

计算机视觉导论

原理算法与实践

张林 沈莹 赵生捷 编著

2022 年 12 月 8 日版本（本材料只供个人学习使用，禁止作为商业用途抄袭、出版或发表）

目录

前言.....	1
第 1 章 绪论.....	2
1.1 什么是计算机视觉?	2
1.2 计算机视觉应用举例.....	3
1.3 本书章节安排.....	8
1.4 符号表示规则.....	8
1.5 习题.....	9
参考文献.....	9
第一篇：图像的全景拼接.....	10
第 2 章 图像全景拼接问题概述.....	11
2.1 问题的定义.....	11
2.2 方案流程.....	11
第 3 章 线性几何变换.....	15
3.1 平面上的线性几何变换.....	15
3.2 变换群与几何学.....	21
3.3 三维空间中的线性几何变换.....	24
3.4 习题.....	26
参考文献.....	26
第 4 章 特征点检测与匹配.....	27
4.1 哈里斯角点及其描述子.....	27
4.2 SIFT 特征点及其特征描述子.....	35
4.3 SURF 特征点及其特征描述子.....	51
4.4 ORB 特征点及其特征描述子.....	51
4.3 特征点匹配.....	51
4.4 习题.....	52
参考文献.....	53
第 5 章 线性最小二乘问题.....	55
5.1 齐次线性最小二乘问题.....	55
5.2 非齐次线性最小二乘问题.....	58
5.3 习题.....	62
参考文献.....	63
第 6 章 射影矩阵的鲁棒估计与图像的插值.....	64
6.1 随机抽样一致算法.....	64
6.2 图像的插值.....	67
6.3 习题.....	70
参考文献.....	71
第二篇：单目测量.....	72
第 7 章 单目测量问题概述.....	73
7.1 问题的定义.....	73
7.2 方案流程.....	74
参考文献.....	75
第 9 章 非线性最小二乘问题.....	76

9.1 无约束优化问题基础.....	76
9.2 非线性最小二乘问题及其解法.....	80
9.3 习题.....	85
参考文献.....	85
第 10 章 相机成像模型与内参标定.....	86
10.1 不考虑镜头畸变的成像模型.....	86
10.2 考虑镜头畸变的成像模型.....	89
10.3 相机内参标定.....	91
10.4 镜头畸变去除.....	108
10.5 习题.....	109
参考文献.....	109
第 11 章 鸟瞰视图.....	110
11.1 基本流程.....	110
11.2 鸟瞰视图坐标系到物理平面坐标系的映射	111
11.3 物理平面坐标系到去畸变图像坐标系的映射	112
11.4 去畸变图像坐标系到原始图像坐标系的映射	113
参考文献.....	113
第三篇：目标检测.....	114
第 12 章 目标检测问题概述.....	115
参考文献.....	115
第 13 章 凸优化基础.....	116
13.1 凸优化问题.....	116
13.2 对偶.....	127
13.3 总结.....	138
13.4 习题.....	139
参考文献.....	140
第 14 章 支持向量机与基于支持向量机的目标检测.....	141
14.1 线性分类问题.....	141
14.2 感知器算法.....	143
14.3 线性可分支持向量机.....	146
14.4 软间隔与线性支持向量机.....	154
14.5 非线性支持向量机与核函数.....	161
14.6 针对多类分类问题的支持向量机.....	165
14.7 习题.....	167
参考文献.....	167
第 15 章 深度神经网络及基于深度神经网络的目标检测.....	168
参考文献.....	168
第四篇：三维立体视觉.....	169
第 16 章 三维重建问题概述.....	170
参考文献.....	170
第 17 章 运动恢复结构.....	171
参考文献.....	171
第 18 章 神经辐射场.....	172
18.1 基于辐射场的体渲染.....	172

18.2 辐射场的隐式表达及其学习	175
参考文献.....	177
附录.....	178
A. 圆锥曲线 ^[1]	178
B. 数字图像导数的近似计算	178
C. 高斯函数的卷积及其傅里叶变换	180
D. 主曲率与海森矩阵	181
E. 拉格朗日乘子法 ^[3]	183
F. 函数或自变量形式为矩阵或向量时的求导运算.....	183
G. 奇异值分解.....	191
H. 函数的极值点、驻点和鞍点	197
I. 罗德里格斯公式	200
参考文献.....	201

前言

计算机视觉是一门研究如何构建具有“视觉”功能的计算机系统的学科，是人工智能研究领域的一个重要分支。从刷脸支付到太空探索，从智能监控到视觉导航，计算机视觉技术正在越来越多的应用领域中影响和改变着我们的生产和生活方式。

近来，随着我国对人工智能领域人才培养支持力度的持续加大，越来越多的高校在本科阶段开设了计算机视觉课程。然而，由于计算机视觉是一门综合性学科，其本身的知识体系和其所涉及到的背景知识都较为庞杂，因此，想编著一本好的读物，能够把该领域内的 important 知识以一种逻辑性强的方式有机组织起来，并不是一件很容易的事。也正是由于这个原因，目前适合于本科层次教学需求的优秀计算机视觉教材还很稀缺。

基于作者十多年的教授计算机视觉课程的经验，在内容组织上，本书遵循了一个新的思路——以具体应用为载体。按照这个原则，作者将全书内容按照“图像的全景拼接”、“单目测量”、“双目立体视觉”和“目标检测与识别”四条主线来组织。对于每一条主线来说，最终目标都是要解决一个明确的具体的问题。我们围绕如何解决这个具体问题，把相关的重要知识点循序渐进地、有机地组织在一起。作者多年教学经验表明，这种形式的内容组织方式很容易为学生所接受，使得初学者更容易从宏观上掌握学科脉络并深刻理解每一个知识点的内涵和作用。

理论与技术并重是本书的一个显著特点。对每一个具体的模型或者算法，本书都尽可能详细地阐述清楚它的来龙去脉，给出必要的数学预备知识以及推导，帮助读者构建起知识的“逻辑大厦”。另一方面，从很大程度上来说，计算机视觉是一门应用科学，读者必须通过编程实践（以及必要的实际动手操作）才能更深刻地理解技术本质。因此，配合理论教学内容，本书提供了丰富了示例程序。这些示例程序可有效帮助读者消化理解相关模型或算法。

为方便使用本教材的老师和同学，我们制作了与教材配套的网站，提供了完整的教学课件和示例程序，网址为 <https://github.com/csLinZhang/CVBook>。

本书可作为人工智能、计算机和软件工程等专业高年级本科生或研究生计算机视觉课程的教材，也可供相关领域的工程技术人员参考。本书内容力求做到“自封闭”，读者只需具有高等数学、线性代数、概率论、解析几何和数字图像处理方面的基本知识即可。

从 2011 年秋季开始，作者在同济大学讲授计算机视觉课程。本书是作者在总结十余年教学实践经验的基础上形成的。计算机视觉学科仍处于蓬勃发展阶段，新理论、新算法、新技术层出不穷，加之作者水平有限，书中难免存在缺陷和不足，殷切希望广大读者批评指正。

作者

2022 年 11 月

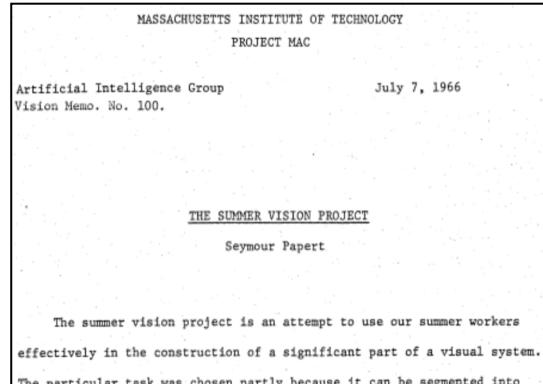
第1章 绪论

1.1 什么是计算机视觉？

计算机视觉（Computer Vision），顾名思义，它是一门研究如何构造具有“视觉”功能的计算机系统的学科。1966年，麻省理工学院的人工智能学家马文·明斯基（Marvin Lee Minsky）^[1]在给学生布置的作业中，要求学生通过编写一个程序让计算机告诉人们它通过摄像头看到了什么，这被认为是计算机视觉最早的任务描述（图1-1）。



(a)



(b)

图1-1：(a) 马文·明斯基（Marvin Minsky，1927年8月9日-2016年1月24日），美国麻省理工学院的人工智能学家；1956年，他和约翰·麦卡锡（John McCarthy）一起发起“达特茅斯会议”并提出人工智能概念；1969年，马文·明斯基获图灵奖。(b) 1966年7月，马文·明斯基在给学生布置的暑期项目中，要求学生通过编写一个程序让计算机告诉人们它通过摄像头看到了什么，这被认为是计算机视觉最早的任务描述。

更进一步，我们要问：具有“视觉”功能的计算机系统到底应该具有哪些具体功能呢？我们不妨做个类比，想一想人类的视觉系统会具有哪些功能。首先，利用自身的视觉系统，我们能够认识家人、朋友，能够识别出苹果、香蕉等目标并且能够知道它们的“边界”在哪里，能够基本准确地判别出所见场景类别（比如，是春天场景还是冬天场景，是雾霾天还是晴天等等）。也就是说，我们的视觉系统具有对场景的理解（understanding）或识别（recognition）能力。另外，想象一下，如果你闭上眼睛一直往前走的话，会发生什么？不是撞到墙上，就是会掉到河里！当然，由于我们有可靠的视觉系统，上述糟糕的场面在正常情况下并不会发生。这是因为，利用自身的视觉系统，我们可以大致测量出目标物体与自己的距离，能够感知到周围三维空间环境及我们自身在该空间中所处的相对位置。这样总结下来，人类的视觉系统应该具备两种基本能力，对场景的理解识别能力和对空间的测量能力。类似的，**计算机视觉领域的研究要解决的基本问题也是对视觉信息的理解识别以及对空间的感知测量**，只不过此时的“数据采集装置”从人眼换成了各类传感器，“数据处理装置”从人脑换成了计

算机。

从上面的介绍可知，在大多数情况下，一个计算机视觉系统包括了视觉传感器和计算平台两个部分。计算平台是运行有为解决某一具体计算机视觉问题而设计的计算程序的计算机，而视觉传感器负责采集感知数据并传送给计算平台。

本书后面大部分内容主要会围绕如何设计解决具体计算机视觉问题的算法或模型展开。这里我们先简要了解一下常用的视觉传感器以及它们所采集的视觉信息的表现形式。计算机视觉系统常用的视觉信息采集装置主要包括各类相机、深度相机、3D 扫描仪等。通过使用具有不同光谱响应特性的相机，我们可以采集到不同电磁波段下的图像（或图像序列），比如 X 光图像、可见光图像、近红外图像、遥感图像等等。使用深度相机，我们可以获得场景的深度图，深度图中每一个像素的值是场景中的对应点到相机成像平面的距离。利用三维扫描仪，我们可以得到被扫描目标物体的三维表示，一般为点云结构或网格结构。图 1-2 展示了计算机视觉系统中常用的视觉传感器。图 1-3 展示了视觉信息的典型表现形式。



