要在集约信息、市场调研预测的基础上，辨识社会的真正需求，进行可行性分析，提出可行性报告和合理的设计要求与设计参数项目表。

集约信息应该是生产单位中包括从情报、设计、制造到社会服务等所有业务部门的任务。市场调研要从市场、技术、社会三个方面进行，预测要按科学的方法进行。辨别需求的可行性分析和可行性报告，应由所有业务部门参加的并行设计组和用户共同完成，而不是设计部门或少数部门完成。

2. 原理方案设计阶段

原理方案设计就是新产品的功能原理设计。用系统化设计法将确定了的新产品总功能按层次分解为分功能直到功能元。用形态学矩阵组合按不同方法求得的各功能元的多个解，得到技术系统的多个功能原理解。经过必要的原理试验，通过评价决策，寻求其中的最优解即新产品的最优原理方案，列表给出原理参数，并作出新产品的功能原理方案图。

3. 技术设计阶段

技术设计是把新产品的最优原理方案具体化。首先是总体设计，按照人—机—环境—社会的合理要求，对产品各部分的位置、运动、控制等进行总体布局。然后分为同时进行的实用化设计和商品化设计两条设计路线，分别经过结构设计（材料、尺寸等）和造型设计（美感、宜人性等）得到若干个结构方案和外观方案，再分别经过试验和评价，得到最优结构方案和最优造型方案。最后分别得出结构设计技术文件、总体布置草图、结构装配草图和造型设计技术文件、总体效果草图、外观构思模型。以上两条设计路线的每一步骤，都经过交流互补，而不是完成了结构设计再进行造型设计。最终完成的图纸和文件所表示的是统一的新产品。

4. 施工设计阶段

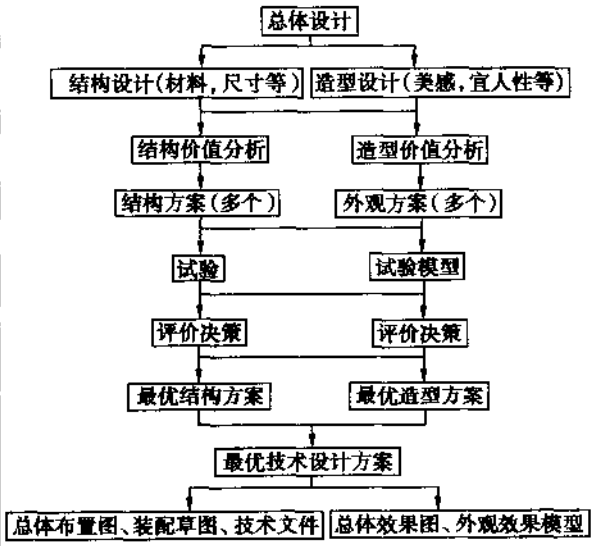
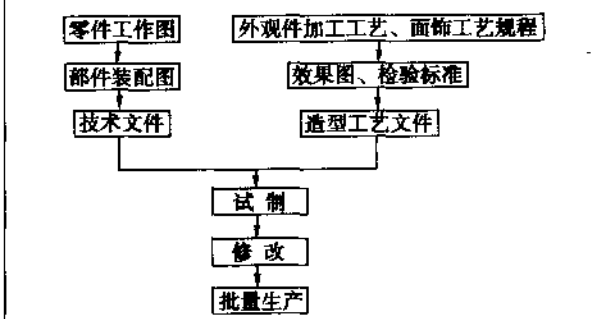
施工设计是把技术设计的结果变成施工的技术文件。一般来说，要完成零件工作图、部件装配图、造型效果图、设计和使用说明书、设计和工艺文件等。

以上机电产品设计的四个工作阶段，应尽可能地采用现代设计方法与技术实现 CAD/CAPP/CAM 一体化。这样可以大大减少工作量，加快设计进度。表 1-1 给出新产品设计一般进程的不同阶段、步骤、使用方法和指导理论，供参考。

表 1-1 新产品设计一般进程

阶段	步骤	方法	主要指导理论
产品设计规划	信息集约（技术造型） ↓ 产品设计任务 ↓ 调研预测（技术造型） ↓ 可行性分析 ↓ 明确任务要求 ↓ 可行性报告、设计要求项目表	设计方法 预测技术	设计方法学 技术预测理论 市场学 信息学
原理方案设计	总功能分析 ↓ 功能分解 ↓ 功能元求解 ↓ 功能载体组合 ↓ 功能原理方案（多个） ↓ 原理试验 ↓ 评价决策 ↓ 最优原理方案 ↓ 原理参数表，方案原理图	系统化设计法 创造技法 评价决策方法	系统工程学 形态学 创造学 思维心理学 决策论 模糊数学

续表

阶段	步骤	方法	主要指导理论
技术设计		价值设计 优化设计 可靠性设计 宜人性设计 产品造型设计 系列化设计 机械性能设计 工艺性设计 自动化设计	价值工程学 最优化方法,工程遗传算法 可靠性理论与实验 人机工程学 工业美学 模块化设计、相似理论 有限元法、动态设计、摩擦学设计、高等机构学 机械设计的工艺基础 控制理论、智能工程、人工神经元计算方法、专家系统
施工设计		各种制造、装配、造型、装饰、检验等方法	各种工艺学

1.2.2 设计方法

设计方法是指达到预定设计目标的途径。在很长的一段时间内,工程设计方法多采用直觉法、类比法及以古典力学、数学和经验数据为基础的半经验设计法,设计中反复多,周期长。20世纪70年代以后,随着计算方法、控制理论、系统工程、价值工程、创造工程等学科理论的发展以及电子计算机的广泛应用,促使许多跨学科的现代设计方法出现,使工程设计进入创新、高质量、高效率的新阶段。设计过程的主要方法与理论如表1-2所示。

表 1-2 设计过程的主要方法与理论

设计阶段	方 法	理论及工具	
明确设计任务 （产品规则）	预测技术与方法	技术预测理论 市场学 信息学	
方案设计	系统化设计法 创造性方法 评价与决策方法	系统工程学 图论 形态学 创造学 思维心理学 决策论 线性代数 模糊数学	计算机
技术设计	构形法 价值设计 优化设计 可靠性设计 宜人性设计 产品造型设计 系列产品设计 模化设计及模型试验	系统工程学 价值工程学 力学 摩擦学 制造工程学 优化理论学 可靠性理论 人机工程学 工业美学 相似理论	计算机
施工设计		工程图学 工艺学	

不管采用哪种技术过程，对每一个具体阶段和步骤都需要应用某种设计方法或技术，各学科、专业中有针对性地解决问题的理论和专门方法，如力学、摩擦学、有限元法等，以及现代设计方法中的计算机辅助设计、优化设计、可靠性设计、人机工程学、工业美学等。

可以把设计的一般程式（纵向主线）和具体设计技术（横向方法）的纵横交叉关系看成是一个三维结构模式，如图 1-3 所示，并可称为“系统工程设计方法”模式。它是一个考虑多因素、多层次的复杂的科学方法体系。

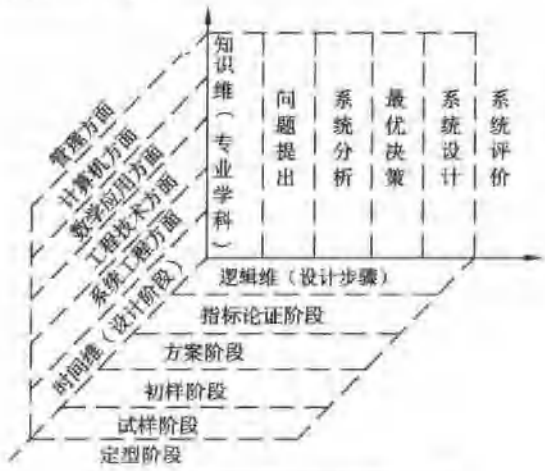


图 1-3 系统工程设计方法模式

1.3 设计类型及设计原则

1.3.1 设计类型

(1) 开发性设计：在设计原理、设计方案全都未知的情况下，根据产品总功能和约束条件，进行全新的创造。这种设计是在国内外尚无类似产品情况下的创新，如专利产品、发明性产品都属于开发性设计。

(2) 适应型设计：在总的方案和原理不变的条件下，根据生产技术的发展和使用部门的要求，对产品结构和性能进行更新改造，使它适应某种附加要求，如电冰箱从单开门变双开门，单缸洗衣机变双缸洗衣机、全自动洗衣机等。

(3) 变参数设计：在功能、原理、方案不变的情况下，只是对结构设置和尺寸加以改变，使之满足功率、速比等不同要求，如不同中心距的减速器系列设计、中心高不同的车床设计、排量不同的发动机设计等。

(4) 测绘和仿制：按照国内外产品实物进行测绘，变成图纸文件，其结构性能不改变，只进行统一标准和工艺性改动。仿制是按照外单位图纸生产，一般只作工艺性变更，以符合工厂的生产特点与技术装备要求。

1.3.2 设计原则

(1) 创新原则：设计本身就是创造性思维活动，只有大胆创新才能有所发明，有所创造。但是，今天的科学技术已经高度发展，创新往往是在已有技术基础上的综合。有的新产品是根据别人研究试验结果而设计，有的是博采众长，加以巧妙的组合。因此，在继承的基础上创新是一条重要原则。

(2) 可靠原则：产品设计力求技术上先进，但更要保证使用中的可靠性，即无故障运行的时间长短，是评价产品质量优劣的一个重要指标。

(3) 效益原则：在可靠的前提下，力求做到经济合理，使产品“价廉物美”，才有较大的竞争能力，创造较高的技术经济效益和社会效益。也就是说，在满足用户提出的功能要求下，有效地节约能源，降低成本。

(4) 审核原则：为减少设计失误，实现高效、优质、经济地设计，必须对每一设计程序的信息，随时进行审核，决不许有错误的信息流入下一道工序。实践证明，产品设计质量不好，其原因往往是审核不严造成的。因此，适时而严细的审核是确保设计质量的一项重要原则。

1.4 部分现代设计方法简介

本书由于受到篇幅所限，只能先部分简单地介绍创新设计、计算机辅助设计、优化设计、可靠性设计、有限元法、动态设计、价值工程、并行工程、模块化设计、相似性设计、虚拟设计、三次设计、反求工程设计、工业艺术造型设计、人机工程等。在运用它们进行工程设计时，一般都以计算机作为分析、计算、综合、决策的工具。这些学科汇集成了一个设计学的新体系，即现代设计方法，它们包含了现代设计理论与方法的各个方面。

1. 创新设计

创新设计和传统设计虽然大多遵循同样的设计过程，但创新设计要求设计者在设计的全过程中不仅仅是重复和模仿，更要求设计者在新成果的基础上，敢于怀疑，有所突破，充分发挥设计者的创造力和想像力，充分利用已有科学技术的理论、原理、方法、技术进行创新构思，追求新奇、新颖、独特和非重复性的创造成果。特别是在总体方案设计中更要充分发挥设计者的主导性和创造性作用，使产品具有更强的竞争力。

产品创新设计的类型主要有三种：开发设计、变异设计和反求设计。

开发设计是一种从提出方案到完成设计全过程都是全新的、探索性的创新设计。

变异设计是在已有产品的基础上针对原有设计的缺点或不足，或针对新提出的要求，从工作原理、机构、结构、参数、尺寸等内容进行变化，开发适应新要求的新产品的创新设计。

反求设计是在消化、吸收了先进产品的关键技术基础上开发同类型新产品的创新设计。但无论是哪种设计，创新都要求设计者在设计的每一个环节上突破常规惯例，追求与前人、众人不同的方案，提出新原理、新方法、新机构、新形式、新材料等，在求异中寻求创新，将设计者的智慧具体物化在整个设计过程中。在创新设计的全过程中，创造性思维将起到至关重要的作用。深刻认识和理解创造性思维的实质、类型和特点，不仅有助于掌握现有的创造原理及由此而创造出的各种创新技法，而且能够推动和促进对新的创造方法的开拓和探索。

2. 计算机辅助设计

计算机辅助设计(Computer Aided Design)，简称 CAD。它是把计算机技术引入设计过程，利用计算机来完成计算、选型、绘图及其他作业的一种现代设计方法。CAD 是设计中应用计算机进行设计信息处理的总称。它应包括产品分析计算和自动绘图两部分功能，甚至扩展到具有逻辑能力的智能 CAD。计算机、自动绘图机及其他外围设备构成 CAD 的系统硬件，而操作系统、文件管理系统、语言处理程序、数据库管理系统和应用软件等构成 CAD 的系统软件。通常所说的 CAD 系统是指由系统硬件和系统软件组成，兼有计算、图形处理、数据库等功能，并能综合地利用这些功能完成设计作业的系统。CAD 是产品或工程的设计系统。CAD 系统应支持设计过程各个阶段，即从方案设计入手，使设计对象模型化；依据提供的设计技术参数进行总体设计和总图设计；通过对结构的静态或动态性能分析，最后确定技术参数；在此基础上，完成详细设计和产品设计。所以，CAD 系统应能支持包括：分析、计算、综合、创新、模拟及绘图等各项基本设计活动。CAD 的基础工作是建立产品设计数据库、图形库、应用程序库。

3. 优化设计

优化设计(Optimal Design)是把最优化数学原理应用于工程设计问题，在所有可行方案中寻求最佳设计方案的一种现代设计方法。进行工程优化设计，首先需将工程问题按优化设计所规定的格式建立数学模型，然后选用合适的优化计算方法在计算机上对数学模型进行寻优求解，得到工程设计问题的最优设计方案。

在建立优化设计数学模型的过程中，把影响设计方案选取的那些参数称为设计变量；设计变量应当满足的条件称为约束条件；而设计者选定来衡量设计方案优劣并期望得到改进的指标表现为设计变量的函数，称为目标函数。设计变量、目标函数和约束条件组成了优化设计问题的数学模型。优化设计需把数学模型和优化算法放到计算机程序中用计算机自动寻优求解。常用的优化算法有 0.618 法、鲍威尔(Powell)法、变尺度法、惩罚函数法等。

4. 可靠性设计

可靠性设计(Reliability Design)是以概率论和数理统计为理论基础,以失效分析、失效预测及各种可靠性试验为依据,以保证产品的可靠性为目标的现代设计方法。可靠性设计的基本内容是:选定产品的可靠性指标及量值,对可靠性指标进行合理的分配,再把规定的可靠性指标设计到产品中去。

5. 有限元法

有限元法(Finite Element Method)是以电子计算机为工具的一种现代数值计算方法。目前,该法不仅能用于工程中复杂的非线性问题、非稳态问题(如结构力学、流体力学、热传导、电磁场等方面问题)的求解,而且还可用于工程设计中进行复杂结构的静态和动力分析,并能准确地计算形状复杂零件(如机架、汽轮机叶片、齿轮等)的应力分布和变形,成为复杂零件强度和刚度计算的有力分析工具。

有限元法的基本思想是:首先假想将连续的结构分割成数目有限的小块体,称为有限单元。各单元之间仅在有限个指定结合点处相联接,用组成单元的集合体近似代替原来的结构。在结点上引入等效结点力以代替实际作用单元上的动载荷。对每个单元,选择一个简单的函数来近似地表达单元位移分量的分布规律,并按弹性力学中的变分原理建立单元结点力与结点位移(速度、加速度)的关系(质量、阻尼和刚度矩阵),最后把所有单元的这种关系集合起来,就可以得到以结点位移为基本未知量的动力学方程。给定初始条件和边界条件,就可求解动力学方程得到系统的动态特性。依据这一思想,有限元法的计算过程是:

(1) 结构离散化(即将连续构件转化为若干个单元);

(2) 单元特性分析与计算(即建立各单元的结点位移和结点力之间的关系式,求出各单元的刚度矩阵);

(3) 单元组求解方程(利用结构力的平衡条件和边界条件,求出结点位移及各单元内的应力值)。所以,有限元法的计算过程思想是"一分一合"。先分是为了进行单元分析,后合则是为了对整个结构进行综合分析。

近些年来,有限元法的应用得到蓬勃发展,国际上不仅研制有功能完善的各类有限元分析通用程序,如 NASTRAN、ANSYS、ASKA、SAP 等,而且还带有功能强大的前处理(自动生成单元网格,形成输入数据文件)和后处理(显示计算结果,绘制变形图、等值线图、振型图并可动态显示结构的动力响应等)程序。由于有限元通用程序使用方便,计算精度高,其计算结果已成为各类工业产品设计和性能分析的可靠依据。

6. 动态设计

结构动态设计是一项正在发展中的新技术,它包含的内容十分丰富,涉及现代动态分析方法、计算机技术、产品结构动力学理论、设计方法学等众多学科范围。目前,还没有形成一套完整的结构动态设计理论、方法和体系。

结构动态设计的大体过程一般是:对满足工作性能要求的产品初步设计图样,或对需要改进的产品结构实物进行动力学建模,并作动态特性分析。然后,根据工程实际情况,给出其动态特性的要求或预定的动态设计目标,再按结构动力学"逆问题"方法直接求解结构设计参数,或按结构动力学"正问题"分析法,进行结构修改设计与修改结构的动态特性预测,其结构的修改与预测过程往往需要反复多次,直到满足各项设计要求,从而得到一个具有良好静、动态特性的产品设计方案。因此,结构动态设计的主要内容包括如下两个方面:

(1) 建立一个切合实际的结构动力学模型。

(2) 选择有效的结构动态优化设计方法。

目前,结构动力学理论建模主要是采用有限元方法(Finite Element Method)建模。这种方法近20年来已有很大的发展,市场上也有许多成熟的软件可供选择,如国际流行的NASTRAN、ADINA、SUP-SAP、PAR Ⅱ等。它们已卓有成效地应用于航空、航天、船舶、汽车、机床等许多工程结构的动态分析。

7. 价值工程

价值工程从产品的功能研究开始,对产品进行设计,或重新审查设计图样文件,剔除那些与用户要求的功能无关的材料、结构、零部件等,代以更新的构思,设计出功能相同而成本更低的产品。

价值指的是事物的用途或积极作用。用户购买商品,最主要是购买商品的功能,设计中经常遇到的是使用功能和美学功能。产品成本是产品的各项生产费用的总和。降低成本要了解产品费用的组成,估算产品的制造费用,研究产品产量与成本、销售量与利润之间的关系,进行盈亏分析。价值=功能/成本。对功能定量赋值的方法很多,如对某产品的主要技术性能从理想值到超差值分为11级,对应于10~0分,即为功能值。又如按产品中各零件的不同功能计算功能成本等。

价值工程设计的基本步骤是:了解设计对象,明确要求的功能价值初评,制订改进方案,获得价值最高的新产品。

8. 并行工程

并行工程是集成地、并行地设计产品及其相关的各种过程(包括制造、后勤等)的系统方法。要求产品开发人员在设计伊始,就考虑产品整个生命周期中,从概念形成到报废处的所有因素,充分利用企业内的一切资源,最大限度地满足市场和用户的需求。

并行工程的目的在于寻求新产品的易制造性、缩短上市周期和增强市场竞争能力。要集中涉及产品寿命的所有部门的工程技术人员,组成并行设计组,共同设计制造产品,对产品的各种性能和制造过程进行计算机动态仿真,生成软样品或快速出样,进行分析评议、设计,取得最优结果,一次成功。利用计算机的数据处理、信息集成和网络通信的能力,发挥并行设计组的集体力量,将新产品开发研究和生产准备等各种工程活动,尽可能并行交叉地进行。这对换代快、批量不大的产品,能显著缩短周期、提高质量。

并行工程的内涵还包含了人的因素和企业文化。如果说,新产品按"设计—试制样机—修改设计—工艺准备—正式投产"的串行工程方法容易造成各自为政、效率低下的结果的话,并行工程则能改变企业组织结构和工作方法,促进人们之间的相互理解,激励积极性,提高协同作战的能力,塑造良好的企业文化氛围,形成一个适合人类发展需要的社会—技术系统。

9. 模块化设计

模块是具有一定功能和特定结合要素的零件、组件或部件。模块化产品是由一组特定模块在一定范围内组成不同功能或功能相同而性能不同的产品。设计模块和模块化产品,可以满足日益增长的多品种、多规格的要求。模块系统的特点是便于发展变型产品,更新换代,缩短设计和供货周期,提高性能价格比,便于维修,但对于结合部位和形体设计有特殊要求。

设计模块系统产品，先要建立模块系列型谱，按型谱的横系列、纵系列、全系列、跨系列或组合系列进行设计，确定设计参数，按功能分析法建立功能模块，设计基本模块、辅助模块、特殊模块和调整模块及其结合部位要素，进行排列组合与编码，设计基型和扩展型产品。模块系统的计算机辅助设计和管理，更显示了模块化设计的优越性。

10. 相似性设计

人们在长期探索自然规律的过程中，逐渐研究形成了自然界和工程中各种相似现象的"相似方法"、"模化设计方法"和相应的相似理论、模拟理论。相似方法是可以把个别现象的研究结果推广到所有相似现象上去的方法。相似理论是现象模拟和研究相似现象的基础。目前在大型复杂设备和结构设计过程中，一般都要在相似理论指导下，通过模化方法和模型试验，使方案取得合理参数，预测设备的性能。当前用计算机辅助进行相似性设计和代替模型试验，取得明显的效果。

解决相似问题的关键是找出相似系统各尺寸参数的相似比。根据各种物理现象的关系式推导出的由物理量组成的无量纲数群为相似准则。与相似准则各参数对应相似比组成的关系式称为相似指标。在基本相似条件和相似三定律的基础上，用相似准则、方程分析、量纲分析列出相似比方程，可求得相似比。

模拟设计是在开发新产品时，在相似的模拟工作条件下设计相似的模型进行试验；通过测定模型性能，预测产品原型性能，分析设计的可行性并进行必要的修改，进一步取得最优参数和结构。

产品系列设计是在基本型设计的基础上，通过相似原理求出系列中其他产品的参数和尺寸。设计步骤是：先设计基本型产品，确定产品系列是几何相似还是半相似，选择计算级差，求得扩展型产品的参数尺寸，确定系列产品的结构尺寸。几何相似的产品还可按相似关系以生产成本进行估算。

11. 虚拟设计

虚拟设计主要表现在设计者可以用不同的交互手段在虚拟环境中对参数化的模型进行修改。

一个虚拟设计系统要具备三个功能：

(1) 3D用户界面：设计者不再用2D鼠标和键盘作交互手段，而用手势、声音、3D虚拟菜单、球标、游杆、触摸屏幕等多种方式进行交互。

(2) 选择参数：设计者用各种交互方式选择或激活一个在虚拟环境中的数据修改原来的数据。参数修改后，在虚拟环境中的模型也随之变成一个新模型。

(3) 数据传送机制：模型修改后所生成的数据要传送到和虚拟环境协同工作的CAD/CAM系统中，有时又要将数据从CAD/CAM系统中返回到虚拟环境中。

这种虚拟设计系统中包含一个独立的CAD/CAM系统，为虚拟环境提供建造模型的功能。在虚拟环境中所修改的模型有时还要返回到CAD/CAM系统中进行精确整理和再输出图形。这种双向数据传送机制在一个虚拟设计系统中是必要的。

另一种虚拟设计系统是在虚拟环境中建造模型，不必要数据传送，称之为VR-CAD系统。

12. 三次设计

三次设计(Taguchi Method)是日本著名质量管理学家田口玄一于20世纪60年代创造

的一种设计方法。该设计法把新产品、新工艺设计分为三个阶段设计，故称三次设计法。第一次设计称为系统设计，即根据市场调查，规划产品的功能，确定产品的基本结构以及组成该产品的各种零部件的参数，提出初始设计方案。系统设计主要依靠专业技术人员的专业知识进行。第二次设计称为参数设计，即在专业人员提出的初始设计方案的基础上，对各零部件参数进行优化组合，求取最优设计方案，使得产品的技术特性合理，稳定性好，抗干扰性强，成本低廉。第三次设计称为容差设计，即在最佳设计方案的基础上，进一步分析导致产品技术特性波动的原因，找出关键零部件，确定合适的容差，进而确定容差，并求得质量和成本二者的最佳平衡。

13. 反求工程设计

反求工程(Reverse Engineering)是消化吸收并改进国内外先进技术的一系列工作方法和技术的总和。它对提高我国的科技和管理水平有着重要的意义。它是通过实物或技术资料对已有的先进产品进行分析、解剖、试验，了解其材料、组成、结构、性能、功能，掌握其工艺原理和工作机理，以进行消化仿制、改进或发展、创造新产品的一种方法和技术。它是针对消化吸收先进技术的系列分析方法和应用技术的组合。反求工程包括设计反求、工艺反求、管理反求等各个方面。

14. 工业产品艺术造型设计

工业产品艺术造型设计是指用艺术手段按照美学法则对工业产品进行造型工作，使产品在保证使用功能的前提下，具有富于表现力的审美特性。

造型设计的三要素中，使用功能是产品造型的出发点和产品赖以生存的主要因素，艺术形象是产品造型的主要成果；物质技术条件是产品功能和外观质量的物质基础。

工业产品艺术造型设计应遵循的原则有：体现高新科技水平的功能美，符合三化的规范美，显示新型材质的肌理美，体现先进加工手段的工艺美，表达各造型因素整体调和统一的和谐美，追求时代精神的新颖美，体现色光新成就的色彩美等。造型设计要有机地运用统一与变化、比例与尺度、均衡与稳定、节奏与韵律等美学法则。造型要寻求线型、平面、立体、色彩、肌理等造型因素的构成规律和变化，以效果图、动画、模型等形式，设计出物质功能与精神功能高度统一的创新产品。

15. 绿色设计的概念

绿色设计(Green Design，GD)，通常也称为生态设计(Ecological Design，ED)、环境设计(Design for Environment，DFE)、生命周期设计(Life Cycle Design，LCD)或环境意识设计(Environmental Conscious Design，ECD)等，由于这些方法的目标基本相同，都是设计和制造生命周期环境影响为最小的产品，因而，经常被互换使用。绿色设计是在产品整个生命周期内，着重考虑产品环境属性(可拆卸性、可回收性、可维护性、可重复利用性等)，并将其作为设计目标，在满足环境目标要求的同时，保证产品应有的功能、使用寿命、质量等。也就是说，要从根本上防止环境污染，节约资源和能源，关键在于设计与制造，不能等产品产生了不良的环境后果再采取防治措施(现行的末端处理方法即是如此)，这就是绿色设计的基本思想。

绿色设计是可以在不同层次上进行的动态设计过程。曾有人将绿色设计过程分为四个动态阶段：第一阶段为产品提高；第二阶段为产品再设计；第三阶段为产品功能创新；第四阶段为产品系统创新。绿色设计与创新是一个从部分到系统、从简单到复杂、从渐进创新到根

本创新的过程。也有人将绿色设计分为三个层次：第一层为治理技术与产品的设计，如“可回收性设计”(Design for Recycling，DFRC)、“为再使用而设计”(Design for Reuse，DFRU)、“可拆卸设计”(Design for Disassembly，DFD)等，其目标是简化、减少或取消产品废弃后的处理处置过程及费用；第二层为清洁预防技术与产品的设计，如“为预防污染而设计”(Design for Pollution Preservation，DFPP)、“为环境而设计”(Design for Environment，DFE)等，目的在于减少生命周期各个阶段的污染；第三层次是为价值而设计，目的在于提高产品的总价值，而这种价值体系是人与环境的共同体。

16. 智能设计

智能设计系统是以知识处理为核心的CAD系统。它将知识系统的知识处理能力同常规CAD系统的计算分析能力、数据库管理与服务能力、图形处理能力等有机地结合起来，从而可以协助设计者完成诸如方案设计、参数选择、性能分析、结构设计、图形处理等不同阶段、不同复杂程度的设计任务。一个完善的智能设计系统一般应包含知识处理、计算分析、数据管理与服务以及图形处理等四大基本功能。

(1) 知识处理功能。设计过程是人的思维过程，是设计者综合运用自己所掌握的知识，并通过分析、计算、推理、判断、决策等思维方式获取满意设计结果的过程。因此，知识处理是智能设计系统的核心，它实现知识的组织、管理及其应用，其主要内容包括：

① 获取领域内的一般知识和领域专家的知识，并将这些知识按特定的形式存放于系统中或外部存储器中，以供设计过程使用。

② 对智能设计系统中的知识实行分层管理和维护。

③ 在设计过程中根据需要提取在外部存储器中的知识，实现知识的推理和应用。

④ 根据知识的应用情况对知识库进行优化。

⑤ 根据推理效果和应用过程学习新的知识，丰富知识库。

对智能设计系统而言是由设计型专家系统实现的。

(2) 分析计算功能。设计过程一般包含大量的分析计算，如设计对象的性能分析、强度校核、动态分析等。分析计算结果可以为设计者提供推理、判断和决策的依据。因此，一个完善的智能设计系统应提供丰富的分析计算方法，主要包括：

① 各种常用数学分析方法；

② 优化设计方法；

③ 有限元分析方法；

④ 可靠性分析方法；

⑤ 各种专用的分析方法。

上述分析方法以程序库的形式集成在智能设计系统中，根据设计者的需要选用，因而可以极大地提高智能设计系统的分析计算能力。

对某个特定的智能设计系统而言，可能并不需要上述所有分析计算方法，但功能强大的分析计算方法库对扩充智能设计系统的设计能力是必不可少的。

(3) 数据服务功能。设计过程实质上是一个信息处理和加工过程。大量的信息，如初始输入信息、中间生成信息、输出结果信息等以不同的数据类型和数据结构形式在系统中存在并根据设计需要进行流动，为设计过程提供服务。随着设计对象复杂度的增加，系统要处理的信息量将大幅度地增加。为了保证系统内庞大的信息能够安全、可靠、高效地存储和流

动，必须引入高效可靠的数据管理与服务功能，为设计过程提供可靠的服务。

(4) 图形处理功能。图形是设计对象量直观的表现形式，尤其是三维实体图形，能更清晰地表达产品的几何形状、结构特征、装配关系。借助于二维、三维或三维实体图形，设计者在设计阶段便可以清楚地了解设计对象的形状和结构特点，还可以通过设计对象的仿真来检查其装配关系、干涉情况和工作情况，从而确认设计结果的有效性和可靠性。

因此，强大的图形处理能力是任何一个 CAD 系统都必须具备的基本功能。图形处理功能由各类图形支撑软件实现。

CAD 系统中广泛使用的图形支撑软件有在微机上使用的 AutoCAD、CADKEY 和在工作站上使用的 I.DEAS、UG Ⅱ、Pro/Engineer 等。这些软件能绘制机械、建筑、电气等领域内的一般性图样。随着计算机软硬件技术的飞速发展，许多微机上使用的通用绘图软件都增加了新的功能，如 AutoCAD 从 12 版本开始增加了三维造型功能(AME)，最近又增加了三维实体功能(MDT)。

MDT 可以直接生成三维实体，并可由三维实体生成二维图样，还具有三维实体装配功能。此外，许多工作站上使用的功能强大的图形软件也移植到微机上，如 LDEAS、Pro/Engineer。

习 题 1

1. 试述设计的含义，工业产品设计具有哪些特征？
2. 何为传统设计、现代设计？
3. 传统设计和现代设计之间的区别在哪几个方面？说明这两种设计之间的正确关系。
4. 简述现代设计方法的主要组成部分及其特点。
5. 简述产品开发的基本阶段和产品设计的基本进程。
6. 试述你知道的现代设计方法。
7. 试述学习现代设计方法这门课程的意义与任务。

第2章　创新设计

本章从科技创新学研究的内容与方法出发，先后对创新思维基本方法、创新思维技法、机电产品创新设计等进行介绍，并通过一些实例来具体地说明机电产品创新设计的原理与应用方法。

2.1　创新学研究的内容与方法

20 世纪是知识不断创新、科技突飞猛进、世界深刻变化的世纪，21 世纪科技创新将进一步成为社会和经济发展的主导力量。世界各国综合国力竞争的核心，是知识创新、技术创新和高新技术产业化。创新是一个国家国民经济可持续发展的基石。对于一个国家而言，拥有持续创新能力和大量的高素质人力资源，就具备了发展知识经济的巨大潜力。缺乏科学储备和创新能力的国家，将失去知识经济带来的机遇。江泽民同志根据世界发展趋势的主要特点，及时指出“创新是一个民族进步的灵魂，是一个国家兴旺发达的不竭动力”。

创新学是研究人类创造发明活动及其一般规律的科学，其宗旨在于通过对人们创新活动和创新发明方法的研究，揭示人类创新发明的一般规律，并用以有效促进人们的各种创造和发明，促进科学技术进步和社会生产力的发展。通过创新学的学习和研究，将会使人们原来误认为十分神秘的、似乎只有科学家或艺术家才独有的创新能力，最终成为每一个普通人也能够被开发出来和运用的一种能力。

2.1.1　创新学研究的内容

创新学的研究内容包括创新活动、创新过程、创新者的人格因素和心理品质、创新力及其开发、创新思维、创新环境、创新性人才培养和创新评价等。

1. 创新的涵义与创新活动

人的实践活动永远离不开创新(也可以说创造)。但是，用科学语言回答什么是创造却极为困难，因为创造包括的外延太多。根据逻辑学原理，它的内涵就很少。所以，《韦氏辞典》英文版干脆将创造定义为“赋予存在”。在我国的《词源》中，“创”字有疮、伤、损、惩的意思，其共同涵义是“破坏”；“造”字有作、为、始、成的意思，共同涵义是“建设”；两字合起来，创造就是“破旧立新”，即“创新”。

国外学者对“创新”一词从不同角度下了许多定义。例如，日本的思田彰认为“创新”是依据异质的信息或事物与至今未有的方法结合起来，产生新的有价值的东西。美国学者欧文·泰勒认为创造包括五个层次，第一层次是表达式创造，这是最初级的创造，像孩子绘画就属于这类；第二层次是生产创造，指的是发展各种技术得到完美产品的活动；第三层次是发明创造，指发明家寻找新方法来解决现存问题，即技术发明；第四层次是创新式的创造，指对各种原理、原则和概念的洞察；第五层次是深奥的创造，指经过长期苦心钻研得到崭新原理的活动。

国内学者对创造也有不同的定义。例如,《发明创造的艺术》一书的作者认为,"创造就是人们利用自己的聪明才智对已有的物质或精神材料进行加工,从而产生前所未有的有价位的物质产品与精神产品"。

作者认为,"创新"或"创造"虽然从语言逻辑定义上有所不同,但是,从科技创新的角度上看,其实就是"破旧立新"。当然从专利的角度上讲又分成"发明创新"(发明专利)、"变异创新"和"组合创新"(新型实用专利)等。所以,所谓创新,是人的主观能动性的高度发挥,是为满足社会物质生活和精神生活的需要,在破旧基础上的立新。创新既是一种有过程的活动,又是一种行为,也是一种具有新颖性、独特性的成果。

例如,美国一家公司在市场调查时发现,人们使用其生产的电熨斗时感到不便的是它拖着一根电线,于是就决心割掉这根"尾巴"。经过研究终于研制出了世界上第一台不带电线的电熨斗。新电熨斗由熨铁和一个连接电源的熨铁架两部分组成。使用前,将熨铁放到熨架上,当加热到一定温度时指示灯就会亮起来,这时就可以使用电熨斗了。而熨铁每加热一次能使用一分钟,而加热一次仅需 6 秒。后来德国一家公司又推出一种感应式电熨斗,即在熨衣板上设有感应电源,熨衣时自然产生温度,且温度可调,使用很方便。这种不断的技术进步就是"创新",或者说"科技创新"。

因此,本书认为,创新其实即是一种活动,一种社会活动。对于主体人而言,创新与创新活动含义是相同的。那么,究竟什么是创新活动呢? 创新活动,是指人们所从事的各种具有"新颖性"的活动。这里所指的"新颖性",有两个不同层次的含义,本文所说的"创新"是指"非重复性"活动,这是一条重要的判断标准。比如一个大学生通过自己的思索而做出了一种别人早已发明成功的"自动鞋刷",这种仅仅对于创新者自己来说是新颖的,即所谓相对新颖性,不属于"创新";而像爱迪生发明电灯,这种"新颖性"对于其他人甚至对于全人类来说都是新颖的,是所谓的绝对新颖性,我们称其为"创新"。

创新活动既然是一种社会活动,那么它就不可能离开社会实践,更不可能不对社会产生一定影响。于是,根据创新活动对社会的影响效果,创造活动大致可划分为正向创新活动和负向创新活动两大类。凡是有利于(或者至少无害于)社会发展、符合社会公德的创新活动,可称为正向创新活动,如哥白尼日心说的创立、核电站的诞生、电视机的问世、拉链的出现等。相反,凡是不利于社会发展、违背社会公德的创新活动,则称为负向创新活动,如从事那些我国专利法中明文规定的"违反国家法律、社会公德或者妨害公共利益的发明创造"的活动以及互联网上各类"黑客"的创新活动等。应该指出,在某些情况下,就创新本身而言是难以明确判断其创新活动的正向性或负向性的,如原子弹及各类武器的发明等。甚至,有些内容完全相同的创新活动在不同时期、不同地点、不同社会背景之下,还可能具有正负向互相转化的趋势。因此,我们极力主张人们做有利于社会发展,造福于人类的"创新活动"。

2. 创新过程

创新过程,有完整的创新过程和不完整的创新过程两大类。

(1) 完整的创新过程——即一般人们所说的创新过程,它包括选题过程、分析思维过程、实施过程和运用各种方法解题过程等。据此,人们已建立了创新过程的程序模式,如杜威(J. Dewey)的创新五阶段模式、沃勒斯(G. Wallas)的四阶段模式以及其他学者的三阶段模式、七阶段模式等。

这里介绍一下人们常用的沃勒斯的四阶段模式。沃勒斯认为,无论是科学的或艺术的

创新，一般都要经过以下四个阶段：

第一阶段（即准备期）：主要指发现问题，收集有关资料，参考别人或前人的知识、经验并从中得到一定启示等；

第二阶段（即酝酿期）：这一阶段主要是冥思苦想，对问题做各种试探性解决；

第三阶段（即明朗期）：是指在上一阶段酝酿成熟的基础上豁然开朗，产生了灵感或顿悟；

第四阶段（即验证期）：即对灵感或顿悟得到的新想法进行检验和证明。

应该说，上述四步与许多学者所称的创新性思维的四个阶段大体相同，这里似乎混淆了创新性思维过程与创新活动过程的界限。实际上创新学中所指的创新过程，还应包括创新的实践过程。

（2）不完整的创新过程——即突发性创新过程，由于这种创新过程非常短暂，所以人们一般很难再将其区分若干个阶段。例如，早年在足球比赛时裁判并未使用哨子。有一次在伦敦比赛，运动员为了一个得分而发生争执，继而观众涌入场内，顿时秩序大乱。当时的裁判恰好是一个警察，处于职业习惯（即联想思维的结果），他灵机一动，从口袋里掏出警笛吹了起来，球场立刻安定下来。由此，人们便发明了足球比赛用哨。其实，我们一般所见的“灵机一动”的创新，很多都反映不完整的创新过程，它们多与人的联想思维关系甚为密切。不完整的创新过程更加增添了创新的神秘色彩，因此，更值得人们去深入研究。

3. 创新者的人格因素和心理品质

创新者的人格因素，是指创新者个人的连续性、持久性心理面貌或心理“格局”。它包括人的性格、品格和体格等等。创新学要研究创新者需要具备什么样的人格和心理品质才更有利于创新，研究一个人应当如何培养自己的创新性人格和心理品质以及如何克服创新过程中的心理障碍因素等。

4. 创新力及其开发

人们几乎天天讲创造力，但真正追究其概念来，往往说不清楚。长期以来对什么是创造力的问题，人们曾经提出各种各样的答案。在 20 世纪初以前，普遍认为创造力是一神秘的现象，是极少数天才人物所具有的特别禀赋。现代当然否定了上述认识，每个正常人都具有一定程度的创造力，只是由于一系列主客观因素的影响，创造力的现实才表现出个体的差异。美国著名学者马斯洛认为，人的创造力有两种层次的表现，一种表现，我们可以在日常生活中看到。例如，任何一位工匠手艺的独特风格；一位家庭主妇的佳肴烹调技术和房间布置的美观雅致。另一种是高层次的创造力，表现为学术思想、科学研究、文学艺术和管理工程等复杂问题的创造。他认为前一种创造是自发的创造，后一种创造是特殊才能的创造。前者是初级的、基本的，后者不仅具有自发的特性，而且需要艰苦的工作，长期的训练等。

美国创造学家阿玛拜尔认为有三种类型的创造力：一是一般的创造才能，指在一切领域都起作用的，最普遍的创造能力，这种能力强的人在许多领域都会表现出创造性；二是特殊的创造才能，指具有与特定活动领域有关的创造才能，如音乐才能、绘画才能等；三是更加专门的创造才能，它往往与人的动机、兴趣等结合紧密，只是对同一领域的某一课题表现出的高度创造力。

奥斯本认为，创造力是通过想像提出新设想、创造新事物、发现和解决新问题的能力。

日本索尼公司研究中心主任菊一在总结索尼公司成功的经验时，也提出了两种不同模

式的创造力。他认为索尼成功的奥秘是日本人的勤奋刻苦工作同一种日本民族特有的“创造力”相结合。这种创造力同美国人所说的创造力概念迥然不同。只有具备这种创造力才能在运用于电子技术上成效特别显著。菊一把爱因斯坦那种天才式的突发灵感称为一种“独立不羁的创造力”。而日本民族最擅长的是“推陈出新的创造力”,即善于集中智慧,解决那些基本性质已有定论的问题。总之,创造力是一种复杂的混合物,是创造者通过创造行为表现出来的各种积极的心理特征的总和,是创造者智、情、意、体、美诸因素的结合。它有不同类型、不同层次,由各种特性或成分构成。其中有些起主导作用,有些起支柱作用,而另外一些则构成有效的创造行为所必需的背景。一种单独分出的孤立的能力要素,如记忆力、想像力、观察力等,即使它们达到非常高的发展水平,表现得非常明显,也不能当做创造力本身。通常创造力受下列因素影响:

(1) 创造力与智力的关系。人们在现实生活中经常谈论智力问题,如说“这个孩子灵”、“那个孩子笨”等。智力、智能和智慧,虽然用词不同,但其含义都是指人的聪明才智。一般而言,智力主要还是指与人学习有关的能力,智力水平往往是一个人学习能力大小的标志。比如智力的五大要素——观察力、注意力、记忆力、想像力、思维力,基本上指学习的速度、深度、广度和精确度;而创造力则主要指干预外界事物、创新和改变存在的能力。创造力更具有主动性、冒险性和灵活性。例如,学生经常被问到这么一道古老的智力题:树上有 10 只鸟,打死 1 只,还剩下几只? 回答“打死 1 只,还有 9 只”的学生被认为是最不聪明的;而回答“打死 1 只,就 1 只也没有了,因为它们都被吓跑了”的学生被认为是最聪明的。这条答案也成了唯一正确的答案,并把它作为考核其他学生的试金石了。表面上看,这种智力问题对学生的智力发展有刺激作用,但实际上,这种追求唯一答案的结果束缚了儿童创造力的发挥。

(2) 创造力与知识的关系。知识是智力的基础,智力又是创造力的基础,所以知识也是创造力的基础。儿童可能在许多方而表现出较强的独创性,但由于缺少知识,有些设想只是空想而无实用性,只能说是前创造力,还没有发展成真正创造力。而我们讲的科技、管理、艺术等方而的创造力,它要依赖于专门知识。然而,知识多又不等于创造力就高,有知识未必顺利完成任务。知识与创造力是不能相互取代的。这是因为,首先,知识只能迁移到与其内容相似的场合中去,与知识适用领域不相似的场合,原知识是无能为力的,而创造为个人特点时,就可以迁移到不同的场合,在极广泛内发挥作用。而往往创造力通常是随一个人的知识的不断积累而加强,但是知识多有时也会阻碍人的创造力发挥。电话的发明人贝尔并不是电气专家,而是一位聋童教育家和语言学教授,与他同时产生电话设想的还有一位叫可格雷的有专业训练的电气专家。可格雷由于头脑传统知识太多,竟莫名其妙地认为一个送话器只能送出单一频率的振荡信号,因此,传送人的话音就必须由许多架不同频率的送话器协同工作,这就太复杂了,所以没有坚持搞下去。相比之下,贝尔却没有那么多顾虑,尽管他走了许多弯路,还是首先发明了电话。美国创造学家道格拉斯曾说过:“孕育了发明能力的小学毕业生,远比扼杀了发明创造能力的哈佛大学毕业生有更多的成功机会”。此话是对创造力与知识之间相互关系的绝好说明。

(3) 创造力与创造动机的关系。研究一个人的创造行为,很自然就要考虑到导致它的直接原因——创造动机,常听到有人问:“是什么力量激励某人去克服重重困难,努力去创造的呢?”这就是在寻找创造动机。所谓创造动机,是指激励着人们去进行某种创造的内在力量,是产生创造行为的动力基础。创造动机推动人从事创造,表现出创造行为,导致创造结

果，因而反映出创造能力的潜在动力。它是一种内隐变量，看不见摸不着，也无法直接测量。但凡创造，总是为某种动机驱使，如有的人是看到生产存在的问题，产生了解题的紧迫感；有的人是对工作本身感兴趣；有的人受榜样力量的鼓舞；有的人是为了谋取个人物质利益；有的人是为提高声誉、威望等。例如，19 世纪以前，天花病是人类疾病中最可怕的一种，死亡率达 10%。那时的医生想不出任何治疗办法时，爱德华·琴纳学医后开办了一家医院，他作为一名医生，为了解救天花病人的痛苦驱使他产生了强烈的创造动机，立志要找到一种治疗和免疫天花的办法。在此动机支配下他勇敢地打破了当时医学界的传统偏见，多次冒危险进行接种实验，终于发明了轰动医学界的接种牛痘的免疫方法，征服了人类的天花病。

此外，经济收入也常常成为人们进行创造的间接动机，这在资本主义社会尤为明显。美国著名电子学家、电子琴的发明人科克在《创造性工程师》一书中，论述了资本主义国家发明家的发明动机，将金钱、声誉放在首位，为祖国作贡献则成了较次要的。当然，在社会主义社会中，经济收入也会成为一些人从事发明创造的动机之一，但我们更提倡树立为国为民造福人类的创造志向，应更多地宣传为祖国建设而去创造发明的模范人物，以产生激励作用。

直接动机是指能直接促使人们去创造的动力，主要包括求知欲、好奇心、挑战心理、创造兴趣和对创造的自豪感等。

求知欲和好奇心都属于人的天性，只不过有的人把这种特性保持下来，而有的人则将其压抑，很多人正是由于强烈的求知欲和好奇心驱使才终生进行创造发明的。物理学家玻恩曾说："我一开始就觉得搞研究工作是很大的乐事，直到今天仍感到是一种享受。"阿基米德在领悟到浮力定律时兴奋得在大街上边跑边喊的故事就是一个典型例子。此外，心理学研究表明，每个人都或多或少都具有挑战心理，即争强好胜，愿意战胜别人，这种挑战心理也常常是发明创造的直接动机，为创造提供一种较大的内在动力。

(4) 创造力与个性的关系。个性是人的态度和行为方面比较稳定的心理特征。个性不是指那种在偶然场合的特殊态度和行为，而是经常性的、习惯化的态度与行为。它是在个人与环境的相互作用过程中形成的，是定型和成熟的。人的个性是通过其外部表现——性格特征而被人们认识的。心理学研究表明，人的个性对其创造力有重要影响，通过大量调查发现，有突出创造性成果的发明家、科学家在个性特征上有许多共同或相似之处。

美国心理学家高夫曾设计过这样一个调查：他编制了有 300 个形容词的测验表，对被各个领域的专家公认具有很高创造力的 12 组共有 1701 位受试者测验。结果发现，他们都普遍选择以下 15 个反映性格特征的形容词，它们是：有才能的、聪明的、有信心的、自我中心的、幽默的、个人主义的、不拘礼节的、有见识的、理智的、兴趣广泛的、创造性的、喜欢沉思的、机智灵活的、自信的、不守常规的。尽管这项研究只是针对美国的创造发明者，并不一定符合我国的情况，但仍然有其代表性。从大量统计资料研究得出的结论是，有利于创造的个性特征普遍有以下几个方面：

① 强烈的好奇心与求知欲。富于创造的人，往往从小到大始终有很强的好奇心与求知欲，见到新东西总要问"为什么"、"怎样"一类的问题，总想动手试试。电话的发明人贝尔在小时候就对大自然产生极大兴趣，他曾解剖过许多小动物，了解它们的身体结构；动手试验当时刚刚出现的照像术；帮助同学的父亲改水磨；甚至在他生命的最后一年，还在研究水翼拖靶船和太阳能热水器等。具有这种个性品质的人，往往有一种谦虚的态度，能倾听别人并承认自己并不是万事皆知的，以不带成见、不过早下结论的态度来看待世界。

② 独立性与自我精神。许多有创造性的发明家、艺术家的独立性与自主精神极其突出，他们产生一个有意义的设想后，往往会主动努力加以实现，既不等待别人的吩咐，也不愿别人过多管束，如果他们无可奈何地落入一种非常拘束的环境，他们会感到心理上的压抑，甚至丧失创造力。这种独立性与自主精神能够形成一种特有的思维方式，往往表现出思维的独创性和自由性，较少地受权威的影响。美国发明家贝利在 20 世纪 50 年代初，曾参加过一个 $6AJ_4$ 射频放大管的发明小组。他们在接受任务的同时，接受了一条命令，经理要求他们任何人不许查看和参阅任何书。结果他们试制成功了这种小功率频率高达 1000MHz 的放大管，用在超高频(UHF)电视波段。成功后他们查看书本倒大吃了一惊，因为书上写着玻璃管子的极限频率是 250MHz。贝利事后说，如果我们事前看了书，一定会怀疑我们是否能造出这种放大管。

③ 喜欢怀疑和冒险。创造力高的人，对传统见解、权威结论或他人的观点常常具有怀疑的精神，但是他们尊重的是事实而不是权威。由于他们具有较强的批判性，也常常会令人难堪或下不了台，因而有时人缘不好，使别人以为他们过于骄傲或清高，在大多数人错时，他们往往是对的。16 世纪以前，大多数人都认为地球是宇宙的不动的中心，太阳等其他星球围绕地球运动。而波兰天文学家哥白尼却大胆地向这种传统观点挑战，提出了日心说，其命运是可想而知的，但最终事实却证明了他是对的。当然，有时也会出现相反情况，即人们是对的，而他们是错的。像爱迪生这样的天才也免不了犯类似的错误。在他发明电影的前身——"西洋镜"后，许多人都劝他把它与幻灯结合起来，变成电影机，可他却认为这是一种没有实用价值和前途的东西。当类似的电影机被别人发明时，他才后悔。据说牛顿曾请瓦匠砌围墙，要求在墙上开一大一小两个猫洞，但瓦匠只开了一个大洞，牛顿很不满意，瓦匠说，小猫也可以由大洞进出，牛顿这才恍然大悟。

另外，创造与冒险总是紧密相联的。创造力强的人常喜欢做一些没把握的事情，那些四平八稳的事对他们没有多少吸引力，他们喜欢智力上的挑战，以解决别人解决不了的问题为极大兴趣，不怕花必要的代价。爱因斯坦建立相对论后，几乎用了后半生的时间研究统一场，终因条件不成熟而失败。对于发明家来说，失败可能意味着丧失自己的财产甚至负债累累。我国木工安全刨的发明人李林森也是在负债累累、变卖家产的冒险精神下完成其发明的。

冒险精神使许多创造发明家获得成功，也使许多尝试者体验了失败的辛酸。人们看到获得巨大成就的人只是冰山的顶而更多的是水面下大得多的冰块——失败者。成功者之所以能成功，是因为他们有更大的冒险精神，并以不计其数的失败为代价换来的，所以，冒险精神和献身精神往往是决定一个人创造力的一项重要品质。

5. *创新性思维*

创新性思维，可以说是创新的核心。现在的创新学主要是从行为学角度来认识人类创新性思维的特点、创新性思维的思维形式、思维方式、思维模式及其思维的规律，同时也研究如何对一般人进行更有效的创新性思维培养和训练。

6. *创新环境*

创新总是需要一定环境条件的，同样都是中华民族的桥梁专家，隋朝的李春只能建造大石拱桥——赵州桥，而现代的茅以升却能主持建造钱塘江大桥，显然，这与他们所处的不同社会环境有关。因此，创新学要研究什么样的环境适合于创新性活动的开展，什么样的环境

最有利于发挥自己的创造才能等。大量事实表明，创造需要一定的环境，而环境又需要人们去创造，只有尽力创造一个适合于创新者创新的环境，才能更好地开展创新活动。由此，可从创新学中分出一支环境创新学来。

7. 创新性人才培养

创新学的最终目的是培养人的各种创新性，由此出发，创新学可同教育学紧密结合，形成独特的教育创新学（一般亦称为创新教育）。因而创新学的研究内容必然还要涉及创新性人才的培养目标、培养条件、培养原则和培养方法等。

8. 对创新的评价

（1）对个人创新性的评价。现在，对于个人创新性的评价主要是通过所谓的“创新力测试”来进行的。然而，目前还缺乏有关这类测试的基础理论，所以至今未见到令人满意的以创新力测试成绩来评价人的创新性的研究成果。

（2）对创新成果的评价。由于人的大脑生来就不喜欢新奇的设想，就像身体不喜欢新奇的蛋白质一样，总会竭力抗拒新生的具有创新性的事物，这就造成了对创新性成果评价的困难，甚至一些专家权威在评价中扼杀杰出创新性成果的现象也屡见不鲜。相比之下，对于一些实物性创新性成果，如遥控彩电的出现、全自动洗衣机的问世等，人们是比较容易通过实践来进行正确评价和肯定的，但是由于对于一种新颖的创新思想、一些新颖的学术观点，如对恐龙灭绝的全新性认识等，评价起来就比较困难了。而最困难的事情，有时则是对创新者个人的正确评价。

（3）对创新者的评价。由于受各种因素的干扰，人们最初对一个创新者的评价往往都不是很高。例如，数学家华罗庚小时侯就曾被人称为“华呆子”；爱迪生也因孩提时代向老师提出为什么三加二等于五之类“显而易见”的问题而被老师称之为“笨蛋”并被开除出校；革命导师恩格斯中学时期的班主任“早就料定恩格斯今后不可能有所作为”；等等。

总之，对于创新的评价是较为困难的。这就使很多优秀的创新成就在一个阶段或者很长时间里难以问世，甚至被扼杀在摇篮之中。因此，如何科学地评价人们的创新性及其创新成果，必然是创新学的重要研究内容。

2.1.2　创新学研究的方法

由于创新学是一门新兴学科，而且又与其他自然科学、社会科学等各种学科横向交叉、内外渗透，因此，创新学的研究方法与其他一些学科的研究方法是密不可分的。目前来看，创新学的研究方法主要有如下一些：

1. 观察法

所谓观察法，是指在一定条件下有目的、有计划、系统地观察一个人的创新过程，并依据其行为、言语、性格特征及情绪变化等诸方面，分析其创新心理、创新性思维和自觉或不自觉地运用某些创新原理的一种研究方法。这种研究方法也包括对已经取得创新成就的科学家、发明家或艺术家进行采访及录像、录音等。观察法也是其他学科普遍采用的一种研究方法。在运用观察法时，应坚持观察的客观性，忌讳猜测，应坚持观察的全面性，防止偏见。当然，在观察时还要注意被观察对象的代表性，即要选择典型事物进行观察，这样可简化观察对象以排除次要和偶然的因素干扰。

2. 传记法

古今中外专门记载人们创新活动、创新性思维和创新方法的文字确实太少，但有关人物特别是有关科学家、发明家的人物传记却能见到。因此，专门研究人物传记遂成为研究创新学的一种很好的方法和途径。通过对人物传记特别是自传的研究，可以了解科学家、发明家或文学艺术家的具体思维过程、创新过程和成长过程，亦可研究他们的创新性人格特征及其所处的环境等。如法国数学家阿达玛(J. S. Hadamard)，他从有关传记中检索出关于数学家成功的事迹后，1945 年出版了专著《数学领域中的发明心理学》。

3. 科学史法

所谓科学史法，是指通过研究某一学科(技术)内部新、旧知识之间的产生、变化、发展和消亡的过程来揭示该学科(技术)自身发展规律的研究方法。由于一门学科(技术)发展的规律在一定程度上可反映创造者的许多创新活动，所以详细研究一门或几门学科(技术)的科学史、发明史，就可以在一定程度上揭示人们的某种创造规律。这一方法也有助于了解不同发明创造之间的内在联系。例如，在地质学中，全球板块构造学说的提出，即经历了“大陆漂移”——“海底扩张”——“全球板块”这样的三步曲，从每一步发展到下一步都蕴含着地质学内部的某些创新规律。

4. 比较研究法

这是一种通过对不同创新过程的比较研究，通过对不同发明家、科学家或艺术家创新人格、创新性思维等的比较研究而深入了解有关创新问题的方法。例如，达尔文(C. R. Darwin)通过生物进化比较研究率先提出了进化论的观点。

5. 调查征询法

调查征询法，就是把创新学要研究的问题分解为详细的纲目、拟成简明易答的问题即《征询表》分发给征询对象以征求答复，然后回收《征询表》并利用各种数学方法进行统计研究而得出结果的研究方法。例如，在 20 世纪 80 年代，我国著名心理学家王极盛就曾在这方面做了极有意义的工作。他有目的地对我国当时的 28 位学部委员和 127 名一般科技工作者的创新性作了分门别类的系统征询和研究，于 1986 年完成了近 40 万字的专著《科学创造心理学》，为我国创新学的研究作出了突出贡献。又如，作者在机电一体化专业开设“机电产品创新设计”课程，为了了解学生对创新设计课程的反应发出了包括 20 多个栏目的“征询表”，从学生中得到了 900 多个反馈信息数据，即为创新性教学和创新学研究积累了宝贵的素材。

6. 测验统计法

创新学研究经常需要进行一些创新性测试，以便对不同的人或者同一个人在不同时期中的创新性作出直观的、定量性的评价，从而探索创新学的某些规律。例如，作者就曾对本校机械专业班的学生进行过两次(刚入学时和毕业前夕)专门的创新性测试，了解入学时和毕业时学生创新水平的变化，取得了部分可供研究的第一手实际资料。然而，由于有关创新性测量的理论研究相对滞后，因此其测试结果只能代表某种倾向，而难以完全加以科学定论。

7. 创新案例研究法

创新案例研究法，是指通过对一些典型创新实例的成果进行全面的创新性分析，从中探寻出创新者的若干创新规律的研究方法。例如，通过对双体自行车实物进行分析可得出同

类组合的创新规律，通过对旧轮胎的充分利用可分析出完满创新原理等。在分析中，虽然研究者对原创新者的创新过程特别是创新性思维的过程知之甚少，甚至连原创新者自己也可能谈不出其中的原委，但这并不影响研究者对其创新成果进行全面的创新性分析。事实表明，从创新实例研究中经常可探索出一些重要的创新规律来。

总之，创新学的研究方法应当是不断改进和发展的，我们提倡用创新的精神灵活运用上述种种创新学的研究方法，并逐步使之更加科学、实用和完善。

2.2　创新思维基本方法

2.2.1　创新思维的基本特征

创造性思维应该具有独立性、求异性、想像性、新颖性、灵感性、潜在性、敏锐性、预见性、连动性等基本特征。

由于人们对于自身的创造性思维往往了解得不多、研究得不够深入，因而对于创造性思维特点的看法也极不一致。从大量科学统计发现，不同研究者在其论著中所表述的创造性思维的“特点”大相径庭。创造性思维的特点就是“思维结果的新颖性”。据此，我们可对创造性思维下这样的定义：“创造性思维就是能产生新颖性和相对新颖性两类。其新颖性程度越高，那么其创造性就越强”，由此可见创造性思维是创造活动的核心。

但是虽然创造性思维是创造活动的核心，是创造活动中不可缺少的部分，然而创造活动毕竟是一个完整的活动过程，因而仅仅依靠创造性思维也是难以完整地完成创造活动全过程的，甚至是难以完整地完成其中的思维活动全过程。这是因为，创造活动的思维过程之中必然有逻辑思维的介入，同时这种逻辑思维也必然要与创造性思维发生一定的联系。

(1) 人类的知识和经验总是不断发展的。在发展的某一阶段，当人类的知识和经验积累到一定程度时，就会导致原有逻辑的矛盾，而要解决这些矛盾，就需要非逻辑思维特别是创造性思维来加以完成。

(2) 创造性思维一旦突破原有的逻辑，就必然会在更高层次上上升为新的逻辑思维，并把新的知识、新的发现纳入到已知体系之中继而作为已有知识形式而保留下来。如果创造性思维不在更高层次上与逻辑思维相融，那么这些新的发现、新的经验、新的突破就不会上升为知识和理论，因而就不可能广泛地指导人们认识世界和利用世界。比如，19 世纪想制造飞机的人，大多数都被人视为“疯子”或“狂人”，而且，当时的物理定律和数学公式似乎也都是为“飞机不可能飞上天”服务的，以致连一些大科学家、大发明家也都“逻辑地”论证飞机不可能飞上天。然而，在飞机发明成功以后，这些“异想天开”的创造性思维很快便上升到了新的逻辑思维高度并继而转化为新的知识。现在的知识和逻辑不仅能够正确地说明飞机可以在天空飞行，而且还可进一步论证其可靠性和稳定性是如何好。

因此，由逻辑思维→创造性思维→更高层次的逻辑思维→更高层次的创造性思维……这样无穷地发展下去，便构成了人类常规活动→创造性活动→更高层次上的常规活动→更高层次上的创造性活动……的螺旋式上升发展。这就是逻辑思维和创造性思维在人类创造性活动中的密切联系，这也许就是人类认识自然和利用自然的一条基本规律。

(3) 既然创造性思维是能产生出新颖性思维结果的思维，那么创造性思维的形式就不可能只限于一种，相反，也不能简单地认定某一种思维形式只能形成或不可能形成创造性思维，比如创造性思维既可由联想思维的形式形成，亦可由灵感思维的形式形成；反之，联想思维和灵感思维所产生的结果也并不全部都是新颖的，因而它们也并不完全都是创造性思维。例如，看到天上飞的鸟后联想到蝴蝶也能飞，这种联想就没有什么新颖性。所以说，并没有哪一种思维形式全部能产生创造性思维，也没有哪一种思维形式全部不可能产生创造性思维。因此，本书在下面只把经常形成创造性思维的最一般的思维形式作一介绍。

2.2.2 直观思维形式

与爱迪生合作的青年数学家阿普顿刚到爱迪生研究所工作时，爱迪生想考考他的能力。于是给了他一只实验用的灯泡，叫他求一下灯泡的容积。一小时后，爱迪生去检查，发现阿普顿正忙着测量计算。爱迪生说"要是我，就往灯泡里灌水，然后将水倒入量杯，就知道灯泡的容积了"。阿普顿的计算才能(逻辑思维能力)无疑是令人钦佩的，然而在这个问题上他所缺少的恰恰就是像爱迪生那样的直观思维(能力)。

所谓直观思维，就是人们不经过逐步分析而迅速对问题的答案作出合理的猜测、设想或顿悟的一种跃进式思维。从上述爱迪生与其助手求灯泡容积的实例可以知道，直观思维着眼于宏观地把注意力放在事物的整体上，它与逻辑思维微观地把注意力放在事物的各个部分上是很不相同的。

直观思维有利于人们从一些偶然事件中抓住问题的实质。例如，古希腊学者阿基米得在澡盆里沐浴时，看到身体入水后水面位置上升并缓缓向外溢出的现象是难以预料的，因而也是难以用通常的逻辑思维解释和判断的，但直观思维却可以发挥作用，其结果常常产生突破、形成飞跃而导致创造。

日本创造学家新崎盛纪曾把直观思维对应于人类的第一信号系统，认为它是建立在人类直观感觉上，通过人的感觉(视觉、听觉、触觉等)而进行的一种思维活动；他把逻辑思维对应于人类的第二信号系统，认为它是建立在人类理性认识(概念、判断及推理等)基础上的思维。简言之，他认为依靠语言进行的思维是逻辑思维，不依靠语言进行的思维则是直观思维。这种简单地一一对应、简单地认为人类思维发展的形式只是一次性地从具体到抽象、从直观到逻辑的看法值得认真商榷。

作者认为，直观思维虽然利用了人们的感性认识(如感觉、知觉、表象等)，但它并没有仅仅停留在这一步上，它很快便发展成为超越其逻辑思维形式的更高层而上的思维。它犹如处于人类的"感性—理性—感性"反复认识中的后一个感性认识阶段，从表而看，同是感性认识，但二者的层次和实质却不同。直观思维在表面上看是不经过逐步分析就迅速找到了问题的症结，其实它在"迅速"中却已经包含了一系列"感性—理性—(逻辑)—感性"的思维过程。因此，其结果虽然仍以直观形式表现出来，但在实际上它完全可能已在头脑中进行了逻辑程序的高度简缩，并迅速地越过了"理性阶段"，只不过是这一整个思维过程难以用语言表述而已。因此，直观思维来源于感性认识，但它又高于感性认识，至少从普遍具有第一信号系统的高等动物来说目前还很难说都具有直观思维的本领。直观思维也决不能与第一信号系统简单地相对应。

由上可知，直观思维是一种重要的创造性思维形式。直观思维虽然能在创造活动中起

很大作用，但由于它是一种跃进式思维，其整个思维过程又是在极短时间内完成的，以致难以用逻辑思维语言逐步加以分析和表述，因此，直观思维的结果往往带有一定局限性和虚假性，并且经常会导致一些错误结论。比如，过去长期在科学的各个领域流行颇广的所谓“中立原理”（即，如果没有充足理由来判定一个事物的真伪，则把它的真、伪的出现各作 0.5 概率考虑）就是一种错误的直观。

2.2.3　联想思维形式

联想，是想像思维的一种形式。所谓联想思维，就是人们通过一件事情的触发而迁移（想）到另一些事情上的思维。联想能够克服两个不同概念在意义上的差距，并在另一种意义上将二者联结起来。由此产生一些新颖的思想。因此，联想思维是创造性思维的重要思维形式，创造工程中的联想发明法就是与联想思维相关的一种创造技法。在科学史上，许多创造发明均发端于人脑的联想思维。

联想思维是人们因一件事物的触发而联想到另一事物的思维，由此，我们把前一事物称为刺激物或触发物，而后一事物则称为联想物。根据联想物与触发物之间的关系，联想思维还可划分为相似联想、对比联想和接近联想三种形式。

1. 相似联想

相似联想，是指联想物和触发物之间存在着一种或多种相同而有明显属性的联想。例如，看到鸟想到飞机（都能飞），看到电灯想到蜡烛、手电筒（都有发光性）等。有位美国发明家一次在理发时看到理发推子的动作，突然与其正在思考中的收割机方案联系起来，遂产生相似联想，从而成功地开发出了利用理发推子动作原理的新型收割机。我国东北制药总厂在研制“脑复康”时，由于其原料性质极有可能发生爆炸，因而长期不敢进行实验。后来该厂总工程师得知我国地下核实验成功的喜讯后，马上联想到他的可能爆炸的实验也可以在地下进行，遂使问题得到了解决。与相似联想关系密切的创造发明即是模仿创造，其中，模仿创造更为常见，如仿天鹅形体的游船、仿动物形体的电话机等。其创造原理可以归属于移植创造原理。

2. 对比联想

对比联想，是指联想物和触发物之间具有明显相反性质的联想。例如，看到白颜色想到黑颜色、看到小的物体想到大的物体等。在传统观念中，玩具的对象一般都是孩子，国外有人通过对比联想想到玩具的对象也可以转化为老人，于是，近年来专门为成人开发的玩具很受大众欢迎，销路很好。与对比联想直接相关的创造原理是逆反创造原理。

3. 接近联想

接近联想，是指联想物和触发物之间存在很大关联或关系极为密切的联想。例如，看到学生想到教室、实验室及课本、书桌等相关事物。研究表明，对于任何两个似乎毫不相干的概念，一般最多只需要经过 4～5 步的联想即可将其之间建立联系。比如，“木质”与“皮球”这两个离得很远的概念，可以联想为：木质——树林；树林——田野；田野——足球场；足球场——皮球。事实上，上述“木质——皮球”联想之所以能够通过四步达到，是因为该联想的最后一环“皮球”是作为这个程序的终点而预先给定的，这种事先给定“目的”的联想，叫做定向联想，或强制联想。因为，创造发明活动总是有目的的，所以定向联想在创造发明中具有特殊重要的意义。当然，对于创造性思维本身而言，它更加提倡的是思想奔放、毫无拘束的

自由联想。这样的自由联想，可以通过相似、对比或接近联想形式的多次重复交叉而形成一系列的“连锁网络”（如举一反三、闻一知十和触类旁通等），从而产生大量的创造性设想。接近联想产物的数量实际即是发散性思维的一种具体表现。

另外，还有人提出因果联想，即触发物和联想物之间存在一定因果关系的联想。例如，看到乌云密布，就会想到马上就要下雨或者下雪等，显然，因果联想具有某些逻辑思维形式的色彩。

综上所述可以看出：进行联想，要有打破砂锅问（纹）到底的精神，联想的范围越广、深度越大，对创造活动就越有益。比如，从落地电扇具有可调节升降性能可联想并发明了升降篮球架；由伞的开合性的联想发明了能开合的菜罩；乃至从小孩玩的粘虫胶联想到火箭燃料的粘合剂等创造发明等都直接与联想思维形式有关。事实上，古往今来，人类一直是在无意或有意地通过各种联想而不断从自然界中获得启迪，从而创造无数工具或方法，为自身生存和发展创造条件。正如日本创造学家高桥一浩所说：“联想是打开沉睡在头脑深处记忆的最简便和最适宜的钥匙。”

当然，联想能力的大小首先决定于一个人的知识积累和经验丰富的程度。一般来说，知识越多、见识越广的人联想的可能性也越大。例如，一个生长在海边的人就经常会与大海发生联想，而一个出生在大平原、从未见过高山的人，一般与“山”的联想能力就会很少或者没有。据说，古代有个穷人一生中吃过的最好东西是芝麻饼，于是他告诉别人说，如果当上皇帝，他就天天吃芝麻饼，由此可见知识和经验对于人们联想能力的限制了。

此外，联想能力的大小同时还与一个人是否具备良好的想问题的习惯有关，即与一个人是否肯“开动脑筋”有关。有的人虽然见多识广，然而他却不愿多动脑筋，因而也不善于联想，很难进入创造境界。因此，养成良好的“想”问题的习惯，是培养联想思维、提高创造能力的一个重要措施。

2.2.4 幻想思维形式

幻想是想像思维的又一种形式。所谓幻想，一般是指与某种愿望相结合并指向未来的一种想像。由于幻想在人们的创造活动中具有重要作用，所以创造学允许并鼓励人们对于事物进行各种各样的幻想。苏联就曾为学生专门开设过“幻想课”，其目的是引导、培养学生进行各种形式的幻想，以提高学生的创造才能。

幻想，因其暂时脱离现实而常不被人们所重视，很多人甚至把“幻想”作为贬义词而将其打入另册。从创造学看来，这是很不公正的，幻想是一种极其可贵的品质。一个科技工作者在认识世界和创造世界活动中，是很需要幻想精神和幻想思维的。大量事实表明，幻想可使人产生创造的欲望，可激发人们的上进心理，可指出人们进取的方向。幻想可以鼓励人们奋发向前，为人类作出贡献。古人的无数幻想（如“上天”、“入地”、“千里眼”、“顺风耳”等），经过人类的世世代代努力和奋斗，有很多已经变为客观现实。由此可见，幻想思维可直接导致创造活动，很多创造活动均离不开幻想，因此我们不宜盲目地反对幻想。

幻想思维的突出表现即是它的“脱离现实性”。幻想是人们从美好目的（希望点）出发而进行的与现实相脱离的一种想像。有人说：幻想虽然是必要的，但必须本着实事求是的精神，必须用科学的态度来对待它。作者认为，如果单从幻想思维来考虑，幻想就是幻想；既然是幻想，就不该过分强调其“实事求是”和“科学态度”。此外，由于对所谓“实事求是”和“科

学态度”的判断总要受人们当时认识深度和科学发展水平的制约，因此过分强调了“实事求是”、“科学态度”，往往就很难使人发挥幻想的重要作用。比如，以前曾被认为有一定科学根据的“科学幻想”中的所谓火星人，现已证明基本不存在；而过去被认为是“纯粹脱离实际”、“毫无科学根据”的幻想——飞机，却恰恰变成了当今的现实。

由此可见，对于一个创造性的问题，在没有充分深入研究的情况之下我们应该大胆地鼓励幻想思维，而不应简单随意地扣以“毫无根据”、“胡思乱想”等罪名。须知，在很多情况下“胡思乱想”中的幻想也并非没有丝毫根据。人们经常可看到，历来有不少“权威”总是以“实事求是”、“科学态度”的大帽子压制不同的学术观点和学派，特别总是以此压制充满好奇心和幻想的、敢说敢干的青年人。这种做法不利于科学的发展，不利于人们创造性思维的启动和发明创造活动的深入，完全违背了科学技术发展的客观规律。我国著名学者郭沫若曾正确地强调过“既异想天开又实事求是”的思想方法，并把异想天开（即幻想思维）放在首位，这是符合创造学原理的。

正因为幻想是“脱离实际”的，所以幻想思维可以在人脑中纵横驰骋，它可以在没有现实干扰的理想状态下向任意方向发展，从而构成创造性思维的重要组成部分。

与幻想思维最为接近的是空想或无稽之谈。日本创造学家高桥一浩认为，空想是人类思想的宝库，他认为天才的一大特点即是空想思维发达。他在《怎样进行创造性思维》一书中写道：“不论是天才还是凡人，他们同样都有着空想力和以现实道理思考问题的能力，不过，大多数人只能以现实的道理去思考问题，因而他们的空想力便逐渐萎缩。反之，天才却乐于运用空想力，在他思考事物时首先求之于空想。”

诚然，幻想越是大胆，它可能包含的错误也越多，不过这并没有什么关系，只要从幻想的天空回到现实，在大地上加以检验，错误就会被发现、被修正，正确就会被充实和发展。据报载，美国有一位心理学家曾根据幻想思维的作用筹办了一家“幻想公司”，其主要业务是把在顾客看来是一些荒诞和不着边际的幻想变成现实。

总之，幻想这种从现实出发而又超越现实的思维活动，可使人思路开阔、思想奔放，因此它在创造中的作用是明显的，尤其是在创造的初期更需要各种各样的幻想。一门学科的某些变革，一项创造的深入实践，往往总是以勇敢的奇异思维作为开路先锋的。德国启蒙思想家莱辛说得好：“缺乏幻想的学者只能是一个好的流动图书馆和活的参考书，他只会掌握知识但不会创造。”法国启蒙思想家狄德罗说得更实际：“没有幻想，一个人既不能成为诗人，也不能成为哲学家、有机智的人、有理性的生物，他也就不成其为人。”

联想思维和幻想思维同属于想像思维，在一般创造学书中提供一些专门进行想像力训练的作业是有一定意义的。

2.2.5　灵感思维形式

灵感思维是创造性思维的又一种表现形式。灵感思维，是人们的创造活动达到高潮后出现的一种最富有创造性的飞跃思维。灵感思维常常以“一闪念”的形式出现，并往往使人们的创造活动进入到一个质的转折点。大量研究表明，灵感思维是由人们的潜意识思维与显意识思维多次叠加而形成的，是人们进行长期创造性思维活动所达到的一个突破阶段，很多创造性成果都是通过灵感思维形式而最后完成的。所以，有人把灵感的到来看做是狭义的“创造”，是有一定道理的。

1. 灵感思维的主要性质

(1) 引发的随机性——所谓灵感思维引发的随机性，是指灵感既不会像具有必然性的逻辑思维那样可以有意识地导出，也不会同想像思维那样可以自觉地进行思索，它完全可能是由创造者事先想不到的原因而诱发产生的一种思维。究竟是什么东西又怎样引起了人们的灵感，目前还难以说得清楚。但可以肯定的是，不同人的灵感往往是在不同的情况下产生的，甚至同一个人的灵感也会在不同的条件下出现。于是，灵感就显得难以预料、难以捉摸，甚至连创造者本人也根本不可能自觉地意识到在何时何地会产生什么样的灵感。这即是灵感的随机性(或者叫偶然性)。例如，爱因斯坦有一次在朋友家饭桌旁与主人讨论问题，忽然间来了灵感，他便立即拿起笔并在衣袋里摸纸，可是没有摸着，于是竟迫不及待地在新桌布上写起公式来。灵感出现的这种随机性，往往给灵感思维抹上了一层神秘的色彩，因而使得人们在研究它时常常陷入不可知论之中。

(2) 出现的瞬时性— 灵感往往是以"一闪念"的形式出现的，它常常瞬息即逝。宋代苏轼的"作诗火急追亡逋，情景一失永难摹"诗句，即是对灵感瞬时性的生动写照。因此，灵感一旦出现，就要立即抓住。前例中爱因斯坦迫不及待地在朋友家新桌布上记下公式，就是在及时捕捉灵感。不少大学生反映，他们对灵感的瞬时性了解甚少，因而当灵感到来之际仍然听之任之、无动于衷，没有采取任何有效方法捕获灵感，致使事后头脑里依旧一片空白，这是值得惋惜的。英国女作家勃朗特年轻时经常在厨房里劳动，她每次都带着纸和笔，随时准备把脑海中涌现出来的思想(灵感)写下来。据说，奥地利作曲家约翰·施特劳斯的世界名曲《蓝色的多瑙河》，就是灵感到来之际作者匆匆写在衬衣袖口上的。可见，随身携带笔和小本子，是一种捕捉灵感的普遍使用的好方法。

(3) 目标的专一性(专注性)——任何灵感都是针对某一问题或某个方面而产生的，这就是灵感的专一性。同一个灵感不可能解决多方面的问题，多方面的问题也不能指望凭借出一次灵感而得到解决。当然，专一的灵感必然来自于之前对于某一专门问题的充分考虑和过量思考。

(4) 结果的新颖性(独创性)——这是灵感思维作为创造性思维形式的关键所在。然而，并不是所有的灵感都能够产生新颖性的结果，所以，并不是所有的灵感思维都属于创造性思维。那些不能产生新颖性结果的所谓"一闪念"，就不属于创造性思维范畴。钱学森在《关于形象思维问题的一封信》中对灵感思维给予了很高评价："光靠形象思维和抽象思维不能创造，不能突破；要创造突破，得有灵感。"古往今来的重大科学发现、技术发明和杰出的文艺创作，无不与灵感的新颖性有关。诗人、文学家的"神来之笔"、军事指挥家的"出奇制胜"、思想战略家的"豁然贯通"和科学家发明家的"茅塞顿开"等，都充分体现了灵感的新颖性。

(5) 内容的模糊性——许多科学家都似乎有一个共同发现，灵感往往出现在人们醒与睡之间的一种中间状态之下，或出现于显意识与潜意识的交叉过渡之中，这便决定了灵感思维的模糊性。所谓灵感的模糊性，是指灵感所产生的新线索、新结论、新想法往往并不很清晰，尚有待于进一步清理。因此，灵感产生后还需要对其进行认真思维和逻辑思考，才能进一步获得明确的结果，这是创造过程中极为重要的一环。例如，德国化学家凯库勒在半醒半睡状态中产生的灵感，是仅仅发现苯分子(C_6H_6)的结构式呈环状，后来经过多次修正，才把模糊的结果上升为清晰的结构图，遂于1865年提出苯分子为环状结构的理论。

2. 灵感思维的普遍存在

灵感是创造性思维由量变发展到质变的一个飞跃(突变关节点)。根据质量互变规律,量变发展到一定程度必然要引起质变,因此,只要能在创造中做到冥思苦索、过量思考,那么灵感就会在人的头脑中出现。可见,灵感的普遍存在是有一定理论根据的。在美国,有人曾向 1000 多位著名学者调查过两个问题:你在解决重要问题时是否借助过灵感?在什么情况下会出现灵感?对于第一个问题,有 80%的人回答说曾借助过灵感;对于第二个问题的回答却多种多样,诸如在换衣、刮脸、开车游玩、整理庭园、钓鱼、打高尔夫球、散步、听音乐等时间都可能会随时产生灵感。

事实表明,除天才和学者以外,一般人的头脑中也常常会出现灵感。比如,我们常听别人说,"我突然想到了……","我灵机一动……","我急中生智……",这些都与灵感思维活动有关。据作者了解,大学生中自己能讲出得到过灵感的人占 20%左右,这些灵感主要出现在解决各种难题、处理日常事务以及一些小发明、小创造的过程之中。由于很多大学生从来也未曾认真地思考过自己的过去,同时过去也从未认真地记下过灵感的出现及其内容,因此真实的情形可能还远远不止这些。上述情况足以表明,灵感思维决不仅仅是某些天才科学家、发明家所独有的,一般人只要科学地进行创造力开发和创造性思维活动,大多数人都可程度不同地产生各种形式的灵感而出现灵感思维的"火花"。

3. 灵感产生的条件和过程

虽然当前人们对灵感思维的本质了解得尚不够充分,但对于灵感产生的过程还是做了若干研究。目前人们一般认为,灵感产生的条件和过程大致有如下几步。

(1) 头脑中有一个待解决的中心问题——这是由灵感的专一性决定的,它是产生灵感的前提。很明显,一个在头脑中并无需要解决问题的人,决不会产生有关问题的灵感。因此,灵感与要解决的问题之间有直接的关系。

(2) 有足够的知识储备或观察资料(信息资料)积累——这是产生灵感的另一个条件。比如,一个不懂得文学的人决不会出现写诗的灵感;一个毫无天文知识的人也不会出现解决什么天文问题的灵感。究其原因,主要在于他们不具备有关知识和不占有相关资料。所以,灵感思维是以一定的知识或经验积累为先决条件的。

(3) 对于渴望解决的中心问题反复地、艰苦地、长时间地进行思考,即是说要进行超出常规的过量思考——这种过量思考是有意识的,在这种有意识的思考中也包含许多无意识(潜意识)的成分。人们对处在这一阶段的创造者往往很不理解,他们常被人们视为"精神失常者"、"疯子"、"狂人"等。如陈景润走路撞电杆;安培在马路上把黑色马车车厢当做黑板解题引起路人的轰笑;爱迪生走进税务局缴税时好半天竟答不出自己的名字;我国化学家曾昭伦在雨中行路却"不知道"打开手中的雨伞等许多事例,都充分地说明了处于这一阶段科学家的过量思考情形。难怪,在美国的一次民意测验中有 40%的人认为科学家是一群"怪里怪气的人"。在这一阶段,创造者头脑中的问题已经达到了挥之不去、驱之不散的程度,有的思想逐渐转化为创造者的一种潜意识。然而尽管这样,有时问题还是得不到解决,在思考达到饱和之后,人的思路常常进入一种僵化状态。

(4) 适当搁置——人们在进行过量思考、思路进入僵化状态后,可把要解决的问题暂时放一放,使大脑放松放松,也可以做一些其他性质的工作,或者玩一玩、散散步,改换一下环境,缓冲一下紧张的思考,使大脑不再受压抑,这样常可以促使头脑中的潜意识积极活动起

来。在搁置阶段，头脑中已形成的潜意识信息一旦遇到相关的信息刺激，使会自然地产生“一闪念”(或顿悟)。

(5) 灵感的产生——人脑的“一闪念”(或顿悟)一旦形成，即表示灵感已经到来。这时关键是要及时抓住灵感，并通过自觉的思维活动对这一突然的“一闪念”进行鉴别，只有对有用的灵感进行有意识的强化并使之清晰以后，灵感才能在创造中起重要作用。这一阶段，往往需要及时地将灵感记录下来，否则，稍有放松灵感就会从脑海中消逝。

当然，灵感的产生并非都要经过上述几个过程。比如，有时不需经过“搁置”阶段就可通过追捕“热线”而直接产生灵感。所谓“热线”，就是由显意识蕴育成熟了的并可与潜意识相沟通的一些思路。“热线”在大脑中形成，是信息量的积累达到质的突破而产生的，大脑中的“热线”一旦闪现，就要尽快追捕，不能中断并迅速将思维活动推向高潮，之后就会妙思泉涌而产生灵感。比如，有些诗人的诗兴一到来，随即便挥笔疾书，甚至连把斜放着的稿纸扶正的时间似乎都没有。

4. 诱发灵感的基本形式

大量事实表明，当科学的思维活动达到高潮但问题仍旧百思不得其解时，为了得到灵感，诱发的因素就成了关键。因此，了解诱发灵感的基本形式，就可以有利于关键时刻主动地诱发灵感，从而有效地进行创造。诱发灵感的基本形式大致有如下几种。

(1) 联想式——当人的思维发展到前述第(3)阶段以后，在久思不得结果的情况下，很可能会因为某一偶然事件的刺激而顿时产生各种联想，从而使问题豁然开朗、迎刃而解。例如，人们早已知道，为了保证内燃机的有效工作，必须使油与空气均匀混合然后再进行燃烧。但是，油与空气如何才能均匀混合呢？美国工程师杜里埃曾为此大伤脑筋，考虑很久也未能解决。1891 年的一天，他偶然看到妻子向头上喷洒香水，顿时便从这个简单的化妆器联想到油的汽化而突发灵感，从而试制成功了内燃机的汽化器。古希腊学者阿基米得解决“金冠之谜”的灵感，也是来自于他洗澡时水面变化的联想。因此，要产生灵感，就应当特别注意周围事物的细微变化，即使是毫不相干的信息也不要轻易放过。