图 1-2：计算机视觉系统中常用的视觉信息采集设备：(a) 手机相机；(b) 监控相机；(c) 广角鱼眼相机；(d) 深度相机；(e) 三维扫描仪。

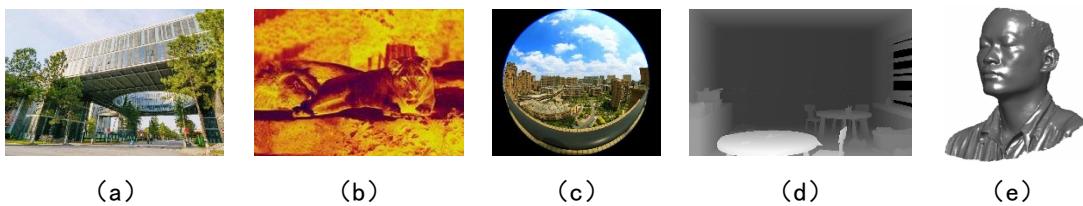


图 1-3：视觉数据的常见表现形式：(a) 可见光图像；(b) 热红外图像；(c) 广角鱼眼图像；(d) 深度图像；(e) 三维模型。

1.2 计算机视觉应用举例

为了使读者对计算机视觉领域的研究范畴有一个更加感性和直观的了解，我们举一些该领域中典型的应用实例。

1.2.1 人脸识别

人脸识别可以说是计算机视觉领域中研究工作开展的最早、应用的最为广泛和最为成熟的技术分支。它是基于人的脸部特征信息进行身份识别的一种生物识别技术。人脸识别系统

用摄像机或摄像头采集含有人脸的图像或视频流，并自动在图像中检测和跟踪人脸，进而对检测到的人脸进行面部识别。

人脸识别系统的研究始于二十世纪六十年代。八十年代后，随着计算机技术和光学成像技术的发展得到提高，而真正进入初级的应用阶段则在九十年代后期。人脸识别系统成功与否的关键在于是否拥有尖端的核心算法，并使识别结果具有实用化的识别率和识别速度。

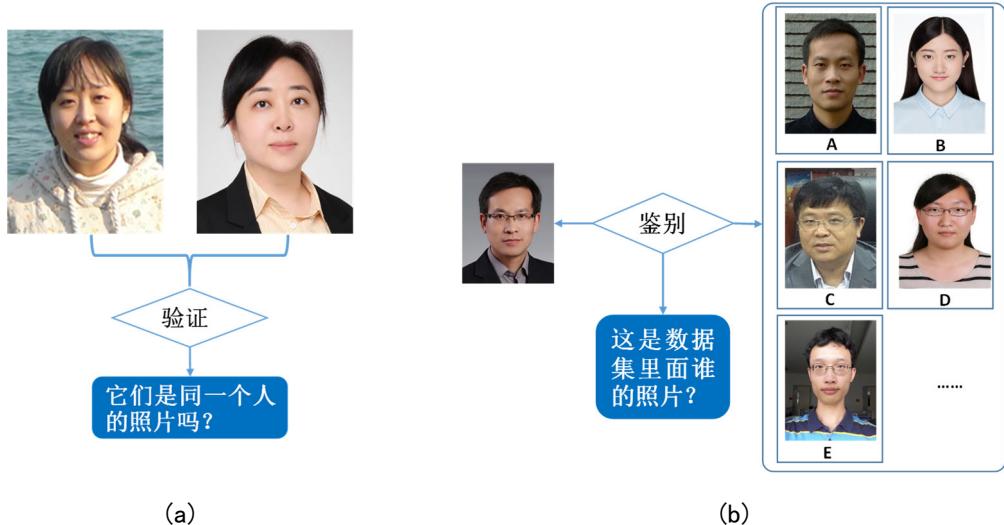


图 1-4：人脸识别领域的两类问题：(a) “一对一”的确认问题；(b) “一对多”的鉴别问题。

人脸识别系统一般包括四个组成部分，分别为：人脸图像采集及检测、人脸图像预处理、人脸图像特征提取以及匹配与识别。根据具体业务逻辑的不同，人脸识别问题可以具体细分为两类问题，“一对一”的确认（Verification）问题和“一对多”的鉴别（Identification）问题。“确认”要解决的问题是：给系统输入两张人脸照片 A 和 B，系统需要“确认” A 和 B 这两张照片是否是采集自同一个人的（图 1-4 (a)）；“鉴别”要解决的问题是：系统连接着一个后台人脸注册数据库 Ω ，对于当前输入的一张人脸照片 F，系统需要回答 F 是 Ω 中哪一个人的照片（图 1-4 (b)）。不难理解，旅客在海关通关时使用的“自助通关”系统、火车站进站口的身份核验系统等，都属于“一对一”的人脸确认系统；公司内部使用的基于人脸的考勤系统、刑侦使用的人脸抓拍照片与重点人员的人脸数据库比对系统，则属于“一对多”的人脸识别系统。

人脸识别产品已广泛应用于金融、司法、军队、公安、边检、政府、航天、电力、工厂、教育、医疗及众多企事业单位等领域。随着技术的进一步成熟和社会认同度的提高，人脸识别技术将应用在更多的领域。

1.2.2 智能监控

视频监控是安保防范的重要手段之一，然而依赖人工手段来检索分析监控视频难以高效地处理海量的视频数据。随着计算机视觉技术的发展，智能监控为视频监控提供了新的解决方案。智能监控技术采用计算机视觉方法对监控数据进行自动化分析，可实现智能化的安保

监控与环境监测。智能监控系统能够更加高效地获取监控数据中的有效信息，实现全自动、全天候的安全监测与智能管理。

为了实现对视频监控数据的智能化分析，智能监控系统一般会涉及多项计算机视觉任务，如图像分割、物体识别、物体追踪以及行为分析等。一般来说，在智能监控系统中，首先需要进行底层的图像处理工作，以得到基础的图像信息，如通过图像分割（Image Segmentation）分离出视频画面中的前景和后景，以提取出监控画面中的重点关注区域。同时，通过物体检测与识别技术对目标物体进行更精确的定位及区分，常见的识别目标包括行人、车辆、车牌、信号灯等等。在定位到目标物体后，智能监控系统需要对物体进行动态目标追踪（Object Tracking），以获取目标的行为特征。比如，可对移动的行人及车辆进行追踪，记录其运动轨迹及速度等信息。图 1-5 展示了对行人目标的检测与追踪。随后，系统即可根据追踪得到的行为特征对目标物体的运动模式进行进一步分析和理解，以判断监控画面中是否存在需要关注的异常事件或行为，如人群聚集、打架斗殴、交通事故等，最终实现对场景的智能化实时监测。

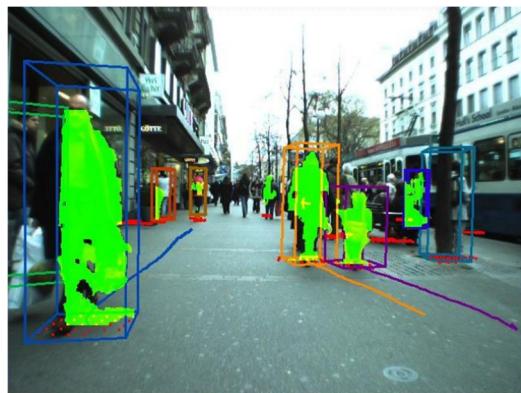


图 1-5：行人的检测与追踪。

智能监控技术能够被应用于交通、农业、军事等领域场景下的安全防范、信息获取与指挥调度等。智能监控既能够为大型公共场所提供安保措施，实现罪犯识别、实时监测和危险事件预警等功能；也为能够服务家庭监控等任务，搭建数字化的家庭监控系统，随时监测居家安全情况和家人的健康状况。随着计算系统软硬件技术的持续发展，智能监控将会在更多的领域内得到应用。

1.2.3 医学图像分割

1.2.4 双目立体视觉

1.2.5 视觉定位

作为学术界和工业界同时发展最活跃的领域之一，智能移动机器人受到了世界各国的普遍重视。随着其性能的不断完善，移动机器人已经成功应用于医疗服务、城市安全、国防和

空间探测等领域。智能移动机器人系统一般包含环境感知、自身定位、决策规划和行为控制等多个功能模块。其中，自身定位作为其中最基本的模块之一，是机器人开展其它工作的基础和前提。

为解决机器人的定位问题，通常会涉及各类传感器的使用。譬如，室外场景中常使用的 GNSS (Global Navigation Satellite System, 全球导航卫星系统)。由于此类卫星定位系统无法在室内场景中使用，因此开发适用于室内场景的定位技术成为了近年来的研究热点。目前，研究较为广泛的室内定位技术主要有：基于无线电信号的定位技术、基于地磁场检测的定位技术及基于 IMU (Inertial Measurement Unit, 惯性导航单元) 的定位技术等。但是，以上室内定位技术在实际使用中仍然存在一些缺陷。基于无线电信号的定位技术依赖于室内场景中布设的信号发射装置，而安装这些装置增加了定位的成本。此外，无线电信号容易受到移动物体的干扰，造成定位精度的下降。基于地磁检测的定位技术容易受到电气设备和金属物体的干扰。而基于 IMU 设备的定位技术容易产生累积误差，不适合用于长距离定位。

由于人类确定自身位置时主要借助视觉，因此通过模仿人类视觉感知功能而设计的视觉定位系统体现出了独特的技术优势。视觉定位系统通过安装在移动机器人上的相机拍摄环境图片，再借助图像中的像素或低/高层次特征的位置变化估计机器人的位姿参数。由于成本低、信息丰富，相机已经成为众多智能移动机器人的标配。同时，因相机在定位方式上的独特优势，利用视觉进行定位已经在诸多领域得到应用。如图 1-6 所示，典型的应用包括美国国家航空航天局的“机遇号”火星探测项目以及大疆科技的 DJI Mavic 3 系列无人机等。

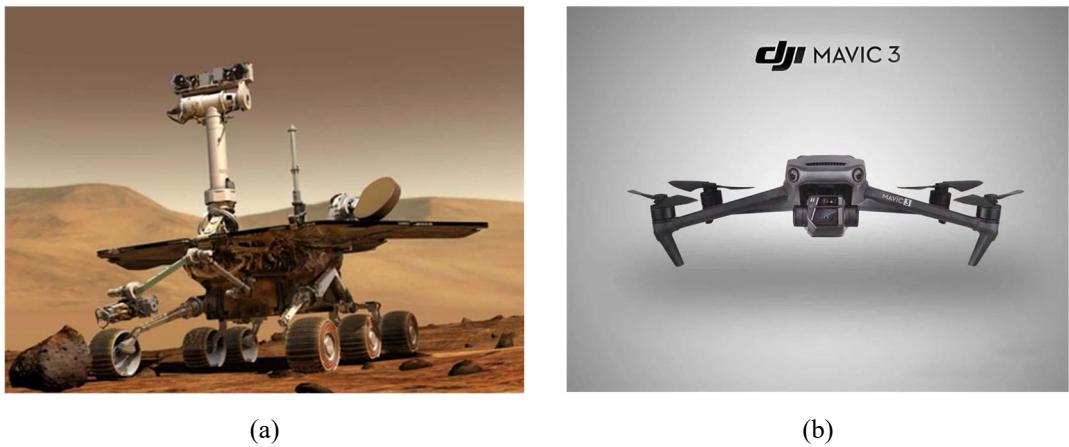


图 1-6：基于视觉定位技术的典型应用：(a) 2003 年美国宇航局发射的“机遇号”火星探测器；
(b) 大疆科技的 DJI Mavic 3 系列无人机。

根据是否使用先验地图，视觉定位技术分为基于先验地图的视觉定位技术与无先验地图的视觉定位技术。前者基于先验地图并借助重定位、图像检索等技术进行视觉定位，能较好地保证定位结果的全局一致性。而由于后者无法依赖先验地图，所以此类技术需要同时估计机器人位姿与周围环境结构，定位结果缺乏全局一致性。根据是否实时运行，无先验地图的视觉定位技术通常为 SLAM (Simultaneous Localization And Mapping, 同时定位与建图) 技术与 SfM (Structure From Motion, 运动恢复结构) 技术。SLAM 技术起源于机器人社区，

更加注重定位和实时性。按照实现方案中相机的个数，SLAM 技术可以分为单目视觉 SLAM 技术、双目视觉 SLAM 技术和多目视觉 SLAM 技术。通常在实际应用中为提高机器人自身定位的精度，视觉 SLAM 技术通常会融合 IMU 和激光雷达等传感器。而 SFM 技术起源于计算机视觉社区，更加注重场景结构恢复的精准度，而非实时性。关于视觉定位技术的分类可简单总结如图 1-7 所示。



图 1-7：视觉定位技术的分类。

如今，视觉定位技术正逐渐从实验室走进人们的生活中，并在教育、医疗和安全等领域扮演着重要角色。对于视觉定位系统，定位精度和实时性是保证用户体验的关键。因此，提高定位精度以及保证系统实时性是视觉定位技术未来发展的主要目标。

1.2.6 三维场景重建

重建现实场景物体的三维模型、在数字空间下呈现真实空间中物体的形状及色彩信息，在许多领域内有着重要的应用价值。在智慧城市应用中，我们使用城市数字模型来展示并管理城市；在文物保护领域，我们可将文物复原为三维模型，构建数字博物馆；在虚拟现实场景中，我们通过数字三维模型来提供沉浸式的用户交互体验；除此之外，在工业制造、游戏开发、机器人导航等任务中，三维重建都有着广泛的应用空间。

三维重建可以通过传统的几何建模方法实现，但是几何建模方法依赖复杂的人工建模过程，操作难度大且难以保证精确度。随着计算机视觉技术及传感器技术的发展，基于图像及点云的三维重建方法成为了三维重建技术领域的主流。三维重建系统使用各种类型的视觉传感器对场景进行扫描，基于多视图几何原理，实现对现实物体的逆向建模，自动化地在数字空间中复原真实物体的结构及纹理特征。

三维重建系统的输入可以是普通的二维图像序列、深度相机采集的深度图以及激光雷达采集得到的点云数据等。对于二维图像序列输入，三维重建系统通过 SFM (Structure from Motion) 技术生成稀疏的点云模型，再通过 MVS (Multi View Stereo) 技术进一步获取稠密的点云模型。而对于深度图像序列及激光雷达数据，系统则可以使用 ICP (Iterative Closest Point, 迭代最近点) 等算法获取传感器的位姿，融合得到相应的点云模型。在得到点云模型后，TSDF (Truncated Signed Distance Function, 距离截断函数) 以及泊松重建 (Poisson Reconstruction) 等方法能够从点云模型中构建三角网格，得到表示物体形状特征的网格模型。随后，可再为网格模型添加纹理贴图，为模型添加真实的色彩信息，即可得到最终的数字三维模型。



图 1-8：使用 iPad 进行三维重建。

目前，三维重建技术正向着实时化、轻量化的方向发展。三维重建系统已经能够做到在设备扫描的同时，实时输出高精度的模型。基于移动设备也已经能够实现三维重建过程，如图 1-8 所示。三维重建技术已经开始逐步普及至日常生活中，普通用户也能方便快捷地生成各种三维模型。伴随着三维重建技术的普及与发展，更多的三维应用场景与需求将会在未来涌现。

1.3 本书章节安排

为了能够让读者在具体情境中深刻理解计算机视觉中各个知识点的作用以及它们之间的逻辑关系，本书以“图像的全景拼接”、“单目测量”、“双目立体视觉”和“目标检测与识别”四个具体问题为主线，分成四篇来组织本书的内容。我们希望能以具体问题为载体，来带领读者系统学习该领域中的基本概念、模型、算法和相关的应用数学知识。

第 1 篇是图像的全景拼接。这部分内容覆盖了本书第 2 至 6 章。在第 2 章中，我们对要解决的图像全景拼接问题给出清晰定义，之后在第 3 至 5 章中依次介绍所需技术。第 3 章介绍线性几何变换，第 4 章介绍特征点检测与匹配，第 5 章介绍线性最小二乘问题及其解法，第 6 章会介绍基于随机采样一致性框架的模型拟合以及图像的双线性插值。

第 2 篇是单目测量。这部分内容覆盖了第 7~11 章，在第 7 章中，我们定义清楚要解决的单目测量问题。第 8 章会介绍射影几何的基本知识。第 9 章会介绍非线性最小二乘问题及其解法。第 10 章介绍针孔相机模型与相机参数标定方案。第 11 章介绍鸟瞰视图的生成方法。

1.4 符号表示规则

除非上下文特殊说明，在数学符号的使用上，本书基本遵从表 1-1 所列规则：

表 1-1 本书符号表示规则

符号说明	范例
小写斜体非粗体，表示标量	f
小写正体粗体，表示列向量	\mathbf{x}

大写斜体非粗体，表示矩阵	A
大写斜体花体，表示集合	\mathcal{R}
大写空心字体，表示几个常见的集合	\mathbb{R} 表示实数集合, \mathbb{Z} 表示整数集合
$(\cdot)^T$ 表示矩阵（也包括向量）的转置	A^T

1.5 习题

- (1) 除了本章所列举的实例之外，请列举几个你在日常生活中遇到的计算机视觉技术的应用场景。
- (2) 安装好 **Matlab** 环境，尝试读取一张磁盘上的图像并显示出来。
- (3) **OpenCV** 是一个跨平台计算机视觉和机器学习软件库，可以运行在 **Linux**、**Windows**、**Android** 和 **Mac OS** 操作系统上。我们后面章节中的部分示例代码需要有 **OpenCV** 库的支持。请在你的实验计算机上安装配置好 **OpenCV** 环境，尝试编写 **C++** 代码，调用 **OpenCV** 库函数，完成读取并显示一张图片的任务。

参考文献

- [1] Marvin Minsky, https://en.wikipedia.org/wiki/Marvin_Minsky

第一篇：图像的全景拼接

第 2 章 图像全景拼接问题概述

2.1 问题的定义

在采集图像时，由于单个相机的视场范围有限，每张图像只能反映有限视场内的信息。如果想获取到更大范围场景的图像信息，可以用全景拼接技术把一组反映同一场景不同局部、相互之间存在一定共视区域的图像拼接合成一张具有更大视场范围的图像。实际上，图像的全景拼接模型和拼接技术有很多种，为了形成完整鲁棒的全景拼接系统，也有很多细节问题需要考虑，但本篇的目的是使读者以这个任务为载体学习和掌握一些重要的计算机视觉知识点，因此我们把该问题简化，把任务限定在有限范围内，不去关注过多的细枝末节。

本篇所要解决的图像全景拼接问题描述如下：

有两张图像 I_1 和 I_2 ，它们所对应的物理场景共面，它们存在共视区域，拍摄它们的相机不存在镜头畸变（这意味着相机的成像平面和物理平面之间对应点的映射关系可以用同一个线性几何变换来刻画）， I_1 和 I_2 之间不存在较大的光照条件变化（这意味着不需要额外考虑拼接图像中可能存在的光照不一致性问题），我们的目标是要把 I_1 和 I_2 根据它们共视区域内图像内容的一致性拼接在一起。如果 I_1 和 I_2 满足上述条件的话，它们对应点（同一物理点在 I_1 和 I_2 上分别所成的像）的像素坐标可以通过同一个线性几何变换 H 联系起来，即 $\forall \mathbf{x} \in I_1$ ，

如果 $\mathbf{x}' \in I_2$ ，且 \mathbf{x} 与 \mathbf{x}' 是对应点的像素坐标，则，

$$\mathbf{x}' = H\mathbf{x} \quad (2-1)$$

2.2 方案流程

2.1 节中所定义的图像全景拼接问题应该如何解决呢？我们通过一个构造的示例来说明解决这个问题的基本思路。假设图 2-1 (a) 是要拍摄的整个物理平面，它包含了 4 个目标，六角星、正方形、三角形和梯形。图 2-1 (b) 是相机拍摄的第 1 张照片 I_1 ，由于视场有限，最右边的景物“梯形”不在 I_1 之上；图 2-1 (c) 是相机拍摄的第 2 张照片 I_2 ，类似地，由于视场所限，最左边的景物“六角星”不在 I_2 之上。不难想象，如果我们把 I_1 和 I_2 拼接在一起，便可以得到物理场景的完整图像。

根据 2.1 节对图像全景拼接问题的界定，可以知道，图像 I_2 上的点 \mathbf{x}' 与图像 I_1 上对应点 \mathbf{x} 之间可以通过线性几何变换 H 联系起来， $\mathbf{x}' = H\mathbf{x}$ 。如果能有办法找到这个 H ，继而对 I_1 中所有像素点施加变换 H ，就会把 I_1 上的所有像素点变换到与 I_2 对应像素点重合的位置上，也就完成了全景拼接任务中的主要部分。想象一下，如果要以手工的方式来完成这件事，我们大概会如何做呢？我们会观察 I_1 和 I_2 ，找到它们各自之上一些非常具有区分性的“特征点”；然后，对 I_1 进行一定程度的缩放、旋转、平移，使 I_1 中的特征点与 I_2 中对应的特征点

都能重合上，即 \mathbf{x}_1 点重合到点 \mathbf{x}'_1 ， \mathbf{x}_2 点重合到 \mathbf{x}'_2 点，...， \mathbf{x}_7 点重合到 \mathbf{x}'_7 点；经历了这些过程之后， I_1 中的“六角星”也已经被变换到了正确的位置，即完成了全景拼接任务。