(2) 触发式——是指人在受到某种刺激、特别是与别人展开讨论或争论并受到别人或自己提出想法的激励而直接迸发出灵感的一种诱发形式。

(3) 省悟式——这种灵感诱发形式的产生不需要借助于外界“触媒”的刺激，而是通过头脑中内在的省悟和内部“思想的闪光”诱发而来。例如，爱因斯坦从 1895 年起就开始思考这个问题，但多年一直没有解决。1905 年的一天早晨，他起床时突然想到：对于一个观察者来说以光速追踪一条光线是同时的两个事件，而对于别的观察者来说就不一定是同时的。他很快地意识到这是个突破口，并牢牢地抓住了这一“思想闪光”的灵感。

从上述诱发灵感的基本形式可知，有意识地暂时消闲(搁置)一下是创造者转移注意、摆脱困境、产生灵感的一个重要方法，如散步、沐浴、听音乐、览读一些与所要解决问题无关的书刊、与外行人员闲谈、入睡前或刚醒时的休息等。据记载，笛卡儿、高斯、庞加莱、爱因斯坦、华莱士、歌德、赫尔姆霍茨等人都曾说过自己有躺在床上休息而得到灵感的体验。日本一家创造力研究所曾于 1983 年 12 月至 1984 年 8 月对 821 名日本发明家进行统计研究，结果表明，有 52%的人曾在枕头上产生过灵感，乘车中产生灵感的有 45%，步行中产生灵感的有 46%，而在工作单位办公桌旁产生灵感的只有 21%。由此可见，在松弛状态下比在工作岗位上产生灵感的机会要大得多。

当然，以上所述只是灵感诱发产生的一般情况，具体灵感的产生过程并非千篇一律，常因人而异。例如，法国物理学家居里(P. Curie)认为，在森林中容易产生激情；美籍意大利物理学家费米(E. Fermi)喜欢躺在寂静的草地上想问题；日本物理学家汤川秀树习惯于夜间在床上思考；法国数学家阿玛边则常在喧哗声中产生灵感；法国剧作家贝克(H. -F. Becque)认为产生灵感最理想的时刻是躺在澡盆之中；而德国物理学家赫尔姆霍茨却认为是一大早或天气晴朗登山之时；著名物理学家杨振宁则认为是在早晨起床后刷牙的时候易来灵感。此外，还有人认为在酒意冲击下会来灵感，法国诗人、作曲家鲁日·德利尔就是这样写出了著名的《马赛曲》。我国李白更有"斗酒诗百篇"的豪兴……可见，每个人均应根据自己的具体情况和习惯，找出诱发自己灵感的最佳方式和最好时机，从而更好地进行创造。许多创造者均有意无意地在利用这一点，大发明家爱迪生就有白天坐在椅子上打盹的习惯，据说许多好念头——灵感就是这样产生的。

应该指出，虽然灵感在创造中具有决定性作用，但这并不意味着所有灵感都是正确的，只要抓住它就可取得创造性成果。其实，无价值的或者不正确的灵感远比成功的要多，只不过人们事后忆及的往往是一些成功的例子而已。

灵感尽管是人们所向往、所追求的目标，但是灵感的到来却是很不容易的，它需要经过大量的、艰苦的劳动和思索。爱迪生认为的"天才就是 1%的灵感加上 99%的汗水"很有道理。即是说，要想得到 1%的灵感，就必须先付出 99%的"汗水"，只有付出了 99%的"汗水"才可能获得 1%的灵感。对此，周恩来总理也有八个字的概括——"长期积累，偶尔得之"。灵感到来的那一瞬间蓦然所得，正是对于创造者长期艰辛而过量思考的回报和奖赏。

总之，灵感的闪现虽然扑朔迷离、犹如幽灵、难以具体捉摸，但是灵感并不神秘，灵感也是可控制的一种思维活动。东京大学名誉教授宫城音弥就"没有努力思考就没有灵感"问题曾说："无论在灵感出现之前还是之后，都需要有意识性的活动。完全脱离意志性的意识活动的灵感，只能在精神病患者身上出现。"

2.3　创新思维技法

创造成果的取得与一个人的创造性密切相关，而主动利用各类创造原理来指导创造技法的实施，则又是创造性表现的重要内容。创新思维技法很多，下面介绍一些常用方法。

2.3.1　智力激励法

智力激励法的原文是 Brain Storming(简称 BS 法)，即头脑风暴之意，故也有人译做"头脑风暴法"或"智暴法"。它是由创造学的奠基人、美国学者奥斯本于 1939 年创立的。该方法最初只用于广告的创造设计，后来很快又在技术革新、管理程序以及社会问题的处理、预测、规划等许多领域得到了广泛应用。智力激励法是能够提出许多创造性设想的有效方法。日本松下公司能在一年内获得 170 万条创造性设想，就是使用了智力激励法。智力激励法的做法大致可分为准备和召开小型会议两步。

1. 准备

因为智力激励法是以召开小型专题讨论会的方式进行的，因此，在会前应先确定好所要攻克的目标，并将其事先通知与会者。如果要解决的问题涉及面太广、包含的因素太多，则

宜先行分解，把大问题分解为若干小问题，然后逐个对每一小问题分别采用智力激励法。

目标确立以后，还要物色好会议的主持人。对于主持人，除要求他必须熟悉该技法以外，还要求能够在具体情境中适当启发和引导与会者，并能与其共同、平等地分析和对待问题。

2. 召开小型会议

小型会议的与会者以5～10人为宜，人多了很难使与会者充分发表意见。如果一定需要更多的人参加，则可分别开几个会。会议除主持人外，可另设1～2名记录员（现在则可使用录音或摄像技术）。参加人员除了熟悉并与该问题有关的以外，还可适当地吸取相近专业人员乃至外行参加。这样做，既能保证所提设想的深度，又利于突破专业习惯思路的束缚，可得到独创性较高的设想。会议时间大约为半小时到1小时。由主持人宣布议题后，即可启发、鼓励大家提出设想。会议进行一般应遵守下列一些原则：

(1) 会议气氛自由奔放—— 解放思想是会议的精髓。会议提倡随便思考、自由畅谈、任意想像、尽量发挥、互相激励。想法越新奇越好，因为有时看上去很"荒唐"的设想却可能很有价值。所以，与会者要善于从多种角度甚至反常角度考虑问题，要暂时抛开头脑中已有的各种准则规定、条条框框，甚至还可故意做一些违背传统、逻辑和一般常识的大胆思考。

(2) 严禁批判—— 在会议上对别人提出的任何想法，都不能批评、不得阻拦。即使自己认为是幼稚的、错误的甚至荒诞离奇的设想，也不宜予以驳斥，同时也不允许自我批判。要真正做到这一点，就要确实在心理上调动每一个与会者的积极性，就要彻底防止出现一些"扼杀语句"和"自我扼杀语句"，诸如："这根本行不通！""你的想法太陈旧了！""道理上也许行，但实际上行吗？""这是不可能的！""这不符合××定律！"以及"我提一个不成熟的看法！""我有一个不一定行得通的想法！"等词句，都不允许在会议上出现。只有这样做，才能保证与会者在充分放松的心境下、在别人所提设想的激励下，集中全部精力，开动脑筋，充分地拓展思路以形成新颖的设想。

(3) 以谋求设想数量为主——在智力激励法的实施会议上，只鼓励和强调与会者提设想，越多越好，会议以谋取设想数量为主要目标。很多事实表明，高质量的设想方案往往多是在后期产生的，而且在同一期限内一个能比别人多提出两倍设想的人，其中有实用价值的设想最终可能比别人要高出10倍。可见，只有设想数量多了，其中好的设想才会更多。

(4) 善于用别人的想法开拓自己的思路——召开智力激励法小型会议的主旨是创设一种与会者相互激励的情境，与会者在这种氛围中善于向别人学习，接受启迪，正是"激励"之关键所在。每个与会者均以他人设想激励自己，或补充他人的设想，或将他人的若干设想加以综合后提出自己新的设想等。总之，要充分利用别人的设想诱发自己的创造性思维，使所有的与会者均可相互诱导、相互启发、相互激励，从而促使提出的设想数量在有限的会议时间内尽量增加。

智力激励法会议"严禁批判"的做法只是暂时的。会议结束以后，人们总要对众多设想进行评议、分类和选择，并从中找出最有可能实施的设想。但是，在会议进行之中则必须"严禁批判"，只有这样做，才会使人们充分发挥想像力，排除各种因素的干扰，以获得"心理安全"和"心理自由"，这样不必担心会被人讽刺为疯子、狂人而框住自己的思路。例如，有一次用智力激励法讨论如何改进饭碗时，很多人都提出了设想。后来，一位平时不干家务的人在他人激励下终于也提出了一种"最好能生产一种不用清洗的碗以免除家务劳动"的设想。后

来经过筛选，发现这种“不用清洗”的碗也是一种社会需要，如在缺水地区、旅游途中、野外勘测等环境中就很有意义。通过研究，一种用多层纸压成、每次吃完饭只需撕去一层的“不用洗的碗”便问世了。

智力激励法是一种有助于集思广益的集体思考方法。当一个人独自思考一件事或一个问题时，其思路常被限制在一定范围而受阻，如果有几个人同时对问题进行思考，各人都以自己的知识经验从各自不同角度认识同一问题，就会有利于互相激励、引出联想，从而产生共振和连锁反应，诱发出更多的设想。该创造技法问世以后，一些学校曾专门开设了智力激励法课程。日本一些大企业也纷纷通过举办训练班大力推广和应用该技法。我国一些工厂运用该技法以后，也收到了明显效果。

在智力激励法的基础上，人们又根据具体情况对其形式做了多种多样的发展，其中，最常见的是默写式智力激励法，又称“635”法，是德国人针对其民族习惯于沉思的性格而发展起来的。按照这一方法，每次会议有 6 人参加，每人首先备有一张卡片，会议要求每人于 5 分钟内在各自的卡片上写出自己的 3 个设想（故名“635”法），然后将卡片传给自己的右邻。每人接到左邻的卡片后，在第二个 5 分钟内参考别人所写的设想后再在其下写出 3 个设想，然后再次把自己填写的卡片传给右邻。……如此多次传递，共传 6 次，半小时即可进行完毕，理论上可产生 108 个设想。对大学生进行创造技法练习的实践情况表明，最初进行创造技法实施时，“635”法有很大的优越性。

不论是智力激励法还是其派生出的“635”法，由于在时间安排上均做了限制，可使人在紧张的气氛中处于高度兴奋状态，通过相互激励而扩大、增多创造性设想，因而，它是一个重要的也是基本的创造技法。

然而，现在创造学界也有一些人认为智力激励法尚存在不少局限之处。比如，有学者认为，智力激励法对于一些具体的、窄而专的科技问题基本无效，因为在运用该技法时非专家对于这些领域了解太少，所以无法提出什么“设想”来，如非电子学专业的专家就不太可能提出有关可控硅快速功率放大问题的设想。因此有人认为，智力激励法应当主要用于开发新产品、扩大产品用途和改进广告等方面。

2.3.2　题目问答法

提出问题是创造发明的第一步，创造力开发较好的人，都具有善于提出问题的能力。有时，如果能够提出一个好的问题，往往意味着成功的一半。但是，如何提出问题、如何通过提问题而达到创造发明的目的呢？设问法，就是通过有关提问的形式去发现事物的症结所在并继而进行发明创造的一类技法。设问法的种类较多，最有代表性的就是奥斯本的“检核表法”。

“检核表法”，是针对创造的目标（或需要发明的对象）从多方面用表格列出一系列思考问题，然后逐个加以讨论、分析和判断，从而获得解决问题的最好方案或设想。一般所说的奥斯本的“检核表法”，多是从以下 9 个方面提问题进行检核的。

（1）现有发明成果有无其他更多的用途？或稍加改变后有无别的用途？

奥斯本认为，创造有两种类型：一种是先确定目标，然后对准目标去寻找方法；另一种是首先发现一种事实，然后想像该事实会有什么作用，即从方法着手引向目的。这一条检核内容是符合后一种创造类型的，是人们常用的一种创造技法。比如，电熨斗还有什么用途呢？

人们可以想像出它的尽可能多的用途，后来有人发现可以用它烙饼，于是将外形稍加改变就发明了一种新烙饼器。此外，有人把理发用的电吹风用于烘干被褥，从而发明了一种新型的被褥烘干机。还有人将水果网兜在棒上绕了几圈，便制成了一种洗瓶器（该小发明使发明者每月收入 40 万日元）。尼龙袜的诞生，也是“用途迁移”的产物。最初的尼龙丝只用于军事，如制造降落伞、做舰艇上的缆绳等，因而尼龙丝的销售量很少。为此，人们开始寻找它的其他用途，后来发现可用结实耐磨的尼龙丝制成袜子以取代纱袜，于是美国的杜邦化学公司制成了尼龙袜并先让纽约的著名舞星和杜邦公司的女秘书试穿做产品广告宣传，随后正式投入市场，仅 1 年时间就销售了 6400 万双。1994 年 4 月，南京人设想出使用消防用的高压水枪扫荡马路旁法国梧桐树上球毛的绝招，解决了多年一直未能解决的犹如雪花飘落的球毛污染问题。可见，每找到一个老事物的新用途，实在不亚于发明一个新产品。

总之，这一设问要求人们对现在物品的固定功能进行怀疑或遐想，只要破除“功能固定”论，就有可能产生新的创造。

（2）过去有无类似的东西？有什么东西可供模仿？能否在现有发明中引入其他创造性设想？

这个提问有助于使某一发明向广度和深度发展，以形成系列发明产品。如从普通火柴到磁性火柴、保险火柴等，都是引入了其他领域的发明才形成的袖珍取火手段的系列产品。泌尿科医生引入微爆破技术消除肾结石，也是借用了其他领域的发明。山西一位建筑工人借用能够烧穿钢板的电弧机烧穿水泥板，打洞又快又好，后经改进终于发明了水泥电弧切割机。作者受超声体外碎石机理的启发，提出了高浓度、难降解有机污水超声聚焦裂解新工艺。

（3）现有发明能否改变形状、颜色、声音、味道或制造方法？

从这些方面提出问题，往往会产生意想不到的发明创造。例如，将蜡烛的形状变为球形，放在玻璃杯中点火非常好看；面包外面裹上一层芳香包装，能增加嗅觉诱惑力；有人将滚柱轴承的滚柱改变成圆球形，遂发明了滚珠轴承；一位制镜商将平面镜的形状改变成多种曲面，制成了哈哈镜；还有彩色大米、彩色棉花、彩色钢铁等仅仅是作了颜色的改变，也都产生了创造发明。

（4）现有东西能否扩大使用范围、增加功能、延长寿命？能否添加部件、增加长度和提高强度？

奥斯本指出，在自我发问的技巧中，研究“再多些”和“再少些”这类有关联的成分，可诱发大量构思和设想。比如，在两块玻璃之间加入某些材料，可制成一种防震、防碎、防弹的新型玻璃；在牙膏中掺入某些药物，可使牙膏增加治疗口腔疾病的功效；日本财阀石桥正二郎曾把袜式胶鞋鞋帮有胶的部位向上加长一些以防止泥水湿透鞋面，这项专利使他在 7 年内销售胶鞋超过 2 亿双，取得了很大的经济效益；美国电影《巨猩》中的主角是一头比一般人个体大几十倍的巨猩，这种“扩大一下”的创造满足了人们的好奇心，影片曾风靡一时。

（5）能否将现有的东西缩小体积、减轻重量？能否省略一些部件？能否进一步细分？

目前，许多产品都出现了由大变小、由重变轻的趋势，其结构也在不减少功能的基础上力求简化，出现了许多小型、微型机器。如袖珍收录机、微型计算机、折叠伞等，都是以缩小体积为目标进行发明的产物；有的造纸厂把大捆的手纸改为小包装，这种“缩小”也打开了产品销路；用微型吸尘器做成的黑板擦也是一种缩小创造；日本的奥塞罗围棋是一般围棋的缩

小，这种棋是在 8×8 即 64 个方格中填入一面白、一面黑的棋子，如果某一个棋子被包围，则该子就要翻过来，这种棋的发明费竟高达 3 亿日元。

据报道，我国留美学生李文杰 1992 年在加利福尼亚大学发明了世界上最小的、只有在显微镜下才能看到的电池，其大小只有红血球细胞的百分之一，如果把这种电池用在集成电路上，可望提高功能 1000 倍。可见，以“缩小”为目标的创造发明往往有其独特优势。1992 年 10 月 18 日，著名提琴家史兹克斯在维也纳公开演奏并引起轰动的、由瑞士提琴制造家史奈得精心制作的袖珍小提琴，其长度只有 3.3cm。

(6) 能否用其他产品、材料或生产工艺、加工方法替代原有的产品或发明？

由于当前世界上某些资源相当紧缺或是其成本昂贵而不易得到，于是人们不得不寻找其他的代用品，这也是一种创造发明。如人造大理石、人造丝等都是很好的例子；此外，还有用汽车中的液压传动代替齿轮、用充氩气的办法代替电灯泡中抽真空等。通过取代和替换途径，可为想像提供广阔的探索领域。

(7) 能否将现有的发明更换一下型号或更换一下顺序？

重新安排、更换位置通常也会带来许多创造性的设想。例如，飞机诞生的初期，螺旋桨均装在头部，后来装到了顶部遂发明了直升飞机；原来的汽车喇叭按钮多装在方向盘的轴心上，每次按喇叭总要把手向上移动到轴心处，既不方便又容易失手肇事，后来有人把喇叭按钮改装在方向盘的下半个圆周上，只要手指轻按一下该半圆上的任何一处，喇叭就响起来；另外，工作时间上的重新调整、城镇建设的合理布局等也都有可能导致更好的创新结果。

(8) 能否将现有的产品、发明或工艺方法颠倒一下？

上下颠倒、内外颠倒、正反颠倒等都可能产生新的效果。例如，大炮一般都是向上发射的，反过来发射行不行呢？苏联发明的“大炮打桩机”，就是用 165mm 口径的大炮向地下发射“炮弹”(即钢桩)，每炮可入地 2.5m，极大地提高了打桩工作效率。有人发现在铁路两侧的油井，其产油量比远离铁路的油井高，追其原因是由于振动提高了产油量，于是人们发明了振动采油强化设备。

(9) 可否将几种发明或产品组合在一起？

组合通常被认为是创造性的动力源泉。如将几种部件组合在一起变成组合机床；把几种金属组合在一起变成性能不同的合金，把几种材料组合成复合材料等。

使用奥斯本检核表法解决一个技术问题，通常可从几个提问中同时受到启发，经过综合后往往可形成最佳方案。

创造学界一般认为，奥斯本的检核表法几乎适合于任何类型和场合的创造劳动，因此享有“创造技法之母”名声。我们认为，正因为它是“母”，奥斯本的检核表法就不宜再屈称其为创造技法了，从它所包含的 9 个方面内容考察，其中大多数均是创造原理，并且都可以归入前一节所讲到的有关创造原理之中。

我国创造学研究者借鉴该检核表法而提出的“12 个一”创造技法，在推广创造学中发挥了很好的作用。此外，有人也根据奥斯本检核表创造技法的原理，结合一些较具体的情况制定了各种各样的“检核表”，共有 10 个方面的内容。

此外，一般的设问法中还有“5W2H”法，这是一种通过为什么(why)、做什么(what)、何人(who)、何时(when)、何处(where)、怎样(how)和多少(how much)等 7 个方面的提问而形成创造性设想的创造技法。

2.3.3 联想组合法

联想组合法的思维基础是联想思维，它所依据的原理主要是组合创造原理。联想组合法又可简称为组合法。即使是一个最简单的组合创造，如铅笔和橡皮的组合，最初也是离不开两者之间的(相近)联想的。

联想组合可划分为自由联想组合和强制联想组合两大类。由于发明创造大多是针对某一目标、为解决某一问题而进行的创造劳动，因而与此相关的强制联想组合在一般发明创造中显得更为重要。为此，本书只介绍强制联想组合法。强制联想组合发明法大致可划分为如下几种具体发法。

1. 查阅产品样本法

查阅产品样本法，是将两个或两个以上的、一般情况下被认为彼此并无关联的产品(或想法)强行联系组合在一起从而产生出新颖性方案的方法。

按照查阅产品样本法，人们可以打开某厂家的产品目录或其他印刷品，随意地将某些项目、某些产品或某些题目逐个挑选出来，并用同样的方法将另一产品目录或印刷品中的某些项目、产品或题目逐个挑选出来，再依次将二者分别进行一一对应的强行组合，以产生出独创性的结果。这时，由于思维随着两件事物的"联系"而产生、跳跃比较大，因此容易克服经验的束缚而启发人的灵感。比如，深受用户欢迎的保温杯就是将暖水瓶的保温胆与杯子强制联想组合而设计成功的。

在进行强制联想组合时，思想一定要解放，对于强制组合法的"新产品"要从创造性角度认真加以分析，不能被表面看来"不可能组合在一起"的框框所限制。比如，酒和西瓜看上去并无什么关联，它们的组合初看似乎是不可能的，但若进行强制组合再仔细思考就可能有所突破，美国一园艺师从这一联想组合出发，就培养出了香味可口的酒味西瓜。为了通过强制联想组合而寻找新的创造目标，加拿大发明家曾将印有几百个产品、项目、题目的小塑料条装进一个特制的容器内，按一下旋钮后容器中的字条被搅拌起来，停下时容器的小窗口上可显示出四五个小条上的字。将这些随即出现在小条上的内容进行强制联想组合，也许就会产生出一些新的创造念头。

2. 二元坐标组合法

二元坐标组合法也是一种强制联想组合发明法。它与查阅产品样本法的不同之处，在于把要组合的对象先列成坐标体系，然后再进行一一对应的强制组合，因而具有系统性和不遗漏性。使用二元坐标组合法的具体步骤如下(以对日历的创造为例)：

第一步，列出有关创造发明目标的元素，然后再任意列出联想组合的元素，其范围可以尽量宽一些。比如，可列出玻璃、扇、气、梯、滑行、日历、清凉、照明、瓶、手摇、管、车、纸、流动、座、三角、笔筒、杯 18 个组合元素，其中日历是要发明的目标元素，其他都是任意所列元素且词性不加限定。

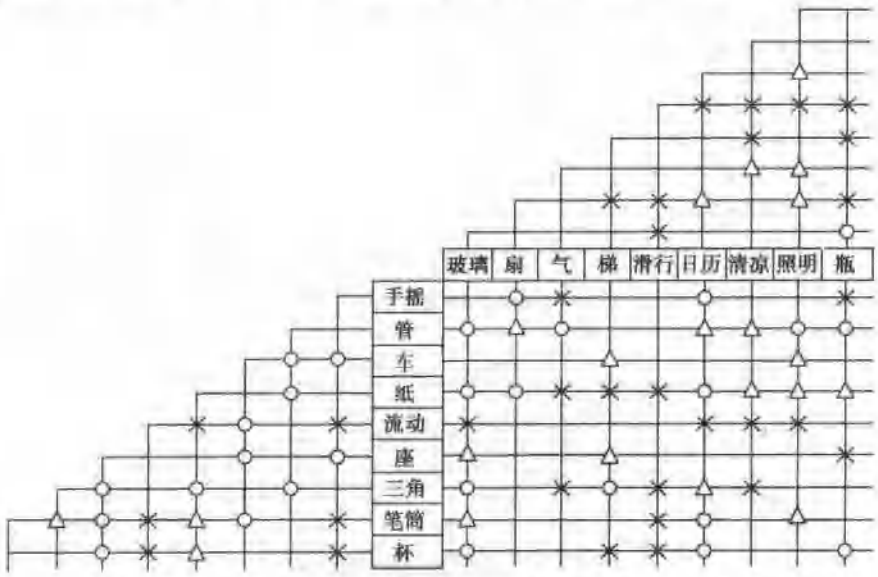
第二步，把这 18 个元素分为相等的两部分，分别排成纵、横行列，然后用组合线强制沟通所有的元素并编制成组合图形，如图 2-1 所示。

第三步，进行联想组合和判断，并将判断的结果按图示标记符号标记在图的组合交点处。在结果判断时，要互换两元素的位置进行。例如，"车"与"手摇车"和"车手摇"，后者是无意义的结果，而前者则是已有的发明。

第四步，从图中找出有意义的结果。比如，在本例中有意义的结果是照明日历（带光源的日历或夜光日历）、日历扇、日历管、三角日历等（其中的三角日历近来已被申请了中国专利）。此外，还能产生一些与目标元素不太相关的其他发明创造的构想，如三角笔筒、玻璃座、纸瓶、照明车等。

第五步，对有意义的结果进行可行性分析。

由于不同的人列出的联想组合元素可能不会相同，如果把若干人所编的这种图表依次互换、取长补短，就可以发挥集体智慧，这种方法又叫做集体应用二元坐标组合法。



图例：×——无意义；○——已有；△——有意义；空白——暂不确定

图 2-1　二元坐标组合图

3. 焦点组合法

焦点组合法过去叫焦点联想法，是联想组合法中最突出的一种创造技法。

查阅产品样本法所选择出来的创造目标是随意的、无定向的，而人们的发明创造目标大多都是预先确定好的。比如，要发明一种新型的打火机，就不能随便地翻阅毫不相关的样品目录。二元坐标组合法虽然包含创造目标的因素在内，但产生的结果对于创造目标来说仍不够集中，大量结果虽然可能很富有创意，却与已有的目标并不相关。因此，对于在没有明确创造目标而帮助创造者选择创造的目标和对象而言，前两种方法是很有作用的，而对于已有明确创造目标的发明创造来说，焦点组合法则显得更为优越。在焦点组合法中，组合的一方是可任意联想的，而组合的另一方则是预先指定的欲创造对象，即所谓的“焦点”。焦点组合法要求创造者紧紧围绕“焦点”进行强制联想，因此，该技法自始至终都紧扣着创造的主题。现以生产椅子为例介绍运用焦点组合法的步骤。

第一步，确定焦点物。要发明新型椅子，则以椅子作为强制联想的“焦点”。

第二步，另外仍选一个物品作为参照物进行联想，联想时该参照物可起一个触发物的作用，如可以选取“灯泡”。

第三步，用发散性思维分析灯泡并将其结果分别与椅子进行强制联想组合。例如：玻璃

灯泡——玻璃做的椅子；球形灯泡——球形椅子；螺口灯泡——螺旋式转椅；电灯泡——电动椅；遥控灯——遥控椅；透明的灯泡——透明质料的坐椅；发光的灯泡——椅背上带灯可供看书的椅子等。

第四步，对于上一步思维发散的结果再次进行联想发散，并将结果再次与椅子之间进行强制组合。例如，以选取最后一个"发光"设想为例，其联想之一为：发光——亮——白天——云彩；云彩一样色彩美丽的椅子；云彩之形——云形的椅子；云彩会变色——变色的椅子；浮云——坐上后有悬浮感的椅子等。又如，从第二个"球形"进行联想，则有：球形——圆形——辐射对称——花；像花一样的椅子；花有玫瑰花、百合花——类似于玫瑰花、百合花的玫瑰椅、百合椅；花有茎和叶部形状；花有香味——能散发香味的椅子等。

第五步，从上述众多方案中选出有商业价值的设想予以试制。

焦点组合法的另一种演变形式是成对特性列举法。它与焦点组合法的区别是，在对任选触发物进行发散性思维的分析以后，还要对焦点物进行发散性分解，然后再把每一个发散性的结果依次与触发物的发散性结果按二元坐标法进行强制联想组合，最后选择出好的目标方案。图 2-2 是以香蕉为触发物、以钢笔为焦点物（发明物）进行的成对特性列举法图解。从上图可得到诸如月芽形（香蕉形）笔杆、香味笔杆、柔软笔帽、香蕉形笔尖许多种可能的发明目标。

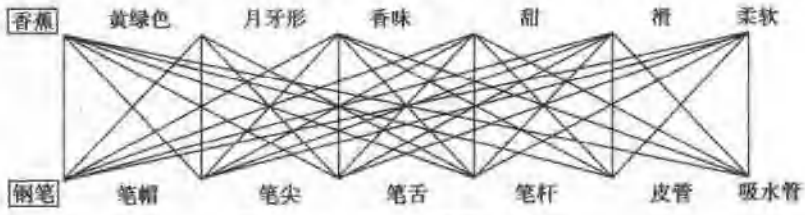


图 2-2 成对特性列举法图解

2.3.4 类比法

类似法，是指用待发明的创造对象与某一具有共同属性的已知事物进行对照类比，以便从中获得启发而进行创造发明。比如，德国物理学家欧姆（G. S. Ohm）在研究电流流动时，将电与热进行类比，把通过导体的电势比做温度、把电流总量比做一定热量，运用傅立叶热传导理论的基本思想再引入电阻概念进行研究，终于在世界上首先提出了著名的欧姆定律。

由此可见，类比发明法需要借助于原有的知识，但又不能受原有知识的过分束缚。这一方法要求人们通过创造性联想思维把两个不同的事物联系起来，把陌生的对象与熟悉的对象联系起来，把未知的东西与已知的东西联系起来，异中求同，从而设想出新的事物。由于世界上所有事物之间都存在着某种程度的相似性，因而类比方法不仅可用于同类事物之间，也可用于不同发展阶段的不同事物之间。所以说，世界上一切事物之间都存在应用类比方法的可能性。

类比发明法的实施大致有以下三个步骤：

第一步，选择类比对象。类比对象的选择应围绕发明创造目标为中心。可以先分析所创造的目标物应该具有什么样的属性，再以此为线索去寻找有关的类比对象；亦可先粗略分析一个已知事物的属性，看其中有哪些属性与所创造的目标物相同，从而择定其为类比的对象。但无论怎样，类比对象都应该是创造者所熟悉的事物。这一步中，联想思维是很重要

的，要善于应用联想把表面上毫不相关的事物联系起来。

第二步，将两者进行分析、比较，从中找出需要的共同属性。

第三步，在第一、二步基础上进行类比联想推理并得出结论。

古埃及人曾用不断转动的链条运送水桶以灌溉农田，1783 年英国人埃文斯运用类比法将该方法用于磨坊以传送谷粒。这一类比法虽然十分简单，但是在长达几千年的时间里却一直没有被人发现。加拿大人通过与机关枪的类比，发明了能连续扫射播种树种的“种树枪”（其子弹由塑料制成，内装树的种子和肥土）。

类比发明法来自于移植创造原理。类比发明法是在两个特定的事物之间进行的，它既不同于从特殊到一般的归纳方法，也不同于从一般到特殊的演绎方法。根据类比的对象和方式不同，类比方法还可以进一步区分为拟人类比、直接类比、反向类比、象征类比、因果类比、对称类比、综合类比等。

2.3.5　列举法

列举法作为一种发明创造技法，是以列举的方式把问题展开，用强制性的分析寻找创造发明的目标和途径。列举法的主要作用是帮助人们克服感知不足和因思想被束缚而引起的障碍，迫使人们带着一种新奇感将事物的细节统统列举出来，迫使人们时时处处去想某一熟悉事物的各种缺陷，迫使人们尽量想到所要达到的具体目的和指标。这样做，比较容易捕捉到所需要的目标从而进行发明创造。

1. 属性列举法

属性列举法，过去称为特性列举法，是由美国创造学家克拉福德（R. P. Crowford）研究总结出来的一种创造技法。运用该技法时先要对创造发明对象的主要属性进行详细分析（即将属性逐一列出），之后再探讨能否进行改革或创新。一般来说，要着手解决或革新的问题越小越容易获得成功。例如，要革新自行车，即便是采用智力激励法也难以得出全新的设想，原因是自行车涉及的面太广，难以把握。如果将自行车分解成若干部分（如车胎、钢圈、钢丝、轴承、链条、齿轮、车身、把手、刹车等）予以分别研究，那么，只要革新其中的一个或几个部分，就可能导致自行车整体性能的改变。这样做，对于自行车的发明创造就容易获得成功。

运用克拉福德属性列举法的一般步骤如下：

第一步，选择一个比较明确的课题，课题宜小不宜大，如果课题较大则应将其分解成若干小课题，课题选定以后，首先要列举出发明或创新对象的属性。一般包括三个方面——名词属性：性质、材料、整体、部分、制造方法等；形容词属性：颜色、形状、大小等；动词属性：有关机能和作用的性质，特别是那些使事物具有存在意义的功能。例如，要改革一只烧水用的水壶，人们可按照属性列举法将水壶的属性分别列出：

名词属性——整体：水壶；部分：壶嘴、壶柄、壶身、壶底、气孔；材料：铝、铁皮、铜皮、搪瓷等；制造方法：冲压、焊接。

形容词属性——颜色：黄色、白色、灰色；体重：轻、重；形状：方、圆、椭圆，大小、高低等。

动词属性——装水、烧水、倒水、保温等。

第二步，从各个属性出发，通过提问诱发出用于革新的新方案（这时亦可参考使用奥斯本的检核表法）。比如，通过名词属性可提出：壶嘴是否太长？壶柄能否改用塑胶？壶盖能否用冲压法以免焊接的麻烦？怎样使焊接处更牢固？除上述材料以外是否还有更廉价的材

料？水开后冒出的蒸汽烫手，气孔能否移到别处？等等。有一种水开后会自动鸣笛，而壶盖上无孔，提壶时不会烫手。当然，如果从形容词上下功夫也可能有所创新，如怎样倒水更方便，怎样烧水节省能源等，同样也可产生受市场欢迎的新产品方案。

2. 缺点列举法

人们常常有一种惰性，对于看惯、用惯了的东西往往很难发现其缺点，也很少主动找它的缺点，因而无形中便"凑合"、"将就"着维持现状，甚至用"理所当然"、"本该如此"等观点对待它，从而使人安于现状，丧失了创造欲望和机会。缺点列举法，是指积极地寻找并抓住、有时甚至需要去挖掘(因为有许多缺点是极不明显的)各种事物的不方便、不得劲、不美观、不实用、不省料、不轻巧、不便宜、不安全、不省力等各种缺点、问题或不足之处，从而确定创造发明目标的一种创造技法。

有些事物的缺点是随着事物的最初出现而出现的。比如，世界上第一台电子计算机的体积太大；第一辆汽车的坐椅不太舒服等。也有许多事物的缺点原本并不是什么"缺点"，后来随着时间的推移和环境的改变而转化成了缺点。例如，一次性塑料饭盒，刚问世时使用得非常好，本身没有什么缺点可言，可是随着用量的增加，它所造成的"白色污染"很快就影响了生态环境，其缺点便显而易见了。

运用缺点列举法没有严格程序，一般可按下列步骤进行：

第一步，确定某一改革、革新的对象。

第二步，尽量列举这一对象事物的缺点和不足(可用智力激励法，也可进行广泛的调查研究、对比分析或征求意见)。

第三步，将众多的缺点加以归类整理。

第四步，针对每一缺点进行分析、改进，或采用缺点逆用法发明出新的产品。

例如，对一双普通的长筒雨靴，可以列出如下一些缺点——材料方面：鞋面弯折处易开裂，鞋后跟易磨损……；外观方面：颜色单调，式样千篇一律……；功能方面：春寒有雨时穿上冻脚，夏天有雨时穿上闷脚，潮气重容易患脚气，走路不跟脚，袜子容易掉下来……只要针对上述某一缺点着手进行改进，就可能创造出更好的新产品。比如，日本有一个叫荒井的人，针对雨靴"夏天穿闷脚、易患脚气"这一缺点在制造方法上加以改进，制成了前后有透气孔的雨靴；还有一个叫野口文雄的人，针对雨靴"脚后跟容易磨损"的缺点研究出了一种浇模时在脚后跟部位埋进一种鞋钉的新式雨鞋，大大提高了耐磨损性能。现在市场上的各种颜色的雨鞋，即是克服"颜色单调"缺点后的创新产品。

缺点列举法简单易行且容易收到效果，很受大中小学生和工厂企业生产一线工作人员的欢迎。据了解，我国在工厂企业中普及创造学最容易出成果的创造技法就是缺点列举法。

与缺点列举法相关的另一技法是缺点逆用法。所谓缺点逆用法，就是针对对象事物中已经发现的缺点不是采用改进缺点的做法，而是从反面考虑如何利用这些缺点从而做到"变害为利"的一种创造技法。例如，日本某纤维公司有一次织错了布，布上的绒毛单向倾斜，因而布卖不出去。这时，有人提出："布的绒毛只向一方倾斜，如果用它来做成刷子不是能刷去衣服上的灰尘吗？"该公司马上派人将其装到刷子把上进行试验，效果很好，连衣服纹理深处的灰尘都能刷净。于是，公司将其定名为"礼节刷子"投入市场，很快便成了畅销品，如图 2-3(a)所示。后来购买这种"礼节刷子"的人又针对其缺点做了改进：只能单方向使用很不方便，如果能使刷子面旋转，改变一下方向就更好了，于是制成了反方向也能用的刷子，它在市场上同样

也很畅销，如图2-3(b)所示。之后，又有人再次运用缺点列举法指出：一次一次地旋转太费事，于是把刷子做成了"V"字型，分别在两面装上绒毛方向相反的布，不仅可不必费事旋转而且可降低成本。这种刷子又是一举成功，颇受顾客的青睐，如图2-3(c)所示。

又如，我国某陶瓷厂因配方下料有误而使其生产的一批陶瓷产品表面釉彩裂开，尽管该产品本身质量并不差，但仍难以销售。此时有人献计道：这些开裂的釉彩看上去形若蟹爪、竹叶、波纹或天上的浮云，各有气势、变化万千，能否作为专门的工艺品投放市场呢？结果，不但销售十分兴旺，而且无意中还开发出了名为"裂纹釉"的新产品。再如，天津某毛纺织厂生产了一种呢料，因其着色不均而出现白点，影响了销路，该厂利用缺点逆用法变消灭"白点"为扩大"白点"，从而开发出了新产品——雪花呢。

(a) 单向礼节刷子　(b) 旋转礼节刷子　(c) "V" 型礼节刷子(两面绒毛方向相反)

图2-3　各种礼节刷子示意图

3. 希望点列举法

缺点列举法直接从社会需要的功能、审美、经济、实用等角度出发研究对象的缺点，提出切实有效的改进方案，因而简便易行，常会取得很好的效果。然而，缺点列举法大多是围绕原来事物的缺陷加以改进，通常不触动原来事物的本质和总体，因而它属于被动型创造方法，一般只适用于对老产品的改造或用于不成熟的新设想、新发明，从而使其趋于完善。希望点列举发明法，则是通过列举希望新的事物具有的属性以寻找新的发明目标的一种创造方法。由于希望点发明法是从人们的意愿出发提出各种希望设想，所以很少或完全不受已有物品的束缚，这便为人们使用该方法提供了广阔的创造思维空间。

希望点列举法的实施步骤是：激发人们的希望(可用智力激励法形成一批希望点)——收集人们的希望——仔细研究人们的希望——创造新产品以满足人们的希望。例如，一家制笔公司用希望点列举法产生了一批改革钢笔的希望点：希望钢笔出水顺利；希望绝对不漏墨水；希望一支笔可写出两种以上颜色的字；希望笔尖不开裂；希望不用吸墨水；希望省去笔套；希望落地时不损坏笔尖等。这家制笔公司后来从"希望省去笔套"希望点出发，研制出一种像圆珠笔一样可以伸缩的钢笔，省去了笔套，打入了市场。又如，株洲车辆厂某车间学习创造学以后，想利用希望点列举发明法改造出钢水箱，遂召开了希望点列举会议，对新型的水箱总结出如下希望点：不会因骤冷骤热产生裂纹而漏水；能够经受钢水的冲刷而不损坏；寿命要长；维修方便；制造简单易行。之后，车间针对这批希望点寻找资料、进行研究。最终采用整体铸造、钢管埋入的方法制成了新型的电炉出钢口的水箱，从而达到了希望的目标。

现在市场上许多新产品都是针对人们的"希望"研制出来的：人们希望电风扇能吹出一阵一阵的风，于是发明了模拟自然风的阵风电扇；人们希望把伞放进提包，于是发明了折叠伞；人们希望夜间开门找钥匙方便，于是发明了带电珠的钥匙圈；人们希望洗衣服不需要费力拧干，于是发明了甩干机；人们希望能不费力地将重物搬上楼，于是发明了能爬楼梯的小车等。

希望人人皆有，但要提出创造性强且又科学可行的希望却不容易。链式传动自行车诞

生于1884年，其实早在1495年达·芬奇就"希望"发明一种靠人力通过链条驱动的自行机械并设计出了有关图纸，然而在当时是无法实现的。这说明，希望总是产生在现实之前的，希望是对现状的冲击和挑战，满足于现状是难以产生希望的。

2.3.6 逆向发明法

在从事发明创造时，有时会遇到难题，绞尽脑汁也想不出好办法来，这时不妨从问题的相反方向，或倒过来去思考，即运用"逆向思维法"或许会使你顿开茅塞。逆向求解的思路和方法是一种新型的求异思维，是辩证思维，是思维开阔、思维灵活的表现，思维没有经过训练的人，用起来比较困难。但是，一旦利用这种方法解决了问题，会使你享受到"柳暗花明又一村"的乐趣和成功的喜悦。

科学史上利用"反面求索发明法"取得成功的事例不胜枚举。不妨把它划成三种类型：

第一种，从逆向去思考。我们知道，在自然界的许多事物和现象中，往往都具有正反两方面的意义。正确认识和理解这些现象，往往会创造出许多新东西来。例如，印刷用凹板不如凸板方便，照相用正片，而医学上的X光用底片更合适。特别是日本的中田解决圆珠笔漏油，更为逆向思考的范例。因为圆珠笔的笔珠被磨损而慢慢漏油，按一般思路是增强笔珠的耐磨性，但新问题又出现了，与笔珠接触的笔头内侧被磨损仍会漏油，甚至笔珠会蹦出。而中田变换思路，从反面思考获得成功：既然圆珠笔写到两万字时笔珠就变小漏油，那就让圆珠笔写到一万五千字时油就全部用完。

第二种，颠倒一下。1819年，丹麦物理学家奥斯特发现通电导体可使磁针转动的磁效应。1820年，法国的安培发现通电的螺线管具有与磁石相同的作用。英国物理学家法拉第想："为什么不能用磁来产生电呢？"于是，法拉第开始做各种各样的试验，用九年的艰苦探索，终于发现了电磁感应现象，为人类进入电气化时代开辟了道路。

有种转动墙板，由轴与碗组成。过去做法是碗口向上，易积灰，转动不便，后来想出一种最简单的方案，就是将轴与碗的方位对换一下，碗口向下，不仅不积灰，而且每转动一次清一次灰，这是不花钱的革新。

第三种，化弊为利。这是一种利用事物的缺点，"以毒攻毒"化弊为利的创造性方法。例如，金属腐蚀是一件坏事，但人们却利用腐蚀原理发明了刻蚀和电化学加工工艺等方法。1974年高温无剂防火涂料出现后，受到大家重视。这种材料涂在蒙古包外，可防火(1200℃下，包内温度仅20℃)。但"防"与"保"也有共性，能否将这种防火涂料用到相反的方向，如用它涂在水泥厂的回转窑外壁，不就可以使热损耗降低了吗？

从反面去思考问题，是一条重要的思维规律，它包含着极为丰富的内容。一个发明家，首先需要的不是常规的想法，而是反常规的想法。日本丰田公司第一位老板丰田章一郎说："我这个人如果说取得了一点成功的话，那因为我什么问题都倒过来思考。"

2.3.7 反求工程法

所谓反求工程是指对已经投入运行证明效果良好的新产品、新技术、新设备进行全面、系统、深入的科学检测、计算、分析和研究，找出其技术关键和控制方法。"全面"是指对尽可能多的同类产品和技术进行详尽具体的分析和评议；"系统"是研究各技术环节的特点及其相互关系；而"深入"是要求运用近代科学测试与计算手段，分析了解其细节和内在联系。因

此反求工程可以认为是开发研究与设计制造的逆过程，也是当前技术市场竞争中的保密技术关键的一种“破译”手段。

实际上，世界各国 70%以上的技术都不是本国独创，而需要由国外引进。要掌握这些技术、了解竞争对手的水平和动向，或者将本国的优秀传统技艺上升为先进技术，正常的、科学的途径都需要通过反求工程方法去研究和探讨。日本的经济振兴就曾经得力于反求工程。如本田公司为了开发新型摩托车，曾对五百多种摩托车进行过反求工程研究，博采众长，开发了有自己特色的摩托车，从而能垄断国际市场。

1. 反求工程研究的内容

反求工程是以引进的先进设备、技术、生产线系统为具体研究对象，在正常、稳定的运转情况下，以各给定的生产条件与操作参数为依据而开展研究工作的。反求工程研究的内容大体包括以下几方面：

(1) 系统工艺设计的指导思想和主要意图；

(2) 关键设备的功能分析和过程原理；

(3) 各子系统的最佳匹配；

(4) 最佳工况和调节控制方法的分析；

(5) 条件变化后生产情况预测；

(6) 根据不同的目标函数，对系统进行科学评议；

(7) 对已取得的关键计算机程序进行“破译”，复原原始模型。

以上内容都要求获得定量结果，以便指导生产与开发工作。

2. 反求工程的研究方法

根据反求的不同要求，可通过以下步骤和方法来实现：

(1) 数据处理的统计规律。大量收集具有代表性数据，包括操作参数、设备结构参数和物理化学特性参数等，通过数理统计的方法或借助其他工程数学方法作为生产控制、理论分析、模型试验和计算机模拟时的参考。

(2) 物理、化学模型试验。为探索设备或系统的工作原理、经验模型、运行规律、确定关联式系数指数或观察实际现象、描述物料运行状态与轨迹乃至化学反应过程等都需要进行专门的模型试验。

(3) 建立子系统乃至全系统的数学模型进行运行参数的反求。根据系统繁简程度和人们对过程认识的深度，可以分别建立经验模型、半经验模型和理论模型。纯经验模型局限性大，理论模型适应性强但难度大。例如，水泥工业窑炉系统至今还未建立起完备的能实际应用的理论模型，大多还需要附以实物模型试验或给以必要的约束条件来推导出半理论模型。

对于复杂的工程问题，在建立数学模型时往往需要根据实际情况作必要的简化假设，这种假设既要使模型简化可解，又要尽可能不失真，因此是很重要的。

(4) 计算机数学模拟模型。利用经过校验修正、比较完善的数学模型可以进行计算机数学模型试验，即人为输入给定的主变量，计算出相应参数的变化值，从而找出变量波动造成的后果，推算出各项指标可能的变化。

(5) 计算机程序的破译。有关生产控制和设计计算的计算机程序，往往是最能集中反映设计思想与方法的重要资料。因此选择关键性的已有程序进行阅读、理解和反复试验，再配合理论分析，复原其程序编制的依据或数学式、数学模型等，对深入认识技术关键也是很重要的。

这个工作有一定难度，并不是所有程序均能深刻理解，这对工作人员素质的要求也很高。

在运用这些反求工程的研究方法解决实际问题时，还会遇到许多困难，这就需要专业人员与计算机应用方面的技术人员相互协作、相互渗透，才可能获得圆满解决。

2.3.8 废物利用法

在马克思生活的年代里，虽还不存在环境污染问题，可他已就废物利用问题在《资本论》中作了如下三点论述：

(1) 废物的大量性——大量废物因大规模生产而产生；

(2) 机械装置的改进——由于机械装置的改进把某些按原来形状无法利用的材料，变成为生产中可以利用的新形状；

(3) 由于科学技术，特别是化学的进步，发现了废物可利用的新性质，因而开发了利用废物的新技术。

从马克思的预言到今天，我们面临着严重的问题——环境污染，看来，开发革新性技术，是既能解决环境污染又能回收废物的好办法。图 2-4 为废物利用法的原理。

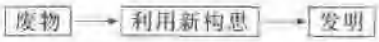


图 2-4 废物利用法原理

通过废物利用法产生发明的例子很多：

由于塑料废弃物不可降解，到处丢弃污染环境，由此想到发明自然性塑料，从而产生了淀粉加聚乙烯醇树脂以及通过在塑料中添加紫外线光敏剂来产生新的塑料替代品。

利用没用的烟灰作为生产水泥的混合原料；发明了以烟灰为原料的轻质砖；将烟灰处理后得到了空心玻璃珠。

为了处理和利用垃圾发明了垃圾压块机、垃圾分类筛选设备，建立垃圾处理工厂甚至创造了垃圾发电的方法和设备。

1979 年，美国政府决定由财政部向全国发出通告，希望造纸厂前去投标，供应制造美钞的专用纸。中标的一家是名不见经传、小得不能再小的克兰造纸公司。这家公司所造的纸张价格低廉，纸型光洁，厚薄均匀，坚固耐磨，受潮不变形。在以后十年中，美国政府几次想用更好的纸代替，但克兰造纸公司却稳操胜券，独揽合同。

这家公司造纸原料来自两个途径：一是纺织行业的纱头碎布、纤维尖角和废弃的丝团垃圾，二是美国家庭丢弃垃圾中的废旧衣物，该公司收集后予以处理利用。

2.4 机电产品创新设计

重视产品的创新设计是增强机电一体化产品竞争力的根本途径，产品的创新设计就是通过设计人员运用创新设计理论和方法设计出结构新颖、性能优良和高效的新产品。照搬照抄是不可能进行创新设计的。设计本身就应该具有创新。当然，创新设计本身也存在着创新多少和水平高低之分。判断创新设计的关键是新颖性，即原理新、结构新、组合方式新。

构思一种新的工作原理就可创造出一类新的产品。例如，激光技术的应用，产生了激光加工机床。创造一种新结构的执行机构就可造就一种新的机器。例如，起重机的抓斗采用

了多自由度差动滑轮组和复式滑轮机构创造发明的"异步抓斗"。采用新的组合方式也可创造出一种新的机器。例如，美国阿波罗13号飞船在没有重新设计和制造一个零部件的情况下，通过选用现有的元器件及零部件组合而成，取得满意的结果。这就是组合的创新。由此可见，创新设计的含义是十分广泛的。产品创新设计的内容一般应包括三个方面：

(1) 功能解的创新设计。这是属于方案设计范畴，其中包括新功能的构思、功能分析和功能结构设计、功能的原理创新、功能元的结构解创新、结构解组成创新等。从机电一体化产品方案创新设计角度来看，其中最核心的部分还是运动和结构方案的创新和构思。所以，有不少设计人员把运动和结构方案的创新设计看做机电产品创新设计的主要内容。

(2) 零部件的创新设计。零部件方案确定以后，零部件的构形设计阶段也有不少内容可以进行设计创新，减小尺寸、重量，采用新材料以提高强度、刚度和使用寿命等，所有这些都是机电产品创新设计的内容。

(3) 工业艺术造型的创新设计。为了增强机电产品的竞争力，我们应该对机电产品的造型、色彩、面饰等进行创新设计。机电产品的工业艺术造型设计得法，可令使用者心情舒畅、爱不释手，同时也可使机电产品的功能得到充分的体现。

由于篇幅限制本节主要介绍功能解的创新设计。

2.4.1　功能的概念

功能分析是方案设计的出发点，是产品设计的第一道工序。机械产品结构如同人体结构。人有头部、胸、腹、四肢等解剖结构件，机器有齿轮、轴、连杆、螺钉、机架等组合结构件；人有消化、呼吸、血液循环等功能件，机器有动力、传动、执行、控制等功能件。这种人—机比较，有助于加深对机器功能的理解。机电一体化产品的常规设计是从结构件开始，而功能分析是从对产品结构的思考转为对它的功能思考，从而做到不受现有结构的束缚，以便形成新的设计构思，提出创造性方案。

功能是抽象地描述机械产品输入量和输出量之间的因果关系，对具体产品来说，功能是指产品的效能、用途和作用。人们购置的是产品功能，人们使用的也是产品功能。比如，运输工具的功能是运物载客；电动机的功能是将电能转换为机械能；减速器的功能是传递转矩，变换转速；机床的功能是把坯料变成零件等。功能还可表述为：功能＝条件×属性。其含义是在不同的条件下利用不同的属性，同一物体可实现不同的功能。

按照功能的重要程度，功能分为两类：基本功能和辅助功能。基本功能是实现产品使用价值必不可少的功能，辅助功能即产品的附加功能。例如，洗衣机的基本功能是去污，其辅助功能是甩干；手表的基本功能是计时，其辅助功能是防水、防震、防碰、夜光等。

采用功能分析法进行方案设计时，按下列步骤进行工作：

(1) 设计任务抽象化，确定总功能。抓

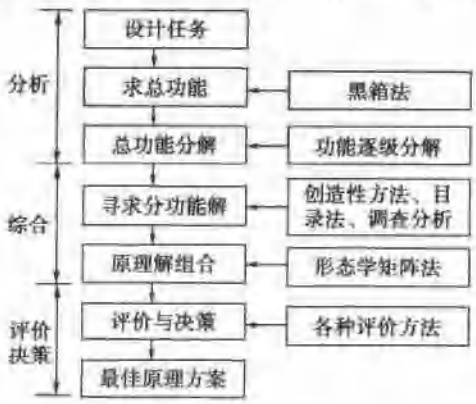


图2-3　原理方案设计步骤

住本质，扩展思路，寻找解决问题的多种方法；

(2) 将总功能逐步分解到不能再分解的功能元，形成功能树为止；

(3) 寻求分功能(功能元)的解；

(4) 原理解组合，形成多种原理设计方案；

(5) 方案评价与决策。

图 2-5 表达了功能分析确定方案的工作步骤。

2.4.2 确定总功能

1. 设计问题抽象化

从抽象到具体、从定性到定量是产品设计的战略思想方法。所谓抽象化，就是将设计要求抽象化，而不是像常规设计那样，一接到任务就开始具体设计。

抽象化是人们认识事物本质的最好途径，无需涉及具体解决方案就能清晰地掌握产品的基本功能，把设计人员思维集中到关键问题上来。通过抽象，抛开头脑里原有的框框和偏见，展开视野，有利于寻求更为理想的设计方案。

抽象化的目的是为了确定产品总功能。例如，采煤机抽象为物料分离和移位的设备；载重汽车抽象为长距离运输物料的工具；洗碗机，抽象为除去餐具上污垢的装置；设计和改进一个密封装置，可抽象为不与轴接触而对元件进行密封。

在设计任务书中，列出了许多要求和愿望，在抽象过程中，要抓住本质，突出重点。淘汰次要条件，将定量参数改为定性描述，对主要部分充分地扩展，只描述任务，不涉及具体解决办法。但是在具体设计时要根据设计任务具体情况对上述步骤作适当删减。举例说明通过问题抽象化获得功能定义能扩大解的范围。例如，丝杠冷轧机的功能树如图 2-6 所示。

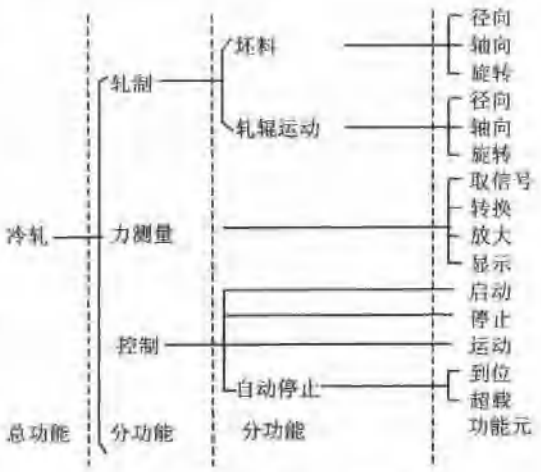


图 2-6 丝杠冷轧机的功能树

又如砸开核桃壳取出果仁的功能描述，若用“砸”则已暗示了解法，而较抽象的表达才可能得到思路更开阔的解答(见表 2-1)。

表 2-1　取核桃仁的功能描述

功能分析	解　法　构　思
砸壳 ↓ 适度抽象化 桃仁与壳分离	外部加压： 砸：利用重力 夹：利用杠杆作用(核桃夹子) 压：螺旋压力机 冲击：水力冲击 射击：把核桃当作为"枪弹"射向硬靶 逆向思维 → 内部加压： 钻孔：向外壳内充气撑破外壳 整体加压：外压骤减，内压撑破外壳(专利发明) 侧向思维： 使壳变脆：冷冻法让壳变脆易碎 把壳溶解，溶解壳而不溶解桃仁

2. 黑箱法

对于要解决的问题，设计人员难以立即认识，犹如对待一个不透明、不知其内部结构的"黑箱"。利用对未知系统的外部观测，分析该系统与环境之间的输入和输出，通过输入和输出的转换关系确定系统的功能、特性，进一步寻求能实现该功能、特性所需具备的工作原理与内部结构，这种方法称为黑箱法。黑箱法要求设计者不要首先从产品结构着手，而应从系统的功能出发设计产品，这是一种设计方法的转变。黑箱法有利于抓住问题的本质、扩大思路、摆脱传统结构的旧框框，获得新颖、较高水平的设计方案。图 2-7 为金属切削机床黑箱示意图。图中左右两边输入和输出都有能量、物料和信号三种形式，图下方为周围环境(灰尘、温度、湿度和地基的震动)对机床工作性能的干扰，图上方为机床工作时对周围环境的影响，如散发热量、产生振动和噪声。通过输入、输出的转换，得到机床的总功能是将毛坯加工成所需零件。

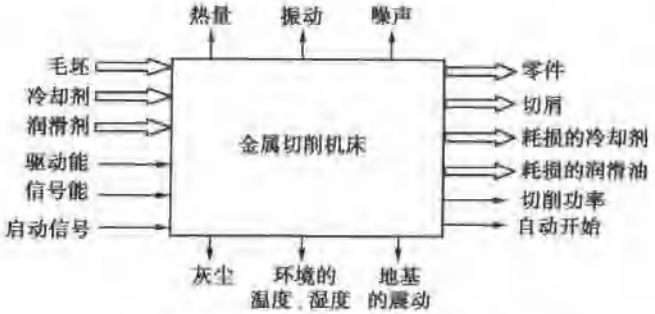


图 2-7　金属切削机床黑箱示意图

2.4.3　总功能分解

系统的总功能可以分解为分功能(或称一级分功能、二级分功能……)，分功能再分解为功能元(最小单位)。所以，功能是有层次的，是能逐层分解的，如图 2-8 所示。

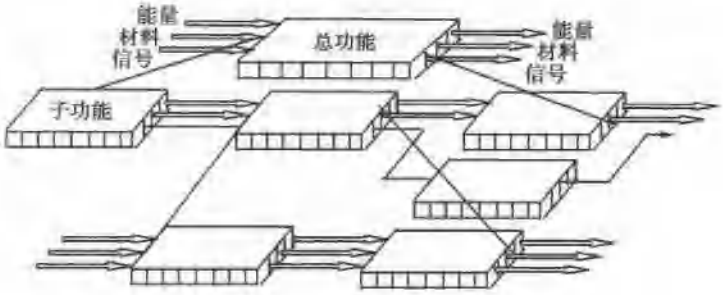


图 2-8　功能分解图

1. 功能元

功能元是功能的基本单位。在机电一体化产品设计中常用的基本功能元有：物理功能元、逻辑功能元和数学功能元。

(1) 物理功能元。它反映系统中能量、物料、信号变化的物理基本动作，常用的有：转变—复原；放大—缩小；连接—分离；传导—绝缘；存储—提取。

"转变—复原"功能元，主要包括各种类型能量之间的转变、运动形式的转变、材料性质的转变、物态的转变及信号种类的转变等。

"放大—缩小"功能元，是指各种能量、信号向量(力、速度等)或物理量的放大与缩小，以及物料性质的缩放(压敏材料电阻随外压力的变化)。

"连接—分离"功能元，包括能量、物料、信号同质或不同质数量上的结合。除物料之间的合并、分离外，流体与能量结合成压力流体泵的功能也属此范围。

"传导—绝缘"功能元，反映能量、物料、信号的位置变化。传导包括单向传导、变向传导，绝缘包括离合器、开关、阀门等。

"存储—提取"功能元，体现一定时间范围内保存的功能。如飞轮、弹簧、电池、电容器等，反映能量的储存；录音带、磁鼓反映声音、信号的储存。

(2) 数学功能元。它反映数学的基本动作，如加和减、乘和除、乘方和开方、积分和微分。数学功能元主要用于机械式的加减机构和除法机构，如差动轮系、计算机、求积仪等。

(3) 逻辑功能元，包括"与"、"或"、"非"三元的逻辑动作，主要用于控制功能。

2. 功能结构

类似电气系统线路图，分功能的关系也可以用图来描述，表达分功能关系的图为功能结构图。功能结构图是结合初步的工作原理或简单的构形设想而建立的。常用功能结构有三种：

(1) 串联结构，又称顺序结构，它反映了分功能之间的因果关系或时间、空间顺序关系，基本形式如图 2-9(a)所示。如台虎钳的施力与夹紧两个分功能就是串联关系，如图 2-9(b)所示。

(2) 并联结构，又称选择结构，几个分功能作为手段共同完成一个目的，或同时完成某些分功能后才能继续执行下一个分功能，则这几个分功能处于并联关系，其一般形式如图 2-10(a)所示。例如，车床需要工件与刀具共同运动来完成加工物料的任务，如图 2-10(b)所示。

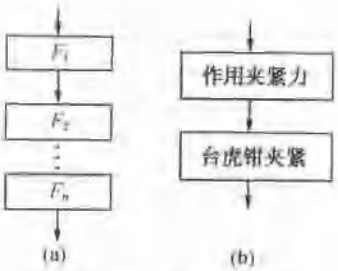


图 2-9　串联结构原理

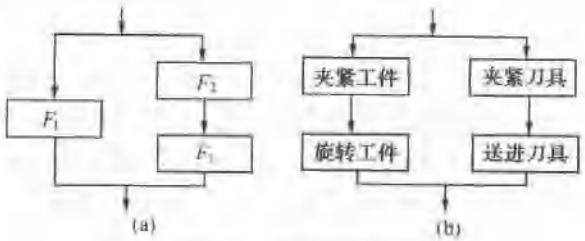


图 2-10　并联结构原理

(3) 环形结构，又称循环结构，输出反馈与输入的结构如图 2-11(a)所示为循环结构。按逻辑条件分析，满足一定条件而循环进行的结构如图 2-11(b)所示。

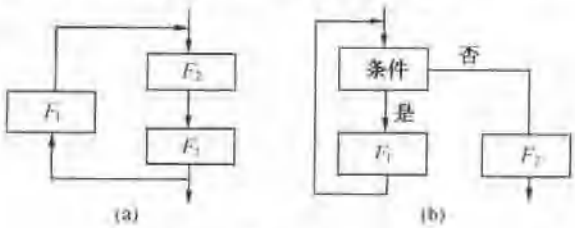


图 2-11　环形结构原理

3. 建立功能结构图的要求

功能结构图的建立是使技术系统从抽象走向具体的重要环节之一。通过功能结构图的绘制，明确实现系统的总功能所需要的分功能、功能元及其顺序关系。这些较简单的分功能和功能元，可以比较容易地与一定物理效应及实现这些效应的实体结构相对应，从而可以得出实现所定总功能需要的实体解答方案来。建立功能结构图时应注意以下要求：

(1) 体现功能元或分功能之间的顺序关系。这是功能结构图与功能分解图之间的区别。

(2) 各分功能或功能元的划分及其排列要有一定的理论依据，物理作用原理或经验支持以确保分功能或功能元有明确解答。

(3) 不能漏掉必要的分功能或功能元，要保证得到预期的结果。

(4) 尽可能简单明了，但要便于实体解答方案的求取。

4. 功能结构图的变化

实现同一功能的功能结构可有多种，改变功能结构常可开发出新的产品。改变的方法有：

(1) 功能的进一步分解或重新组合。

(2) 顺序的改变。能量进入系统以后，其转换与传递顺序不同，实体解答方案亦将不同。

(3) 分功能联接形式的改变。

(4) 系统边界的改变。必要时可扩大或缩小系统的功能，以求得更合理的解答方案。提高系统机械化、自动化程度是其重要方面。

5. 建立功能结构图的步骤

(1) 通过技术过程分析，划定技术系统的边界，定出总功能。

(2) 划分分功能及功能元。通常首先考虑所应完成的主要工作过程的动作和作用,具体做法可参见功能分解。

(3) 建立功能结构图。根据其物理作用原理、经验或参照已有的类似系统,首先排定与主要工作过程有关的分功能或功能元的顺序,通常先提出一个粗略方案,然后检验并完善其相互关系,补充其他部分。为了选出较优的方案,一般应同时考虑几个不同功能结构。

(4) 评比选出最佳的功能结构方案。进行评比的方面是:实现的可能性、复杂程度、是否获得解答方案、是否满足特定要求。通常可取少数较好的方案进一步具体化,直至实体解答完全确立。建立功能结构图的流程可归纳如图 2-12 所示。

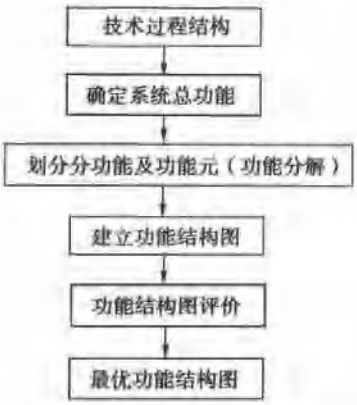


图 2-12 建立功能结构图的流程

例 2-1 利用上述原理试建立波轮式洗衣机分功能及功能结构图。

解:① 确定系统总功能:洗衣机的总功能是洗涤衣物,包括容纳衣物和水、搅动衣物和水、定时、排水、能量转换、联接和支承。

② 划分分功能及功能元(图 2-13)。

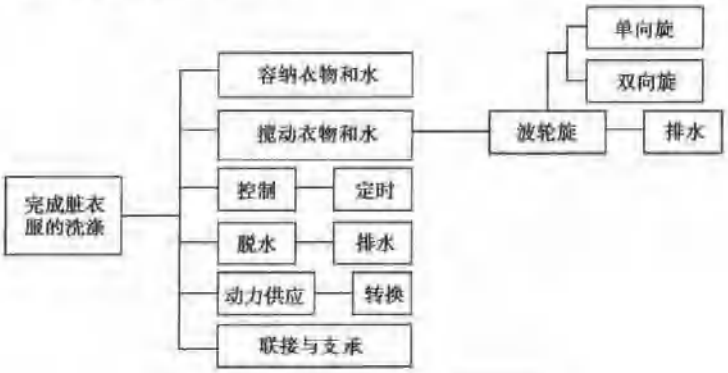


图 2-13 划分分功能及功能元

③ 建立功能结构图(图 2-14)。

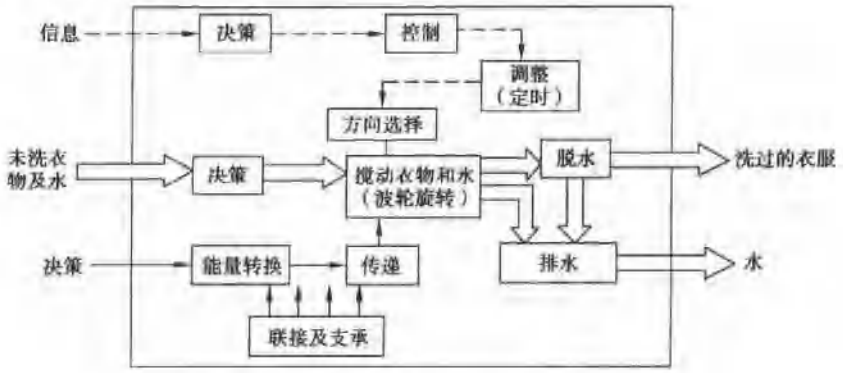


图 2-14 功能结构图

例 2-2　建立材料拉伸试验机的功能结构图。

解：① 用黑箱法求总功。通过分析输入和输出关系，得到材料拉伸机的总功能：测量试件受力与变形，如图 2-15 所示。

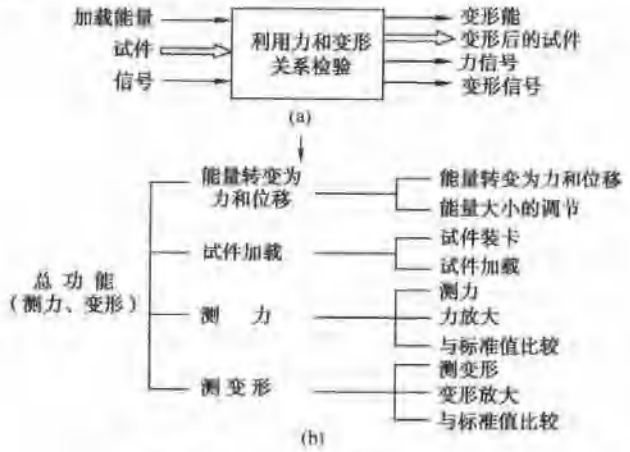


图 2-15　方案总功能分解图

② 总功能分解。总功能分解为一级分功能：能量转换为力和位移、力测量、变形测量、试件加载。然后考虑到各分功能的实现尚需满足其他要求，如输入能量大小要调节、力和变形测量值需放大，试件加载拉伸要装卡，在调节和测量时均需与标准值进行比较等，因此将一级分功能再分解为二级分功能，其具体内容如图 2-16 所示。

③ 建立功能结构。有了总功能图后，接着建立一级分功能结构图，如图 2-16(a)所示。最后建立二级分功能结构图，如图 2-16(b)所示。

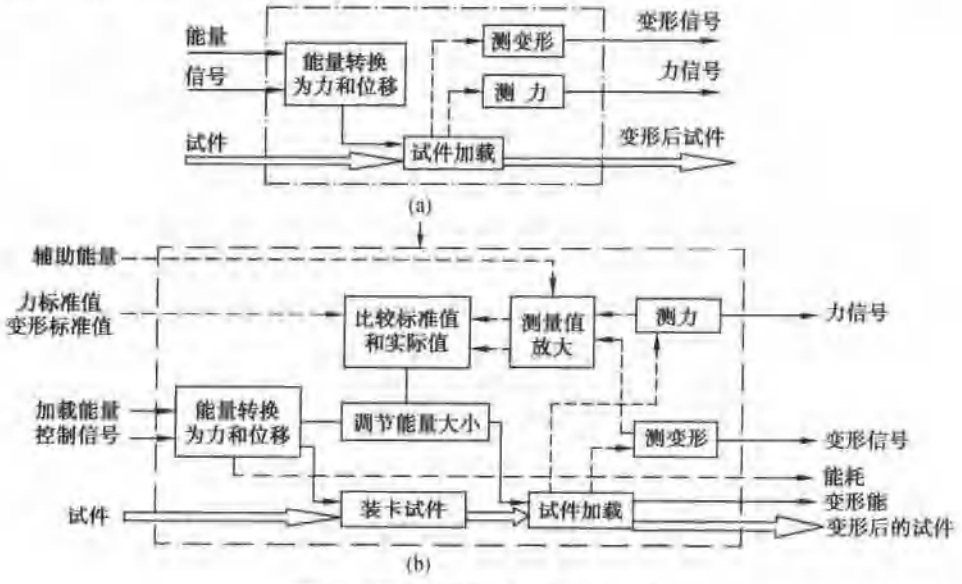


图 2-16　材料拉伸机功能结构

2.4.4 功能元(分功能)求解

功能(功能元)求解是原理方案设计中的关键步骤。功能元求解就是将所需执行动作，用合适的执行机构型式来实现。功能元载体的求解可根据解法目录找到，再通过运动链抽象、变异得到更佳的机构解，也可由创新技法构思出一个新型机构来实现。将机构按运动转换的种类进行分类，就是起到解法目录的作用。下面介绍几种求解方法：

1. 直觉法

直觉法是设计师凭借个人的智慧、经验和创造能力，包括采用后面将要论述的几种创造性思维方法，如质量功能配置、智暴法、类比法和综合法等，充分调动设计师的灵感思维，来寻求各种分功能的原理解。

直觉思维是人对设计问题的一种自我判断，往往是非逻辑的、快速的直接抓住问题的实质，但它又不是神秘或无中生有的，而是设计者长期思考而突然获得解决的一种认识上的飞跃。日本富士通用电气公司职工小野，一次雨后散步，在路旁发现一张湿淋淋的展开的卫生纸，由此激发了他的灵感：天晴时，废纸是一团团的，而被雨水淋湿后，都自动伸展开来。后来，他利用"废纸干湿卷伸原理"，研制成功了"纸型自动控制器"，获得一项日本专利。

2. 调查分析法

设计师要了解当前国内外技术发展状况，大量查阅文献资料和专业书刊、专利资料、学术报告、研究论文等，掌握多种专业门类的最新研究成果。这是解决设计问题的重要源泉。

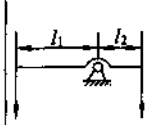
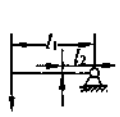
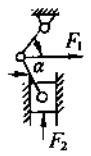
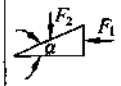
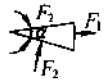
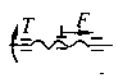
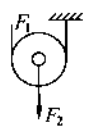
我们的知识来源于大自然，设计师有意识地研究大自然的形状、结构变化过程，对动植物生态特点深入研究，必将得到更多的启示，诱发出更多新的、可应用的功能解，或技术方案。通过对生物学和工程技术方面的关系的研究，开辟了仿生学或生物工程学科。利用自然现象来解决工程技术问题。例如，雷达与声呐的发明就是模仿蝙蝠的"导航系统"，机器人的出现就是模仿人的听觉、视觉和部分思维及动作面产生。

调查分析同类机电产品对其进行功能和结构分析，研究哪些是先进可靠的，哪些是陈旧落后的、需要更新改进的，这都对开发新产品、构思新方案、寻找功能原理解法有益处。

3. 设计目录法

设计目录法是设计工作的一种有效工具，是设计信息的存储器、知识库。它以清晰的表格形式把设计过程中所需的参考解决方案加以分类、排列，供设计者查找和调用。设计目录不同于传统的设计和标准手册，它提供给设计师的不是零件的设计计算方法，而是提供分功能或功能元的原理解，给设计者具体启发，帮助设计者具体构思，如表 2-2、表 2-3 所示。

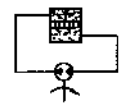
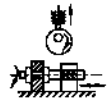
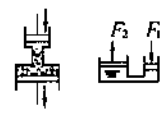
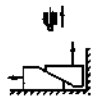
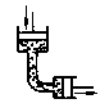
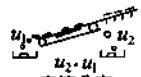
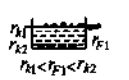
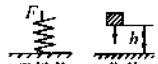
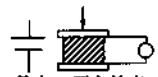
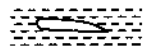
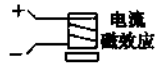
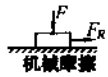
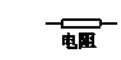
表 2-2 机械一次增力功能元解法目录

机构名称	杠	杆	肘杆(曲杆)	斜 面	楔	螺 旋	动滑动
机构简图							

续表

机构名称	杠　　杆		肘杆(曲杆)	斜　面	楔	螺　旋	动滑动
计算公式	$F_2=\frac{l_1}{l_2}F_1$ $l_1>l_2$	$F_2=\frac{l_1}{l_2}F_1$	$F_2=F_1\tan\alpha$ $\tan\alpha>1$	$F_2=\frac{F_1}{\tan\alpha}$	$F_2=\frac{F_1}{2\sin\frac{\alpha}{2}}$	$F=-\frac{2T}{g(\lambda+\rho)}$ λ—螺纹升角 ρ—当量摩擦角	$F_2=\frac{F_1}{2}$

表 2-3　部分常用物理基本功能元解法目录

功能 \ 原理解		机　　械				液　　气	电　　磁
转变		凸轮传动	联杆传动	齿轮传动	拉伸/压缩方式传动		
							
缩小(放大)							
变向							
分离		 摩擦分离					 磁分离
力产生	静力	 弹性能　位能				 液压能	 静电　压电效应
	动力	 离心力				 液体压力效应	 电流磁效应
	摩擦力	 机械摩擦				 毛细管	 电阻

2.5　创新设计实例

2.5.1　缆索机器人设计

“缆索机器人设计”项目是 2003 年江苏省大学生创新设计大赛的参赛项目。该项目的创意来自于猴子的爬绳动作，利用机电控制系统实现对猴子爬绳动作的模仿。该设计的现

实意义在于：目前，城市建筑不断向高空发展，高层楼宇玻璃幕墙比比皆是，外墙清洁是劳动强度高且危险性极大的工作，此外，管道监测维修、隧道掘进、高层建筑顶制件安装、墙面施工等工作都需要用机器代替人力。设计出能够满足这类高空作业要求的机器人是该项目的设计目标。该设计初步实现了缆索机器人的爬绳动作，要成为一种实用的产品还需进一步的研究开发设计。

2.5.2 总体设计方案

首先，我们来分析猴子的爬树动作，仔细观察猴子的动作可以发现，它是用双手和双脚分别夹住树干，配合身体的伸缩来实现灵活的上下爬行动作，这样的爬行动作可以采用多种结构来实现，如曲柄连杆机构、凸轮机构等。该方案采用汽缸作为执行元件，通过汽缸推动滑块机构来实现上下爬行动作。

机器人由三个汽缸担任执行机构，它们分别模拟双手和双脚的抓放动作和身体的伸缩动作。由空压机提供压力动力。汽缸的运动由电磁阀来控制，电磁阀的动作间隔和时间由PC机来控制，可实现多个速度的爬行调整，并可通过参数变化调整爬行的高度。

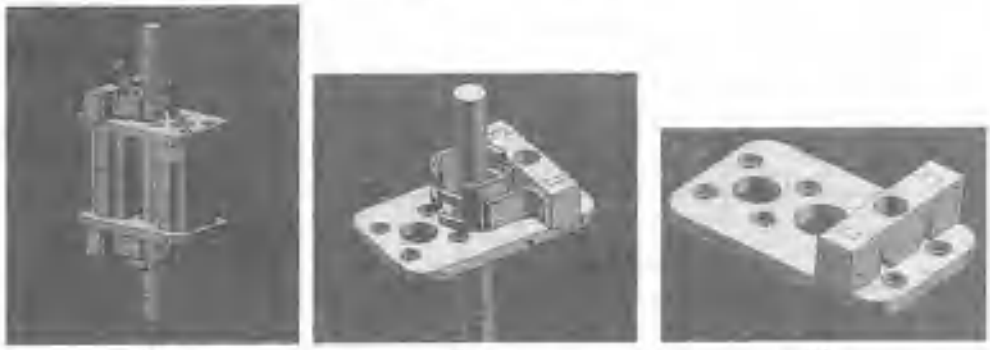
2.5.3 详细设计

整个设计工作由机械部分设计和控制部分设计所组成。

1. 机械部分设计

在设计过程中，由于采用了AUTODESK INVENTOR三维设计软件，能在整个机械结构和机构设计过程中直观地看到各个部分的形状、配合情况、装配位置的干涉情况等，省去了不必要的试制和修改过程而在短时间之内完成设计和制作。

机器人的结构部分是由相同的两组机构分别模仿猴子的双手和双脚的抱树动作和一个模仿猴子身体的伸缩动作的汽缸所组成，模仿抱树动作的机构是由四个零件组成的滑块机构[图2-17(a)、(b)、(c)]和一个汽缸组成。



(a) (b) (c)

图2-17 缆索机器人三维结构图

滑块机构由左右滑块、中间滑块和导向滑块所组成，它们由燕尾槽相互连接，左右滑块分别和一个手爪连接，中间滑块与下方的汽缸活塞通过螺钉连接，当活塞上下运动时，通过中间滑块的上下运动带动左右滑块进行左右运动，从而实现手爪的抓紧和放松动作。

手爪以及手爪的抱绳动作设计时，采用了模仿猴子抱树的设计理念。由于上面一双

“手”和下面一双“脚”是分别抓住绳索的，而绳索和树干是不同的，它有柔软性，为了避免绳索摇摆离开手爪的位置，在导板上设计了使绳索穿过的孔，并且在设计手爪的运动时考虑了极限位置，使手爪放开绳索时，绳索不离开手的范围，确保在下一次能够可靠地抓住绳索。实现爪的双向运动，采用了燕尾槽滑块机构，由爪连接件带动爪实现抓紧绳子的动作。

2. 气动系统控制部分设计

气动原理如图 2-18 所示。

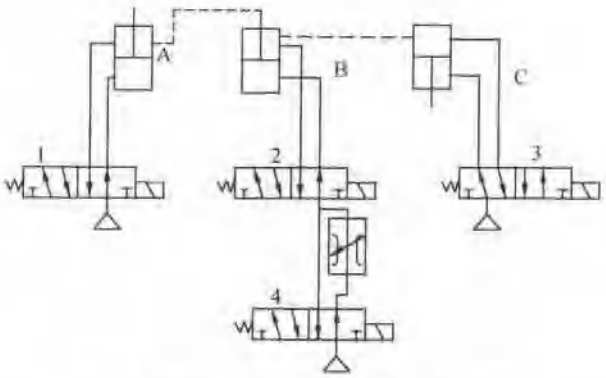


图 2-18　缆索机器人气动控制原理图

控制部分是由一台 PC 机、一台 AT89C2051 单片机以及控制软件组成，控制过程由 PC 机发出命令实现对电磁阀的实时控制。

电磁阀动作分析如表 2-4、表 2-5、表 2-6、表 2-7、表 2-8 所示(通电为＋，断电为－)。

表 2-4　上绳初始状态

动作步骤 \ 电磁阀	1	2	3
1	－	－	－
2	＋	－	＋

表 2-5　下绳状态

动作步骤 \ 电磁阀	1	2	3
1	－	－	－

表 2-6　上行过程

动作步骤 \ 电磁阀	1	2	3
1 初始	＋	－	＋
2 上抓松	－	－	＋
3 中间伸	－	＋	＋
4 上抓紧	＋	＋	＋
5 下抓松	＋	＋	－
6 中间收	＋	－	－

表 2-7 下行过程

动作步骤 \ 电磁阀	1	2	3
1 初始	+	—	+
2 下抓松	+	—	—
3 中间伸	+	+	—
4 下抓紧	+	+	+
5 上抓松	—	+	+
6 中间收	—	—	+

表 2-8 调速过程

动 作 \ 电磁阀	4
快	—
慢	+

3. 控制系统设计

本系统的微处理器采用 Atmel 公司的 AT89C2051(其外观及管脚排列如图 2-19 所示)。AT89C2051 是一个 2k 字节可编程 EPROM 的高性能微控制器。AT89C2051 有以下特点:2k 字节 EPROM、128 字节 RAM、15 根 I/O 线、2 个 16 位定时/计数器、5 个向量二级中断结构、1 个全双向的串行口,该微处理器内含精密模拟比较器和片内振荡器,具有4.25V至5.5V的电压工作范围和 12MHz/24MHz 工作频率,同时还具有加密阵列的二级程序存储器加锁、掉电和时钟电路等。此外,AT89C2051 还支持两种软件可选的电源节电方式。空闲时,CPU 停止,而让 RAM、定时/计数器、串行口和中断系统继续工作。可掉电保存 RAM 内容,使振荡器停振以禁止芯片所有的其他功能直到下一次硬件复位。

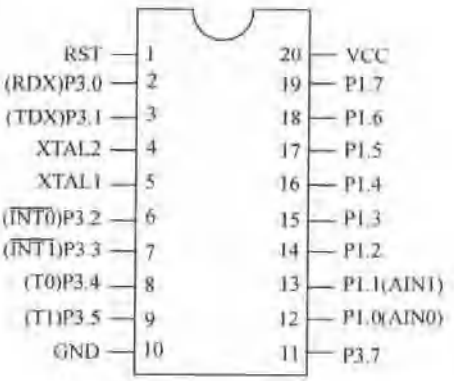


图 2-19 单片机引脚

AT89C2051 有两个 16 位计时/计数器寄存器 Timer0 和 Timer1。作为一个定时器,每个机器周期寄存器增加 1,这样寄存器即可计数机器周期。因为一个机器周期有 12 个振荡器周期,所以计数率是振荡器频率的 1/12。作为一个计数器,该寄存器在相应的外部输入脚 P3.4/T0 和 P3.5/T1 上出现从 1 至 0 的变化时增 1。由于需要两个机器周期来辨认一次 1 到 0 的变化,所以最大的计数率是振荡器频率的 1/24。

PC 机传送 8 位数据给单片机。用 C 语言编写界面,可以选择:1. 慢速上行、2. 慢速下行、3. 快速上行、4. 快速下行、5. 急停、6. 运动高度。PC 机实现与单片机的串行通信。PC 机中有标准 RS-232C 总线接口留给用户使用。RS-232C 是电子工业协会 EIA(electronic

industries association)从国际电话与电报顾问委员会 CCITT 远程通信控制标准中导出的一个标准。RS－232C 的位串行方式、数据格式与单机 UART 兼容，但是两者之间逻辑电平不同，单片机 UARMAX232 实现电平转换，使用 RS232 串行线。由于单片机的电流小，所以单片机与电磁阀之间加一三极管(型号：IRF540)。电磁阀选用的是 12V 三位两通型(4V210－08)。由单片机给信号实现电磁阀的自动切换。