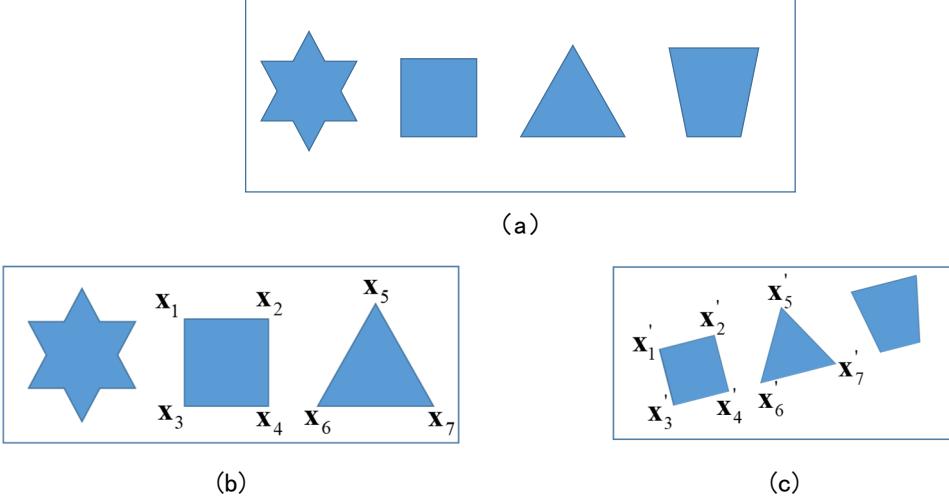


图 2-1：图像全景拼接问题示意图。（a）整体场景；（b）图像 I_1 ；（c）图像 I_2 。我们的任务是把 I_1 和 I_2 拼接在一起，来得到整体场景的图像信息。

把上述手工过程的每一步以计算机算法的形式实现出来，就会得到图像全景拼接的算法流程。给定两张图像 I_1 和 I_2 ，假设它们满足 2.1 节所述的图像全景拼接问题定义中的限定条件，则可按照下述步骤完成 I_1 和 I_2 的拼接任务：

1) 特征点检测

利用特征点检测算法在 I_1 和 I_2 中检测出具有较高区分性的特征点，检测到的图像特征点最终的表达形式为图像中的二维位置坐标。

2) 创建特征点描述子

当从 I_1 和 I_2 中检测出特征点以后，为了估计 I_1 和 I_2 之间的几何变换 H ，我们必须要知道 I_1 和 I_2 中特征点的对应关系。显然，如果仅有特征点的位置信息，我们是无法准确获得这个对应关系的。为了能进行特征点匹配从而得到 I_1 和 I_2 特征点间的对应关系，需要为每一个特征点构造它的描述子。一个特征点 \mathbf{x} 的描述子 \mathbf{d} 是一个基于 \mathbf{x} 的局部图像信息所构造出来的向量。从理论上来说，我们希望所构造出来的描述子能具有如下特性：如果 I_1 中的特征点 \mathbf{x} 和 I_2 中的特征点 \mathbf{x}' 是对应的特征点（即，它们是物理场景中同一个点的像），那么

\mathbf{x} 的描述子（基于 I_1 中 \mathbf{x} 周围的局部图像信息构造）和 \mathbf{x}' 的描述子（基于 I_2 中 \mathbf{x}' 周围的局部图像信息构造）应该是相同的；反之，则不同。

3) 特征点匹配

当在 I_1 和 I_2 中分别检测了特征点并为每个特征点构建了描述子之后，接下来需要设计基于描述子信息的特征点匹配算法，以得出 I_1 与 I_2 中特征点对应点对集合 $\mathcal{S} = \{\mathbf{x}_i \leftrightarrow \mathbf{x}'_i\}_{i=1}^p$ ，

其中 \mathbf{x}_i 是来自 I_1 的特征点， \mathbf{x}'_i 是来自 I_2 的特征点， $\mathbf{x}_i \leftrightarrow \mathbf{x}'_i$ 表示 \mathbf{x}_i 与 \mathbf{x}'_i 是一对对应的特征点，

p 为 I_1 和 I_2 中具有对应关系的特征点对的个数。

4) 几何变换估计

经过特征点匹配以后，得到了特征点对应点对集合 $\mathcal{S} = \{\mathbf{x}_i \leftrightarrow \mathbf{x}'_i\}_{i=1}^p$ 。根据全景拼接问题的假定，我们知道 $\mathbf{x}'_i = H\mathbf{x}_i$ ， H 是一个表达线性变换的矩阵。这样，我们便可从特征点对应点对集合 \mathcal{S} 中得到一个关于 H 的线性方程组，

$$\begin{cases} \mathbf{x}'_1 = H\mathbf{x}_1 \\ \mathbf{x}'_2 = H\mathbf{x}_2 \\ \vdots \\ \mathbf{x}'_p = H\mathbf{x}_p \end{cases} \quad (2-2)$$

通过解方程组 (2-2)，便可以得到几何变换 H 。

在这个过程中，有一个细节问题需要考虑一下。由于特征点检测、描述子构建、特征点匹配等算法潜在的局限性，步骤“3) 特征点匹配”中得到的特征点对应点对集合中可能会存在个别对应点对关系是错误的情况。我们应该如何处理这个问题呢？换言之，我们是否有办法能在 \mathcal{S} 中存在部分错误点对关系的情况下依然能够鲁棒地估计出 H ？为了应对这个问题，可以使用随机采样一致性算法，这是一个能从存在外点（错误观测）的观测数据集合中鲁棒地拟合出模型的算法框架，它可以有效地对抗外点所带来的干扰。

5) 坐标变换

当得到了 H 以后，我们便可以把 I_1 中的每个像素点 \mathbf{x}_i 变换到新的位置 $H\mathbf{x}_i$ ，以对齐 I_1 和 I_2 。再经过一些后处理操作，便完成了 I_1 和 I_2 的全景拼接。在这个过程中，也有一个细节问题需要我们考虑，就是如何具体实现对 I_1 施加几何变换 H 的操作。如果按照“正向”思路， I_1 中的点 \mathbf{x}_i 变换之后的位置应该是 $H\mathbf{x}_i$ ，因此只需要把 $H\mathbf{x}_i$ 位置的像素值设置成像素值 $I_1(\mathbf{x}_i)$ 不就可以了？但需要注意，数字图像的像素坐标都是用整数表示的，也就是说 \mathbf{x}_i 是个整数，那么目的坐标 $H\mathbf{x}_i$ 几乎肯定是个浮点数，那么 $H\mathbf{x}_i$ 这个位置在图像上就没办法唯一确定了。因此，对 I_1 施加几何变换 H ，在具体实现上需要使用图像的插值技术。

接下来的第 3 至 6 章将详细阐述图像全景拼接算法的全部细节。

我们在前面提到，在全景拼接问题中，如果 \mathbf{x} 和 \mathbf{x}' 是 I_1 和 I_2 中对应的特征点，那么它们可以被线性几何变换 H 联系起来， $\mathbf{x}'_i = H\mathbf{x}_i$ 。但我们尚未说明 \mathbf{x} 和 \mathbf{x}'_i 的表达是什么样子的、 H 这个矩阵是什么样子的、 H 具有哪些属性等等。我们会在第 3 章“线性几何变换”中把这些问题交代清楚。

在第 4 章中，我们将介绍目前计算机视觉领域中应用最为广泛的几种特征点检测算法、

描述子构建算法以及描述子匹配算法。

解方程组（2-2）的问题实际上是一个线性最小二乘问题。我们将在第 5 章中详细阐述线性最小二乘问题的解法。

“如何能在 \mathcal{S} 中存在部分错误点对关系的情况下依然能够鲁棒地估计出 H ？”这个问题将在第 6 章中解决。那时，我们将学习能从观测数据中鲁棒拟合出模型的随机采样一致性算法框架。在第 6 章结束时，针对图像的线性几何变换这个问题，我们也会介绍一下双线性差值算法。

第3章 线性几何变换

在第2章中提到，在我们所定义的全景拼接问题中，图像 I_1 和 I_2 能够拼在一起的前提是它们的对应像素点的坐标关系可以通过统一的线性几何变换 H 来表达，其中 H 是表达坐标变换的矩阵。首先来说一下什么是线性几何变换。在 n 维向量空间 \mathbb{R}^n 中，对其中的元素进行的几何变换 T 为线性几何变换的充要条件是存在可逆矩阵 H 使得，

$$\forall \mathbf{x} \in \mathbb{R}^n, T(\mathbf{x}) = H\mathbf{x} \quad (3-1)$$

因此，能够表达线性几何变换的矩阵 H 必须是一个可逆矩阵。

由于本篇的主题是图像的全景拼接，因此我们会主要讨论二维平面上的线性几何变换，并还会以群论的视角来重新看待线性几何变换。之后，会把在二维情况下线性几何变换的有关结论推广到三维情况。在三维空间中的几何变换相关结论会在后续的单目测量、双目立体视觉等章节中用到。

3.1 平面上的线性几何变换

3.1.1 旋转变换（Rotation transformation）

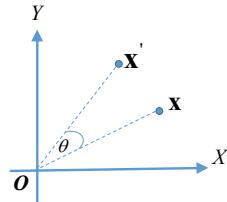


图 3-1：平面内一点 \mathbf{x} 绕坐标原点旋转到 \mathbf{x}' 关系示意图。

假设平面上有一点 $\mathbf{x}=(x,y)^T$ ，该点绕原点逆时针方向旋转 θ 角后得到点 $\mathbf{x}'=(x',y')^T$ ，则

\mathbf{x} 与 \mathbf{x}' 之间的关系（如图 3-1 所示）可表达为，

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \quad (3-2)$$

记矩阵 $R_{2 \times 2} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$ ，则显然 $R_{2 \times 2}$ 可以刻画平面内两点之间的旋转关系。这个矩阵 $R_{2 \times 2}$

的特点是：它是一个正交矩阵且它的行列式为 1。实际上，这个结论反过来也成立：如果一个矩阵 $R_{2 \times 2}$ 是行列式为 1 的正交矩阵，它可以用来表达一个平面内的**保持方向（Orientation Preserving）的旋转**。

我们强调表达保持方向旋转的正交矩阵 $R_{2 \times 2}$ 的行列式要为 1。根据线性代数^[1]的知识可知，正交矩阵的行列式要么是 1，要么是 -1。那么，行列式为 -1 的二维正交矩阵表达的几何变换是什么呢？这类正交矩阵表达的几何变换是平面内的旋转再复合一个反射变换。我们通过一个示例来理解一下。假设图 3-2 (a) 是变换之前的原始图像。图 3-2 (b) 是图像 3-2 (a)

经过了由矩阵 $\begin{bmatrix} \cos \frac{\pi}{6} & -\sin \frac{\pi}{6} \\ \sin \frac{\pi}{6} & \cos \frac{\pi}{6} \end{bmatrix}$ (其行列式为 1) 定义的绕图像中心的几何变换得到的结果；

图 3-2 (c) 是图像 3-2 (a) 经过了由矩阵 $\begin{bmatrix} -\cos \frac{\pi}{6} & -\sin \frac{\pi}{6} \\ -\sin \frac{\pi}{6} & \cos \frac{\pi}{6} \end{bmatrix}$ (其行列式为 -1) 定义的几何变

换得到的结果。在图像 3-2 (a) 中，花坛前的石头在“同济大学”这个矢量的顺时针一侧。在图像 3-2 (b) 中，该石头依然是在“同济大学”这个矢量的顺时针一侧。但在图像 3-2 (c) 中，该石头在“同济大学”这个矢量的逆时针一侧。因此，我们说图像图 3-2 (a) 到图像 3-2 (b) 的变换是保持方向的，而它到图像 3-2(c) 的变换并没有保持方向。通过这个例子我们看到，行列式为 1 的正交矩阵可以表达平面内的旋转，而行列式为 -1 的正交矩阵则不可以。在本书中，我们所定义的旋转变换为保向旋转变换，要求表达旋转变换的矩阵为行列式为 1 的正交矩阵。因此，旋转变换也可称为特殊正交变换（在机器人学中这个称呼用的比较多），其特殊性就在于表达旋转的正交矩阵的行列式必须为 1。

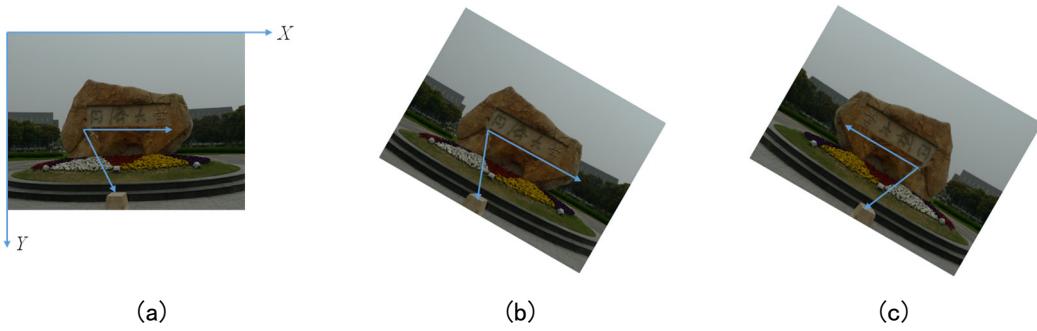


图 3-2：平面内的旋转变换与旋转和反射复合变换。(a) 原始图像；(b) 图像 (a) 经过一个平面内的旋转变换之后得到的结果；(c) 图像 (a) 经过旋转复合反射之后得到的结果。

为了便于表达形式的统一和扩展，我们使用齐次坐标的方式来表达点的位置。对于二维平面上的点，其齐次坐标的表示为一个三维向量 $(x_1, x_2, x_3)^T$ 。如果 $x_3 = 0$ ，则说明这个点为一个无穷远点；如果 $x_3 \neq 0$ ，则说明该点为一个正常点。点的齐次坐标从形式上来说不具有唯一性：如果一个点的齐次坐标为 $(x_1, x_2, x_3)^T$ ，则 $k(x_1, x_2, x_3)^T$ (k 为任意实数且 $k \neq 0$) 也是该

点的齐次坐标。对于一个正常点，给定它的齐次坐标 $(x_1, x_2, x_3)^T$ ，可以得出它的规范化齐次坐标 $(x_1/x_3, x_2/x_3, 1)^T$ 。显然，对于一个正常点来说，虽然它的齐次坐标形式不唯一，但它有唯一的规范化齐次坐标形式。

正常点的齐次坐标与非齐次坐标可以相互转换。如果一个平面正常点的坐标为 $(x_1, x_2)^T$ ，则它的规范化齐次坐标为 $(x_1, x_2, 1)^T$ ，它的齐次坐标为 $k(x_1, x_2, 1)^T, k \neq 0$ 。如果一个平面正常点的齐次坐标为 $(x_1, x_2, x_3)^T$ ，则它的非齐次坐标表示为 $(x_1/x_3, x_2/x_3)^T$ 。

一般情况下，对于旋转变换，我们只考虑针对正常点的情况。设旋转变换之前的点 $(x, y)^T$ 的规范化齐次坐标为 $\mathbf{x} = (x, y, 1)^T$ ，变换之后的点 $(\dot{x}, \dot{y})^T$ 的规范化齐次坐标为 $\mathbf{\dot{x}} = (\dot{x}, \dot{y}, 1)^T$ ，则旋转变换的表达式（3-2）可以重新表示为，

$$\begin{bmatrix} \dot{x} \\ \dot{y} \\ 1 \end{bmatrix} = \begin{bmatrix} \cos \theta & -\sin \theta & 0 \\ \sin \theta & \cos \theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \quad (3-3)$$

简记为，

$$\mathbf{\dot{x}} = \begin{bmatrix} R_{2 \times 2} & \mathbf{0}_{2 \times 1} \\ \mathbf{0}_{1 \times 2} & 1 \end{bmatrix} \mathbf{x} \quad (3-4)$$

这样，我们便知，表达平面上的旋转关系的变换矩阵 H 应该具有如下形式，

$$H_{3 \times 3} = \begin{bmatrix} R_{2 \times 2} & \mathbf{0}_{2 \times 1} \\ \mathbf{0}_{1 \times 2} & 1 \end{bmatrix} \quad (3-5)$$

其中， $R_{2 \times 2}$ 为正交矩阵且 $\det(R_{2 \times 2})=1$ 。容易知道，平面内的旋转变换有1个自由度。

3.1.2 欧氏变换 (Euclidean transformation)

在数学类书籍中，一般把同时考虑了旋转、反射与平移的几何变换称为欧氏变换。但在计算机视觉与机器人领域，一般不会考虑反射变换。因此，本书所讲的欧氏变换是由旋转和平移复合而成的，不考虑反射的情况。在机器人领域，这种复合了旋转和平移、而不考虑反射的几何变换也被称为**特殊欧氏变换**^[2]。

对于欧氏变换，一般也只考虑针对正常点的情况。设变换之前点的规范化齐次坐标为 $\mathbf{x} = (x, y, 1)^T$ ，该点经历了一个绕原点的逆时针旋转，旋转角度为 θ ，之后又经历了一次平移，平移量为 $(t_x, t_y)^T$ 。设变换之后点的规范化齐次坐标为 $\mathbf{\dot{x}} = (\dot{x}, \dot{y}, 1)^T$ ，则 \mathbf{x} 与 $\mathbf{\dot{x}}$ 的关系为，

$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = \begin{bmatrix} \cos\theta & -\sin\theta & t_x \\ \sin\theta & \cos\theta & t_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \quad (3-6)$$

简记为,

$$\mathbf{x}' = \begin{bmatrix} R_{2 \times 2} & \mathbf{t}_{2 \times 1} \\ \mathbf{0}_{1 \times 2} & 1 \end{bmatrix} \mathbf{x} \quad (3-7)$$

其中, $R_{2 \times 2}$ 为正交矩阵且 $\det(R_{2 \times 2})=1$, $\mathbf{t}_{2 \times 1}=(t_x, t_y)^T$ 。

这样, 我们便知, 表达平面上的欧氏变换的矩阵 H 应该具有如下形式,

$$H_{3 \times 3} = \begin{bmatrix} R_{2 \times 2} & \mathbf{t}_{2 \times 1} \\ \mathbf{0}_{1 \times 2} & 1 \end{bmatrix} \quad (3-8)$$

其中的 R 和 \mathbf{t} 与公式 (3-7) 中相同。同旋转变换相比, 欧氏变换多了两个刻画平移量的自由度, 因此平面内的欧氏变换有三个自由度。同时, 不难注意到, 旋转是欧氏变换的一个特例。

3.1.3 相似变换 (Similarity transformation)

在式 (3-6) 所表达的欧氏变换 (依然不考虑反射) 的基础上再复合上一个各向同性 (isotropic) 的缩放, 就得到了相似变换,

$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = \begin{bmatrix} s\cos\theta & -s\sin\theta & t_x \\ s\sin\theta & s\cos\theta & t_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \quad (3-9)$$

其中, $s>0$ 是刻画缩放程度的标量。相应地, 式 (3-9) 可简记为,

$$\mathbf{x}' = \begin{bmatrix} sR_{2 \times 2} & \mathbf{t}_{2 \times 1} \\ \mathbf{0}_{1 \times 2} & 1 \end{bmatrix} \mathbf{x} \quad (3-10)$$

其中, $R_{2 \times 2}$ 为正交矩阵且 $\det(R_{2 \times 2})=1$, $\mathbf{t}_{2 \times 1}=(t_x, t_y)^T$ 。

因此, 表达平面上的相似变换的矩阵 H 应该具有如下形式,

$$H_{3 \times 3} = \begin{bmatrix} sR_{2 \times 2} & \mathbf{t}_{2 \times 1} \\ \mathbf{0}_{1 \times 2} & 1 \end{bmatrix} \quad (3-11)$$

其中的 s 、 R 和 \mathbf{t} 与公式 (3-10) 中相同。同欧氏变换相比, 相似变换多了一个控制缩放比例的自由度, 因此平面内的相似变换有四个自由度。同时注意到, 欧氏变换是相似变换的一个特例。

3.1.4 仿射变换 (Affine transformation)

可以看到, 相似变换相较于欧氏变换来说, 我们对变换矩阵左上角 2×2 的子矩阵的要求放松了: 在欧氏变换中, 要求这个 2×2 的子矩阵是个能表达旋转的矩阵 (正交且行列式为 1), 而在相似变换中, 只要求这个 2×2 的子矩阵是旋转矩阵的常数倍即可。如果我们继续放松对

这个子矩阵的要求，只要求它是一个 2×2 的非奇异矩阵，那么得到的相应矩阵所能刻画的线性几何变换就称为仿射变换。

我们也只考虑针对正常点来定义仿射变换。设变换之前点的规范化齐次坐标为 $\mathbf{x}=(x,y,1)^T$ ，该点经历了一个仿射变换，设变换之后点的规范化齐次坐标为 $\mathbf{x}'=(x',y',1)^T$ ，

则 \mathbf{x} 与 \mathbf{x}' 的关系为，

$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & t_x \\ a_{21} & a_{22} & t_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \quad (3-12)$$

其中，左上角矩阵 $A_{2\times 2}=\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$ 非奇异。式(3-12)可简记为，

$$\mathbf{x}' = \begin{bmatrix} A_{2\times 2} & \mathbf{t}_{2\times 1} \\ \mathbf{0}_{1\times 2} & 1 \end{bmatrix} \mathbf{x} \quad (3-13)$$

其中， $\mathbf{t}_{2\times 1}=(t_x, t_y)^T$ 。

因此，表达平面上的仿射变换的矩阵 H 应该具有如下形式，

$$H_{3\times 3} = \begin{bmatrix} A_{2\times 2} & \mathbf{t}_{2\times 1} \\ \mathbf{0}_{1\times 2} & 1 \end{bmatrix} \quad (3-14)$$