4. **控制程序的编制**

本设计采用单片机编程控制。程序如下：

```
        MID1    BIT     P1.0
        MID2    BIT     P2.0
        DT1     BIT     P1.1
        DT2     BIT     P1.2
        DT3     BIT     P1.3
        DT4     BIT     P1.4
        SINGLE BIT   20H
        ORG     0000H
        AJMP    MAIN
        ORG     0003H
        AJMP    INT0_SER
        ORG     000BH
        AJMP    T0_SER
        ORG     0013H
        AJMP    INT1_SER
        ORG     001BH
        AJMP    T1_SER
        ORG     0023H
        AJMP    SERIAL_SER
MAIN:
        MOV     SP,#3FH
        MOV     P1,#00H
        CLR     SINGLE
START:
        ACALL   KEY
        JNB     SINGLE,START
        CLR     SINGLE
START1:
        ACALL   CONTROL
        AJMP    START1
KEY:
        ACALL   SCAN
```

```
        ACALL   DDEL
        JNZ     KEY0
        RET
KEY0:
        ACALL   DDEL
        ACALL   DDEL
        ACALL   SCAN
        JNZ     NEXT
        RET
NEXT:
        SETB    SINGLE
BACK:
        ACALL   SCAN
        JNZ     BACK
        RET
SCAN:
        MOV     A,P1
        ANL     A,#10H
        XRL     A,#10H
        RET
DDEL:
        MOV     R3,#5FH
DDEL0:
        MOV     R4,#0FFH
        DJNZ    R4,$
        DJNZ    R3,DDEL0
        RET
CONTROL:
        SETB    MID1,MID2
        ACALL   DEL
        CLR     MID1
        SETB    DT2
        ACALL   DEL
        CLR     MID2
        SETB    DT3
        ACALL   DEL
        CLR     DT2,DT3
        SETB    MID1,MID2
        ACALL   DEL
```

```
        CLR     MID1
        SETB    DT1
        ACALL   DEL
        CLR     MID2
        SETB    DT4
        ACALL   DEL
        CLR     DT1,DT4
        RET
DEL:
        MOV     R5,#100
DEL1:
        MOV     R6,#50
DEL2:
        MOV     R7,#100
DEL3:
        DJNZ    R7,DEL3
        DJNZ    R6,DEL2
        DJNZ    R5,DEL1
        RET
INT0_SER:
        RETI
T0_SER:
        RETI
INT1_SER:
        RETI
T1_SER:
        RETI
SERIAL_SER:
        RETI
END
```

习　题　2

1. 设计方法学研究的主要内容包括哪几个方面？
2. 何为设计系统？
3. 试述技术过程及其影响因素、技术系统及其结构。
4. 什么是功能分析法？试述其用途。
5. 试述“黑箱法”及其用途。
6. 试述创造力的构成及其影响因素、创造过程的基本阶段。

7. 何为创造性思维？试述创造性思维的基本特点及形式。
8. 试述常用的创造技法及其所适用的方面。
9. 创造性的思维活动具有哪些特点？列举出 2～3 个创造性的方法。
10. 说明在产品设计过程中，创造性设计主要体现在哪些环节上？
11. 用你身边的实例说明其创造性体现在哪些方面？

第 3 章　优化设计

本章首先介绍了优化设计基本概念、优化设计的数学基础，在此基础上进一步介绍了优化设计常用的几种基本方法：一维优化搜索方法、无约束优化方法及约束优化方法等。

3.1　概　　述

3.1.1　优化设计基本概念

在传统的设计中，很早就存在着“选优”的思想。设计人员可以根据需要同时提出几种不同的设计方案，通过分析评价后，选出较好的方案加以采用。这种选优的方案，在很大程度上带有经验性，即具有一定的局限性。传统的选优思想，受到了时间、条件（经费等）和经验的限制。这种选优的方案，只能称之为认可的方案。

在计算机应用之前，人们曾用经典的函数极小化概念，处理简单结构的优化设计问题。对于工程问题的复杂性，这种理论在实际上的应用，受到了限制。自计算机问世后，设计才从传统的设计方法走上了优化设计方法。概括地说，优化设计就是以数学规划理论为基础，以计算机为工具的一种优化设计参数的现代设计方法。

由于最优化技术和计算机技术在设计领域的应用，优化设计为工程设计提供了一种重要的科学设计方法，使得在解决复杂设计问题时，能从众多的设计方案中寻到尽可能完善的或最适宜的设计方案。采用这种设计方法能大大提高设计效率和设计质量。

目前优化设计方法不仅用于结构设计、化工系统设计、电气传动设计、制造工艺设计，也用于运输路线的确定、商品流通量的调配、产品配方的配比等，而且取得了不少成绩。优化设计理论与方法最大的特点是把经验的、感性的、类比的传统设计方法转变为科学的、理性的，立足于计算分析的设计方法。特别是近年来，随着有限元法、可靠性设计、计算机辅助设计理论和方法的发展与优化设计方法的结合应用，使整个设计过程逐步向自动化、集成化、智能化方向发展。因而，学习、掌握和应用优化设计理论与方法对工程设计人员来说是十分必要的。

下面通过几个简单例子来说明优化设计的基本概念。

例 3-1　如图 3-1 所示，有一空心等截面简支柱，两端承受轴向压力 $P=22680\text{N}$，柱高 $l=254\text{cm}$，材料为铝合金，弹模模量 $E=7.03\times10^{4}\text{MPa}$，密度 $\rho=2.768\times10^{3}\text{kg/m}^3$，许用应力 $[\sigma]=140\text{MPa}$。截面的平均直径 $D=(D_0+D_1)/2$，并不应大于 8.9cm，壁厚 δ 不小于 0.1cm。现要求设计最小质量的柱子，问 D 与 δ 值应多少？

若以柱子的结构参数 D 为横坐标，δ 为纵坐标，由于 D 与 δ 只允许取正值，因此可用直角坐标的第一象限来表示它们的设计关系。如图 3-2 所示直线 a-a 以左是 $D\leqslant8.9\text{cm}$ 的区域，直线 b-b 以上是 $\delta\geqslant0.1\text{cm}$ 的区域，这些是结构参数选择的边界限制区域。

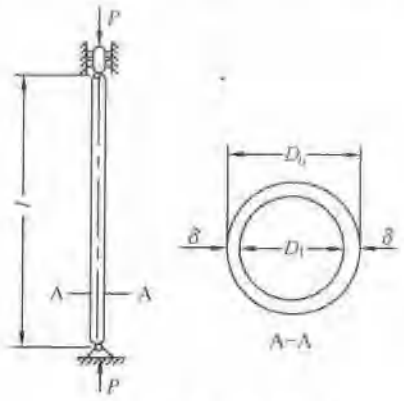


图 3-1 空心简支柱

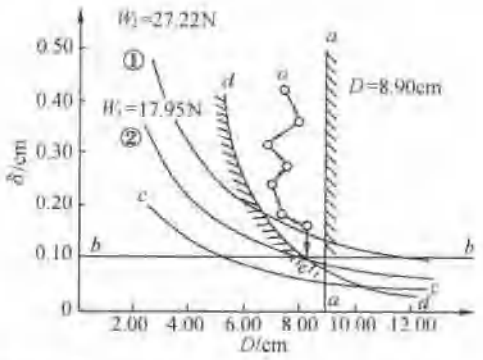


图 3-2 空心简支柱限制条件的几何图形

除上述限制条件以外，还应保证管柱有足够的承压强度和足够的整体稳定性。为此应使柱内的工作压应力 $\sigma=\dfrac{P}{\pi D\delta}$ 小于许用应力[σ]，即

$$\frac{22680}{\pi D\delta\times 10^2}-140\leqslant 0 \tag{3-1}$$

或

$$D\delta-0.516\geqslant 0$$

这个方程式的函数曲线为 c-c，所以要使管柱有足够的强度，其 D 和 δ 应在曲线 c-c 的右上方取值。

根据压杆稳定要求，柱内的工作压应力 σ 应小于管柱的稳定临界应力 $\sigma_c=\dfrac{\pi^2E}{8l^2}(D^2+\delta^2)\approx\dfrac{\pi^2E}{8l^2}D^2$（假定壁厚 δ 远远小于平均直径 D）。于是得

$$\frac{22680}{\pi D\delta}-\frac{7.03\times 10^4}{\dfrac{8\times 254^2}{10^2}}\pi^2D^2\leqslant 0 \tag{3-2}$$

或

$$1.35D^2-\frac{72.2}{D\delta}\geqslant 0$$

这个方程式的函数曲线为图 3-2 中的 d-d，位于此曲线右上方的 D 和 δ 值均可满足压杆稳定性要求。至此，满足几何尺寸限制条件，以及管柱强度、整体稳定性应力限制条件的区域已全部用几何图形清楚地表示出来，一切满足上述各项限制条件的结构参数 D 和 δ 的组合应在阴影线的区域内。

由于要求设计最小重量的压柱，而它的重量 W 可表示为结构参数 D、δ 的函数，即

$$W=\rho l\pi D\delta=2.2D\delta \tag{3-3}$$

所以，若将它赋予不同的重量，如 $W=17.95$N、27.22N、…则可以在图上画出等重曲线①、②等，这样就可以发现，在上述可行区域内，其最轻的等重曲线与压杆稳定的上限曲线、管子壁厚 δ 的下限曲线交于 e 点。由于问题比较简单，这点的值是很容易求得的，只要将壁厚的下限值代入稳定性限制条件式，即可求出最优的压柱设计方案：

$$W=1.786\text{kg},\ D=8.117\text{cm},\ \delta=0.1\text{cm}$$

当然，实际上求最优设计方案并不如此简单，一般需要用优化设计方法利用电子计算机

来解，即从一个不是很好的设计方案（o 点）开始，沿着使压柱重量减轻的方向，不断进行搜索，直至找到压柱重量最轻而又不违反所有限制条件的最优方案（即 e 点或接近 e 点）为止。其搜索路线如图 3-2 所示。

例 3-2　机床主轴结构的优化设计。如图 3-3 所示是一个机床主轴的典型结构原理图。对于这类问题，目前可以采用有限元法，利用状态方程来计算轴端变形 y 和固有频率 ω。

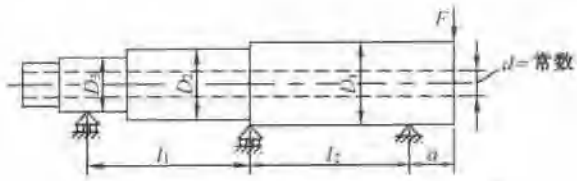


图 3-3　机床主轴的典型结构原理图

优化设计的任务是确定 D_i、l_i 和 a，保证 y 和 ω 在允许范围内，使结构的质量最轻。这时，问题归结为：求 D_i、l_i、a 的值，使质量 $f(D_i,l_i)=\rho\pi[\Sigma(D_i^2-d^2)l_i+(D_n^2-d^2)a]$ 为最小，并满足

$$y\leqslant[y]$$

$$\omega^2\geqslant\omega_0^2$$

$$D_{i\min}\leqslant D_i\leqslant D_{i\max}\quad(i=1,2,\cdots,n)$$

$$l_{i\min}\leqslant l_i\leqslant l_{i\max}$$

$$a_{\min}\leqslant a\leqslant a_{\max}$$

$$N_{\min}\leqslant\frac{l_i}{a}\leqslant N_{\max}$$

式中：ρ ——材料的密度；

D_i、l_i ——阶梯形主轴的外径和对应的长度；

D_n ——与 a 对应的外径。

在主轴结构动力优化设计时，也可取由振型和质量确定的能耗为目标函数。约束条件可以取激振力频率避开$(1+20\%)\omega$ 的禁区范围。

通过上述两个例子，我们可以获得“优化设计”的最基本的概念，即解决设计方案参数的最佳选择问题。这种选择不仅保证多参数的组合方案满足各种设计要求，而且又使设计指标达到最优值。因此，求解优化设计问题需要采用优化方法。简言之，就是在一些等式或不等式约束条件下求多变量函数的极小值或极大值。

这样，一个优化设计的过程及其相互关系可以用图 3-4 来表示。

由上述框图可以看出整个优化设计过程可以分为两部分，虚线框 I 表示首先对机械设计问题建立优化设计的数学模型，然后选用某一种优化方法；虚线框 I 表示利用电子计算机的自动计算过程，包括程序的编制、数据的准备与结果的分析和整理。框(3)是机器的内评估，它是利用电子计算机的逻辑判断能力来评估机械设计的优化目标是否达到最优值。框(7)为优化途径，即优选设计参数，它利用优化方法按照一定的方向与步长一步一步来搜索最优值 x^*（或 x_{opt}）。因此优化设计是一个反复循环的计算过程。这样一种设计计算，不借助电子计算机的自动计算是很难完成的。

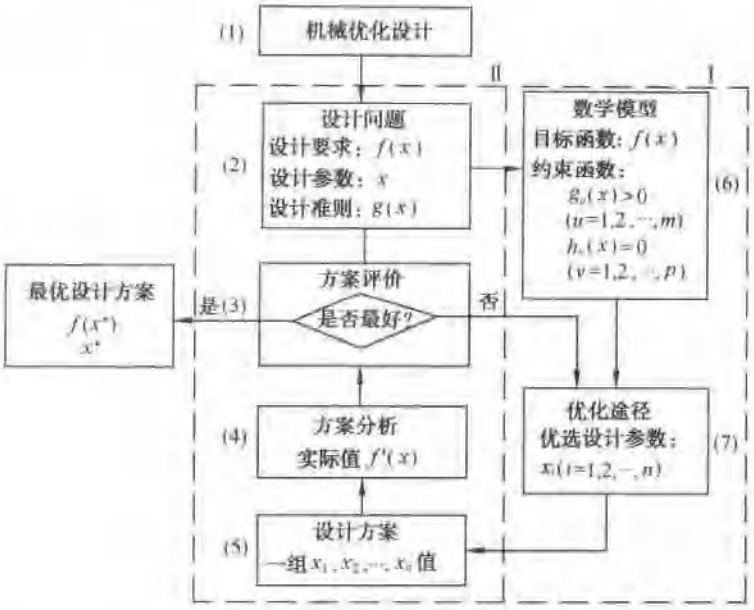


图 3-4 优化设计的过程框图

应用计算机进行优化设计，与传统设计相比，具有如下三个特点：

(1) 设计的思想是最优设计，需要建立一个能正确反映设计问题的数学模型；

(2) 设计的方法是优化方法，一个方案参数的调整是计算机沿着使方案更好的方向自动进行的，从而选出最优方案；

(3) 设计的手段是计算机，由于计算机的运算速度快，分析和计算一个方案只要几秒以至千分之一秒，因而可以从大量的方案中选出"最优方案"。这种设计是设计方法上的一个很大的变革，它使许多复杂的问题得到最完善的解决，它不但可以提高设计效率、缩短设计周期，而且还可以为设计人员提供大量的设计分析数据，有助于考察设计结果，从而可以提高产品的设计质量。

同时，通过上述例子我们可看出，对于任何实际优化设计问题，我们可采用先将其进行数学抽象化，即建立优化数学模型，通过对数学模型的求解获得优化设计的结果。下面我们先来探讨如何建立优化数学模型问题。

3.1.2 优化设计的数学模型

进行实际问题的优化设计，首先需要建立其数学模型。优化设计的数学模型需要用设计变量、设计约束和目标函数等基本概念才能予以完整地描述。

1. 设计变量

在工业及产品设计中，一个零部件或一台机器的设计方案，常用一组基本参数来表示。概括起来参数可分为两类：一类是按照具体设计要求事先给定，且在设计过程中保持不变的参数，称为设计常量；另一类是在设计过程中须经不断调整，以确定其最优值的参数，称为设计变量。也就是说，设计变量是优化设计要优选的量。优化设计的任务就是确定设计变量的最优值以得到最优设计方案。

由于设计对象不同，选取的设计变量也不同。它可以是几何参数，如零件外形尺寸、截面尺寸、机构的运动尺寸等；也可以是某些物理量，如零部件的重量、体积、力与力矩、惯性矩等；还可以是代表工作性能的导出量，如应力、变形等。总之，设计变量必须是对该项设计性能指标优劣有影响的参数。

设计变量是一组相互独立的基本参数。可用向量 $\boldsymbol{X}$ 来表示

$$\boldsymbol{X}=[x_1 \quad x_2 \quad \cdots \quad x_n]^{\mathrm{T}} \tag{3-4}$$

式中 x_i 表示 n 维向量 $\boldsymbol{X}$ 的第 i 个分量。

设计变量的每一个分量都是相互独立的。以 n 个设计变量为坐标轴所构成的实数空间称为设计空间，或称 n 维欧氏空间，用 $\boldsymbol{R}^n$ 表示。当 $n=2$ 时，$\boldsymbol{X}=[x_1 \quad x_2]^{\mathrm{T}}$，是二维设计向量；当 $n=3$ 时，$\boldsymbol{X}=[x_1 \quad x_2 \quad x_3]^{\mathrm{T}}$ 为三维设计向量，设计变量 x_1、x_2、x_3 组成一个三维空间；当 $n>3$ 时，设计空间是一个想像的超越空间，称 n 维实数空间。其中二维和三维设计空间如图 3-5 所示。

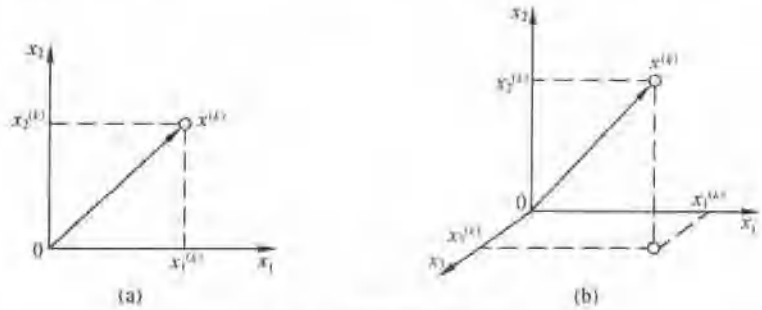


图 3-5 设计空间

设计空间是所有设计方案的集合，用符号 $\boldsymbol{X}\in\boldsymbol{R}^n$ 表示。任何一个设计方案，都可以看做是从设计空间原点出发的一个设计向量 $\boldsymbol{X}^{(k)}$，该向量端点的坐标值就是这一组设计变量 $\boldsymbol{X}^{(k)}=[x_1^{(k)} \quad x_2^{(k)} \quad \cdots \quad x_n^{(k)}]^{\mathrm{T}}$。因此，一组设计变量表示一个设计方案，它与一个向量的端点相对应，也称设计点。而设计点的集合即构成了设计空间。

根据设计变量的多少，一般将优化设计问题分为三种类型：当设计变量数目 $n<10$ 的称为小型优化问题；$n=10\sim50$ 的称为中型优化问题；$n>50$ 的称为大型优化问题。

在工程优化设计中，根据设计要求，设计变量常有连续量和离散量之分。大多数情况下，设计变量是有界连续变化型量，称为连续设计变量。但在一些情况下，有些设计变量是离散型量，则称离散设计变量，如齿轮的齿数、模数，钢管的直径，钢板的厚度等。对于离散设计变量，在优化设计过程中常是先把它视为连续量，在求得连续量的优化结果后再进行圆整或标准化，以求得一个实用的最优设计方案。

2. 目标函数

优化设计的目的就是要在所有的可行设计方案中寻求一个最优方案。那么，什么是最优呢？这里必然有一个评价优劣的标准。对于不同的优化设计问题，评价优劣的标准各不相同，也就是追求的目标不相同。例如，机械优化设计的目标常常可用重量最轻、体积最小、成本最低、用料最省、利润最高、产值最大、寿命最长、可靠性最好、机械技术性最佳等来标志。若把这个"标准"表示为设计变量的可计算函数，优化这个函数，则可以取得最优设计方

案。在优化设计中，这一用于评选设计方案的函数，称为目标函数，它是设计变量的函数，记做

$$f(\boldsymbol{X})=f(x_1,x_2,\cdots,x_n) \tag{3-5}$$

目标函数是一个标量函数。由于目标函数值的大小可以评价设计质量的优劣，所以也称为评价函数。

优化设计就是要寻求一个最优设计方案，即最优点 X^*，从而使目标函数达到最优值 $f(\boldsymbol{X}^*)$。在优化设计中，一般取最优值为目标函数的最小值，记做

$$f(\boldsymbol{X}^*)=\min f(\boldsymbol{X}),\boldsymbol{X}\in\boldsymbol{R}^n \tag{3-6}$$

确定目标函数是优化设计中最重要的决策之一。因为这不仅直接影响优化方案的质量，而且还影响到优化过程。目标函数可以根据工程问题的要求从不同角度来建立，如成本、重量、几何尺寸、运动轨迹、功率、应力、动力特性等。

一个优化问题，可以用一个目标函数来衡量，称之为单目标优化问题；也可以用多个目标函数来衡量，称之为多目标优化问题。单目标优化问题，由于指标单一，易于衡量设计方案的优劣，求解过程比较简单明确；而多目标优化问题求解比较复杂，但可获得更佳的最优设计方案。

目标函数可以通过等值线（面）在设计空间中表现出来。所谓目标函数的等值线（面），就是当目标函数 $f(\boldsymbol{X})$ 的值依次等于一系列常数 $c_i(i=1,2,\cdots)$ 时，设计变量 $\boldsymbol{X}$ 取得一系列值的集合。以二维为例，如图 3-6所示，二维目标函数 $f(x_1,x_2)$ 在 x_1、x_2 和以 $f(\boldsymbol{X})$ 为坐标轴的空间内是一个曲面。显然在二维的设计平面 x_1Ox_2 中，每一个点 (x_1,x_2) 都有一个相应的目标函数值 $f(x_1,x_2)$，它在图中表示沿 $f(\boldsymbol{X})$ 轴方向的高度。将 $f(x_1,x_2)$ 曲面上具有相同高度点投影到设计平面 x_1Ox_2 上，则得到 $f(x_1,x_2)=c$ 的平面曲线，称此是目标函数的等值线。显然，当给定一系列不同的 c 值时，可以得到一组平面曲线：$f(x_1,x_2)=c_1,c_2,\cdots$ 这组曲线就构成了目标函数的等值线族。

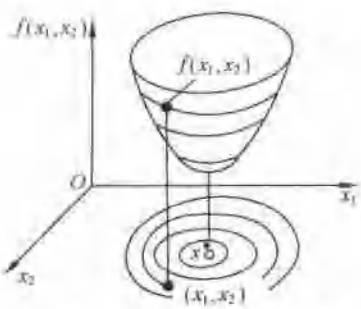


图 3-6 二维目标函数等值线

等值线有以下特点：

(1) 不同值的等值线不相交；

(2) 除极值点外，在设计空间内，等值线不会中断；

(3) 等值线反映了目标函数值的变化规律，愈内层的等值线，其函数值愈小，等值线族的中心点就是目标函数的极值点，因此，求目标函数的极值点也就是求等值线族的共同形心问题；

(4) 等值线的间隔愈密，表示该处函数值的变化率愈大，否则变化愈小；

(5) 一般在极值点附近，等值线近似地呈同心椭圆族，极值点就是椭圆的中心点。

在设计空间内，目标函数值相等点的连线，对二维问题，构成了等值线；对三维问题构成了等值面；对四维以上问题，则构成了等值超曲面。

3. 设计约束

在优化设计中，对设计变量选取的限制条件，称为设计约束（或约束条件）。其形式常常

是用不等式或等式来表示，所以设计约束有不等式约束和等式约束两种形式，即

不等式约束　　$g_u(\boldsymbol{X})\leqslant 0, u=1,2,\cdots,m$

等式约束　　$h_v(\boldsymbol{X})=0, v=1,2,\cdots,p$

设计约束若按性质又可分为两类：

(1) 边界约束，它是规定设计变量的取值范围，即取值许可范围的上、下限，如要求构件长度 l 应满足：$l_{\min}\leqslant l\leqslant l_{\max}$ 等；

(2) 性能约束，指对设计对象的某种设计性能或指标提出的约束条件，如对零件工作应力、变形的限制，以及对运动学参数如移位、速度、加速度值的限制等。设计约束的几何意义是它将设计空间一分为二，形成了可行域和非可行域。以二维问题为例，如图 3-7 所示，每一个不等式约束[如 $g_i(\boldsymbol{X})\leqslant 0$]都把设计空间划分为两部分：一部分是满足该不等式约束条件的，即 $g_i(\boldsymbol{X})<0$；另一部分则不满足，即 $g_j(\boldsymbol{X})>0$。两部分的分界面称为约束面(或约束边界)，即由 $g_j(\boldsymbol{X})=0$ 的点集构成。在二维设计空间中约束面是一条曲线或直线，在三维及以上的设计空间中则是一个曲面或超曲面。一个优化设计问题的所有不等式约束的边界将组成一个复合约束边界(图 3-7)。其约束边界所包围的区域(图中阴影线内)是设计空间中满足所有不等式约束条件的部分，在这个区域中所选择的设计变量是允许采用的，称该区域为可行设计域，简称可行域，记做

$$D=\{\boldsymbol{X}\mid g_u(\boldsymbol{X})\leqslant 0, u=1,2,\cdots,m\} \tag{3-7}$$

当优化设计问题除有 m 个不等式约束条件外，还应满足 p 个等式条件时，即对设计变量的选择又增加了限制，如图 3-7 所示。当有一个等式约束条件 $h(x_1,x_2)=0$ 时，其可行设计方案只允许在 D 域内的等式约束函数曲线 AB 段上选择，因此，在一般情况下，其可行设计域可表示为

$$D=\{\boldsymbol{X}\mid g_u(\boldsymbol{X})\leqslant 0, u=1,2,\cdots,m;\ h_v(\boldsymbol{X})=0, v=1,2,\cdots,p\} \tag{3-8}$$

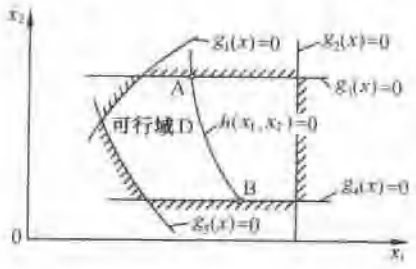


图 3-7　二维问题可行域

与此相反，除去可行域以外的设计空间称为非可行设计区域，简称非可行域。据此，凡在可行域内的任一设计点都代表了一允许采用的设计方案，这样的设计点称可行设计方案，简称可行点(或内点)。处于不等式约束边界上[即不等式约束的极限条件 $g_j(\boldsymbol{X})=0$]的设计点，称为边界设计点。边界设计点也是可行点，不过它是一个为该项约束所允许的极限设计方案，所以又称极限设计点。除此之外，在可行域外的点，称为非可行点，或称外点。非可行点即为不允许采用的非可行设计方案。

4. 数学模型的表达式

优化设计问题的数学模型是实际工程设计问题的具体抽象，是反映各主要因素之间内在联系的一种数学形态。在确定出设计变量、设计约束和列出目标函数后，即可建立优化设计问题的数学模型。

设某优化设计问题有 n 个设计变量，即 $\boldsymbol{X}=[x_1\quad x_2\quad \cdots\quad x_n]^{\mathrm{T}}$ 在满足 $g_u(\boldsymbol{X})\leqslant 0(u=1,2,\cdots,m)$ 和 $h_v(\boldsymbol{X})=0(v=1,2,\cdots,p)$ 的约束条件下，使目标函数 $f(\boldsymbol{X})$ 值达到最小。可简记为

$$\begin{cases} \min f(\boldsymbol{X}) & \boldsymbol{X}\in \boldsymbol{R}^n \\ \text{s.t.} \quad g_u(\boldsymbol{X})\leqslant 0, u=1,2,\cdots,m \\ h_v(\boldsymbol{X})\leqslant 0, v=1,2,\cdots,p \end{cases} \tag{3-9}$$

上式就是优化设计数学模型的一般表达式。其中 s.t. 为 subjected(受约束于)的缩写。这一优化设计数学模型，称为约束优化设计问题。

若式(3-9)所列数学模型内 $m=p=0$，则成为

$$\begin{cases} \min f(\boldsymbol{X}) \\ \boldsymbol{X}\in \boldsymbol{R}^n \end{cases} \tag{3-10}$$

即这一优化问题不受任何约束，称为无约束优化设计问题。式(3-10)即为无约束优化设计问题的数学模型表达式。

5. 优化问题的最优解

按已建立的数学模型，可求得优化问题的最优解

最优方案　　$\boldsymbol{X}^*=[x_1^*\quad x_2^*\quad \cdots\quad x_n^*]^{\mathrm{T}}$

最优值　　$f(\boldsymbol{X}^*)$

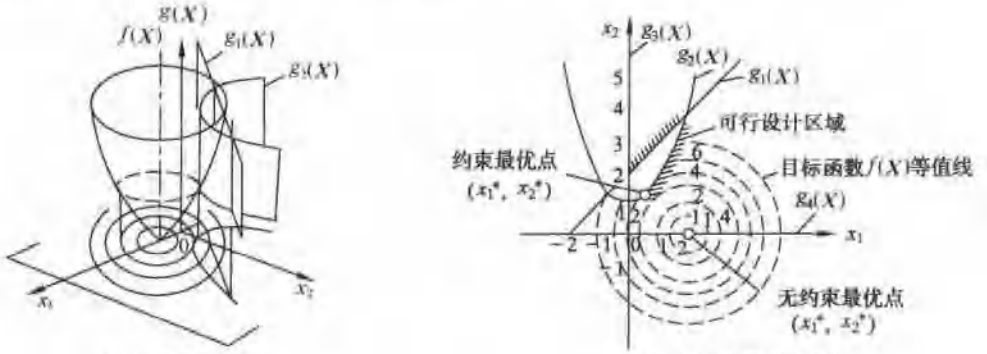
所以，优化问题的最优解由两部分组成：一是最优点 $\boldsymbol{X}^*$，即最优设计点，或称最优设计方案；二是最优值 $f(\boldsymbol{X}^*)$，是最优点 $\boldsymbol{X}^*$ 代入目标函数 $f(\boldsymbol{X})$ 所求得的最优函数值，它是评价设计点优劣程度的一个标量值。

3.1.3　优化设计的迭代算法

对于优化问题数学模型的求解，目前可采用的求解方法有三种：即数学解析法、图解法和数值迭代法。

数学解析法就是把优化对象用数学模型描述出来后，用数学解析法(如微分、变分法等)来求出最优解，如高等数学中求函数极值或条件极值的方法。数学解析法是优化设计的理论基础，但它仅限于维数较少且易求导的优化问题的求解。

图解法就是直接利用作图的方法来求解优化问题，通过画出目标函数和约束函数的图形，求出最优解。此法的特点是简单直观，但仅限于 $n\leqslant 2$ 的低维优化问题的求解。图 3-8 所示即为采用图解法来求解如下二维优化问题。



(a) 问题的立体图　　(b) 设计空间关系图

图 3-8　二维优化问题的几何解

图 3-8(a)所示为该问题的目标函数、约束函数的立体图；图 3-8(b)所示为该问题的设计空间关系图，阴影线部分即为由所有约束边界围成的可行域。该问题的约束最优点就是约束边界 $g_2(\boldsymbol{X})=0$ 与目标函数等值线的切点，即图中的 $\boldsymbol{X}^*$ 点，$\boldsymbol{X}^*=[x_1^* \quad x_2^*]^{\mathrm{T}}=[0.58 \quad 1.34]^{\mathrm{T}}$，其目标函数极小值 $f(x)=0.38$。

数值迭代法完全是依赖于计算机的数值计算特点而产生的，它是具有一定逻辑结构并按一定格式反复迭代计算、逐步逼近优化问题最优解的一种方法。采用数值迭代法可以求解各种优化问题(包括数学解析法和图解法不能适用的优化问题)。

1. 数值迭代法的迭代格式

数值迭代法的基本思想是：搜索、迭代、逼近。为了寻找目标函数 $f(\boldsymbol{X})$ 的极小点 $\boldsymbol{X}^*$，首先在设计空间中给出一个估算的初始设计点 $\boldsymbol{X}^{(0)}$，然后从该点出发，按照一定的规则确定适当的搜索方向 $\boldsymbol{S}^{(0)}$ 和搜索步长 $\alpha^{(0)}$，求得第一个改进设计点 $\boldsymbol{X}^{(1)}$，它应满足条件：$f(\boldsymbol{X}^{(1)})<f(\boldsymbol{X}^{(0)})$，至此完成第一次迭代。之后，又以 $\boldsymbol{X}^{(1)}$ 为新的初始点，重复上述步骤，求得 $\boldsymbol{X}^{(2)}$，…，如此反复迭代，从而获得一个不断改进的点列$\{\boldsymbol{X}^{(k)},k=0,1,2,\cdots\}$以及一相应的递减函数值数列$\{f(\boldsymbol{X}^{(k)}),k=0,1,2,\cdots\}$。这一迭代过程若用数学式子来表达，即得数值迭代法的基本迭代算式为

$$\begin{cases}\boldsymbol{X}^{(k+1)}=\boldsymbol{X}^{(k)}+\alpha^{(k)}\boldsymbol{S}^{(k)} & (k=0,1,2,\cdots)\\ f(\boldsymbol{X}^{(k+1)})<f(\boldsymbol{X}^{(k)}) & \\ g_u(\boldsymbol{X}^{(k+1)})\leqslant 0 & (u=1,2,\cdots,m)\end{cases} \tag{3-11}$$

式中：$\boldsymbol{X}^{(k)}$——前一步已取得的设计方案(迭代点)；

$\boldsymbol{X}^{(k+1)}$——新的改进设计方案(新的迭代点)；

$\boldsymbol{S}^{(k)}$——第 k 次迭代计算的搜索方向；

$\alpha^{(k)}$——第 k 次迭代计算的步长因子。

这样一步一步地重复数值计算，不断地用改进了的新设计点迭代前次设计点，逐步改进目标函数值并最终逼近极值点——即极小点 $\boldsymbol{X}^*$。数值迭代法的迭代过程如图 3-9 所示。

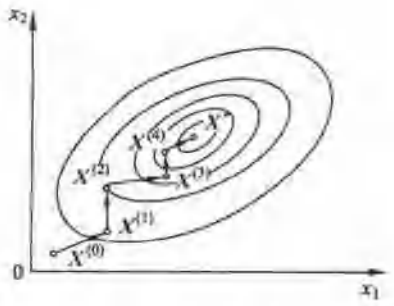


图 3-9　二维优化问题的迭代过程

从以上分析可见，优化迭代过程所得一系列迭代点都是以同一基本迭代格式进行重复运算所获得的，因而在计算机上很容易实现，而且由于每一次迭代取得的新的可行迭代点之目标函数值有所下降，于是迭代点不断地向约束最优点靠拢，最后必将到达十分逼近理论最优点的近似最优点 $\boldsymbol{X}^*$。

通过以上分析及其图 3-9 可知，要运用数值迭代法寻找到目标函数的极小值 $\boldsymbol{X}^*$，这里关键要解决三个问题：一是如何确定迭代步长 $\alpha^{(k)}$；二是怎样选定搜索方向 $\boldsymbol{S}^{(k)}$；三是如何判断是否找到了最优点，以终止迭代。由于在优化技术中，关于迭代方法有多种，它们之间的区别就在于确定 $\alpha^{(k)}$ 和 $\boldsymbol{S}^{(k)}$ 的方式不同。特别是 $\boldsymbol{S}^{(k)}$ 的确定，在各种方法中起着关键性的作用。关于前两个问题的确定我们将在以后各节介绍。

2. 迭代计算的终止准则

目标函数下降，但总应有个停止迭代的标准，标准就是终止准则。常用的迭代终止准则

有如下几种。

(1) 点距足够小准则：

$$\| \boldsymbol{X}^{(k-1)} - \boldsymbol{X}^{(k)} \| \leqslant \varepsilon \tag{3-12}$$

式中 ε 表示给定的计算精度，它是一个足够小的正数。

$\| \boldsymbol{X}^{(k+1)} - \boldsymbol{X}^{(k)} \|$ 为向量 $X^{(k+1)} - X^{(k)}$ 的模。它可用下式计算

$$\| \boldsymbol{X}^{(k+1)} - \boldsymbol{X}^{(k)} \| = \{ \sum_{i=1}^{n} [\boldsymbol{X}_i^{(k+1)} - \boldsymbol{X}_i^{(k)}]^2 \}^{1/2} \tag{3-13}$$

(2) 函数下降量足够小准则：

$$| f(\boldsymbol{X}^{(k+1)}) - f(\boldsymbol{X}^{(k)}) | < \varepsilon \tag{3-14}$$

(3) 函数梯度充分小准则：

$$\| \nabla f(\boldsymbol{X}^{(k+1)}) \| \leqslant \varepsilon \tag{3-15}$$

这是由于函数极值点的必要条件是函数在这一点的梯度值的模为零。因此当迭代点的函数梯度的模已充分小时，则认为迭代可以终止。

只要满足以上三准则中之一，就可以认为目标函数值 $f(\boldsymbol{X}^{(k+1)})$ 已收敛于其极小值。

3.2 工程优化的数学基础

工程优化设计中绝大多数是多变量有约束的非线性规划问题，即是求解多变量非线性函数的极值问题。由此可见，优化设计是建立在多元函数的极值理论基础上的，对于无约束优化问题为数学上的无条件极值问题，而对于约束优化问题则为数学上的条件极值问题。为便于后续优化方法的学习，有必要研究这些非线性函数的性质和变化规律。

3.2.1 函数的方向导数与梯度

1. 函数的方向导数

我们知道，一个二元函数 $f(x_1, x_2)$ 在点 $\boldsymbol{X}^{(0)}(x_1^{(0)}, x_2^{(0)})$ 处的偏导数，其定义是

$$\frac{\partial f(\boldsymbol{X}^{(0)})}{\partial x_1} = \lim_{\Delta x_1 \to 0} \frac{f(x_1^{(0)} + \Delta x_1, x_2^{(0)}) - f(x_1^{(0)}, x_2^{(0)})}{\Delta x_1}$$

$$\frac{\partial f(\boldsymbol{X}^{(0)})}{\partial x_2} = \lim_{\Delta x_2 \to 0} \frac{f(x_1^{(0)}, x_2^{(0)} + \Delta x_2) - f(x_1^{(0)}, x_2^{(0)})}{\Delta x_2}$$

而 $\frac{\partial f(\boldsymbol{X}^{(0)})}{\partial x_1}$ 和 $\frac{\partial f(\boldsymbol{X}^{(0)})}{\partial x_2}$ 分别是函数 $f(x_1, x_2)$ 在点 $\boldsymbol{X}^{(0)}$ 处沿坐标轴 x_1 和 x_2 方向的变化率。因此，函数 $f(x_1, x_2)$ 在点 $\boldsymbol{X}^{(0)}$ 处沿某一方向 $\boldsymbol{S}$ 的变化率如图 3-10 所示，且 $\rho = \sqrt{(\Delta x_1)^2 + (\Delta x_2)^2}$，其定义式应为

$$\frac{\partial f(\boldsymbol{X}^{(0)})}{\partial \boldsymbol{S}} = \lim_{\rho \to 0} \frac{f(x_1^{(0)} + \Delta x_1, x_2^{(0)} + \Delta x_2) - f(x_1^{(0)}, x_2^{(0)})}{\rho}$$

称它为该函数沿此方向的导数。据此，偏导数 $\frac{\partial f(\boldsymbol{X}^{(0)})}{\partial x_1}$ 和 $\frac{\partial f(\boldsymbol{X}^{(0)})}{\partial x_2}$ 也可看成是函数 $f(x_1, x_2)$ 分别沿 x_1 和 x_2

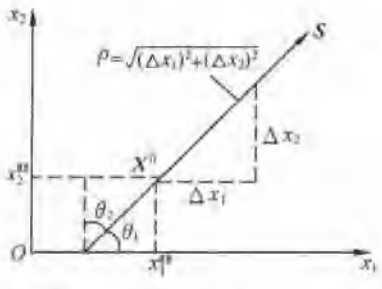


图 3-10 二维空间中的方向

坐标轴方向的方向导数。所以，方向导数是偏导数概念的推广，偏导数是方向导数的特例。

方向导数与偏导数之间的数量关系，可从下述推导中求得

$$\begin{aligned}\frac{\partial f(\boldsymbol{X}^{(0)})}{\partial \boldsymbol{S}}&=\lim_{\rho\to 0}\frac{f(x_1^{(0)}+\Delta x_1,x_2^{(0)}+\Delta x_2)-f(x_1^{(0)},x_2^{(0)})}{\rho}\\&=\lim_{\rho\to 0}\frac{f(x_1^{(0)}+\Delta x_1,x_2^{(0)})-f(x_1^{(0)},x_2^{(0)})}{\Delta x_1}\frac{\Delta x_1}{\rho}\\&+\lim_{\rho\to 0}\frac{f(x_1^{(0)}+\Delta x_1,x_2^{(0)}+\Delta x_2)-f(x_1^{(0)}+\Delta x_1,x_2^{(0)})}{\Delta x_2}\frac{\Delta x_2}{\rho}\\&=\frac{\partial f(\boldsymbol{X}^{(0)})}{\partial x_1}\cos\theta_1+\frac{\partial f(\boldsymbol{X}^{(0)})}{\partial x_2}\cos\theta_2\end{aligned}\tag{3-16}$$

同样，一个三元函数 $f(x_1,x_2,x_3)$ 在 $\boldsymbol{X}^{(0)}(x_1^{(0)},x_2^{(0)},x_3^{(0)})$ 点处沿 $\boldsymbol{S}$ 方向的方向导数 $\dfrac{\partial f(\boldsymbol{X}^{(0)})}{\partial \boldsymbol{S}}$ 如图 3-11 所示，可类似地表示为如下形式

$$\begin{aligned}\frac{\partial f(\boldsymbol{X}^{(0)})}{\partial \boldsymbol{S}}&=\frac{\partial f(\boldsymbol{X}^{(0)})}{\partial x_1}\cos\theta_1\\&+\frac{\partial f(\boldsymbol{X}^{(0)})}{\partial x_2}\cos\theta_2+\frac{\partial f(\boldsymbol{X}^{(0)})}{\partial x_3}\cos\theta_3\end{aligned}\tag{3-17}$$

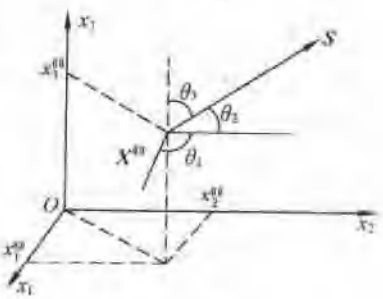


图 3-11　三维空间中的方向

依此类推，可得到 n 元函数 $f(x_1,x_2,\cdots,x_n)$ 在点 $\boldsymbol{X}^{(0)}(x_1^{(0)},x_2^{(0)},\cdots,x_n^{(0)})$ 处沿 $\boldsymbol{S}$ 方向的方向导数

$$\frac{\partial f(\boldsymbol{X}^{(0)})}{\partial \boldsymbol{S}}=\frac{\partial f(\boldsymbol{X}^{(0)})}{\partial x_1}\cos\theta_1+\frac{\partial f(\boldsymbol{X}^{(0)})}{\partial x_2}\cos\theta_2+\cdots+\frac{\partial f(\boldsymbol{X}^{(0)})}{\partial x_n}\cos\theta_n=\sum_{i=1}^{n}\frac{\partial f(\boldsymbol{X}^{(0)})}{\partial x_i}\cos\theta_i\tag{3-18}$$

式中，$\cos\theta_i$ 为 $\boldsymbol{S}$ 方向与坐标轴 x_i 方向之间夹角的余弦。

2. 函数的梯度

函数 $f(\boldsymbol{X})$ 在某点 $\boldsymbol{X}$ 的方向导数表明函数沿某一方向 $\boldsymbol{S}$ 的变化率。一般来说，函数在某一确定点沿不同方向的变化率是不同的。为求得函数在某点 $\boldsymbol{X}$ 的方向导数为最大的方向，特引入梯度的概念。

仍以二元函数 $f(\boldsymbol{X})=f(x_1,x_2)$ 为例来讨论，函数 $f(\boldsymbol{X})$ 沿 $\boldsymbol{S}$ 方向的方向导数式为

$$\frac{\partial f(\boldsymbol{X}^{(0)})}{\partial \boldsymbol{S}}=\frac{\partial f(\boldsymbol{X}^{(0)})}{\partial x_1}\cos\theta_1+\frac{\partial f(\boldsymbol{X}^{(0)})}{\partial x_2}\cos\theta_2=\begin{bmatrix}\dfrac{\partial f(\boldsymbol{X})}{\partial x_1} & \dfrac{\partial f(\boldsymbol{X})}{\partial x_2}\end{bmatrix}\begin{bmatrix}\cos\theta_1\\ \cos\theta_2\end{bmatrix}$$

令

$$\nabla f(\boldsymbol{X})=\begin{bmatrix}\dfrac{\partial f(\boldsymbol{X})}{\partial x_1} & \dfrac{\partial f(\boldsymbol{X})}{\partial x_2}\end{bmatrix}^{\mathrm{T}}\tag{3-19}$$

称 $\nabla f(\boldsymbol{X})$ 为函数 $f(\boldsymbol{X})$ 在点 $\boldsymbol{X}$ 处的梯度，记做 $\mathrm{grad}f(\boldsymbol{X})$。同时，设 $\boldsymbol{S}$ 为单位向量，即

$$\boldsymbol{S}=[\cos\theta_1,\cos\theta_2]^{\mathrm{T}}$$

于是，可将方向导数 $\dfrac{\partial f(\boldsymbol{X})}{\partial \boldsymbol{S}}$ 表示为

$$\frac{\partial f(\boldsymbol{X})}{\partial \boldsymbol{S}}=[\nabla f(x)]^{\mathrm{T}}\boldsymbol{S}=\|\nabla f(\boldsymbol{X})\|\,\|\boldsymbol{S}\|\cos(\nabla f(\boldsymbol{X}),\boldsymbol{S})\tag{3-20}$$

式中，$\|\nabla f(\boldsymbol{X})\|$ 代表梯度向量 $\nabla f(\boldsymbol{X})$ 的模；$\|\boldsymbol{S}\|$ 代表单位向量的模，即为 1；$\cos(\nabla f(\boldsymbol{X}),\boldsymbol{S})$ 表示梯度向量与 $\boldsymbol{S}$ 方向夹角的余弦。

此式表明，函数 $f(\boldsymbol{X})$ 沿方向 $\boldsymbol{S}$ 的方向导数等于向量 $\nabla f(\boldsymbol{X})$ 在方向 $\boldsymbol{S}$ 上的投影，且当 $\cos(\nabla f(\boldsymbol{X}),\boldsymbol{S})=1$，即向量 $\nabla f(\boldsymbol{X})$ 与 $\boldsymbol{S}$ 的方向相同时，向量 $\nabla f(\boldsymbol{X})$ 在方向 $\boldsymbol{S}$ 上的投影最大，其值为 $\|\nabla f(\boldsymbol{X})\|$。这表明梯度 $\nabla f(\boldsymbol{X})$ 是点 $\boldsymbol{X}$ 处方向导数最大的方向，也就是函数变化率最大的方向。

同理，上述梯度的定义和运算可以推广到 n 元函数中去，即对于 n 元函数 $f(x_1,x_2,\cdots,x_n)$，其梯度的定义可写做

$$\nabla f(\boldsymbol{X})=\left[\frac{\partial f(\boldsymbol{X})}{\partial x_1}\quad\frac{\partial f(\boldsymbol{X})}{\partial x_2}\quad\cdots\quad\frac{\partial f(\boldsymbol{X})}{\partial x_n}\right]^{\mathrm{T}} \tag{3-21}$$

由此可见，梯度是一个向量，梯度方向是函数具有最大变化率的方向，即梯度 $\nabla f(\boldsymbol{X})$ 方向是指函数 $f(\boldsymbol{X})$ 的最速上升方向，负梯度 $-\nabla f(\boldsymbol{X})$ 则为函数 $f(\boldsymbol{X})$ 的最速下降方向。

例 3-3 求二元函数 $f(x_1,x_2)=x_1^2+x_2^2-4x_1-2x_2+5$ 在 $\boldsymbol{X}^{(0)}=[2,2]^{\mathrm{T}}$ 处的梯度及梯度的模。

解：由梯度的定义式(3-21)可求得

$$\nabla f(\boldsymbol{X})=\begin{bmatrix}\dfrac{\partial f(\boldsymbol{X})}{\partial x_1}\\[2ex]\dfrac{\partial f(\boldsymbol{X})}{\partial x_2}\end{bmatrix}=\begin{bmatrix}2x_1-4\\2x_2-2\end{bmatrix}$$

将 $\boldsymbol{X}^{(0)}=[2,2]^{\mathrm{T}}$ 代入上式得

$$\nabla f(\boldsymbol{X}^{(0)})=\begin{bmatrix}2x_1-4\\2x_2-2\end{bmatrix}=\begin{bmatrix}0\\2\end{bmatrix}$$

$\nabla f(\boldsymbol{X}^{(0)})$ 的模为

$$\|\nabla f(\boldsymbol{X}^{(0)})\|=\sqrt{0^2+2^2}=2$$

梯度的单位向量为

$$\boldsymbol{P}=\frac{\nabla f(\boldsymbol{X}^{(0)})}{\|\nabla f(\boldsymbol{X}^{(0)})\|}=\frac{1}{2}\begin{bmatrix}0\\1\end{bmatrix}=\begin{bmatrix}0\\2\end{bmatrix}$$

可用图 3-12 来表示点 $\boldsymbol{X}^{(0)}=[2,2]^{\mathrm{T}}$ 处的梯度方向。从图中可以看出，点 $\boldsymbol{X}^{(0)}$ 处函数的梯度方向 $\boldsymbol{P}$ 是点 $\boldsymbol{X}^{(0)}$ 处函数变化率最大的方向，即是等值线的法线方向，也就是同心圆的半径方向。

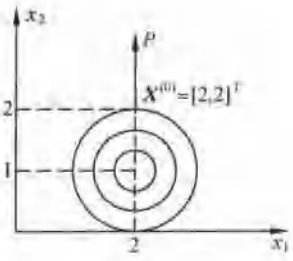


图 3-12 函数的梯度

3.2.2 多元函数的泰勒展开式与海森矩阵

为讨论复杂函数的极值问题，常用泰勒展开式得到目标函数在所讨论点的近似表达式，最常用的是线性近似和二次近似。

设 n 元函数 $f(\boldsymbol{X})$ 在 $\boldsymbol{X}^{(k)}$ 点至少有二阶连续的偏导数，则在这一点邻近的泰勒(Taylor)展开式取到二次项时为

$$f(\boldsymbol{X})=f(\boldsymbol{X}^{(k)})+\sum_{i=1}^{n}\frac{\partial f(\boldsymbol{X}^{(k)})}{\partial x_i}(x_i-x_i^{(k)})+\frac{1}{2}\sum_{i,j=1}^{n}\frac{\partial f(\boldsymbol{X}^{(k)})}{\partial x_i\partial x_j}(x_i-x_i^{(k)})(x_j-x_j^{(k)})$$

写成矩阵形式为

$$f(\boldsymbol{X})\approx f(\boldsymbol{X}^{(k)})+[\nabla f(\boldsymbol{X}^{(k)})]^{\mathrm{T}}(\boldsymbol{X}-\boldsymbol{X}^{(k)})+\frac{1}{2}(\boldsymbol{X}-\boldsymbol{X}^{(k)})^{\mathrm{T}}\nabla^2 f(\boldsymbol{X}^{(k)})(\boldsymbol{X}-\boldsymbol{X}^{(k)}) \tag{3-22}$$

式中

$$\nabla^2 f(\boldsymbol{X})=H(\boldsymbol{X})=\begin{bmatrix}\dfrac{\partial^2 f(\boldsymbol{X})}{\partial x_1^{\,2}} & \cdots & \dfrac{\partial^2 f(\boldsymbol{X})}{\partial x_1 \partial x_n}\\ \vdots & \cdots & \vdots\\ \dfrac{\partial^2 f(\boldsymbol{X})}{\partial x_n \partial x_1} & \cdots & \dfrac{\partial^2 f(\boldsymbol{X})}{\partial x_n^{\,2}}\end{bmatrix} \tag{3-23}$$

为 $f(\boldsymbol{X})$的二阶导数矩阵，称为 $f(\boldsymbol{X})$的海森(Hessian)矩阵，海森矩阵是一个 $n\times n$ 的对称矩阵，常用 $H(\boldsymbol{X})$表示。

例 3-4　一般二元二次函数 $f(\boldsymbol{X})=\frac{1}{2}\boldsymbol{X}^{\mathrm{T}}A\boldsymbol{X}+B^{\mathrm{T}}\boldsymbol{X}+C$，求 $H(\boldsymbol{X})$。

解：

$$H(\boldsymbol{X})=\nabla^2 f(\boldsymbol{X})=\nabla^2\left(\frac{1}{2}\boldsymbol{X}^{\mathrm{T}}A\boldsymbol{X}\right)+\nabla^2(B^{\mathrm{T}}\boldsymbol{X})+\nabla^2 C$$

而

$$\nabla^2\left(\frac{1}{2}\boldsymbol{X}^{\mathrm{T}}A\boldsymbol{X}\right)=\nabla^2\left[\frac{1}{2}(a_{11}x_1^{\,2}+2a_{12}x_1x_2+a_{22}x_2^{\,2})\right]=\begin{bmatrix}a_{11} & a_{12}\\ a_{21} & a_{22}\end{bmatrix}$$

$$\nabla^2(B^{\mathrm{T}}\boldsymbol{X})=\nabla^2(b_1x_1+b_2x_2)=0$$

$$\nabla^2 C=0$$

所以

$$H(\boldsymbol{X})=\nabla^2 f(\boldsymbol{X})=\begin{bmatrix}a_{11} & a_{12}\\ a_{21} & a_{22}\end{bmatrix}=A$$

3.2.3　无约束问题的最优性条件

由微分学知，连续可微的一元函数 $f(\boldsymbol{X})$在给定区间内某点 $\boldsymbol{X}^*$ 处有极值的必要条件是

$$f'(\boldsymbol{X}^*)=0$$

满足上式的点称为驻点。驻点不一定就是极值点，即使是极值点，也不能断定是极小点还是极大点，因此还需用函数在该点的二阶导数来判断。驻点为极值点的充分条件是：若 $f''(\boldsymbol{X}^*)>0$，则点 $\boldsymbol{X}^*$ 为极小点；若 $f''(\boldsymbol{X}^*)<0$，则点 $\boldsymbol{X}^*$ 为极大点。

下面讨论多元函数的情况。

1. *必要条件*

n 元函数在 R^n 中极值点 $\boldsymbol{X}^*$ 存在的必要条件为

$$\nabla f(\boldsymbol{X}^*)=\left[\frac{\partial f(\boldsymbol{X}^*)}{\partial x_1}\quad\frac{\partial f(\boldsymbol{X}^*)}{\partial x_2}\quad\cdots\quad\frac{\partial f(\boldsymbol{X}^*)}{\partial x_n}\right]^{\mathrm{T}}=0 \tag{3-24}$$

即在极值点处函数 $f(\boldsymbol{X})$的梯度为 n 维零向量。

2. *充分条件*

当 $\boldsymbol{X}^*$ 为驻点(梯度为零)时，将式(3-24)代入式(3-22)得

$$f(\boldsymbol{X})-f(\boldsymbol{X}^*)\approx\frac{1}{2}(\boldsymbol{X}-\boldsymbol{X}^*)^{\mathrm{T}}\nabla^2 f(\boldsymbol{X}^*)(\boldsymbol{X}-\boldsymbol{X}^*) \tag{3-25}$$

欲使 $\boldsymbol{X}^*$ 为极小点，即在 $\boldsymbol{X}^*$ 附近，有 $f(\boldsymbol{X})-f(\boldsymbol{X}^*)>0$，于是有

$$(\boldsymbol{X}-\boldsymbol{X}^*)^{\mathrm{T}}\nabla^2 f(\boldsymbol{X}^*)(\boldsymbol{X}-\boldsymbol{X}^*)=(\boldsymbol{X}-\boldsymbol{X}^*)^{\mathrm{T}}H(\boldsymbol{X}^*)(\boldsymbol{X}-\boldsymbol{X}^*)>0 \tag{3-26}$$

或者说，在 $\boldsymbol{X}^*$ 点海森矩阵 $H(\boldsymbol{X}^*)$应是正定的。此即 $\boldsymbol{X}^*$ 为极小点的充分条件。

同理，当 $\boldsymbol{X}$ 为 $\boldsymbol{X}^*$ 附近的任意一点，且恒有 $(\boldsymbol{X}-\boldsymbol{X}^*)^{\mathrm{T}}H(\boldsymbol{X}^*)(\boldsymbol{X}-\boldsymbol{X}^*)<0$ 时，即

$H(\boldsymbol{X}^*)$应是负定的，则 $\boldsymbol{X}^*$ 为极大值点；当$(\boldsymbol{X}-\boldsymbol{X}^*)^{\mathrm{T}}H(\boldsymbol{X}^*)(\boldsymbol{X}-\boldsymbol{X}^*)=0$，即 $H(\boldsymbol{X}^*)$为不定的，则 $\boldsymbol{X}^*$ 点为鞍点。

需要指出，充分条件式(3-25)并不是必要的，即有这样的情况，尽管 $\boldsymbol{X}^*$ 是 $f(\boldsymbol{X})$的极小点，但却不满足条件式(3-25)。例如，$f(\boldsymbol{X})=\boldsymbol{X}^4$，它的极小点是 $\boldsymbol{X}^*=0$，但 $f''(\boldsymbol{X}^*)=0$，这不满足式(3-25)。

此外，下述条件可用来判定海森矩阵是否为正定或负定。

一个 n 阶对称矩阵为正定的充要条件是其各阶顺序主子式均大于 0，即

$$a_{11}>0,\ \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}>0,\cdots,\ \begin{vmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ & \cdots & \cdots & \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{vmatrix}>0$$

一个 n 阶对称矩阵为负定的充要条件是其各阶顺序主子式负正相间，即

$$a_{11}>0,\ \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}>0,\ \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix}<0,\cdots,(-1)^n \begin{vmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ & \cdots & \cdots & \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{vmatrix}>0$$

例 3-5 试求 $f(x_1,x_2)=2x_1^2-8x_1+2x_2^2-4x_2+20$ 的极值及极值点。

解：由极值点存在的必要条件

$$\nabla f(\boldsymbol{X})=\begin{bmatrix} \dfrac{\partial f}{\partial x_1} \\ \dfrac{\partial f}{\partial x_2} \end{bmatrix}=\begin{bmatrix} 4x_1-8 \\ 4x_2-4 \end{bmatrix}=\begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

得驻点 $\boldsymbol{X}^*[2\quad 1]^{\mathrm{T}}$，在 $\boldsymbol{X}^*$ 点海森矩阵为

$$H(\boldsymbol{X}^*)=\begin{bmatrix} \dfrac{\partial^2 f}{\partial x_1^2} & \dfrac{\partial^2 f}{\partial x_1 \partial x_2} \\ \dfrac{\partial^2 f}{\partial x_2 \partial x_1} & \dfrac{\partial^2 f}{\partial^2 x_2^2} \end{bmatrix}=\begin{bmatrix} 4 & 0 \\ 0 & 4 \end{bmatrix}$$

由子其各阶主子行列式

$$|4|>0,\quad \begin{vmatrix} 4 & 0 \\ 0 & 4 \end{vmatrix}>0$$

可知在 $\boldsymbol{X}^*$ 点海森矩阵正定，因此 $\boldsymbol{X}^*$ 为极小点，其极小值为

$$f(\boldsymbol{X}^*)=f(2,1)=2\times 2^2-(8\times 2)+(2\times 1^2)-(4\times 1)+20=10$$

3.2.4 凸集、凸函数与凸规划

如果函数在整个可行域中有两个或两个以上的极值点，则称每一个极值点为局部极值点。在整个可行域中，函数值最小的点为全域极值点。为求得全域极值点，以获得最好的可行设计方案，就需要进一步讨论局部最小点和全域最小点的关系，因而涉及到凸集、凸函数与凸规划问题。

1. 凸集

设 D 是 n 维欧氏空间 $\boldsymbol{R}^n$ 的一个点集，即 $D\subset\boldsymbol{R}^n$。若任意两点 $\boldsymbol{X}^{(1)}\in D$、$\boldsymbol{X}^{(2)}\in D$ 的连

线上的一切点，$a\boldsymbol{X}^{(1)}+(1-a)\boldsymbol{X}^{(2)}\in D(0<a<1)$，则称 D 为凸集。

从直观上讲，凸集的内部没有“孔洞”，边界上没有凹陷部分。凸集的几何特征是：其任意两点连线上的一切点都位于这个集合内。图 3-13 中的(a)、(c)是凸集，(b)不是凸集。

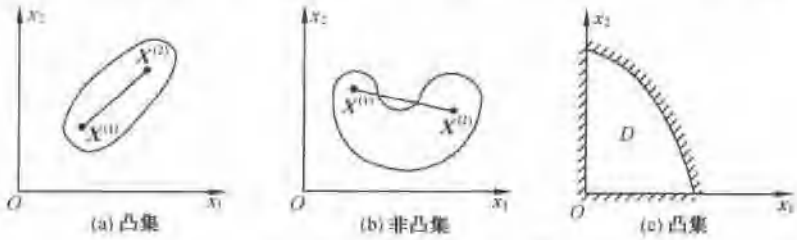


图 3-13　二维空间的凸集和非凸集

2. 凸函数

设 D 为 $\boldsymbol{R}^n$ 中的一个凸集，$f(\boldsymbol{X})$为定义在 D 上的一个函数，若对 D 内任意两个点 $\boldsymbol{X}^{(1)}$ 和 $\boldsymbol{X}^{(2)}$ 及任意 $a\in[0,1]$，恒有

$$f(a\boldsymbol{X}^{(2)}+(1-a)\boldsymbol{X}^{(1)})\leqslant af(\boldsymbol{X}^{(2)})+(1-a)f(\boldsymbol{X}^{(1)}) \tag{3-27}$$

则称 $f(\boldsymbol{X})$为凸函数。其几何意义为，这两个点的连线完全处在 $f(\boldsymbol{X})$曲线(曲面)的上方，或在 $f(\boldsymbol{X})$曲线(曲面)上。图 3-14(a)和图 3-14(b)分别给出了一维和二维凸函数的图像。可见，凸函数都是向上弯曲的。显然，凸函数的局部极小亦是全局极小。

同理可定义凹函数，当式(3-27)中的“≤”改为“≥”时，则称 $f(\boldsymbol{X})$为凹函数；图 3-15(a)和图 3-15(b)分别给出了一维和二维凹函数的图像。凹函数都是向下弯曲的，故局部极大亦是全局极大。

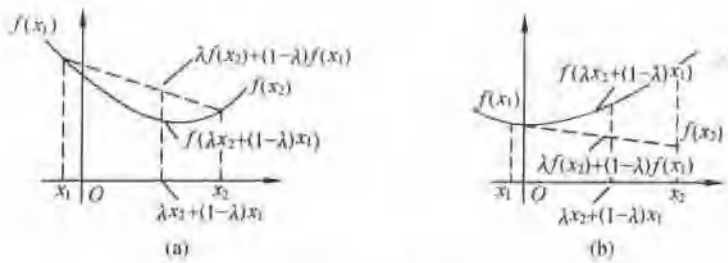


图 3-14　单变量函数

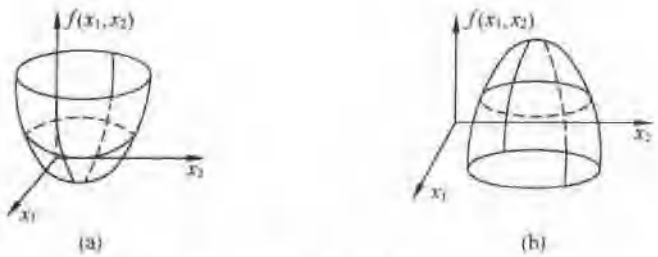


图 3-15　双变量函数

判定一个函数的凸性，可利用以下性质：

(1) 设 $f(\boldsymbol{X})$为定义在 D_1 上的具有连续一阶导数的函数，而 D 为 D_1 内部的一个凸集，则 $f(\boldsymbol{X})$在 D 上为凸函数的充分必要条件为：对任意两点 $\boldsymbol{X}^{(1)}$ 和 $\boldsymbol{X}^{(2)}$，有

$$f(\boldsymbol{X}^{(2)}) \geqslant f(\boldsymbol{X}^{(1)}) + [\nabla f(\boldsymbol{X}^{(1)})]^{\mathrm{T}}(\boldsymbol{X}^{(2)} - \boldsymbol{X}^{(1)})$$

恒成立。

(2) 设 $f(\boldsymbol{X})$为定义在 D_1 上的具有连续二阶导数的函数，而 D 为 D_1 内部的一个凸集，则 $f(\boldsymbol{X})$在 D 上为凸函数的充分必要条件是，对一切 $\boldsymbol{X} \in D$，$f(\boldsymbol{X})$的海森矩阵 $H(\boldsymbol{X})$为半正定矩阵；若 $H(\boldsymbol{X})$是正定的，则 $f(\boldsymbol{X})$在 D 上为严格凸函数，即在式(3-27)中"$\leqslant$"改为"$<$"。

3. 凸规划

对于非线性规划

$$\begin{cases} \min f(\boldsymbol{X}) & \boldsymbol{X} \in R^n \\ \text{s.t.} \quad g_i(\boldsymbol{X}) \leqslant 0 & i = 1, 2, \cdots, m \end{cases}$$

若其中 $f(\boldsymbol{X})$和 $g_i(\boldsymbol{X})(i=1,2,\cdots,m)$均为凸函数(对于 $g_i(\boldsymbol{X}) \geqslant 0$ 约束则为凹函数)，则这样的规划问题称为凸规划。凸规划的可行域为凸集。

凸规划的局部极小点一定是全局极小点，这是凸规划的一个重要性质。

3.2.5 约束问题的最优性条件

有约束优化问题比无约束优化问题复杂得多。在优化理论中，判断约束极值点存在的条件，可用库恩-塔克(Kulm-Tucker)条件，简称 K-T 条件。该条件指出，某个点 $\boldsymbol{X}^{(k)}$ 为约束极值点的必要条件是：目标函数在该点的负梯度 $-\nabla f(\boldsymbol{X}^{(k)})$应为在该点起作用约束条件梯度 $\nabla g_i(\boldsymbol{X}^{(k)})$的线性组合，即

$$-\nabla f(\boldsymbol{X}^{(k)}) = \sum_{i=1}^{q} \lambda_i \nabla g_i(\boldsymbol{X}^{(k)})$$

式中：q——为在点 $\boldsymbol{X}^{(k)}$ 处的起作用约束数；

λ_i——为拉格朗日乘子，是一非负乘子。

对于等式约束，理解起来容易。对于不等式约束，分为最优点在约束曲线切点上和交点上两种情况，如图 3-16 所示。最优点 $\boldsymbol{X}^*$ 在约束曲线 $g_1(\boldsymbol{X})$与目标函数等值线的切点处，则 $g_1(\boldsymbol{X})$为起作用约束，其他两个约束为不起作用约束。对于图 3-17，最优点 $\boldsymbol{X}^*$ 在两约束曲线 $g_1(\boldsymbol{X})$和 $g_2(\boldsymbol{X})$的交点上，$\boldsymbol{X}^*$ 处目标函数的负梯度 $-\nabla f(\boldsymbol{X}^{(k)})$夹于两约束条件梯度 $\nabla g_1(\boldsymbol{X}^{(k)})$和$\nabla g_2(\boldsymbol{X}^{(k)})$之间，此两约束均为起作用约束。

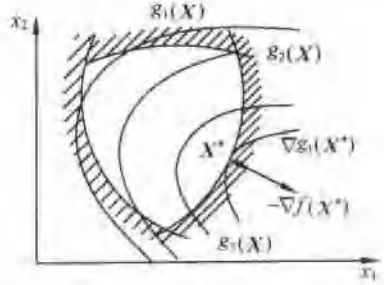


图 3-16 最优点在切点处的不等式约束情况

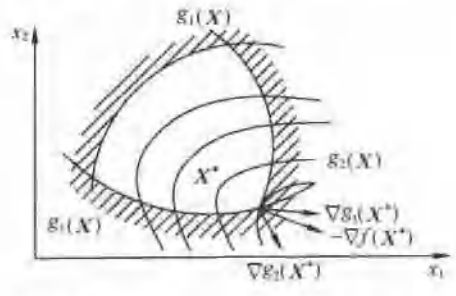


图 3-17 最优点在交点处的不等式约束情况

K-T 条件可简述为

$$\begin{cases}\min f(\boldsymbol{X}) & \boldsymbol{X}\in \boldsymbol{R}^n \\ \text{s.t.}\quad g_i(\boldsymbol{X})\leqslant 0 & i=1,2,\cdots,m \\ h_j(\boldsymbol{X})=0 & j=1,2,\cdots,p,\ p<n\end{cases}$$

如果 $\boldsymbol{X}^*$ 是一个局部最优点，且各函数梯度组成线性关系，那么，必存在一组非负乘子（拉格朗日乘子）λ_i^* 和另一组乘子（无非负要求）μ_j^*，使得

$$\begin{cases}\nabla f(\boldsymbol{X}^*)+\sum\limits_{i=1}^{m}\lambda_i^*\nabla g_i(\boldsymbol{X}^*)+\sum\limits_{j-1}^{p}\mu_j^*\nabla h_j(\boldsymbol{X}^*)=0 \\ \lambda_i^* g_i(\boldsymbol{X}^*)=0,\ i=1,2,\cdots,m\end{cases}$$

成立。

K-T 条件是判别约束最优点的必要条件，而不是充分条件。只有当优化问题属凸规划问题，即目标函数为凸函数，可行域为凸集时，K-T 条件才是有约束优化问题最优解的充要条件，这种情况下的局部最优解必为问题的全局最优解。

例 3-6　对于约束优化问题

$$\begin{cases}\min f(\boldsymbol{X})=(x_1-2)^2+x_2^2 \\ \text{s.t.}\quad g_1(\boldsymbol{X})=x_1^2+x_2-1\leqslant 0 \\ g_2(\boldsymbol{X})=-x_2\leqslant 0 \\ g_3(\boldsymbol{X})=-x_1\leqslant 0\end{cases}$$

它的当前迭代点为 $\boldsymbol{X}^{(k)}=[1\quad 0]^{\mathrm{T}}$，试用 K-T 条件判别它是否为约束最优点。

解：(1) 当前迭代点 $\boldsymbol{X}^{(k)}=[1\quad 0]^{\mathrm{T}}$ 为可行点，因为满足各约束条件，即

$$g_1(\boldsymbol{X}^{(k)})=1^2-1=0$$
$$g_2(\boldsymbol{X}^{(k)})=0$$
$$g_3(\boldsymbol{X}^{(k)})=-1<0$$

(2) $\boldsymbol{X}^{(k)}$ 点的起作用约束为 $g_1(\boldsymbol{X})$ 和 $g_2(\boldsymbol{X})$，因为 $\boldsymbol{X}^{(k)}$ 点不在 $g_3(\boldsymbol{X})$ 上，而在 $g_1(\boldsymbol{X})$、$g_2(\boldsymbol{X})$ 交点上，即

$$g_1(\boldsymbol{X}^{(k)})=x_1^2+x_2-1=0$$
$$g_2(\boldsymbol{X}^{(k)})=-x_2=0$$
$$g_3(\boldsymbol{X}^{(k)})=-x_1\neq 0$$

(3) $\boldsymbol{X}^{(k)}$ 点处各函数的梯度为

$$\nabla f(\boldsymbol{X}^{(k)})=\begin{bmatrix}2x_1-4\\ 2x_2\end{bmatrix}^{x^{(k)}}=\begin{bmatrix}-2\\ 0\end{bmatrix}$$

$$\nabla g_1(\boldsymbol{X}^{(k)})=\begin{bmatrix}2x_1\\ 1\end{bmatrix}^{x^{(k)}}=\begin{bmatrix}2\\ 1\end{bmatrix}$$

$$\nabla g_2(\boldsymbol{X}^{(k)})=\begin{bmatrix}0\\ -1\end{bmatrix}$$

(4) 求拉格朗日乘子 λ_1、λ_2。

按 K-T 条件应有

$$\nabla f(\boldsymbol{X}^{(k)})+\lambda_1\nabla g_1(\boldsymbol{X}^{(k)})+\lambda_2\nabla g_2(\boldsymbol{X}^{(k)})=0$$

$$\begin{bmatrix}-2\\0\end{bmatrix}+\lambda_1\begin{bmatrix}2\\1\end{bmatrix}+\lambda_2\begin{bmatrix}0\\-1\end{bmatrix}=0$$

$$\begin{cases}-2+2\lambda_1=0\\ \lambda_1-\lambda_2=0\end{cases}$$

解线性方程组得　　$\lambda_1=1>0,\quad \lambda_2=1>0$

由于拉格朗日乘子均为非负，这说明 $\boldsymbol{X}^{(k)}=[1\quad 0]^{\mathrm{T}}$ 是一个局部最优点，因为它满足K-T条件，即 $-\nabla f(\boldsymbol{X}^{(k)})$ 为 $\nabla g_1(\boldsymbol{X})$ 与 $\nabla g_2(\boldsymbol{X})$ 的线性组合。

由图 3-18 知，$\boldsymbol{X}^{(k)}=[1\quad 0]^{\mathrm{T}}$ 确实是该约束优化问题的局部最优点，而且，由于目标函数 $f(\boldsymbol{X})$ 是凸函数，可行域为凸集，所以点 $X^{(k)}=[1\quad 0]^{\mathrm{T}}$ 也是该问题的全局最优点。

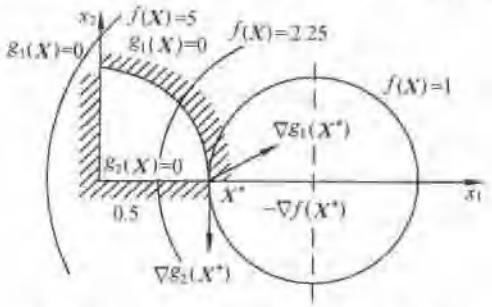


图 3-18　局部最优点和全局最优点

3.2.6　优化问题的数值迭代法

1. 数值迭代法的基本思想和迭代格式

机械优化问题比较复杂，难以用数学中的微分法来求解。为了适应电子计算机的运算，常用的优化方法多采用数值迭代法求解。数值迭代法的基本思想是搜索、迭代和逼近。现以二维约束优化问题为例来说明，如图 3-19 所示。

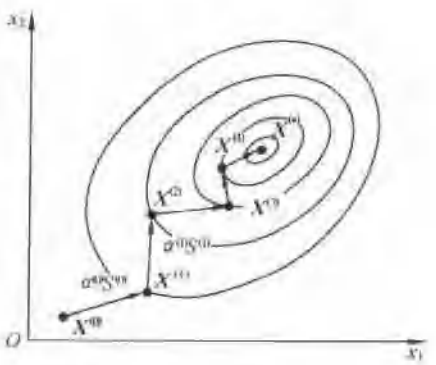


图 3-19　迭代过程

从选定的初始点 $\boldsymbol{X}^{(0)}$ 出发，按照选用的优化方法所规定的搜索方向 $\boldsymbol{S}^{(0)}$ 和步长 $\alpha^{(0)}$ 进行搜索，求得一个目标函数值有所下降的新设计点 $\boldsymbol{X}^{(1)}$，然后以 $\boldsymbol{X}^{(1)}$ 点作为新迭代点，重复上述过程，又获得另一个目标函数值有所下降的新设计点 $\boldsymbol{X}^{(2)}$，如此循环往复不断搜索，最后可求得满足设计精度要求的逼近理想最优点的近似最优点 $\boldsymbol{X}^{*}$。这就是数值迭代法的过程。

数值迭代法的迭代格式可写为

$$\boldsymbol{X}^{(k+1)}=\boldsymbol{X}^{(k)}+\alpha^{(k)}\boldsymbol{S}^{(k)},\quad k=0,1,2,\cdots \tag{3-28}$$

式中：$\boldsymbol{X}^{(k)}$——第 k 次迭代初始点；

$\boldsymbol{X}^{(k+1)}$——第 $k+1$ 次迭代产生的新点；

$\alpha^{(k)}$——第 k 次迭代步长（或步长因子），是标量；

$\boldsymbol{S}^{(k)}$——第 k 次迭代搜索方向，是向量。

2. 迭代计算的终止准则

从理论上讲，优化迭代计算过程可以产生一个无穷的点列$\{\boldsymbol{X}^{(k)}, k=0,1,2,\cdots\}$，一直计算到目标函数理论上的极小点$\boldsymbol{X}^*$。但是这种无穷计算实际上既办不到，也无这个必要。因此，一般是根据迭代计算终止准则，求得足够近似的最优点$\boldsymbol{X}^*$时，即可终止迭代计算。

数值迭代计算常用的迭代终止准则有三种：

(1) 点距准则。当相邻两次优化迭代点$\boldsymbol{X}^{(k)}$和$\boldsymbol{X}^{(k+1)}$之间的距离已达到充分小时，迭代计算可以终止，即

$$\| \boldsymbol{X}^{(k+1)} - \boldsymbol{X}^{(k)} \| \leqslant \varepsilon \tag{3-29}$$

或

$$\sqrt{\sum_{i=1}^{n} (x_i^{(k+1)} - x_i^{(k)})^2} \leqslant \varepsilon \tag{3-30}$$

(2) 函数值下降量准则。当相邻两次优化迭代点$\boldsymbol{X}^{(k)}$和$\boldsymbol{X}^{(k+1)}$所对应的目标函数值的下降量或相对下降量已达到充分小时，迭代则可终止，即

$$|f(\boldsymbol{X}^{(k+1)}) - f(\boldsymbol{X}^{(k)})| \leqslant \varepsilon \tag{3-31}$$

或

$$\frac{|f(\boldsymbol{X}^{(k+1)}) - f(\boldsymbol{X}^{(k)})|}{|f(\boldsymbol{X}^{(k)})|} \leqslant \varepsilon \tag{3-32}$$

(3) 梯度准则。当优化迭代点所对应的目标函数梯度已达到充分小时，迭代则可终止，即

$$\| \nabla f(\boldsymbol{X}^{(k)}) \| \leqslant \varepsilon \tag{3-33}$$

以上各式中的ε是根据设计要求预先给定的迭代精度。

在优化设计中，一般只要满足以上终止准则之一，则可认为设计点收敛于极值点。应该指出，有时为了防止当函数变化剧烈时，点距准则虽已满足，求得的最优值$f(\boldsymbol{X}^{(k+1)})$与真正的最优值$f(\boldsymbol{X}^*)$仍相差较大；或当函数变化缓慢时，目标函数值下降量准则虽已得到满足，而所求得的最优点$\boldsymbol{X}^{(k+1)}$与真正的最优点$\boldsymbol{X}^*$仍相距较远，往往将前两种终止准则结合起来使用，要求同时成立。至于梯度准则，仅用于需要计算目标函数梯度的最优化方法中。

3.3　一维搜索的优化方法

求解一维目标函数$f(\alpha)$的极小点和极小值的数值迭代方法称为一维搜索方法。

机械优化设计大都是多维问题，一维问题的情况很少。但是一维优化方法是优化方法中最基本的方法。它不仅用来解决一维目标函数的求优问题，而且更常用于多维优化问题中在既定方向上寻求最优步长的一维搜索。

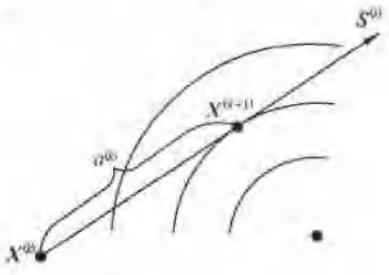


图 3-20　一维搜索

由优化算法的基本迭代公式

$$\boldsymbol{X}^{(k+1)} = \boldsymbol{X}^{(k)} + \alpha^{(k)} \boldsymbol{S}^{(k)}, \quad k=0,1,2,\cdots$$

和图 3-20 可知，当已知迭代初始点$\boldsymbol{X}^{(k)}$及搜索方向$\boldsymbol{S}^{(k)}$确定后，迭代所得的新点$\boldsymbol{X}^{(k+1)}$取决于步长$\alpha^{(k)}$，不

同的 $\alpha^{(k)}$ 会得到不同的 $\boldsymbol{X}^{(k+1)}$ 和不同的目标函数值 $f(\boldsymbol{X}^{(k+1)})$。因此，在多维优化问题中，一维优化的目的是在既定的 $\boldsymbol{X}^{(k)}$ 和 $\boldsymbol{S}^{(k)}$ 下寻求最优步长 $\alpha^{(k)}$，使迭代产生的新点 $\boldsymbol{X}^{(k+1)}$ 的函数值为最小，即

$$f(\boldsymbol{X}^{(k)})+\alpha^{(k)}\boldsymbol{S}^{(k)}=\min f(\boldsymbol{X}^{(k)}+\alpha\,\boldsymbol{S}^{(k)})$$

上述极小化问题实质上就是求单变量 α 的一维函数极小化问题，即 $\min f(\alpha)$。

一维搜索的优化方法很多，本章主要介绍常用的黄金分割法和二次插值法两种。

一维优化一般分为两大步骤：

(1) 确定初始搜索区间$[a,b]$，该区间应是包括一维函数极小点在内的单峰区间。

(2) 在搜索区间$[a,b]$内寻找极小点。

3.3.1　确定搜索区间的方法——进退法

进退法的基本思路是：由单峰函数性质可知，在极小点 α^* 左边函数值应严格下降，而在极小点右边函数值应严格上升。因此，可从某一个给定的初始点 α_0 出发，以初始步长 h_0 沿着目标函数值的下降方向，逐步前进(或后退)，直至找到相继的三个试点的函数值按“大—小—大”变化为止。

进退法确定搜索区间的步骤如下：

(1) 给定初始点 α_0 和初始步长 h_0。

(2) 令 $\alpha_1=\alpha_0$，$h=h_0$，$\alpha_2=\alpha_1+h$，得两试点 α_1、α_2，计算它们的函数值 $f_1=f(\alpha_1)$，$f_2=f(\alpha_2)$。

(3) 比较 f_1 和 f_2，存在两种情况：

① 若 $f_1>f_2$，如图 3-21(a)、(b)所示，则作前进运算。取第三个试点 $\alpha_3=\alpha_2+h$，计算函数值 $f_3=f(\alpha_3)$，并比较 f_2 与 f_3：

若 $f_2\leqslant f_3$，如图 3-21(a)所示，则找到了相继三个试点的函数值按“大—小—大”变化，故有搜索区间$[a,b]=[\alpha_1,\alpha_3]$；

若 $f_2>f_3$，如图 3-21(b)所示，则将步长加倍，即令 $h=2h$，$\alpha_1=\alpha_2$，$\alpha_2=\alpha_3$，$\alpha_3=\alpha_2+h$，如此重复该过程，总能找到相继三个试点的函数值符合“大—小—大”变化的要求。取左端点为 a，右端点为 b，从而找到了搜索区间$[a,b]$。

② 若 $f_2\geqslant f_1$，如图 3-21(c)、(d)所示，则作后退计算。令 $h=-h$，将 α_1、f_1 与 α_2、f_2 对调，并取第三个试点 $\alpha_3=\alpha_2+h$，计算其函数值 $f_3=f(\alpha_3)$，比较对调后的 f_2 与 f_3：

若 $f_2\leqslant f_3$，如图 3-21(c)所示，则搜索区间$[a,b]=[\alpha_3,\alpha_1]$；

若 $f_2>f_3$，如图 3-21(d)所示，则步长加倍，继续作后退运算，令 $h=2h$，$\alpha_1=\alpha_2$，$\alpha_2=\alpha_3$，$\alpha_3=\alpha_2+h$，继续比较 f_2 与 f_3，直至找到相继三个试点的函数值按“大—小—大”变化为止，相应的区间为$[\alpha_3,\alpha_1]$。

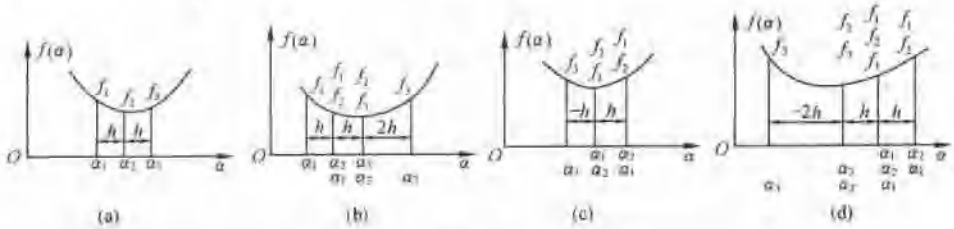


图 3-21　进退法确定搜索区间

上述用进退法确定搜索区间的算法框图如图 3-22 所示。搜索区间找到后，便可运用下述一维优化算法在区间内找到极小点。

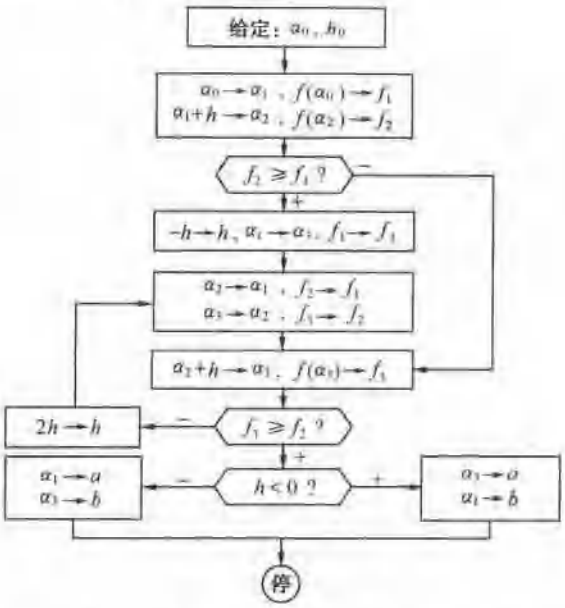


图 3-22　进退法确定搜索区间的算法框图

3.3.2　黄金分割法

1. 黄金分割法的基本原理

黄金分割法又称 0.618 法，它是通过不断缩短搜索区间的长度来寻求一维函数 $f(\alpha)$ 的极小点。这种方法的基本原理是，在搜索区间 $[a,b]$ 内按如下规则对称地取两点 α_1 和 α_2：

$$\alpha_1 = a + 0.382(b-a),\quad \alpha_2 = a + 0.618(b-a) \tag{3-34}$$

计算它们的函数值 $f_1 = f(\alpha_1)$，$f_2 = f(\alpha_2)$，比较 f_1 与 f_2 的大小，有两种可能：

(1) 若 $f_1 > f_2$，如图 3-23(a)所示。极小点必在区间 $[\alpha_1, b]$ 内，消去区间 $[a, \alpha_1)$，令 $a = \alpha_1$，产生新区间 $[a,b]$，到此，区间收缩了一次。值得注意的是新区间的 α_1 点与原区间的 α_2 点重合，可令 $\alpha_1 = \alpha_2$，$f_1 = f_2$，这样可少找一个新点和节省一次函数值计算。