其中的 A 和 \mathbf{t} 与公式(3-13)中相同。由于矩阵 A 包括了4个独立的元素， \mathbf{t} 是一个二维向量，所以平面内的仿射变换总共有6个自由度。同时注意到，相似变换是仿射变换的一个特例。

我们可以进一步来理解一下矩阵 A 所带来的四个自由度的几何意义。由于 A 是二阶非奇异矩阵，它必然具有如下的奇异值分解形式，

$$A=UDV^T \quad (3-15)$$

其中， U 和 V 为正交矩阵， $D=\begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}$ ， $\lambda_1>0, \lambda_2>0$ 。因此 $A=UV^TVDV^T=UV^T(VDV^T)$ 。

由于 V^T 是正交矩阵，从几何上来说，它表示某个旋转角为 ϕ 的旋转，记为 $R(\phi)$ 。那么，相

应地， V 所表示的旋转一定是 $R(-\phi)$ 。 UV^T 也是正交矩阵，它表示某个旋转角为 θ 的旋转，

记为 $R(\theta)$ 。这样， $A=R(\theta)R(-\phi)\begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}R(\phi)$ 。如果把 A 作用在几何图形上，相当于先把该

图形绕原点旋转角度 ϕ ，之后再沿X和Y方向进行缩放，缩放系数分别为 λ_1 和 λ_2 ，之后旋转回原来的位置，最后再旋转角度 θ 。

我们通过一个例子来感受一下仿射变换。在图3-3中，(a)是原始图像，(b)是经由矩

阵 $\begin{bmatrix} 0.9 & 0.3 & 0 \\ 0.3 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$ 所定义的仿射变换得到的结果。可以看到，原来几乎是正圆形的盘子在这个

仿射变换之后，变成了椭圆形。由此可见，在仿射变换之下，角度有可能是会被保持的。

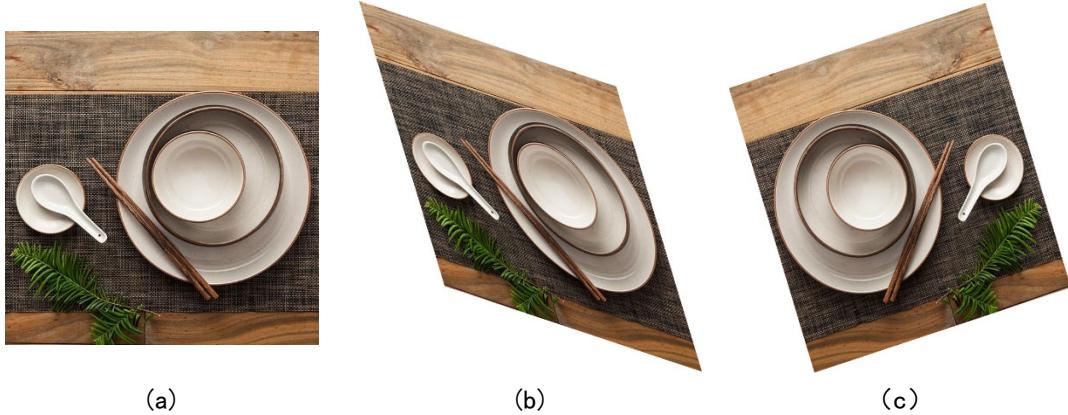


图 3-3：仿射变换举例。(a) 原始图像；(b) 和 (c) 是对图像 (a) 进行仿射变换之后得到的结果，(b) 图像保持了原始图像的方向性，而 (c) 图像中方向发生了翻转。

在前面谈到旋转变换、欧氏变换和相似变换时，我们都强调了变换要保持方向性的问题。对于这三类变换来说，只有当其表达矩阵（式 3-5、式 3-8 和式 3-11）中的 R 的行列式为 1 的时候，变换才会保持图像的方向性。对于仿射变换，也有类似的结论。如果不对式 3-13 中的 A 加以约束，得到的仿射变换可能会改变图形的方向性。只有当 $\det(A)>0$ 时，对应的仿射变换才会保持图形的方向性；当 $\det(A)<0$ 时，图像的方向会改变。事实上，可以证明^[3]：在平面上有两个方向不同的矢量 \mathbf{a} 、 \mathbf{b} ，当平面上发生了由式 3-13 所表达的仿射变换后， \mathbf{a} 、 \mathbf{b} 两个矢量相应地变换为矢量 \mathbf{a}' 和 \mathbf{b}' ，那么以矢量 \mathbf{a}' 、 \mathbf{b}' 为邻边所围成的平行四边形的定向面积与以矢量 \mathbf{a} 、 \mathbf{b} 为邻边所围成的平行四边形的定向面积之比为 $\det(A)$ 。所以，当 $\det(A)<0$

时，变换之后图形的方向性会发生变化。比如，在图 3-3 中，对 (a) 图像施加由 $\begin{bmatrix} -0.9 & 0.3 & 0 \\ 0.3 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$

（其行列式小于零）所定义的仿射变换后得到 (c) 图像，可以看到 (c) 图像与 (a) 图像相比，图形的方向性发生了改变。因此，如果只考虑保持方向的仿射变换，需要要求 $\det(A)>0$ 。

3.1.5 射影变换 (Projective transformation)

在之前的讨论中，表达线性几何变换的矩阵 H 都有一个共同的特点，那就是最后一行是 $(0, 0, 1)$ 。如果继续放松对矩阵 H 的要求，只要求它是一个非奇异的 3×3 的矩阵，那么此时 H 所能表达的线性几何变换称为射影变换。与前面讨论的几种变换不同，射影变换不但可以定义在正常点上，也可以定义在无穷远点上。在射影变换下，不再区分正常点和无穷远

点，它可以把正常点变换到无穷远点，也可以把无穷远点变换到正常点。因此，我们对点的坐标表达就不再限定为规范化齐次坐标了（因为无穷远点没有规范化齐次坐标），而是使用一般化的齐次坐标表达。

假设变换之前点的齐次坐标为 $\mathbf{x} = (x_1, x_2, x_3)^T$ ，经过射影变换 H 之后，该点就变为了 $H\mathbf{x}$ 。

我们注意到，由于点的齐次坐标不具有唯一性， $H\mathbf{x}$ 与 $kH\mathbf{x}$, $\forall k \neq 0$ 代表的都是同一个平面点。这也就意味着 H 与 kH , $\forall k \neq 0$ 表达的实际上是同一个射影变换。因此，尽管从形式上看，射影变换矩阵 H 有 9 个元素，但实际上它只有 8 个自由度。

如果点 $\mathbf{x} = (x_1, x_2, x_3)^T$ 与点 $\mathbf{x}' = (x'_1, x'_2, x'_3)^T$ 可以由射影变换 H 对应起来，那么 $H\mathbf{x}$ 与 \mathbf{x}' 之间是一个常数倍的关系，即必存在一个数 c ，使得 $c\mathbf{x}' = H\mathbf{x}$ 。也就是说，如果点 \mathbf{x} 经过射影

变换 $H = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix}$ 变换到了 \mathbf{x}' ，那么它们满足关系，

$$c \begin{pmatrix} x'_1 \\ x'_2 \\ x'_3 \end{pmatrix} = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \quad (3-16)$$

其中， c 是一个与点 \mathbf{x}' 有关的数。

最后再来谈一下射影变换对图形方向性的保持问题。我们在前面提到，对于由式 3-13 所定义的仿射变换来说，可以根据 $\det(A)$ 的符号来判断该变换是否能够保持图形的方向性。但对于射影变换来说，我们无法判定一个射影变换是否会保持图形的方向性^[4]。

3.2 变换群与几何学

德国数学家费利克斯·克莱因 (Felix Klein) 在 1872 年运用变换群的思想来区分各种几何学，他提出，每一种几何都是在研究图形在一定的变换群下不变的性质^[3]。这就是著名的爱尔兰根纲领 (Erlangen Program)。



图 3-4：费利克斯·克莱因 (Felix Klein, 1849 年 4 月 25 日—1925 年 6 月 22 日)，德国数学家。克莱因生于德国杜塞多夫，在爱尔兰根、慕尼黑和莱比锡当过教授，最后在哥廷根教授数学。他的主要课题是非欧几何、群论和复变函数论。他发布的爱尔兰根纲领将各种几何用它

们的基础对称群来分类，是对当时多个数学分支的一个综合导向，影响深远。

3.2.1 群的定义

群是一个代数学的概念。由于几何与代数的密切关系，这个概念对于几何学的研究不但重要的而且产生了深远的影响。

定义 3.1 群^[5]。设有一个集合 \mathcal{G} ，在其上元素之间定义操作“ \circ ”，如果集合 \mathcal{G} 关于运算 \circ 满足下列条件：

- 1) 封闭性： $\forall g_1, g_2 \in \mathcal{G}, \exists g_3 \in \mathcal{G}$ ，使得 $g_3 = g_1 \circ g_2$ ；
- 2) 结合性： $\forall g_1, g_2, g_3 \in \mathcal{G}, g_1 \circ (g_2 \circ g_3) = (g_1 \circ g_2) \circ g_3$ ；
- 3) 存在单位元： $\exists e \in \mathcal{G}$ ，使得 $\forall g \in \mathcal{G}, e \circ g = g \circ e = g$ ， e 称为 \mathcal{G} 中的单位元；
- 4) 每个元素存在逆元： $\forall g \in \mathcal{G}, \exists g^{-1} \in \mathcal{G}$ ，使得 $g \circ g^{-1} = g^{-1} \circ g = e$ 。

则称 \mathcal{G} 在运算 \circ 之下构成一个群。

3.2.2 线性几何变换群

根据群的定义不难验证，在 3.1 节中定义的 5 种表达线性几何变换的矩阵元素在普通矩阵乘法运算之下均构成群，分别称为旋转变换群（也称为特殊正交群^[2]，Special orthogonal group）、欧氏变换群（Euclidean group）¹、相似变换群、仿射变换群和射影变换群。在机器人学中，刻画机器人的“保向”刚体运动是最基本的问题。因此，在该领域中，最常见的特殊正交群和特殊欧氏变换群都有着通用的表达记号。在二维空间（欧氏平面）中，特殊正交群被记为 SO(2)，特殊欧氏变换群被记为 SE(2)；在三维空间中，特殊正交群被记为 SO(3)，特殊欧氏变换群被记为 SE(3)。

作为例子，我们一起来验证一下表达平面内旋转变换的矩阵集合，

$$\mathcal{G} = \{R \in \mathbb{R}^{2 \times 2} : |RR^T = I, \det(R) = 1\}$$

在普通矩阵乘法之下构成群。

1) 验证封闭性。假设 $g_1 = R_1 \in \mathcal{G}, g_2 = R_2 \in \mathcal{G}$ ，则 $g_1 g_2 = R_1 R_2$ ，则有

$$(g_1 g_2)(g_1 g_2)^T = (R_1 R_2)(R_1 R_2)^T = R_1 R_2 R_2^T R_1^T = I, \text{ 且 } \det(g_1 g_2) = \det(R_1 R_2) = \det(R_1) \det(R_2) = 1, \text{ 则}$$

$$g_1 g_2 \in \mathcal{G}$$

¹ 一般数学类书籍中所说的欧氏变换会包含反射的情况。但本书中所说的欧氏变换不考虑反射的情况，这类欧氏变换群在机器人学中也称为特殊欧氏群（special Euclidean group）。

2) 验证结合性。在普通矩阵乘法之下，结合性显然成立。

3) 验证存在单位元。单位元为二阶单位矩阵 $I_{2 \times 2}$ 。

4) 验证每个元素存在逆元。设 $g = R \in \mathcal{G}$ 。我们验证 R 的逆矩阵 R^{-1} 也属于 \mathcal{G} :

$$R^{-1}(R^{-1})^T = R^T R = I, \text{ 且 } \det(R^{-1}) = \frac{1}{\det(R)} = 1, \text{ 则 } R^{-1} \in \mathcal{G}。另，由于 } g R^{-1} = R R^{-1} = I,$$

$$R^{-1}g = R^{-1}R = I, \text{ 则 } g \text{ 的逆元为 } g^{-1} = R^{-1}。$$

我们现在知道了前面提到的 5 种几何变换都构成群，那么就可以得到一个重要推论：

推论 3.1 由于群的封闭性，两个同类型的几何变换复合在一起，得到的复合变换依然还是这个类型的几何变换。

比如，平面内两个欧氏变换复合在一起，得到的复合变换依然是平面内的欧氏变换，该复合变换的自由度依然是 3 个，它不会变成具有 4 个自由度的相似变换，也不会变成具有 6 个自由度的仿射变换。

另外，不难理解，我们提到的这 5 个变换群具有如下包含关系：

推论 3.2 旋转变换群 \subset 欧氏变换群 \subset 相似变换群 \subset 仿射变换群 \subset 射影变换群。

如克莱因指出的，每一种几何都是在研究图形在一定的变换群下不变的性质。那么接下来看一下，在我们所定义的 5 种变换群下，图形会具有哪些不变的几何性质。

显然，在旋转变换与欧氏变换之下，两点之间的距离是保持不变的。从距离这个基本不变量出发，我们可以推导出其他的不变量，比如两条线之间的夹角、图形的面积等。

在相似变换下，基本的几何不变量是相似比。假设变换之前有任意点 $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$ 和 \mathbf{x}_4 ，变

换之后它们的对应点分别为 $\mathbf{x}'_1, \mathbf{x}'_2, \mathbf{x}'_3$ 和 \mathbf{x}'_4 。相似比不变指的是 $\frac{\|\mathbf{x}_1 \mathbf{x}_2\|}{\|\mathbf{x}_3 \mathbf{x}_4\|} = \frac{\|\mathbf{x}'_1 \mathbf{x}'_2\|}{\|\mathbf{x}'_3 \mathbf{x}'_4\|}$ 。由相似比这个

基本不变量，我们也可以推导出相似变换下的其他不变量，比如两条线之间的夹角、直线之间的平行关系等。

定义 3.2 简单比值^[3]。设 $\mathbf{a}, \mathbf{b}, \mathbf{c}$ 是共线三点，在此直线上取定一个单位向量 \mathbf{e} ，若 $\overrightarrow{\mathbf{ab}} = \lambda \mathbf{e}$ ，则称 λ 是线段 \mathbf{ab} 的代数长，就用 \mathbf{ab} 表示线段 \mathbf{ab} 的代数长。称 $\frac{\mathbf{ab}}{\mathbf{bc}}$ 为共线三点 $\mathbf{a}, \mathbf{b}, \mathbf{c}$ 的简单比值，记作 $(\mathbf{a}, \mathbf{b}, \mathbf{c})$ ，即 $(\mathbf{a}, \mathbf{b}, \mathbf{c}) = \frac{\mathbf{ab}}{\mathbf{bc}}$ 。

在仿射变换下，基本的几何不变量是简单比值，即仿射变换保持共线三点的简单比值不变。由简单比值这个基本不变量，我们也可以推导出仿射变换下的其他不变量：直线之间的平行关系在变换前后保持不变；两个图形的面积比在变换前后保持不变；若 \mathbf{c} 是有向线段 $\overrightarrow{\mathbf{ab}}$ 的中点，则变换之后它的对应点 \mathbf{c}' 也是对应有向线段 $\overrightarrow{\mathbf{a}'\mathbf{b}'}$ 的中点。需要格外注意的是，仿射变换不会保持角度，比如一个矩形在仿射变换之下可能会变成平行四边形。

在射影变换下，基本的几何不变量是交比。由于交比在本书中其他地方不会再涉及，我们就不再详加介绍了。相对于前面几种变换群来说，射影变换群是最大的，同时它能够保持的几何不变量是最少的。比如，在仿射变换下，直线的平行关系是可以被保持的，但一般的射影变换并不能保持直线间的平行关系，这就意味着一个矩形在射影变换之后可能会变成一个一般的四边形。但射影变换毕竟是线性几何变换，一些最基本的几何关系还是能被保持的，比如：它会把直线变换到直线，变换前是不重合的两点在射影变换后依然不重合等等。

3.3 三维空间中的线性几何变换

在 3.1 节中讲述的二维平面上的线性几何变换与在 3.2 节中讲述的关于变换群与几何学的有关结论，可以直接推广到三维空间。为了便于读者查阅，我们把在三维空间中的线性几何变换的有关结论总结在本节。

与二维情况一样，三维空间中的旋转变换、欧氏变换、相似变换和仿射变换都是针对正常点（非无穷远点）进行的。设变换之前三维空间点的规范化齐次坐标为 $\mathbf{x} = (x, y, z, 1)^T$ ，变换之后点的规范化齐次坐标为 $\mathbf{x}' = (x', y', z', 1)^T$ 。

在旋转变换下， \mathbf{x} 与 \mathbf{x}' 的关系为，

$$\mathbf{x}' = \begin{bmatrix} R_{3 \times 3} & \mathbf{0}_{3 \times 1} \\ \mathbf{0}_{1 \times 3} & 1 \end{bmatrix} \mathbf{x} \quad (3-17)$$

其中， $R_{3 \times 3}$ 为正交矩阵且 $\det(R_{3 \times 3}) = 1$ 。表达三维空间旋转变换的矩阵元素在普通矩阵乘法运算之下构成群，称为三维空间下的旋转变换群。

在欧氏变换下， \mathbf{x} 与 \mathbf{x}' 的关系为，

$$\mathbf{x}' = \begin{bmatrix} R_{3 \times 3} & \mathbf{t}_{3 \times 1} \\ \mathbf{0}_{1 \times 3} & 1 \end{bmatrix} \mathbf{x} \quad (3-18)$$

其中， $R_{3 \times 3}$ 为正交矩阵且 $\det(R_{3 \times 3}) = 1$ ， $\mathbf{t}_{3 \times 1} = (t_x, t_y, t_z)^T$ 为平移向量。表达三维空间欧氏变换的矩阵元素在普通矩阵乘法运算之下构成群，称为三维空间下的欧氏变换群。在三维旋转变换群与欧氏变换群下，空间点之间的距离保持不变。

在相似变换下， \mathbf{x} 与 \mathbf{x}' 的关系为，

$$\mathbf{x}' = \begin{bmatrix} sR_{3 \times 3} & \mathbf{t}_{3 \times 1} \\ \mathbf{0}_{1 \times 3} & 1 \end{bmatrix} \mathbf{x} \quad (3-19)$$

其中， $R_{3 \times 3}$ 为正交矩阵且 $\det(R_{3 \times 3}) = 1$ ， $\mathbf{t}_{3 \times 1} = (t_x, t_y, t_z)^T$ 为平移向量， $s > 0$ 为尺度缩放系数。表达三维空间相似变换的矩阵元素在普通矩阵乘法运算之下构成群，称为三维空间下的相似变换群。在三维相似变换群下，空间点之间距离的相似比保持不变。

在仿射变换下， \mathbf{x}' 与 \mathbf{x} 的关系为，

$$\mathbf{x}' = \begin{bmatrix} A_{3 \times 3} & \mathbf{t}_{3 \times 1} \\ \mathbf{0}_{1 \times 3} & 1 \end{bmatrix} \mathbf{x} \quad (3-20)$$

其中， $A_{3 \times 3}$ 为非奇异矩阵且 $\det(A) > 0$ ， $\mathbf{t}_{3 \times 1} = (t_x, t_y, t_z)^T$ 为平移向量。表达三维空间仿射变换的矩阵元素在普通矩阵乘法运算之下构成群，称为三维空间下的仿射变换群。在三维仿射变换群下，空间共线三点之间的简单比值保持不变。另外，与二维情况类似，可以证明：在三维空间中有三个不共面矢量 \mathbf{a} 、 \mathbf{b} 、 \mathbf{c} ，当该空间发生了由式 3-20 所表达的仿射变换后， \mathbf{a} 、 \mathbf{b} 、 \mathbf{c} 三个矢量相应地变换为矢量 \mathbf{a}' 、 \mathbf{b}' 和 \mathbf{c}' ，那么以矢量 \mathbf{a}' 、 \mathbf{b}' 和 \mathbf{c}' 为邻边所围成的平行六面体的定向体积与以矢量 \mathbf{a} 、 \mathbf{b} 和 \mathbf{c} 为邻边所围成的平行六面体的定向体积之比为 $\det(A)$ ，因此 $\det(A)$ 也被形象地称为仿射变换的“变积系数”^[5]。

表 3-1：二维空间与三维空间下的线性几何变换（ n 为空间维度， $n=2, 3$ ）

变换名称	矩阵表达式	二维情况下 自由度个数	三维情况下 自由度个数	不变量
旋转变换	$\begin{bmatrix} R_{n \times n} & \mathbf{0}_{n \times 1} \\ \mathbf{0}_{1 \times n} & 1 \end{bmatrix}$, R 为正交矩阵且 $\det(R)=1$	1	3	长度，角度，面积（体积）
欧氏变换	$\begin{bmatrix} R_{n \times n} & \mathbf{t}_{n \times 1} \\ \mathbf{0}_{1 \times n} & 1 \end{bmatrix}$, R 为正交矩阵且 $\det(R)=1$	3	6	长度，角度，面积（体积）
相似变换	$\begin{bmatrix} sR_{n \times n} & \mathbf{t}_{n \times 1} \\ \mathbf{0}_{1 \times n} & 1 \end{bmatrix}$, R 为正交矩阵且 $\det(R)=1, s>0$	4	7	相似比，角度，面积（体积）比
仿射变换	$\begin{bmatrix} A_{n \times n} & \mathbf{t}_{n \times 1} \\ \mathbf{0}_{1 \times n} & 1 \end{bmatrix}$, A 为非奇异矩阵且 $\det(A)>0$	6	12	简单比，面积（体积）比，平行关系
射影变换	$H_{(n+1) \times (n+1)}$, H 为非奇异矩阵	8	15	交比，共线关系