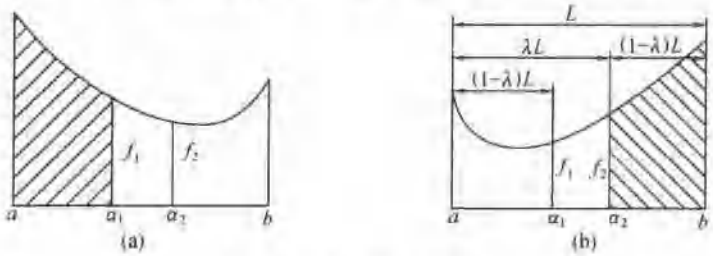


图 3-23　黄金分割法搜索区间

(2) 若 $f_1 \leqslant f_2$，如图 3-23(b)所示。极小点必在区间 $[a, \alpha_2]$ 内，消去区间 $[\alpha_2, b]$，令 $b = \alpha_2$，产生新区间 $[a,b]$，到此，区间缩短了一次。同样，新区间 α_2 点与原区间的 α_1 点重合，令

$a_2=a_1$，$f_2=f_1$。

当缩短的新区间长度小于等于某一精度 ε，即 $b-a\leqslant\varepsilon$ 时，则取 $a^*=\dfrac{a_1+a_2}{2}$ 为近似极小点。

2. 黄金分割法的区间收缩率 λ

每次缩小所得的新区间长度与缩小前区间长度之比，称为区间收缩率，以 λ 表示，如图 3-23(b)所示。为加快区间收缩应保证区间收缩率不变，因此，必须在搜索区间$[a,b]$内对称地取计算点 a_1、a_2。设初始区间长度为 L，则第一次和第二次收缩得到的新区间长度分别为 λL 和$(1-\lambda)L$。根据收缩率相等的原则，可得

$$\lambda L : L=(1-\lambda)L : \lambda L$$

即

$$\lambda^2+\lambda-=0$$

该方程的正根为 $\lambda=\dfrac{\sqrt{5}-1}{2}\approx 0.618$。这就是在区间内按式(3-34)取两对称点的原因。

黄金分割法程序结构简单，容易理解，可靠性好，但计算效率偏低，适用于低维优化的一维搜索。它的算法框图如图 3-24 所示。

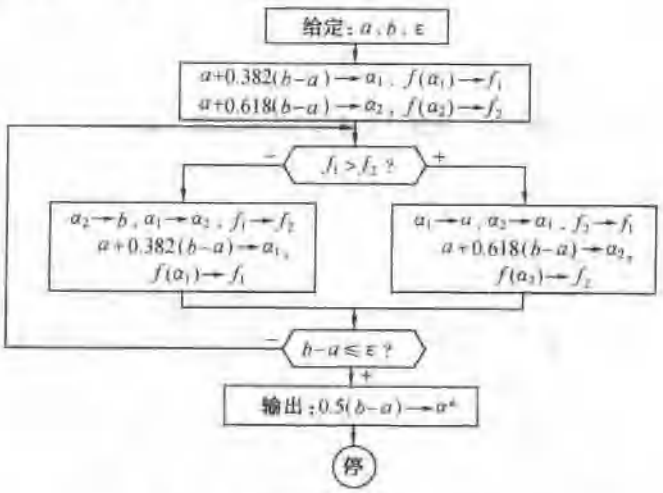


图 3-24　黄金分割法算法框图

例 3-7　用黄金分割法求 $f(a)=a^2-7a+10$ 的最优解。设初始点 a_0，初始步长 $h=1$，取迭代精度 $\varepsilon=0.35$。

解：首先用进退法确定搜索区间，按图 3-24 所示的算法框图进行计算。

$$a_1=a_0=0,\quad f_1=f(a_1)=10$$

$$a_2=a_1+h=1,\quad f_2=f(a_2)=4$$

比较 f_2、f_1，因 $f_2<f_1$，作前进运算：

$$a_3=a_2+h=2,\quad f_3=f(a_3)=0$$

比较 f_2、f_3，因 $f_2>f_3$，再作前进运算：

$$h=2\times 1=2$$

$$a_1=a_2=1,\quad f_1=f_2=4$$

$$\alpha_2=\alpha_3=2,\quad f_2=f_3=0$$
$$\alpha_3=\alpha_2+h=4,\quad f_3=f(\alpha_3)=-2$$

比较 f_2、f_3，因 $f_2>f_3$，再作前进运算：

$$h=2\times 2=4$$
$$\alpha_1=\alpha_2=2,\quad f_1=f_2=4$$
$$\alpha_2=\alpha_3=4,\quad f_2=f_3=-2$$
$$\alpha_3=\alpha_2+h=8,\quad f_3=f(\alpha_3)=18$$

此时，α_1、α_2 和 α_3 三点的函数值出现了“大—小—大”变化，故 $a=\alpha_1=2$，$b=\alpha_3=8$，初始搜索区间$[a,b]=[2,8]$。下面按图 3-24 所示的黄金分割算法框图进行求优。

(1) 在初始区间$[a,b]=[2,8]$中取两计算点并计算其函数值

$$\alpha_1=a+0.382(b-a)=4.292,\quad f_1=f(\alpha_1)=-1.622736$$
$$\alpha_2=a+0.618(b-a)=5.708,\quad f_2=f(\alpha_2)=2.625264$$

(2) 比较函数值，缩短搜索空间。因有 $f_1<f_2$，则

$$b=\alpha_2=5.708$$
$$\alpha_2=\alpha_1=4.292,\qquad f_2=f(\alpha_2)=-1.622736$$
$$\alpha_1=a+0.382(b-a)=3.416456,\quad f_1=f(\alpha_1)=-2.243020$$

(3) 判断迭代终止条件

$$b-a=5.708-2=3.708<\varepsilon$$

不满足迭代终止条件，比较函数值 f_1、f_2，继续缩短区间。

将各次缩短区间的有关计算数据列于下表：

区间缩短次数	a	b	α_1	α_2	f_1	f_2	$b-a$
原区间	2	8	4.292	5.708	−1.622736	2.625264	6
1	2	5.708	3.416456	4.292	−2.24302	−1.622736	3.708
2	2	4.929	2.975544	3.416456	−1.97496	−2.24302	2.292
3	2.975544	4.929	3.41456	3.789114	−2.24302	−2.166413	1.316456
4	2.975544	3.789114	3.286328	3.416456	−2.204344	−2.24302	0.81357
5	3.286328	3.789114	3.416456	3.59705	−2.24302	−2.240581	0.502786
6	3.286328	3.59705	3.40523	3.416456	−2.24098	−2.24302	0.310722

可见区间缩短 6 次后，区间长度为

$$b-a=3.59705-3.286328=0.310722<\varepsilon$$

迭代即可终止，近似最优解为

$$a^*=\frac{b+a}{2}=3.441689,\quad f^*=f(a^*)=-2.2466$$

3.3.3　二次插值法

二次插值法又称抛物线法。它的基本思路是：在寻求目标函数 $f(a)$极小点的搜索区间内，取三个点的函数值来构造一个二次插值多项式 $p(a)$，用它的极小点近似地作为原目标函数的极小点。若近似程度不满足精度要求时，可以反复使用此法，随着区间的缩短，二次

插值多项式的极小点就逼近原目标函数的极小点。

1. 二次插值法的基本原理

设一元函数 $f(a)$在搜索区间$[a,b]$内取三点：$a_1=a$，$a_2=0.5(a+b)$，$a_3=b$。计算它们的函数值 $f_1=f(a_1)$，$f_2=f(a_2)$，$f_3=f(a_3)$，且满足 $f_1>f_2<f_3$，即满足函数值呈"大—小—大"变化。于是可通过原函数曲线上的三个点 $P_1(a_1,f_1)$、$P_2(a_2,f_2)$和 $P_3(a_3,f_3)$，作一条二次曲线(抛物线)，如图 3-25 所示。此二次函数可表示为

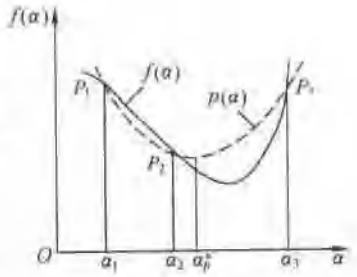


图 3-25 二次插值法基本原理

$$p(a)=a_0+a_1a+a_2a^2$$

对 $p(a)$求导数，并令其为零，即

$$p'(a)=a_1+2a_2a=0$$

解得二次函数极小点

$$a_p^*=-\frac{a_1}{2a_2} \tag{3-35}$$

为求得 a_p^* 应求出上式(3-35)中的待定系数 a_1 和 a_2。

根据插值条件，插值函数 $p(a)$与原函数 $f(a)$在插值结点 P_1、P_2 和 P_3 处函数值相等，得

$$\begin{aligned}p(a_1)&=a_0+a_1a_1+a_2a_1^2=f_1\\p(a_2)&=a_0+a_1a_2+a_2a_2^2=f_2\\p(a_3)&=a_0+a_1a_3+a_2a_3^2=f_3\end{aligned} \tag{3-36}$$

解方程组得 a_1 和 a_2，并代入式(3-35)，即得二次插值函数极小点 a_p^* 的计算公式

$$a_p^*=\frac{1}{2}\left[\frac{(a_2^2-a_3^2)f_1+(a_3^2-a_1^2)f_2+(a_1^2-a_2^2)f_3}{(a_2-a_3)f_1+(a_3-a_1)f_2+(a_1+a_2)f_3}\right] \tag{3-37}$$

为便于计算，可将上式改写为

$$a_p^*=0.5\left(a_1+a_3-\frac{C_1}{C_2}\right) \tag{3-38}$$

式中 $C_1=\dfrac{f_3-f_1}{a_3-a_1}$，$C_2=\dfrac{\dfrac{f_2-f_1}{a_2-a_1}-C_1}{a_2-a_3}$

若将只用一回二次插值计算所得的 a_p^* 作为函数的极小点 a_p 的近似解往往达不到精度要求，为此需缩短区间，进行多次插值计算，使 a_p^* 不断逼近原函数的极小点 a^*。

比较 a_2、a_3 两点的函数值的大小，在区间$[a_1,a_3]$内的四个点中选取三个点，使它们的函数值在呈现"大—小—大"变化的前提下缩短搜索区间，然后再重复上述方法进行二次插值计算，直至相继两次插值函数极小点之间的距离小于某一精度要求时为止。

2. 二次插值法的迭代过程与程序框图

二次插值法的迭代过程如下：

(1) 给定初始搜索区间$[a,b]$和精度 ε。

(2) 在区间$[a,b]$内取三点：$a_1=a$，$a_2=0.5(a+b)$，$a_3=b$，计算函数值 $f_1=f(a_1)$，$f_2=f(a_2)$，$f_3=f(a_3)$构成三个插值结点 $P_1(a_1,f_1)$、$P_2(a_2,f_2)$和 $P_3(a_3,f_3)$。

(3) 按式(3-38)计算二次插值函数的极小点 a_p^*，并将 a_p^* 记做 a_4，计算 $f_4=f(a_4)$。

若本步骤为第一次插值或 a_3 点仍为初始给定点时，说明 a_2 和 a_4 不代表前后两次插值函数的极小点，不能进行终止判断，故进行下一步骤(4)；否则转步骤(5)。

(4) 缩短搜索区间。缩短搜索区间的原则是：比较 f_2、f_4，取其小者所对应的点作为新的 a_2 点，并以此点左右邻点分别取作为新的 a_1 和 a_3，这样构成了缩短后的新搜索区间 $[a_1, a_3]$。根据原区间中 a_4 与 a_2 的相对位置和函数值 f_2 与 f_4 的比较，区间缩短有四种情况，如图 3-26 所示。图中阴影线部分表示丢去的区间，在对新区间的三个新点的代号作依次为 a_1、a_2、a_3 的一般化处理后，计算函数值 $f_1=f(a_1)$、$f_2=f(a_2)$、$f_3=f(a_3)$，返回步骤(3)。

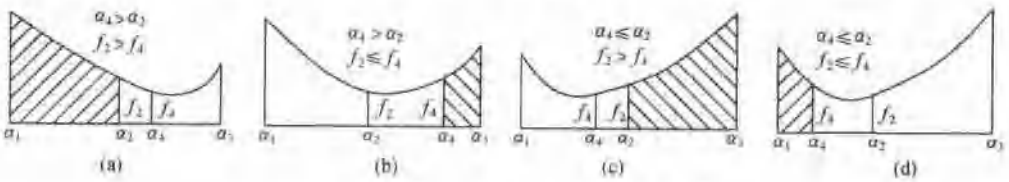


图 3-26　二次插值法区间缩短的四种情况

(5) 判断是否满足精度要求。当满足 $|a_1-a_2|\leqslant\varepsilon$ 时，停止迭代，把 a_2 与 a_4 中原函数值较小的点作为极小点；否则，返回步骤(4)，再次缩短搜索区间，直到满足精度要求为止。

按上述步骤设计的二次插值法的程序框图如图 3-27 所示。

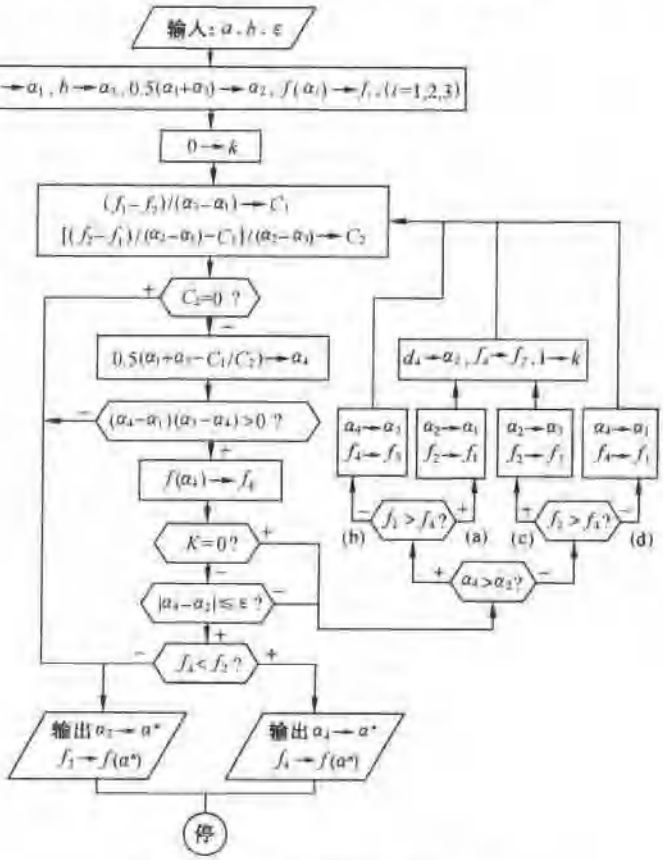


图 3-27　二次插值法的程序框图

在算法框图中有以下三点需作些说明：

(1) 判别框：$C_2=0$?

若不成立，按式(3-36)和式(3-38)，则有

$$\frac{f_2-f_1}{\alpha_2-\alpha_1}=C_1=\frac{f_3-f_1}{\alpha_3-\alpha_1}$$

说明三个插值结点 P_1、P_2 和 P_3 在一条直线上。

(2) 判别框：$(\alpha_4-\alpha_1)(\alpha_3-\alpha_4)>0$?

若不成立，说明 α_1 落在区间$[\alpha_1,\alpha_3]$之外。

以上两种情况只是在区间已缩得很小，三个插值结点已十分接近，由于计算机的舍入误差才可能使其发生。此时取 α_2 和 f_2 作为最优解是合理的。

(3) 开关 $k=0$ 和 $k=1$ 分别表示初始点 α_2 第一次被 α_4 代替前和代替后的状态。

例 3-8 用二次插值法求 $f(a)=a^2-7a+10$ 的最优解，已知初始区间为$[2,8]$，取终止迭代点距精度 $\varepsilon=0.01$。

解：按图 3-27 所示算法框图进行计算

(1) 确定初始插值结点

$$\alpha_1=a=2,\quad f_1=f(\alpha_1)=0$$
$$\alpha_3=b=8,\quad f_3=f(\alpha_3)=18$$
$$\alpha_2=\frac{a+b}{2}=5,\quad f_2=f(\alpha_2)=0$$

(2) 计算插值函数极小点

$$C_1=\frac{f_3-f_1}{\alpha_3-\alpha_1}=3,\quad C_2=\frac{\dfrac{f_2-f_1}{\alpha_2-\alpha_1}-C_1}{\alpha_2-\alpha_3}=1>0$$

$$\alpha_4=0.5(\alpha_1+\alpha_3-\frac{C_1}{C_2})=3.5$$

$$(\alpha_4-\alpha_1)(\alpha_3-\alpha_4)=6.75>0$$

$$f_4=f(\alpha_4)=-2.25$$

(3) 缩短搜索区间。

因 $\alpha_4<\alpha_2$，$f_2>f_4$，属于图 3-26(c)的情况，故

$$\alpha_3=\alpha_2=5,\quad f_3=0$$
$$\alpha_2=\alpha_4=3.5,\quad f_2=-2.25$$
$$\alpha_1=2,\quad f_1=0$$

开关 $k\neq0$，返回步骤(3)计算得

$$C_1=0,\quad C_2=1>0$$
$$\alpha_4=3.5,\quad f_4=-2.25$$
$$(\alpha_4-\alpha_1)(\alpha_3-\alpha_4)=2.25>0$$

(4) 判断迭代终止条件。

$$|\alpha_4-\alpha_2|=|3.5-3.5|=0<\varepsilon$$

满足迭代终止条件，得最优解

$$a^*=\alpha_4=3.5,\quad f^*=f(\alpha_4)=-2.25$$

由本例可见，对于二次函数用二次插值法求优，只需一次插值计算即可；对于非二次函数，随着区间的缩短使函数的二次性态加强，因而收敛也是较快的。

二次插值法收敛速度快，有效性好，但程序较复杂，可靠性稍差，适用于多维优化的一维搜索迭代。

3.4　无约束优化方法

所谓无约束优化问题，即对设计变量的取值范围不加任何限制，其数学模型为

$$\min f(\boldsymbol{X}),\quad \boldsymbol{X}\in \boldsymbol{R}^n$$

无约束优化问题的基本思路是从选定的某一初始点 $\boldsymbol{X}^{(0)}$ 出发，沿着按一定规律产生的搜索方向 $\boldsymbol{S}^{(k)}$ 寻求使函数值下降的新迭代点，使得

$$\boldsymbol{X}^{(k+1)}=\boldsymbol{X}^{(k)}+\alpha^{(k)}\boldsymbol{S}^{(k)}$$

且满足关系

$$f(\boldsymbol{X}^{(k+1)})<f(\boldsymbol{X}^{(k)})$$

随着迭代次数 k 的不断增加，无约束优化问题将逐步逼近最优点。

各种无约束优化方法的区别，主要在于搜索方向的不同。无约束优化问题的求解方法大致分为两类：

第一类：直接解法。这种方法中只用到函数 $f(\boldsymbol{X})$，而不涉及其导数，如坐标轮换法、鲍威尔法(Powell 法)、随机搜索法、单纯形法等。

第二类：间接解法。它要用到 $f(\boldsymbol{X})$ 的导数，如用到一阶导数的方法有梯度法、共轭梯度法和变尺度法；用到二阶导数的方法以牛顿法为代表。间接解法也称为解析法。

在实际工程问题中，无约束条件的优化问题是比较少见的。但是，无约束优化方法是解有约束优化问题的基础，许多约束优化问题往往是通过对约束条件的处理而转化为无约束优化问题来求解的。因此，无约束优化方法是优化方法中的基本方法。

3.4.1　梯度法

函数的负梯度方向是函数值下降最快的方向(称最速下降方向)，梯度法就是采用负梯度方向作为搜索方向的方法，即

$$\boldsymbol{S}^{(k)}=-\nabla f(\boldsymbol{X}^{(k)})$$

或

$$S^{(k)}=-\frac{\nabla f(\boldsymbol{X}^{(k)})}{\|\nabla f(\boldsymbol{X}^{(k)})\|}$$

则下一个迭代点为

$$\boldsymbol{X}^{(k+1)}=\boldsymbol{X}^{(k)}-\alpha^{(k)}\nabla f(\boldsymbol{X}^{(k)}) \tag{3-39}$$

这就是梯度法的迭代公式。

为了使目标函数在搜索方向 $-\nabla f(\boldsymbol{X}^{(k)})$ 下降最多，常采用沿该方向做一维搜索，求得最优步长 $\alpha^{(k)}$，

$$f(\boldsymbol{X}^{(k)}+\alpha^{(k)}\boldsymbol{S}^{(k)})=\min_{\alpha}\boldsymbol{X}^{(k)}+\alpha^{(k)}\boldsymbol{S}^{(k)}$$

梯度法迭代过程简单，直观易懂，且对初始点的选取要求不严。梯度法又称最速下

降法。

1. 梯度法的迭代步骤

(1) 给定初始点 $\boldsymbol{X}^{(0)}$ 及收敛精度 ε，令 $k=0$；

(2) 求目标函数的梯度向量 $-\nabla f(\boldsymbol{X}^{(k)})$；

(3) 若 $\|\nabla f(\boldsymbol{X}^{(k)})\| \leqslant \varepsilon$，停止迭代，否则转下步；

(4) 用一维搜索方法确定步长 $\alpha^{(k)}$；

(5) 按 $\boldsymbol{X}-\alpha\nabla f(\boldsymbol{X})$ 求得下一个点，且 $k=k+1$，转步骤(2)。

梯度法的算法框图如图 3-28 所示。

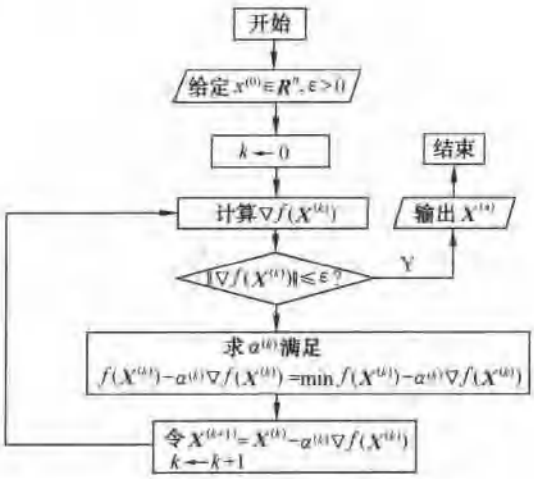


图 3-28 梯度法程序框图

例 3-9 试用梯度法求函数 $f(\boldsymbol{X})=x_1^2+25x_2^2$ 的极小点。

解：取初始点 $\boldsymbol{X}^{(0)}=[2 \quad 2]^{\mathrm{T}}$，此时

$$f(\boldsymbol{X}^{(0)})=104$$

函数的梯度表达式是

$$\nabla f(\boldsymbol{X})=\nabla f(x_1,x_2)=[2x_1 \quad 50x_2]^{\mathrm{T}}$$

由于 $\boldsymbol{X}^{(0)}=[2 \quad 2]^{\mathrm{T}}$，计算出 $\nabla f(\boldsymbol{X}^{(0)})=[4 \quad 100]^{\mathrm{T}}$。

沿 $-\nabla f(\boldsymbol{X}^{(0)})$ 方向求一维极小：

$$q(\alpha)=f(\boldsymbol{X}^{(0)}-\alpha^{(0)}\nabla f(\boldsymbol{X}^{(0)}))=f(\boldsymbol{X}^{(1)})=$$
$$[\boldsymbol{X}_1^{(1)}]^2+25[\boldsymbol{X}_2^{(1)}]^2=(2-4\alpha)^2+25(2-100\alpha)^2$$

由 $\dfrac{\mathrm{d}q(\alpha)}{\mathrm{d}\alpha}$，得

$$\alpha^{(0)}=0.02003$$

从而得

$$\boldsymbol{X}^{(1)}=\boldsymbol{X}^{(0)}-\alpha^{(0)}\nabla f(\boldsymbol{X}^{(0)})=[1.92 \quad -0.003]^{\mathrm{T}}$$

此时函数值为

$$f(\boldsymbol{X}^{(1)})=3.686$$

再从 $\boldsymbol{X}^{(1)}$ 点出发，重复上述迭代过程，可求得 $\boldsymbol{X}^{(2)}$。如此迭代下去，经几次迭代后，便可得到和精确极小点$[0 \quad 0]^{\mathrm{T}}$非常接近的近似解。实际上，这是一个无限的迭代过程。下面给出前三次迭代的结果，列于表3-1。迭代过程示于图3-29。

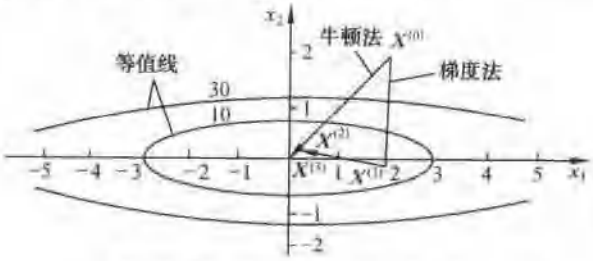


图3-29　例3-9的迭代过程

表3-1

点号 k	$x_1^{(k)}$	$x_2^{(k)}$	$f(\boldsymbol{X}^{(k)})$	$\dfrac{\partial f(\boldsymbol{X}^{(k)})}{\partial x_1}$	$\dfrac{\partial f(\boldsymbol{X}^{(k)})}{\partial x_2}$	$\alpha^{(k)}$
0	2	2	104	4	100	0.02
1	1.920	−0.003	3.686	3.840	−0.154	0.482
2	0.071	0.071	0.131	0.142	3.544	0.02
3	0.068	−0.0001	0.0046	0.0136	−0.0054	0.482

2. *梯度法的特点*

由于在迭代公式(3-39)中的 $\alpha^{(k)}$ 是函数 $g(\alpha)=f(\boldsymbol{X}^{(k)}-\alpha\nabla f(\boldsymbol{X}^{(k)}))$ 的一维极小点，故必有

$$q'(\alpha^{(k)})=-\nabla f(\boldsymbol{X}^{(k)}-\alpha^{(k)}\nabla f(\boldsymbol{X}^{(k)}))^{\mathrm{T}}\nabla f(\boldsymbol{X}^{(k)})=0$$

即

$$\nabla f(\boldsymbol{X}^{(k+1)})^{\mathrm{T}}\nabla f(\boldsymbol{X}^{(k)})=0$$

因此，在梯度法中，相邻两次搜索方向(即相邻两迭代点的梯度方向)是正交的。

结合例3-4并参照图3-30所示的二维目标函数的搜索过程，说明梯度法的特点：

(1) 梯度法理论明确，程序简单，计算量和存储量较少，对初始点的要求不严格。

(2) 由于负梯度方向的最速下降性，容易使人误认为负梯度方向是理想的搜索方向，梯度法是一种理想的方法，但实际上并非如此。梯度法的收敛速度并不快。这是因为最速下降方向仅仅是指某点的一个局部性质，一旦离开这点，就不能保证仍是最速下降方向了。

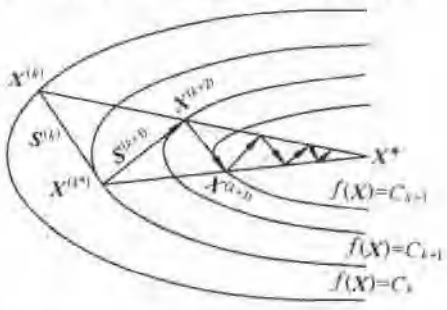


图3-30　二维目标函数的搜索过程

(3) 由于梯度法的相邻两次搜索方向的正交性，决定了迭代全过程的搜索路线呈锯齿状，如图3-30所示，故前几次迭代函数值下降较快。但以后的迭代下降越来越慢。当目标函数的等值线(面)是一系列很扁平的椭圆(椭球)或类似的图形时，收敛更慢，尤其在接近极值点时。

所以，梯度法常与其他方法联合使用，即在迭代的第一步或前几步使用，当接近极小点时，改为用其他算法，以此加快收敛速度。

3.4.2 牛顿法及其改进

1. 牛顿法

牛顿法的基本思想是在 $X^{(k)}$ 邻域内用一个二次函数 $\phi(X)$[将 $f(X)$ 在 $X^{(k)}$ 作泰勒展开]来近似代替原目标函数 $f(X)$，并将 $\phi(X)$ 的极小点 X_ϕ^* 作为对原目标函数 $f(X)$ 求优的下一个迭代点 $X^{(k+1)}$，经过多次迭代，使之逐步逼近目标函数 $f(X)$ 的极小点 X^*。

$$f(X)=f(X^{(k)})+\nabla f(X^{(k)})^{\mathrm{T}}(X-X^{(k)})+[X-X^{(k)}]^{\mathrm{T}}\cdot\nabla^2 f(X^{(k)})(X-X^{(k)})=\phi(X)$$

求 $\phi(X)$ 的极小点 X_ϕ^*，即令其梯度为零，

$$\nabla\phi(X)=\nabla f(X^{(k)})+\nabla^2 f(X^{(k)})(X-X^{(k)})=0$$

则

$$X_\phi^*=X^{(k)}-[\nabla^2 f(X^{(k)})]^{-1}\nabla f(X^{(k)})$$

由此

$$X^{(k+1)}=X^{(k)}-[\nabla^2 f(X^{(k)})]^{-1}\nabla f(X^{(k)})$$

或者为

$$X^{(k+1)}=X^{(k)}-H(X^{(k)})^{-1}\nabla f(X^{(k)}) \tag{3-40}$$

上式中 $H(X^{(k)})$ 为在点 $X^{(k)}$ 处的海森矩阵，这就是牛顿法的迭代公式。

由上述迭代公式可以看出，牛顿法的搜索方向为 $S^{(k)}=-H(X^{(k)})^{-1}\nabla f(X^{(k)})$，该方向称牛顿方向。式中没有步长因子 $\alpha^{(k)}$，或者说步长恒等于 1。

如果 $f(X)$ 本身是二次函数，则 $H(X)$ 是一个常量矩阵，其元素均为常量，这时逼近式是准确的，即 $f(X)=\phi(X)$，因此从任一点出发，由式(3-40)迭代，只需一步即达到极小点 X^*。对于非二次函数，若函数的二次性态较强或迭代点已进入最优点的邻域，则其收敛速度也是很快的。

例 3-10 对例 3-9 试用牛顿法求解。

解：仍取初始点 $X^{(0)}=[2\quad 2]^{\mathrm{T}}$，则有 $\nabla f(X^{(0)})=[4\quad 100]^{\mathrm{T}}$

$$H(X^{(0)})=\begin{bmatrix}2 & 0\\ 0 & 50\end{bmatrix},\qquad H(X^{(0)})^{-1}=\begin{bmatrix}1/2 & 0\\ 0 & 1/50\end{bmatrix}$$

由式(3-40)得下一迭代点为

$$X^{(1)}=X^{(0)}-H(X^{(0)})^{-1}\nabla f(X^{(0)})=\begin{bmatrix}2\\ 2\end{bmatrix}-\begin{bmatrix}1/2 & 0\\ 0 & 1/50\end{bmatrix}\begin{bmatrix}4\\ 100\end{bmatrix}=\begin{bmatrix}0\\ 0\end{bmatrix}$$

这就是函数 $f(X)$ 的精确极小点 $X^*=[0\quad 0]^{\mathrm{T}}$。将其搜索过程也画在图 3-29 上，可以和梯度法进行比较。

2. 阻尼牛顿法

由于牛顿法中步长因子恒取 $\alpha^{(k)}=1$，会造成在迭代过程中函数值有所增大的情况，因此，对初始点的选取有严格要求，尽管用牛顿法收敛很快，但初始点不能离极小点太远，否则可能不收敛。例如，对于求 $\min f(X)=100(x_2-x_1^2)^2+(1-x_1)^2$ 这样的问题，当初始点选 $X^{(0)}=[0.5\quad 0.5]^{\mathrm{T}}$ 时迭代后目标函数值下降，迭代有意义。但如果取初始点为 $X^{(0)}=[0.5\quad 0.5]^{\mathrm{T}}$，

$$\nabla f(\boldsymbol{X}^{(0)})=\begin{bmatrix}-2\\0\end{bmatrix},\ H(\boldsymbol{X}^{(0)})=\begin{bmatrix}2&0\\0&200\end{bmatrix}$$

由式(3-38)得

$$\boldsymbol{X}^{(1)}=\begin{bmatrix}0\\0\end{bmatrix}-\frac{1}{400}\begin{bmatrix}200&0\\0&2\end{bmatrix}\begin{bmatrix}-2\\0\end{bmatrix}=\begin{bmatrix}1\\0\end{bmatrix}$$

代入目标函数,得 $f(\boldsymbol{X}^{(1)})=100$,因 $f(\boldsymbol{X}^{(0)})=1$,可见迭代后函数值反而增大。

为了摆脱由于初始点选择不当而造成的不收敛的情况,人们将牛顿法改进为阻尼牛顿法。

阻尼牛顿法的迭代公式为

$$\boldsymbol{X}^{(k+1)}=\boldsymbol{X}^{(k)}-\alpha^{(k)}H(\boldsymbol{X}^{(k)})^{-1}\nabla f(\boldsymbol{X}^{(k)})$$

式中 $\alpha^{(k)}$ 仍为步长因子,由一维搜索方法确定,即

$$f(\boldsymbol{X}^{(k)}-\alpha^{(k)}H^{-1}\nabla f(\boldsymbol{X}^{(k)}))=\min_{\alpha} f(\boldsymbol{X}^{(k)}-\alpha^{(k)}H^{-1}\nabla f(\boldsymbol{X}^{(k)}))$$

阻尼牛顿法保持了牛顿法收敛快的特点,且对初始点无特别要求。虽然计算量多了一些,但实用性更好。

3. 阻尼牛顿法迭代步骤

(1) 取初始点 $\boldsymbol{X}^{(0)}$,$\varepsilon>0$,令 $k=0$;

(2) 计算 $\nabla f(\boldsymbol{X}^{(k)})$;

(3) 若 $\|\nabla f(\boldsymbol{X}^{(k)})\|\leqslant\varepsilon$,停止,$\boldsymbol{X}^{*}=\boldsymbol{X}^{(k)}$,否则转(4);

(4) 计算 $H(\boldsymbol{X}^{(k)})^{-1}$,令 $\boldsymbol{S}^{(k)}=-H(x^{(k)})^{-1}\nabla f(\boldsymbol{X}^{(k)})$;

(5) 求 $\alpha^{(k)}$,使 $f(\boldsymbol{X}^{(k)}+\alpha^{(k)}\boldsymbol{S}^{(k)})=\min_{\alpha} f(\boldsymbol{X}^{(k)}+\alpha\boldsymbol{S}^{(k)})$;

(6) 令 $\boldsymbol{X}^{(k+1)}=\boldsymbol{X}^{(k)}+\alpha^{(k)}\boldsymbol{S}^{(k)}$,$k=k+1$,转步骤(2)。

阻尼牛顿法算法框图如图 3-31 所示。牛顿法和阻尼牛顿法虽具有收敛快的优点,但它们的最大缺点是每次迭代都要计算二阶导数矩阵(海森阵)及其逆矩阵。一般来讲,求逆矩阵是比较麻烦的,应尽量避免。此外,牛顿法还要求海森阵 $H(\boldsymbol{X})$ 正定,且为非奇导的,否则无法计算其逆阵 $H(\boldsymbol{X})^{-1}$,所以对于多变量复杂目标函数的优化问题,牛顿法几乎不实用。

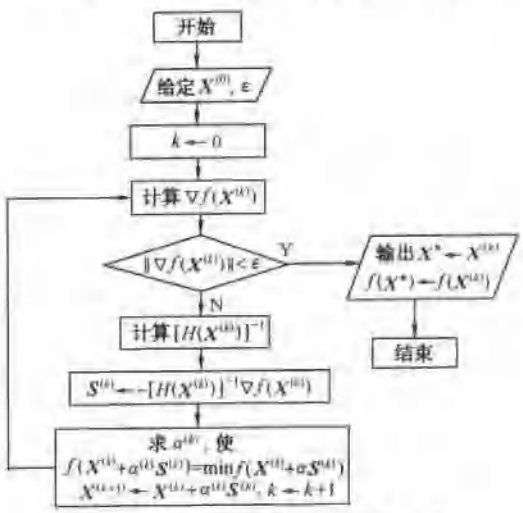


图 3-31 阻尼牛顿法算法框图

3.4.3 变尺度法

变尺度法又称拟牛顿法，它是在牛顿法的基础上进行了重大改进的一类方法。变尺度法所含内容极为丰富，DFP 法便是其中最具有代表性的一种。DFP 法是 1959 年由 Davidon 提出的，1963 年 Fletcher 和 Powell 进行了改进，故称 DFP 法。由于这种方法对解高维问题具有显著的优越性，使其获得了很高的声誉，至今仍被公认为求解无约束优化问题最有效的算法之一。

1. 基本思想

DFP 与梯度法和牛顿法有密切关系，其优化迭代的一般公式表示为

$$\boldsymbol{X}^{(k+1)}=\boldsymbol{X}^{(k)}-\alpha^{(k)}\boldsymbol{A}^{(k)}\nabla f(\boldsymbol{X}^{(k)}) \tag{3-41}$$

式中 $\boldsymbol{A}^{(k)}$ 是需要构造的一个 $n\times n$ 对称方阵，它随迭代点位置的变化而变化，形成一个矩阵序列。其搜索方向 $\boldsymbol{S}^{(k)}=-\boldsymbol{A}^{(k)}\nabla f(\boldsymbol{X}^{(k)})$ 称为拟牛顿方向。观察这一公式，如令 $\boldsymbol{A}^{(k)}=\boldsymbol{I}$（单位矩阵），则得到梯度法；如令 $\boldsymbol{A}^{(k)}=[\nabla^2 f(\boldsymbol{X}^{(k)})]^{-1}$（海森矩阵的逆），则得到阻尼牛顿法；又特别地，当 $\alpha^{(k)}=1$，则得到牛顿法。

由于梯度法构造简单，只用到一阶偏导数，计算量小，迭代初期收敛速度快，但当迭代到最优点附近时收敛速度极慢；而牛顿法虽收敛很快，对二次函数只需迭代一次便达到最优点，对非二次函数也能较快迭代到最优点，但牛顿法要计算海森矩阵及其逆阵，对维数较高的优化问题，其计算工作量和存储量都太大，而且对某些函数可能根本无法计算二阶偏导数矩阵及其逆阵。为此需要综合改进这两种方法，扬长避短，于是产生了变尺度法的基本思想。所谓变尺度是指 $\boldsymbol{A}^{(k)}$ 矩阵序列的变化性，称为尺度矩阵。当矩阵 $\boldsymbol{A}^{(k)}$，通过不断地迭代而能很好地逼近 $[\nabla^2 f(\boldsymbol{X}^{(k)})]^{-1}$（或 $H(\boldsymbol{X}^{(k)})^{-1}$）时，我们就不再需要计算二阶导数，从而避免了牛顿法的繁琐，收敛速度也能很快。

变尺度法的关键在于尺度矩阵 $\boldsymbol{A}^{(k)}$ 的产生。

2. 构造 $\boldsymbol{A}^{(k)}$ 尺度矩阵

对 $\boldsymbol{A}^{(k)}$ 的构造可从初始矩阵 $\boldsymbol{A}^{(k)}=\boldsymbol{I}$（单位矩阵）开始，通过对公式

$$\boldsymbol{A}^{(k+1)}=\boldsymbol{A}^{(k)}+\Delta\boldsymbol{A}^{(k)} \tag{3-42}$$

中称之为修正矩阵的 $\Delta\boldsymbol{A}^{(k)}$ 的不断修正，在迭代中逐步逼近于 $H(\boldsymbol{X}^{(k)})^{-1}$。因此，一旦达到最优点附近，就可望达到牛顿法收敛速度，避免了求矩阵 $H(\boldsymbol{X}^{(k)})^{-1}$。此外，修正矩阵 $\Delta\boldsymbol{A}^{(k)}$ 取不同的形式，就构成了不同的变尺度法。

DFP 法中的修正矩阵 $\Delta\boldsymbol{A}^{(k)}$ 为

$$\Delta\boldsymbol{A}^{(k)}=\frac{\Delta\boldsymbol{X}^{(k)}\Delta\boldsymbol{X}^{(k)\mathrm{T}}}{\Delta\boldsymbol{X}^{(k)\mathrm{T}}\Delta\boldsymbol{g}^{(k)}}-\frac{\boldsymbol{A}^{(k)}\Delta\boldsymbol{g}^{(k)}\Delta\boldsymbol{g}^{(k)\mathrm{T}}\boldsymbol{A}^{(k)}}{\Delta\boldsymbol{g}^{(k)\mathrm{T}}\boldsymbol{A}^{(k)}\Delta\boldsymbol{g}^{(k)}} \tag{3-43}$$

式中

$$\begin{cases}\Delta\boldsymbol{g}^{(k)}=\boldsymbol{g}^{(k+1)}-\boldsymbol{g}^{(k)}=\nabla f(\boldsymbol{X}^{(k+1)})-\nabla f(\boldsymbol{X}^{(k)})\\ \Delta\boldsymbol{X}^{(k)}=\boldsymbol{X}^{(k+1)}-\boldsymbol{X}^{(k)}\end{cases} \tag{3-44}$$

$\boldsymbol{A}^{(k)}$ 为 $n\times n$ 阶对称正定矩阵。

有关 $\Delta\boldsymbol{A}^{(k)}$ 递推公式的推导，可参看有关书籍。

3. DFP 法的迭代步骤

（1）任取初始点 $\boldsymbol{X}^{(0)}$，初始矩阵 $\boldsymbol{A}^{(0)}=\boldsymbol{I}$，$\varepsilon>0$，令 $k=0$。

(2) 计算$\nabla f(\boldsymbol{X}^{(k)})$，若$\|\nabla f(\boldsymbol{X}^{(k)})\| \leqslant \varepsilon$，则$\boldsymbol{X}^{(k)}$即为所求，停止迭代；否则转步骤(3)。

(3) 搜索方向$\boldsymbol{S}^{(k)} = -\boldsymbol{A}^{(k)}\nabla f(\boldsymbol{X}^{(k)})$，沿$\boldsymbol{S}^{(k)}$方向作一维搜索，求最优步长$\alpha^{(k)}$，即

$$f(\boldsymbol{X}^{(k)} + \alpha^{(k)}\boldsymbol{S}^{(k)}) = \min_{\alpha} f(\boldsymbol{X}^{(k)} + \alpha\,\boldsymbol{S}^{(k)})$$

得到新的迭代点$\boldsymbol{X}^{(k+1)} = \boldsymbol{X}^{(k)} + \alpha^{(k)}\boldsymbol{S}^{(k)}$。

(4) 若$\|\nabla f(\boldsymbol{X}^{(k)})\| \leqslant \varepsilon$，则停止迭代。输出$\boldsymbol{X}^* = \boldsymbol{X}^{(k+1)}$，$f^* = f(\boldsymbol{X}^*)$；否则转步骤(5)。

(5) 若$k=n$(维数)，则重置，即$\boldsymbol{S}(k)$从负梯度方向重新开始，取$\boldsymbol{X}^{(0)} = \boldsymbol{X}^{(k+1)}$，$k=0$转步骤(3)；否则转步骤(6)。

(6) 由式(3-43)和式(3-44)去修正尺度矩阵$\boldsymbol{A}^{(k)}$，由式(3-42)得到$\boldsymbol{A}^{(k+1)}$，令$k=k+1$，转步骤(3)。

DFP法的迭代框图如图3-32所示。

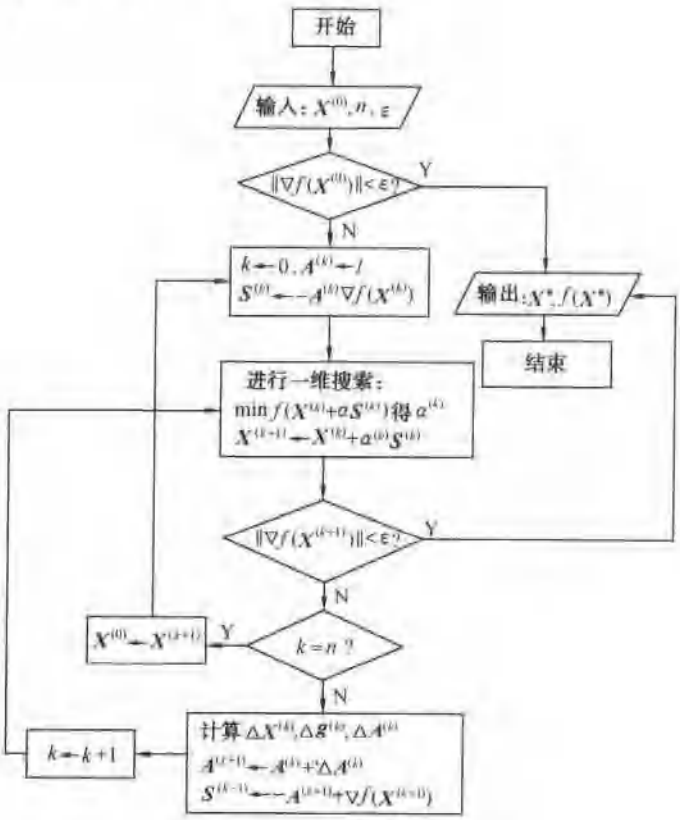


图3-32　DFP法的迭代框图

例3-11　用DFP法解$\min f(\boldsymbol{X}) = 60 - 10x_1 - 4x_2 + x_1^2 + x_2^2 - x_1x_2$。初始点为$\boldsymbol{X}^{(0)}[0 \quad 0]^{\mathrm{T}}$，$\varepsilon=0.0001$。

解：(1) 令$k=0$，$\boldsymbol{A}^{(0)}\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}$；

(2) 计算目标函数的梯度$\nabla f(\boldsymbol{X}^{(0)})$。

$$\nabla f(\boldsymbol{X}^{(0)})=\begin{bmatrix}-10+2x_1-x_2\\-4+2x_2-x_1\end{bmatrix}_{[0\ \ 0]}=\begin{bmatrix}-10\\-4\end{bmatrix}$$

因为 $\|\nabla f(\boldsymbol{X}^{(k)})\|=\sqrt{10^2+4^2}>\varepsilon$ 继续迭代。

(3) 搜索方向为

$$\boldsymbol{S}^{(0)}=-\boldsymbol{A}^{(0)}\nabla f(\boldsymbol{X}^{(0)})=-\begin{bmatrix}1&0\\0&1\end{bmatrix}\begin{bmatrix}-10\\-4\end{bmatrix}=\begin{bmatrix}10\\4\end{bmatrix}$$

此时搜索方向实质为负梯度方向。沿此方向进行一维搜索，求最优步长因子 $\alpha^{(k)}$，将 $\boldsymbol{X}^{(1)}=\boldsymbol{X}^{(0)}+\alpha\boldsymbol{S}^{(0)}$ 代入目标函数得

$$f(\boldsymbol{X}^{(1)})=60-10(10\alpha)-4(4\alpha)+(10\alpha)^2+(4\alpha)^2-(10\alpha)(4\alpha)=60-116\alpha+76\alpha^2=q(\alpha)$$

为求极小值，将上式对 α 求导，并令 $q'(\alpha)=0$，即 $\dfrac{\mathrm{d}q}{\mathrm{d}\alpha}=-116+152\alpha=0$，解得

$$\alpha^{(0)}=0.7631$$

于是得新点

$$\boldsymbol{X}^{(1)}=\boldsymbol{X}^{(0)}+\alpha^{(0)}\boldsymbol{S}^{(0)}=\begin{bmatrix}0\\0\end{bmatrix}+0.7631\begin{bmatrix}10\\4\end{bmatrix}=\begin{bmatrix}7.631\\3.052\end{bmatrix}$$

(4) 收敛性判别，

$$\nabla f(\boldsymbol{X}^{(1)})=\begin{bmatrix}-10+2x_1-x_2\\-4+2x_2-x_1\end{bmatrix}_{(7.631\ \ 3.052)}=\begin{bmatrix}2.211\\-5.526\end{bmatrix}$$

因为 $\|\nabla f(\boldsymbol{X}^{(k)})\|=\sqrt{(2.211)^2+(-5.526)^2}>\varepsilon$，所以继续下一步迭代。

(5) 此时 $k<n=2$，所以计算

$$\Delta\boldsymbol{X}^{(k)}=\Delta\boldsymbol{X}^{(0)}=\boldsymbol{X}^{(1)}-\boldsymbol{X}^{(0)}=\begin{bmatrix}7.631\\3.052\end{bmatrix}$$

按式(3-42)、式(3-43)和式(3-44)，计算尺度矩阵 $\boldsymbol{A}^{(k+1)}$

$$\boldsymbol{A}^{(k+1)}=\boldsymbol{A}^{(1)}=\boldsymbol{A}^{(0)}-\Delta\boldsymbol{A}^{(0)}=$$

$$\begin{bmatrix}1&0\\0&1\end{bmatrix}+\frac{\begin{bmatrix}7.631\\3.052\end{bmatrix}[7.631\ \ 3.052]}{[7.631\ \ 3.052]\begin{bmatrix}12.211\\-1.526\end{bmatrix}}-\frac{\begin{bmatrix}1&0\\0&1\end{bmatrix}\begin{bmatrix}12.211\\-1.526\end{bmatrix}[12.211\ \ -1.526]\begin{bmatrix}1&0\\0&1\end{bmatrix}}{[12.211\ \ -1.526]\begin{bmatrix}1&0\\0&1\end{bmatrix}\begin{bmatrix}12.211\\-1.526\end{bmatrix}}=$$

$$\begin{bmatrix}1&0\\0&1\end{bmatrix}+\begin{bmatrix}0.658&0.263\\0.263&0.105\end{bmatrix}-\begin{bmatrix}0.985&-0.123\\-0.123&0.0153\end{bmatrix}=\begin{bmatrix}0.673&0.386\\0.386&1.0897\end{bmatrix}$$

由此可见，它是一个对称正定矩阵。

(6) $k\leftarrow k+1$，构造新的搜索方向(拟牛顿方向)为

$$\boldsymbol{S}^{(k+1)}=\boldsymbol{S}^{(1)}=-\boldsymbol{A}^{(1)}\nabla f(\boldsymbol{X}^{(1)})=-\begin{bmatrix}0.637&0.386\\0.386&1.0897\end{bmatrix}\begin{bmatrix}2.211\\-5.526\end{bmatrix}=\begin{bmatrix}0.646\\5.169\end{bmatrix}$$

沿 $\boldsymbol{S}^{(1)}$ 方向作一维搜索求 $\alpha^{(1)}$，方法与求 $\alpha^{(0)}$ 相同，得 $\alpha^{(1)}=0.5701$，新的迭代点为

$$\boldsymbol{X}^{(2)}=\boldsymbol{X}^{(1)}+\alpha^{(1)}\boldsymbol{S}^{(1)}=\begin{bmatrix}7.631\\3.052\end{bmatrix}+0.5701\begin{bmatrix}0.646\\5.169\end{bmatrix}=\begin{bmatrix}7.999\\5.999\end{bmatrix}\approx\begin{bmatrix}8\\6\end{bmatrix}$$

(7) 收敛性判断：$\|\nabla f(\boldsymbol{X}^{(k)})\|=\sqrt{(0)^2+(0)^2}<\varepsilon$，停止迭代。输出最优解

$$\boldsymbol{X}^*=\boldsymbol{X}^{(2)}=[8\quad 6]^{\mathrm{T}},\quad f(\boldsymbol{X}^*)=f(\boldsymbol{X}^{(2)})=8$$

4. DFP 变尺度法的几点说明

(1) 在变尺度法中对 $\boldsymbol{A}^{(k)}$ 矩阵是有要求的。因为对于极小化问题，为了使搜索方向 $\boldsymbol{S}^{(k)}=-\boldsymbol{A}^{(k)}\nabla f(\boldsymbol{X}^{(k)})$ 为下降方向，需要使搜索方向 $\boldsymbol{S}^{(k)}$ 与负梯度方向的夹角为锐角，即

$$[-\nabla f(\boldsymbol{X}^{(k)})]^{\mathrm{T}}\boldsymbol{S}^{(k)}>0$$

所以有

$$[\nabla f(\boldsymbol{X}^{(k)})]^{\mathrm{T}}\boldsymbol{A}^{(k)}\nabla f(\boldsymbol{X}^{(k)})>0$$

故 $\boldsymbol{A}^{(k)}$ 必须为正定矩阵。

(2) 对于对称正定矩阵 $\boldsymbol{A}$ 和非零向量 $\boldsymbol{S}^{(1)},\boldsymbol{S}^{(2)},\cdots,\boldsymbol{S}^{(n)}$ 如满足下式：

$$\begin{cases}[\boldsymbol{S}^{(i)}]^{\mathrm{T}}\boldsymbol{A}\boldsymbol{S}^{(j)}=0 & i\neq j;i,j=1,2,\cdots,n\\ [\boldsymbol{S}^{(i)}]^{\mathrm{T}}\boldsymbol{A}\boldsymbol{S}^{(i)}=0 & i=1,2,\cdots,n\end{cases}$$

则称非零向量 $\boldsymbol{S}^{(1)},\boldsymbol{S}^{(2)},\cdots,\boldsymbol{S}^{(n)}$ 是关于 A 的两两共轭。利用这种共轭向量作为搜索方向的优化方法称为共轭方向法。在 DFP 变尺度法中，由于所构成的搜索方向序列 $\boldsymbol{S}^{(0)},\boldsymbol{S}^{(1)},\cdots,\boldsymbol{S}^{(n+1)}$ 实为一组关于海森矩阵共轭的向量，所以 DFP 变尺度法属于共轭方向法。

当 $\boldsymbol{A}=\boldsymbol{I}$(单位矩阵)时，$[\boldsymbol{S}^{(i)}]^{\mathrm{T}}\boldsymbol{S}^{(j)}=0$，此即为通常的正交条件。共轭方向是正交方向的推广。

共轭方向法的一大特点：对于 n 维正定二次目标函数，若依次沿 n 个共轭方向进行一维搜索，则 n 步就可收敛到极小点。

(3) 每次迭代都能使目标函数值单调下降即为所谓的算法稳定性。DFP 变尺度法在稳定性方面存在一些问题，由于每次迭代的一维搜索只能获得一定的精确度，且存在计算机舍入误差，所以易使构造矩阵的正定性遭到破坏，以至算法不稳定。为了提高实际计算中的稳定性，通常采用“重置”的方法，即在 n 次迭代后重新设置单位矩阵。

DFGS 变尺度法对于维数较高问题具有更好的稳定性，公式为

$$\Delta\boldsymbol{A}^{(k)}=\frac{1}{\Delta\boldsymbol{X}^{(k)\mathrm{T}}\Delta\boldsymbol{g}^{(k)}}\left(\Delta\boldsymbol{X}^{\mathrm{T}}\Delta\boldsymbol{X}^{(k)T}+\frac{\Delta\boldsymbol{X}^{(k)}\Delta\boldsymbol{X}^{(k)\mathrm{T}}\Delta\boldsymbol{g}^{(k)}\boldsymbol{A}^{(k)}\Delta\boldsymbol{g}^{(k)}}{\Delta\boldsymbol{X}^{(k)}\Delta\boldsymbol{g}^{(k)}}\right.$$

$$\left.-\boldsymbol{A}^{(k)}\Delta\boldsymbol{g}^{(k)}\Delta\boldsymbol{X}^{(k)\mathrm{T}}-\Delta\boldsymbol{X}^{(k)}\Delta\boldsymbol{g}^{(k)T}\boldsymbol{A}^{(k)}\right)$$

式中各符号的意义与 DFP 法相同。

3.5　约束优化方法

3.5.1　概述

机械优化设计中的问题，大多数属于约束优化设计问题，其数学模型为

$$\begin{aligned}&\min f(\boldsymbol{X})=f(x_1,x_2,\cdots,x_n)\\ \text{s,t.}\quad &g_j(\boldsymbol{X})=g_j(x_1,x_2,\cdots,x_n)\leqslant 0 \qquad (j=1,2,\cdots,m)\\ &h_k(\boldsymbol{X})=h_k(x_1,x_2,\cdots,x_n)=0 \qquad (k=1,2,\cdots,l)\end{aligned}\tag{3-45}$$

求解式(3-45)的方法称为约束优化方法。根据求解方式的不同，可分为直接解法和间接解法等。

直接解法通常适用于仅含不等式约束的问题，它的基本思路是在 m 个不等式约束条件所确定的可行域内，选择一个初始点 $\boldsymbol{X}^0$，然后决定可行搜索方向 d，且以适当的步长 α，沿 d

方向进行搜索，得到一个使目标函数值下降的可行的新点 $\boldsymbol{X}^{(1)}$，即完成一次迭代(图 3-33)。再以新点为起点，重复上述搜索过程，满足收敛条件后，迭代终止。每次迭代计算均按以下基本迭代格式进行

$$\boldsymbol{X}^{(k+1)}=\boldsymbol{X}^{(k)}+a_k\boldsymbol{S}^{(k)},\quad k=1,2,\cdots \tag{3-46}$$

式中：a_k——步长；

$\boldsymbol{S}^{(k)}$——可行搜索方向。

所谓可行搜索方向是指，当设计点沿该方向作微量移动时，目标函数值将下降，且不会越出可行域。产生可行搜索方向的方法将由直接解法中的各种算法决定。

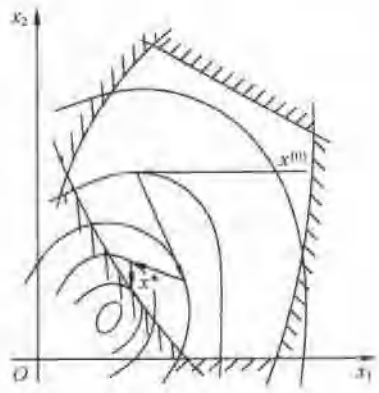


图 3-33　直接解法的探索路线

直接解法的原理简单，方法实用，其特点是：

(1) 由于整个求解过程在可行域内进行，因此，迭代计算不论何时终止，都可以获得一个比初始点好的设计点。

(2) 若目标函数为凸函数，可行域为凸集，则可保证获得全域最优解。否则，因存在多个局部最优解，当选择的初始点不相同时，可能搜索到不同的局部最优解。为此，常在可行域内选择几个差别较大的初始点分别进行计算，以便从求得的多个局部最优解中选择更好的最优解。

(3) 要求可行域为有界的非空集，即在有界可行域内存在满足全部约束条件的点，且目标函数有定义。

间接解法有不同的求解策略，其中一种解法的基本思路是将约束优化问题中的约束函数进行特殊的加权处理后，和目标函数结合起来，构成一个新的目标函数，即将原约束优化问题转化成为一个或一系列的无约束优化问题。再对新的目标函数进行无约束优化计算，从而间接地搜索到原约束问题的最优解。

间接解法的基本迭代过程是，首先将式(3-45)所示的约束优化问题转化成新的无约束目标函数

$$\phi(\boldsymbol{X},\mu_1,\mu_2)=f(\boldsymbol{X})+\sum_{j=1}^{m}\mu_1\boldsymbol{G}[g_j(\boldsymbol{X})]+\sum_{k=1}^{l}\mu_2\boldsymbol{H}[h_k(\boldsymbol{X})] \tag{3-47}$$

式中：$\phi(\boldsymbol{X},\mu_1,\mu_2)$——转换后的新目标函数；

μ_1、μ_2——加权因子；

$\sum_{j=1}^{m}\mu_1\boldsymbol{G}[g_j(\boldsymbol{X})]$ 和 $\sum_{k=1}^{l}\mu_2\boldsymbol{H}[h_k(\boldsymbol{X})]$——约束函数 $g_j(\boldsymbol{X})$ 和 $h_k(\boldsymbol{X})$ 经过加权处理后构成的某种形式的复合函数或泛函数。

然后对 $\phi(\boldsymbol{X},\mu_1,\mu_2)$ 进行无约束极小化计算。由于在新目标函数中包含了各种约束条件，在求极值的过程中还改变加权因子的大小，因此可以不断地调整设计点，使其逐步逼近约束边界，从而间接地求得原约束问题的最优解。图 3-34 所示的框图表示了这一基本迭代过程。

间接解法是目前在机械优化设计中得到广泛应用的一种有效方法，其特点是：

(1) 由于无约束优化方法的研究日趋成熟，已经研究出不少有效的无约束最优化方法

和程序，使得间接解法有了可靠的基础。目前，这类算法的计算效率和数值计算的稳定性也都有较大的提高。

(2) 可以有效地处理具有等式约束的约束优化问题。

(3) 间接解法存在的主要问题是，选取加权因子较为困难。加权因子选取不当，不但影响收敛速度和计算精度，甚至会导致计算失败。

求解约束优化设计问题的方法很多，本节将着重介绍属于直接解法的复合形法、可行方向法、广义简约梯度法，属于间接解法的惩罚函数法和增广乘子法。另外，还将对约束优化方法的另一类解法——二次规划法等作简要介绍。

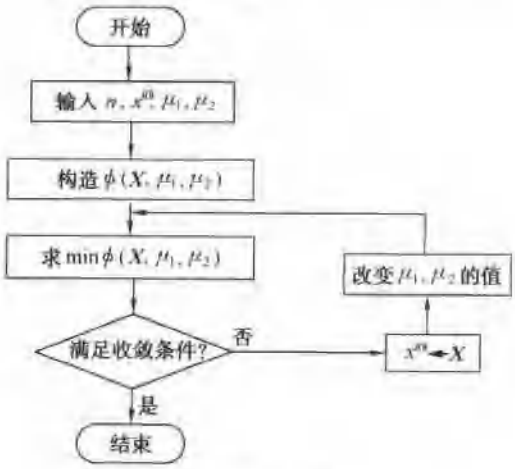


图 3-34　间接解法

3.5.2　复合形法

复合形法是求解约束优化问题的一种重要的直接解法。它的基本思路是在可行域内构造一个具有 k 个顶点的初始复合形。对该复合形各顶点的目标函数值进行比较，找到目标函数值最大的顶点(称最坏点)，然后按一定的法则求出目标函数值有所下降的可行的新点，并用此点代替最坏点，构成新的复合形。复合形的形状每改变一次，就向最优点移动一步，直至逼近最优点。

由于复合形的形状不必保持规则的图形，对目标函数及约束函数的性状又无特殊要求，因此该法的适应性较强，在机械优化设计中得到广泛应用。

1. 原理

复合形法是求解约束优化问题的一种重要的直接解法。它源自无约束优化问题的单纯形法，是单纯形法在约束优化问题中的发展。它与单纯形法的不同点在于，初始复合形的各顶点要满足约束条件(为可行点)，在随后的复合形顶点的选择与替换中，要同时满足函数值下降要求和约束条件。此外，复合形法需要在设计空间内构造的复合形的顶点数为 k 个 $(n+1\leqslant k\leqslant 2n)$。图 3-35 为复合形法的原理示意图。

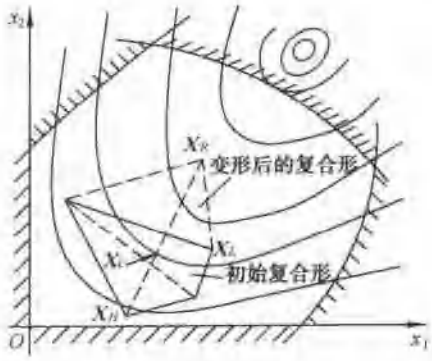


图 3-35 复合形法的原理示意图

2. 算法

(1) 形成初始复合形。

① 在设计变量少、约束函数简单的情况下，可由设计者决定 k 个可行点，构成初始复合形。

② 当设计变量较多或约束函数复杂时，由设计者决定 k 个可行点常常很困难，这时可采用以下方法生成初始复合形。

选定一个可行点作为初始顶点 $x_1^{(0)}$（控制初始复合形的位置），其余的 $k-1$ 个可行点用随机法产生。各顶点按下式计算：

$$\boldsymbol{X}_i^{(0)}=a+q_i(b-a) \quad (i=2,3,\cdots,k) \tag{3-48}$$

式中：a、b——设计变量的上、下限；

q_i——(0,1)区间内的伪随机数。

用式(3-44)计算得到的 $k-1$ 个随机点不一定都在可行域内，因此要设法将非可行点移到可行域内。通常采用的方法是先求出可行域内 q 个顶点 $\boldsymbol{X}_1^{(0)}$，$\boldsymbol{X}_2^{(0)}$，$\cdots$，$\boldsymbol{X}_q^{(0)}$ 的中心 X_c，即

$$\boldsymbol{X}_c=\frac{1}{q}\sum_{i=1}^{q}\boldsymbol{X}_i \tag{3-49}$$

然后将非可行点 $\boldsymbol{X}_{q+1}^{(0)}$，$\boldsymbol{X}_{q+2}^{(0)}$，$\cdots$，$\boldsymbol{X}_k^{(0)}$ 向中心点 $\boldsymbol{X}_c$ 移动，得新点

$$\begin{aligned}\boldsymbol{X}_{q+1}^{(0)}&=\boldsymbol{X}_{q+1}^{(0)}+\beta(\boldsymbol{X}_{q+1}^{(0)}-\boldsymbol{X}_c)\\&\cdots\\\boldsymbol{X}_k^{(0)}&=\boldsymbol{X}_k^{(0)}+\beta(\boldsymbol{X}_k^{(0)}-\boldsymbol{X}_c)\end{aligned} \tag{3-50}$$

一般取 $\beta=0.5$。若某一点仍为不可行点，则利用上式，使其继续向中心点移动。只要中心点为可行点，$k-1$ 个点经过上述处理后，最终全部成为可行点，并构成初始复合形。

③ 由计算机自动生成初始复合形的全部顶点。其方法是首先随机产生一个可行点，然后按②产生其余的 $k-1$ 个可行点。这种方法对设计者来说最为简单，但因初始复合形在可行域内的位置不能控制，可能会给以后的计算带来困难。

(2) 复合形法的搜索方法和计算步骤。

① 计算各顶点函数值，$F_i=F(\boldsymbol{X}_i)$；比较函数值的大小，确定最好点 $\boldsymbol{X}_L$、最差点 $\boldsymbol{X}_H$、次差点 $\boldsymbol{X}_G$，即

$$\begin{aligned} F_L &= F(\boldsymbol{X}_L) = \min \boldsymbol{F}(\boldsymbol{X}_i) \quad (i=1,2,\cdots,k) \\ F_H &= F(\boldsymbol{X}_H) = \max \boldsymbol{F}(\boldsymbol{X}_i) \quad (i=1,2,\cdots,k) \\ F_G &= F(\boldsymbol{X}_G) = \max \boldsymbol{F}(\boldsymbol{X}_i) \quad (i=1,2,\cdots,k;i\neq H) \\ F_H &> F_G > F_L \end{aligned} \tag{3-51}$$

② 计算 $\boldsymbol{X}_H$ 点之外各点的"重心"$\boldsymbol{X}_c$，即

$$\boldsymbol{X}_c = \frac{1}{k-1}\left(\sum_{i=1}^{k}\boldsymbol{X}_i - \boldsymbol{X}_H\right) \tag{3-52}$$

③ 如果 $\boldsymbol{X}_c$ 在可行域内，则沿 $\boldsymbol{X}_H\boldsymbol{X}_c$ 方向上作 $\boldsymbol{X}_H$ 点相对于 $\boldsymbol{X}_c$ 点的反射点 $\boldsymbol{X}_R$，

$$\boldsymbol{X}_R = \boldsymbol{X}_c + \alpha(\boldsymbol{X}_c - \boldsymbol{X}_H)$$

式中：α——反射系数，一般取 $\alpha=1.3$。

判别反射点 $\boldsymbol{X}_R$ 是否为可行点，如在可行域外，则将 α 减半，重新计算反射点，直至满足全部约束。

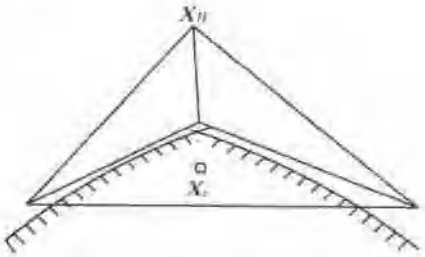


图 3-36 中心点 X_c 不在可行域之内

④ 如果中心点 $\boldsymbol{X}_c$ 不在可行域之内，可行域则可能为非凸集(图 3-36)。为了将 $\boldsymbol{X}_c$ 移至可行域内，以 $\boldsymbol{X}_c$ 和 $\boldsymbol{X}_L$ 为界的超正方体，重新利用伪随机数产生 k 个新的顶点，构成新的复合形(算法和计算步骤见形成初始复合形)。此时，变量的上、下限修改如下：

若 $x_{Li} < x_{ci}$ $(i=1,2,\cdots,n)$，则

$$\left.\begin{aligned} a_i &= x_{Li} \\ b_i &= x_{ci} \end{aligned}\right\} \quad (i=1,2,\cdots,n) \tag{3-53}$$

否则相反。重复①、②，直至 $\boldsymbol{X}_c$ 及 $\boldsymbol{X}_H$ 点相对于 $\boldsymbol{X}_c$ 点的反射点 $\boldsymbol{X}_R$ 都进入可行域为止。

⑤ 计算 $F(\boldsymbol{X}_R)$，如果 $F(\boldsymbol{X}_R) < F(\boldsymbol{X}_H)$，则用 $\boldsymbol{X}_R$ 代替 $\boldsymbol{X}_H$ 构成新复合形，转入①开始下一轮搜索；否则继续⑥。

⑥ 如果 $F(\boldsymbol{X}_R) > F(\boldsymbol{X}_H)$，则将 α 减半，重新计算 $\boldsymbol{X}_R$，直至 $F(\boldsymbol{X}_R) < F(\boldsymbol{X}_H)$。若 $F(\boldsymbol{X}_R) < F(\boldsymbol{X}_H)$ 且 $\boldsymbol{X}_R$ 为可行点，转⑤；若经过若干次 α 减半的计算，使 α 值小于给定的很小的正数 $\zeta(\zeta=10^{-3})$ 时，仍不能找到正确的反射点，将最差点 $\boldsymbol{X}_H$ 换为次差点 $\boldsymbol{X}_G$ 并转入②。当复合形收缩到很小时

$$\max_{1\leqslant i\leqslant k} \| \boldsymbol{X}_i - \boldsymbol{X}_c \| < \varepsilon_1 \tag{3-54}$$

或各顶点目标函数值满足

$$\left\{\frac{1}{k}\sum_{i=1}^{k}[F(\boldsymbol{X}_i) - F(\boldsymbol{X}_c)]^2\right\}^{\frac{1}{2}} < \varepsilon_2 \tag{3-55}$$

时，停止迭代，$\boldsymbol{X}^* = \boldsymbol{X}_L$ 即为最优解。

以上①～⑥是只含反射的基本复合形法及其迭代计算步骤。反射是复合形法的一种主要寻优搜索算法。除反射算法外，在复合形寻优搜索中还可采用将反射点扩张、压缩，复合形向最好点收缩，将 $\boldsymbol{X}_H$ 在 $\boldsymbol{X}_H$、$\boldsymbol{X}_L$、$\boldsymbol{X}_c$ 决定的平面内绕最好点 $\boldsymbol{X}_L$ 旋转某一角度并向 $\boldsymbol{X}_L$ 靠拢等算法。

3. 讨论

(1) 复合形法无需计算目标函数的一、二阶导数，也无需进行一维搜索优化，因此对目标函数和约束函数无特殊要求，对约束优化问题的适应性较强，算法较简单，但随着问题维数和约束条件的增多，计算效率显著降低。当 $n\leqslant 5$ 时，可取复合形的顶点数 $k=2n$；当 $n>5$ 时，可适当减少顶点数，但不能少于 $n+1$ 个。

(2) 复合形法一般仅适用于不等式约束。

例 3-12 求解约束优化问题：

$$\min F(\boldsymbol{X})=(x_1-5)^2+4(x_2-6)^2$$

$$\text{s.t. } g_1(\boldsymbol{X})=64-x_1^2-x_2^2\leqslant 0$$

$$g_2(\boldsymbol{X})=x_2-x_1-10\leqslant 0$$

$$g_3(\boldsymbol{X})=x_1-10\leqslant 0$$

初始点取$[8 \quad 14]^{\mathrm{T}}$(给定的初始复合形第一个顶点)，迭代 67 次后的近似最优解为 $\boldsymbol{X}^*=[5.21975 \quad 6.06253]^{\mathrm{T}}$，$F(\boldsymbol{X}^*)=0.06393$，计算结果见表 3-2。问题图解如图 3-37 所示。

表 3-2 计算结果

k	x_1	x_2	$F(\boldsymbol{X})$	k	x_1	x_2	$F(\boldsymbol{X})$
0	8	14	100	40	5.25561	6.06049	0.07997
10	4.43521	6.90164	3.570084	50	5.20952	6.07303	0.06523
20	5.35314	6.68238	1.98728	60	5.21975	6.06253	0.06402
30	6.58604	6.06063	0.35813	67	5.21975	6.06253	0.06393

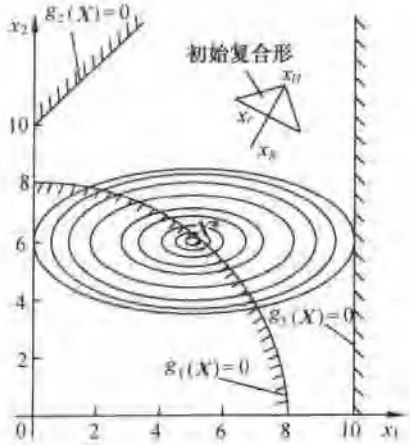


图 3-37 例 3-12 问题图解

3.5.3 可行方向法

约束优化问题的直接解法中，可行方向是最大的一类，它也是求解大型约束优化问题的主要方法之一。这种方法的基本原理是在可行域内选择一个初始点 $\boldsymbol{X}^{(0)}$，当确定了一个可行方向 $\boldsymbol{S}$ 和适当的步长后，按下式

$$\boldsymbol{X}^{(k+1)}=\boldsymbol{X}^{(k)}+\alpha\boldsymbol{S}^{(k)}, \quad k=1,2,\cdots \tag{3-56}$$

进行迭代计算。在不断调整可行方向的过程中，使迭代点逐步逼近约束最优点。

1. 可行方向法的搜索策略

可行方向法的第一步迭代都是从可行的初始点 $\boldsymbol{X}^{(0)}$ 出发，沿点 $\boldsymbol{X}^{(0)}$ 的负梯度方向 $\boldsymbol{S}^{(0)}=-\nabla f(\boldsymbol{X}^{(0)})$，将初始点移动到某一个约束面（只有一个起作用的约束时）上或约束面的交集（有几个起作用的约束时）上。然后根据约束函数和目标函数的不同性状，分别采用以下几种策略继续搜索。

第一种情况如图 3-38 所示，在约束面上的迭代点 $\boldsymbol{X}^{(k)}$ 处，产生一个可行方向 $\boldsymbol{S}^{(k)}$，沿此方向作一维最优化搜索，所得到的新点 $\boldsymbol{X}$ 在可行域内，即令 $\boldsymbol{X}^{(k+1)}=\boldsymbol{X}$，再沿点 $\boldsymbol{X}^{(k+1)}$ 的负梯度方向 $\boldsymbol{S}^{(k+1)}=-\nabla f(\boldsymbol{X}^{(k+1)})$ 继续搜索。

第二种情况如图 3-39 所示，沿可行方向 $\boldsymbol{S}^{(k)}$ 作一维最优化搜索，所得到的新点 $\boldsymbol{X}$ 在可行域外，则设法将点 $\boldsymbol{X}$ 移动到约束面上，即取 $\boldsymbol{S}^{(k)}$ 和约束面的交点作为新的迭代点 $\boldsymbol{X}^{(k+1)}$。

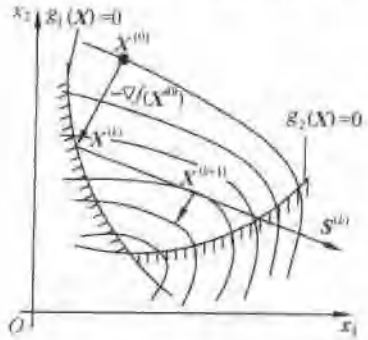


图 3-38　新点 $\boldsymbol{X}$ 在可行域内的情况

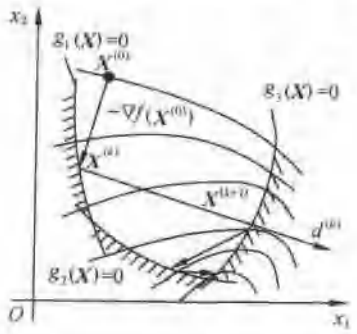


图 3-39　新点 $\boldsymbol{X}$ 在可行域外的情况

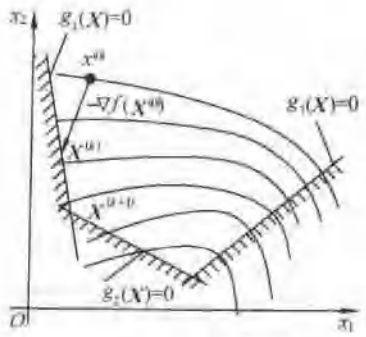


图 3-40　沿线性约束面搜索

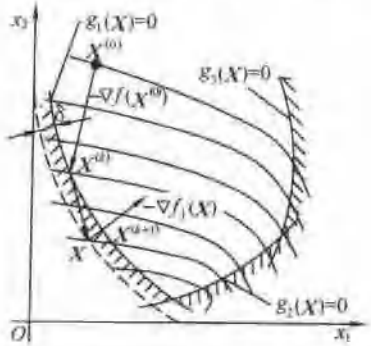


图 3-41　沿非线性约束面搜索

第三种情况是沿约束面搜索。对于只具有线性约束条件的非线性规划问题（图 3-40），从点 $\boldsymbol{X}^{(k)}$ 出发，沿约束面移动，在有限的几步内即可搜索到约束最优点；对于非线性约束函数（图 3-41），沿约束面移动将会进入非可行域，使问题变得复杂得多。此时，需将进入非可行域的新点 $\boldsymbol{X}$ 设法调整到约束面上，然后才能进行下一次迭代。调整的方法是先规定约束面容差 δ，建立新的约束边界，如图 3-41 的虚线所示。然后将已离开约束面的点 $\boldsymbol{X}$，沿起作用约束函数的负梯度方向 $-\nabla g(\boldsymbol{X})$ 返回到约束面上。其计算公式为

$$\boldsymbol{X}^{(k+1)}=\boldsymbol{X}+\alpha_1\nabla g(\boldsymbol{X}) \tag{3-57}$$

式中 a_t 表示调整步长，可用试探法决定，或用下式估算

$$a_t=\left|\frac{g(\boldsymbol{X})}{\nabla g(\boldsymbol{X})^{\mathrm{T}}\nabla g(\boldsymbol{X})}\right| \tag{3-58}$$

2. *产生可行方向的条件*

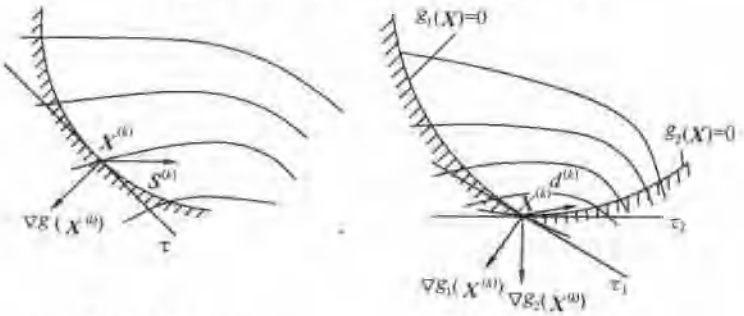
可行方向是指沿该方向作微小移动后，所得到的新点是可行点，且目标函数值有所下降。显然，可行方向应满足可行和下降两个条件。

(1) 可行条件。方向的可行条件是指沿该方向作微小移动后，所得到的新点为可行点。如图 3-42(a)所示。若点 $\boldsymbol{X}^{(k)}$ 在一个约束面上，过点 $\boldsymbol{X}^{(k)}$ 作约束面 $g(\boldsymbol{X})=0$ 的切线 τ，显然满足可行条件的方向 $\boldsymbol{S}^{(k)}$ 应与起作用的约束函数在点 $\boldsymbol{X}^{(k)}$ 的梯度 $\nabla g(\boldsymbol{X}^{(k)})$ 的夹角大于或等于 90°。用向量关系式可表示为

$$[\nabla g(\boldsymbol{X}^{(k)})]^{\mathrm{T}}d^{(k)}\leqslant 0 \tag{3-59}$$

若点 $\boldsymbol{X}^{(k)}$ 在 J 个约束面的交集上，如图 3-42(b)所示，为保证方向 $\boldsymbol{S}^{(k)}$ 可行，要求 $\boldsymbol{S}^{(k)}$ 和 J 个约束函数在点 $\boldsymbol{X}^{(k)}$ 的梯度 $\nabla g_j(\boldsymbol{X}^{(k)})(j=1,2,\cdots,J)$ 的夹角均大于等于 90°。其向量关系可表示为

$$[\nabla g_j(\boldsymbol{X}^{(k)})]^{\mathrm{T}}\boldsymbol{S}^{(k)}\leqslant 0,\quad j=1,2,\cdots,J \tag{3-60}$$



(a) 一个起作用的约束　　(b) 两个起作用的约束

图 3-42　方向的可行条件

(2) 下降条件。方向的下降条件是指沿该方向作微小移动后，所得新点的目标函数值是下降的。如图 3-43 所示，满足下降条件的方向 $\boldsymbol{S}^{(k)}$ 应和目标函数在点 $\boldsymbol{X}^{(k)}$ 的梯度 $\nabla f(\boldsymbol{X}^{(k)})$ 的夹角大于 90°。其向量关系可表示为

$$[\nabla f(\boldsymbol{X}^{(k)})]^{\mathrm{T}}\boldsymbol{S}^{(k)}<0 \tag{3-61}$$

满足可行和下降条件，即式(3-60)、(3-61)同时成立的方向称可行方向。如图 3-44 所示，它位于约束曲面在点 $\boldsymbol{X}^{(k)}$ 的切线和目标函数等值线在点 $\boldsymbol{X}^{(k)}$ 的切线所围成的扇形区内，该扇形区称为可行下降方向区。

综上所述，当点 $\boldsymbol{X}^{(k)}$ 位于 J 个起作用的约束面上时，满足

$$\begin{cases}[\nabla g_j(\boldsymbol{X}^{(k)})]^{\mathrm{T}}\boldsymbol{S}^{(k)}\leqslant 0, j=1,2,\cdots,J\\ [\nabla f(\boldsymbol{X}^{(k)})]^{\mathrm{T}}\boldsymbol{S}^{(k)}<0\end{cases} \tag{3-62}$$

的方向 $\boldsymbol{S}^{(k)}$ 称可行方向。

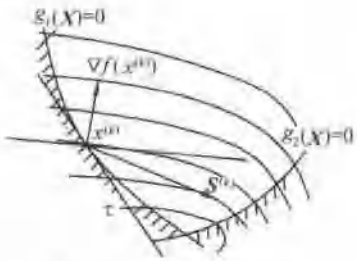


图 3-43　方向的下降条件

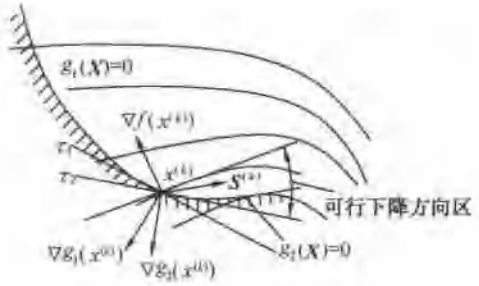


图 3-44　可行下降方向区

3. 可行方向的产生方法

如上所述，满足可行、下降条件的方向位于可行下降扇形区内，在扇形区内寻找一个最有利的方向作为本次迭代的搜索方向，其方法主要有优选方向法和梯度投影法两种。

(1) 优选方向法。在由式(3-62)构成的可行下降扇形区内选择任一方向 $\boldsymbol{S}$ 进行搜索，可得到一个目标函数值下降的可行点。现在的问题是如何在可行下降扇形区内选择一个能使目标函数下降最快的方向作为本次迭代的方向。显然，这是一个以搜索方向 $\boldsymbol{S}$ 为设计变量的约束优化问题，这个新的约束优化问题的数学模型可写成

$$\begin{aligned} &\min[\nabla f(\boldsymbol{X}^{(k)})]^{\mathrm{T}}\boldsymbol{S} \\ &[\nabla g_j(\boldsymbol{X}^{(k)})]^{\mathrm{T}}\boldsymbol{S}^{(k)}\leqslant 0 \quad j=1,2,\cdots,J \\ \text{s.t.}\quad &[\nabla f(\boldsymbol{X}^{(k)})]^{\mathrm{T}}\boldsymbol{S}<0 \\ &\|\boldsymbol{S}\|\leqslant 1 \end{aligned} \tag{3-63}$$

由于 $\nabla f(\boldsymbol{X}^{(k)})$ 和 $\nabla g_j(\boldsymbol{X}^{(k)})(j=1,2,\cdots,J)$ 为定值，上述各函数均为设计变量 $\boldsymbol{S}$ 的线性函数，因此式(3-63)为一个线性规划问题。用线性规划法求解后，求得的最优解 $\boldsymbol{S}^*$ 即为本次迭代的可行方向，即 $\boldsymbol{S}^{(k)}=\boldsymbol{S}^*$。

(2) 梯度投影法。当点 $x^{(k)}$ 目标函数的负梯度方向 $-\nabla f(\boldsymbol{X}^{(k)})$ 不满足可行条件时，可将 $-\nabla f(\boldsymbol{X}^{(k)})$ 方向投影到约束面(或约束面的交集)上，得到投影向量 $\boldsymbol{S}^{(k)}$。从图 3-45 中可看出，该投影向量显然满足方向的可行和下降条件。梯度投影法就是取该方向作为本次迭代的可行方向。可行方向的计算公式为

$$\boldsymbol{S}^{(k)}=\frac{-\boldsymbol{P}\nabla f(\boldsymbol{X}^{(k)})}{\|\boldsymbol{P}\nabla f(\boldsymbol{X}^{(k)})\|} \tag{3-64}$$

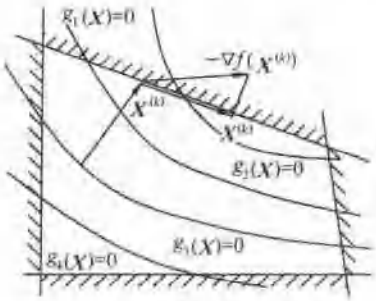


图 3-45　约束面上的梯度投影方向

式中：$\nabla f(\boldsymbol{X}^{(k)})$——点 $\boldsymbol{X}^{(k)}$ 的目标函数梯度；

$\boldsymbol{P}$——投影算子，为 $n\times n$ 阶矩阵，其计算公式为

$$\boldsymbol{P}=\boldsymbol{I}-\boldsymbol{G}[\boldsymbol{G}^{\mathrm{T}}\boldsymbol{G}]^{-1}\boldsymbol{G}^{\mathrm{T}} \tag{3-65}$$

式中：$\boldsymbol{I}$——单位矩阵，$n\times n$ 阶矩阵；

$\boldsymbol{G}$——起作用约束函数的梯度矩阵，$n\times j$ 阶矩阵

$$\boldsymbol{G}=[\nabla g_1(\boldsymbol{X}^{(k)}),\nabla g_2(\boldsymbol{X}^{(k)}),\cdots,\nabla g_j(x^{(k)})] \tag{3-66}$$

其中 j 为起作用的约束函数个数。

4. 步长的确定

可行方向 $S^{(k)}$ 确定后，按下式计算新的迭代点

$$X^{(k+1)}=X^{(k)}+a_k S^{(k)} \tag{3-67}$$

由于目标函数及约束函数的性状不同，步长 a_k 的确定方法也不同，不论是用何种方法，都应使新的迭代点 $X^{(k+1)}$ 为可行点，且目标函数具有最大的下降量。确定步长 a_k 的常用方法有以下两种：

(1) 取最优步长。如图 3-46 所示，从点 $X^{(k)}$ 出发，沿 $S^{(k)}$ 方向进行一维最优化搜索，取得最优步长 a^*，计算新点 x 的值

$$X=X^{(k)}+a^* S^{(k)}$$

若新点 X 为可行点，则本次迭代的步长取 $a_k=a^*$。

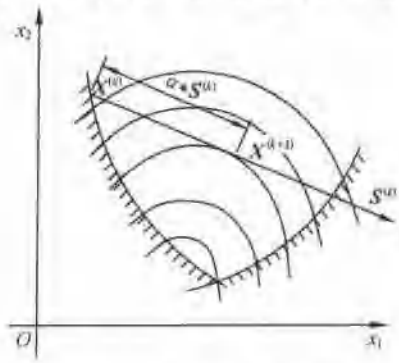


图 3-46 按最优步长确定新点

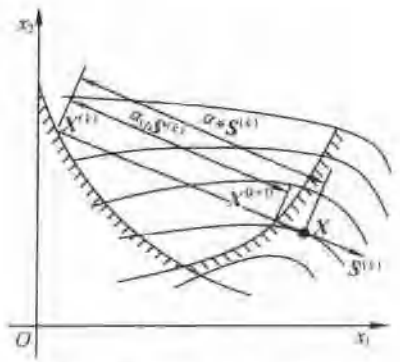


图 3-47 按最大步长确定新点

(2) a_k 取到约束边界的最大步长。如图 3-47 所示，从点 $X^{(k)}$ 沿 $S^{(k)}$ 方向进行一维最优化搜索，得到的新点 X 为不可行点，根据可行方向法的搜索策略，应改变步长：使新点 X 返回到约束面上来。使新点 X 恰好位于约束面上的步长称为最大步长，记为 a_M。则本次迭代的步长取 $a_k=a^*$。

5. 收敛条件

按可行方向法的原理，将设计点调整到约束面上后，需要判断迭代是否收敛，即判断该迭代点是否为约束最优点。常用的收敛条件有以下两种：

(1) 设计点 $X^{(k)}$ 及约束允差满足

$$\begin{cases} \left|[\nabla f(X^{(k)})]^{\mathrm{T}} S^{(k)}\right| \leqslant \varepsilon \\ \delta \leqslant \varepsilon_2 \end{cases} \tag{3-68}$$

条件时，迭代收敛。

(2) 设计点 $X^{(k)}$ 满足库恩-塔克条件

$$\begin{cases} \nabla f(X^{(k)})+\sum\limits_{j=1}^{J}\lambda_j \nabla g_j(X^{(k)})=0 \\ \lambda_j \geqslant 0 \quad j=1,2,\cdots,J \end{cases} \tag{3-69}$$

时，迭代收敛。

3.5.4 惩罚函数法

惩罚函数法是一种使用很广泛、很有效的间接解法。它的基本原理是将约束优化问题

中的不等式和等式约束函数经过加权转化后，和原目标函数结合形成新的目标函数——惩罚函数。求解该新目标函数的无约束极小值，以期得到原问题的约束最优解。为此，按一定的法则改变加权因子 r_1 和 r_2 的值，构成一系列的无约束优化问题，求得一系列的无约束最优解，并不断地逼近原约束优化问题的最优解。因此惩罚函数法又称序列无约束极小化方法，常称 SUMT 法。

$$\begin{cases} & \min f(\boldsymbol{X}) \\ \text{s.t.} & g_j(\boldsymbol{X}) \leqslant 0 \quad j=1,2,\cdots,m \\ & h_k(\boldsymbol{X}) = 0 \quad k=1,2,\cdots,l \end{cases}$$

$$\phi(x,r_1,r_2) = f(\boldsymbol{X}) + r_1 \sum_{j=1}^{m} G[g_j(\boldsymbol{X})] + r_2 \sum_{k=1}^{l} H[h_k(\boldsymbol{X})] \tag{3-70}$$

式(3-70)中的 $r_1 \sum_{j=1}^{m} G[g_j(\boldsymbol{X})]$ 和 $r_2 \sum_{k=1}^{l} H[h_k(\boldsymbol{X})]$ 称为加权转化项。根据它们在惩罚函数中的作用，又分别称为障碍项和惩罚项。障碍项的作用是当迭代点在可行域内时，在迭代过程中将阻止迭代点越出可行域。惩罚项的作用是当迭代点在非可行域或不满足等式约束条件时，在迭代过程中将迫使迭代点逼近约束边界或等式约束曲面。

根据迭代过程是否在可行域内进行，惩罚函数法又可分为内点惩罚函数法、外点惩罚函数法和混合惩罚函数法三种。

1. 内点惩罚函数法

内点惩罚函数法简称内点法，这种方法将新目标函数定义于可行域内，序列迭代点在可行域内逐步逼近约束边界上的最优点。内点法只能用来求解具有不等式约束的优化问题。

对于只具有不等式约束的优化问题

$$\begin{cases} & \min f(\boldsymbol{X}) \\ \text{s.t.} & g_j(\boldsymbol{X}) \leqslant 0 \quad j=1,2,\cdots,m \end{cases}$$

转化后的惩罚函数形式为

$$\phi(\boldsymbol{X},r) = f(\boldsymbol{X}) - \sum_{j=1}^{m} \frac{1}{g_j(\boldsymbol{X})} \tag{3-71}$$

或

$$\phi(\boldsymbol{X},r) = f(\boldsymbol{X}) - \sum_{j=1}^{m} \ln[-g_j(\boldsymbol{X})] \tag{3-72}$$

式中：　r——惩罚因子，它是由大到小且趋近于 0 的数列，即 $r^{(0)} > r^{(1)} > r^{(2)} \cdots \to 0$；

$\sum_{j=1}^{m} \frac{1}{g_j(\boldsymbol{X})}$ 或 $\sum_{j=1}^{m} \ln[-g_j(\boldsymbol{X})]$——障碍项。

由于内点法的迭代过程在可行域内进行，障碍项的作用是阻止迭代点越出可行域。由障碍项的函数形式可知，当迭代点靠近某一约束边界时，其值趋近于 0，而障碍项的值陡然增加，并趋近于无穷大，好像在可行域的边界上筑起了一道“围墙”，使迭代点始终不能越出可行域。显然，只有当惩罚因子 $r \to 0$ 时，才能求得在约束边界上的最优解。

2. 外点惩罚函数法

外点惩罚函数法简称外点法。这种方法和内点法相反，新目标函数定义在可行域之外，序列迭代点从可行域之外逐渐逼近约束边界上的最优点。外点法可以用来求解含不等式和等式约束的优化问题。

对于约束优化问题

$$\begin{cases} \min f(\boldsymbol{X}) \\ \text{s.t.} \quad g_j(\boldsymbol{X}) \leqslant 0 \quad j=1,2,\cdots,m \\ \quad\quad h_k(\boldsymbol{X}) = 0 \quad k=1,2,\cdots,l \end{cases}$$

转化后的外点惩罚函数的形式为

$$\phi(\boldsymbol{X},r) = f(\boldsymbol{X}) + r\sum_{j=1}^{m}\{\max[0,g_j(\boldsymbol{X})]\}^2 + r\sum_{k=1}^{l}[h_k(\boldsymbol{X})]^2 \tag{3-73}$$

式中：r——惩罚因子，它是由小到大，且趋近于∞的数列，即 $r^{(0)}<r^{(1)}<r^{(2)}\cdots\to\infty$；

$\sum_{j=1}^{m}\{\max[0,g_j(\boldsymbol{X})]\}^2$，$\sum_{k=1}^{l}[h_k(\boldsymbol{X})]^2$—— 对应于不等式约束和等式约束函数的惩罚项。

由于外点法的迭代过程在可行域之外进行。惩罚项的作用是迫使迭代点逼近约束边界或等式约束曲面。由惩罚项的形式可知，当迭代点 $\boldsymbol{X}$ 不可行时，惩罚项的值大于 0，使得惩罚函数 $\phi(\boldsymbol{X},r)$ 大于原目标函数，这可看成是对迭代点不满足约束条件的一种惩罚。当迭代点离约束边界愈远，惩罚项的值愈大，这种惩罚愈重。但当迭代点不断接近约束边界和等式约束曲面时，惩罚项的值减小，且趋近于 0，惩罚项的作用逐渐消失，迭代点也就趋近于约束边界上的最优点了。

3. 混合惩罚函数法

混合惩罚函数法简称混合法，这种方法是把内点法和外点法结合起来，用来求解同时具有等式约束和不等式约束函数的优化问题。

$$\begin{cases} \min f(\boldsymbol{X}) \\ \text{s.t.} \quad g_j(\boldsymbol{X}) \leqslant 0 \quad j=1,2,\cdots,m \\ \quad\quad h_k(\boldsymbol{X}) = 0 \quad k=1,2,\cdots,l \end{cases}$$

转化后混合惩罚函数的形式为

$$\phi(\boldsymbol{X},r) = f(\boldsymbol{X}) - r\sum_{j=1}^{m}\frac{1}{g_j(\boldsymbol{X})} + \frac{1}{\sqrt{r}}\sum_{k=1}^{l}[h_k(\boldsymbol{X})]^2 \tag{3-74}$$

式中：$r\sum_{j=1}^{m}\frac{1}{g_j(\boldsymbol{X})}$ 惩罚项，惩罚因子为 r 按内点法选，取 $r^{(0)}>r^{(1)}>r^{(2)}\cdots\to 0$。

$\frac{1}{\sqrt{r}}\sum_{k=1}^{l}[h_k(\boldsymbol{X})]^2$—— 惩罚项，惩罚因子为 $\frac{1}{\sqrt{r}}$，$r\to 0$ 时，$\frac{1}{\sqrt{r}}\to\infty$ 满足外点法对惩罚因子的要求。

混合法具有内点法的求解特点，即迭代过程在可行域内进行，因而初始点 $\boldsymbol{X}^{(0)}$、惩罚因子的初值 $r^{(0)}$ 均可参考内点法选取。计算步骤及程序框图也与内点法相近。

习 题 3

1. 某工厂生产一批金属工具箱，要求工具箱的体积为 0.5m^3，高度不低于 0.8m，试写出耗费金属板面积为最小的优化设计数学模型。

2. 已知函数 $f(\boldsymbol{X})=x_1^2+2x_1x_2+x_2^2$，试用矩阵形式表达该函数 $f(\boldsymbol{X})$，并说明该函数 $f(\boldsymbol{X})$ 是否为凸函数。

3. 试求函数 $f(\boldsymbol{X})=x_1^2+x_1x_2+x_2^2-60x_1-3x_2$ 的极值点，并判断该点是极大点还是极小点。

4. 试用进退法确定函数 $f(\boldsymbol{X})=3x^2-8x+9$ 的初始单峰区间。设给定的初始点 $\boldsymbol{X}_0=0$，初始步长 $h=0.1$。

5. 试用0.618法和二次插值法分别求解目标函数 $f(\boldsymbol{X})=8x^3-2x^2-7x+3$ 的最优解。已知初始单峰区间为[0,2]，迭代精度 $\varepsilon=0.01$。

6. 试用鲍威尔法从 $\boldsymbol{X}^{(0)}=[2,2]^{\mathrm{T}}$ 开始求目标函数 $f(\boldsymbol{X})=2x_1^2+x_2^2-x_1x_2$ 的最优解，并用表格列出各次搜索方向。

7. 试用DFP变尺度法求目标函数 $f(\boldsymbol{X})=x_1^2+2x_2^2-2x_1x_2-4x_1$ 的最优解。设给定的初始点为 $\boldsymbol{X}^{(0)}=[1,1]^{\mathrm{T}}$，初始构造矩阵 $\boldsymbol{A}^{(0)}=\boldsymbol{I}$，迭代精度 $\varepsilon=0.01$。

第4章 可靠性设计

本章从可靠性基本概念出发，先后介绍了可靠度、产品失效形式、可靠性常用分布函数等，在此基础上，进一步介绍了可靠性设计原理、机械静强度可靠性设计及机械系统可靠性设计等原理及方法。

4.1 可靠性基本概念和理论

可靠性是产品在规定条件下和规定时间内，完成规定功能的能力。这其中的两个规定具有数值的概念。一个数值是“规定时间”内，它具有一定寿命的数值概念，不能认为寿命愈长愈好，要有一个最经济有效的使用寿命。当然，这个规定的时间指的是产品出厂后的一段时间，这一段时间可以叫做产品的保险期。另一个数值是“规定功能”，它说的是保持功能参数在一定界限值之内的能力，不能任意扩大界限值的范围。

产品丧失规定的功能称为出故障，对不可修复或不予修复的产品而言，它又称为失效。为保持或恢复产品能完成规定功能的能力而采取的技术管理措施称为维修。可以维修的产品在规定条件下使用，在规定的时间内按规定的程序和方法进行维修时，保持或恢复到能完成规定功能的能力，称为产品的维修性或维修度。我们把可以维修的产品在某时刻所具有的，或能维持规定功能的能力称为可用性或可利用度或有效度。

当所考虑的产品是由部件或子系统所组成的系统时，我们不能期望它的组成部件或子系统都是等寿命的。因为影响各组成部件或子系统的因素是复杂的，所以研究可靠性目前都应当考虑应用概率和统计的数学方法。

可靠性的数值标准常用以下的指标(或称特征值)：

(1) 可靠度(reliability)；

(2) 累积失效概率或不可靠度(failure rate)；

(3) 平均寿命(mean life)；

(4) 有效寿命(useful life)；

(5) 维修度(maintainability)；

(6) 有效度(availability)；

(7) 重要度(importance)等。

它们统称为“可靠性尺度”。有了尺度，则在设计和生产时就可用数学方法来计算和预测，也可以用试验方法来评定产品或系统的可靠性。

4.1.1 可靠度 $R(t)$ 和累积失效概率 $F(t)$

产品的可靠度是时间的函数，用概率来表示。如果产品的寿命为 T(随机变量)，则产品在 t 时刻的可靠度 $R(t)$($T>t$)为这个随机事件的概率，即

$$R(t)=P(T>t),\ t>0 \tag{4-1}$$

由概率定义得

$$0 \leqslant R(t) \leqslant 1$$

如有一批数量为 n 的相同产品，在时刻 $t=0$ 开始工作，随着时间的推移，失效（或故障）的件数 $n_f(t)$ 在增大，而正常工作的件数 $n_s(t)$ 在减小，则产品在任意时刻 t 可靠度的观测值为

$$\overline{R}(t)=\frac{n_s(t)}{n} \tag{4-2}$$

这里 $\overline{R}(t)$ 表示完好产品在 n 件产品中出现的频率，则有

$$R(t)=\lim_{n\to\infty}\overline{R}(t)=\lim_{n\to\infty}\left[\frac{n_s(t)}{n}\right]$$

若某种产品工作至 2000h 的可靠度 $R(t)=0.95$，则表明有 95%的产品可以工作 2000h 以上，或对一件产品而言，它工作 2000h 以上的可能性为 95%。显然，可靠度 $R(t)$ 是评价产品可靠性的最重要的定量指标。

由于产品的失效与正常为对立事件，因而产品从时刻 $t=0$ 开始，工作至任意时刻 t 的累积失效概率 $F(t)$，即不可靠度

$$F(t)=1-R(t) \tag{4-3}$$

或

$$F(t)=P(T\leqslant t),\ t>0 \tag{4-4}$$

其观测值为

$$\overline{F}(t)=\frac{n_f(t)}{n} \tag{4-5}$$

当 $n\to\infty$ 时，就可以用频率来近似表示概率，即 $\overline{F}(t)\to F(t)$。

由上而的定义可知，$F(t)$ 就是产品寿命 T 的分布函数。

4.1.2　失效密度 $f(t)$

失效密度 $f(t)$ 的观测值为产品在 t 到 $t+\Delta t$ 的时间间隔内，单位时间内的失效频率，即

$$\overline{f}(t)=\frac{n_f(t+\Delta t)-n_f(t)}{n\Delta t}=\frac{\Delta n_f(t)}{n\Delta t} \tag{4-6}$$

式中：$n_f(t)$——n 个产品工作到 t 时刻的失效数；

$\Delta n(t)$——Δt 时间间隔内产品的失效数。

当 $\Delta t\to 0$，$n\to\infty$ 时，产品在 t 时刻的失效密度 $f(t)$，即概率密度函数为

$$f(t)=\lim_{\substack{n\to\infty\\ \Delta t\to 0}}\frac{\Delta n_f(t)}{n\Delta t}=\frac{\mathrm{d}n_f(t)}{n\mathrm{d}t}=\frac{\mathrm{d}F(t)}{\mathrm{d}t} \tag{4-7}$$

上式可改写为

$$F(t)=\int_0^t f(t)\mathrm{d}t \tag{4-8}$$

因此

$$R(t)=1-F(t)=\int_t^\infty f(t)\mathrm{d}t \tag{4-9}$$

显然 $F(t)$ 随着时间的增大而增大，$R(t)$ 随着时间 t 的增大而减小。由式(5-3)可得

$$f(t)=-\frac{\mathrm{d}R(t)}{\mathrm{d}t} \tag{4-10}$$

图 4-1 表示了 $R(t)$、$F(t)$和 $f(t)$三者之间的关系。从图中可以看出，$R(t)$和 $F(t)$分别为失效密度曲线 $f(t)$下面的两块面积，其和等于 1，$F(t)$为 $T\leqslant t$ 部分，$R(t)$为 $T>t$ 部分。

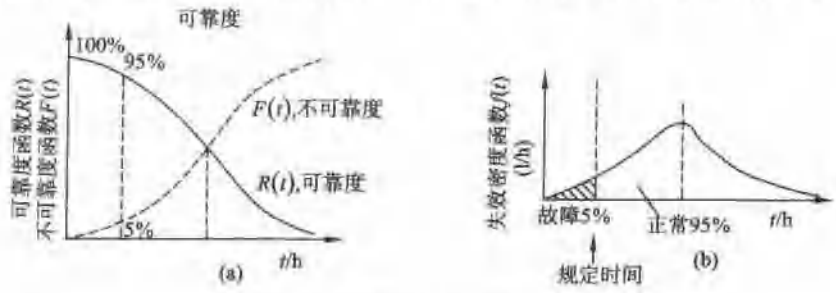


图 4-1 $R(t)$、$F(t)$和 $f(t)$的关系

例 4-1 某批电子器件有 1000 个，开始工作至 500h 内有 100 个损坏，工作至 1000h 共有 500 个损坏，求该批电子器件工作到 500h 和 1000h 的可靠度。

解：$n=1000$，$n_f(500)=100$，$n_f(1000)=500$，由式(4-2)得

$$R(t)=\frac{n-n_f(t)}{n}$$

$$R(500)=\frac{1000-100}{1000}=0.9,\ R(1000)=\frac{1000-500}{1000}=0.5$$

4.1.3 失效率 $\lambda(t)$

失效率又称为故障率，表示当产品已工作到 t 时刻的条件下，在下阶段 $\Delta t+t$ 的单位时间内发生失效的条件概率(取 $\Delta t\to 0$ 的极限值)。其数学表达式为

$$\lambda(t)=\lim_{\Delta t\to 0}\frac{P(t<T\geqslant t+\Delta t\mid T>t)}{\Delta t} \tag{4-11}$$

失效率 $\lambda(t)$的统计观测值为

$$\bar{\lambda}(t)=\frac{n_f(t+\Delta t)-n_f(t)}{[n-n_f(t)]\Delta t}=\frac{\Delta n_f(t)}{n_s(t)\Delta t} \tag{4-12}$$

因此式(4-11)可改写为

$$\lambda(t)=\lim_{\substack{n\to\infty\\ \Delta t\to 0}}\frac{n_f(t+\Delta t)-n_f(t)}{[n-n_f(t)]\Delta t}$$

即

$$\lambda(t)=\frac{1}{n-n_f(t)}\frac{\mathrm{d}n(t)}{\mathrm{d}t}=\frac{1}{\frac{n_s(t)}{n}}\frac{\frac{\mathrm{d}n(t)}{n}}{\mathrm{d}t}=\frac{1}{R(t)}\frac{\mathrm{d}F(t)}{\mathrm{d}t}=\frac{f(t)}{R(t)} \tag{4-13}$$

或

$$\lambda(t)=\frac{f(t)}{R(t)}=-\frac{1}{R(t)}\frac{\mathrm{d}R(t)}{\mathrm{d}t}$$

将该式从 0 到 t 进行积分，则得

$$\int_0^t \lambda(t)\mathrm{d}t = -\ln R(t)$$

于是得

$$R(t) = \mathrm{e}^{-\int_0^t \lambda(t)\mathrm{d}t} \tag{4-14}$$

上式称为可靠度函数 $R(t)$的一般方程，当 $\lambda(t)$为恒定值时，就是我们常用到的指数分布可靠度函数表达式。

以上所述表明，产品的可靠性指标：$R(t)$、$F(t)$、$f(t)$、$\lambda(t)$都是相互联系的，已知其中之一，便可以推算出其余三个指标。

特别指出，$R(t)$和 $F(t)$均为无量纲值，以小数或百分数(%)表示，而 $f(t)$和 $\lambda(t)$均为有量纲值(1/h)，常用的失效率 $\lambda(t)$单位还有 $1/(10^3\mathrm{h})$、$1/(10^6\mathrm{h})$。例如，某型号滚动轴承失效率 $\lambda=0.05/(10^3\mathrm{h})=5\times10^{-5}/\mathrm{h}$，表示 10^5个轴承中每小时有 5 个失效，它反映了轴承失效的变化速度。

4.1.4　平均寿命 m

在讨论产品的可靠性时，人们总要把它和产品的寿命联系起来。平均寿命指的是一批类型、规格相同的产品从投入运行到发生失效(或故障)的平均工作时间。由于产品投入运行后出现失效的时间(或寿命 T)是个随机变量，具有确定的统计分布规律，因此，平均寿命实际上是这个随机变量 T 的数学期望 $E(T)$。

设有 n 个产品从开始使用到发生失效的时间为 t_1、t_2、…、t_n，则平均寿命的观测值为

$$m = \frac{1}{n}\sum_{i=1}^{n} t_i \tag{4-15}$$

若产品的失效密度为 $f(t)$，则产品的平均寿命，即数学期望为

$$m = E(T) = \int_0^\infty t f(t)\mathrm{d}t \tag{4-16}$$

平均寿命 m，对于不可修复的产品是指从开始使用到发生失效的平均时间，用 MTTF (Mean Time To Failure)表示；对可修复的产品是相邻两次故障间工作时间的平均值，用 MTBF(Mean Time Between Failures)表示。若只考虑首次故障，则指的是产品从开始使用到第一次发生故障的平均时间，用 MTTFF(Mean Time To First Failure)表示。对可修复产品，人们不仅关心 MTBF，有时则更关心 MTTFF。

平均寿命 m 也可通俗地表达为

$$m = \frac{\text{所有产品的总工作时间}}{\text{总故障数}} \tag{4-17}$$

将式(4-10)代入式(4-16)得平均寿命 m 与可靠度 $R(t)$有如下关系：

$$m = \int_0^\infty t\left[-\frac{\mathrm{d}R(t)}{\mathrm{d}t}\right]\mathrm{d}t = -\int_0^\infty t\mathrm{d}R(t)$$

用部分积分法可解得

$$m = \int_0^\infty R(t)\mathrm{d}t \tag{4-18}$$

上式表明，平均寿命 m 的几何意义是：可靠度 $R(t)$曲线与时间轴所夹的面积。

4.2 产品的失效率曲线

4.2.1 电子产品的失效率曲线

电子产品的失效率曲线如图 4-2 所示。这个曲线常被形象地称为浴盆曲线，该曲线分为三段。

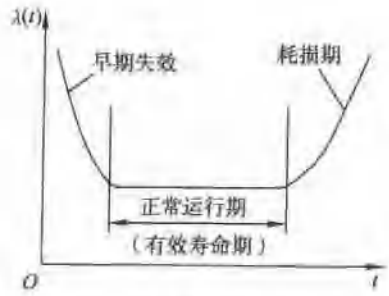


图 4-2 电子产品的失效率曲线

1. 早期失效期

早期失效期一般为产品试车跑合阶段。在这一阶段中，失效率由开始很高的数值急剧地降到某一稳定的数值。引起这一阶段失效率特别高的原因是由于材料缺陷、制造工艺缺陷、验差错等引起。因此，为了提高可靠性，产品在出厂前应进行严格的测试，查找失效原因，并采取各种措施发现隐患和纠正缺陷，使失效率下降且逐渐趋于稳定。

2. 正常运行期

正常运行期又称为有效寿命期。在此阶段内如果发生失效，一般都是由于偶然的原因而引起的，因此这一阶段也称为偶然失效期。其失效的特点是随机的。例如，个别产品由于使用过程中工作条件发生不可预测的突然变化而导致失效。这个时期的失效率低且稳定，近似为常数，是产品最佳状态时期。产品、系统的可靠度通常以这一时期为代表，通过提高可靠性设计质量、改进设备使用管理、加强监视诊断和维护保养等工作，使产品的失效率降低到最低水平，延长产品的使用寿命。

3. 耗损失效期

耗损失效期出现在产品使用的后期。其特点是失效率随工作时间的增加而上升。耗损失效主要是产品经长期使用后，由于疲劳、磨损、老化等原因，已渐近衰竭，从而处于频发失效状态，使失效率随时间推移而上升，最终会导致产品的功能终止。改进耗损失效的办法是不断提高产品的工作寿命，对整机应制定一套预防维修和更新措施。在到达耗损失效期前就及时予以维修或更换某些易损件，这样就可以把上升的失效率拉下来，也就是说，采取这种办法可延长产品的使用寿命。

上述典型的电子产品失效率曲线变化的三个阶段，宛如人从幼年经壮年而进入老年的寿命过程一样。对于人类来说，应尽力增强体质，减少病伤事故，使死亡率（失效率）尽可能降低，以延长寿命。对产品而言，也是如此。

4.2.2 机械零件的失效率曲线

图 4-3 是典型的机械零件失效率曲线。由图可见，机械零件的失效率曲线不同于电子产品。一般情况下，它没有电子产品那么长的有效寿命期，而且失效率不等于常数。因为，机械零件主要失效形式如疲劳、磨损、腐蚀及蠕

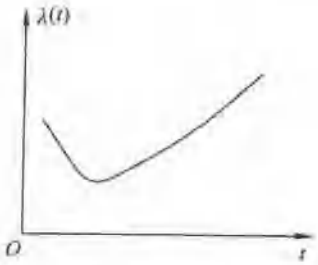


图 4-3 典型的机械零件失效率曲线

变等，都属于典型的损伤累积失效，而且影响失效的偶然因素很复杂，所以，随着时间的推移，失效率是递增的。在调试期或运行的初期，少数零件由于材料的严重缺陷，或者在制造工艺过程中（如铸造、焊接、热处理等）造成内部缺陷，使得少数零件一旦承受载荷就很快失效，因而出现一定的失效率，但和电子元件相比，它要小很多。随后，零件进入正常使用期，但由于损伤不断积累，所以失效率不断增大。

4.3　可靠性常用分布函数

4.3.1　离散型随机变量的分布

1. 二项分布

二项分布适用于描述只有两种状态的事物，如检验产品是合格还是不合格、判定一个产品或系统是正常工作还是失效等。

设有 A 和 B 两台设备。其正常工作的概率（可靠度）和失效概率分别以 $R(A)$、$R(B)$和 $F(A)$、$F(B)$表示。对每台设备而言，只有正常工作和失效两种状态。对这两台设备，只有四种可能性：① 两台同时无失效，概率为 $R(A)R(B)$；② A 失效 B 不失效，概率为 $F(A)R(B)$；③ B 失效 A 不失效，概率为 $R(A)F(B)$；④ 两台同时失效，概率为 $F(A)F(B)$。

以上所有可能的概率之和等于 1，即

$$R(A)R(B)+F(A)R(B)+R(A)F(B)+F(A)F(B)=1$$

若设备 A 和 B 具有相同的可靠度和相同的失效概率时，即 $R(A)=R(B)=R$，$F(A)=F(B)=F$，则上式成为

$$R^2+2FR+F^2=1$$

即

$$(R+F)^2=1$$

同理，可推出三台设备的情况为

$$(R+F)^3=1$$

具有相同可靠度的 N 台设备的情况为

$$(R+F)^N=1 \tag{4-19}$$

将上式展开，得

$$R^N+NR^{N-1}F+\frac{N!}{(N-2)!\ 2!}R^{N-2}F^2+\cdots+$$

$$\frac{N!}{r!\ (N-r)!}R^{N-r}F^r+\cdots+F^N=1 \tag{4-20}$$

式中：R^N——无失效概率；

$NR^{N-1}F$——只有一台设备失效的概率；

$\dfrac{N!}{(N-r)!\ r!}R^{N-r}F^r$——有 r 台失效的概率；

F^N——N 台设备都失效的概率，上式称为二项式分布。

设某一系统由 N 个相同元件组成，每个元件可靠度为 R，失效概率为 $F=1-R$。如果系统中全部元件均不失效系统才能正常工作，则上式左端第一项便是系统的可靠度。如允

许一个元件失效，即只要不发生失效或只有一个失效，系统便是成功的，则系统的可靠度为上式左端前两项之和。若允许 r 个失败，则系统可靠度为前 $r+1$ 项之和。由此，得二项分布的可靠度函数表达式为

$$R(r)=\sum_{i=1}^{r}C_N^iR^{N-i}F^i \tag{4-21}$$

其中任意一项，记为 $P(i)=C_N^iR^{N-i}F^i$，可以看做有 N 个元件投入试验时出现 i 个失效的概率；$C_N^i=\dfrac{N!}{i!\ (N-i)!}$，为 N 个中取 i 个的组合数。

例 4-2 某控制系统中有 12 个相互串联的指示灯，使用中发现，平均每月要换一个灯泡才能使用，问需要储备多少个备用灯泡才能保证指示灯总是亮的？

解：根据题意，系统中任一灯泡失效的概率为 $F=1/12=0.083$，无失效的概率为 $R=1-F=0.917$，可由二项分布计算，即

$$(R+F)^N=(0.917+0.083)^{12}$$

由式(4-21)可算得，当允许同时失效的灯泡个数 r 分别为 0、1、2、3 时，即储备的备用灯泡个数分别为 0、1、2、3 时，指示灯总发亮的可靠度分别为 0.354、0.739、0.930、0.987。由此可见，若备用三个灯泡，指示灯总是发亮的可靠度已达 0.987，已是足够可靠了。

2. 泊松分布

在可靠性工程中，常会遇到试验次数 N 很大，每次试验失效概率 F 很小，而 NF 为常数的情况。这时若仍然用二项分布进行计算是非常复杂的，采用泊松分布可使计算得以简化。

当 N 很大，F 很小时，N 次试验出现 r 次失效的概率[即二项分布 $P(r)=C_N^rR^{N-r}F^r$]将接近一个极限，这个极限称为泊松分布，其表达式为

$$P(r)=\frac{(NF)^r}{r!}e^{-NF} \tag{4-22}$$

式中，$NF=\lambda t$，为平均失效数，这里 λ 为失效率，t 为时间，因此乘积 λt 为在时间 t 内发生的平均失效数。在很多情况下，不知道 N 和 F 的数值，却可能知道 λ 和 t 的值，即可进行计算。

如果 N 次试验均不出现失效，即 $r=0$，其可靠度为

$$R(0)=P(0)=\frac{(NF)^{(0)}}{0!}e^{-NF}=e^{-NF}$$

如果 N 次试验只出现一次失效算合格，这时的可靠度为失效次数 $r=0$ 和 $r=1$ 时的概率之和，即

$$R(1)=P(0)+P(1)=e^{-NF}+(NF)e^{-NF}=(1+NF)e^{-NF}$$

表 4-1

N 次试验允许的失效数 r	按式(4-21)计算的可靠度 $R(r)$				按近似公式(4-23)计算的 $R(r)$
	$N=10$ $F=0.1$	$N=20$ $F=0.05$	$N=30$ $F=0.025$	$N=40$ $F=0.01$	
0	0.349	0.358	0.363	0.366	0.368
1	0.734	0.735	0.736	0.736	0.736
2	0.928	0.924	0.922	0.921	0.920
3	0.985	0.984	0.983	0.982	0.981
4	0.996	0.997	0.997	0.997	0.996

依次类推，当 N 次试验允许 r 次失效时，其可靠度函数为

$$R(r)=\sum_{i=1}^{r}\frac{(NF)^i}{i!}e^{-NF} \tag{4-23}$$

泊松分布是二项分布的近似表达式。表 4-1 给出当 $\lambda=NF=1$ 为常数时，该泊松分布近似的可靠度计算公式 (4-23)和对不同的 N 与 F 的取值按二项分布可靠度公式(4-21)直接计算所得结果的比较。从中可直观地看出式 (4-23)的近似程度。在实际计算中，当 $N\geqslant 10$，$F\leqslant 0.1$ 时，就可用式(4-23)作为式(4-21)的近似。随着 N 趋大和 F 趋小，近似程度将趋好。

例 4-3　某种零件的失效率 λ 为常数，其平均值为 0.010/(1000h)，现仅有两个备件，且半年内不能再进备件，实际工作需保证设备运转 50000h，问这种情况设备不能正常工作的概率为多大？

解：在 50000h 内的平均失效零件数为

$$\lambda t=NF=0.010/1000\times 50000=0.5$$

由式(4-22)，出现三个零件失效的概率为

$$P(r)=P(3)=\frac{(0.5)^3}{3!}e^{-0.5}=0.0126$$

同理，出现四个零件失效的概率为

$$P(4)=\frac{(0.5)^4}{4!}e^{-0.5}=0.00157$$

由此可见，出现五个零件失效的概率已非常小，可不必再计算下去。因此，设备不能工作的概率为

$$P(r\geqslant 3)=P(3)+P(4)+P(5)+\cdots\approx 0.0126+0.00157=0.01417$$

保证设备正常工作的可靠性为

$$R(2)=P(0)+P(1)+P(2)=1-P(r\geqslant 3)\approx 1-0.0141=0.9859$$

4.3.2　连续型随机变量的分布

前面所讨论的离散型分布，针对的是成败型产品。以下所讨论的连续型分布，所描述的是某一产品的某种特性，尽管它在某一极小范围内取值的概率非常小，但仍是可能出现的。

1. 指数分布

指数分布是可靠性工程中最常用的分布类型之一。当产品工作进入浴盆曲线(图 4-2)的偶然失效期后，失效率 λ 接近为常数，此时，可靠度函数 $R(t)$、失效概率函数 $F(t)$、失效概率密度函数 $f(t)$都是指数分布。由式 (4-13)、式(4-14)有

$$\begin{cases}R(t)=\exp\left[-\int_0^t\lambda(t)\mathrm{d}t\right]=e^{-\lambda t}\\ f(t)=\lambda e^{-\lambda t}\\ F(t)=1-R(t)=1-e^{-\lambda t}\end{cases} \tag{4-24}$$

由式 (4-18)，此时产品平均寿命 m 为

$$m=\int_0^\infty R(t)\mathrm{d}t=\int_0^\infty e^{-\lambda t}\mathrm{d}t=\frac{1}{\lambda}\int_0^\infty e^{-\lambda t}\mathrm{d}(-\lambda t)=\frac{1}{\lambda} \tag{4-25}$$

由此可见，当随机变量为指数分布时，平均寿命 m 与失效率 λ 互为倒数。当产品的工

作时间 $t=m=\frac{1}{\lambda}$ 时，$R(t)=e^{-1}=0.37$。这表明，对于失效规律服从指数分布的一批产品而言，能够工作到平均寿命的仅占 37%左右，约有 63%的产品将在达到平均寿命前失效。因此，要提高产品的可靠性，必须在远小于平均寿命的时间内工作；同时必须提高产品的平均寿命，产品才有实际应用价值。

例 4-4 某仪器的寿命 T 服从指数分布，其平均无故障连续工作时间 MTBF 为 25h，试求其失效率为多少？若要求有 90%不出故障的可靠性，问应如何选择连续工作时间？

解：对于指数分布，其失效率为

$$\lambda=\frac{1}{m}=\frac{1}{25}=0.04/\mathrm{h}$$

因要求 90%把握不出故障，则有 $R(t)=e^{-\lambda t}=90\%$，由此解出

$$t=-\ln(0.9)/\lambda=2.634\mathrm{h}$$

即为了有 90%的把握不出故障，该仪器连续工作时间应不超过 2.634h。

2. *正态分布*

(1) 正态分布的定义。正态分布(也称高斯分布)是数理统计中的经典分布，应用范围极广，几乎渗透到每一个工程技术领域。

设连续型随机变量的分布密度函数为

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right],-\infty<x+\infty \tag{4-26}$$

于是，失效概率 $F(x)$ 和可靠度 $R(x)$ 分别为

$$F(x)=\int_{-\infty}^{x}\frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]\mathrm{d}x \tag{4-27}$$

$$R(x)=1-F(x)=\int_{x}^{+\infty}\frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]\mathrm{d}x \tag{4-28}$$

可见正态分布是一个双参数连续分布，记为 $N(\mu,\sigma)$。参数 μ 称为随机变量的数学期望(均值)；参数 σ 称为随机变量的标准差(均方差)，是对测试数据分散性的度量。分布参数 μ 和 σ 的估计量可按下列公式计算：

$$\mu\leftarrow\bar{x}=\frac{1}{N}\sum_{i=1}^{n}x_i$$

$$\sigma=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})^2}$$

式中：x_i——第 i 次测试值；

N——总测试次数；

$\bar{x}$——算术平均值，当 N 足够大时，$\bar{x}$ 将稳定到一个确定的值 μ，即变量的数学期望。若变量代表产品的寿命，则寿命服从正态分布的产品，其平均寿命 $m=\mu$。

(2) 正态曲线的特征。如图 4-4 所示，正态分布的分布密度 $f(x)$ 是以均值 μ 为中心的对称曲线。在 $x=\mu$ 处达到最大值，等于 $\frac{1}{\sigma\sqrt{2\pi}}$；在 $x=\mu\pm\sigma$ 处有拐点；当 $x\to\pm\infty$ 时，曲线以 x 轴为其渐近线。变量落入以 μ 为中心、$\pm\sigma$ 区间的概率为 68.27%；落入 $\pm2\sigma$ 区间的概率为 95.45%；落入 $\pm3\sigma$ 区间的概率为 99.73%；落入 $\pm3\sigma$ 之外的概率已很小，为 0.27%，在一

般工程计算中可忽略不计。

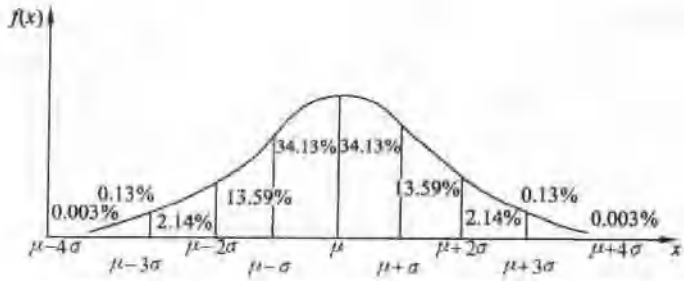


图 4-4　在标准离差的整数倍之间正态分布的百分数

正态分布的均值 μ 确定了曲线的位置，标准差 σ 确定了曲线的形状，如图 4-5 所示。不论 μ 和 σ 取值如何，$f(x)$ 曲线与横坐标围成的面积恒等于 1。

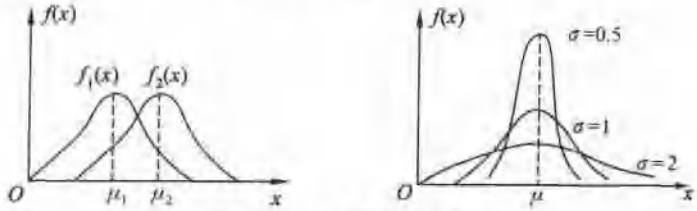


图 4-5　正态分布曲线

(3) 标准正态分布。正态分布的累积分布函数 $F(x)$ 或可靠度 $R(x)$ 的计算比较复杂，为了便于数学处理和制出统一的标准正态分布表，可将正态分布规一化。这只需将随机变量 x 做如下变换：

$$z=\frac{x-\mu}{\sigma} \tag{4-29}$$

则式 (4-26)，式 (4-27) 成为

$$f(z)=\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}z^2\right),\quad -\infty<z<+\infty$$

$$F(z)=\int_{-\infty}^{z}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}z^2\right)\mathrm{d}z \tag{4-30}$$

这种 $\mu=0$，$\sigma=1$ 时的正态分布，称为标准正态分布，记为 $N(0,1)$，此时的累积分布函数 $F(z)$ 一般记为 $\phi(z)$，它与 z 值的对应数值，可由标准正态分布表 4-2 查出，简化了计算。如图 4-6 所示，表 4-2 给出的是

$$\phi(-z)=P(x<-z)=\frac{1}{2\pi}\int_{-\infty}^{-z}\mathrm{e}^{-\frac{z^2}{2}}\mathrm{d}z$$

的值。表中所列出的 z 的范围为 0～5，即 $(-z)$ 为负值。当 $(-z)$ 为正值时，其累积分布函数的值可按下式计算：

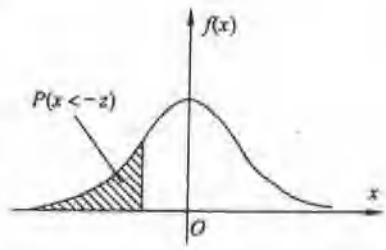


图 4-6　标准状态分布

$$\phi(-z)=1-(\text{表 4-2 中所列的相应数值})$$

许多自然现象和产生的许多特性大都服从正态分布。例如，测量误差、工艺误差、加工误差、射击误差、材料特性、应力分布等都十分接近于正态分布。因此正态分布广泛应用于机械与结构的可靠性计算和寿命分析。此外，产品在使用后期的磨损失效一般具有均匀耗损性质，通常可用正态分布来描述。

例 4-5 某机器生产的螺栓长度服从正态分布 $N(10.05,0.06)$，规定长度在 $10.05^{+0.12}_{-0.06}$ 范围内为合格品，求一个螺栓为合格品的可靠度。

解：由已知得 $\mu=10.05$，$\sigma=0.06$，合格螺栓的长度尺寸上限为 $x_u=10.17$，下限为 $x_d=9.99$，由式(4-29)得

$$z_u=\frac{x_u-\mu}{\sigma}=\frac{10.17-10.05}{0.06}=2$$

$$z_d=\frac{x_d-\mu}{\sigma}=\frac{9.99-10.05}{0.06}=-1$$

由标准正态分布表 4-2 查得

$$\phi(z_d)=\phi(-1)=0.1587$$

$$\phi(z_m)=\phi(2)=1-\phi(-2)=1-0.0228=0.9772$$

故该机器生产的任一螺栓为合格品的可靠度为

$$R=\phi(z_u)-\phi(z_d)=0.9772-0.1587=0.8185$$

表 4-2 标准正态分布表(对应于图 4-6)

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641
0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4401	0.4364	0.4325	0.4286	0.4247
0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
0.7	0.2420	0.2389	0.2358	0.2327	0.2297	0.2266	0.2236	0.2206	0.2177	0.2148
0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.09853
1.3	0.09680	0.09510	0.09342	0.09176	0.09012	0.08851	0.08691	0.08534	0.08379	0.08226
1.4	0.08076	0.07927	0.07780	0.07636	0.07493	0.07353	0.07215	0.07078	0.06944	0.06811
1.5	0.06681	0.06552	0.06426	0.06301	0.06178	0.06057	0.05938	0.05821	0.05705	0.05592
1.6	0.05480	0.05370	0.05262	0.05155	0.05050	0.04947	0.04846	0.04746	0.04648	0.04551
1.7	0.04457	0.04363	0.04272	0.04182	0.04093	0.04006	0.03920	0.03836	0.03754	0.03673
1.8	0.03593	0.03515	0.03438	0.03362	0.03288	0.03216	0.03144	0.03074	0.03005	0.02938
1.9	0.02872	0.02807	0.02743	0.02680	0.02619	0.02559	0.02500	0.02442	0.02385	0.02330

续表

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
2.0	0.02275	0.02222	0.02169	0.02118	0.02068	0.02018	0.01970	0.01923	0.01876	0.01831
2.1	0.01786	0.01743	0.01700	0.01659	0.01618	0.01578	0.01539	0.01500	0.01463	0.01426
2.2	0.01390	0.01355	0.01321	0.01287	0.01255	0.01222	0.01191	0.01160	0.01130	0.01100
2.3	0.01072	0.01044	0.01017	$0.0^{2}9903$	$0.0^{2}9642$	$0.0^{2}9387$	$0.0^{2}9117$	$0.0^{2}8894$	$0.0^{2}8656$	$0.0^{2}8424$
2.4	$0.0^{2}8219$	$0.0^{2}7976$	$0.0^{2}7800$	$0.0^{2}7549$	$0.0^{2}7344$	$0.0^{2}7143$	$0.0^{2}6947$	$0.0^{2}6756$	$0.0^{2}6569$	$0.0^{2}6387$
2.5	$0.0^{2}6210$	$0.0^{2}6037$	$0.0^{2}5857$	$0.0^{2}5703$	$0.0^{2}5543$	$0.0^{2}5386$	$0.0^{2}5234$	$0.0^{2}5085$	$0.0^{2}4940$	$0.0^{2}4799$
2.6	$0.0^{2}4661$	$0.0^{2}4527$	$0.0^{2}4396$	$0.0^{2}4269$	$0.0^{2}4145$	$0.0^{2}4025$	$0.0^{2}3907$	$0.0^{2}3793$	$0.0^{2}3681$	$0.0^{2}3573$
2.7	$0.0^{2}3467$	$0.0^{2}3364$	$0.0^{2}3264$	$0.0^{2}3167$	$0.0^{2}3072$	$0.0^{2}2980$	$0.0^{2}2890$	$0.0^{2}2803$	$0.0^{2}2718$	$0.0^{2}2635$
2.8	$0.0^{2}2555$	$0.0^{2}2477$	$0.0^{2}2401$	$0.0^{2}2327$	$0.0^{2}2256$	$0.0^{2}2186$	$0.0^{2}2118$	$0.0^{2}2052$	$0.0^{2}1988$	$0.0^{2}1926$
2.9	$0.0^{2}1866$	$0.0^{2}1807$	$0.0^{2}1750$	$0.0^{2}1695$	$0.0^{2}1641$	$0.0^{2}1589$	$0.0^{2}1538$	$0.0^{2}1489$	$0.0^{2}1441$	$0.0^{2}1359$
z	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
3.0	$0.0^{3}1350$	$0.0^{3}9676$	$0.0^{3}6871$	$0.0^{3}4834$	$0.0^{3}3369$	$0.0^{3}2326$	$0.0^{3}1591$	$0.0^{3}1078$	$0.0^{4}7235$	$0.0^{4}4810$
4.0	$0.0^{4}3167$	$0.0^{4}2066$	$0.0^{4}1335$	$0.0^{5}8540$	$0.0^{5}5413$	$0.0^{5}3398$	$0.0^{5}2112$	$0.0^{5}1301$	$0.0^{5}7933$	$0.0^{6}4792$

3. 对数正态分布

若随机变量 x 本身并不服从状态分布，但对其取对数后，$\ln x$ 服从正态分布，则称 x 为服从对数正态分布的随机变量。对数正态分布常用来描述某些非负的随机变量（如寿命）的分布规律。

对数正态分布中含有两个参数 μ 和 σ，其概率密度函数 $f(x)$ 呈“不对称的山包型”，图 4-7 所示为对数均值 $\mu=1$、对数标准差 σ 取不同值时对数正态分布的密度函数和失效率函数的曲线。与正态分布情况相类似，服从对数正态分布的产品，在“山顶”附近失效密度较高。

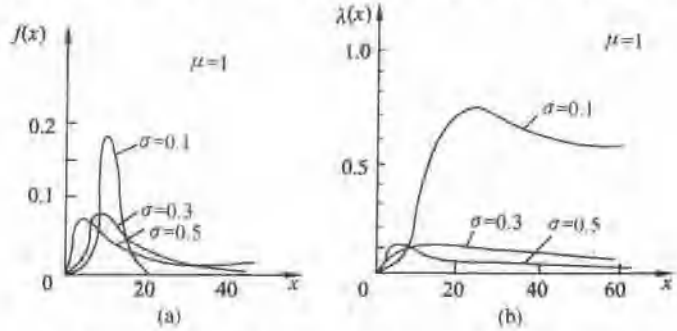


图 4-7　对数正态分布的密度函数和失效率函数曲线

对数正态分布的计算方法与正态分布相同，只要将随机变量 x 变换成 $\ln x$ 即可。对数正态分布的分布密度 $f(x)$ 和分布函数 $F(x)$ 分别为

$$f(x)=\frac{1}{x\sigma\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{\ln x-\mu}{\sigma}\right)^{2}\right],\ x\geqslant 0 \tag{4-31}$$

$$F(x)=\int_{0}^{x}\frac{1}{x\sigma\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{\ln x-\mu}{\sigma}\right)^{2}\right]\mathrm{d}x \tag{4-32}$$

对数正态分布通过变换

$$z=\frac{\ln x-\mu}{\sigma} \tag{4-33}$$

可将其化为标准正态分布，由式(4-33)，有 $dx=x\sigma dz$，代入式(4-32)，则有

$$F(x)=\int_0^x \frac{1}{x\sigma\sqrt{2x}}e^{\frac{(\ln x-\mu)^2}{2\sigma^2}}dx=\int_{-\infty}^{\frac{\ln x-\mu}{\sigma}} \frac{1}{\sqrt{2x}}e^{\frac{z^2}{2}}dz=\phi\left(\frac{\ln x-\mu}{\sigma}\right)=\phi(z) \tag{4-34}$$

于是便可利用标准正态分布表查值了。

对数正态分布实际上是一种偏态分布，能较好地符合一般零部件失效过程的时间分布。在机械零部件的疲劳寿命、疲劳强度、耐磨寿命研究中，大量应用了对数正态分布。电子元器件中有很多产品接近对数正态分布，如半导体器件的寿命分布就属于这种类型。

例 4-6 某产品的寿命(h)服从 $\mu=5,\sigma=1$ 的对数正态分布，求 $t=150$h 时的可靠度。

解：寿命为 T，则 $\ln T$ 服从正态分布 $N(5,1)$。由式(4-33)，对数正态分布的标准正态变量为 $z=\frac{\ln t-\mu}{\sigma}$。

当 $t=150$h 时，有

$$z=\frac{\ln 150-5}{1}=\frac{5.01-5}{1}=0.01$$

查标准正态分布表得

$$\phi(z)=\phi(0.01)=0.504$$

即 $t=150$h 时的累积失效概率 F 为 0.504，所以，其可靠度为

$$R(150)=1-\phi(z)=1-\phi(0.01)=1-0.504=0.496$$

4. *威布尔分布*

威布尔分布是瑞典人 Weibull 构造的一种分布函数。凡属于局部失效(如某一最薄弱环节失效)而导致整体机能失效的模型(串联模型)，一般都能采用这种分布函数来描述，因此在可靠性工程中应用十分广泛。

威布尔分布的分布函数 $F(t)$ 和概率密度函数 $f(t)=F'(t)$ 分别为

$$F(t)=1-e^{-\frac{(t-r)^m}{a}},\quad t\geqslant r \tag{4-35}$$

$$f(t)=\frac{m}{a}(t-r)^{m-1}e^{-\frac{(t-r)^m}{a}},\quad t\geqslant r \tag{4-36}$$

当 $t<r$，$f(t)=F(t)=0$。可见威布尔分布是具有三个参数的连续分布。其中，m 称为形状参数，r 称为位置参数，a 称为尺度参数。下面分别讨论这三个参数的影响。

(1) 形状参数 m。形状参数 m 决定了分布密度曲线 $f(t)$ 的形状。例如，m 取不同的值，便构成了不同的图形(为明显起见，取 $a=1,r=0$)，如图 4-8 所示。此时的失效率曲线如图 4-9 所示。

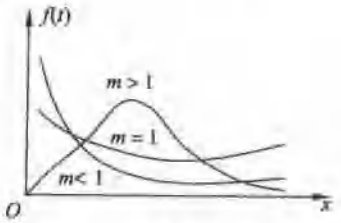


图 4-8 m 参数的几何形状

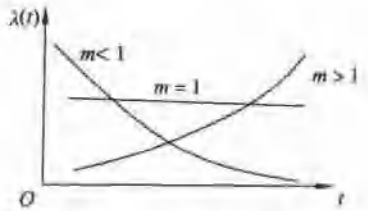


图 4-9 m 取不同值时的失效率曲线

由图 4-9 可见，$m<1$ 时，曲线呈下降型，反映了失效率随时间递减的情况，类似于早期失效(图 4-2)，故可用于描述早期失效过程；当 $m=1$ 时，$\lambda(t)$ 曲线呈常数型，此时威布尔分布变为指数分布，可用于描述偶然失效过程；当 $m>1$ 时，$\lambda(t)$ 曲线呈上升型，与耗损失效相符。由图 4-8 还可看出，随着 m 的增大，$f(x)$ 曲线逐渐趋于对称，当 $m=3.5$ 时，已极为接近正态分布密度曲线。由此可见，当根据试验数据求得 m 值后，就可大致判断该产品的失效类型。

(2) 位置参数 r。位置参数 r 决定分布密度曲线 $f(t)$ 的位置，而不影响 $f(t)$ 的形状。在 m 和 a 不变情况下，不同的 r 相当于把曲线沿横坐标作整体移动：当 $r>0$ 时，曲线由 $r=0$ 时的位置向右平移，移动距离为 r；当 $r<0$ 时，曲线则向左平移，移动距离为 $|r|$。图 4-10 给出了取定 $a=1$，$m=2$，而 r 取不同值时，曲线 $f(t)$ 的变化情况。r 为负值时，表示某些产品在开始工作($t=0$)前就已失效，即在存储期间失效；r 为正时，表示存在一段不失效的时间；$r=0$ 时，表示使用前是好的，失效的可能性随开始使用而产生。显然，大部分产品具有这种属性。

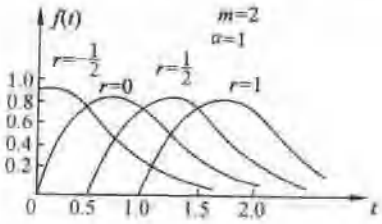


图 4-10　位置参数变化

当 $r=0$ 时，三参数威布尔分布退化为两参数的分布，这时，其失效密度函数 $f(t)$、失效概率函数 $F(t)$ 和失效率函数 $\lambda(t)$ 分别为

$$f(t)=\frac{m}{a}t^{m-1}e^{-\frac{t^m}{a}} \tag{4-37}$$

$$F(t)=1-e^{-\frac{t^m}{a}} \tag{4-38}$$

$$\lambda(t)=\frac{m}{a}t^{m-1} \tag{4-39}$$

(3) 尺度参数 a。尺度参数 a 变化会放大或缩小概率密度函数 $f(t)$ 的横坐标 t 的标尺，即会引起 $f(t)$ 曲线在横轴方向伸长或压缩，而分布的形状仍是类似的，如图 4-11 所示。

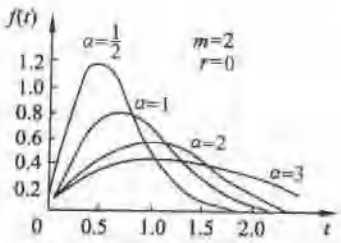


图 4-11　尺度参数变化

如令 $\eta=a^{\frac{1}{m}}$，则 η 称为真尺度参数，或称特征寿命。威布尔分布的平均寿命为

$$\begin{aligned}\text{MTBF(或 MTTFF)}&=\int_0^{+\infty}tf(t)\mathrm{d}t\\&=a^{\frac{1}{m}}\Gamma\left(1+\frac{1}{m}\right)\\&=\eta\Gamma\left(1+\frac{1}{m}\right)\end{aligned} \tag{4-40}$$

式中

$$\Gamma(x)=\int_0^{+\infty}t^{x-1}e^{-t}\mathrm{d}t$$

由于威布尔分布具有前述这样一些特点，调整某个参数的值，可方便地改变分布的形状，从而适用于不同的分布状态，因此它是应用最为灵活的一种经验分布函数，具有普遍意

义并得到广泛应用。对于滚珠轴承、电子管、超高频器件、机械零件、结构部件等许多产品的寿命分布规律以及金属材料的疲劳寿命，都可用威布尔分布函数进行描述。

例 4-7 某种飞机部件的寿命服从威布尔分布。已知 $m=2, \alpha=40000\text{h}, r=0$。试计算此部件的平均寿命、可靠度为 $R=95\%$ 时的寿命和 100h 之内的最大失效率。

解：由 Γ 函数表，查得 $\Gamma\left(1+\frac{1}{2}\right)=0.886$。由式(4-40)得

$$\text{平均寿命}=\alpha^{\frac{1}{m}}\Gamma\left(1+\frac{1}{m}\right)=(40000)^{\frac{1}{2}}\times 0.886=177.2(\text{h})$$

由 $R=\mathrm{e}^{-\frac{(t-r)^m}{\alpha}}=0.95$，可算出可靠度为 95%时的寿命为

$$t=r+\alpha^{\frac{1}{m}}\left(\ln\frac{1}{R}\right)^{\frac{1}{m}}=0+40000^{\frac{1}{2}}\left(\ln\frac{1}{0.95}\right)^{\frac{1}{2}}=45.3(\text{h})$$

100h 之内的最大失效率为

$$\lambda(100)=\frac{mt^{m-1}}{\alpha}=\frac{2\times 100}{40000}=0.5\%$$

4.4 可靠性设计原理

4.4.1 概率设计的基本概念

在常规的机械设计中，通常采用安全系数法或许用应力法。其出发点是使作用在危险截面上的工作应力 s 小于等于许用应力$[s]$，而许用应力$[s]$是由极限应力 $s_{\lim}$除以大于 1 的安全系数 n 而得到的；也可以使机械零件的计算系数 n 大于预期或许用安全系数$[n]$，即

$$s\leqslant[s]=\frac{s_{\lim}}{n}$$

$$n=\frac{s_{\lim}}{s}\geqslant[n] \tag{4-41}$$

这种常规设计方法沿用了许多年，只要安全系数选用适当，便是一种可行的设计方法。但是，随着产品日趋复杂，对可靠性要求愈来愈高，常规方法就显得不够完善。首先，大量的实验表明，现实的设计变量如负荷、极限应力以及材料硬度、尺寸等大都是随机变量，都呈现或大或小的离散性，都应该依概率取值。不考虑这一点，设计出来的结果难免与实际脱节。其次，常规设计方法的关键是选取安全系数。安全系数过大，造成浪费；过小，影响正常使用。但在选取安全系数时，常常没有确切的选择尺度，其结果是使设计极易受局部经验的影响。实际上，不考虑变量离散性的安全系数是不能正确反映设计的安全裕度的。许多时候，安全系数大，未必可靠；反之，也不一定危险。表 4-3 列出了不同情况下安全系数和可靠度的比较。表中 r 表示强度，相当于承载能力，s 表示承受的工作应力。作为随机变量的 r、s 有它本身的均值 μ_r、μ_s 和标准差 σ_r、σ_s。μ_n 表示安全系数，$\mu_n=\frac{\mu_r}{\mu_s}$。可靠度 R 是按正态分布计算的。由表 4-1 可以看出，只要强度 r 和应力 s 的均值保持相同比值，平均安全系数就不会改变，但当标准差不同时，可靠度就有较大区别。表中 A 与 B 的安全系数均为 1.5，但 B 的标准差比 A 的小，可靠度由 0.7823 提高到 0.9998。C、D 的安全系数均达 3，但 C 的标准

差大，可靠度只有0.9406。因为当σ_r较大时，选用的材料强度处于低限的机会就比较多，σ_s较大时，所受的应力处于高限的机会也大，可靠度必然降低。因此，使设计更符合实际，应该在常规设计的基础上进行概率设计。

表4-3　几种情况下μ_n和R的比较

序　号	强度均值 μ_r	强度标准差 σ_r	应力均值 μ_s	应力标准差 σ_s	平均安全系数 μ_n	可靠度 R
A	300	100	200	80	1.5	0.7823
B	300	20	200	20	1.5	0.9998
C	300	100	100	80	3	0.9406
D	300	20	100	20	3	1.0

注：μ_r、σ_r和μ_s、σ_s的单位为N/mm^2。

概率设计基本观点如下：

(1) 认为零件的强度是服从于概率密度为$f_r(r)$的随机变量，加在零件上的应力占服从概率密度为$f_s(s)$的随机变量。

(2) 零件的强度r随时间推移而退化，即强度的均值随时间的推移而减小，而均方差σ随时间推移而增大，如图4-12所示。加在零件上的应力s对时间而言是稳态的，即其概率密度上(s)不随时间推移而变化。

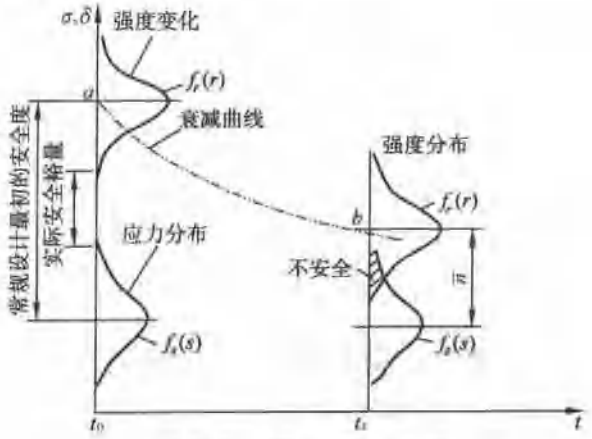


图4-12　强度-应力关系

(3) 当零件强度r大于加在零件上的应力s时，零件是可靠的，其可靠度表示为

$$R(t)=P(r>s)$$

常规设计与概率设计存在以下不同：

① 设计变量的性质不同。常规设计的设计变量是确定数值的单值变量，而概率设计中所涉及的变量为具有多值的随机变量，要以统计数据为基础，它们都服从一定的概率分布。

② 设计变量运算方法不同。常规设计中变量运算为实数域的代数运算，得到的是确定的单值实数，但在可靠性设计中，随机的设计变量间的运算要用概率及其分布函数的数字特征(均值和标准差)的概率运算法则进行。

③ 设计准则的含义不同。常规设计中，判断一个零件是否安全，应用安全系数来判断，在计算中未考虑影响零件应力和强度的许多非确定性因素。而在概率设计中，综合考虑了各个设计变量的统计分布特征，定量地用概率表达了所设计产品的可靠程度，因而更能反映实际情况，更科学合理。

4.4.2 应力-强度干涉模型

概率设计所依据的模型主要是应力-强度干涉模型。当应力超过强度时就会发生失效。这里的应力和强度具有广义的概念。应力表示导致失效的任何因素，如机械应力、电压或温度引起的内应力等。强度是指阻止失效发生的任何因素，如硬度、机械强度、加工精度、电器元件的击穿电压等。

机械产品的"可靠度"实质上就是零件在给定的运行条件下抵抗失效的能力，也就是"应力"与"强度"相互作用的结果，或者说是"应力"与"强度"干涉的结果。

令应力和强度的概率密度函数分别为 $f_s(s)$ 和 $f_r(r)$。一般情况下，应力和强度是相互独立的随机变量，且在机械设计中应力和强度具有相同的量纲，因此，可以把 $f_s(s)$ 和 $f_r(r)$ 表示在同一坐标系中。

由统计分布函数的性质可知，机械工程中常用的分布函数的概率密度曲线都是以横坐标为渐进线的，这样绘于同一坐标系中的两条概率密度曲线 $f_s(s)$ 和 $f_r(r)$ 必定有相交的区域，该区域称为干涉区，这个区域表示产品可能发生失效。图 4-13 称为应力-强度分布的干涉模型。

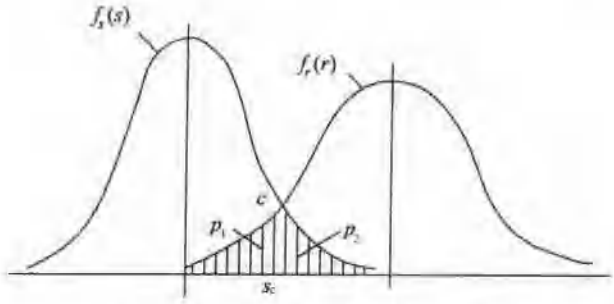


图 4-13 应力-强度分布的干涉模型

4.4.3 可靠度的确定方法

从干涉模型可以看到，要确定可靠度或失效概率必须研究一个随机变量超过另一个随机变量的概率，其推导过程如下：

(1) 令 E_1 表示应力随机变量 s 落在某一假定应力 s_0 附近一微区间 $\mathrm{d}s$ 内占的事件，如图 4-14 所示，则 E_1 出现的概率为

$$P(E_1)=P\left(s_0-\frac{\mathrm{d}s}{2}\leqslant s\leqslant s_0+\frac{\mathrm{d}s}{2}\right)=F_s\left(s_0-\frac{\mathrm{d}s}{2}\right)-F_s\left(s_0-\frac{\mathrm{d}s}{2}\right)=f_s(s_0)\mathrm{d}s$$

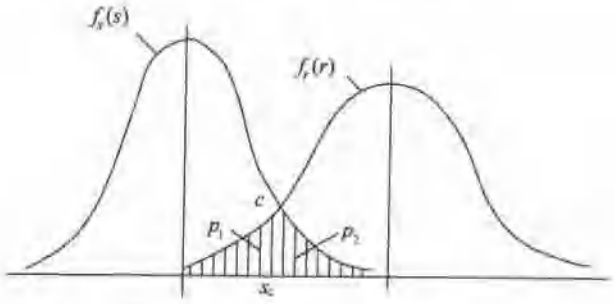


图4-14　应力-强度干涉模型的可靠性分析

(2) E_2 表示强度随机变量 r 大于 s_0 的事件，其出现概率为

$$P(E_2)=P(r>s_0)=\int_{s_0}^{\infty}f_r(r)\mathrm{d}r$$

(3) 可以认为事件 E_1、E_2 是互相独立的，所以 E_1、E_2 同时出现的概率为

$$P(E_1\cap E_2)=P(E_1)P(E_2)=f_s(s_0)\mathrm{d}s\int_{s_0}^{\infty}f_r(r)\mathrm{d}r$$

至此，求得 s 落在 s_0 区域而 r 又大于 s_0 的概率。

(4) 考虑到某一假定应力可能为 s_1、s_2 包括 s 所有可能出现的值，而只要 s 落在 s_i 某个区域而 r 又大于 s_i，产品就可靠，所以能够应用概率加法或积分，最终导出可靠度的表达式为

$$R=P(r>s)=\int_{-\infty}^{\infty}f_s(s)\left[\int_{s}^{\infty}f_r(r)\mathrm{d}r\right]\mathrm{d}s \tag{4-42}$$

式(4-42)即为在已知强度和应力的分布密度函数后，计算零件可靠度的一般方程式。

4.4.4　应力-强度均服从正态分布时的可靠度计算

通常，只要随机变量受多种因素影响而无一种因素起着显著且具有决定性作用时，可以认为该变量服从正态分布。在概率设计中常常将设计变量看做是正态变量。

根据应力-强度干涉模型导出的可靠度计算公式完全可以用于正态变量，但是，由于正态变量本身所具备的一些性质，如正态变量之和或差也服从正态分布，从而可以导出一组简单实用的概率设计公式。

假设应力与强度随机变量均服从正态分布，则它们的概率密度函数分别为

$$\begin{aligned}f_r(r)&=\frac{1}{\sigma_r\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{r-\mu_r}{\sigma_r}\right)^2\right]\quad(-\infty<r<\infty)\\ f_s(s)&=\frac{1}{\sigma_s\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{s-\mu_s}{\sigma_s}\right)^2\right]\quad(-\infty<s<\infty)\end{aligned} \tag{4-43}$$

式中：μ_r——强度的均值；

σ_r——强度的标准差；

μ_s——应力的均值；

σ_s——应力的标准差。

引进变量 y，令 $y=r-s$。因为 r 和 s 均为服从正态分布的随机变量，故其差 y 也是服

从正态分布的随机变量。

因此

$$f_y(y)=\frac{1}{\sigma_y\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{y-\mu_y}{\sigma_y}\right)^2\right] \quad (-\infty<y<\infty)$$

式中：μ_y——y 的均值，$\mu_y=\mu_r-\mu_s$；

σ_y——y 的标准差，$\sigma_y^2=\sigma_r^2+\sigma_s^2$。

那么可靠度

$$R=P[(r-s)>0]=P(y>0)=\int_0^\infty\frac{1}{\sigma_y\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{y-\mu_s}{\sigma_y}\right)^2\right]\mathrm{d}y \quad (4\text{-}44)$$

如果将随机变量 y 标准化，令

$$z=\frac{y-\mu_y}{\sigma_y}$$

则有 $\sigma_y\mathrm{d}z=\mathrm{d}y$，将其代入式(4-44)得

$$R=\int_z^\infty\frac{1}{\sigma_y\sqrt{2\pi}}\exp\left(-\frac{z^2}{2}\mathrm{d}y\right)=\frac{1}{\sqrt{2\pi}}\int_z^\infty\exp\left(\frac{z^2}{2}\right)\mathrm{d}z \quad (4\text{-}45)$$

当 $y=0$ 时，$z=\frac{0-\mu_y}{\sigma_y}=-\frac{\mu_y}{\sigma_y}$ 或 $z=-\frac{\mu_r-\mu_s}{\sqrt{\sigma_r^2+\sigma_s^2}}$。

将 z 值代入式(4-45)得

$$R=\frac{1}{\sqrt{2\pi}}\int_{-\frac{\mu_r-\mu_s}{\sqrt{\sigma_r^2+\sigma_s^2}}}^\infty\exp\left(-\frac{z^2}{2}\right)\mathrm{d}z \quad (4\text{-}46)$$

由上式可以看出，可靠度只明显地与积分下限有关。如把积分下限的负值表示为

$$z_0=\frac{\mu_r-\mu_s}{\sqrt{\sigma_r^2+\sigma_s^2}} \quad (4\text{-}47)$$

根据正态分布的对称性，可靠度 R 的计算式可写为

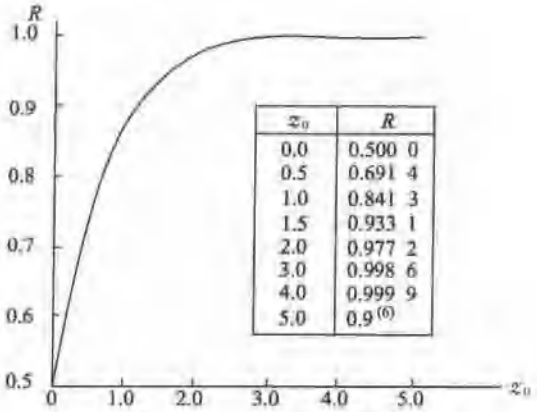
$$R=\frac{1}{\sqrt{2\pi}}\int_{-z_0}^\infty\exp\left(-\frac{z^2}{2}\right)\mathrm{d}z=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z_0}\exp-\left(\frac{z^2}{2}\right)\mathrm{d}z \quad (4\text{-}48)$$

式(4-47)将强度、应力和可靠度三者联系起来，故称它为“联结方程”或“耦合方程”，z_0 称为联结系数或可靠系数。

在已知 μ_r、μ_s、σ_r、σ_s 的条件下，利用联结方程可直接计算出 z_0 值，根据 z_0 值从标准正态分布表(附表)中查出可靠度值，也即

$$R=P(y>0)=\phi\left(\frac{\mu_r-\mu_s}{\sqrt{\sigma_r^2+\sigma_s^2}}\right) \quad (4\text{-}49)$$

可靠度 R 和 z_0 函数关系可用图 4-15 表示。



z_0	R
0.0	0.500 0
0.5	0.691 4
1.0	0.841 3
1.5	0.933 1
2.0	0.977 2
3.0	0.998 6
4.0	0.999 9
5.0	$0.9^{(6)}$

图 4-15　可靠度作为 z_0 函数关系图

4.5　机械静强度可靠设计

进行机械强度可靠性设计，首先要搞清楚载荷(应力)及零件强度的分布规律，合理地建立应力与强度之间的数学模型。应用应力-强度干涉理论，严格控制失效概率，以满足设计要求。整个设计过程可用图 4-16 表示。

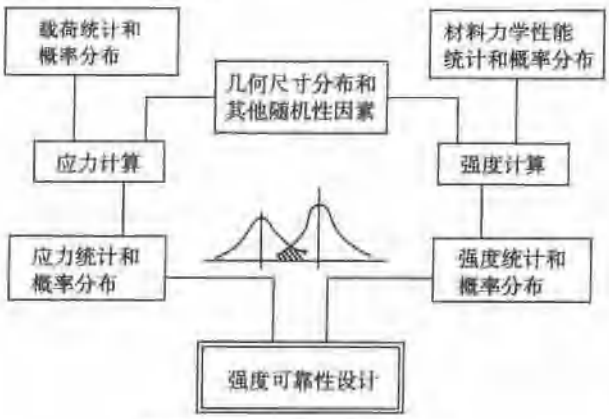


图 4-16　可靠性设计过程

4.5.1　材料力学性能统计处理

材料力学性能项目比较多，最常用的指标有强度极限、屈服极限、疲劳极限、硬度、延伸率、断裂韧性及弹性模量等，这些变量一般符合正态分布或近似等于正态分布，目前手册中给出的性能数据，一般是给出一个确定值，或是一个范围，尺寸数据一般是给出公称尺寸或公差。在概率设计中要应用这些数据时，需要从中得出某一参数的均值和标准差。

当材料性能数据给出范围时，设 max、min 分别为某一性能数据的上下限，则均值 μ 和标准差 σ 分别可取为

$$\mu=\frac{1}{2}(\max+\min) \tag{4-50}$$

$$\sigma=\frac{1}{6}(\max-\min) \tag{4-51}$$

上述两式是正态分布，设可靠度 99.7%是按 3σ 法则确定的。当给出材料性能数据确定值时，作统计量处理时，可以将此值作为该参量的均值，标准用变异系数（亦称变差系数）来求取。

材料性能变异系数是描述该性能参量相对的离散程度，一般用 V 表示，有

$$V=\frac{\text{标准差}(\sigma)}{\text{均值}(\mu)} \tag{4-52}$$

常用材料性能的变异系数 V 值如表 4-4 所示。

由式(4-52)得

$$\sigma=V\times\mu \tag{4-53}$$

表 4-4 常用材料变异系数

性　　能	V
金属材料的屈服强度	0.07(0.5—0.10)
金属材料的抗拉强度	0.05(0.05—0.10)
钢的疲劳持久极限	0.08(0.05—0.10)
钢的布氏硬度	0.05
金属材料的断裂韧性	0.07(0.05—0.13)
零件的疲劳强度	0.08—0.15
钢、铝的弹性模量	0.03
铸铁的弹性模量能	0.04

例 4-8 45＃调质钢，屈服强度 $\sigma_s=353\text{MPa}$，对这一数据作出统计处理时，可写为均值

$$\mu_{\sigma_s}=353\text{MPa}$$

取屈服强度的 V 值为 0.07，其标准差为

$$\sigma_{\sigma_s}=0.07\times353\text{MPa}=24.71\text{MPa}$$

4.5.2 工作载荷的统计分析

作用在机械或构件上的外载称为载荷，这些载荷可以是力、力矩、应力、功率、温度等，载荷有静载、动载、稳定、不稳定等多种类型。

通过实测，得到工作载荷的一系列数据，根据数据统计原理进行分析，确定其分布类型与参数，给出数学模型，为可靠性设计提供载荷参数。

对于动载荷，目前常用的记录方法有功率谱法和循环计数法。功率谱法借助于富氏变换，将复杂的随机载荷分解成有限个具有各种频率的简谐变化之和，以获得功率谱密度函数。循环计数法是把载荷-时间历程离散成一系列峰谷值，然后计算其峰谷值或幅值等发生的频率，从而找出概率密度函数及参数。

4.5.3 几何尺寸的分布与统计偏差

由于加工误差的原因，零件几何尺寸也随机变化。加工尺寸是多个随机因素综合影响

的结果，通常也符合正态分布。一般尺寸都给出规定的公差，这时可按 3σ 法则处理。若尺寸 D 的实际尺寸为 $D^{\pm T}$，则有 $3\sigma=T$，所以标准差为

$$\sigma=\frac{T}{3} \tag{4-54}$$

若尺寸的极限偏差对公称尺寸不是对称的(如单边的)，则由 $D^{=T}$ 可得

$$\sigma=\frac{T-0}{6}=\frac{T}{6} \tag{4-55}$$

4.5.4　随机变量函数的统计特征值

在机械强度可靠性设计中，经常应用正态分布函数，并需进行各随机变量间的代数运算。例如，应力 $s=\dfrac{P}{A}$，尺寸链 $L=L_1+L_2$，这里载荷 P、零件截面积 A、尺寸 L_1 和 L_2 均为随机变量，显然应力 s 和尺寸 L 仍是随机变量，称为随机变量函数。当各随机变量为正态分布时，随机变量函数仍为正态分布。下面介绍多个随机变量函数的统计特征值的求法。

设 $x_1,x_2,\cdots,x_n$ 是相互独立的随机变量，其均值和标准差分别为 $\mu_1,\mu_2,\cdots,\mu_n$ 和 $\sigma_1,\sigma_2,\cdots,\sigma_n$，则 $y=f(x_1,x_2,\cdots,x_n)$ 也是随机变量，其均值 μ_y 和标准差 σ_y 可按下式计算

$$\mu_y=f(\mu_1,\mu_2,\cdots,\mu_n) \tag{4-56}$$

$$\sigma_y^2=\left(\frac{\partial y}{\partial x_1}\right)_{x=\mu}\sigma_1^2+\left(\frac{\partial y}{\partial x_2}\right)_{x=\mu}\sigma_2^2+\cdots+\left(\frac{\partial y}{\partial x_n}\right)_{x=\mu}\sigma_n^2 \tag{4-57}$$

式中，$\left(\dfrac{\partial y}{\partial x_i}\right)_{x=\mu}$ 为计算 $\dfrac{\partial y}{\partial x_i}=\dfrac{\partial y}{\partial x_i}f(x_1,x_2,\cdots,x_n)$ 时，代入 $x_i=\mu_i$ 后的值。

例 4-9　当 $y=x_1+x_2+\cdots+x_n$ 时，求 y 的均值和标准差。

解：由式(4-56)和式(4-57)，有

$$\mu_y=f(\mu_1,\mu_2,\cdots,\mu_n)$$

$$\frac{\partial y}{\partial x_1}=\frac{\partial y}{\partial x_2}=\cdots=\frac{\partial y}{\partial x_n}=1$$

故有　　$\sigma_y^{\,2}=\sigma_1^{\,2}+\sigma_2^{\,2}+\cdots+\sigma_n^{\,2}$

即　　$\sigma_y=\sqrt{\sigma_1^{\,2}+\sigma_2^{\,2}+\cdots\sigma_n^{\,2}}$

为应用方便，将常用的正态分布随机变量函数 z 的统计特征值求解公式列于表 4-5。表中，x 和 y 为相互独立的服从正态分布的随机变量，a 为任意常数。

表 4-5

序号	函数形式	均值	标准离差 σ_z
1	$z=ax$	$a\mu_x$	$a\sigma_x$
2	$z=x+a$	μ_x+a	σ_x
3	$z=x\pm y$	$\mu_x\pm\mu_y$	$\sqrt{\sigma_x^{\,2}+\sigma_y^{\,2}}$
4	$z=xy$	$\mu_x\mu_y$	$\sqrt{\mu_x^{\,2}\sigma_y^{\,2}+\mu_y^{\,2}\sigma_x^{\,2}}$
5	$z=\dfrac{x}{y}$	$\dfrac{\mu_x}{\mu_y}$	$\dfrac{1}{\mu_y^{\,2}}\sqrt{\mu_x^{\,2}\sigma_y^{\,2}+\mu_y^{\,2}\sigma_x^{\,2}}$
6	$z=x^n$	μ_x^n	$n\mu_x^{\,n-1}\sigma_x$
7	$z=x^{\frac{1}{2}}$	$\mu_x^{\frac{1}{2}}$	$\dfrac{1}{2}\mu_x^{-\frac{1}{2}}\sigma_x$

4.5.5 零件静强度可靠性设计

零部件的静强度可靠性设计思路与步骤同常规设计相类似。不同之处仅在于可靠性设计把各设计变量，如载荷、材料强度、零件尺寸及其他影响因素都视为随机变量，并服从某一分布。这里仅仅讨论最常见的正态分布情况。对于零部件疲劳强度的可靠性设计问题，在此不加以讨论，因为其方法与静强度可靠性设计一样，不同之处仅在于应力应采用交变应力，强度应采用疲劳强度，且要考虑对不同零部件的疲劳极限系数的修正。

1. 受拉零件的静强度可靠性设计

在机械设计中受拉零件较多。作用在零件上的拉伸载荷 $P(\overline{P},\sigma_p)$、零件的计算截面积 $A(\overline{A},\sigma_A)$、零件材料的抗拉强度 $\delta(\overline{\delta},\sigma_\delta)$ 均为随机变量，且一般呈正态分布。若载荷的波动很小，则可按静强度问题处理。失效模式为拉断。其静强度可靠性设计步骤如下：

(1) 选定可靠度 R；

(2) 计算零件发生强度破坏的概率 F：

$$F=1-R$$

(3) 由 F 值查附表取 z 值后，得 $z_R=-z$；

(4) 确定零件强度的分布参数 μ_δ、σ_δ，在未给定又无统计资料的情况下可用近似计算式计算：

$$\mu_{b_t}=\frac{\varepsilon_1}{\varepsilon_2}\sigma_\delta$$

① 对塑性材料：

$$\sigma_\delta=0.1\mu_\delta=0.1\left(\frac{\varepsilon_1}{\varepsilon_2}\right)\sigma_s \tag{4-58}$$

② 对脆性材料：

$$\sigma_\delta=0.1\mu_\delta=0.1\left(\frac{\varepsilon_1}{\varepsilon_2}\right)\sigma_{b_t} \tag{4-59}$$

(5) 列出应力 s 的表达式；

(6) 计算工作应力；

由于截面积尺寸 A 是要求的未知量，因此工作应力可表达为 A 的函数。

(7) 将应力、强度、z_R 均代入联结方程

$$z_R=\frac{\mu_\delta-\mu_s}{\sqrt{\sigma_\delta^2+\sigma_s^2}}$$

求得截面积参数的均值。

为了对计算结果进行分析、比较和检验，有时还加进某些参数值的变化对可靠度影响的分析，有时还与常规设计结果进行比较。有的还将联结方程中的 μ_s 值乘以强度储备系数 n ($n\geqslant 1$，如取 $n=1.25$)以增强强度储备。

例 4-10 要设计一拉杆，所承受的拉力 $P\sim N(\mu_P,\sigma_P^2)$，其中 $\mu_P=40000\text{N}$，$\sigma_P=1200\text{N}$；取 45＃钢为制造材料，求拉杆的截面尺寸。

解：设拉杆取圆截面，其半径为 $r(\text{mm})$，求 μ_r、σ_r。查表知 45＃碳素钢的抗拉强度数据为 $\mu_\delta=667\text{MPa}$，$\sigma_\delta=25.3\text{MPa}$，也服从正态分布。

解题步骤如下：

(1) 选定可靠度为 $R=0.999$。

(2) 计算零件发生强度破坏的概率

$$F=1-R=1-0.999=0.001$$

(3) 查附表，得 $z_R=-z=3.09$。

(4) 查得强度的分布参数为

$$\mu_\delta=667\text{MPa},\ \sigma_\delta=25.3\text{MPa}$$

(5) 列出应力表达式：$s=\dfrac{P}{A}=\dfrac{P}{\pi r^2}\text{MPa}$，由表 4-5 知

$$\mu_A=\pi\mu_r^2,\ \sigma_A=\pi\cdot 2\mu_r\sigma_r=2\pi\mu_r\sigma_r$$

取拉杆圆截面半径的公差为 $\pm\Delta_r=\pm 0.015\mu_r$，则按式(4-54)可求得

$$\sigma_r=\frac{\Delta_r}{3}=\frac{0.015}{3}\mu_r=0.005\mu_r\text{mm}$$

$$\sigma_A=2\pi\mu_r\sigma_r=0.01\pi\mu_r^2\text{mm}^2$$

$$\mu_s=\frac{\mu_P}{\mu_A}=\frac{\mu_P}{\pi\mu_r^2}=\frac{40000}{\pi\mu_r^2}\text{MPa}$$

$$\sigma_s=\frac{1}{\mu_A^2}\sqrt{\mu_P^2\sigma_A^2+\mu_A^2\sigma_P^2}=\frac{1}{(\pi\mu_r^2)^2}\sqrt{(0.01\pi\mu_r^2)\mu_P^2+(\pi\mu_r^2)^2\sigma_P^2}$$

$$=\frac{1}{\pi\mu_r^2}\sqrt{(0.01)^2\mu_P^2+\sigma_P^2}\text{MPa}$$

(6) 计算工作应力，得

$$\mu_s=\frac{40000}{\pi\mu_r^2}=12732.406\,\frac{1}{\mu_r^2}\text{MPa}$$

$$\sigma_s=\frac{1}{\pi\mu_r^2}\sqrt{(0.01)^2\cdot(40000)^2+(1200)^2}=402.634\,\frac{1}{\mu_r^2}\text{MPa}$$

(7) 将应力、强度及 z_R 代入联结方程：

$$z_R=\frac{\mu_\delta-\mu_s}{\sqrt{\sigma_\delta^2+\sigma_s^2}}=\frac{667-\dfrac{12732.406}{\mu_r^2}}{\sqrt{(25.3)^2+\dfrac{(402.634)^2}{\mu_r^4}}}=3.09$$

或

$$\frac{667\mu_r^2-12732.406}{\sqrt{(25.3)^2\mu_r^4+(402.634)^2}}=3.09$$

化简后得

$$\mu_r^4-38.710\mu_r^2+365.940=0$$

解得

$$\mu_r^2=22.301\text{mm}^2\ \text{和}\ \mu_r^2=16.410\text{mm}^2$$

或

$$\mu_r=4.722\text{mm}\ \text{和}\ \mu_r=4.050\text{mm}$$

代入联结方程验算，取 $\mu_r=4.722$，舍去 $\mu_r=4.050$

$$\sigma_r=0.005\mu_r=0.005\times 4.722=0.0236$$

$$r=\mu_r\pm\Delta_r=4.722\pm 3\sigma_r=(4.722\pm 0.0708)(\text{mm})$$

因此，为保证拉杆的可靠度为 0.999，其半径应为(4.722±0.0710)mm。

为进一步分析设计计算结果，可把它与常规设计作一比较。

(8) 与常规设计作比较。为了比较，因此拉杆的材料不变，仍用圆截面，取安全系数 $n=3$，则有

$$\sigma=\frac{P}{\pi r^2}\leqslant[\sigma]=\frac{\mu_S}{n}=\frac{667}{3}=222.333\text{MPa}$$

即有

$$\frac{40000}{\pi r^2}\leqslant 222.333,\ r^2\geqslant\frac{40000}{\pi\times 222.333}=57.267$$

得拉杆圆截面的半径为 $r\geqslant 7.568$mm。

显然，常规设计结果比可靠性设计结果大了许多。如果在常规设计中采用拉杆半径为 $r=4.722$mm，即可靠性设计结果，则其安全系数变为

$$n\leqslant\frac{\mu_S\pi r^2}{F}=\frac{667\times\pi\times(4.722)^2}{40000}=1.168$$

这从常规设计来看是不敢采用的，而可靠性设计采用这一结果，其可靠度竟达到 0.999，即拉杆破坏的概率仅有 0.1%。但从联结方程可以看出，要保证这一高的可靠度必须使 μ_S、σ_S、μ_r、σ_r 值保持稳定不变，即可靠性设计的先进性是要以材料制造工艺的稳定性及对载荷测定的准确性为前提条件。

(9) 敏感度分析。如果本例题的其他条件不变，而载荷及强度的标准差即 σ_r、σ_S 值均增大，通过具体计算就可以明显看出，由于载荷和强度值分散性的增加，可靠度将迅速下降。因此，当载荷及强度的均值不变时，只有严格控制载荷和强度的分散性才能保证可靠性设计结果能更好地应用。

2. 梁的静强度可靠性设计

例 4-11 受集中载荷力 P 作用的简支梁，如图 4-17 所示。显然，力 P、跨度 l、力作用点位置 a 均为随机变量。当载荷 $P(\overline{P},\sigma_P)$、梁的跨度 $l(\overline{l},\sigma_l)$ 时力作用点位置的均值标准差为 $a(\overline{a},\sigma_a)$。

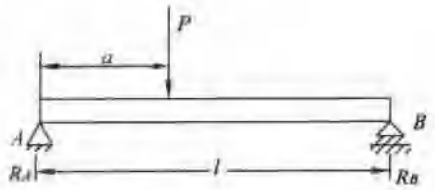


图 4-17 受集中载荷力 P 作用的简支梁

梁的静强度可靠性设计步骤与上面介绍的拉杆类似。

(1) 选定可靠度 R；

(2) 计算 $F=1-R$；

(3) 按 F 值查附表，取 z 值后得 $z_R=-z$；

(4) 确定强度分布参数 $\overline{\delta}$、σ_δ；

(5) 列出应力 s 的表达式：

梁的最大弯矩发生在载荷力 P 的作用点处，其值为

$$M=\frac{Pa(l-a)}{l}\tag{4-60}$$

式中 P、l、a，如图 4-17 所示。

最大弯曲应力则发生在该截面的底面和顶面，其值为

$$s=\frac{MC}{I}$$

式中：s——应力，单位为 MPa；

M——弯矩，单位为 N·mm，见式(4-60)；

C——截面中性轴至梁的底面或顶面的距离，单位为 mm；

I——梁截面对中性轴的惯性矩，单位为 mm^4。

(6) 计算工作应力。将已知量代入上述应力公式，其中包括待求的梁截面的尺寸参数，如梁截面的高度。

(7) 将应力、强度的分布参数代入联结方程，求未知量。

(8) 敏感度分析。

例 4-12　今要设计一工字钢简支梁，已知参数如下：

跨距：$l=3048\pm3.175\text{mm}$，$\bar{l}=3048\text{mm}$，$\sigma_l=1.058\text{mm}$；

梁上受力点至梁一端支承的距离：

$a=1828.8\pm3.175\text{mm}$，$\bar{a}=1828.3\text{mm}$，$\sigma_a=1.058\text{mm}$；

载荷：$\bar{P}=27011.5\text{N}$，$\sigma_P=890\text{N}$；

工字钢强度：$\bar{\delta}=1171.2\text{MPa}$，$\sigma_\delta=32.794\text{MPa}$。试用可靠性设计方法，在保证 $R=0.9990$ 的条件下确定工字钢的尺寸。

解：工字钢的尺寸符号如图 4-18 所示。给定其尺寸关系有：$\frac{b}{t}=8.88$，$\frac{h}{d}=15.7$，$\frac{b}{h}=0.92$，因此

$$\frac{I}{C}=\frac{bh^3-(b-d)(h-2t)^3}{6h}=0.0822h^3$$

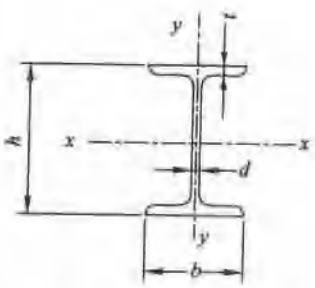


图 4-18　工字钢截面

由表 4-3，并令 $\sigma_h=0.01\bar{h}$，则 $(\overline{I/C})=0.0822\bar{h}^3$ 和 $\sigma_{(I/C)}=0.002466\bar{h}^3$。

按以下步骤进行：

(1) 给定 $R=0.9990$；

(2) 求 $F=1-R=0.0010$；

(3) 按 F 值查附表求得 $z_R=-z=3.09$；

(4) 强度分布参数已给定：

$$\bar{\delta}=1171.2\text{MPa}，\sigma_\delta=32.794\text{MPa}$$

(5) 列出应力表达式：

$$\left.\begin{aligned}\bar{s}&=\frac{\overline{M}}{(\overline{I/C})}\\ \sigma_s&=\left\{\left[\frac{2}{(\overline{I/C})}\right]^2\sigma_M{}^2+\left[\frac{-\overline{M}}{(\overline{I/C})^2}\right]^2\sigma^2_{(I/C)}\right\}^{\frac{1}{2}}\end{aligned}\right\}\tag{4-61}$$

(6) 计算工作应力：由表 4-3 所列公式，可求得

$$\overline{M}=\bar{P}\,\bar{a}\left(1-\frac{\bar{a}}{\bar{l}}\right)$$

$$=27011.5\times1828.8\left(1-\frac{1828.8}{3048}\right)$$

$$=19759452.48\text{N}\cdot\text{mm}$$

因而
$$s=\frac{19759452.48}{0.0822\bar{h}^3}=\frac{240382633.6}{\bar{h}^3}\text{MPa}$$

按式(4-57)求 σ_M^2：

$$\sigma_y^2=\left(\frac{\partial y}{\partial x_1}\right)^2_{x=\mu}\sigma_1^2+\left(\frac{\partial y}{\partial x_2}\right)^2_{x=\mu}\sigma_2^2+\cdots+\left(\frac{\partial y}{\partial x_n}\right)^2_{x=\mu}\sigma_n^2$$

$$=\left[\frac{a(l-a)}{l}\right]^2\sigma_P^2+\left(P-\frac{2Pa}{l}\right)^2\sigma_a^2+\left(\frac{Pa^2}{l^2}\right)\sigma_l^2$$

$$=\left[\frac{1828.8(3048-1828.8)}{3048}\right]^2(890)^2+\left[27011.5-\frac{2\times27011.5\times(1828.8)}{(3048)^2}\right]^2\times(1.058)^2$$

$$=424008262692\approx4.240\times10^{11}(\text{N}\cdot\text{mm}^2)$$

$\sigma_y=651160\text{N}\cdot\text{mm}$

$\sigma_M=651160\text{N}\cdot\text{mm}$

将以上有关值代入式(4-61)，得

$$\sigma_s=\left\{\left(\frac{1}{0.0822\bar{h}^3}\right)^2\times4.240\times10^{11}+\left[\frac{-19759452.48}{(0.0822\bar{h}^3)^2}\right]^2\times(0.002466\bar{h}^3)^2\right\}^{\frac{1}{2}}$$

$$=\frac{10712453.33}{\bar{h}^3}$$

(7) 将应力、强度分布参数代入联结方程，求未知量 h：

$$z_R=\frac{\bar{\delta}-\bar{s}}{\sqrt{\sigma_\delta^2+\sigma_s^2}}$$

$$3.09=\frac{1171.2-\left(\frac{240382633.6}{\bar{h}^3}\right)}{\sqrt{(32.794)^2+\left(\frac{10712453.33}{\bar{h}^3}\right)^2}}$$

解上式可求得 $h=62.154\text{mm}$，这时可靠度满足 $R=0.9990$。

(8) 敏感度分析。将 $h=62.154$ 代入上式，并令 σ_δ 及 z_R 作为待定量，这样就可以研究材料强度对 z_R 进而对可靠度 R 的影响，即研究可靠性对于材料强度变化的敏感度。对于 σ_δ 取不同值时的 R 值见表 4-6。

表 4-6 σ_δ 与 R 间的关系

σ_δ/MPa	z_R	R	σ_δ/MPa	z_R	R
34.447	3.035	0.998797	75.783	1.945	0.974110
48.226	2.604	0.995393	89.562	1.709	0.956276
62.005	2.239	0.987418	103.341	1.519	0.935614

3. 承受转矩的轴的静强度可靠性设计

研究一端固定而另一端承受转矩的实心轴的可靠性设计，如汽车的扭杆弹簧，假定其应力、强度均呈正态分布，则其静强度可靠性设计步骤与前述步骤完全相同，仅应力表达式有差别。

设轴的直径为 $d$mm，单位长度的扭转角 θ(rad)，轴的材料的剪切弹性模量为 G(MPa)，

则在转矩

$$T=G\theta I_P$$

的作用下，产生的剪切应力为　　$\tau=\frac{1}{2}G\theta d=\frac{Td}{2I_P}$

式中 I_P 为轴横截面的极惯性矩。对于实心轴 $I_P=\frac{\pi d^4}{32}$，因此有

$$\tau=\frac{16T}{\pi d^3}=\frac{2T}{\pi r^3} \tag{4-62}$$

例 4-13　要求设计一个一端固定另一端受扭的轴，设计随机变量的分布参数为：

作用转矩：$T\sim N(\overline{T},\sigma_T^2)$

$$\overline{T}=11303000\text{N}\cdot\text{mm},\ \sigma_T=1130300\text{N}\cdot\text{mm}$$

许用剪切应力：$\delta\sim N(\overline{\delta},\sigma_\delta^2)$

$$\overline{\delta}=344.47\text{MPa},\ \sigma_\delta=34.447\text{MPa}$$

轴半径的变化为

$$\sigma_r=\frac{\alpha}{3}\overline{r}$$

式中：α 为偏差系数。

解：可靠性设计计算：

(1) 给定可靠度 $R=0.999$；

(2) 求 $F=1-R=0.001$；

(3) 按 F 值查附表，得 $z_R=3.09$；

(4) 强度分布参数：

$\overline{\delta}=344.47\text{MPa},\ \sigma_\delta=34.447\text{MPa}$；

(5) 列出应力表达式，如式(4-62)所示；按表 4-3，于是有

$$\overline{\tau}=\frac{2\overline{T}}{\pi\times\overline{r}^3}(\text{MPa}) \tag{4-63}$$

按式(4-57)求得

$$\sigma_\tau^2=\frac{4\sigma_T^2}{\pi^2\overline{r}^6}+\frac{36\overline{T}^2\sigma_r^2}{\pi^2\overline{r}^8} \tag{4-64}$$

(6) 计算工作应力

$$\overline{\tau}=\frac{2\overline{T}}{\pi\times\overline{r}^3}=\frac{2\times11303000}{\pi\times r^3}=\frac{7195719.365}{\overline{r}^3}(\text{MPa})$$

$$\sigma_\tau^2=\frac{4\sigma_T^2}{\pi^2\overline{r}^6}+\frac{36\overline{T}^2\sigma_r^2}{\pi^2\overline{r}^8}=\frac{4(1130300)^2}{\pi^2\overline{r}^6}+\frac{36(11303000)^2\times\frac{\alpha^2}{9}\overline{r}^2}{\pi^2\overline{r}^8}$$

$$=\frac{4(1130300)^2[1+(10\alpha)^2]}{\pi^2\overline{r}^6}$$

$$\sigma_\tau=\frac{2\times1130300}{\pi\overline{r}^3}\sqrt{1+(10\alpha)^2}=\frac{719571.9365}{\overline{r}^3}\sqrt{1+(10\alpha)^2}(\text{MPa})$$

(7) 将应力、强度的分布参数代入联结方程，求未知量半径 r：

$$z_R=3.09=\frac{\bar{\delta}-\bar{\tau}}{\sqrt{\sigma_\delta^2+\sigma_\tau^2}}=\frac{344.47-\left(\frac{7195719.365}{\bar{r}^3}\right)}{\sqrt{(344.47)^2+\left(\frac{7195719.365}{\bar{r}^3}\right)^2(1+100\alpha^2)}}$$

设 $\alpha=0.03$，代入上式，可解得

$$\bar{r}=32.13\text{mm}$$，并可满足 $R=0.999$。

(8) 敏感度分析。将 $\bar{r}=32.13$mm 代入上式并改变 α 值，计算相应的 z_R 值以及可靠度，分析半径的偏差对可靠度的影响，结果如表 4-7 所示。

表 4-7 半径偏差 α 对可靠度的影响

$\bar{r}$ 的偏差 α	z_R	R	$\bar{r}$ 的偏差 α	z_R	R
0.010	3.136	0.99916	0.040	3.072	0.99890
0.020	3.123	0.99910	0.050	3.035	0.99880
0.030	3.099	0.99903	0.100	2.772	0.99740

如果取 $\bar{r}=32.13$ 和 $\alpha=0.03$，而改变上述联结方程中的 σ_δ 值，计算相应的 z_R 值及可靠度 R，则 σ_δ 值对可靠度 R 的影响如表 4-8 所示。

表 4-8 σ_δ 对可靠度 R 的影响

剪切强度标准差/MPa	z_R	R	剪切强度标准差/MPa	z_R	R
13.779	4.825	0.99999	41.336	6.712	0.99664
27.558	3.585	0.99983	55.115	2.145	0.98422
34.447	3.090	0.99903	68.894	1.763	0.96080

当 α 和 σ_δ 具有前述给定值时，利用联结方程可计算 $\bar{r}$ 对可靠度的影响，结果如表 4-9 所示。

表 4-9 平均半径 $\bar{r}$ 对可靠度 R 的影响

轴的半径 $\bar{r}$/mm	z_R	R	轴的半径 $\bar{r}$/mm	z_R	R
25.40	−1.642	0.05050	40.64	6.555	0.99999
30.48	2.086	0.98169	45.72	7.621	0.99999
35.56	4.824	0.99999	50.80	8.736	1.00000

对于传递转矩并由钢管制成的汽车传动轴或其他传递转矩的转轴来说，上述可靠性设计方法也是适用的。

4.6 机械系统可靠性设计

系统是由零件、部件、子系统等组成的。系统的可靠性，不仅取决于组成系统零部件的可靠性，而且也取决于各组成零部件的相互组合方式。

机械系统可靠性设计的目的，就是要使系统在满足规定可靠性指标，完成预定功能的前

提下，使系统的技术性能、重量、成本、时间等各方面取得协调，求得最佳设计；或是在性能、重量、成本、时间和其他要求的约束下，设计能得到实际高可靠性的系统。

系统可靠性设计方法可归纳为两种类型：

(1) 按已知零部件的可靠性数据，计算系统的可靠性指标，进行几种设计方案的比较，以选择最佳设计方案。

(2) 按规定的系统可靠性指标，对各组成零部件进行可靠性分配，进行几种设计方案的比较，以选择最佳设计方案。

下面分别就如何确定元件的可靠性、如何计算系统的可靠性、系统可靠性如何在各组成元件之间进行合理分配等问题加以讨论。

4.6.1　可靠性预测

可靠性预测是一种预报方法，它是从所得的失效率数据预报一个元件、部件、子系统或系统实际可能达到的可靠度，即预报这些元件或系统在特定的应用中完成规定功能的概率。

可靠性预测的目的是：

(1) 协调设计参数及指标，提高产品可靠性。零部件和系统的设计，必须考虑性能、重量、成本、时间和其他要求等参数及指标，这些参数及指标互相制约着，可靠性预测可以用来协调参数及指标，以求得合理的高可靠度为准绳。

(2) 对比设计方案，以选择最佳系统。为完成某一任务，可以提出几种设计方案，在设计开始阶段，一般总要进行方案分析及比较，以期在几个设计方案中选择最佳方案。分析及比较某一系统、子系统或某一设备的几个设计方案时，左右其选择的因素之一是这些方案的相对可靠度。可靠性预测可以用来分析及比较几种设计方案的可靠度，以便从中选择最佳方案。

(3) 预示薄弱环节，以采取改进措施。可靠性预测可以发现哪些零部件或子系统是造成系统失效的主要因素。找出薄弱环节之后，便可采取必要的改进措施，以减小整个系统的失效率，提高系统的可靠性。

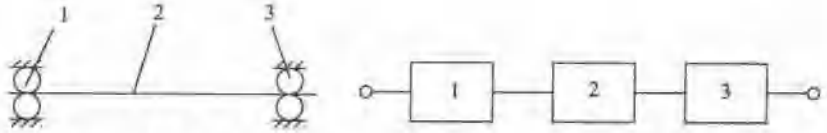
可靠性预测是可靠性设计的重要内容，它包括元件可靠性预测和系统可靠性预测。预测系统的可靠度通常是以预测系统中的元件或组件的可靠度为基础。所有元件的可靠度确定以后，把这些元件的可靠度适当地组合起来就可得出系统的可靠度，元件的可靠性是在一定的使用(或试验)条件和环境条件下得出的，设计时可从可靠性手册上查得，各国均有专门的组织进行收集、整理、提供和管理各可靠性数据，有条件时可通过可靠性试验求得。

系统(或称设备)的可靠性是与组成系统的零部件数量、零部件的可靠性以及零部件之间的相互关系有关。前面已讨论了关于零部件可靠性预测的基本方法，现在讨论各零部件在系统中的相互关系，以便对系统的可靠性进行预测。这里所说的零部件的相互关系主要是指功能关系，而不是物理关系。为此，下面先讨论逻辑图。

1. 逻辑图

当我们研究一个系统时，特别是研究一个大的复杂系统时，首先必须了解组成该系统的各单元或子系统的功能、相互关系和对所研究系统的影响。一个系统小则由两个子系统组成，大则由几百或上千个子系统组成。为了清晰地研究它们，在可靠性工程中往往用系统图和逻辑图来描述，进而对系统及其组成零部件进行定量的设计与计算。

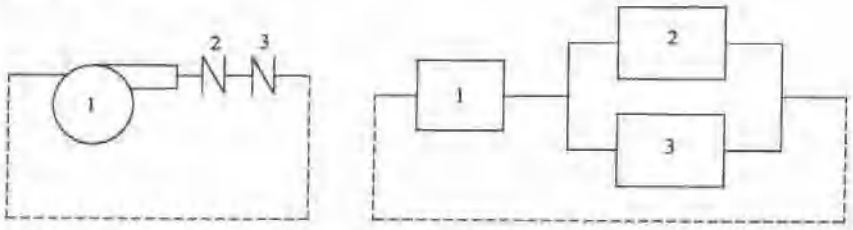
逻辑图通常由一组框图或单元连成一个或几个子系统所构成。一个子系统包含的单元数可以少到一个，也可以多到上百个，每个子系统都表示系统完成某一功能或几个功能的逻辑关系。图 4-19 所示是由三个零件组成的最简单的轴系，由三个零件串联的子系统构成一个系统逻辑图。



(a) 由三个零件组成的最简轴系　　(b) 最简轴系逻辑

图 4-19　由三个零件构成的一个串联系统逻辑图

值得注意的是，往往有一些单元，它们在系统结构图中是并联的，而它们的功能关系却是任一单元失效都将引起系统不能完成功能。因此，这种单元在逻辑图中应用串联表示。同样，有一些单元，它们在结构图中是串联的，而它们的功能却是任一单元失效并不引起系统不能完成功能。因此，这种单元在逻辑图中应用并联表示。还有另外一些单元其逻辑关系要在设计计算后才能确定其串并联关系。例如，有一流体系统，由一个泵和两个抑制阀串联组成。图 4-20(a)为这部分装置的系统结构图。图中的两个抑制阀是当泵不工作并且倒流压力超过顺流压力时阻止倒流的冗余系统。它的逻辑图却应如图 4-20(b)所示，其中两个抑制阀应用并联表示。



(a)　　(b)

1-泵　2,3-抑制阀

图 4-20　流体泵压系统逻辑

逻辑图的作用，一是反映零部件之间的功能关系；二是为计算系统的可靠度提供数学模型。在下面介绍的系统可靠性预测的第一种方法，即数学模型法中可以清楚看出这一点。

下面分别讨论系统可靠性的预测方法。

2. 数学模型法

(1) 串联系统的可靠性计算。在构成一个系统的元件中，只要有一个失效，该系统就失效的话，这种系统称为串联系统。串联系统的逻辑图如图 4-21 所示。

例如，齿轮减速器是由齿、轴、键、轴承、箱体、螺栓、螺母等组成，从功能关系来看，它们中任一部分失效，都会使减速器不能正常工作，因此，它们的逻辑图是串联的。又如起重机的升降机构是由电动机、联轴器、制动器、减速器、卷筒、钢丝绳、滑轮组、吊钩装置等部件组成。它们中任一部分失效，都会使升降机构不能工作，因此，它们的逻辑图也是串联的。

图 4-21　串联系统逻辑

设各单元的可靠度分别为 R_1、R_2、…、R_n，如果这个单元的失效相互独立，则由 n 个单元组成的串联系统的可靠度，可根据概率乘法定理按下式计算

$$R_s = R_1R_2R_3\cdots R_n = \prod_{i=1}^{n} R_i \tag{4-65}$$

由此可见，串联系统的可靠度 R_s 与串联单元的数量 n 及可靠度 R_i 有关。在串联系统中，随着单元可靠度的减小和单元数量的增加，串联系统的可靠度将迅速降低。必要时应采取措施来提高系统的可靠度。

设各单元的失效分布服从指数分布，并且失效率分别为 $\lambda_1(t)$、$\lambda_2(t)$、…、$\lambda_n(t)$，则

$$R_1(t) = \exp\left[-\int_0^t \lambda_1(t)\mathrm{d}t\right]$$

$$R_2(t) = \exp\left[-\int_0^t \lambda_2(t)\mathrm{d}t\right]$$

$$R_n(t) = \exp\left[-\int_0^t \lambda_n(t)\mathrm{d}t\right]$$

代入式(4-65)，得

$$R_s(t) = \exp\left\{-\int_0^t [\lambda_1(t) + \lambda_2(t) + \cdots + \lambda_n(t)]\mathrm{d}t\right\}$$

于是

$$\lambda_s(t) = \lambda_1(t) + \lambda_2(t) + \cdots \lambda_n(t) \tag{4-66}$$

$$R_s(t) = \exp\left[-\int_0^t \lambda_s(t)\mathrm{d}t\right] \tag{4-67}$$

上式说明对串联系统来说，系统失效率 $\lambda_s(t)$ 是各单元失效率 $\lambda_i(t)$ 之和。

由于可靠性预测主要是针对产品的正常期，因此可以认为各单元的失效率基本上为常量，所以式(4-66)和式(4-67)在实际应用时，可改写为

$$\lambda_s = \lambda_1 + \lambda_2 + \cdots \lambda_n = \sum_{i=1}^{n} \lambda_i \tag{4-68}$$

$$R_s = \mathrm{e}^{-\lambda_s t} = \mathrm{e}^{-\sum_{i=1}^{n}\lambda_i t} \tag{4-69}$$

$$\theta_s = \frac{1}{\lambda} = \frac{1}{\sum_{i=1}^{n}\lambda_i} \tag{4-70}$$

式中：θ_s 表示系统工作的平均寿命。

(2) 并联系统的可靠性计算。构成系统的元件，只有在全部发生故障后，整个系统才不能工作。这种系统称为并联系统，由于并联系统有单元的重复，而且只要有一个单元不失效就能维持整个系统工作，所以又称为工作冗余系统。并联系统的逻辑图如图 4-22 所示。

设各单元的可靠度分别为 R_1、R_2、…、R_n，则各单元的失效概率分别为 $1-R_1$、$1-R_2$、…、$1-R_n$。如果各单元的失效互相独立，则由 n 个单元组成的并联系统的失效概率 F_s 可根据乘法定理按下式计算

$$F_s=(1-R_1)(1-R_2)\cdots(1-R_n)=\prod_{i=1}^{n}(1-R_i) \tag{4-71}$$

所以并联系统的可靠度为

$$R_s=1-F_s=1-\prod_{i=1}^{n}(1-R_i) \tag{4-72}$$

对于并联系统，R_s 随 n 及 R 的增加而增大。当提高元件的可靠度受到限制的情况下，用并联系统，可以提高系统的可靠度。在机械系统中实际上用得较多的是 $n=2$ 的情况。并联系统工作的平均寿命 θ_s 可用下式计算

$$\theta_s=\int_0^{\infty}R(t)\mathrm{d}t \tag{4-73}$$

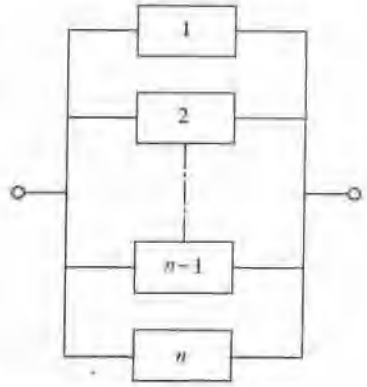


图 4-22 并联系统的逻辑

(3) 后备系统的可靠性计算。这种系统也是并联系统，但是有的单元并不工作，当某一个工作单元失效后，原来未参与工作的单元开始工作，而将失效单元换下、修理或更换，故又称为后备冗余系统，也称非工作后备系统。后备系统的逻辑图如图 4-23 所示。

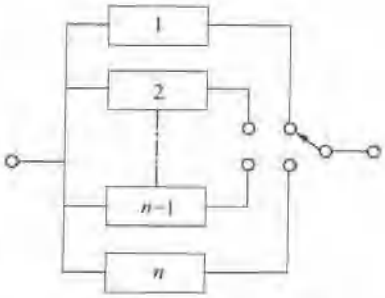


图 4-23 后备系统逻辑图

由 n 个元件构成的后备系统，在给定的时间 t 内，只要失效元件数不多于 $n-1$ 个，系统均处于可靠状态。设元件的失效率为 $\lambda_1(t)=\lambda_2(t)=\cdots=\lambda_n(t)=\lambda$，则系统的可靠度按下列普阿松分布的部分求和公式来求

$$R_s(t)=\mathrm{e}^{-\lambda t}\left[1+\lambda t+\frac{(\lambda t)^2}{2!}+\frac{(\lambda t)^3}{3!}+\frac{(\lambda t)^4}{4!}+\cdots+\frac{(\lambda t)^{n-1}}{(n-1)!}\right] \tag{4-74}$$

例如，当 $n=2$，则

$$R_s=\mathrm{e}^{-\lambda t}(1+\lambda t)$$

$$\lambda_s=-\frac{1}{R_s}\frac{\mathrm{d}R_s}{\mathrm{d}t}=\frac{\lambda^2 t}{1+\lambda t}$$

$$\theta_s=\int_0^{\infty}R_s\mathrm{d}t=\int_0^{\infty}\mathrm{e}^{-\lambda t}\mathrm{d}t+\int_0^{\infty}\lambda t\mathrm{e}^{-\lambda t}\mathrm{d}t$$

$$=\frac{1}{\lambda}+\frac{1}{\lambda}=\frac{2}{\lambda}$$

(4) 表决系统的可靠性计算。一个由 n 个元件两两组成的并联系统，只要其中任意 m 个不失效，则系统就不会失效，这就是 n 中取 m 个表示系统。如图 4-24 所示，为三中取二表决系统逻辑图。此系统要求失效不多于一个元件，故有四种成功的工况：即没有失效，只有第一个元件失效，只有第二个元件失效，只有第三个元

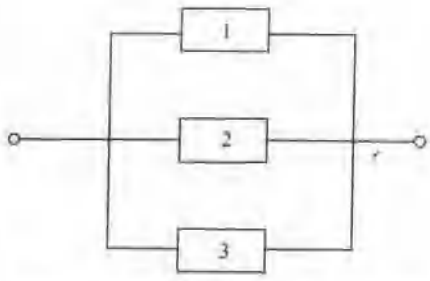


图 4-24 三中取二表决系统逻辑图

件失效。按概率乘法定理和加法定理，可求得系统的可靠度。

$$R_s=R_1R_2R_3+(1-R_1)R_2R_3+R_1(1-R_2)R_3+R_1R_2(1-R_3)$$

当各元件相同时，即 $R_1=R_2=R_3$，则

$$\theta_s=\int_0^{\infty}R_s\mathrm{d}t=\int_0^{\infty}(3\mathrm{e}^{-2\lambda t}-2\mathrm{e}^{-\lambda t})\mathrm{d}t=\frac{3}{2\lambda}-\frac{2}{3\lambda}=\frac{5}{6\lambda}$$

(5) 串并联系统的可靠性计算。串并联系统是一种串联系统和并联系统组合起来的系统。如图 4-25(a)所示为一复杂的并联系统，其处理办法如下：

① 先求出串联单元 3、4 和 5、6 两个系统 R_{34}、R_{56}的可靠度分别为

$$R_{34}=R_3R_4$$

$$R_{56}=R_5R_6$$

② 求出 R_{34}和 R_{56}以及 7 和 8 并联的子系统的可靠度分别为

$$R_{3456}=[1-(1-R_{34})(1-R_{56})]$$

$$R_{78}=[1-(1-R_7)(1-R_8)]$$

③ 于是得到一个等效串联系统，如图 4-25(c)所示，其可靠度为

$$R_s=R_1R_2R_{3456}R_{78}$$

$$=R_1R_2[1-(1-R_{34})(1-R_{56})][1-(1-R_7)(1-R_8)]$$

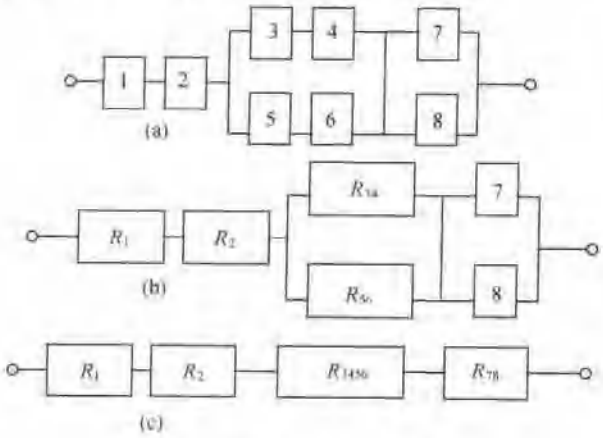


图 4-25　串并联系统的简化

3. 布尔真值表法

有很多复杂的系统不能简化为下面所述的数学模型加以计算，而只能分析其成功和失效的各种状态，然后加以计算。

例如，图 4-26 表示一复杂系统。元件 A 可以通到 C_1 和 C_2，但由 B_1 到 C_1 或由 B_2 到 C_2 是没有路的。这种系统可靠度的计算虽有几种方法，但可靠的方法是运用布尔真值表的方法。

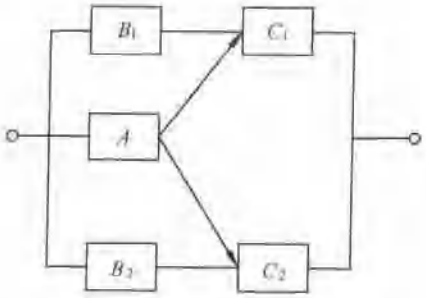


图 4-26　特殊的冗余系统

该系统有 A、B_1、B_2、C_1、C_2 五个元件，每个元件都有“正常”和“故障”两种状态，因此，该系统的状态有 $2^5=32$ 种。利用表 4-10 对这 32 种

状态进行全面调查，将该系统正常的概率全部加起来即可。

系统的状态号码是从 1 到 32。五个元件下面的数字 0 和 1 对应此元件的“故障”和“正常”状态(即 0 为故障，1 为正常)。状态号码为 1 时，各元件为 0，全系统属于故障状态，故正常或故障项下记入 F(故障)。状态号码 2、3 时只有一个元件是 1，其他都不正常，依次记入 F。在状态号码 4 时，C_2 和 A 是 1，由图 4-26 看出，系统是正常的，记入 S(正常)。这样，在 32 行中都有 F 或 S 的记载，因此只计算 S 所处的行就行了。例如，在第 4 行中，$B_1=0$，$B_2=0$，$C_1=0$，$C_2=1$，$A=1$，使对应于 0 的状态为 $1-R_i$，对应于 1 的状态为 R_i，则 $R_{s4}=(1-R_{B1})\times(1-R_{B2})(1-R_{C1})R_{C2}R_A$，将其计算结果计入 R_{si} 栏内，表中是以 $R_A=0.9$，$R_{B1}=R_{B2}=0.85$，$R_{C1}=R_{C2}=0.8$ 来计算的。求各 R_s 的总和，即得系统的可靠度 R_s。

表 4-10 布尔真值表(2^5 时)

系统状态序号	单元及其工作状态					系统状态正常或故障	正常概率 R_{si}
	B_1	B_2	C_1	C_2	A		
1	0	0	0	0	0	F	—
2	0	0	0	0	1	F	—
3	0	0	0	1	0	F	—
4	0	0	0	1	1	S	0.00324
5	0	0	1	0	0	F	—
6	0	0	1	0	1	S	0.00324
7	0	0	1	1	0	F	—
8	0	0	1	1	1	S	0.01296
9	0	1	0	0	0	F	—
10	0	1	0	0	1	F	—
11	0	1	0	1	0	S	0.00204
12	0	1	0	1	1	S	0.01836
13	0	1	1	0	0	F	—
14	0	1	1	0	1	S	0.01836
15	0	1	1	1	0	S	0.00816
16	0	1	1	1	1	S	0.07344
17	1	0	0	0	0	F	—
18	1	0	0	0	1	F	—
19	1	0	0	1	0	F	—
20	1	0	0	1	1	S	0.01836
21	1	0	1	0	0	S	0.00204
22	1	0	1	0	1	S	0.01836
23	1	0	1	1	0	S	0.00816
24	1	0	1	1	1	S	0.07344
25	1	1	0	0	0	F	—
26	1	1	0	0	1	F	—
27	1	1	0	1	0	S	0.01156
28	1	1	0	1	1	S	0.10404
29	1	1	1	0	0	S	0.01156
30	1	1	1	0	1	S	0.10404
31	1	1	1	1	0	S	0.04624
32	1	1	1	1	1	S	0.41816

$$\sum R_s = 0.95376$$

$$R_s = \sum_{i=1}^{n} R_s = 0.95376$$

4. 卡诺图法

卡诺图法是在布尔真值表的基础上，应用概率图将处于S状态的各行列分组隔开，得出计算系统可靠度的简化方法。

当系统由 n 个元件组成时，系统的所有可能状态有 2^n 个，我们可以把这 2^n 个状态分别用 n 位的二进制数来表示，每位二进制数代表一个元件所处的状态(0表示故障状态，1表示正常状态)，绘成如图4-27所示的概率图。图中每个方格代表系统的一个状态。如 $n=5$ 中，标有"*"号的方格为01110，即表示 $\bar{A}BCD\bar{E}$ 这样一个事件(系统的一个状态)。在构造图4-27概率图时，要求表头相邻两格的二进制数仅在一位上有差别。因此，图中相邻方格的二进制表示只在一个二进制位上有差别。如 $n=5$ 中，标有"*"号和"△"号的相邻方格，它们的二进制表示分别为01110、11110。这样编排的表，使我们比较容易地从图中求出"系统正常"这一事件的不交和。

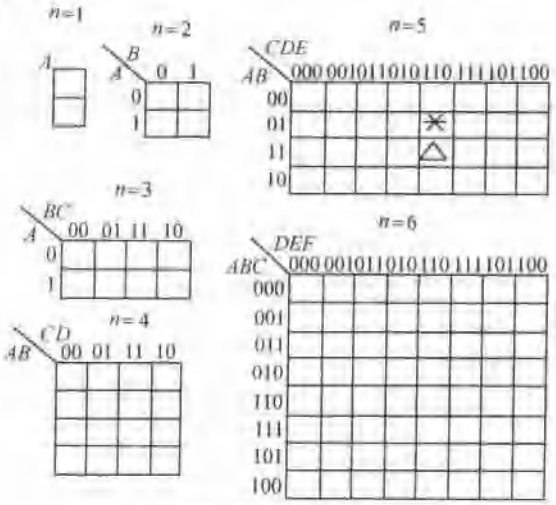


图4-27　概率图

有了上述概率图，可以将布尔真值表中处于S状态的各行转移到概率图的相应方格中，标以"*"号。然后将这些标有"*"号的方格，按相邻的行列分成若干组，并用方框隔开，即可得到图4-28所示的卡诺图。图中各组分别代表系统处于S状态时的相应概率，将各组的概率值相加，即可求得系统可靠度。

现仍以图4-26为例。此时由表4-7可以画出卡诺图4-28。由此图可见，系统正常这一事件 S 可以表示为下列不交和

$$S=\bar{C}_1C_2A+B_2C_2\bar{A}+B_1\bar{B}_2C_1C_2\bar{A}+C_1A+B_1C_1\bar{C}_2\bar{A}$$

故系统可靠度为

$$R_s=F_{C1}R_{C2}R_A+R_{B2}R_{C2}F_A+R_{B1}F_{B2}R_{C1}R_{C2}F_A+R_{C1}R_A+R_{B1}R_{C1}F_{C2}F_A$$

将各元件的可靠度 $R_A=0.9$，$R_{B1}=R_{B2}=0.85$，$R_{C1}=R_{C2}=0.8$ 代入上式得

$$R_s=0.95376$$

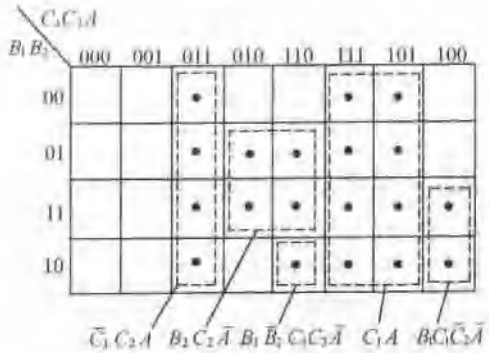


图 4-28 对应于图 4-25 系统的卡诺图

由于概率图中划分方块的方式可以不同，故 S 可以有不同的表达式，但其结果是相同的。由上述可知，当系统的元件数 n 不太大时（一般小于 6 或 7），布尔真值表法与卡诺图都能比较直观地求得系统的可靠度。特别是卡诺图法，它把系统正常这一事件表示成比较简单的不交事件和，使计算得到简化。

5. 条件概率法

对于不能直接用串并联法计算的复合系统，可先将其按一定条件分解成可用串并联法计算的几个子系统，分别算出各子系统的可靠度，然后按条件概率原理，用下列公式将各子系统综合起来，求出原系统的可靠度

$$R_s = R_{s(x好)} R_x + R_{s(x坏)} F_x \tag{4-75}$$

式中，$R_{s(x好)}$——元件 x 好条件下系统的条件可靠度；

$R_{s(x坏)}$——元件 x 坏条件下系统的条件可靠度；

R_x——元件 x 的可靠度；

F_x——元件 x 的失效概率。

例如，在图 4-29 中根据元件 E 是好或是坏为条件，将原系统分解成两个子系统。

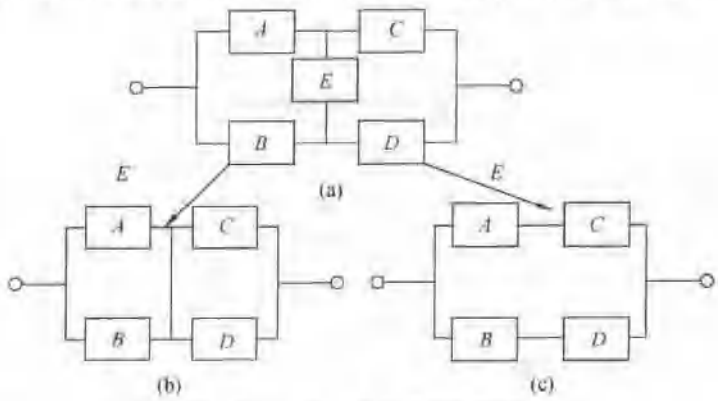


图 4-29 桥式网络系统分解图

（1）以 E 好为条件，则系统的条件可靠度为

$$R_{s(E好)} = (1 - F_A F_B)(1 - F_C F_D)$$

（2）以 E 坏为条件，则系统的条件可靠度为

$$R_{s(E)}=1-(1-R_AR_C)(1-R_BR_D)$$

按式(4-75)可求得该系统的可靠度为

$$\begin{aligned}R_s&=R_{s(E)}R_{(E)}+R_{s(E)}F_E\\&=(1-F_AF_B)(1-F_CF_D)R_E+[1-(1-R_AR_C)(1-R_BR_D)]F_E\end{aligned}$$

4.6.2　可靠性分配

可靠性分配是将设计任务书上规定的系统可靠度指标，合理地分配给系统各单元的一种设计方法。

分配的主要目的是，确定每个单元合理的可靠度指标，作为单元设计的一个重要参数。

可靠性分配的方法有多种，常用的有平均分配法、加权分配法、代数分配法、拉格朗日乘数分配法等。下面主要讨论平均分配法和再分配法。

1. **平均分配法**

对系统中的全部单元分配以相等的可靠度的方法称为“等分配法”或“等同分配法”。

(1) 串联系统可靠度分配。当系统中，n 个单元具有近似的复杂程度、重要性以及制造成本时，则可用等分配法分配系统各单元的可靠度。这种分配法的另一出发点是考虑到串联系统的可靠度往往取决于系统中的最弱单元，因此，对其他单元分配以高的可靠度无实际意义。

当系统的可靠度为 R_s，各单元分配的可靠度为 R_i 时，系统可靠度 R_s 为

$$R_s=\prod_{i=1}^{n}R_i=R_i^n \tag{4-76}$$

因此，单元的可靠 R_i 为

$$R_i=(R_s)^{\frac{1}{n}}\quad(i=1,2,\cdots,n) \tag{4-77}$$

(2) 并联系统可靠度分配。当系统的可靠度指标要求很高(如 $R_s>0.99$)而选用已有的单元又不能满足要求时，则可选用 n 个相同单元的并联系统，这时单元的可靠度 R_i 可大大低于系统的可靠度 R_s，则

$$R_s=1-(1-R_i)^n$$

故单元的可靠度 R_i 应分配为

$$R_i=1-(1-R_s)^{\frac{1}{n}}\quad(i=1,2,\cdots,n) \tag{4-78}$$

(3) 串并联系统可靠度分配。利用等分配法对串并联系统进行可靠性分配时，可先将串并联系统化简为“等效串联系统”和“等效单元”，再给同级等效单元分配以相同的可靠度。

例如，对于图 4-30(a)所示的串并联系统作两步化简后，则可先从最后的等效串联系统图 4-30(c)开始按等分配法对各单元分配可靠度：

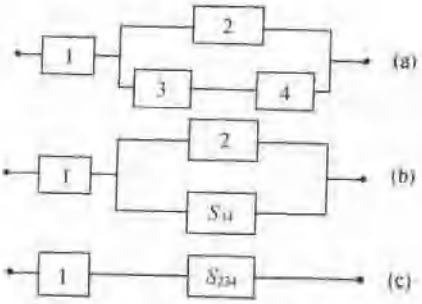
$$R_1=R_{s234}=R_s^{1/2}$$

再由图 4-30(b)分得

$$R_2=R_{s34}=1-(1-R_{s234})^{1/2}$$

最后再求图 4-30(a)中的 R_3 及 R_4：

$$R_3=R_4=R_{34}^{1/2}$$



(a) 串并联系统 (b) 中间等效系统 (c) 等效系统

图 4-30 串并联系统的可靠性分配

2. 再分配法

如果已知串联系统(或串并联系统的等效串联系统)各单元的可靠度预测值为$\hat{R}_1$，$\hat{R}_2$，…，$\hat{R}_n$，则系统的可靠度预测值为

$$\hat{R}_s = \prod_{i=1}^{n} \hat{R}_i \quad (i = 1,2,\cdots,n)$$

若设计规定的系统可靠度指标$R_s > \hat{R}_s$，表示预测值不能满足要求，需改进单元的可靠度指标并按规定的R_i值作再分配计算。显然，提高低可靠性单元的可靠度，效果要好些且容易些，因此，可提高低可靠性单元的可靠度并按等分配法进行再分配。为此，先将各单元的可靠度预测值按由小到大的次序排列，则有

$$\hat{R}_1 < \hat{R}_2 < \cdots < \hat{R}_m < \hat{R}_{m+1} < \cdots \hat{R}_n$$

令

$$R_1 = R_2 = \cdots = R_m = R_0 \tag{4-79}$$

并找出m值使

$$\hat{R}_m < R_0 = \left(\frac{R_s}{\prod_{i=m+1}^{n} \hat{R}_i} \right)^{\frac{1}{m}} < \hat{R}_{m-1} \tag{4-80}$$

则单元可靠度的再分配可按下式进行：

$$R_1 = R_2 = \cdots = R_m = \left(\frac{R_s}{\prod_{i=m+1}^{n} \hat{R}_i} \right)^{\frac{1}{m}} \tag{4-81}$$

$$R_{m+1} = \hat{R}_{m+1}, R_{m+2} = \hat{R}_{m+2}, \cdots, R_n = \hat{R}_n$$

例 4-14 设串联系统四个单元的可靠度预测值由小到大的排列为$\hat{R}_1 = 0.9507$，$\hat{R}_2 = 0.9570$，$\hat{R}_3 = 0.9856$，$\hat{R}_4 = 0.9998$。若设计规定串联系统的可靠度$R_s = 0.9560$，试进行可靠度再分配。

解 由于系统的可靠性预测值($\hat{R}_s = 0.8965$)不能满足设计指标，因此需提高单元的可靠度，并进行可靠度再分配。

设 $m=1$，则由式(4-80)得

$$R_0=\left(\frac{R_s}{\hat{R}_2\ \hat{R}_3\ \hat{R}_4}\right)^{\frac{1}{1}}=\left(\frac{0.9560}{0.9570\times0.9856\times0.9998}\right)^{1}=1.0138>\hat{R}_2$$

因此需另设 m 值。

设 $m=2$，则有

$$R_0=\left(\frac{R_s}{\hat{R}_3\ \hat{R}_4}\right)^{\frac{1}{2}}=\left(\frac{0.9560}{0.9856\times0.9998}\right)^{\frac{1}{2}}=0.9850$$

$$\hat{R}_2=0.9570<R_0=0.9850<\hat{R}_3=0.9856$$

因此，分配有效，再分配的结果为

$$R_1=R_2=0.9850,\ R_3=\hat{R}_3=0.9856,\ R_4=\hat{R}_4=0.9998$$

4.6.3　系统可靠性最优化

系统可靠性最优化是指利用最优化方法去解决系统的可靠性问题，又称为可靠性最优化设计。例如，在满足系统最低限度可靠性要求的同时使系统的“费用”为最小，或者在满足每个单元或子系统的可靠性最低限度要求的同时，使系统的费用为最小；通过对单元或子系统可靠度值的优化分配使系统的可靠度最大；通过合理设置单元或子系统的冗余部件使系统可靠度最大，等等。这里的所谓“费用”，不仅指为提高系统可靠度所需要的花费，除价格属于直接的花费外，单元或子系统质量或体积的大小都影响费用。下面仅就系统花费最小的最优化分配方法作一些介绍。

若串联系统 n 个单元的预计可靠度(现有可靠度水平)按非减序列排列为 $R_1,R_2,\cdots,R_n$，则系统的预计可靠度为

$$R_s=\prod_{i=1}^{n}R_i=R_i^n$$

如果要求的系统可靠度指标 $R_{sd}>R_s$，则系统中至少有一个单元的可靠度必须提高，即单元的分配可靠度 R_{id} 要大于单元的预计可靠度 R_i。为此，必须花费一定的研制开发费用。令 $G(R_i,R_{id})$，$i=1,2,\cdots,n$ 表示费用函数，即为使第 i 个单元的可靠度由 R_{id} 提高到 R_i 需要的花费总量。显然，$R_{id}-R_i$ 值愈大，即可靠度值提高的幅度愈大，则费用函数 $G(R_i,R_{id})$ 值也就愈大，费用也就愈高；另外，R_i 值愈大，则提高 $R_{id}-R_i$ 值所需的费用也愈高。

要使系统可靠度由 R_i 提高到 R_{id} 的总花费为 $\sum_{i=1}^{n}G(R_i,R_{id})$，$i=1,2,\cdots,n$。希望总花费为最小，于是构成一个最优化设计问题，其数学模型为

目标函数为
$$\min\sum_{i=1}^{n}G(R_i,R_{id}) \tag{4-82}$$

约束条件为
$$\prod_{i=1}^{n}R_{id}\geqslant R_{sd} \tag{4-83}$$

令 j 表示系统中需要提高可靠度的单元序号，显然应从可靠度最低的单元开始提高其

可靠度，即 j 从 1 开始，按需要可依次增大。

令
$$R_{0j}=\left(\frac{R_{sd}}{\prod_{i=j+1}^{n-1}R_i}\right)^{\frac{1}{j}}\quad j=1,2,\cdots,n \tag{4-84}$$

式中 $R_{n+1}=1$，则有

$$R_{0j}=\left(\frac{R_{sd}}{\prod_{i=j+1}^{n+1}R_i}\right)^{\frac{1}{j}}>R_j \tag{4-85}$$

上式表明，想要获得所要求的系统可靠度指标 R_{sd}，则 $j=1,2,\cdots,n$ 各单元的可靠度均应提高到 R_{0j}。若继续增大 j，当达到某子值（如 $j+1$）后使得

$$R_{0,j+1}=\left(\frac{R_{sd}}{\prod_{i=j+1}^{n+1}R_i}\right)^{\frac{1}{j+1}}>R_{j+1} \tag{4-86}$$

即第 $j+1$ 号单元的预计可靠度 R_{j+1} 已比提高到 $R_{0,j+1}$ 值为大，因此，j 为需要提高可靠度的单元的序号的最大值，今命它为 k_0，则说明：为使系统可靠度指标达到 R_{sd}，令 $j=k_0$，$i=1,2,\cdots,k_0$ 的各单元的分配可靠度 R_{sd} 均应提高到

$$R_{k0}=\left(\frac{R_{sd}}{\prod_{i=j-1}^{n+1}R_i}\right)^{\frac{1}{k_0}}>R_d \tag{4-87}$$

即序号为 $i=1,2,\cdots,k_0$ 的各单元的分配可靠度皆为 R_d，而序号为 $i=k_0+1,\cdots,n$ 的各单元的分配可靠度可各保持原预计可靠度值 $R_i(i=k_0+1,k_0+2,\cdots,n)$ 不变，即最优化问题的唯一最优解为

$$R_{id}=\begin{cases}R_d, & i\leqslant k_0\\ R_i, & i>k_0\end{cases} \tag{4-88}$$

提高有关单元的可靠度后，系统的可靠度指标为

$$R_{sd}=R_d^{k_0}\prod_{i=k_0+1}^{n+1}R_i \tag{4-89}$$

例 4-15 汽车驱动桥双级主减速器第一级螺旋锥齿轮主从动齿轮的预计可靠度为：$R_A=0.85$，$R_B=0.85$；第二级斜齿圆柱齿轮的预计可靠度为：$R_C=0.96$，$R_D=0.97$，若它们的费用函数相同，要求齿轮系统的可靠度指标为 $R_{sd}=0.80$，试用花费最小的原则对四个齿轮作可靠度分配。

解：(1) 系统的预计可靠度为

$$R_s=R_AR_BR_CR_D=0.85\times0.85\times0.96\times0.97=0.67279<0.8,$$

故应提高系统的可靠度，为此必须重新分配齿轮的可靠度。

(2) 将各单元（齿轮）的预计可靠度按非减顺序排列为

$$R_1=R_A=0.85,\ R_2=R_B=0.85,\ R_3=R_C=0.96,\ R_4=R_D=0.97$$

(3) 求 j 的最大值 k_0，由式(4-84)有

当 $j=1$ 时：

$$R_{01}=\left(\frac{R_{sd}}{\prod_{i=1+1}^{4+1}R_i}\right)^{\frac{1}{1}}=\left(\frac{0.80}{0.85\times0.96\times0.97\times1}\right)^{1}=1.01071>0.85=R_1$$

当 $j=2$ 时：

$$R_{02}=\left(\frac{R_{sd}}{\prod_{i=2+1}^{4+1}R_i}\right)^{\frac{1}{2}}=\left(\frac{0.80}{0.96\times0.97\times1}\right)^{\frac{1}{2}}=0.92688>0.85=R_2$$

当 $j=3$ 时：

$$R_{03}=\left(\frac{R_{sd}}{\prod_{i=3+1}^{4+1}R_i}\right)^{\frac{1}{3}}=\left[\frac{0.80}{0.97\times1}\right]^{\frac{1}{3}}=0.93779>0.96=R_3$$

因此，$k_0=2$，而以上各式中取 $R_i=1$。

(4) 由式(4-87)，得

$$R_d=\left(\frac{R_{sd}}{\prod_{i=k_0+1}^{4+1}R_i}\right)^{\frac{1}{k_0}}=\left(\frac{0.80}{0.96\times0.97\times1}\right)^{\frac{1}{2}}=0.92688$$

故四个齿轮的分配可靠度分别为

$R_{1d}=R_d=R_{Ad}=0.92688$　　$R_{2d}=R_d=R_{Bd}=0.92688$

$R_{3d}=R_3=R_{Cd}=0.96$　　$R_{4d}=R_4=R_{Dd}=0.97$

(5) 验算系统可靠度指标 R_{sd}，由式(4-89)：

$$R_{sd}=R_d^{k_0}\prod_{i=k_0+1}^{n+1}R_i=0.92688^2\times0.96\times0.97\times1=0.800000004>0.80$$ 满足要求。

习　题　4

1. 何为机械产品的可靠性？研究可靠性有何意义？

2. 何为可靠度？如何计算可靠度？

3. 何为失效率？如何计算？失效率与可靠度有何关系？

4. 可靠性分布有哪几种常用分布函数？试写出它们的表达式。

5. 试述浴盆曲线的失效规律和失效机理？如果产品的可靠性提高，那么，浴盆曲线将如何变化？

6. 可靠性设计与常规静强度设计有何不同？可靠性设计的出发点是什么？

7. 为什么按静强度设计法分析为安全零件，而按可靠性分析后会出现不安全的情况？试举例说明。

8. 已知零件受应力 $g(s)$ 作用，零件强度为 $f(r)$，如何计算该零件的强度安全可靠度。

9. 某机械零件承受的应力为服从正态分布的随机变量，其均值为 196MPa，标准差为 29.4MPa，该零件的强度也服从正态分布，其均值为 392MPa，标准偏差为 39.2MPa，求该零件的可靠度。

10. 有一方形截面的拉杆，它承受集中载荷的均值为 150kN，标准偏差为 1kN。拉杆材料的拉伸强度的均值为 800MPa，标准偏差为 20MPa，试求保证可靠度为 0.999 时杆件截面的最小边长（设公差为名义尺寸的 0.015 倍）。

第 5 章　弹性力学与有限元

为了使读者更好地学习掌握好有限元法，本章先介绍了与有限元法密切相关的弹性力学基础知识(小位移弹性理论的基本方程、小位移理论中的能量概念)，在此基础上介绍了有限元分析法等。

5.1　小位移弹性理论的基本方程

小位移弹性理论的主要任务是，分析受外力作用并处于平衡状态的弹性体内的应力、应变和位移状态及其相互关系等。这里不对小位移弹性理论作详细的分析，只给出一些基本方程并加以必要的说明，为后续内容做准备。

5.1.1　平衡微分方程

从受力平衡的弹性体内部取一六面微分体，它也应保持力和力矩平衡。如图 5-1 所示，微分体上各点的应力是坐标的函数，设 BCD、ACD、ABD 三个平面上的应力分量为

(1) 垂直于该面的正应力 σ_x(x 代表该面，即平行于 yOz 平面的法线方向)。

(2) 作用在该面的切应力。它分解为平行于坐标轴 y 的 τ_{xy} 和平行于坐标轴 z 的 τ_{xz}。

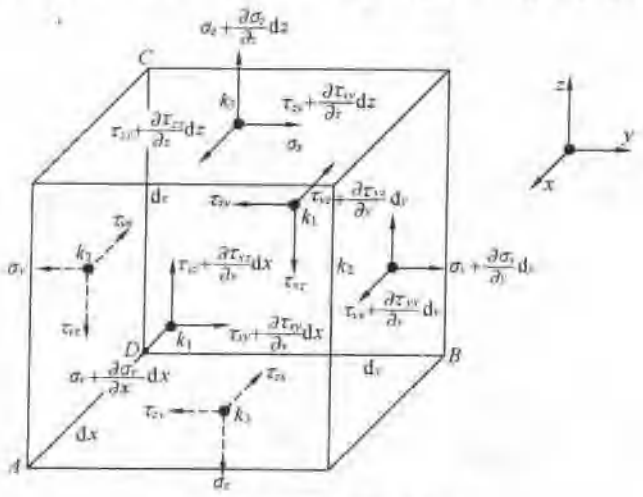


图 5-1　微分体平衡分析

考虑到通过弹性体中的一点总可以作出三个互相垂直的坐标平面，所以总共可以得到九个应力分量，它们是：σ_x、τ_{xy}、τ_{xz}、σ_y、τ_{yx}、τ_{yz}、σ_z、τ_{zx}、τ_{zy}。

应力的正负方向可以采用如下的符号规则：如果正应力是拉应力，则为正值，即正应力的方向是背离其作用面的；如果是朝向作用面的压应力，则为负值。这样，$A'B'C'D'$ 面上，正应力的正方向和坐标轴的方向一致，而在 $ABCD$ 面上，正应力的正方向则和坐标轴的方

向相反。切应力的符号和正应力的正、负方向有关：如果正应力的正方向和坐标轴的正方向相同，则和另外两个坐标轴的正方向相同的切应力规定为正的切应力；如果正应力的正方向和坐标轴方向相反，正的切应力方向也应该和其他两个坐标轴的方向相反。

弹性体中一点的九个应力分量是既有大小、方向，又有其作用面的向量。它们可以写成矩阵的形式

$$\boldsymbol{\sigma}=\begin{bmatrix}\sigma_x & \tau_{xy} & \tau_{xz}\\ \sigma_y & \tau_{yx} & \tau_{yz}\\ \sigma_z & \tau_{zx} & \tau_{zy}\end{bmatrix} \tag{5-1}$$

由 $\sum M_{k_1k_1}=0$，有

$$\tau_{yz}\mathrm{d}x\mathrm{d}z\frac{\mathrm{d}y}{2}+\left(\tau_{yz}+\frac{\partial\tau_{yz}}{\partial y}\mathrm{d}y\right)\mathrm{d}x\mathrm{d}z\frac{\mathrm{d}y}{2}=\tau_{zy}\mathrm{d}x\mathrm{d}y\frac{\mathrm{d}z}{2}+\left(\tau_{zy}+\frac{\partial\tau_{zy}}{\partial z}\mathrm{d}z\right)\mathrm{d}x\mathrm{d}y\frac{\mathrm{d}z}{2} \tag{5-2}$$

经化简得 $\tau_{yz}=\tau_{zy}$。同理由微分体另两轴线力矩平衡，得 $\tau_{zx}=\tau_{xz}$，$\tau_{xy}=\tau_{yx}$，即剪应力互等定律。

再由微分体在三个坐标轴方向上作用力分量之和为零，沿 x 轴有

$$\left(\sigma_x+\frac{\partial\sigma_x}{\partial x}\mathrm{d}x\right)\mathrm{d}y\mathrm{d}z+\left(\tau_{yx}+\frac{\partial\tau_{yx}}{\partial y}\mathrm{d}y\right)\mathrm{d}x\mathrm{d}z+\left(\tau_{zx}+\frac{\partial\tau_{zx}}{\partial z}\mathrm{d}z\right)\mathrm{d}x\mathrm{d}y+X\mathrm{d}x\mathrm{d}y\mathrm{d}z-\sigma_x\mathrm{d}y\mathrm{d}z-\tau_{yx}\mathrm{d}x\mathrm{d}z-\tau_{zx}\mathrm{d}x\mathrm{d}y=0 \tag{5-3}$$

y、z 轴同理建立方程，化简并应用剪应力互等定律可得

$$\begin{cases}\dfrac{\partial\sigma_x}{\partial x}+\dfrac{\partial\tau_{xy}}{\partial y}+\dfrac{\partial\tau_{xz}}{\partial z}+X=0\\ \dfrac{\partial\sigma_y}{\partial y}+\dfrac{\partial\tau_{xy}}{\partial x}+\dfrac{\partial\tau_{yz}}{\partial z}+Y=0\\ \dfrac{\partial\sigma_z}{\partial z}+\dfrac{\partial\tau_{zx}}{\partial x}+\dfrac{\partial\tau_{yz}}{\partial y}+Z=0\end{cases} \tag{5-4}$$

式(5-4)称为平衡微分方程，它表明微分体内力与外力的关系。

在一般三维应力状态的情况下，分别取作用在单元体上的力对 z、y 和 x 轴的力矩，便得

$$\begin{cases}\tau_{xy}=\tau_{yx}\\ \tau_{xz}=\tau_{zx}\\ \tau_{yz}=\tau_{zy}\end{cases} \tag{5-5}$$

这表明剪应力分量是两两相等的。

这样，式(5-1)是一个对称矩阵。因此，一点的应力状态用六个应力分量表示即可。在以后的讨论中，我们把这六个应力分量写成列矩阵的形式，即

$$\boldsymbol{\sigma}=\begin{bmatrix}\sigma_x\\ \sigma_y\\ \sigma_z\\ \tau_{xy}\\ \tau_{yz}\\ \tau_{zx}\end{bmatrix}\text{ 或 }\boldsymbol{\sigma}^{\mathrm{T}}=[\sigma_x\quad\sigma_y\quad\sigma_z\quad\tau_{xy}\quad\tau_{yz}\quad\tau_{zx}]$$

还可以证明，在三维应力状态的一般情况下，如果已知某一点的应力分量 $\boldsymbol{\sigma}$，则作用在任一平面上该点的应力分量可由下式表示

$$\begin{cases} X_N = l\sigma_x + m\tau_{yx} + n\tau_{zx} \\ Y_N = l\tau_{xy} + m\sigma_y + n\tau_{zy} \\ Z_N = l\tau_{xz} + m\tau_{yz} + n\sigma_z \end{cases} \tag{5-6}$$

式中 X_N、Y_N、Z_N 分别是作用在某一个任意平面上的沿 x、y、z 方向的应力分量。这个平面的法线为 N，且方向余弦为 $l=\cos(N,x)$、$m=\cos(N,y)$、$n=\cos(N,z)$，如图 5-2 所示。

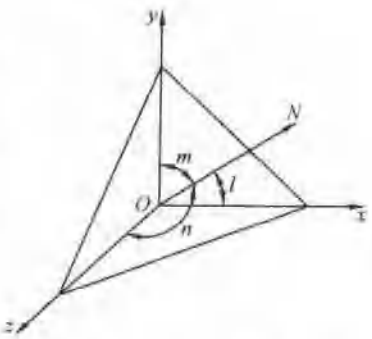


图 5-2　弹性体任意平面的几何表示法

5.1.2　几何方程

当物体中各点的相对位置改变时，则该物体即处于形变状态。令形变前，物体中任一点 A 的坐标为 (x,y,z)。形变后，该点在三个方向发生了位移 u、v、w，并且有新的坐标 $(x+u,y+v,z+w)$，如图 5-3 所示（图中未画出 z 方向坐标和位移）。一般来说，物体中的位移 u、v、w 是逐点变化的，所以 u、v、w 是 x、y、z 的函数。

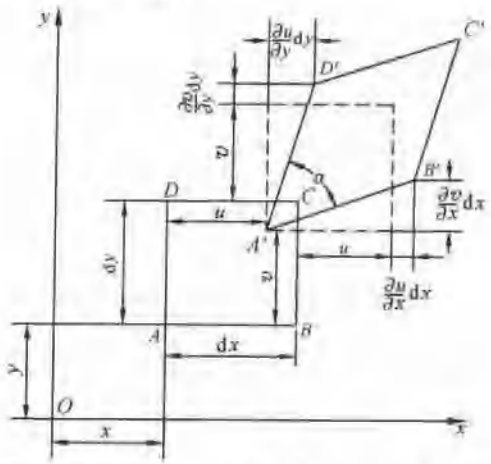


图 5-3　形变物体形变前后相应点的坐标变化

先研究二维的平面应变的情况。所谓平面应变，即原来在同一平面内的质点，形变后，仍位于同一平面内。若取 x 和 y 坐标轴在变形平面内，则 $w=0$，而 u、v 与 z 无关。

现在研究形变前物体中一个以 $\mathrm{d}x$ 和 $\mathrm{d}y$ 为边长的微小单元体 $ABCD$。形变后，它移到新位置 $A'B'C'D'$ 上，如图 5-3 所示。这样，就可以看到两种基本的几何变形，即一种是在某一方向上原来直线长度的变化称为线应变；另一种是所给定夹角的变化，称为切应变。

线应变用直线段的变化量和它的原长度之比 ε 代表；切应变用形变前原来为直角的角度改变量 γ 代表。

形变前，AB 的长度为 $\mathrm{d}x$；形变后，A 移到 A'。点 A 的位移在 x、y 方向的分量分别是 u、v，因为物体中的 u、v 是逐点变化的，所以把它们展成泰勒级数，略去高阶微量，则由点 B

到点 B' 的位移可写成 $u+\left(\frac{\partial u}{\partial x}\right)\mathrm{d}x$ 和 $v+\left(\frac{\partial v}{\partial x}\right)\mathrm{d}x$，则线段 $A'B'$ 在 x 轴上的投影为 $\mathrm{d}x+\left(\frac{\partial u}{\partial x}\right)\mathrm{d}x$，而在 y 轴上的投影为 $\frac{\partial v}{\partial x}\mathrm{d}x$。线段 $A'B'$ 长度的平方为

$$(\overline{A'B'})^2=\left(\mathrm{d}x+\frac{\partial u}{\partial x}\mathrm{d}x\right)^2+\left(\frac{\partial v}{\partial x}\mathrm{d}x\right)^2$$

x 方向上的线应变分量 ε_x 定义为单元体在 x 方向(形变前的方向)上的线应变。因此，得

$$\varepsilon_x=\frac{\overline{A'B'}-\overline{AB}}{\overline{AB}}$$

或

$$\overline{A'B'}=(1+\varepsilon_x)\overline{AB}=(1+\varepsilon_x)\mathrm{d}x$$

把它代入前面 $(\overline{A'B'})^2$ 的表达式中，并用 $(\mathrm{d}x)^2$ 除全式，则得

$$2\varepsilon_x+\varepsilon_x^2=2\frac{\partial u}{\partial x}+\left(\frac{\partial u}{\partial x}\right)^2+\left(\frac{\partial v}{\partial x}\right)^2$$

略去二阶微量，即得

$$\varepsilon_x=\frac{\partial u}{\partial x}$$

同理，y 方向上的线应变分量为

$$\varepsilon_x=\frac{\partial v}{\partial y}$$

为了求切应变，要研究原来是直角的角度变化。某点的切应变 γ_{xy} 可用该点上原来平行于 x、y 轴的两微小线段间的夹角的变化量表示。因此，点 A 的 γ_{xy} 等于线段 AB 和 AD 之间夹角的变化量。点 B' 在 y 方向上的位移等于 $v+\left(\frac{\partial v}{\partial x}\right)\mathrm{d}x$。而点 D' 在 x 方向上的位移等于 $u+\left(\frac{\partial u}{\partial y}\right)\mathrm{d}y$。

由于产生了位移，直线 AB 在形变后成为 $A'B'$，它和原来方向的微小转角为 $\frac{\partial v}{\partial x}$，而 $A'D'$ 与 AD 间的微小转角为 $\frac{\partial u}{\partial y}$。由此可见，线段 AB 与 AD 间的直角 DAB 改变了 $\frac{\partial u}{\partial y}+\frac{\partial v}{\partial x}$，所以

$$\gamma_{xy}=\frac{\partial u}{\partial y}+\frac{\partial v}{\partial x}$$

对于三维的一般情况，各应变分量可用上述方法求得，即

$$\begin{bmatrix}\varepsilon_x=\frac{\partial u}{\partial x}\\ \varepsilon_y=\frac{\partial v}{\partial y}\\ \varepsilon_z=\frac{\partial w}{\partial z}\\ \gamma_{xy}=\frac{\partial u}{\partial y}+\frac{\partial v}{\partial x}\\ \gamma_{yz}=\frac{\partial v}{\partial z}+\frac{\partial w}{\partial y}\\ \gamma_{zx}=\frac{\partial w}{\partial x}+\frac{\partial u}{\partial z}\end{bmatrix}=\begin{bmatrix}\frac{\partial u}{\partial x} & 0 & 0\\ 0 & \frac{\partial}{\partial y} & 0\\ 0 & 0 & \frac{\partial}{\partial z}\\ \frac{\partial}{\partial y} & \frac{\partial}{\partial x} & 0\\ 0 & \frac{\partial}{\partial z} & \frac{\partial}{\partial y}\\ \frac{\partial}{\partial z} & 0 & \frac{\partial}{\partial x}\end{bmatrix}=\begin{bmatrix}u\\ v\\ w\end{bmatrix} \tag{5-7}$$

上述这六个分量，称为应变分量。式(5-7)同时也给出位移 u、v、w 和应变分量之间的关系，即六个应变分量可以用三个位移分量表示，这种关系可称为几何关系。

5.1.3　应力和应变的关系

当弹性单元体上只作用有拉伸或压缩的应力 σ_x 时，则应力 σ_x 和应变 ε_x 成比例，其比值称为拉伸弹性模量，用 E 表示，即此时有(虎克定律)

$$\varepsilon_x=\frac{\sigma_x}{E}$$

大多数工程材料的拉伸弹性模量和压缩弹性模量相等，因而简称为弹性模量。当单元体在 x 方向拉伸时，在 y 和 z 两个方向必伴随着横向收缩。因此有

$$\varepsilon_y=\varepsilon_z=-\mu\varepsilon_x=-\mu\frac{\sigma_x}{E}$$

式中 μ 为泊松比。

切应力分量和其对应的切应变之比，称为剪切弹性模量，用 G 表示，即

$$G=\frac{\tau}{\gamma}$$

G 和 E 之间的关系是

$$G=\frac{E}{2(1+\mu)}$$

对于各向同性的材料，在三维情况下，应力和应变之间的关系可写成

$$\begin{cases}
\sigma_x=2G\left[\varepsilon_x+\dfrac{\mu}{1-2\mu}(\varepsilon_x+\varepsilon_y+\varepsilon_z)\right]=\dfrac{E}{1+\mu}\left(\dfrac{1-\mu}{1-2\mu}\varepsilon_x+\dfrac{\mu}{1-2\mu}\varepsilon_y+\dfrac{\mu}{1-2\mu}\varepsilon_z\right)\\
\sigma_y=2G\left[\varepsilon_y+\dfrac{\mu}{1-2\mu}(\varepsilon_x+\varepsilon_y+\varepsilon_z)\right]=\dfrac{E}{1+\mu}\left(\dfrac{\mu}{1-2\mu}\varepsilon_x+\dfrac{1-\mu}{1-2\mu}\varepsilon_y+\dfrac{\mu}{1-2\mu}\varepsilon_z\right)\\
\sigma_z=2G\left[\varepsilon_z+\dfrac{\mu}{1-2\mu}(\varepsilon_x+\varepsilon_y+\varepsilon_z)\right]=\dfrac{E}{1+\mu}\left(\dfrac{\mu}{1-2\mu}\varepsilon_x+\dfrac{\mu}{1-2\mu}\varepsilon_y+\dfrac{1-\mu}{1-2\mu}\varepsilon_z\right)\\
\tau_{xy}=G\gamma_{xy}=\dfrac{E}{2(1+\mu)}\gamma_{xy}\\
\tau_{yz}=G\gamma_{yz}=\dfrac{E}{2(1+\mu)}\gamma_{yz}\\
\tau_{zx}=G\gamma_{xy}=\dfrac{E}{2(1+\mu)}\gamma_{zx}
\end{cases}\tag{5-8}$$

或

$$\begin{cases}\varepsilon_x=\dfrac{1}{E}[\sigma_x-\mu(\sigma_y+\sigma_z)]\\ \varepsilon_y=\dfrac{1}{E}[\sigma_y-\mu(\sigma_z+\sigma_x)]\\ \varepsilon_z=\dfrac{1}{E}[\sigma_z-\mu(\sigma_x+\sigma_y)]\\ \gamma_{xy}=\dfrac{1}{G}\tau_{xy}\\ \gamma_{yz}=\dfrac{1}{G}\tau_{yz}\\ \gamma_{zx}=\dfrac{1}{G}\tau_{zx}\end{cases}\tag{5-9}$$

式(5-8)或(5-9)表示了三维情况下应力和应变之间的关系，称为广义虎克定律。这种关系又可称为物理关系。式(5-8)和式(5-9)可以写成矩阵形式

$$\boldsymbol{\sigma}=\frac{E}{(1+\mu)(1-2\mu)}\begin{bmatrix}1-\mu & \mu & \mu & 0 & 0 & 0\\ \mu & 1-\mu & \mu & 0 & 0 & 0\\ \mu & \mu & 1-\mu & 0 & 0 & 0\\ 0 & 0 & 0 & \dfrac{1-2\mu}{2} & 0 & 0\\ 0 & 0 & 0 & 0 & \dfrac{1-2\mu}{2} & 0\\ 0 & 0 & 0 & 0 & 0 & \dfrac{1-2\mu}{2}\end{bmatrix}\boldsymbol{\varepsilon}$$

或

$$\boldsymbol{\sigma}=\boldsymbol{D}\boldsymbol{\varepsilon}\tag{5-10}$$

$$\boldsymbol{\varepsilon}=\frac{1}{E}\begin{bmatrix}1 & -\mu & -\mu & 0 & 0 & 0\\ -\mu & 1 & -\mu & 0 & 0 & 0\\ -\mu & -\mu & 1 & 0 & 0 & 0\\ 0 & 0 & 0 & 2(1+\mu) & 0 & 0\\ 0 & 0 & 0 & 0 & 2(1+\mu) & 0\\ 0 & 0 & 0 & 0 & 0 & 2(1+\mu)\end{bmatrix}\boldsymbol{\sigma}$$

或

$$\boldsymbol{\varepsilon}=\boldsymbol{\Phi}\boldsymbol{\sigma}\tag{5-11}$$

式中 $\boldsymbol{\Phi}=\boldsymbol{D}^{-1}$；$\boldsymbol{D}$ 为弹性矩阵。

5.2 小位移弹性理论中的能量概念

当弹性体受到外力作用时，它就发生变形，因而外力在变形(位移)方向上对弹性体做功。如果不计弹性体在加载和卸载时能量的损失，即当结构系统是保守系统，则对这样的物体在变形时所做的功，可以看成是储存在物体中的能量，称为应变能。因此，应变能可以看成是弹性体变形时，它所吸收的能量。在讨论弹性系统的能量时，不仅要考虑外力所做的功(对应于系统位能的，称为外力位能)，还要考虑和变形相对应的应力(内力)所做的功，即应变能(对应于系统位能的，有时称为内力位能)。这样，当系统能量以位能表述时，则总位能

包括外力位能和应变能。

5.2.1　应变能

假定从弹性体中截出一个边长为 dx、dy、dz 的单元平行六面体，下面计算作用在单元体边界上的应力所做的功。

首先，假定此单元体上只作用有 σ_x，如图 5-4 所示。

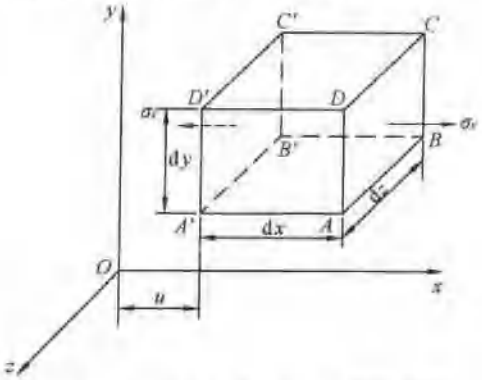


图 5-4　单元体受一维正应力作用

若 $A'B'C'D'$ 面在 x 方向的位移分量为 u，则 $ABCD$ 面在 x 方向的位移分量为 $u+\frac{\partial u}{\partial x}dx$。在 $A'B'C'D'$ 面上的力是 $\sigma_x dydz$，它的作用方向和 x 轴正方向相反，即和位移 u 的方向相反；而作用在 $ABCD$ 面上的力则和位移方向相同。变形过程中，σ_x 从零增加到某一数值 σ_x，位移 u 也是从零增加到 u 的。于是，作用在 $ABCD$ 面上的力和作用在 $A'B'C'D'$ 面上的力都做了功。所以，单元体所做的净功或储存于单元体中的能量为

$$dU=\int\sigma_x d\left(u+\frac{\partial u}{\partial x}dx\right)dydz-\int\sigma_x u\,dydz$$

或

$$dU=\int\sigma_x d\left(\frac{\partial u}{\partial x}\right)dxdydz$$

把 $\frac{\partial u}{\partial x}=\varepsilon_x=\frac{\sigma_x}{E}$ 代入，得

$$dU=\int_{\sigma_x=0}^{\sigma_x=\sigma_x}\frac{\sigma_x}{E}d\sigma_x dxdydz=\frac{\sigma_x^2}{2E}dxdydz=\frac{1}{2}\sigma_x\varepsilon_x dxdydz$$

现在，假设单元体受 σ_x 和 σ_y 两个应力作用，其作用次序为：首先，当 $\sigma_y=0$ 时，σ_x 由零增至 σ_x；然后，σ_x 保持不变，而 σ_y 由零增至 σ_y。

当 $\sigma_y=0$ 时，所做的功就是上面的结果，令其为 dU_1，即

$$dU_1=\frac{1}{2E}\sigma_x^2 dxdydz$$

当 σ_y 由零增至 σ_y 时，ε_y 也由零增至 $-\mu\frac{\sigma_x}{E}$。但此时 $\sigma_y=0$，所以对应于这部分的 ε_y 做的功等于零。

当 σ_x 由零增至 σ_y 时，产生相应的应变 $\varepsilon_y=\frac{\sigma_y}{E}$，它们产生的应变能为

$$dU_2=\frac{1}{2E}\sigma_y^2\,dx\,dy\,dz$$

同时，由于 σ_y 的作用，ε_x 由 $\frac{\sigma_x}{E}$ 变成 $\frac{\sigma_x}{E}-\mu\frac{\sigma_y}{E}$。但是当 ε_x 发生变化时，σ_x 是保持不变的。因此，σ_x 对 ε_x 这时并不做功，而仅对变化的 $-\mu\frac{\sigma_y}{E}$ 做功。所以，这时由 σ_x 这个常数值所做的功为

$$dU_3=\frac{\sigma_x dy dz(-\mu\sigma_y dx)}{E}=-\frac{\mu}{E}\sigma_x\sigma_y\,dx\,dy\,dz$$

这里没有 $\frac{1}{2}$ 是因为当 ε_x 变化时，σ_x 仍保持为常量。于是，储存在单元体中总的应变能为

$$dU=dU_1+dU_2+dU_3=\frac{1}{2E}(\sigma_x^2+\sigma_y^2-2\mu\sigma_x\sigma_y)\,dx\,dy\,dz$$

由于 $\varepsilon_x=\frac{1}{E}(\sigma_x-\mu\sigma_y)$，$\varepsilon_y=\frac{1}{E}(\sigma_y-\mu\sigma_x)$，则得

$$dU=\frac{1}{2}(\sigma_x\varepsilon_x+\sigma_y\varepsilon_y)\,dx\,dy\,dz$$

可见，应变能只和最终的应力状态有关，和应力作用的先后次序无关。

再来讨论在切应力 τ_{xy} 作用下的单元体。从图 5-5 看出，作用在单元体边界上的力等于 $\tau_{xy}\,dx\,dz$。在此力作用方向上的位移为 $\gamma_{xy}\,dy$。于是，应变能为

$$dU=\frac{1}{2}(\tau_{xy}\,dx\,dz)(\gamma_{xy}\,dy)=\frac{1}{2}\tau_{xy}\gamma_{xy}\,dx\,dy\,dz$$

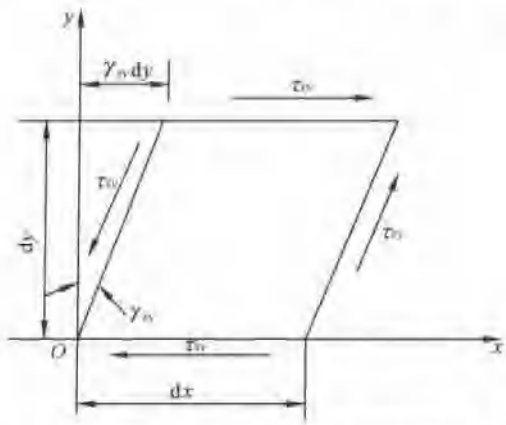


图 5-5 单元体在平面受切应力作用

由虎克定律可以看出，正应力将不产生任何切应变，而切应力也不引起任何线应变。所以，如果单元体 $dx\,dy\,dz$ 同时受到 σ_x、σ_y 和 τ_{xy} 的作用，则储存于单元体中总应变能为

$$dU=\frac{1}{2}(\sigma_x\varepsilon_x+\sigma_y\varepsilon_y+\tau_{xy}\gamma_{xy})\,dx\,dy\,dz$$

在一般的三维应力状态下，储存于单元体中的应变能可用同样的方法求得

$$dU=\frac{1}{2}(\sigma_x\varepsilon_x+\sigma_y\varepsilon_y+\sigma_z\varepsilon_z+\tau_{xy}\gamma_{xy}+\tau_{yz}\gamma_{yz}+\tau_{zx}\gamma_{zx})\,dV$$

$$\phi = U - W = \frac{1}{2}\int_V \boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{\sigma}\mathrm{d}V - \boldsymbol{F}^{\mathrm{T}}\boldsymbol{q} \tag{5-18}$$

应该特别指出的是：外力位能和机构加载过程中外力所做的功不仅符号不同，而且其绝对值也是不等的。加载过程外力$\{\boldsymbol{F}\}_i$的值自零逐渐增大到最终值。对于线性情况，它对位移q_i所做的功是$\frac{1}{2}\sum \boldsymbol{F}_i^{\mathrm{T}}\boldsymbol{q}_i = -\boldsymbol{F}\boldsymbol{q}^{\mathrm{T}}$。而外力位能却是其最终值从结构系统的最终位置恢复到参考状态时所做的功，其值是$-\sum \boldsymbol{F}_i^{\mathrm{T}}\boldsymbol{q}_i = -\boldsymbol{F}\boldsymbol{q}^{\mathrm{T}}$。据此，可以认为外力位能的值是不变力$\{\boldsymbol{F}\}_i$在位移$\{\boldsymbol{q}_i\}$上所做的功。在有些参考资料中，在写系统总位能时也有把用$-\sum \boldsymbol{F}_i^{\mathrm{T}}\boldsymbol{q}_i$表示的外力位能称为外力功。这时就应从上述含义上来理解。

为了便于应用位移法，可以根据式(5-15)改写成如下的形式

$$\begin{aligned} U(u,v,w) = \frac{1}{2}\int_V \Bigg\{ & \frac{\mu E}{(1+\mu)(1-2\mu)}\left(\frac{\partial u}{\partial x}+\frac{\partial v}{\partial y}+\frac{\partial w}{\partial z}\right)^2 + \\ & 2G\left[\left(\frac{\partial u}{\partial x}\right)^2+\left(\frac{\partial v}{\partial y}\right)^2+\left(\frac{\partial w}{\partial z}\right)^2\right]+ \\ & \frac{G}{2}\left[\left(\frac{\partial w}{\partial y}+\frac{\partial v}{\partial z}\right)^2+\left(\frac{\partial u}{\partial z}+\frac{\partial w}{\partial x}\right)^2+\left(\frac{\partial v}{\partial x}+\frac{\partial u}{\partial y}\right)^2\right]\mathrm{d}V \end{aligned} \tag{5-19}$$

这样，式(5-17)又可以简写为

$$\phi = U(u,v,w) - \boldsymbol{F}^{\mathrm{T}}\boldsymbol{q} \tag{5-20}$$

式中的位移u、v、w是满足S_u上几何边界条件的任意单值连续函数。

式(5-19)或式(5-20)表述了弹性体变形时的能量关系。对它应用变分方法可以导出能量原理中的一些重要原理。例如，有限元法中很有用的最小位能原理和最小余能原理等。

5.3 有限元分析法概述

5.3.1 有限元分析法的基本概念

在机械设计中，对工程结构进行强度、刚度和稳定性分析，所依据的理论主要是材料力学，结构力学和弹、塑性力学等。随着生产的发展，新材料不断出现，工程结构的形状和载荷日益复杂，原有的理论体系逐渐受到挑战，对一些问题传统的解析解几乎无法解决。因此在20世纪初，人们开始搜索近似解法。

对于通常的力学问题或场问题，一般可以建立它们所遵循的基本方程(即常微分方程或偏微分方程)和相应的边界条件，利用有限差分法或变分法求出近似解。有限差分法其实质是将由经离散化建立的相应的差分方程组来代替由物理模型建立的微分方程及其相应的边界条件，求得近似数值解。变分法本是研究泛函极值问题的一种方法。由变分原理可知，微分方程的边值问题的解等价于相应泛函极值问题的解，因此可以将微分方程的边值问题转化为泛函的变分问题求解。

当遇到几何形状复杂、边界条件复杂时，上述方法求解会发生困难，解的精度较低。随着计算机应用技术的发展，进一步的数值近似计算方法——有限元法出现了。有限元法的基本思想是“先分后合”，即将连续体或结构先人为地分割成许多单元，并认为单元与单元之

间只通过节点相联结，力也只通过节点作用，如图 5-6 所示。在此基础上，根据分片近似的思想，假定单元位移函数，利用力学原理推导建立每个单元的平衡方程组；再将所有单元的方程组组集成表示整个结构力学特性的代数方程组，并引入边界条件求解。

图 5-6　网格划分

从数学的角度来看，有限元分析法是将一个偏微分方程化成一个代数方程组，利用计算机求解的方法。在计算机出现以前，求解微分方程的方法大都表示为级数展开的形式，通称为解析法，它们只能求解极其简单的微分方程和极其规则的区域问题。但是，绝大多数实际的微分方程的求解问题，方程和区域都是十分复杂的，因此解析法远远不能满足实际的需要。有限元分析法原则上可以求解任何复杂的偏微分方程和任何复杂的求解区域问题，它是一种使复杂工程解获得近似解的数值分析技术。有限元分析理论所涉及的内容非常广，如结构静力分析的有限元分析法、结构动力学问题的有限元分析法、温度场问题的有限元分析法、流体流动有限元分析法、电磁场有限元分析法等，但无论是何种有限元分析法，其基本思想不外乎都是把一个连续体人为地分割成有限个单元，即把一个复杂结构看成由若干通过节点相连的单元组成的整体，先进行单元分析，然后再把这些单元组合起来代表原来的结构。可以说，有限元分析法的实质就是先化整为零、再积零为整的方法；或先拆再搭的方法。

如图 5-7 所示的柱形曲面板梁组合结构，是由铝板弯成的曲面，曲面背后有纵向梁和横向梁加固。这样一个结构可以用许多梁单元和壳块单元的组合来代替，这些单元是工程技术人员都很熟悉的构件，其力学性质简单明了。对每个单元加以分析，只要用有限个参数（如单元端点处的轴力、弯矩、位移、转角、……）即可加以描述，而整个结构又是由有限个单元组合而成的。这是一种从部分到整体的方法，在“一分一合”、“拆了再搭”的过程中，把复杂的结构计算问题转化为简单构件的分析与综合问题，因而使分析大为简化。

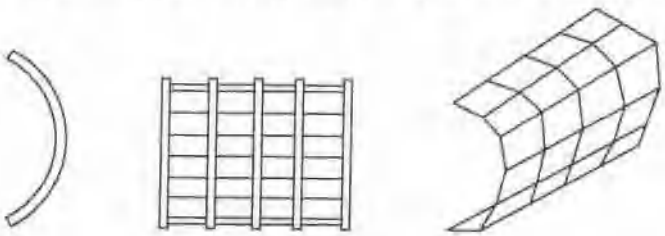


图 5-7　柱形曲面板梁组合结构

对于有限元分析法的分析过程可概括如下：

1. 连续体的离散化

连续体是指所分析的工程系统（物件或结构），离散化是将拟分析的连续体分割成由有

限个单元组成的有限元模型。如图 5-8(a)所示的桁架，这种分割十分明显，可以取每根杆作为一个单元。由于这种单元是细长杆，只沿轴线方向有拉压变形，故称之为一维单元。对于二维连续体，单元的形状可以是三角形、四边形，如图 5-8(b)所示 L 形薄板，可被分割为 32 个小三角平板，当这种单元仅受到板平面内的载荷时，可称之为二维单元。对于三维问题，单元可以是四面体、六面体，如图 5-8(c)所示。在同一个问题中，可以应用不同类型的单元。划分单元是建立有限元模型的一个重要步骤，需要工程经验。

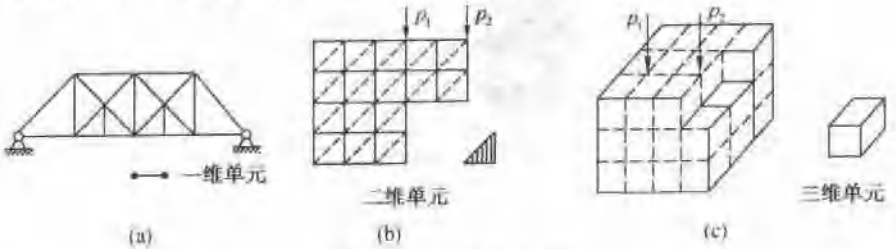


图 5-8 结构与单元

2. 分析单元的特性，建立单元刚度矩阵

在建立有限元模型时，首先要选择单元节点处的基本未知量。在结构有限元分析法中，基本未知量可以选择节点位移，也可以选择节点力。本书主要介绍在应用上最为普遍的有限元位移法(或称为有限元刚度法)，即选择节点的位移作为基本未知量，称为位移模式。有限元模型一经建立(即选择好单元和它们的位移模式)，就可以进行单元的力学特性分析，并导出单元的刚度矩阵。刚度将节点处的位移(节点位移)同节点处施加的力(节点力)联系起来。单元刚度矩阵取决于位移模式、单元的形状尺寸、单元的材料性质和构成关系。

3. 组成结构和总体刚度方程

这个过程包括将各个单元的刚度矩阵集成为总体刚度矩阵，以及将各单元的节点力向量集成总的力向量。集成过程所依据的原理是节点变形协调条件和平衡条件。

4. 求解方程组

由上述所形成的总体刚度方程是一组线性代数方程，在求解之前，必须考虑问题的边界条件，适当修改这些方程，以进行求解，即可解出各节点的位移。

5. 计算单元的内力、应力及应变

在某些情况下，节点位移就是工程分析所需要的未知量，但通常还需要由位移计算导出其他量，如内力、应力、应变等，最后可将计算结果进行整理或绘制成必要的表格或图形，作为工程设计的依据。

5.3.2 杆系结构的有限元分析

在本节中，我们将以结构静力学中的杆系结构的有限元分析为例，说明有限元分析法的分析思路和求解步骤，以帮助我们掌握有限元分析法的基本原理，并应用于工程实践中。

杆系结构包括桁架结构、刚架结构及其组合。杆系结构由杆件构成。桁架结构的杆件之间用铰链联结，所受外力都集中作用在铰接点上，杆件只受沿轴向的拉力或压力，称为杆单元；刚架结构的杆件之间为刚性固结，杆件除受有轴向力外，还有剪力和弯矩，一般把这一类杆件称为梁。因此杆系结构被离散化为由有限个杆单元或梁单元组成。有时也可把一个

杆件分成数个单元，如对于变截面杆件的处理。确定单元是进行有限元分析的第一步。

桁架结构是工程中经常采用的结构。本书中我们选取研究平面桁架结构来说明杆系结构的有限元分析方法，其概念、方法同样也适用于刚架结构及其他形式的结构。

杆件之间的连接处称为节点。节点一定具有某种程度的自由度，以表示该工程系统受到外力后的反应结果。以三维空间结构系统而言，节点的自由度含三个方向位移变形及三个方向角变形。在杆系结构有限元分析中，整个结构的变形状态用各节点位移来表示，节点位移就是杆单元端点的线位移，如对于平面桁架，在每一节点处有 x 和 y 方向两个线位移分量。对于梁单元还包括角位移(转角)。

在有限元分析法中，结构的内力状态可由各节点力来表示，结构的内力都是通过节点进行传递的。节点力包括外界加于结构的载荷、支座点反力及结构各元件相互间的作用力。桁架结构的外载荷一般是作用在节点上的，并由节点将载荷传至整个结构；而对于刚架则可能有分布载荷，或非节点处的集中载荷，这时需要运用静力等效的原理，使其等效为作用在节点上的集中力和力矩。

1. *杆单元的刚度及应力矩阵*

由于有限元分析法把整个结构看成是单元的组合体。为此，首先讨论单元刚度矩阵的建立。

(1) 杆单元刚度矩阵。

对于平面桁架结构，图 5-9 给出处在 xOy 平面内任意位置的一个杆单元 e，用 A、E 和 L 分别表示它的横截面积、弹性模量和单元长度。单元的两个节点编号分别为 i 和 j，用 θ 表示杆单元轴线与 x 坐标轴的夹角。

在平面内，对任意一根杆件变形后的位置需用节点位移 u_i、v_i、u_j、v_j 来确定(对于桁架节点没有角位移)，其中 u 代表 x 方向位移，v 代表 y 方向位移，因此它具有四个自由度。设结构在外载作用下产生变形如图 5-10 所示，杆单元的节点 j 在 x 方向产生一个位移，$u_j \neq 0$，而其他位移均为零，即 $u_i = v_i = v_j = 0$。这时杆单元比原来(图 5-9)伸长量为 ΔL，

$$\Delta L = u_j \cos\theta$$

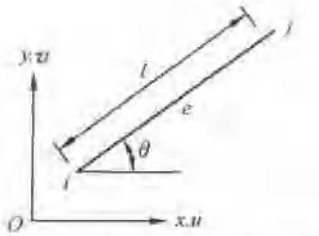


图 5-9　平面内的一个杆单元

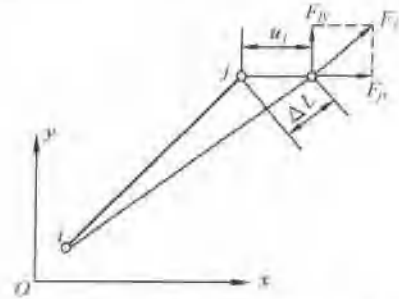


图 5-10　杆单元产生变形

由此产生的轴向力为

$$F_j = \frac{u_j \cos\theta}{L} EA$$

F_j 在 x 方向的分量为

$$F_{jx} = F_j \cos\theta = \frac{EA}{L} \cos^2\theta u_j$$

F_j 在 y 方向的分量为

$$F_{jy}=F_j\sin\theta=\frac{EA}{L}\cos\theta\sin\theta u_j$$

由杠杆的平衡可得

$$F_{ix}=-F_{jx}=-\frac{EA}{L}\cos^2\theta u_j$$

$$F_{iy}=-F_{jy}=-\frac{EA}{L}\cos\theta\sin\theta u_j$$

同样，如果设节点 j 在 y 方向产生一个位移 $v_j\neq0$，而其他位移均为零，即 $u_i=v_j=u_j=0$（图 5-11），将会得到另一组节点力。这时杆件比原来伸长了 $\Delta L=v_j\sin\theta$，由此产生的轴向力为

$$F_j=\frac{v_j\sin\theta}{L}EA$$

在 x 方向的分量为　$F_{jx}=F_j\cos\theta=\dfrac{EA}{L}\cos\theta\sin\theta v_j$

在 y 方向的分量为　$F_{jy}=F_j\sin\theta=\dfrac{EA}{L}\sin^2\theta v_j$

根据杠杆的平衡有　$F_{ix}=-F_{jx}=-\dfrac{EA}{L}\cos\theta\sin\theta v_j$

$$F_{iy}=-F_{jy}=-\frac{EA}{L}\sin^2\theta v_j$$

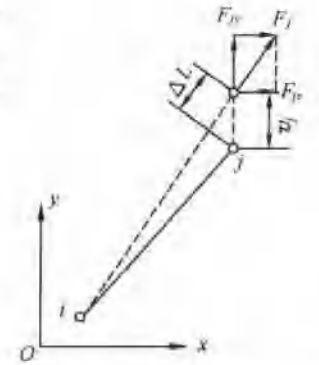


图 5-11　杆单元节点 j 在 y 方向产生位移

同上述关于节点 j 产生位移的讨论相类似，当节点 i 在 x 和 y 方向有位移时，即 $u_i\neq0$，其他位移为零；$v_i\neq0$，其他位移为零时，我们可分别求得杆的各节点力分量。根据叠加原理，将所得方程（共 16 个）都综合在一起有

$$\begin{cases}F_{ix}=\dfrac{EA}{L}(\cos^2\theta u_i+\cos\theta\sin\theta v_i-\cos^2\theta u_j-\cos\theta\sin\theta v_j)\\ F_{iy}=\dfrac{EA}{L}(\cos^2\theta\sin\theta u_i+\sin^2\theta v_i-\cos\theta\sin\theta u_j-\sin^2\theta v_j)\\ F_{jx}=\dfrac{EA}{L}(-\cos^2\theta u_i-\cos\theta\sin\theta v_i+\cos^2\theta u_j+\cos\theta\sin\theta v_j)\\ F_{jy}=\dfrac{EA}{L}(-\cos^2\theta\sin\theta u_i-\sin^2\theta v_i+\cos\theta\sin\theta u_j+\sin^2\theta v_j)\end{cases}$$

写成矩阵形式

$$\begin{bmatrix}F_{ix}\\F_{iy}\\F_{jx}\\F_{jy}\end{bmatrix}=\frac{EA}{L}\left[\begin{array}{cc:cc}\cos^2\theta&\cos\theta\sin\theta&-\cos^2\theta&-\cos\theta\sin\theta\\ \cos\theta\sin\theta&\sin^2\theta&-\cos\theta\sin\theta&-\sin^2\theta\\ \hdashline -\cos^2\theta&-\cos\theta\sin\theta&\cos^2\theta&\cos\theta\sin\theta\\ -\cos\theta\sin\theta&-\sin^2\theta&\cos\theta\sin\theta&\sin^2\theta\end{array}\right]\begin{bmatrix}u_i\\v_i\\u_j\\v_j\end{bmatrix}$$

或

$$\begin{bmatrix}F_{ix}\\F_{iy}\\F_{jx}\\F_{jy}\end{bmatrix}=\frac{EA}{L}\left[\begin{array}{cc:cc}l^2&lm&-l^2&-lm\\ lm&m^2&-lm&-m^2\\ \hdashline -l^2&-lm&l^2&lm\\ -lm&-m^2&lm&m^2\end{array}\right]\begin{bmatrix}u_i\\v_i\\u_j\\v_j\end{bmatrix}$$

上式中 $l=\cos\theta, m=\sin\theta$。

上式还可简写为

$$[\boldsymbol{F}]^e=[\boldsymbol{K}]^e[\delta]^e \tag{5-21}$$

其中，$[\boldsymbol{F}]^e$ 为单元节点力列阵；$[\delta]^e$ 为单元节点位移列阵；e 表示为杆单元的单元号。

式(5-21)为用单元节点位移 $[\delta]^e$ 来表示单元节点力的表达式，称为单元刚度方程。$[\boldsymbol{K}]^e$ 称为杆单元的刚度矩阵，它是一个对称矩阵。由于对任何一个平面杆单元刚度矩阵可写成

$$[\boldsymbol{K}]^e=\frac{EA}{L}\begin{array}{c}\begin{matrix}\quad i & \quad & \quad j & \quad\end{matrix}\\\left[\begin{array}{cc:cc}l^2 & lm & -l^2 & -lm\\ lm & m^2 & -lm & -m^2\\ \hdashline -l^2 & -lm & l^2 & lm\\ -lm & -m^2 & lm & m^2\end{array}\right]\begin{matrix}i\\ \\ j\\ \\ \end{matrix}\end{array}=\begin{bmatrix}[\boldsymbol{K}_{ii}^e] & [\boldsymbol{K}_{ij}^e]\\ [\boldsymbol{K}_{ji}^e] & [\boldsymbol{K}_{jj}^e]\end{bmatrix} \tag{5-22}$$

这里，上标表示杆单元，下标 i 和 j 表示杆单元的两个节点号，这些子块有下述关系

$$[\boldsymbol{K}_{ii}^e]=[\boldsymbol{K}_{jj}^e]=-[\boldsymbol{K}_{ij}^e]=-[\boldsymbol{K}_{ji}^e]=\frac{EA}{L}\begin{bmatrix}l^2 & lm\\ lm & m^2\end{bmatrix}$$

于是式(5-21)可写成

$$\begin{bmatrix}[\boldsymbol{F}_i^e]\\ [\boldsymbol{F}_j^e]\end{bmatrix}=\begin{bmatrix}[\boldsymbol{K}_{ii}^e] & [\boldsymbol{K}_{ij}^e]\\ [\boldsymbol{K}_{ji}^e] & [\boldsymbol{K}_{jj}^e]\end{bmatrix}\begin{bmatrix}[\delta_i^e]\\ [\delta_j^e]\end{bmatrix} \tag{5-23}$$

在上面分块矩阵形式的单元刚度方程中，$[\boldsymbol{F}_i^e]$和$[\delta_i^e]$为二维(对平面结构)或三维(对空间结构)列向量。对于平面杆单元而言，每个节点有两个自由度，因此其刚度矩阵中每一个子块$[\boldsymbol{K}_{ij}]$均为 2×2 阶的(对于空间杆单元则为 3×3 的)子块。

对于空间桁架中的杆单元刚度矩阵，可用同样的方法求得。对于空间节点来说，它有位移分量，即 $u、v、\omega$，与此相对应，它有三个轴向力分量 $F_x、F_y、F_z$。因此其单元刚度矩阵为 6×6 的方阵，其刚度矩阵表达式为

$$[\boldsymbol{K}]^e=\frac{EA}{L}\left[\begin{array}{ccc:ccc}l^2 & lm & ln & -l^2 & -lm & -ln\\ lm & m^2 & mn & -lm & -m^2 & -mn\\ ln & -mn & -n^2 & -ln & -mn & n^2\\ \hdashline -l^2 & -lm & -ln & l^2 & lm & ln\\ -lm & -m^2 & -mn & lm & m^2 & mn\\ -ln & -mn & -n^2 & ln & mn & n^2\end{array}\right]$$

式中 $l=\cos\theta_x、m=\cos\theta_y、n=\cos\theta_z$ 为杆件的方向余弦，它们的计算可按下式进行

$$\begin{cases}l=\cos\theta_x=\dfrac{x_j-x_i}{L}\\ m=\cos\theta_y=\dfrac{y_j-y_i}{L}\\ n=\cos\theta_z=\dfrac{z_j-z_i}{L}\\ L=[(x_j-x_i)^2+(y_j-y_i)^2+(z_j-z_i)^2]^{\frac{1}{2}}\end{cases}$$

式中 $\theta_x、\theta_y、\theta_z$ 分别表示杆轴线与坐标轴 $x、y、z$ 的夹角；$x_i、y_i、z_i$ 为杆端 i 点在总体坐标中的位置；$x_j、y_j、z_j$ 为杆端 j 点的坐标位置。

需要注意，所谓节点力，是广义力的概念，相应的节点位移也是广义位移的概念，因为对

于梁单元来说，节点力还包括弯矩、扭矩，而节点位移还包括转角。

节点位移的符号规定为：位移方向与坐标轴方向一致为正，反之为负，这里与材料力学中的规定不同。

从以上推导我们可以看出，对于任意一个桁架杆件，只要给出该杆的几何及物理特性，即截面积 A、弹性模量 E 及杆件两端点的节点坐标，便可以直接得出该杆在总体坐标下的单元刚度矩阵。

类似以上所介绍的这种导出单元刚度矩阵的方法，称为直接法。针对不同类型的单元建立刚度矩阵的方法还有：虚功原理法、能量变分原理法等。

(2) 杆单元应力矩阵

平面杆件内力 N 由节点力分量 F_{jx} 和 F_{jy} 表示为

$$N=F_{jx}\cos\theta_x+F_{jy}\cos\theta_y$$

杆件应力表达式为

$$\sigma=\frac{N}{A}=\frac{1}{A}(F_{jx}\cos\theta_x+F_{jy}\cos\theta_y)=\frac{1}{A}[\cos\theta_x \quad \cos\theta_y]\begin{bmatrix}F_{jx}\\F_{jy}\end{bmatrix}$$

$$=\frac{E}{L}[\cos\theta_x\cos\theta_y]\begin{bmatrix}-\cos^2\theta_x & -\cos\theta_x\cos\theta_y & \cos^2\theta_x & \cos\theta_x\cos\theta_y\\ -\cos\theta_y\cos\theta_y & -\cos^2\theta_y & \cos\theta_x\cos\theta_y & \cos^2\theta_y\end{bmatrix}[\boldsymbol{\delta}]^e$$

或写为

$$\sigma=[S][\boldsymbol{\delta}]^e \tag{5-24}$$

其中

$$[S]=\frac{E}{L}[-\cos\theta_x \quad -\cos\theta_y \quad \cos\theta_x \quad \cos\theta_y] \tag{5-25}$$

式(5-25)称为平面杆单元应力矩阵。式(5-24)是以杆件两端节点位移表示的杆件应力，而杆件内力表达式为

$$N=\sigma A=\frac{EA}{L}[-\cos\theta_x \quad -\cos\theta_y \quad \cos\theta_x \quad \cos\theta_y][\boldsymbol{\delta}]^e$$

同理，不难推得空间桁架结构杆单元的应力矩阵

$$[S]=\frac{E}{L}[-\cos\theta_x \quad -\cos\theta_y \quad -\cos\theta_z \quad \cos\theta_x \quad \cos\theta_y \quad \cos\theta_z]$$

以上，内力 N 以轴向拉力为正，应力 σ 以拉应力为正。

2. 桁架结构整体刚度矩阵

前面讨论了如何建立杆单元的刚度矩阵。现在，需要将单元组合起来进行结构整体分析，即建立用节点位移表示的整个离散体系的平衡方程组。其中重要的工作是拼装结构整体刚度矩阵。结构整体刚度矩阵是基于下述两个基本原则建立的。

结构的各节点必须满足连续性条件。整个结构受载变形后，各单元仍必须在各节点处协调地联系在一起。例如，当有 n 个单元在某个节点 i 处相联，则这 n 个单元在该节点处都必须具有相同的节点位移，即

$$[\boldsymbol{\delta}_i^1]=[\boldsymbol{\delta}_i^2]=\cdots=[\boldsymbol{\delta}_i^n]=[\boldsymbol{\delta}_i] \tag{5-26}$$

结构的各节点必须满足平衡条件。也就是说，对于结构的任一节点，汇集于该节点的所有单元作用于其上的节点力，应与施加于该节点上的载荷保持平衡，即

$$\sum_e[F_i^e]=[\boldsymbol{P}_i] \tag{5-27}$$

式中 $\sum_e$ ——对与节点 i 相联的所有单元求和；

$[\boldsymbol{F}_i^e]$——在节点 i 处单元 e 的节点力向量；

$[\boldsymbol{P}_i]$——节点 i 的载荷向量。

对于杆单元，$[\boldsymbol{P}_i]$为直接作用于节点 i 的集中力；对于梁单元，$[\boldsymbol{P}_i]$除了集中力外，还应包括所有的与节点 i 相联的各单元的等效节点载荷。

下面将以平面桁架结构来说明结构整体分析基本方法和结构整体刚度矩阵拼装方法。这些方法具有普遍性，适用于任何单元，而不论它的类型、大小、形状和节点数如何。

(1) 整体刚度矩阵的形成。

例 5-1　如图 5-12 所示为一平面桁架，杆截面积及弹性模量分别为 A 和 E，它由四个节点、三个杆件组成，节点载荷及单元和节点的编号示于图中。在各杆端以铰链联结，当受到外载荷 P_{4x} 和 P_{4y} 作用时结构将发生变形且杆件产生内力。图中虚线所示为变形后的结构。

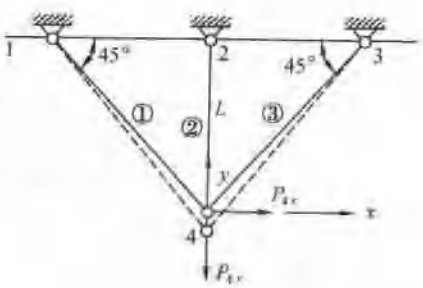


图 5-12　平面桁架

根据式(5-23)，各单元的刚度方程分别为

单元①　$(i=1, j=4)$

$$\begin{bmatrix}[\boldsymbol{F}_1^1]\\ [\boldsymbol{F}_4^1]\end{bmatrix}=\begin{bmatrix}[\boldsymbol{K}_{11}^1] & [\boldsymbol{K}_{14}^1]\\ [\boldsymbol{K}_{41}^1] & [\boldsymbol{K}_{44}^1]\end{bmatrix}\begin{bmatrix}[\boldsymbol{\delta}_1^1]\\ [\boldsymbol{\delta}_4^1]\end{bmatrix}$$

得相应的节点力向量为

$$\begin{cases}[\boldsymbol{F}_1^1]=[\boldsymbol{K}_{11}^1][\boldsymbol{\delta}_1^1]+[\boldsymbol{K}_{14}^1][\boldsymbol{\delta}_4^1]\\ [\boldsymbol{F}_4^1]=[\boldsymbol{K}_{41}^1][\boldsymbol{\delta}_1^1]+[\boldsymbol{K}_{44}^1][\boldsymbol{\delta}_4^1]\end{cases}\tag{5-28}$$

单元②　$(i=2, j=4)$

$$\begin{bmatrix}[\boldsymbol{F}_2^2]\\ [\boldsymbol{F}_4^2]\end{bmatrix}=\begin{bmatrix}[\boldsymbol{K}_{22}^2] & [\boldsymbol{K}_{24}^2]\\ [\boldsymbol{K}_{42}^2] & [\boldsymbol{K}_{44}^2]\end{bmatrix}\begin{bmatrix}[\boldsymbol{\delta}_2^2]\\ [\boldsymbol{\delta}_4^2]\end{bmatrix}$$

相应的节点力向量为

$$\begin{cases}[\boldsymbol{F}_2^2]=[\boldsymbol{K}_{22}^2][\boldsymbol{\delta}_2^2]+[\boldsymbol{K}_{24}^2][\boldsymbol{\delta}_4^2]\\ [\boldsymbol{F}_4^2]=[\boldsymbol{K}_{42}^2][\boldsymbol{\delta}_2^2]+[\boldsymbol{K}_{44}^2][\boldsymbol{\delta}_4^2]\end{cases}\tag{5-29}$$

单元③　$(i=3, j=4)$

$$\begin{bmatrix}[\boldsymbol{F}_3^3]\\ [\boldsymbol{F}_4^3]\end{bmatrix}=\begin{bmatrix}[\boldsymbol{K}_{33}^3] & [\boldsymbol{K}_{34}^3]\\ [\boldsymbol{K}_{43}^3] & [\boldsymbol{K}_{44}^3]\end{bmatrix}\begin{bmatrix}[\boldsymbol{\delta}_3^3]\\ [\boldsymbol{\delta}_4^3]\end{bmatrix}$$

相应的节点力向量为

$$\begin{cases}[\boldsymbol{F}_3^3]=[\boldsymbol{K}_{33}^3][\boldsymbol{\delta}_3^3]+[\boldsymbol{K}_{34}^3][\boldsymbol{\delta}_4^3]\\ [\boldsymbol{F}_4^3]=[\boldsymbol{K}_{43}^3][\boldsymbol{\delta}_3^3]+[\boldsymbol{K}_{44}^3][\boldsymbol{\delta}_4^3]\end{cases}\tag{5-30}$$

由各节点的平衡条件式(5-31)(图 5-13)，可分别得到各节点平衡的矩阵方程，即

$$\begin{gathered}[\boldsymbol{F}_1^1]=[\boldsymbol{P}_1]\quad [\boldsymbol{F}_2^2]=[\boldsymbol{P}_2]\quad [\boldsymbol{F}_3^3]=[\boldsymbol{P}_3]\\ [\boldsymbol{F}_4^1]+[\boldsymbol{F}_4^2]+[\boldsymbol{F}_4^3]=[\boldsymbol{P}_4]\end{gathered}\tag{5-31}$$

将各节点力向量表达式(5-28)、式(5-29)和式(5-30)代入式(5-31)中，并利用各节点位移的连续性条件式(5-26)，它们是

$$[\boldsymbol{\delta}_1^1]=[\boldsymbol{\delta}_1]\quad [\boldsymbol{\delta}_2^2]=[\boldsymbol{\delta}_2]\quad [\boldsymbol{\delta}_3^3]=[\boldsymbol{\delta}_3]$$

$$[\boldsymbol{\delta}_1^1]=[\boldsymbol{\delta}_1^2]=[\boldsymbol{\delta}_1^3]=[\boldsymbol{\delta}_1]$$

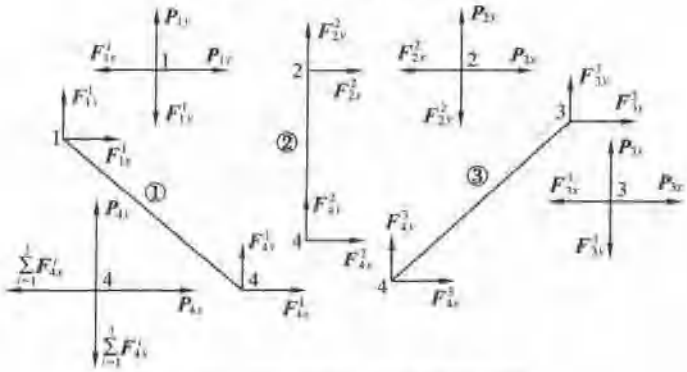


图 5-13 例 5-1 各节点力情况

则可得到用节点位移表示的各节点平衡方程为

$$[\boldsymbol{K}_{11}^1][\boldsymbol{\delta}_1]+[\boldsymbol{K}_{14}^1][\boldsymbol{\delta}_4]=[\boldsymbol{P}_1]$$

$$[\boldsymbol{K}_{22}^2][\boldsymbol{\delta}_2]+[\boldsymbol{K}_{24}^2][\boldsymbol{\delta}_4]=[\boldsymbol{P}_2]$$

$$[\boldsymbol{K}_{33}^3][\boldsymbol{\delta}_3]+[\boldsymbol{K}_{34}^3][\boldsymbol{\delta}_4]=[\boldsymbol{P}_3]$$

$$[\boldsymbol{K}_{41}^1][\boldsymbol{\delta}_1]+[\boldsymbol{K}_{44}^1][\boldsymbol{\delta}_4]+[\boldsymbol{K}_{42}^2][\boldsymbol{\delta}_2]+[\boldsymbol{K}_{43}^2][\boldsymbol{\delta}_3]+[\boldsymbol{K}_{43}^3][\boldsymbol{\delta}_3]+[\boldsymbol{K}_{44}^3][\boldsymbol{\delta}_4]=[\boldsymbol{P}_4]$$

写成矩阵形式，则有

$$\begin{bmatrix}[\boldsymbol{K}_{11}] & 0 & 0 & [\boldsymbol{K}_{14}]\\ 0 & [\boldsymbol{K}_{22}] & 0 & [\boldsymbol{K}_{24}]\\ 0 & 0 & [\boldsymbol{K}_{33}] & [\boldsymbol{K}_{34}]\\ [\boldsymbol{K}_{41}] & [\boldsymbol{K}_{42}] & [\boldsymbol{K}_{43}] & [\boldsymbol{K}_{44}]\end{bmatrix}\begin{bmatrix}[\boldsymbol{\delta}_1]\\ [\boldsymbol{\delta}_2]\\ [\boldsymbol{\delta}_3]\\ [\boldsymbol{\delta}_4]\end{bmatrix}=\begin{bmatrix}[\boldsymbol{P}_1]\\ [\boldsymbol{P}_2]\\ [\boldsymbol{P}_3]\\ [\boldsymbol{P}_4]\end{bmatrix}$$

或简写为

$$[\boldsymbol{K}][\boldsymbol{\delta}]=[\boldsymbol{P}] \tag{5-32}$$

式中 $[\boldsymbol{K}_{44}]=[\boldsymbol{K}_{44}^1]+[\boldsymbol{K}_{44}^2]+[\boldsymbol{K}_{44}^3]$，其余子块为$[\boldsymbol{K}_{ij}]=[\boldsymbol{K}_{ij}^e]$。

式(5-32)为联系节点位移与外载荷的联立线性方程组，其方程数目正好等于结构的自由度数，这就是所谓的结构整体刚度方程，或称为结构的总刚度方程。其中，$[\delta]$称为结构的节点位移列阵，即特求未知量列阵，$[\boldsymbol{P}]$称为结构的节点载荷列阵，当某节点为支承点(固定点)时，该点的节点载荷即为支座反力；$[\boldsymbol{K}]$称为结构的整体刚度矩阵，或称为总刚度矩阵。显而易见，结构整体刚度矩阵是单元刚度矩阵叠加而成的。

整体刚度矩阵可由每个单元刚度矩阵拼装而成。下面我们介绍一种拼装整体刚度矩阵的方法，称为“单元扫描法”，即按单元刚度矩阵子块“对号入座”。

仍以图 5-12 所示的结构为例。首先准备原始数据，为清楚起见，可列表进行这一工作。一般按照小节点号在前，大节点号在后的规则进行(表 5-1)。表中的 i 表示杆单元的小节点号，j 表示大节点号。

表 5-1

杆单元号	x_j-x_i	y_j-y_i	杆长 L	$l=\frac{x_j-x_i}{L}$	$m=\frac{y_j-y_i}{L}$	l^2	lm	m^2
①	L	$-L$	$\sqrt{2}L$	$\frac{1}{\sqrt{2}}$	$-\frac{1}{\sqrt{2}}$	$\frac{1}{2}$	$-\frac{1}{2}$	$-\frac{1}{2}$
②	0	$-L$	L	0	-1	0	0	1
③	$-L$	$-L$	$\sqrt{2}L$	$-\frac{1}{\sqrt{2}}$	$-\frac{1}{\sqrt{2}}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$

各单元刚度矩阵如下　$[\boldsymbol{K}]^{①}=\dfrac{\sqrt{2}EA}{4L}\begin{matrix}\\1\\\\4\end{matrix}\begin{bmatrix}\overset{1}{1} & -1 & -1 & \overset{4}{1}\\ -1 & 1 & 1 & -1\\ -1 & 1 & 1 & -1\\ 1 & -1 & -1 & 1\end{bmatrix}\begin{matrix}1\\\\\\4\end{matrix}$

在刚度矩阵的上边和侧边所标的数字，是为了进行结构总刚度矩阵元素叠加时方便而标注的，仅作参考用。同理

$$[\boldsymbol{K}]^{②}=\frac{EA}{L}\begin{matrix}\\2\\\\4\end{matrix}\begin{bmatrix}\overset{2}{0} & 0 & 0 & \overset{4}{0}\\ 0 & 1 & 0 & -1\\ 0 & 0 & 0 & 0\\ 0 & -1 & 0 & 1\end{bmatrix}\begin{matrix}2\\\\\\4\end{matrix}$$

$$[\boldsymbol{K}]^{③}=\frac{\sqrt{2}EA}{4L}\begin{matrix}\\3\\\\4\end{matrix}\begin{bmatrix}\overset{3}{1} & 1 & \overset{4}{-1} & -1\\ 1 & 1 & -1 & -1\\ -1 & -1 & 1 & 1\\ -1 & -1 & 1 & 1\end{bmatrix}\begin{matrix}3\\\\\\4\end{matrix}$$

单元扫描法是按单元的序号依次生成单元刚度矩阵。首先生成第一个单元的刚度矩阵$[\boldsymbol{K}]^{①}$，然后单元刚度矩阵的各个子块按其在总刚度矩阵中的位置“对号入座”进入总刚度矩阵。例如，单元刚度矩阵$[\boldsymbol{K}]^{①}$所对应的节点号为 1 和 4，则$[\boldsymbol{K}]^{①}$的各子块应进入整体刚度矩阵$[\boldsymbol{K}]$中与节点号 1 和 4 对应的位置，即

$$[\boldsymbol{K}]^{①}=\begin{matrix}\\1\\\\4\end{matrix}\begin{bmatrix}\overset{1}{[\boldsymbol{K}_{11}^{①}]} & \overset{4}{[\boldsymbol{K}_{14}^{①}]}\\ [\boldsymbol{K}_{41}^{①}] & [\boldsymbol{K}_{44}^{①}]\end{bmatrix}\begin{matrix}1\\\\4\end{matrix}$$

$$[\boldsymbol{K}]=\begin{matrix}\\1\\2\\3\\4\end{matrix}\begin{bmatrix}\overset{1}{[\boldsymbol{K}_{11}^{①}]} & \overset{2}{\phantom{[\boldsymbol{K}_{11}^{①}]}} & \overset{3}{\phantom{[\boldsymbol{K}_{11}^{①}]}} & \overset{4}{[\boldsymbol{K}_{14}^{①}]}\\ \phantom{[\boldsymbol{K}_{11}^{①}]} & \phantom{[\boldsymbol{K}_{11}^{①}]} & \phantom{[\boldsymbol{K}_{11}^{①}]} & \phantom{[\boldsymbol{K}_{11}^{①}]}\\ \phantom{[\boldsymbol{K}_{11}^{①}]} & \phantom{[\boldsymbol{K}_{11}^{①}]} & \phantom{[\boldsymbol{K}_{11}^{①}]} & \phantom{[\boldsymbol{K}_{11}^{①}]}\\ [\boldsymbol{K}_{41}^{①}] & \phantom{[\boldsymbol{K}_{11}^{①}]} & \phantom{[\boldsymbol{K}_{11}^{①}]} & [\boldsymbol{K}_{44}^{①}]\end{bmatrix}\begin{matrix}1\\2\\3\\4\end{matrix}$$

按照同样的方法将第二个单元刚度矩阵的子块也“对号入座”进入总刚度矩阵。如此按单元的序号一直进行到最后一个单元，即可得到结构总体刚度矩阵

$$[K]=\sum_{e=1}^{3}[K]^e=\begin{matrix}\begin{matrix}1 & 2 & 3 & & 4\end{matrix}\\ \begin{bmatrix}[K_{11}^1] & & & & [K_{14}^1] \\ & [K_{22}^2] & & & [K_{24}^2] \\ & & [K_{33}^3] & & [K_{34}^3] \\ [K_{41}^1] & [K_{42}^2] & [K_{43}^3] & [K_{44}^1]+[K_{44}^2]+[K_{44}^3] & \end{bmatrix}\begin{matrix}1\\2\\3\\4\end{matrix}\end{matrix}$$

从上述结构整体刚度矩阵$[K]$不难看出，单元刚度矩阵$[K]^e$中的各子块加入到$[K]$中的位置仅依赖于结构的节点编号方式。而整体刚度矩阵$[K]$的各子矩阵$[K_{ij}]$乃是与第i个节点直接相连的各单元矩阵中出现的相应子矩阵$[K_{ij}^e]$叠加而成，即

$$[K_{ij}]=\sum_{e}[K_{ij}^e]$$

当两个节点i和j之间没有一个单元相连时，整体刚度矩阵中的子块$[K_{ij}]$为零。换言之，只有与节点i通过单元直接相连的节点j，在整体刚度矩阵中的子块$[K_{ij}]$才不为零子块。在实际工程结构中，与一个节点通过单元直接相连的节点数目比之结构全部节点数是非常有限的，故整体刚度矩阵中含有大量的零子块。

(2) 刚度矩阵的性质。一般说来，不论何种类型的结构单元，它们的刚度矩阵都具有下述共同的性质。

① 对称性。单元刚度矩阵和总体刚度矩阵都是对称方阵。其第i行、第j列元素也就是第j行、第i列元素，即$[K_{ij}]=[K_{ji}]$(如$[K_{41}]=[K_{14}]$)。这种关系表明：由第j个单位位移分量引起的第i个杆端力分量等于由第i个单位位移分量引起的第j个杆端力分量。这实际上就是根据弹性结构的一个普遍定理(反力互等定理)得出的结论。这个性质十分有用，在计算机运算时，可不必存储刚度矩阵的全部元素，只须存储刚度矩阵的上三角(或下三角)部分，其余元素由$[K_{ij}]=[K_{ji}]$决定，这样可减少内存量。

② 奇异性。单元刚度矩阵和总刚度矩阵都是奇异矩阵。也就是说，刚度矩阵的行列式等于零。这样刚度矩阵的逆矩阵也就不存在。这一性质说明需要对总刚度矩阵按照边界条件进行处理(即消除矩阵的奇异性)之后才能进行求解。结构的边界条件，要排除结构的刚体位移，在排除结构的刚体位移之后，总刚度矩阵是一个正定阵。

③ 稀疏性。总刚度矩阵是零元素非常多的矩阵。一个矩阵如果存在大量的零元素，这种矩阵称做稀疏矩阵。结构越大，这种情况越加明显，大型结构的总刚度矩阵一般都是非常稀疏矩阵。这种性质使我们能够考虑各种有效的存储总刚度的方法来充分利用计算机有限的内存量。

3. *结构整体刚度方程的求解*

结构总刚度方程是根据节点力的平衡条件建立的。总刚度方程为

$$[P]=[K][\delta] \tag{5-33}$$

在建立结构整体刚度方程时，需要注意以下两点：

节点位移列阵中的元素与节点力列阵中的元素必须是对应的，不能任意排列，即同一节点的同一方向的节点位移和节点力的序号相同。

刚度矩阵的各元素必须与节点力相对应，即当节点力序号需要变动时，相应刚度矩阵中的元素位置也要与之对应进行变动。

由于总刚度矩阵的奇异性，要求解总刚度方程，就一定要消除其奇异性。因此，必须考

虑结构的边界约束条件，引入边界条件（支座条件）等。$[P]$为结构外载荷列阵，若某一节点在某一方向上没有外载荷，那么在该节点该方向上各杆轴力之和必等于零。因此当结构外载荷已知时，可作为数据直接输入。$[\delta]$为结构位移列阵，分为未知节点位移和已知节点位移两类。如果在刚度方程的位移列阵$[\delta]$中出现零值，这种零位移对应于没有位移的边界约束点。在总刚度矩阵中，这种位移为零的节点所对应的行与列的元素，对求其他节点位移将不起作用，因而可以从矩阵$[K]$中划去，当然，方程组的阶数也就相应地降低了。因此，在这种情况下引入边界条件的间距就归结为划去矩阵$[K]$中对应的行与列，从而使总刚度矩阵降阶。

结构总刚度方程式(5-33)可写为

$$\begin{bmatrix}[P_\alpha]\\ [K_\beta]\end{bmatrix}=\begin{bmatrix}[K_{\alpha\alpha}] & [K_{\alpha\beta}]\\ [P_{\beta\alpha}] & [K_{\beta\beta}]\end{bmatrix}\begin{bmatrix}[\delta_\alpha]\\ [\delta_\beta]\end{bmatrix} \tag{5-34}$$

式中：$[\delta_\alpha]$——未知位移列阵；

$[\delta_\beta]$——已知位移列阵；

$[P_\alpha]$——未知位移对应的已知节点载荷列阵；

$[P_\beta]$——已知位移对应的未知约束反力列阵。

结构刚度方程若不是按上述排列的，则需按此排列作相应的调整。

由式(5-34)得

$$[P_\alpha]=[K_{\alpha\alpha}][\delta_\alpha]+[K_{\alpha\beta}][\delta_\beta]$$
$$[P_\beta]=[K_{\beta\alpha}][\delta_\alpha]+[K_{\beta\beta}][\delta_\beta]$$

若$[\delta_\beta]=[0]$，则有

$$[P_\alpha]=[K_{\alpha\alpha}][\delta_\alpha] \tag{5-35}$$
$$[P_\beta]=[K_{\beta\alpha}][\delta_\alpha] \tag{5-36}$$

由式(5-35)可求得全部未知的节点位移。如欲求全部约束力，可将已求出的节点位移$[\delta_\alpha]$代入式(5-36)，则可得到全部未知约束反力$[P_\beta]$，再将已知节点位移代入单元刚度方程式或求出欲求的或全部的节点力，从而得到单元的内力。

上述这种改变原方程组排列次序，并经划行划列改变方程组阶次的处理方法，适宜于结构划分单元少或采用手算的情况。但对使用计算机计算时，这种方法就不适宜了。

以上所述边界条件处理方法不仅适用于桁架结构，同样也适用于刚架、平面连续体、空间连续体、板壳及其他各种结构。

在边界条件修正后得到的总体刚度方程$[P]=[K][\delta]$是一组线性代数方程，其中总体刚度矩阵$[K]$是对称的正定矩阵。由此可解出结构全部的节点位移$[\delta]$为

$$[\delta]=[K]^{-1}[P]$$

由于高阶矩阵求逆$[K]^{-1}$的计算过于麻烦，在程序中通常采用高斯消去法、三角分解法及迭代解法等直接求解线性方程组。

根据求得的基本未知量——结构各节点的位移，由单元的应力矩阵，可计算出各单元的内力和应力。

4. 桁架计算简例

例 5-2　设桁架结构和载荷如图 5-14 所示，试进行结构静力分析。

解：(1) 将结构离散化为若干杆单元，并分别注出单元和节点号码，如图 5-14 所示。

(2) 各单元的数据列于表 5-2。

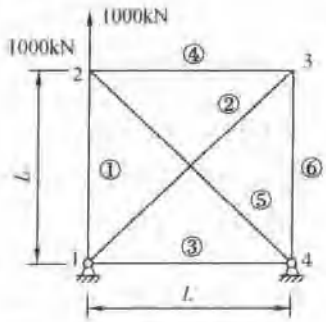


图 5-14 例 2 桁架结构

表 5-2

单元号	i 点	j 点	截面积	杆长	弹性模量	倾角	l	m	l^2	m^2	lm
①	1	2	A	L	E	90°	0	1	0	1	0
②	1	3	A	$\sqrt{2}L$	E	45°	$\frac{1}{\sqrt{2}}$	$\frac{1}{\sqrt{2}}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$
③	1	4	A	L	E	0°	1	0	1		0
④	2	3	A	L	E	0°	1	0	1	0	0
⑤	2	4	A	$\sqrt{2}L$	E	315°	$\frac{1}{\sqrt{2}}$	$-\frac{1}{\sqrt{2}}$	$\frac{1}{2}$	$\frac{1}{2}$	$-\frac{1}{2}$
⑥	3	4	A	L	E	270°	0	−1	0	1	0

(3) 单元刚度矩阵。由杆单元刚度矩阵公式(5-22)得

$$[\boldsymbol{K}]^{①}=\frac{AE}{L}\begin{bmatrix} & 1 & & 2 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 1 \end{bmatrix}\begin{matrix} \\ 1 \\ \\ 2 \end{matrix}$$

$$[\boldsymbol{K}]^{②}=\frac{AE}{2\sqrt{2}L}\begin{bmatrix} & 1 & & & 3 \\ 1 & 1 & -1 & -1 \\ 1 & 1 & -1 & -1 \\ -1 & -1 & 1 & 1 \\ -1 & -1 & 1 & 1 \end{bmatrix}\begin{matrix} \\ 1 \\ \\ 3 \end{matrix}$$

$$[\boldsymbol{K}]^{③}=\frac{AE}{L}\begin{bmatrix} & 1 & & & 4 \\ 1 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}\begin{matrix} \\ 1 \\ \\ 4 \end{matrix}$$

$$[\boldsymbol{K}]^{1}=\frac{AE}{L}\begin{bmatrix} & 2 & & 3 \\ 1 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}\begin{matrix}2\\ \\3\\ \end{matrix}$$

$$[\boldsymbol{K}]^{5}=\frac{AE}{2\sqrt{2}L}\begin{bmatrix} & 2 & & 4 \\ 1 & -1 & -1 & 1 \\ -1 & 1 & 1 & -1 \\ -1 & 1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix}\begin{matrix}2\\ \\ \\4\end{matrix}$$

$$[\boldsymbol{K}]^{6}=\frac{AE}{L}\begin{bmatrix} & 3 & & 4 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 1 \end{bmatrix}\begin{matrix}3\\ \\ \\4\end{matrix}$$

(4)建立结构总刚度矩阵。

$$[\boldsymbol{K}]^{1}=\frac{AE}{2\sqrt{2}L}\begin{bmatrix} & 1 & & 2 & & 3 & & 4 & \\ 3.8284 & 1 & 0 & 0 & -1 & -1 & -2.8284 & 0 & 1 \\ 1 & 3.8284 & 0 & -2.8284 & -1 & -1 & 0 & 0 & \\ 0 & 0 & 3.8284 & -1 & -2.8284 & 0 & -1 & 1 & 2 \\ 0 & -2.8284 & -1 & 3.8284 & 0 & 0 & 1 & -1 & \\ -1 & -1 & -2.8284 & 0 & 3.8284 & 1 & 0 & 0 & 3 \\ -1 & -1 & 0 & 0 & 1 & 3.8284 & 0 & -2.8284 & \\ -2.8284 & 0 & -1 & 1 & 0 & 0 & 3.8284 & -1 & 4 \\ 0 & 0 & 1 & -1 & 0 & -2.8284 & -1 & 3.8284 & \end{bmatrix}$$

(5)考虑边界条件后的总刚度矩阵。如指定节点 1 和 4 设有固定铰链支承，则给定的边界条件为

$$u_1=v_1=v_4=u_4=0$$

现在，我们按已知位移与未知位移，连同与它们对应的载荷，列成分块矩阵，即

$$\begin{bmatrix}[\boldsymbol{P}_\alpha]\\ \cdots \\ [\boldsymbol{P}_\beta]\end{bmatrix}=\begin{bmatrix}\boldsymbol{K}_{\alpha\alpha} & \boldsymbol{K}_{\alpha\beta}\\ \boldsymbol{K}_{\beta\alpha} & \boldsymbol{K}_{\beta\beta}\end{bmatrix}\begin{bmatrix}[\boldsymbol{\delta}_\alpha]\\ \cdots \\ [\boldsymbol{\delta}_\beta]\end{bmatrix}$$

式中各子块的意义与式(5-25)同。

如设节点 1 和 4 的支座反力为 R_{1x}、R_{1y}、R_{4x}、R_{4y}，有

$$[\boldsymbol{P}_\alpha]=\begin{bmatrix}P_{2x}=1000\\ P_{2y}=1000\\ P_{3x}=0\\ P_{4x}=0\end{bmatrix}\begin{matrix}2\\ \\3\\ \end{matrix}\qquad [\boldsymbol{\delta}_\alpha]=\begin{bmatrix}u_2\\ v_2\\ u_3\\ v_3\end{bmatrix}\begin{matrix}2\\ \\3\\ \end{matrix}$$

$$[\boldsymbol{P}_\beta]=\begin{bmatrix}P_{1x}=R_{1x}\\ P_{1y}=R_y\\ P_{4x}=R_{4x}\\ P_{4x}=R_{4y}\end{bmatrix}\begin{matrix}1\\ \\ \\4\end{matrix}\qquad [\boldsymbol{\delta}_\beta]=\begin{bmatrix}u_1=0\\ v_1=0\\ u_4=0\\ v_4=0\end{bmatrix}\begin{matrix}1\\ \\ \\4\end{matrix}$$

于是按已知位移和未知位移重新排列，得出如下形式的总刚度方程

$$\begin{bmatrix} P_{2x} \\ P_{2y} \\ P_{3x} \\ P_{3y} \\ P_{1x} \\ P_{1y} \\ P_{4x} \\ P_{4y} \end{bmatrix} = \frac{AE}{2\sqrt{2}L} \left[\begin{array}{cccc:cccc} 3.8284 & -1 & -2.8284 & 0 & 0 & 0 & -1 & 1 \\ -1 & 3.8284 & 0 & 0 & 0 & -2.8284 & 1 & -1 \\ -2.8284 & 0 & 3.8284 & 1 & -1 & -1 & 0 & 0 \\ 0 & 0 & 1 & 3.8284 & -1 & -1 & 0 & -2.8284 \\ 0 & 0 & -1 & -1 & 3.8284 & 1 & -2.8284 & 0 \\ \hdashline 0 & -2.8284 & -1 & -1 & & 3.8284 & 0 & 0 \\ -1 & 1 & 0 & 0 & & 0 & 3.8284 & -1 \\ 1 & -1 & 0 & -2.8284 & & 0 & -1 & 3.8284 \end{array}\right] \begin{bmatrix} u_2 \\ v_2 \\ u_3 \\ v_3 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

K_{aa}　　$K_{a\beta}$

$K_{\beta a}$　　$K_{\beta\beta}$

考虑边界条件后的总刚度方程为

$$\begin{bmatrix} P_{2x} \\ P_{2y} \\ P_{3x} \\ P_{3y} \end{bmatrix} = \frac{AE}{2\sqrt{2}L} \begin{bmatrix} 3.8284 & -1 & -2.8284 & 0 \\ -1 & 3.8284 & 0 & 0 \\ -2.8284 & 0 & 3.8284 & 1 \\ 0 & 0 & 1 & 3.8284 \end{bmatrix} \begin{bmatrix} u_2 \\ v_2 \\ u_3 \\ v_3 \end{bmatrix}$$

(6) 求节点位移。求解上式，得全部未知节点位移：

$$\begin{bmatrix} u_2 \\ v_2 \\ u_3 \\ v_3 \end{bmatrix} = \frac{L}{AE} \cdot 1000 \begin{bmatrix} 2.6927 \\ 1.4420 \\ 2.150 \\ -0.5577 \end{bmatrix}$$

(7) 求支座反力。由总刚度方程可得

$$\begin{bmatrix} P_{1x} \\ P_{1y} \\ P_{4x} \\ P_{4y} \end{bmatrix} = \frac{1000}{2\sqrt{2}} \begin{bmatrix} 0 & 0 & -1 & -1 \\ 0 & -2.8284 & -1 & -1 \\ -1 & 1 & 0 & 0 \\ 1 & -1 & 0 & -2.8284 \end{bmatrix} \begin{bmatrix} 2.6927 \\ 1.4420 \\ 2.1350 \\ -0.5577 \end{bmatrix} = \frac{1000}{2\sqrt{2}} \begin{bmatrix} -1.5774 \\ -5.6576 \\ -1.2519 \\ 2.8315 \end{bmatrix}$$

所以，支座反力为

$$\begin{bmatrix} P_{1x}=R_{1x} \\ P_{1y}=R_{1y} \\ P_{4x}=R_{4x} \\ P_{4y}=R_{4y} \end{bmatrix} = \begin{bmatrix} -557.5 \\ -2000 \\ -442.5 \\ 1000 \end{bmatrix}$$

(8) 求单元节点力。现在，节点位移已全部求出，利用单元刚度方程，可容易地求得单元节点力。这里仅以杆单元④为例(图 5-15)，其他杆可按同解求得，其结果示于图 5-16。

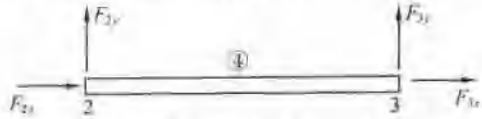


图 5-15　杆单元④

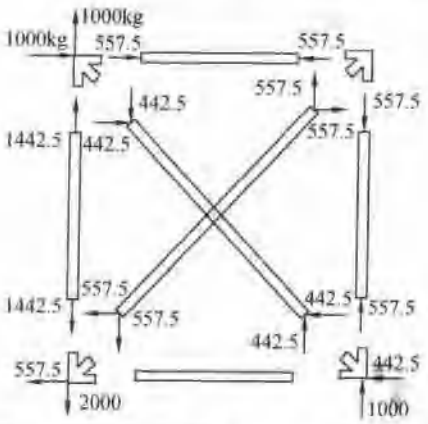


图 5-16　所有杆单元节点力(单位:kN)

$$\begin{bmatrix} F_{2x} \\ F_{2y} \\ F_{3x} \\ F_{3y} \end{bmatrix} = \frac{AE}{L} \begin{bmatrix} 1 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} u_2 \\ v_2 \\ u_3 \\ v_3 \end{bmatrix} = 1000 \begin{bmatrix} 1 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 2.6927 \\ 1.4420 \\ 2.1350 \\ -0.5577 \end{bmatrix} = \begin{bmatrix} 557.5 \\ 0 \\ -557.5 \\ 0 \end{bmatrix}$$

5. 深梁计算

图 5-17(a)为一受均匀载荷 $q=100\text{MPa}$、两端固结的深梁，其厚度 $h=100\text{mm}$。下面采用有限元法计算该构件的变形及应用。

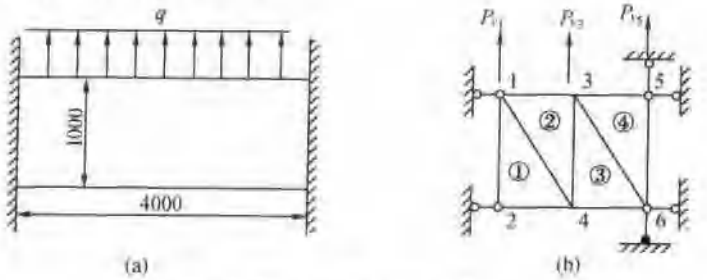
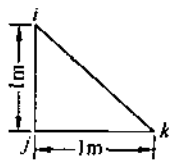
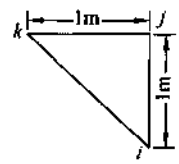


图 5-17　深梁及有限元分析示意图

考虑到问题的对称性，只研究右半段。将右半段划分为四个单元，单元编号及节点编号见图 5-17(b)。将分布载荷 q 等效成三个节点载荷 P_{y1}、P_{y3}、P_{y5}。由于梁的中间截面上的水平位移为零，因此用水平链杆将节点 1 和 2 的水平方向位移约束(即这两个节点的水平位移为零)，而梁的右端显然水平和垂直方向上的位移都不允许，所以节点 5 和 6 在水平和垂直方向上都加了链杆予以约束。这些节点某一分量为定值的条件就称为边界条件或约束条件。

(1) 单元刚度矩阵计算，见表 5-3。

表 5-3 单元刚度矩阵计算

单元图形					
单元编号		①	③	②	④
单元信息	i	1	3	4	6
	j	2	4	3	5
	k	4	6	1	3
$[B]$		$\begin{bmatrix} 0 & 0 & -1 & 0 & 1 & 0 \\ 0 & 1 & 0 & -1 & 0 & 0 \\ 1 & 0 & -1 & -1 & 0 & 1 \end{bmatrix}$		$\begin{bmatrix} 0 & 0 & 1 & 0 & -1 & 0 \\ 0 & -1 & 0 & 1 & 0 & 0 \\ -1 & 0 & 1 & 1 & 0 & -1 \end{bmatrix}$	
$[D]$ $\mu=0.3$ $E=2\times10^5$MPa		$\begin{bmatrix} 2.20 & 0.66 & 0 \\ 0.66 & 2.20 & 0 \\ 0 & 0 & 0.77 \end{bmatrix}\times10^5$		$\begin{bmatrix} 2.20 & 0.66 & 0 \\ 0.66 & 2.20 & 0 \\ 0 & 0 & 0.77 \end{bmatrix}\times10^5$	
$[D][B]$		$\begin{bmatrix} 0 & 0.66 & -2.20 & -0.66 & 2.20 & 0 \\ 0 & 2.20 & -0.66 & -2.20 & 0.66 & 0 \\ 0.77 & 0 & -0.77 & -0.77 & 0 & 0.77 \end{bmatrix}\times10^5$		$\begin{bmatrix} 0 & -0.66 & 2.20 & 0.66 & -2.20 & 0 \\ 0 & -2.20 & 0.66 & 2.20 & -0.66 & 0 \\ 0.77 & 0 & 0.77 & 0.77 & 0 & -0.77 \end{bmatrix}\times10^5$	
$[K]=[B]^T[D][B]\Delta$ $h=0.1$m $\Delta=0.5$m^2		$\begin{bmatrix} 0.77 & 0 & -0.77 & -0.77 & 0 & 0.77 \\ & 2.20 & -0.66 & -2.20 & 0.66 & 0 \\ & & 2.97 & 1.43 & -2.20 & -0.77 \\ & 对称 & & 2.97 & -0.66 & -0.77 \\ & & & & 2.20 & 0 \\ & & & & & 0.77 \end{bmatrix}\times5\times10^3$		$\begin{bmatrix} 0.77 & 0 & -0.77 & -0.77 & 0 & 0.77 \\ & 2.20 & -0.66 & -2.20 & 0.66 & 0 \\ & & 2.97 & 1.43 & -2.20 & -0.77 \\ & 对称 & & 2.97 & -0.66 & -0.77 \\ & & & & 2.20 & 0 \\ & & & & & 0.77 \end{bmatrix}\times5\times10^3$	

（2）总刚度矩阵计算：

$$K=[K]_{①}+[K]_{②}+[K]_{③}+[K]_{④}$$

$$=\begin{bmatrix} 2.97 & 0 & -0.77 & 0.77 & -2.20 & -0.66 & 0 & 1.43 & 0 & 0 & 0 & 0 \\ & 2.97 & -0.66 & -2.20 & -0.77 & -0.77 & 1.43 & 0 & 0 & 0 & 0 & 0 \\ & & 2.97 & 1.43 & 0 & 0 & -2.20 & -0.77 & 0 & 0 & 0 & 0 \\ & & & 2.97 & 0 & 0 & -0.66 & -0.77 & 0 & 0 & 0 & 0 \\ & & & & 5.94 & -1.43 & -1.54 & -1.43 & -2.20 & -0.66 & 0 & 1.43 \\ & & & & & 5.94 & -1.43 & -4.40 & -0.77 & -0.77 & 1.43 & 0 \\ & & & & & & 5.94 & 1.43 & 0 & 0 & -2.20 & -0.77 \\ & & & & & & & 5.94 & 0 & 0 & -0.66 & -0.77 \\ & & & 对称 & & & & & 2.97 & 1.43 & -0.77 & -0.66 \\ & & & & & & & & & 2.97 & -0.77 & -2.20 \\ & & & & & & & & & & 2.97 & 0 \\ & & & & & & & & & & & 2.97 \end{bmatrix}\times5\times10^3$$

(3) 总节点载荷列阵 F 及总节点位移列阵

$$q=[u_1 \quad v_1 \quad u_2 \quad v_2 \quad u_3 \quad v_3 \quad u_4 \quad v_4 \quad u_5 \quad v_5 \quad u_6 \quad v_6]^{\mathrm{T}}$$

其中，$u_1=u_2=u_5=v_5=u_6=v_6=0$，这些已知值成为问题的边界条件(约束条件)。

将支点(节点 1,2,5,6)的约束反力用 N_{x1}、N_{x2}、N_{x5}、N_{y5}、N_{x6}、N_{y6} 表示，则总节点载荷列阵

$$F=[N_{x1} \quad P_{y1} \quad N_{x2} \quad 0 \quad 0 \quad P_{y3} \quad 0 \quad 0 \quad N_{x5} \quad N_{y5}+P_{y5} \quad N_{x6} \quad N_{y6}]^{\mathrm{T}}$$

0 元素表示节点上的内力互相抵消，而分布载荷 q 向节点移置后可得 $P_{y1}=5000\mathrm{kN}$，$P_{y3}=10000\mathrm{kN}$，$P_{y5}=5000\mathrm{kN}$。

(4) 建立有限元方程式：根据边界条件 $u_1=u_2=u_5=v_5=u_6=v_6=0$，整体有限元方程式如下

$$5\times10^3\times\begin{bmatrix}
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\
 & 2.97 & 0 & -2.20 & -0.77 & -0.77 & 1.43 & 0 & 0 & 0 & 0 & 0\\
 & & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\
 & & & 2.97 & 0 & 0 & -0.66 & -0.77 & 0 & 0 & 0 & 0\\
 & & & & 5.94 & -1.43 & -1.54 & -1.43 & 0 & 0 & 0 & 0\\
 & & & & & 5.94 & -1.43 & -4.4 & 0 & 0 & 0 & 0\\
 & & & \text{对称} & & & 5.94 & 1.43 & 0 & 0 & 0 & 0\\
 & & & & & & & 5.94 & 0 & 0 & 0 & 0\\
 & & & & & & & & 1 & 0 & 0 & 0\\
 & & & & & & & & & 1 & 0 & 0\\
 & & & & & & & & & & 1 & 0\\
 & & & & & & & & & & & 1
\end{bmatrix}\begin{bmatrix}u_1\\v_1\\u_2\\v_2\\u_3\\v_3\\u_4\\v_4\\u_5\\v_5\\u_6\\v_6\end{bmatrix}=\begin{bmatrix}0\\5000\\0\\0\\0\\10000\\0\\0\\0\\0\\0\\0\end{bmatrix}$$

上式已根据边界条件作了缩减，因为 $u_1=u_2=u_5=v_5=u_6=v_6=0$ 是已知值，可将 K 中对应的行上元素除对角线元素置为 1 外，其余均置为 0，而同时需将 F 中对应的元素置为 0。这意味着，已知结果的边界上的节点位移分量不必再去计算，因而上述方程被缩减为 $12-6=6$ 阶线性方程组。除此之外，还应注意到，以上过程并不用求出各支反力 $N_{x1}, N_{x2}, \cdots, N_{y6}$。

(5) 计算结果：采用高斯消元法解上述线性方程组，可得出节点位移，再求应变和应力。计算结果列于表 5-4。

表 5-4　计算结果

节点号	节点位移/mm		单元号	单元应力/MPa		
	X 向	Y 向		σ_x	σ_y	τ_{xy}
1	0	3.5764	②	−63.872	64.021	−64.021
2	0	3.1604	①	78.378	25.924	−35.979
3	0.35301	2.3402	④	92.090	30.038	−119.983
4	−0.4154	2.3282	③	−77.584	−23.275	−180.017
5	0	0				
6	0	0				

习 题 5

1. 简述有限元法的基本思路及其数学基础。它的解题方法是解析性的还是数值性的，为什么？

2. 有限元分析法的解题精度与单元的位移函数或形状函数的设定有直接关系。请说明它们之间存在什么关系；为使它们合理，应满足哪些条件(通过某一种单元的一种位移函数和形状函数形式进行说明)？

3. 单元刚度矩阵在有限元分析法中起什么作用？它的物理意义是什么？

4. 当结构划分出单元后，根据单元刚度矩阵如何组成总刚度矩阵？如何从物理意义上去理解这种组成方法(可以用两个单元为例)？

第 6 章　设计中的评价与决策

本章首先介绍了设计中的评价方法，主要包括评价内容、标准和方法，然后介绍了设计中决策方法，主要包括决策基本原则、分析决策方法。最后又进一步介绍了模糊评价方法，为工程设计铺垫了评价与决策基础。

工程设计过程是一个由发散至收敛，由搜索至筛选的多次反复过程。对设计工作中所获得的多个设计方案，必须通过方案评价与决策，才能优选出拟采用的最佳设计方案。

工程设计中的评价与决策是两种不同的概念。所谓评价是对各方案的价值进行比较评定。所谓决策是依据价值的高低选择并确定最终方案。但是，它们两者之间又紧密相关。这就是评价是决策的依据，而决策是评价的最终目的。

6.1　设计中的评价

工程设计具有约束性、多解性、相对性这样三个特征，尤其是它的多解性，即解答方案不是惟一的，这就要求先对某问题提出尽可能多的解决方案，然后从众多满足要求的方案中，优选出拟采用的方案来。

为了选出拟采用的方案，首先要对各候选方案进行评价。所谓评价即对方案的质量、价值或就其某一性质作出说明。例如，方案完成预定功能的程度、外形美观的程度等。有了各候选方案的评价结果，即可作出决策。所谓决策就是对评价结果或对所提供的某些情况，根据预定目标作出选择或决定，决策的结果就是拟采用的方案。

在设计中进行评价和决策时应注意以下几点：

(1) 评价的原始依据是设计要求。这些要求，有些是定量的，如生产率要求、速度要求等，有些要求却是定性的，如外形美观、结构简单等，评价中必须注意定性问题的处理。

(2) 评价中一个重要的要求是评价结果要符合评价对象的实际。但因这一工作总是由人进行的，不可避免地会引入主观因素的影响。评价中必须注意其客观性的增强。

(3) 设计的要求总是多方面的，它也为决策提供了一定目标，正确的评价、决策应综合考虑多种要求，注意全面、适当，应该指出，局部最优不一定全局最优，短期最优不一定长期最优，单项最优不一定整体最优，最终的决策常是多方面要求的折中。

实际上，人们在工作过程中，总是自觉或不自觉地对可以设想的方案进行评价并作出决策，随着科学技术的发展和设计对象的复杂化，有必要采用先进理论和方法使评价过程更自觉、更科学地进行。评价不应仅理解为对方案的科学分析和评定，还应针对方案的技术、经济弱点加以改进和完善。广义的评价实质上是产品开发的优化过程。

6.1.1　评价的内容

工程设计的评价内容一般包括三个方面：技术评价、经济评价和社会评价。

(1) 技术评价：主要围绕预定功能要求进行，评价系统(或设计方案)能否满足预定技

术性能要求及其满足程度，如产品的各项性能指标、寿命、可靠性、安全性和能源消耗等。

(2) 经济评价：主要围绕经济效益进行评价，包括方案的成本、利润、各项经济耗费等，其目的是降低工程造价和产品成本，提高经济效益。

(3) 社会评价：主要评定方案实施后可能产生的社会效益和影响，如能否推动科技进步、能否促进生产发展、能否减少环境污染或有助于生态平衡、是否符合国家政策或有利于资源的综合利用等。

6.1.2 评价标准

进行产品或设计方案的评价，首先应该确定评价标准，否则很难进行评价工作。

评价标准又称评价目标或评价指标，它来源于设计所要达到的目的，它可从设计任务书或要求明细表中找到。

评价标准分为定性和定量两种指标，如美观程度，只能定性描述，属于定性指标；而成本、重量、产量等可以用数值表示，称为定量指标。在评价标准中，有时定量和定性指标是可以相互转化的。

工业产品设计要求有单项的，也有多项的，因此，评价标准可以是单个的，也可以是多个的。

由于实际的评价标准(评价目标)常不止一个，其重要程度亦不相同，因此需建立评价目标系统。所谓评价目标系统，就是依据系统论观点，把评价目标看成系统。评价目标系统常用评价目标树来表达。评价目标树就是依据系统可以分解的原则，把总评价目标分解为一级、二级、……等子目标，形成倒置的树状。图 6-1 为评价目标树的示意图。图中，Z 为总目标；z_1、z_2 为第一级子目标；z_{11}、z_{12} 为 z_1 的子目标，也就是 Z 的第二级子目标；z_{111}、z_{112} 是 z_{11} 的子目标，也是 Z 的第三级子目标。最后一级的子目标即为总目标的各具体评价目标(评价标准)。

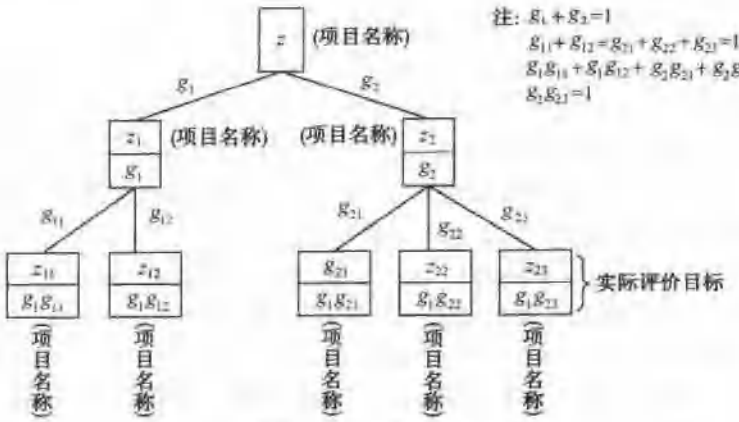


图 6-1 评价目标树

建立评价目标树是将产品的总目标具体化，使之便于定性或定量评价。定量评价时应根据各目标的重要程度设置加权系数。加权系数是反映目标重要程度的量化系数，加权系数大，意味着重要程度高。为便于分析计算，一般取各评价目标加权系数 $g_i \leqslant 1$，且 $\sum g_i = 1$。

评价目标加权系数在目标树中是逐级分配的(图 6-2)。同级目标加权系数之和等于 1。同级子目标的加权系数之和等于它上一级目标的加权系数。对目标系数评价时,使用最末一级子目标的加权系数,用 g_i 表示,且 g_i 之和应为 1。

加权系数值可用经验确定或用判别表法列表计算。判别表法是将评价目标的重要程度两两加以比较并给予分值。两目标同等重要时各给 2 分;某一项比另一项重要可分别给 3 分和 1 分;某一项比另一项重要得多,则分别给 4 分和 0 分。最后将各评价目标的给分汇列成表(表 6-1),并计算出各评价目标的加权系数,

$$g_i = \frac{W_i}{\sum_{i=1}^{n} W_i} \tag{6-1}$$

式中:W_i——各评价目标的总分;

n——评价目标数。

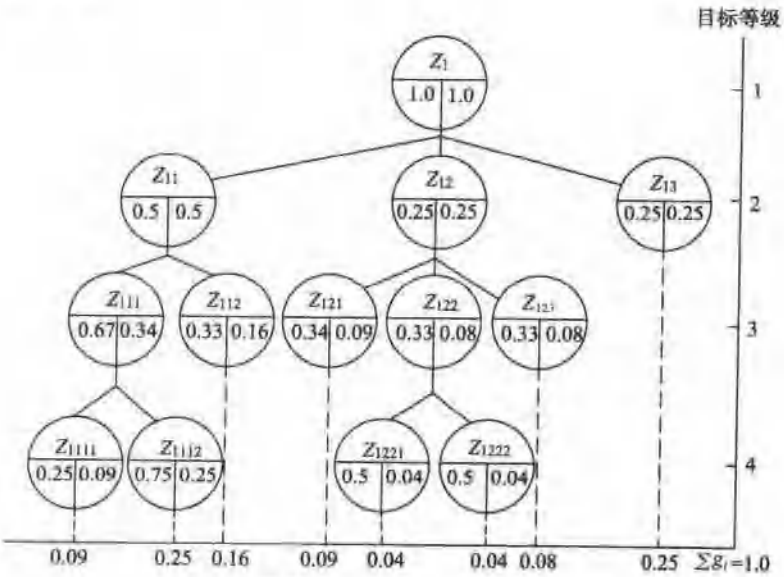


图 6-2　评价树与加权系数

表 6-1　加权系数判别计算表

评价目标 \ 比较目标	价格	维修性	舒适性	寿命	外观	W_i	g_i
价　格		4	3	3	4	14	0.35
维修性	0		1	1	3	5	0.125
舒适性	1	3		2	4	10	0.25
寿　命	1	3	2		4	10	0.25
外　观	0	1	0	0		1	0.025

$$\sum W_i = 40, \sum g_i = 1$$

例 6-1　试对某种自行车进行评价。五个评价目标的重要程度依次为价格、舒适性、寿命、维修性、外观,试定加权系数。

解：按判别表法所判定出的各评价目标的加权系数如表 6-1 所示。

判定价格、维修性、舒适性、寿命和外观五个评价目标的加权系数分别为 0.35、0.125、0.25、0.25 和 0.025。

通过上述目标树的分析，使人们对总目标、各评价目标及其重要程度一目了然，使用起来也很方便。

6.1.3 评价方法

设计中用于评选方案的评价方法有很多种。以下介绍常用的三种：简单评价法、评分法和技术经济评价法。

1. 简单评价法

简单评价法可用来对有关方案作定性的粗略评价和优劣排序，不反映评价目标的重要程度和方案的理想程度。该方法中最常用的如点评价法，它是对各方案按评价目标逐项作粗略评价，并用行(+)、不行(−)、信息不足(?)三种符号打分，最后总评并作出决策。这种点评价法示例如表 6-2 所示。

表 6-2 点评价法

设计方案 / 评价目标	A	B	C	D
满足功能要求	+	+	+	+
成本在规定范围内	−	+	+	+
加工装配可行	+	?	+	+
使用维护方便	+	−	+	+
满足人机学要求	−	+	+	−
总　评	+	? ++	5+	3+

由表 6-2 得出结论：C 方案较好，其次为 D、B、A 方案。

2. 评分法

评分法用分值作为衡量方案优劣的尺度，对方案进行定量评价，如有多个评价目标则先分别对各目标进行评分，再经处理求得方案的总分。

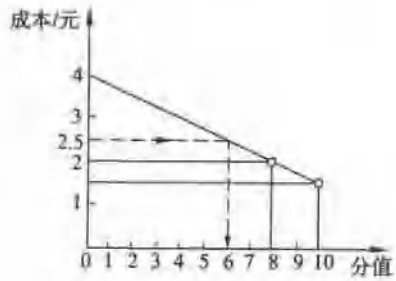


图 6-3 评分曲线

评分可用 10 分制或 5 分制对方案进行打分，如果方案为理想状态取最高分，不能用则取 0 分。评分标准见表 6-3。各评价目标的参数值与分值的关系，可用评分系数估算。先根据评价目标的允许值、要求值和理想值分别给 0 分、8 分和 10 分(10 分制)或 0 分、4 分和 5 分(5 分制)，用三点定曲线的方法求出评分函数曲线，由该曲线再求各参数值对应的分值。如某产品成本 1.6 元为理想值(10 分)，2 元为要求值(8 分)，4 元为极限值(0 分)，可根据这三点求出该产品的评分函数曲线如图 6-3 所示。若此产品的某种方案成本价为 2.5 元，则由该产品的评分曲线求得其分值为 6 分。

为减少个人主观因素对评分的影响，一般都采用集体评分法，即由几个评分者以评价目标为序对各方案评分，取平均值或去除最大、最小值后的平均值作为方案的分值。

对于多评价目标的方案其总分可用分值相加法、分值连乘法或加权计分法(有效值)等方法进行计算。其中加权计分法在总分计算中由于综合考虑了各评价目标的分值及其加权系数的影响,使总分计算更趋合理,应用也最广泛。

表 6-3　评分标准

10 分制	0	1	2	3	4	5	6	7	8	9	10
	不能用	缺陷多	较差	勉强可用	可用	基本满意	良	好	很好	超目标	理想
5 分制	0		1		2		3		4		5
	不能用		勉强可用		可用		良好		很好		理想

加权计分法(有效值法)的评分计分与过程如下:

(1) 确定评价目标。整个设计评价目标系统可视为一个集合,评价目标集合可表示为

$$Z=\{z_1,z_2,\cdots,z_n\}$$

(2) 确定各评价目标的加权系数。$g_i\leqslant 1$,$\sum g_i=1,i=1,2,\cdots,n$,各评价目标的加权系数矩阵为

$$\boldsymbol{G}=[g_1 \quad g_2 \quad \cdots \quad g_n]$$

(3) 确定评分制式(采用 10 分制或 5 分制),列出评分标准;

(4) 对各评价目标评分(可用评分曲线或集体评分法),最后用矩阵形式列出 m 个方案 n 个评价目标的评分值矩阵

$$\boldsymbol{W}=\begin{bmatrix}w_1\\w_2\\\vdots\\w_i\\\vdots\\w_m\end{bmatrix}=\begin{bmatrix}w_{11}&w_{12}&\cdots&w_{1i}&\cdots&w_{1n}\\w_{21}&w_{22}&\cdots&w_{2i}&\cdots&w_{2n}\\\vdots&&&&&\\w_{j1}&w_{j2}&\cdots&w_{ji}&\cdots&w_{jn}\\\vdots&&&&&\\w_{m1}&w_{m2}&\cdots&w_{mi}&\cdots&w_{mn}\end{bmatrix}$$

(5) 计算 m 个方案 n 个评价目标的加权分值(有效值)矩阵

$$\boldsymbol{R}=\boldsymbol{W}\boldsymbol{G}^{\mathrm{T}}=\begin{bmatrix}w_{11}&w_{12}&\cdots&\cdots&\cdots&w_{1n}\\w_{21}&w_{22}&\cdots&\cdots&\cdots&w_{2n}\\\vdots&&&&&\vdots\\w_{i1}&w_{i2}&\cdots&\cdots&\cdots&w_{in}\\\vdots&&&&&\\w_{m1}&w_{m2}&\cdots&\cdots&\cdots&w_{mn}\end{bmatrix}\begin{bmatrix}g_1\\g_2\\\vdots\\g_i\\\vdots\\g_n\end{bmatrix}=\begin{bmatrix}R_1\\R_2\\\vdots\\R_i\\\vdots\\R_m\end{bmatrix}$$

其中,第 j 个方案的加权总分值(有效值)

$$R_j=\boldsymbol{W}\boldsymbol{G}_j^{\mathrm{T}}=w_{j1}g_1+w_{j2}g_2+\cdots+w_{jn}g_n \tag{6-2}$$

(6) 比较各方案的加权总分值(有效值),评选最佳方案。R_j 的数值越大,表示此方案的综合性能越好,故 R_j 值大者为最佳方案。

例 6-2　试用加权计分法对某种空调的三种设计方案进行评价。

解:(1) 根据设计要求建立评价目标树,如图 6-4 所示。

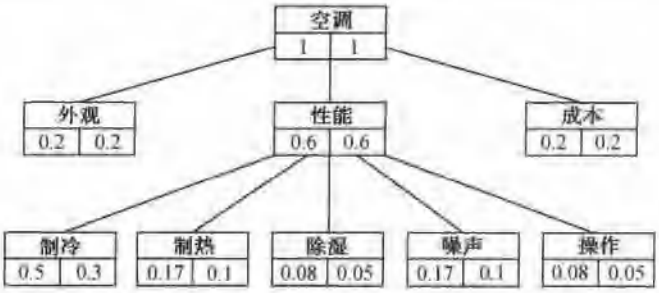


图 6-4 某种空调评价目标树及加权系数

(2) 评分及计算总分 R:

① 根据建立的评价目标树,该空调设计方案的评价目标为制冷效果、制热效果、除湿效果、噪声、操作性能、外观美、价格低廉共 7 项。故评价目标矩阵为

$$\boldsymbol{Z}=[z_1 \quad z_2 \quad z_3 \quad z_4 \quad z_5 \quad z_6 \quad z_7]$$

② 经分析和确定出的相应的加权系数集为

$$\boldsymbol{G}=[g_1 \quad g_2 \quad g_3 \quad g_4 \quad g_5 \quad g_6 \quad g_7]=[0.3 \quad 0.1 \quad 0.05 \quad 0.1 \quad 0.05 \quad 0.2 \quad 0.2]$$

③ 按照表 6-3 的 10 分制标准打分。

④ 对各评价目标评分结果,如表 6-4 所示。

表 6-4 某空调评分结果

评价目标 / 方案	制冷效果	制热效果	除湿效果	噪 声	操作性能	外观美	价格低廉
1	9	8	8	9	0	9	9
2	8	7	8	8	7	7	7
3	7	7	8	7	0	9	10
加权系数	0.3	0.1	0.05	0.1	0.05	0.2	0.2

因而该空调的三种方案七个评价目标的评分值矩阵为

$$\boldsymbol{W}=\begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix}=\begin{bmatrix} 9 & 8 & 8 & 9 & 0 & 9 & 9 \\ 8 & 7 & 8 & 8 & 7 & 7 & 7 \\ 7 & 7 & 8 & 7 & 0 & 9 & 10 \end{bmatrix}$$

⑤ 求加权分值矩阵:

$$\boldsymbol{R}=\boldsymbol{W}\boldsymbol{G}^{\mathrm{T}}=\begin{bmatrix} 9 & 8 & 8 & 9 & 0 & 9 & 9 \\ 8 & 7 & 8 & 8 & 7 & 7 & 7 \\ 7 & 7 & 8 & 7 & 0 & 9 & 10 \end{bmatrix}=\begin{bmatrix} 0.3 \\ 0.1 \\ 0.05 \\ 0.1 \\ 0.05 \\ 0.2 \\ 0.2 \end{bmatrix}=\begin{bmatrix} 8.4 \\ 7.45 \\ 7.7 \end{bmatrix}$$

(3) 比较各方案加权总分值(有效值),决策选用方案。由上可知各方案加权总分值为 $\boldsymbol{R}^{\mathrm{T}}=[R_1 \quad R_2 \quad R_3]=[8.4 \quad 7.45 \quad 7.7]$。由于 $R_1>R_3>R_2$,故方案 1 为最佳,方案 3 次之,方案 2 最差;所以决策方案 1。

3. 技术经济评价法

技术经济评价法是将总目标分为两个子目标，即技术目标和经济目标，求出技术价 W_t 和经济价 W_w，然后按一定方式进行综合，求出总价值 W，方案中 W 高者优胜。技术经济评价法是德国工程师协会技术准则 VDI2225 中推荐的评价方法。

技术评价法的评价过程如下：

(1) 对方案进行技术评价。技术评价的目的是求方案的技术价 W_t，其步骤为：

① 确定技术评价目标，建立评价目标树；

② 评分：对各方案的技术评价目标打分(参照表 6-3)，分值为 W_{1j}、W_{2j}、…、W_{nn}；

③ 确定加权系数：根据各目标的重要程度，给予不同的系数 g_1、g_2、…、g_n；

④ 求方案的技术价 W_t，即

$$W_t = \frac{\sum_{j=1}^{n} W_j q_i}{W_{\max} \sum_{i=1}^{n} g_i} = \frac{\sum_{i=1}^{n} W_j g_i}{W_{\max}} \tag{6-3}$$

式中：W_j——各技术评价指标的评分值；

q_i ——各技术评价指标的加权系数，取 $\sum q_i = 1, i = 1, 2, \cdots, n$；

$W_{\max}$——最高分值(10 分制的 10 分，5 分制的 5 分)。

技术价 W_t 值越高，说明方案的技术性能越好。理想方案的技术价为 1，若 $W_t < 0.6$ 表示方案在技术上不合格，必须加以改进才能考虑选用。

(2) 对方案进行经济评价。经济评价的目的是求方案的经济价 W_w，即计算理想生产成本与实际生产成本的比值，其计算公式为

$$W_w = \frac{H_1}{H} = \frac{0.7 H_z}{H} \tag{6-4}$$

式中：H——实际生产成本(元)；

H_1——理想生产成本(元)；

H_z——设计任务书中允许生产成本(元)，应低于市场最低有效价格，一般取 $H_1 = 0.7 H_z$。

经济价 W_w 值越高，表示该方案的经济效益越好。当 $W_w = 1$ 时，为理想的经济价，即实际生产成本与理想成本相等。W_w 的许用值为 0.7，此时实际生产成本等于允许生产成本。

(3) 技术经济综合评价。在求出方案的技术价和经济价后，可利用计算或图示方法进行方案的技术经济综合评价。

① 计算法。计算方案总价值 W 有两种方法：

直线法(均值法)

$$W = \frac{1}{2}(W_t + W_w) \tag{6-5}$$

抛物线法

$$W = (W_t + W_w)^{\frac{1}{2}} \tag{6-6}$$

总价值 W 值越大，表明方案的技术经济综合性能越好，一般应取 $W \geqslant 0.65$。直线法的特点是：W_t 与 W_w 相差较大时，所得 W 值仍较大；而用抛物线法时，只要 W_t、W_w 两项有一

项数值小，就会使总价值 W 值降低很多，所以，使用抛物线法更便于方案评价与决策。

② 作图法。通过作优度图来表达方案的技术经济综合性能。优度图如图 6-5 所示。图中横坐标表示技术价 W_t，纵坐标代表经济价 W_w。在 W_t 与 W_w 构成的平面坐标系中，每个方案的 W_t 和 W_w 值构成点 S_i，S_i 的位置反映此方案的优良程度(优度)。坐标系中 $W_w=1$、$W_t=1$ 构成的点 S^* 为理想优度，它是技术经济综合指标的理想值。OS^* 连线称为"开发线"，线上各点 $W_w=W_t$。S_i 点离 S^* 越近，表示方案技术经济指标高；而离开发线越近方案技术经济综合性能越好。因此，用优度图可形象地看出设计方案的技术与经济综合性能，且便于提出改进方法。

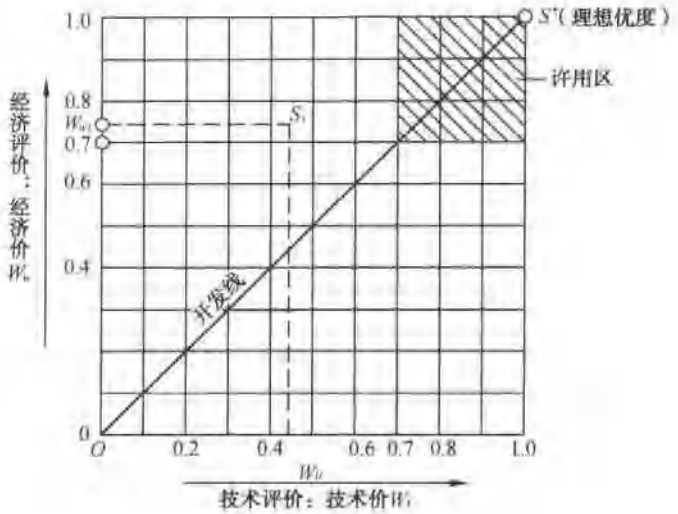


图 6-5 优度图

6.2 设计中的决策

在工程及产品设计中，有了各候选方案的评价结果，即可作出决策。所谓决策就是对方案评价结果或对所提供的某些情况，根据预定目标作出选择或决定。决策的结果就是拟采用的方案。

6.2.1 决策的基本原则

由于设计要求总是多方面的，所以设计中的决策应注意以下各原则：

(1) 系统原则。从系统观点来看，任何一个设计方案都是一个系统，可用各种性能指标来描述。但方案本身又会与制造、检验、销售等其他系统发生关系，所以决策时不能仅从方案本身或方案中某一性能指标出发，还应考虑以整个方案的总体目标为核心的有关系统的综合平衡，达到企业总体最佳的决策。

(2) 可行性原则。要使所作出的决策具有确实的可行性。成功的决策不仅要考虑需要，还要考虑可能。要估计有利因素和成功机会，还要估计不利因素和失败的风险。既要考虑当前状态和需要，也要估计今后的变化和发展。

(3) 满意原则。由于设计工作的复杂性，不仅设计要求涉及很多方面，而且很多方面本身就无法准确评价，所以在设计中追求十全十美的最佳方案既不可能也无意义，只能在众多的方案中寻求一个或几个相对比较满意的方案来。

(4) 反馈原则。设计过程中的决策是否正确应通过实践来检验。要根据实践过程中各因素的发展变化所反馈的信息，及时作出调整，作出正确的决策。

(5) 多方案原则。设计过程中各设计方案逐步具体化，人们对它的认识也逐渐深刻。为了保证设计质量，特别是在方案设计阶段，决策可以是多方案的。几个选出的方案同时发展，直到确能分出各方案优劣之后再作出新的决策。

6.2.2　决策类型及其分析方法

根据各设计方案的有关情况，决策可有不同的类型。不同类型的决策应采用不同的决策方法。下面就确定型决策、风险型决策和非确定型决策简述其分析方法。

1. 确定型决策

所谓确定型决策是指在决策问题的未来状态已经完全确定情况下的决策。在评价中对各设计方案的未来状态包括其成本估算、销售前景等都可作出较准确的估价时，方案的选定便属于确定型决策。这时，可按评价结果直接作出决策。

2. 风险型决策

所谓风险型决策是指在对决策问题的未来状态不能完全肯定却知道其发生概率情况下对问题作出的决策。在选择设计方案时，如果以产品销售量或市场占有率等未来状态不能完全肯定的条件决定方案的取舍时，这种决策即风险型决策。进行风险型决策的分析方法是：首先根据已知条件绘出决策树；利用决策树，计算各个方案的损益期望值；根据损益期望值作出决策。我们通过例 6-3 说明这类决策的分析方法。

例 6-3　某产品有两种设计方案，年产量都是 100 台，预计生产 5 年，估计产品销路好的可能性是 0.7，销路差的可能性是 0.3。据测算方案 A 每台成本为 0.5 万元，销路好时每年盈利 100 万元，销路差时每年亏损 20 万元。方案 B 每台成本为 0.25 万元，销路好时每年盈利 40 万元，销路不好时每年盈利 5 万元，要求选定产品的方案。

解：(1) 由以上已知条件作决策树如图 6-6 所示。图中 a_0 是决策点，从 a_0 出发，画若干支线，每条支线代表一个方案，称作方案枝。方案枝的末端 a_1、a_2 称为自然状态点。再由自然状态点引出的分支线代表相应方案某种自然状态的概率，称为概率枝。在各概率枝的末端记下各方案在相应自然状态下可能盈利或损失的数值。至此决策树绘制完成。

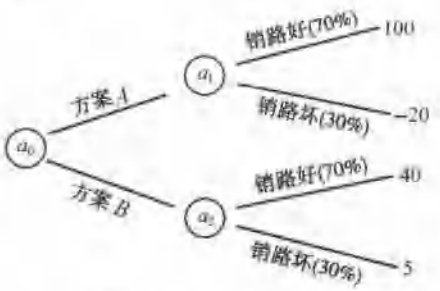


图 6-6　决策树示意图

(2) 计算各方案的损益期望值，其计算公式为

$$\text{损益期望值 } W = \text{盈利值} \times \text{盈利的概率} + \text{亏损值} \times \text{亏损的概率} \tag{6-7}$$

因此，方案 A、B 的损益期望值为

$$W_A = (100 \times 5 \times 0.7) + (-20 \times 5 \times 0.3) = 320(\text{万元})$$

$$W_B = (40 \times 5 \times 0.7) + (5 \times 5 \times 0.3) = 147.5(\text{万元})$$

(3) 根据各方案的损益期望值作出决策。由于各种方案的损益期望值,代表今后可能损失或盈利的数值,当然损益期望值愈大愈好。依据两种方案的计算结果,建议采用方案 A,因为它的盈利较多。

由于这种决策是建立在一定概率基础上的,它只代表了一种可能性,总有一定的风险,故称风险型决策。

6.2.3 非确定型决策

所谓非确定型决策是指决策问题的未来状态及其发生的概率均不能肯定时所进行的决策。对于这种类型决策,工程上常用等概率法、最大最小收益值法、乐观系数法来进行分析决策,选定最佳设计方案。下面通过例 6-4 来说明非确定型决策的原理。

例 6-4 设某产品有 A、B、C 三种设计方案,由于缺乏资料,对这种产品的市场需求量只能作大致估计,分为较高、一般、较低三种情况。该产品计划生产 6 年,根据计算,这三种方案在 6 年内的损益额计算值列于表 6-5。现用等概率法、乐观系数法等来进行该产品设计方案的非确定型决策。

表 6-5 某产品各方案损益额计算值 单位:万元

方案 / 市场估计	方案 A	方案 B	方案 C
需求量较高	600	800	400
需求量一般	400	250	250
需求量较低	0	−250	90

解:(1) 等概率法,即认为各种可能状态发生的概率均相等。本例中市场需求按三种情况考虑,故概率均为 $\frac{1}{3}$。各方案可能的损益值

方案 A $\quad \frac{1}{3}\times(600+400+0)=333.3$(万元)

方案 B $\quad \frac{1}{3}\times(800+350-250)=300$(万元)

方案 C $\quad \frac{1}{3}\times(400+250+90)=246.7$(万元)

按等概率法考虑,方案 A 的损益值最高,即 6 年后可能收益为 333.3 万元,故为最佳方案。

(2) 乐观系数法,即选取一个小于 1 的 α 值称为乐观系数,而 $1-\alpha$ 则称为悲观系数,用这两个系数分别对各方案的最大和最小收益值进行修正,估计可能取得的收益,以求得收益值最大的方案为最佳方案。

本例设取乐观系数 α 为 0.3,则悲观系数 $1-\alpha=0.7$,故各方案可能的收益值为

方案 A $\quad 0.3\times600+0.7\times0=180$(万元)

方案 B $\quad 0.3\times800+0.7\times(-250)=65$(万元)

方案 C $\quad 0.3\times400+0.7\times90=183$(万元)

由于方案 C 的可能收益值为最大,故为最佳方案。

综上所述,可见决策类型很多,方法也各不相同。设计中应根据评价结果及其主要目

标，正确确定决策类型，选用相应方案，才能获得预期的结果，作出正确决策。

6.3　模糊评价法

科学技术的发展不断深化，研究对象越来越复杂，日常生活中，人们常说：高、矮、胖、瘦，老、中、青；在评论产品质量时，人们常说：好、中、差；在产品设计中，作方案评价时，有些评价目标如舒适、美观、安全、便于加工等无法定量分析，也只能用好、一般、差来描述。这都是一些含义不确切、边界不清楚、没有量化的模糊概念。这里介绍的模糊评价法就是利用模糊数学将模糊信息数值化，再进行定量评价的方法。

6.3.1　模糊集合

既然模糊现象是事物客观存在的一种属性，因此是可以描述的，是有它自身规律的。1965 年，美国控制论专家查德(Zadeh)首先提出了模糊集合的概念，给出了模糊现象的定量描述方法，诞生了模糊数学。

1. 模糊集合

模糊集合是定量描述模糊概念的工具，是精确性与模糊性之间的桥梁，是普通集合的推广。模糊集合可表示为：

$$\underset{\sim}{A}=\frac{\mu_{\underset{\sim}{A}}(u_1)}{U_1}+\frac{\mu_{\underset{\sim}{A}}(u_2)}{U_2}+\cdots+\frac{\mu_{\underset{\sim}{A}}(u_n)}{U_n}=\sum_{i=1}^{n}\frac{\mu_{\underset{\sim}{A}}(u_i)}{U_i}$$

几点说明：

(1) $\mu_{\underset{\sim}{A}}(u_i)$为论域 U 中第 i 个元素 u_i，隶属于模糊集合 $\underset{\sim}{A}$ 的程度，简称为元素 u_i 的隶属度；$\mu_{\underset{\sim}{A}}(u)$为模糊集合 $\underset{\sim}{A}$ 的隶属函数，显然，隶属函数的值为隶属度。

(2) 符号“+”不是加号，“$\sum$”也不是求和，而是表示各元素与其隶属度对应关系的一个总括。

(3) $\frac{\mu_{\underset{\sim}{A}}(u_i)}{U_i}$不是分式，仅是一种约定的记号，“分母”是论域 U 中第 i 个元素，“分子”是相应元素的隶属度。

(4) $0\leqslant\mu_{\underset{\sim}{A}}(u_i)\leqslant 1$。

(5) 模糊集合完全由隶属函数决定。

(6) 论域 U 无限时，模糊集合可表示为

$$\underset{\sim}{A}=\int_{u\in U}\frac{\mu_{\underset{\sim}{A}}(u)}{U}$$

符号“$\int$”亦不表示积分。

通常还可把模糊集合简单地表示为：

$$\underset{\sim}{A}=(\mu_1,\mu_2,\cdots,\mu_n)$$

其中 μ_i 为第 i 个元素的隶属度。

2. 模糊集合的运算

模糊集合的运算很多就是普通集合运算的推广，常用到的有：

(1) 相等。对所有元素 x，若有 $\mu_{\underset{\sim}{A}}(x)=\mu_{\underset{\sim}{B}}(x)$，则称模糊集合 $\underset{\sim}{A}$ 与 $\underset{\sim}{B}$ 相等，记为 $\underset{\sim}{A}=\underset{\sim}{B}$。

(2) 包含。对所有元素 x，若有 $\mu_{\underset{\sim}{A}}(x)\leqslant\mu_{\underset{\sim}{B}}(x)$，则称模糊集合 $\underset{\sim}{B}$ 包含 $\underset{\sim}{A}$，记为 $\underset{\sim}{A}\subset\underset{\sim}{B}$。

(3) 并集。两个模糊集合 $\underset{\sim}{A}$ 和 $\underset{\sim}{B}$ 的并集 $\underset{\sim}{C}$ 仍为一模糊集合，其隶属函数为

$$\mu_{\underset{\sim}{C}}(x)=\max[\mu_{\underset{\sim}{A}}(x),\mu_{\underset{\sim}{B}}(x)]$$

也可表示为

$$\mu_{\underset{\sim}{C}}(x)=\mu_{\underset{\sim}{A}}(x)\vee\mu_{\underset{\sim}{B}}(x)$$

式中“∨”表示取大运算，记为 $\underset{\sim}{C}=\underset{\sim}{A}\cup\underset{\sim}{B}$。

(4) 交集。两个模糊集合 $\underset{\sim}{A}$ 与 $\underset{\sim}{B}$ 的交集 $\underset{\sim}{D}$ 仍为一模糊集合，其隶属函数为

$$\mu_{\underset{\sim}{D}}(x)=\min[\mu_{\underset{\sim}{A}}(x),\mu_{\underset{\sim}{B}}(x)]$$

或

$$\mu_{\underset{\sim}{D}}(x)=\mu_{\underset{\sim}{A}}(x)\wedge\mu_{\underset{\sim}{B}}(x)$$

式中“∧”表示取小运算，记为 $\underset{\sim}{D}=\underset{\sim}{A}\cap\underset{\sim}{B}$。

(5) 补集。模糊集合 $\underset{\sim}{A}$ 补集 $\overline{\underset{\sim}{A}}$ 仍为一模糊集合，其隶属函数为

$$\mu_{\overline{\underset{\sim}{A}}}(x)=1-\mu_{\underset{\sim}{A}}(x)$$

(6) 空集与全集。对所有 x，若有 $\mu_{\underset{\sim}{A}}(x)=0$，则称 $\underset{\sim}{A}$ 为空模糊集合，记为 $\varnothing$。

对所有 x，若有 $\mu_{\underset{\sim}{A}}(x)=1$，则称 $\underset{\sim}{A}$ 为全集合。

空集与全集互为补集。

6.3.2 隶属度及隶属函数

1. 隶属度

在模糊数学中，把隶属于或者从属于某个事物的程度叫隶属度，比如某方案对“操作安全”有七成符合，那么称此方案对“操作安全”的隶属度为 0.7。由于模糊概念对事物一般不是简单的肯定(1)或否定(0)，而是“亦此亦彼”，因此隶属度就可以用 0 到 1 之间的一个实数来表示。“1”表示完全隶属，“0”表示完全不隶属。

2. 隶属函数

描述从完全隶属到完全不隶属的渐变过程的函数叫隶属函数。模糊信息定量化，是通过隶属函数来实现的。确定隶属函数是较复杂和困难的。它既要反映设计参数的变化、设计实施的难易程度及变化规律，同时还要考虑实施的可能性及有关标准、规范等因素。

隶属函数种类很多。函数形式有直线式、曲线式。根据不同的评价对象选择合适的函数形式。现行使用的多为半矩形、半梯形、直线形。它们虽然只能近似地反映评价标准的隶属关系，但具有直观性、处理方便等优点。

3. 求隶属度的方法

(1) 通过抽样调查统计求隶属度。例如，对在市场上大量销售的某名牌电视机的图像显示清晰度进行评价，通过对 500 户抽样调查、统计结果，65%的用户反映图像很清晰、20%认为清晰、10%评价为一般，而 5%的用户反映不清晰，由此就得到对电视机图像显示四种评价的隶属度，它分别为 0.65、0.2、0.1 和 0.05。

(2) 通过隶属函数求隶属度。根据评价对象选择隶属函数，从中求得规定条件下的隶

属度。

例 6-5　一机械产品设计方案的成本 x，要求 $x\leqslant 2000$ 元为优，$x=2500$ 元为中等，$x\geqslant 3500$ 元为差，方案设计后，估算成本 2150 元，求模糊评价的隶属度。

解：根据题意，对于此类简单的计算，可采用梯形分布的隶属函数，如图 6-7 所示。函数表达式 $u(x)$ 如下所示。

优：
$$u(x)=\begin{cases}1 & (x\leqslant 2000)\\ \dfrac{2500-x}{2500-2000} & (2000<x<2500)\\ 0 & (x\geqslant 2500)\end{cases}$$

中：
$$u(x)=\begin{cases}0 & (x\leqslant 2500)\\ \dfrac{x-2000}{2500-2000} & (2000<x<2500)\\ 1 & (x=2500)\\ \dfrac{3500-x}{3500-2500} & (2500<x<3500)\\ 0 & (x\geqslant 3500)\end{cases}$$

差：
$$u(x)=\begin{cases}0 & (x\leqslant 2500)\\ \dfrac{x-2500}{3500-2500} & (2500<x<3500)\\ 1 & (x\geqslant 3500)\end{cases}$$

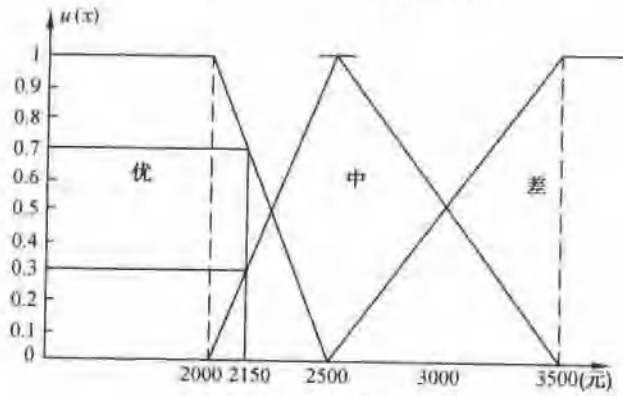


图 6-7　成本隶属函数

当方案估算成本为 2150 时，在图 6-7 中，求得 $u(x)_{优}=0.7$，$u(x)_{中}=0.3$，$u(x)_{差}=0$。

6.3.3　模糊评价方法及步骤

根据评价目标的数量，模糊评价分为单目标和多目标两种。

1. 单目标评价

(1) 建立评价集。评价者对评价对象可能作出的各种评判结果的集合叫评价集，用 u 表示，$u=\{u_1,u_2,\cdots,u_i,\cdots,u_m\}$。例如，前面对电视机图像显示清晰度评价，评价集 $u=\{u_1,u_2,u_3,u_4\}=\{$很清晰，清晰，一般，不好$\}$。

(2) 模糊评价集的表达式

$$R=\left\{\frac{r_1}{u_1},\frac{r_2}{u_2},\cdots,\frac{r_i}{u_i},\cdots,\frac{r_m}{u_m}\right\}$$

或者简写为

$$R=\{r_1,r_2,\cdots,r_i,\cdots,r_m\}$$

式中 r_i 为隶属度。

电视机图像模糊评价集的表达式

$$R=\{0.65,0.2,0.1,0.05\}$$

2. 多目标评价

(1) 建立评价目标集

$$\boldsymbol{X}=\{x_1,x_2,\cdots,x_i,\cdots,x_n\},\quad n=\text{目标数}$$

(2) 建立权重系数矩阵

$$\boldsymbol{G}=[g_1\quad g_2\quad\cdots\quad g_i\quad\cdots\quad g_n],\ \sum_{i=1}^{n}g_i=1$$

(3) 建立评价集

$$u=\{u_1,u_2,\cdots,u_i,\cdots,u_m\},\ m=\text{评价数}$$

(4) 建立一个方案对 n 个评价目标的模糊评价矩阵

$$\boldsymbol{R}=\begin{bmatrix}R_1\\R_2\\\cdots\\R_j\\\cdots\\R_n\end{bmatrix}=\begin{bmatrix}r_{11}&r_{12}&\cdots&r_{1j}&\cdots&r_{1m}\\r_{21}&r_{22}&\cdots&r_{2j}&\cdots&r_{2m}\\\cdots&&&&&\\r_{i1}&r_{i2}&\cdots&r_{ij}&\cdots&r_{im}\\\cdots&&&&&\\r_{n1}&r_{n2}&\cdots&r_{nj}&\cdots&r_{nm}\end{bmatrix}$$

考虑权重系数的模糊综合评价矩阵

$$\boldsymbol{B}=\boldsymbol{GR}=[g_1\quad g_2\quad\cdots\quad g_i\quad\cdots\quad g_n]\begin{bmatrix}r_{11}&r_{12}&\cdots&r_{1j}&\cdots&r_{1m}\\r_{21}&r_{22}&\cdots&r_{2j}&\cdots&r_{2m}\\\cdots&&&&&\\r_{i1}&r_{i2}&\cdots&r_{ij}&\cdots&r_{im}\\\cdots&&&&&\\r_{n1}&r_{n2}&\cdots&r_{nj}&\cdots&r_{nm}\end{bmatrix}$$

$$=[b_1\quad b_2\quad\cdots\quad b_i\quad\cdots\quad b_n]$$

b_i 是模糊综合评价集中的第 j 个隶属度，其计算是采用模糊矩阵合成的多种数学模型，现介绍常用的两种运算方法。

模型Ⅰ：$M(\wedge,\vee)$，按先取小($\wedge$)、后取大($\vee$)进行矩阵合成计算。

式中：M——模型；

"$\wedge$"、"$\vee$"——合成运算方式符号，若 $a\wedge b$ 取小者，若 $a\vee b$ 取大者。

$$b_j=\bigvee_{i=1}^{n}(g_i\wedge r_{ij})\quad(j=1,2,\cdots,m)\tag{6-8}$$

上式计算展开为下面算式

$$b_j=(g_1\wedge r_{1j})\vee(g_2\wedge r_{2j})\vee(g_3\wedge r_{3j})\vee\cdots(g_m\wedge r_{mn})\quad(j=1,2,\cdots,m)$$

取小取大运算，由于突出了 g_i 与 r_{ij} 中主要因素的影响，因此运算简单明确。但是在计

算中丢失了很多的 g_i 与 r_{ij} 的值，即丢失了很多评价信息，所以模型Ⅰ对于评价目标多、g_i 值很小，或者评价目标很少、g_i 值又较大的两种情况不适用。

模型Ⅱ：$M(\cdot,+)$：按先乘后加进行矩阵合成计算。

$$b_j = \sum_{i=1}^{n} g_i r_{ij} \quad (j = 1,2,\cdots,m) \tag{6-9}$$

该模型综合考虑了 g_i、r_{ij} 的影响，保留了全部信息，这是最显著的优点。由于评价实际效果好，故常用于机械产品的模糊综合评价和模糊优化设计。

3. 多方案的比较和决策

(1) 按各方案模糊综合评价中最高一级隶属度的数值大小定级，这称为最大隶属度法。

(2) 方案排队时，一方面以同级中隶属度高者为先，同时还要依据本级隶属度与更高一级隶属度之和的大小，排出方案先后。

例 6-6　对下面五个方案进行排队。

已知：评价集＝{最好，好，一般，差}

各方案模糊综合评价 B：

$$\boldsymbol{B}_1 = (0.33, 0.27, 0.4, 0)$$

$$\boldsymbol{B}_2 = (0.32, 0.3, 0.3, 0.08)$$

$$\boldsymbol{B}_3 = (0.45, 0.2, 0.3, 0.05)$$

$$\boldsymbol{B}_4 = (0.65, 0.2, 0.1, 0.05)$$

$$\boldsymbol{B}_5 = (0.5, 0.2, 0.1, 0.2)$$

解：(1) 按最大隶属度法，方案排序为 4、5、3、1、2。

(2) 按最高一级隶属度与第二级隶属度之和排出方案的各次顺序是：4、5、3、2 和 1。

例 6-7　对某型号推土机，三个设计方案的性能、使用进行模糊综合评价和决策。

解：(1) 分析和确定评价目标和权重系数，建立目标树。如图 6-8 所示为推土机评价目标树及权重系数分布。

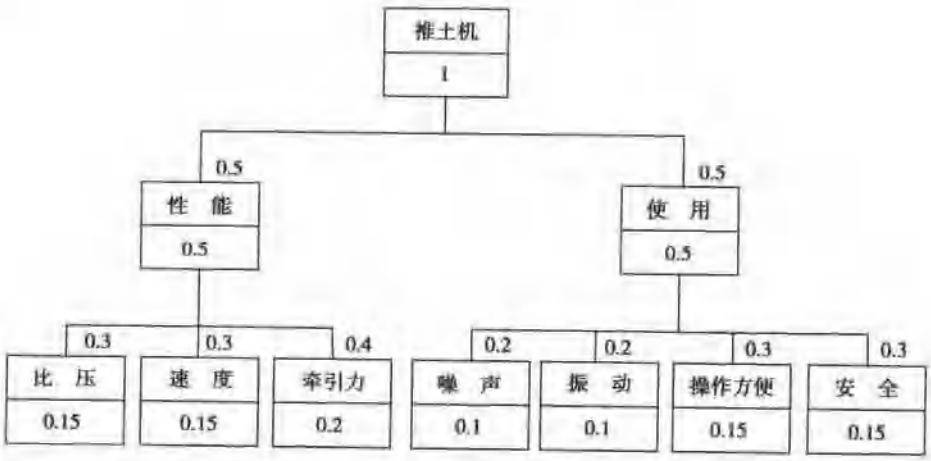


图 6-8　推土机评价目标树及权重系数分布

(2) 各方案评价目标的初步评语见表 6-6。

表 6-6

目标评语 / 方案	1	2	3	4	5	6	7
	比压	速度	牵引力	噪声	振动	操作方便	安全性
方案Ⅰ	差	一般	一般	差	差	一般	差
方案Ⅱ	一般	较好	一般	好	好	一般	好
方案Ⅲ	好	较好	差	好	好	一般	好

(3) 模糊评价。

① 评价目标集 $\boldsymbol{X}=\{x_1,x_2,x_3,x_4,x_5,x_6,x_7\}$

② 权重系数矩阵 $\boldsymbol{G}=[0.15\quad 0.15\quad 0.2\quad 0.1\quad 0.1\quad 0.15\quad 0.15]$

③ 评价集 $u=\{u_1,u_2,u_3,u_4\}=\{$优,良,中,差$\}$

④ 通过专家评审给分求得三个方案的隶属度矩阵

$$\boldsymbol{R}_{\mathrm{I}}=\begin{bmatrix}0 & 0 & 0.5 & 0.5\\ 0 & 0.25 & 0.5 & 0.25\\ 0 & 0.25 & 0.5 & 0.25\\ 0 & 0 & 0.5 & 0.5\\ 0 & 0 & 0.5 & 0.5\\ 0 & 0.25 & 0.5 & 0.25\\ 0 & 0 & 0.5 & 0.5\end{bmatrix}\qquad \boldsymbol{R}_{\mathrm{II}}=\begin{bmatrix}0 & 0.25 & 0.5 & 0.25\\ 0.25 & 0.5 & 0.25 & 0\\ 0 & 0.25 & 0.5 & 0.25\\ 0.6 & 0.25 & 0.15 & 0\\ 0.6 & 0.25 & 0.15 & 0\\ 0 & 0.25 & 0.5 & 0.25\\ 0.6 & 0.25 & 0.15 & 0\end{bmatrix}$$

$$\boldsymbol{R}_{\mathrm{III}}=\begin{bmatrix}0.6 & 0.25 & 0.15 & 0\\ 0.25 & 0.5 & 0.25 & 0\\ 0 & 0 & 0.5 & 0.5\\ 0.6 & 0.25 & 0.15 & 0\\ 0.6 & 0.25 & 0.15 & 0\\ 0 & 0.25 & 0.5 & 0.25\\ 0.6 & 0.25 & 0.15 & 0\end{bmatrix}$$

⑤ 求各方案模糊综合评价,按 M(∧,∨):

$\boldsymbol{B}_{\mathrm{I}}=\boldsymbol{G}\boldsymbol{R}_{\mathrm{I}}=(0.15,0.15,0.2,0.1,0.1,0.15,0.15)\cdot\boldsymbol{R}_{\mathrm{I}}=(b_1,b_2,b_3,b_4)$

根据(6-8)式得

$b_1=(0.15\wedge 0)\vee(0.15\wedge 0)\vee(0.2\wedge 0)\vee(0.1\wedge 0)\vee(0.1\wedge 0)\vee(0.15\wedge 0)\vee(0.15\wedge 0)=0$

同理 $b_2=0.2,b_3=0.2,b_4=0.2$

$\boldsymbol{B}_{\mathrm{II}}=\boldsymbol{G}\boldsymbol{R}_{\mathrm{II}}=(b_1,b_2,b_3,b_4)$

根据(6-8)式得

$b_1=(0.15\wedge 0)\vee(0.15\wedge 0.25)\vee(0.2\wedge 0)\vee(0.1\wedge 0.6)\vee(0.1\wedge 0.6)\vee(0.15\wedge 0)\vee(0.15\wedge 0.6)=0.15$

同理 $b_2=0.2;b_3=0.2;b_4=0.2$

$\boldsymbol{B}_{\mathrm{III}}=\boldsymbol{G}\boldsymbol{R}_{\mathrm{III}}=(b_1,b_2,b_3,b_4)$

根据(6-8)式得

$b_1=(0.15\wedge 0.6)\vee(0.15\wedge 0.25)\vee(0.2\wedge 0)\vee(0.1\wedge 0.6)\vee(0.1\wedge 0.6)\vee(0.15\wedge 0)\vee(0.15\wedge 0.6)=0.15$

同理　$b_2=0.15, b_3=0.2, b_4=0.2$

⑥ 各方案综合评价指标 B 的比较

$$\boldsymbol{B}_{\mathrm{I}}=(0,0.2,0.2,0.2)$$

$$\boldsymbol{B}_{\mathrm{II}}=(0.15,0.2,0.2,0.2)$$

$$\boldsymbol{B}_{\mathrm{III}}=(0.15,0.15,0.2,0.2)$$

为便于各方案的比较，将评价指标归一化，即 $\boldsymbol{B}=\left(\dfrac{b_1}{\sum_{j=1}^{m}b_j},\dfrac{b_2}{\sum_{j=1}^{m}b_j},\cdots,\dfrac{b_m}{\sum_{j=1}^{m}b_j}\right)$，得到三个方案模糊综合评价指标

$$\boldsymbol{B}'_{\mathrm{I}}=\left(\frac{0}{0.6},\frac{0.2}{0.6},\frac{0.2}{0.6},\frac{0.2}{0.6}\right)=(0,0.33,0.33,0.33)$$

$$\boldsymbol{B}'_{\mathrm{II}}=\left(\frac{0.15}{0.75},\frac{0.2}{0.75},\frac{0.2}{0.75},\frac{0.2}{0.75}\right)=(0.2,0.27,0.27,0.27)$$

$$\boldsymbol{B}'_{\mathrm{III}}=\left(\frac{0.15}{0.7},\frac{0.15}{0.7},\frac{0.2}{0.7},\frac{0.2}{0.7}\right)=(0.21,0.21,0.28,0.28)$$

三个方案按优劣排队顺序为Ⅲ、Ⅱ、Ⅰ，故选用第Ⅲ方案。

6.3.4　设计方案的三级模糊综合评判方法

目的是在已知问题后从众多的参评方案中选出较优的解答方案。工程上的方案各式各样，对于同一个问题，可以有不同原理的方案参评，也可以有同一原理下不同结构的方案候选。为此可将方案分成原理相同或原理不同两种。反映在评价中，原理不同的方案评价准则不同，原理相同的评价准则相同。

1. 评价准则相同

评价准则是决定方案评价成败的一个重要因素，将其适当分类，以不同方法处理，使评价客观、准确。为此，采用三级模糊综合评判。

(1) 三级模糊综合评判。这种方法的思路是评价时先按每一评价准则的各个等级进行一级模糊综合评判，再按每一类型的各个评价准则进行二级模糊综合评判，最后再在类型之间进行三级模糊综合评判。在已有三级模糊综合评判中，评判准则是按其性质进行分类的，分类后各类评价准则按二级模糊综合评判法处理。

实际上，方案评价时，评价准则有多有少，有时不一定需要分类处理。从形式上看评价准则有量化指标，也有非量化的模糊指标，若均以二级模糊综合评价法处理，对于量化指标就反而会引入更多的主观因素，因此本节把评价准则分为量化指标和非量化指标两大类，对量化指标以规范指标的评价法处理，非量化指标进行二级模糊综合评判，再综合求优评价出各个方案对等级——优、良、差的隶属度，其中以对优的隶属度最大的方案为最优方案。具体参见图 6-9。

图 6-9 设计方案模糊综合评判程序

(2) 量化指标的评判。在工程设计领域中，有些评价指标是有标准可循的，有些则可以根据经验或设计者的期望给出一个合理的量化范围。对于这两类指标采用基于量化指标的模糊综合评价。其主要步骤如下：

① 获取评判指标，建立评判指标矩阵。设参与评判的方案有 m 个，评价指标有 n 个，建立评判指标矩阵

$$U=(U_i)_{m\times n}$$

式中：U_i 表示第 i 个评判指标。

通过实测或其他方法，获取每个方案、每个评判指标的实测值，建立评价指标实测值矩阵

$$K=(k_{ij})_{m\times n}$$

式中 k_{ij} 为第 i 个方案的第 j 个指标的实测值。

② 建立量化指标集。由量化指标为元素组成集合

$$K=\{k_1,k_2,\cdots,k_n\}$$

其中 k_j 为评判指标 U_j 的希望值。对于具体指标 U_j 来说，其希望值 k_j 或为其上界 k_j，或为其下界 k_j，或为某一区间 $[\underline{k_j},\overline{k_j}]$。一般来说，各量化指标的希望值还具有不同程度的模糊性。因此视希望值为模糊量：

$$\boldsymbol{K}=(k_1,k_2,\cdots,k_n)$$

每一期望值 k_j，其区间如图 6-10 所示，用梯形函数来确定其隶属度，亦即用直线来近似过渡区间的隶属度，以便于计算机处理。图中 $\underline{k_j^U}$ 及 $\overline{k_j^U}$ 的值可根据参数的性质和问题的要求决定。

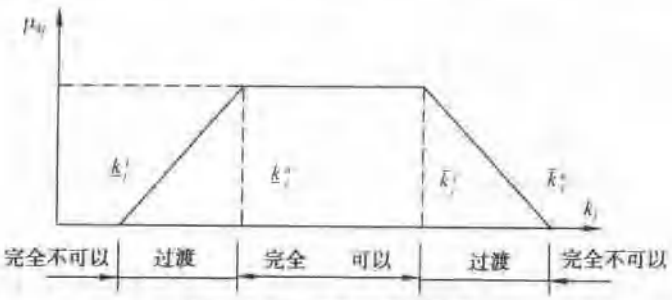


图 6-10　k_j 的隶属度

③ 建立满意度矩阵。设 μ_{ij} 为第 i 个评判对象的第 j 个评判指标的满意度($i=1,2,\cdots,m;j=1,2,\cdots,n$)，它反映评价指标的希望值 k_j 与其实测值 k_{ij} 之间的满意程度，以它为元，则组成满意度矩阵：

$$M=(\mu_{ij})_{m\times n}$$

根据 k_j 的不同情况，对满意度 μ_{ij} 分别计算如下：

(a) 当 $\underset{\sim}{k}_j$ 为模糊上界 $\overline{\underset{\sim}{k}}_j$ 时，令 $k_j=\overline{k}_j^L$，

$$\mu_{ij}=\begin{cases}1+\dfrac{\overline{k}_j^L-k_{ij}}{\overline{k}_j^L} & (k_{ij}\leqslant\overline{k}_j^U)\\ \mu_i & \\ 0 & (k_{ij}>\overline{k}_j^U)\end{cases}$$

式中 $\mu_i=\max\limits_i\left(1+\dfrac{\overline{k}_j^L-k_{ij}}{\overline{k}_j^L}\right)\quad(i=1,2,\cdots,m)$

(b) 当 $\underset{\sim}{k}_j$ 为模糊下界 $\underline{\underset{\sim}{k}}_j$ 时，令 $k_j=\underline{k}_j^U$

$$\mu_{ij}=\begin{cases}1+\dfrac{\underline{k}_j^U+k_{ij}}{\underline{k}_j^U} & (k_{ij}\geqslant\underline{k}_j^L)\\ \mu_i & \\ 0 & (k_{ij}<\underline{k}_j^L)\end{cases}$$

式中 $\mu_i=\max\limits_i\left(1+\dfrac{k_{ij}-\underline{k}_j^U}{\underline{k}_j^U}\right)\quad(i=1,2,\cdots,m)$

(c) 当 $\underset{\sim}{k}_j$ 为模糊区间$[\underline{\underset{\sim}{k}}_j,\overline{\underset{\sim}{k}}_j]$时，令 $\underline{k}_j=\underline{k}_j^U$，$\overline{k}_j=\overline{k}_j^L$

$$\mu_{ij}=\begin{cases}1 & (\underline{k}_j^U\leqslant k_{ij}\leqslant\underline{k}_j^L)\\ 1-\dfrac{k_{ij}-\overline{k}_j^L}{\overline{k}_j^L} & (\overline{k}_j^L\leqslant k_{ij}\leqslant\overline{k}_j^U)\\ 1-\dfrac{\underline{k}_j^U-k_{ij}}{\underline{k}_j^U} & (\underline{k}_j^L\leqslant k_{ij}\leqslant\underline{k}_j^U)\\ 0 & (k_{ij}<\underline{k}_j^L\text{ 或 }k_{ij}>\overline{k}_j^U)\end{cases}$$

应指出的是，某些实际问题，若某些希望值不具有模糊性，只需去掉过渡区间，令 $\overline{k}_j^L=\overline{k}_j^U$，$\underline{k}_j^L=\underline{k}_j^U$ 即可。

④ 建立权重集。给予权数 w_j 来反映各个指标 $U_j=(j=1,2,\cdots,n)$ 的重要程度，并以 w_j 为元组成权重集

$$\underset{\sim}{w}=\{w_1,w_2,w_3,\cdots,w_n\}=(w_j)_{1\times n}$$

各权数应满足归一性和非负性条件：

$$\sum_{j=1}^{n}w_j=1,\quad w_j\geqslant 0\quad (j=1,2,\cdots,n)$$

⑤ 综合评判。综合评判矩阵

$$A=\underset{\sim}{w}\cdot\underset{\sim}{M}=(w_j)_{1\times n}g(\mu_{ij})^{\mathrm{T}}_{m\times n}=(a_i)_{1\times m}$$

式中 $a_i=\sum_{j=1}^{n}lw_j\mu_{ij}$ 表示第 i 个方案的综合评判指标，其值越大，表示其质量越好。

(3) 非量化指标的评价。这里利用二级模糊评判法处理非量化指标。

① 建立评价指标集。设解决某确定问题的参评方案有 m 个，将影响方案评价的指标如设计水平高低、使用条件好坏、制造工艺性、维修性等组成评价指标集

$$U_{\mathrm{I}}=\{U_1,U_2,U_3,\cdots,U_n\}$$

其中 U_i 为第 i 个评价指标。

每一评价指标按其性质和程度细分为若干等级。一般说来，分的等级越多，越易于反映实际情况，但相应地给其隶属度的确定带来困难，本节将指标分为五等。将所有评价指标的等级组成集合，得到评价指标等级矩阵

$$\boldsymbol{U}_{\mathrm{II}}=(U_{ij})_{n\times 5}$$

其中 U_{ij} 为第 i 个指标的第 j 个等级。

各指标等级之间，很难有一个明确的界限，任一指标等级都在其后相邻的两等级之间处于某种模糊的分布状态。由于评价指标的模糊性及等级的模糊性，很难甚至不可能把某一评价指标，具体地规定为它的某一等级，因此，各模糊评价指标应视为等级论域上的模糊子集，即

$$\underset{\sim}{U_i}=\frac{\mu_{i1}}{U_{i1}}+\frac{\mu_{i2}}{U_{i2}}+\cdots+\frac{\mu_{iS}}{U_{iS}}$$

其中 $0\leqslant\mu_{ij}\leqslant 1$ 为第 i 个评价指标的第 j 个等级该指标的隶属度。

② 建立备择集。方案评价的备择集应是方案的优劣程度，即优、良、中、差等。方案评价主要是找出每种方案对优的隶属度，将方案的优劣程度划分太细无实际意义，分三等即可，由此备择集为

$$V=\{\text{优},\text{中},\text{差}\}$$

③ 一级模糊综合评判。设参评方案按第 i 个评价指标的第 j 个等级 U_{ij} 评判，对备择集中第 k 个元素的隶属度为 $r^i_{jk}(k=1,2,3)$，则第 i 个评价指标的等级评判集可表示为

$$\underset{\sim}{R^i_j}=\frac{r^i_{j1}}{(U_{ij},V_1)}+\frac{r^i_{j2}}{(U_{ij},V_2)}+\frac{r^i_{j3}}{(U_{ij},V_3)}$$

以上述等级评判集的隶属度为行组成矩阵便是第 i 个评价指标的等级评判矩阵

$$\underset{\sim}{\boldsymbol{R}_i}=(r^i_{jk})_{5\times 3}$$

把第 i 个指标的第 j 个等级对该指标的隶属度 $\mu_{ij}(i=1,2,\cdots,n;j=1,2,\cdots,5)$ 归一化后的值

$$w_{ij}=\frac{\mu_{ij}}{\sum_{j=1}^{5}\mu_{ij}}\quad (i=1,2,\cdots,n)$$

取作为该等级的权数，便可得到第 i 个指标的等级权重集为

$$\underset{\sim}{w}_i=(w_{i1},w_{i2},w_{i3},\cdots,w_{in})$$

至此，按第 i 个指标的各等级模糊子集进行综合评判，便得一级模糊综合评判如下：[按 $M(\cdot,+)$ 计算]

$$\underset{\sim}{\boldsymbol{A}}=\underset{\sim}{\boldsymbol{w}}_i\underset{\sim}{\boldsymbol{R}}_i=(a_{i1},a_{i2},a_{i3})$$

式中 $a_{ik}=\sum_{j=1}^{5}w_{ij}\cdot r_{jk}^{i}(i=1,2,\cdots,n;k=1,2,3)$ 即为综合考虑第 i 个评价指标的各个等级贡献时，参评方案对优、中、差三个等级的隶属度。以 a_{ik} 为元素便得到一级模糊综合评判矩阵：

$$\underset{\sim}{\boldsymbol{A}}_i=\underset{\sim}{\boldsymbol{w}}_i\underset{\sim}{\boldsymbol{R}}_i=(a_{i1},a_{i2},a_{i3})$$

④ 二级模糊综合评判。一级模糊综合评判仅反映了一个评价指标的影响，矩阵 A 即为二级模糊综合评判的单因素评判矩阵。

设 w_i 为第 i 个评价指标的权数，则反映各评价指标重要程度的权重为

$$\underset{\sim}{w}=(w_1,w_2,w_3,\cdots,w_n)=(w_i)_{1\times n}$$

w 为评价指标集上的模糊子集，各权数应满足

$$\sum_{i=1}^{n}w_i=1,w_i\geqslant 0(i=1,2,\cdots,n)$$

于是按所有评价指标进行综合评判便得二级模糊综合评判矩阵如下：

$$\underset{\sim}{\boldsymbol{B}}=\underset{\sim}{w}\underset{\sim}{\boldsymbol{A}}=(b_1,b_2,b_3)$$

式中 $b_k=\sum_{i=1}^{n}w_iw_{ik}(k=1,2,3)$ 为综合考虑所有评价指标时，参评方案对优、中、差三个等级的隶属度。

若有 m 个方案，便得矩阵

$$\underset{\sim}{\boldsymbol{B}}=(b_{mk})_{m\times 3}$$

(4) 综合求优。若参评方案既有量化指标，又有非量化指标，并且已获得量化指标和非量化指标的权数分别为 w_A、w_B。

由前述知，量化指标的满意度矩阵为 $\boldsymbol{A}(a_i)_{1\times m}$，$m$ 个方案的非量化指标对优、中、差的隶属度矩阵为 $\underset{\sim}{\boldsymbol{B}}=(b_{mk})_{m\times 3}$ 进行综合评判，便得三级模糊综合评判矩阵

$$\underset{\sim}{\boldsymbol{C}}=w_A\boldsymbol{A}+w_B\boldsymbol{B}=(C_i)_{1\times m}$$

式中 $C_i=w_Aa_i+w_Bb_i$ 第 i 个方案对优的隶属度。找出 $C_k=\max_i\{C_i\}(i=1,2,\cdots,m)$ 即为最优方案。

2. 评价准则不同

当参评方案的评价准则不同时，除以下两点外，方法、步骤与评价准则相同时一样。

(1) 评价准则不同时，只能逐个方案地进行评价，每个方案依自己的评价准则计算出对优的隶属度，最后再找出对优隶属度最大的方案。

(2) 评价准则不同时，对量化指标中求满意度的分式修改为：

① 当 $\underset{\sim}{k}_j$ 为模糊上界 $\overline{\underset{\sim}{k}}_j$ 时，令 $k_j=\overline{k}_j^{L}$

$$\mu_{ij}=\begin{cases}1 & (k_{ij}\leqslant\overline{k}_j^L)\\ \dfrac{\overline{k}_j^U-k_{ij}}{\overline{k}_j^U-\overline{k}_j^L} & (\overline{k}_j^L\leqslant k_{ij}\leqslant\overline{k}_j^U)\\ 0 & (k_{ij}>\overline{k}_j^U)\end{cases}$$

② 当 $\underset{\sim}{k}_j$ 为模糊下界 $\underline{\underset{\sim}{k}}_j$ 时，令 $k_j=\underline{k}_j^U$

$$\mu_{ij}=\begin{cases}1 & (k_{ij}\geqslant\underline{k}_j^U)\\ \dfrac{k_{ij}-\underline{k}_j^L}{\underline{k}_j^U-\underline{k}_j^L} & (\underline{k}_j^L\leqslant k_{ij}\leqslant\underline{k}_j^U)\\ 0 & (k_{ij}>\underline{k}_j^L)\end{cases}$$

③ 当 $\underset{\sim}{k}_j$ 为模糊区间 $[\underline{\underset{\sim}{k}}_j,\overline{\underset{\sim}{k}}_j]$ 时

$$\mu_{ij}=\begin{cases}1 & (\underline{k}_j^U\leqslant k_{ij}\leqslant\overline{k}_j^L)\\ \dfrac{\overline{k}_j^U-k_{ij}}{\overline{k}_j^U-\overline{k}_j^L} & (\overline{k}_j^L\leqslant k_{ij}\leqslant\overline{k}_j^U)\\ \dfrac{k_{ij}-\underline{k}_j^L}{\underline{k}_j^U-\underline{k}_j^L} & (\underline{k}_j^L\leqslant k_{ij}\leqslant\underline{k}_j^U)\\ 0 & (k_{ij}\geqslant\overline{k}_j^U\ or\ k_{ij}<\underline{k}_j^L)\end{cases}$$

3. 在方案评价中应尽量减少主观因素的影响

方案评价归根结底要由人来完成，要使评价结果客观、公正，应尽量减少主观因素的影响。

(1) 评价准则的选定。评价准则是方案评价的依据，应由给定的问题和具体方案给出评价准则，应力求全面、丰富、具体、客观、准确地反映实际情况。一般由专家咨询或专家小组会议商定。

(2) 权数的确定。权数可由专家给定，也可用判别表法计算。

(3) 隶属度。在三级模糊评价中，受到主观影响的隶属度主要是等级评判矩阵。等级评判矩阵可由专家确定数值，也可由隶属度函数等距离移动求值得到。应尽量使等级评判矩阵近似对称矩阵。

合理的评价使多种方案得到充分比较，为科学的决策创造有利的条件，是设计中选取最佳方案不可缺少的步骤。

根据任务要求确定评价目标项目并判别各目标的重要程度(以加权系数定量表达)，这是方案评价前必须做的准备工作。

针对不同的评价对象和目的可采用相应的评价方法。例如，只要求作定性评价，对各方案排列顺序时，可采用点评价法和名次计分法。一般评分法应用较多，普通的工程方案用加权计分的有效值法既不复杂又很实用。技术-经济评价法求出对于理想方案的相对评价值，在评价过程中有利于找出方案的弱点加以改进。模糊评价法可使各种模糊评价概念定量化，便于计算机在评价中的全面应用。

例 6-8 四、六级英语考试通过率评价。

(1) 评价方法与步骤。对于全国大学英语四、六级考试通过率预测评价问题，其评价方法与步骤如下：

① 影响因素集　$X=\{x_1,x_2,x_3\}=$｛平时水平，摸底测试水平，同等水平历年正式考试通过率｝。其中：≥85 分为优，75～85 分为良，60～75 分为中，<60 分为差。

② 权重系数矩阵　$G=[g_1\quad g_2\quad g_3]=$[平时水平权重　摸底测试水平权重　同等水平历年正式考试通过率权重]。其中：平时水平权重＝0.2，摸底测试水平权重＝0.5，同等水平历年正式考试通过率权重＝0.3。

③ 评价集　$U=\{u_1,u_2,u_3,u_4\}=$｛通过率高，通过率较高，有希望通过，不能通过｝。其中：≥0.85 为通过率高，>0.75～0.85 为通过率较高，0.60～0.75 为有希望通过，<0.60 为不能通过。

④ 模糊综合评价：

建立一个方案对 n 个评价目标的模糊评价矩阵；

考虑权重系数的模糊综合评价矩阵；

模糊综合评价值通过专家数据库对应得分。

(2) 预测评价实验。按照上述建立的全国大学英语四、六级考试通过率预测评价模型，作者从 1999 年开始进行实验评价研究，下面举例说明该四、六级考试通过率预测评价实验研究情况。

表 6-7　四、六级考试通过率评价目标

目标评语 / 考生	1	2	3
	平时水平	摸底测试水平	同等水平历年考试通过率
考生Ⅰ	86(优)	82(良)	0.85～0.91
考生Ⅱ	74(中)	76(良)	0.62～0.67
考生Ⅲ	71(中)	59(差)	0.41～0.50
考生Ⅳ	79(良)	77(良)	0.62～0.67

(3) 预测评价。通过专家评审给分求得四个考生的隶属度矩阵

$$\boldsymbol{R}_{\mathrm{I}}=\begin{bmatrix}0.6 & 0.25 & 0.15 & 0\\ 0.25 & 0.5 & 0.25 & 0\\ 0.25 & 0.5 & 0.25 & 0\end{bmatrix}\qquad \boldsymbol{R}_{\mathrm{II}}=\begin{bmatrix}0 & 0.25 & 0.5 & 0.25\\ 0.25 & 0.5 & 0.25 & 0\\ 0 & 0.25 & 0.5 & 0.25\end{bmatrix}$$

$$\boldsymbol{R}_{\mathrm{III}}=\begin{bmatrix}0 & 0.25 & 0.5 & 0.25\\ 0 & 0 & 0.5 & 0.5\\ 0 & 0 & 0.5 & 0.5\end{bmatrix}\qquad \boldsymbol{R}_{\mathrm{IV}}=\begin{bmatrix}0.25 & 0.5 & 0.25 & 0\\ 0.25 & 0.5 & 0.25 & 0\\ 0 & 0.25 & 0.5 & 0.25\end{bmatrix}$$

模型Ⅰ求各考生模糊综合评价：

$$\boldsymbol{B}_{\mathrm{I}}=\boldsymbol{G}\boldsymbol{R}_{\mathrm{I}}=(0.2,0.5,0.3)\boldsymbol{R}_{\mathrm{I}}=(b_1,b_2,b_3,b_4)$$

从而有　$b_1=(0.2\wedge 0.6)\vee(0.5\wedge 0.25)\vee(0.3\wedge 0.25)=0.25$

同理　$b_2=0.5,b_3=0.25,b_4=0$

$$\boldsymbol{B}_{\mathrm{II}}=\boldsymbol{G}\boldsymbol{R}_{\mathrm{II}}=(b_1,b_2,b_3,b_4)$$

$$b_1=0.25,b_2=0.5,b_3=0.3,b_4=0.25$$

$$\boldsymbol{B}_{\mathrm{III}}=\boldsymbol{G}\boldsymbol{R}_{\mathrm{III}}=(b_1,b_2,b_3,b_4)$$

$$b_1=0,b_2=0.2,b_3=0.5,b_4=0.5$$

$$\boldsymbol{B}_{\mathrm{IV}}=\boldsymbol{G}\boldsymbol{R}_{\mathrm{IV}}=(b_1,b_2,b_3,b_4)$$

$$b_1=0.25, b_2=0.5, b_3=0.3, b_4=0.25$$

各考生综合评价指标 $\boldsymbol{B}$ 的比较

$$\boldsymbol{B}_{\mathrm{I}}=(0.25,0.5,0.25,0)$$
$$\boldsymbol{B}_{\mathrm{II}}=(0.25,0.5,0.3,0.25)$$
$$\boldsymbol{B}_{\mathrm{III}}=(0,0.2,0.5,0.5)$$
$$\boldsymbol{B}_{\mathrm{IV}}=(0.25,0.5,0.3,0.25)$$

为便于各考生的比较，将评价指标归一化，即 $\boldsymbol{B}=\left(\frac{b_1}{\sum_{j=1}^{m}b_j},\frac{b_2}{\sum_{j=1}^{m}b_j},\cdots,\frac{b_m}{\sum_{j=1}^{m}b_j}\right)$，得到四个考生模糊综合评价指标

$$\boldsymbol{B}_{\mathrm{I}}'=\left(\frac{0.25}{1},\frac{0.5}{1},\frac{0.25}{1},\frac{0}{1}\right)=(0.25,0.5,0.25,0)$$
$$\boldsymbol{B}_{\mathrm{II}}'=\left(\frac{0.25}{1.3},\frac{0.5}{1.3},\frac{0.3}{1.3},\frac{0.25}{1.3}\right)=(0.19,0.38,0.23,0.19)$$
$$\boldsymbol{B}_{\mathrm{III}}'=\left(\frac{0}{1.2},\frac{0.2}{1.2},\frac{0.5}{1.2},\frac{0.5}{1.2}\right)=(0,0.17,0.42,0.42)$$
$$\boldsymbol{B}_{\mathrm{IV}}'=\left(\frac{0.25}{1.3},\frac{0.5}{1.3},\frac{0.3}{1.3},\frac{0.25}{1.3}\right)=(0.19,0.38,0.23,0.19)$$

这样四个考生按成功可能性的排队顺序就为：Ⅰ、Ⅱ和Ⅳ、Ⅲ，这里考生Ⅱ和Ⅳ并列第二。

模型Ⅱ求各考生模糊综合评价：

$$\boldsymbol{B}_{\mathrm{I}}=\boldsymbol{G}\boldsymbol{R}_{\mathrm{I}}=(0.2,0.5,0.3)\boldsymbol{R}_{\mathrm{I}}=(b_1,b_2,b_3,b_4)$$

从而有 $b_1=0.2\times0.6+0.5\times0.25+0.3\times0.25=0.32$

同理 $b_2=0.45, b_3=0.23, b_4=0$

$$\boldsymbol{B}_{\mathrm{II}}=\boldsymbol{G}\boldsymbol{R}_{\mathrm{II}}=(b_1,b_2,b_3,b_4)$$
$$b_1=0.125, b_2=0.375, b_3=0.375, b_4=0.125$$
$$\boldsymbol{B}_{\mathrm{III}}=\boldsymbol{G}\boldsymbol{R}_{\mathrm{III}}=(b_1,b_2,b_3,b_4)$$
$$b_1=0, b_2=0.05, b_3=0.5, b_4=0.45$$
$$\boldsymbol{B}_{\mathrm{IV}}=\boldsymbol{G}\boldsymbol{R}_{\mathrm{IV}}=(b_1,b_2,b_3,b_4)$$
$$b_1=0.175, b_2=0.425, b_3=0.325, b_4=0.075$$

由此，四个考生模糊综合评价指标为

$$\boldsymbol{B}_{\mathrm{I}}=(0.32,0.45,0.23,0)$$
$$\boldsymbol{B}_{\mathrm{II}}=(0.125,0.375,0.375,0.125)$$
$$\boldsymbol{B}_{\mathrm{III}}=(0,0.05,0.5,0.45)$$
$$\boldsymbol{B}_{\mathrm{IV}}=(0.175,0.425,0.325,0.075)$$

这样四个考生按成功可能性的排队顺序就为：Ⅰ、Ⅳ、Ⅱ、Ⅲ。

(4) 预测结果及分析。从上述研究可以看出：模糊Ⅰ预测成功可能性的排队顺序为：Ⅰ、Ⅳ和Ⅱ、Ⅲ(Ⅱ和Ⅳ并列)，模型Ⅱ预测成功可能性的排队顺序为：Ⅰ、Ⅳ、Ⅱ、Ⅲ(Ⅱ在Ⅳ之后)。比较两个模型的预测结果可知，模型Ⅱ的预测精度较高。

模型Ⅰ由于突出了主要因素的影响，因此运算(取小、取大运算)简单。但是在计算中丢

失了很多 g_i 与 r_{ij} 的值，即丢失了很多的评价信息，所以模型Ⅰ对于评价目标多、g_i 值很小，或者评价目标很少、g_i 值又较大的两种情况不适用。模型Ⅱ综合考虑了 g_i、r_{ij} 的影响，保留了全部信息，这是最显著的优点，故评价实际效果较模型Ⅰ更好。

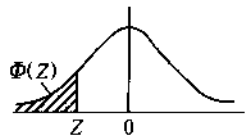
(5) 结论。① 把模糊理论用于预测四、六级英语水平考试的通过率，经过几年的连续评价实验证明，使用该评价方法可较好地对考生考试通过率进行评价，克服了考生盲目报名、造成考试通过率低、影响考生考试的信心等问题。

② 本文通过同一组实验数据分别用模型Ⅰ和模型Ⅱ预测后的结果比较，指出模型Ⅱ对四、六级英语水平考试通过率的预测比较合适。

习　题　6

1. 试述设计中的评价与决策研究的内容与意义。
2. 试述评价目标加权系数的作用。
3. 试述设计中的决策的基本原则及其分析方法。
4. 试述评价中的模糊隶属度及隶属函数。
5. 试述模糊评价方法及步骤。

附表 标准正态分布表 $\Phi(z)=\int_{-\infty}^{z}\frac{1}{\sqrt{2\pi}}e^{-z^2/2}dZ=P\{Z\leqslant z\}$



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	z
−0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641	−0.0
−0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247	−0.1
−0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859	−0.2
−0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483	−0.3
−0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121	−0.4
−0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776	−0.5
−0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451	−0.6
−0.7	0.2420	0.2389	0.2358	0.2327	0.2297	0.2266	0.2236	0.2206	0.2177	0.2148	−0.7
−0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867	−0.8
−0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611	−0.9
−1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379	−1.0
−1.1	0.1357	0.1335	0.1311	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170	−1.1
−1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.09853	−1.2
−1.3	0.09680	0.09510	0.09342	0.09176	0.09012	0.03851	0.08691	0.08534	0.08379	0.08226	−1.3
−1.4	0.08076	0.07927	0.07780	0.07636	0.07493	0.07353	0.07215	0.07078	0.06944	0.06811	−1.4
−1.5	0.06681	0.06552	0.06426	0.06301	0.06178	0.06057	0.05938	0.05821	0.05705	0.05592	−1.5
−1.6	0.05480	0.05370	0.05262	0.05155	0.05050	0.04947	0.04846	0.04746	0.04648	0.04551	−1.6
−1.7	0.04457	0.04363	0.04272	0.04182	0.04093	0.04006	0.03920	0.03836	0.03754	0.03673	−1.7
−1.8	0.03593	0.03515	0.03438	0.03362	0.03288	0.03216	0.03144	0.03074	0.03005	0.02938	−1.8
−1.9	0.02872	0.02807	0.02743	0.02680	0.02619	0.02559	0.02500	0.02442	0.02385	0.02330	−1.9
−2.0	0.02275	0.02222	0.02169	0.02118	0.02068	0.02018	0.01970	0.01923	0.01876	0.01831	−2.0
−2.1	0.01786	0.01743	0.01700	0.01659	0.01618	0.01578	0.01539	0.01500	0.01463	0.01426	−2.1
−2.2	0.01390	0.01355	0.01321	0.01287	0.01255	0.01222	0.01191	0.01160	0.01130	0.01101	−2.2
−2.3	0.01072	0.01044	0.01017	$0.0^2$9903	$0.0^2$9642	$0.0^2$9387	$0.0^2$9137	$0.0^2$8894	$0.0^2$8656	$0.0^2$8424	−2.3
−2.4	$0.0^2$8198	$0.0^2$7976	$0.0^2$7760	$0.0^2$7549	$0.0^2$7344	$0.0^2$7143	$0.0^2$6947	$0.0^2$6756	$0.0^2$6569	$0.0^2$6387	−2.4
−2.5	$0.0^2$6210	$0.0^2$6037	$0.0^2$5868	$0.0^2$5703	$0.0^2$5543	$0.0^2$5386	$0.0^2$5234	$0.0^2$5085	$0.0^2$4940	$0.0^2$4799	−2.5
−2.6	$0.0^2$4661	$0.0^2$4527	$0.0^2$4396	$0.0^2$4269	$0.0^2$4145	$0.0^2$4025	$0.0^2$3907	$0.0^2$3793	$0.0^2$3681	$0.0^2$3573	−2.6
−2.7	$0.0^2$3467	$0.0^2$3364	$0.0^2$3264	$0.0^2$3167	$0.0^2$3072	$0.0^2$2930	$0.0^2$2890	$0.0^2$2803	$0.0^2$2718	$0.0^2$2635	−2.7
−2.8	$0.0^2$2555	$0.0^2$2477	$0.0^2$2401	$0.0^2$2327	$0.0^2$2256	$0.0^2$2186	$0.0^2$2118	$0.0^2$2052	$0.0^2$1938	$0.0^2$1926	−2.8
−2.9	$0.0^2$1866	$0.0^2$1807	$0.0^2$1750	$0.0^2$1695	$0.0^2$1641	$0.0^2$1589	$0.0^2$1538	$0.0^2$1489	$0.0^2$1441	$0.0^2$1395	−2.9
−3.0	$0.0^2$1350	$0.0^2$1306	$0.0^2$1264	$0.0^2$1223	$0.0^2$1183	$0.0^2$1144	$0.0^2$1107	$0.0^2$1070	$0.0^2$1035	$0.0^2$1001	−3.0
−3.1	$0.0^3$9676	$0.0^3$9354	$0.0^3$9043	$0.0^3$8740	$0.0^3$8447	$0.0^3$8164	$0.0^3$7888	$0.0^3$7622	$0.0^3$7364	$0.0^3$7114	−3.1
−3.2	$0.0^3$6871	$0.0^3$6637	$0.0^3$6410	$0.0^3$6190	$0.0^3$5976	$0.0^3$5770	$0.0^3$5571	$0.0^3$5377	$0.0^3$5190	$0.0^3$5009	−3.2
−3.3	$0.0^3$4834	$0.0^3$4665	$0.0^3$1501	$0.0^3$4342	$0.0^4$4189	$0.0^3$4041	$0.0^3$3897	$0.0^3$3758	$0.0^3$3624	$0.0^3$3495	−3.3
3.4	$0.0^3$3369	$0.0^3$3248	$0.0^3$3131	$0.0^3$3018	$0.0^3$2909	$0.0^3$2803	$0.0^3$2701	$0.0^3$2602	$0.0^3$2507	$0.0^3$2415	−3.4
−3.5	$0.0^3$2326	$0.0^3$2241	$0.0^3$2158	$0.0^3$2078	$0.0^4$2001	$0.0^3$1926	$0.0^3$1854	$0.0^3$1785	$0.0^3$1718	$0.0^3$1653	−3.5
−3.6	$0.0^3$1591	$0.0^3$1531	$0.0^3$1473	$0.0^3$1417	$0.0^3$1363	$0.0^3$1311	$0.0^3$1261	$0.0^3$1213	$0.0^3$1166	$0.0^3$1121	−3.6

续表

x	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	x
−3.7	$0.0^{3}1078$	$0.0^{3}1036$	$0.0^{4}9961$	$0.0^{4}9574$	$0.0^{4}9201$	$0.0^{4}3842$	$0.0^{4}8496$	$0.0^{4}8162$	$0.0^{4}7841$	$0.0^{4}7532$	−3.7
−3.8	$0.0^{4}7235$	$0.0^{4}6948$	$0.0^{4}6673$	$0.0^{4}6407$	$0.0^{4}6152$	$0.0^{4}5906$	$0.0^{4}5669$	$0.0^{4}5442$	$0.0^{4}5223$	$0.0^{4}5012$	−3.8
−3.9	$0.0^{4}4810$	$0.0^{4}4615$	$0.0^{4}4427$	$0.0^{4}4247$	$0.0^{4}4074$	$0.0^{4}3908$	$0.0^{4}3747$	$0.0^{4}3594$	$0.0^{4}3446$	$0.0^{4}3304$	−3.9
−4.0	$0.0^{4}3167$	$0.0^{4}3036$	$0.0^{4}2910$	$0.0^{4}2789$	$0.0^{4}2673$	$0.0^{4}2561$	$0.0^{4}2454$	$0.0^{4}2351$	$0.0^{4}2252$	$0.0^{4}2157$	−4.0
−4.1	$0.0^{4}2066$	$0.0^{4}1978$	$0.0^{4}1894$	$0.0^{4}1814$	$0.0^{4}1737$	$0.0^{4}1662$	$0.0^{4}1591$	$0.0^{4}1523$	$0.0^{4}1458$	$0.0^{4}1395$	−4.1
−4.2	$0.0^{4}1335$	$0.0^{4}1277$	$0.0^{4}1222$	$0.0^{4}1168$	$0.0^{4}1118$	$0.0^{4}1069$	$0.0^{4}1022$	$0.0^{3}9774$	$0.0^{3}9345$	$0.0^{3}8934$	−4.2
−4.3	$0.0^{3}8540$	$0.0^{3}8163$	$0.0^{3}7801$	$0.0^{3}7455$	$0.0^{3}7124$	$0.0^{3}6807$	$0.0^{3}6503$	$0.0^{3}6212$	$0.0^{3}5934$	$0.0^{3}5668$	−4.3
−4.4	$0.0^{3}5413$	$0.0^{3}5169$	$0.0^{3}4935$	$0.0^{3}4712$	$0.0^{3}4498$	$0.0^{3}4294$	$0.0^{3}4098$	$0.0^{3}3911$	$0.0^{3}3732$	$0.0^{3}3561$	−4.4
−4.5	$0.0^{3}3398$	$0.0^{3}3241$	$0.0^{3}3092$	$0.0^{3}2949$	$0.0^{3}2813$	$0.0^{3}2682$	$0.0^{3}2558$	$0.0^{3}2439$	$0.0^{3}2325$	$0.0^{3}2216$	−4.5
4.6	$0.0^{3}2112$	$0.0^{3}2013$	$0.0^{3}1919$	$0.0^{3}1828$	$0.0^{3}1742$	$0.0^{3}1660$	$0.0^{3}1581$	$0.0^{3}1506$	$0.0^{3}1434$	$0.0^{3}1366$	−4.6
−4.7	$0.0^{3}1301$	$0.0^{3}1239$	$0.0^{3}1179$	$0.0^{3}1123$	$0.0^{3}1069$	$0.0^{3}1017$	$0.0^{6}9680$	$0.0^{6}9211$	$0.0^{6}8765$	$0.0^{6}8339$	−4.7
−4.8	$0.0^{6}7933$	$0.0^{6}7547$	$0.0^{6}7178$	$0.0^{6}6827$	$0.0^{6}6492$	$0.0^{6}6173$	$0.0^{6}5869$	$0.0^{6}5580$	$0.0^{6}5304$	$0.0^{6}5042$	−4.8
−4.9	$0.0^{6}4792$	$0.0^{6}4554$	$0.0^{6}4327$	$0.0^{6}4111$	$0.0^{6}3906$	$0.0^{6}3711$	$0.0^{6}3525$	$0.0^{6}3348$	$0.0^{6}3179$	$0.0^{6}3019$	−4.9
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359	0.0
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753	0.1
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141	0.2
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517	0.3
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879	0.4
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224	0.5
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549	0.6
0.7	0.7580	0.7611	0.7642	0.7673	0.7703	0.7734	0.7764	0.7794	0.7823	0.7852	0.7
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133	0.8
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389	0.9
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621	1.0
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830	1.1
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.90147	1.2
1.3	0.90320	0.90490	0.90658	0.90824	0.90988	0.91149	0.91309	0.91466	0.91621	0.91774	1.3
1.4	0.91924	0.92073	0.92220	0.92364	0.92507	0.92647	0.92785	0.92922	0.93056	0.93189	1.4
1.5	0.93319	0.93448	0.93574	0.93699	0.93822	0.93943	0.94062	0.94179	0.94295	0.94408	1.5
1.6	0.94520	0.94630	0.94738	0.94845	0.94950	0.95053	0.95154	0.95254	0.95352	0.95449	1.6
1.7	0.95543	0.95637	0.95728	0.95818	0.95907	0.95994	0.96080	0.96164	0.96246	0.96327	1.7
1.8	0.96407	0.96485	0.96562	0.96638	0.96712	0.96784	0.96856	0.96926	0.96995	0.97062	1.8
1.9	0.97128	0.97193	0.97257	0.97320	0.97381	0.97441	0.97500	0.97558	0.97615	0.97670	1.9
2.0	0.97725	0.97778	0.97831	0.97882	0.97932	0.97982	0.98030	0.98077	0.98124	0.98169	2.0
2.1	0.98214	0.98257	0.98300	0.98341	0.98382	0.98422	0.98461	0.98500	0.98537	0.98574	2.1
2.2	0.98610	0.98645	0.98679	0.98713	0.98745	0.98778	0.98809	0.98840	0.98870	0.98899	2.2
2.3	0.98928	0.98956	0.98983	$0.9^{2}0097$	$0.9^{2}0358$	$0.9^{2}0613$	$0.9^{2}0863$	$0.9^{2}1106$	$0.9^{2}1344$	$0.9^{2}1576$	2.3
2.4	$0.9^{2}1802$	$0.9^{2}2024$	$0.9^{2}2240$	$0.9^{2}2451$	$0.9^{2}2656$	$0.9^{2}2857$	$0.9^{2}3053$	$0.9^{2}3244$	$0.9^{2}3431$	$0.9^{2}3613$	2.4
2.5	$0.9^{2}3790$	$0.9^{2}3963$	$0.9^{2}4132$	$0.9^{2}4297$	$0.9^{2}4457$	$0.9^{2}4614$	$0.9^{2}4766$	$0.9^{2}4915$	$0.9^{2}5060$	$0.9^{2}5201$	2.5
2.6	$0.9^{2}5339$	$0.9^{2}5473$	$0.9^{2}5604$	$0.9^{2}5731$	$0.9^{2}5855$	$0.9^{2}5975$	$0.9^{2}6093$	$0.9^{2}6207$	$0.9^{2}6319$	$0.9^{2}6427$	2.6

续表

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	z
2.7	$0.9^{2}6533$	$0.9^{2}6636$	$0.9^{2}6736$	$0.9^{2}6833$	$0.9^{2}6928$	$0.9^{2}7020$	$0.9^{2}7110$	$0.9^{2}7197$	$0.9^{2}7282$	$0.9^{2}7365$	2.7
2.8	$0.9^{2}7445$	$0.9^{2}7523$	$0.9^{2}7599$	$0.9^{2}7673$	$0.9^{2}7744$	$0.9^{2}7814$	$0.9^{2}7882$	$0.9^{2}7948$	$0.9^{2}8012$	$0.9^{2}8074$	2.8
2.9	$0.9^{2}8134$	$0.9^{2}8193$	$0.9^{2}8250$	$0.9^{2}8305$	$0.9^{2}8359$	$0.9^{2}8411$	$0.9^{2}8462$	$0.9^{2}8511$	$0.9^{2}8559$	$0.9^{2}8605$	2.9
3.0	$0.9^{2}8650$	$0.9^{2}8694$	$0.9^{2}8736$	$0.9^{2}8777$	$0.9^{2}8817$	$0.9^{2}8856$	$0.9^{2}8893$	$0.9^{2}8930$	$0.9^{2}8965$	$0.9^{2}8999$	3.0
3.1	$0.9^{3}0324$	$0.9^{3}0646$	$0.9^{3}0957$	$0.9^{3}1260$	$0.9^{3}1553$	$0.9^{3}1836$	$0.9^{3}2112$	$0.9^{3}2378$	$0.9^{3}2636$	$0.9^{3}2886$	3.1
3.2	$0.9^{3}3129$	$0.9^{3}3363$	$0.9^{3}3590$	$0.9^{3}3810$	$0.9^{3}4024$	$0.9^{3}4230$	$0.9^{3}4429$	$0.9^{3}4623$	$0.9^{3}4810$	$0.9^{3}4991$	3.2
3.3	$0.9^{3}5166$	$0.9^{3}5335$	$0.9^{3}5499$	$0.9^{3}5658$	$0.9^{3}5811$	$0.9^{3}5959$	$0.9^{3}6103$	$0.9^{3}6242$	$0.9^{3}6376$	$0.9^{3}6505$	3.3
3.4	$0.9^{3}6631$	$0.9^{3}6752$	$0.9^{3}6869$	$0.9^{3}6982$	$0.9^{3}7091$	$0.9^{3}7197$	$0.9^{3}7299$	$0.9^{3}7398$	$0.9^{3}7493$	$0.9^{3}7585$	3.4
3.5	$0.9^{3}7674$	$0.9^{3}7759$	$0.9^{3}7842$	$0.9^{3}7922$	$0.9^{3}7999$	$0.9^{3}8074$	$0.9^{3}8146$	$0.9^{3}8215$	$0.9^{3}8282$	$0.9^{3}8347$	3.5
3.6	$0.9^{3}8409$	$0.9^{3}8469$	$0.9^{3}8527$	$0.9^{3}8583$	$0.9^{3}8637$	$0.9^{3}8689$	$0.9^{3}8739$	$0.9^{3}8787$	$0.9^{3}8834$	$0.9^{3}8879$	3.6
3.7	$0.9^{3}8922$	$0.9^{3}8964$	$0.9^{4}0039$	$0.9^{4}0426$	$0.9^{4}0799$	$0.9^{4}1158$	$0.9^{4}1504$	$0.9^{4}1838$	$0.9^{4}2159$	$0.9^{4}2468$	3.7
3.8	$0.9^{4}2765$	$0.9^{4}3052$	$0.9^{4}3327$	$0.9^{4}3593$	$0.9^{4}3848$	$0.9^{4}4094$	$0.9^{4}4331$	$0.9^{4}4558$	$0.9^{4}4777$	$0.9^{4}4988$	3.8
3.9	$0.9^{4}5190$	$0.9^{4}5385$	$0.9^{4}5573$	$0.9^{4}5753$	$0.9^{4}5926$	$0.9^{4}6092$	$0.9^{4}6253$	$0.9^{4}6406$	$0.9^{4}6554$	$0.9^{4}6696$	3.9
4.0	$0.9^{4}6833$	$0.9^{4}6964$	$0.9^{4}7090$	$0.9^{4}7211$	$0.9^{4}7327$	$0.9^{4}7439$	$0.9^{4}7546$	$0.9^{4}7649$	$0.9^{4}7748$	$0.9^{4}7843$	4.0
4.1	$0.9^{4}7934$	$0.9^{4}8022$	$0.9^{4}8106$	$0.9^{4}8186$	$0.9^{4}8263$	$0.9^{4}8338$	$0.9^{4}8409$	$0.9^{4}8477$	$0.9^{4}8542$	$0.9^{4}8605$	4.1
4.2	$0.9^{4}8665$	$0.9^{4}8723$	$0.9^{4}8778$	$0.9^{4}8832$	$0.9^{4}8882$	$0.9^{4}8931$	$0.9^{4}8978$	$0.9^{5}0226$	$0.9^{5}0655$	$0.9^{5}1066$	4.2
4.3	$0.9^{5}1460$	$0.9^{5}1837$	$0.9^{5}2199$	$0.9^{5}2545$	$0.9^{5}2876$	$0.9^{5}3193$	$0.9^{5}3497$	$0.9^{5}3788$	$0.9^{5}4066$	$0.9^{5}4332$	4.3
4.4	$0.9^{5}4587$	$0.9^{5}4831$	$0.9^{5}5065$	$0.9^{5}5288$	$0.9^{5}5502$	$0.9^{5}5706$	$0.9^{5}5902$	$0.9^{5}6089$	$0.9^{5}6268$	$0.9^{5}6439$	4.4
4.5	$0.9^{5}6602$	$0.9^{5}6759$	$0.9^{5}6908$	$0.9^{5}7051$	$0.9^{5}7187$	$0.9^{5}7318$	$0.9^{5}7442$	$0.9^{5}7561$	$0.9^{5}7675$	$0.9^{5}7784$	4.5
4.6	$0.9^{5}7888$	$0.9^{5}7987$	$0.9^{5}8081$	$0.9^{5}8172$	$0.9^{5}8258$	$0.9^{5}8340$	$0.9^{5}8419$	$0.9^{5}8494$	$0.9^{5}8566$	$0.9^{5}8634$	4.6
4.7	$0.9^{5}8699$	$0.9^{5}8761$	$0.9^{5}8821$	$0.9^{5}8877$	$0.9^{5}8931$	$0.9^{5}8983$	$0.9^{6}0320$	$0.9^{6}6789$	$0.9^{6}1235$	$0.9^{6}1661$	4.7
4.8	$0.9^{6}2067$	$0.9^{6}2453$	$0.9^{6}2822$	$0.9^{6}3173$	$0.9^{6}3508$	$0.9^{6}3827$	$0.9^{6}4131$	$0.9^{6}4420$	$0.9^{6}4696$	$0.9^{6}4958$	4.8
4.9	$0.9^{6}5208$	$0.9^{6}5446$	$0.9^{6}5673$	$0.9^{6}5889$	$0.9^{6}6094$	$0.9^{6}6289$	$0.9^{6}6475$	$0.9^{6}6652$	$0.9^{6}6821$	$0.9^{6}6981$	4.9

参考文献

1. Ramskumar R. Engincedng Reliability: Fundamentals and Applications Englewood Clifrs, New Jersey: Prentice Hall, 1993
2. Rao SS. Reliability-Design, New York: McGraw-Hill, Inc, 1992
   Approach, New York:
3. 刘京生，陈屹，谢华. 现代设计方法及其应用. 成都：电子科技大学机械电子工程学院讲义，1997
4. (日)田村坦之. 系统工程. 北京：科学出版社，2001
5. 张鄂. 现代设计方法. 西安：西安交通大学出版社，1999
6. 叶元烈. 机械优化理论与设计. 北京：中国计量出版社，2001
7. 李景谓. 有限元法. 北京：北京邮电大学出版社，1999
8. 杜平安. 结构有限元分析建模方法. 北京：机械工业出版社，1998
9. 孙新民等. 现代设计方法实用教程. 北京：人民邮电出版社，1999 年
10. 廖林清等. 现代设计方法. 重庆：重庆大学出版社，2000 年
11. 居滋培. 可靠性工程. 北京：原子能出版社，2000 年
12. 万耀青，阮宝湘. 机电工程现代设计方法. 北京：北京理工大学出版社，1994 年
13. 赵松年等. 现代设计方法. 北京：机械工业出版社，1996 年
14. 黄清远. 优势设计. 北京：机械工业出版社，1999 年
15. 刘惟信. 机械最优化设计(第 2 版). 北京：清华大学出版社，1994
16. 孙靖民主编. 机械结构优化设计. 哈尔滨：哈尔滨工业大学出版社，2003
17. 孙国正主编. 优化设计及应用(机械类专业用). 北京：人民交通出版社，2000
18. 高健编. 机械优化设计基础. 北京：科学出版社，2000
19. 王安麟主编. 机械工程现代最优化设计方法与应用. 上海：上海交通大学出版社，2000
20. 粟塔山，彭维杰，周作益等编著. 最优化计算原理与算法程序设计. 长沙：国防科技大学出版社，2001
21. 孙靖民主编. 机械优化设计. 第 3 版. 北京：机械工业出版社，2003
22. 王凤歧，张连洪，邵宏宇编著. 现代设计方法. 天津：天津大学出版社，2004
23. 陈屹，谢华编著. 现代设计方法及其应用. 北京：国防工业出版社，2004