如果三维空间点（齐次坐标表示） $\mathbf{x}=(x_1, x_2, x_3, x_4)^T$ 与点 $\mathbf{x}'=(\dot{x}_1, \dot{x}_2, \dot{x}_3, \dot{x}_4)^T$ 可以由射影变

换 $H = \begin{bmatrix} h_{11} & h_{12} & h_{13} & h_{14} \\ h_{21} & h_{22} & h_{23} & h_{24} \\ h_{31} & h_{32} & h_{33} & h_{34} \\ h_{41} & h_{42} & h_{43} & h_{44} \end{bmatrix}$ 对应起来，那么它们满足关系，

$$c \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{bmatrix} h_{11} & h_{12} & h_{13} & h_{14} \\ h_{21} & h_{22} & h_{23} & h_{24} \\ h_{31} & h_{32} & h_{33} & h_{34} \\ h_{41} & h_{42} & h_{43} & h_{44} \end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} \quad (3-21)$$

其中， c 是一个与点 \mathbf{x} 有关的数。表达三维空间射影变换的矩阵元素在普通矩阵乘法运算之下构成群，称为三维空间下的射影变换群。在三维射影变换群下，空间共线四点之间的交比值保持不变。

表 3-1 总结了二维空间与三维空间下的线性几何变换的主要相关结论。

3.4 习题

(1) 请证明形如式 3-8 所定义的表达平面内欧氏变换的矩阵元素集合构成群。

参考文献

- [1] 李世栋, 乐经良, 冯卫国, 王纪林, 线性代数, 科学出版社, 2000 年。
- [2] 高翔, 张涛等, 视觉 SLAM 十四讲: 从理论到实践 (第二版), 电子工业出版社, 2019 年。
- [3] 丘维声, 解析几何 (第二版), 北京大学出版社, 1996 年。
- [4] Richard Hartley, Andrew Zisserman, Multiple View Geometry in Computer Vision (2nd Edition), Cambridge University Press, 2004.
- [5] 方德植, 陈奕培, 射影几何, 高等教育出版社, 1983 年。

第4章 特征点检测与匹配

在这一章中，我们将学习如何在给定图像中检测出特征点、如何对特征点构建特征描述子向量以及如何根据特征描述子建立起两幅图像中特征点的匹配关系。

图像特征点检测与匹配算法是很多高层计算机视觉应用系统的基石。因此，从上世纪 70 年代开始到本世纪初，该问题一直是计算机视觉领域的研究热点。在讲述具体的图像特征点检测与匹配算法之前，我们首先来定性说明一下一个好的特征点检测算法以及描述子构造算法应该具有哪些性质。对于图像特征点来说，我们希望它们具有如下性质：

- **局部性。**特征点的位置需要是容易准确定位的。比如，图像中的边缘点就不具备很好的局部性，因为沿着边缘方向移动，所经过的点的形态都高度相似。
- **稀疏性。**图像上的特征点相对于图像上全体像素点来说，其数量应该是比较稀少。如果太过于稠密，会显著提升后续处理过程的计算代价。
- **对光照变化的稳定性。**当环境的光照条件发生了变化之后，我们希望特征点检测算法依然能够找到相同的特征点。
- **对几何变换的稳定性。**当相机拍摄视角发生了改变之后，图像平面会发生相应的几何变换，我们希望特征点检测算法依然能够检测出对应的特征点。

对于特征描述子构建算法来说，我们希望它们具有如下性质：

- **高判别性。**设 \mathbf{x}_1 、 \mathbf{x}_2 为两个图像特征点， \mathbf{d}_1 和 \mathbf{d}_2 分别为它们的特征描述子。若 \mathbf{x}_1 与 \mathbf{x}_2 对应于物理场景中的同一点，我们期望 \mathbf{d}_1 和 \mathbf{d}_2 相同；若 \mathbf{x}_1 与 \mathbf{x}_2 对应于物理场景中不同的点，我们期望 \mathbf{d}_1 和 \mathbf{d}_2 距离较大。
- 当环境的光照条件发生了变化之后，我们希望对相应特征点所构建的特征描述子能够保持不变。
- **对光照变化的稳定性。**当环境的光照条件发生了变化之后，我们希望对相应特征点所构建的特征描述子能够保持不变。
- **对几何变换的稳定性。**当相机拍摄视角发生了改变之后，图像平面会发生相应的几何变换，我们希望对相应特征点所构建的特征描述子能够保持不变。

4.1 哈里斯角点及其描述子

4.1.1 哈里斯角点检测算法设计思路

哈里斯角点（Corners）检测算法是由英国学者 Chris Harris 和 Mikes Stephens 于 1988 年提出来的^[1]，是一种经典的常用的图像特征点检测算法。哈里斯等认为，图像中的角点是一类非常稳定的、稀疏的、特殊的点，可以作为图像的特征点。那么应该如何来判断一个点是不是角点呢？

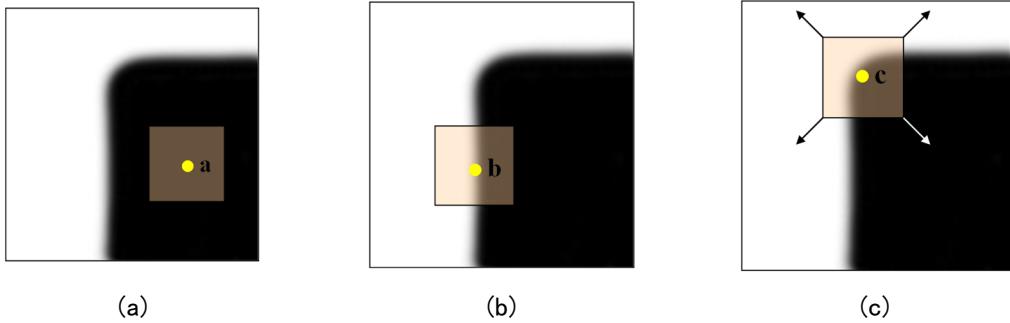


图 4-1：哈里斯角点检测算法设计思路。(a) **a** 点位于平坦图像区域之上，不是角点；(b) **b** 点位于图像边缘之上，不是角点；(c) **c** 点为图像角点，其覆盖窗口无论沿任何方向移动，新窗口所覆盖的图像块与原窗口所覆盖的图像块相比，像素值都会发生较大的变化。

如图 4-1 所示，我们可以从一个简单的理想模型出发来进行定性分析。在图 4-1 (a)、(b) 和 (c) 中，看看被考察点 **a**、**b**、**c** 是否是角点。通过直观观察，不难理解，只有图 4-1 (c) 中的 **c** 点是角点，其他两个都不是。那图 4-1 (c) 中的 **c** 点与图 4-1 (a)、(b) 中的被考察点 **a**、**b** 相比，有什么特点呢？考虑在被考察点周围取一个邻域窗口 W ，如果我们将 W 移动一个小量，就会到达一个新窗口 W' 。我们来观察一下 W' 与 W 所覆盖的图像块的像素值的变化 s_W 。在图 4-1 (a) 中无论朝哪个方向移动 W ，所引起的 s_W 都会很小，这说明被考察点 **a** 位于图像平坦区域上，不会是角点。在图 4-1 (b) 中，如果 W 是沿垂直方向移动到达 W' 的话， s_W 会很小，而当 W 是沿水平方向移动到达 W' 的话， s_W 会比较大，这说明被考察点 **b** 位于边缘 (edge) 上，也不是角点。而在图 4-1 (c) 中，不论 W' 是由 W 沿何方向移动得到的， s_W 都会比较大。基于上述分析，图像中的角点被定义为：在点 \mathbf{x} 周围取一个邻域窗口 W ，无论沿哪个方向移动 W ，新窗口 W' 所覆盖的图像区域与旧窗口 W 所覆盖的图像区域在像素值上都会有很大变化，那么点 \mathbf{x} 即为角点。

4.1.2 哈里斯角点检测算法实现

哈里斯角点检测算法就是按照 4.1.1 节中对角点属性的定性分析来设计的。对于图像 f 上某点 $\mathbf{x}=(x, y)$ ，考察该点是否为角点。在图像 f 上，以 \mathbf{x} 为中心取窗口 W ， W 移动小量 $(\Delta x, \Delta y)$ 之后，新旧窗口所覆盖图像区域的像素值的差异可表达为，

$$s_W(\Delta x, \Delta y) = \sum_{(x_i, y_i) \in W} (f(x_i, y_i) - f(x_i + \Delta x, y_i + \Delta y))^2 \quad (4-1)$$

由于 $(\Delta x, \Delta y)$ 很小，我们可以对 $f(x_i + \Delta x, y_i + \Delta y)$ 进行一阶泰勒近似，

$$f(x_i + \Delta x, y_i + \Delta y) \approx f(x_i, y_i) + \left(\frac{\partial f}{\partial x} \Big|_{(x_i, y_i)}, \frac{\partial f}{\partial y} \Big|_{(x_i, y_i)} \right) \begin{pmatrix} \Delta x \\ \Delta y \end{pmatrix} \quad (4-2)$$

将式 4-2 带入式 4-1 可得,

$$\begin{aligned}
s_W(\Delta x, \Delta y) &\approx \sum_{(x_i, y_i) \in W} \left(f(x_i, y_i) - f(x_i, y_i) - \left(\frac{\partial f}{\partial x} \Big|_{(x_i, y_i)}, \frac{\partial f}{\partial y} \Big|_{(x_i, y_i)} \right) \begin{pmatrix} \Delta x \\ \Delta y \end{pmatrix} \right)^2 \\
&= \sum_{(x_i, y_i) \in W} \left(\left(\frac{\partial f}{\partial x} \Big|_{(x_i, y_i)}, \frac{\partial f}{\partial y} \Big|_{(x_i, y_i)} \right) \begin{pmatrix} \Delta x \\ \Delta y \end{pmatrix} \right)^2 \\
&= \sum_{(x_i, y_i) \in W} (\Delta x, \Delta y) \begin{pmatrix} \frac{\partial f}{\partial x} \Big|_{(x_i, y_i)} \\ \frac{\partial f}{\partial y} \Big|_{(x_i, y_i)} \end{pmatrix} \begin{pmatrix} \frac{\partial f}{\partial x} \Big|_{(x_i, y_i)}, \frac{\partial f}{\partial y} \Big|_{(x_i, y_i)} \end{pmatrix} \begin{pmatrix} \Delta x \\ \Delta y \end{pmatrix} \\
&= (\Delta x, \Delta y) \left\{ \sum_{(x_i, y_i) \in W} \begin{pmatrix} \frac{\partial f}{\partial x} \Big|_{(x_i, y_i)} \\ \frac{\partial f}{\partial y} \Big|_{(x_i, y_i)} \end{pmatrix} \begin{pmatrix} \frac{\partial f}{\partial x} \Big|_{(x_i, y_i)}, \frac{\partial f}{\partial y} \Big|_{(x_i, y_i)} \end{pmatrix} \right\} \begin{pmatrix} \Delta x \\ \Delta y \end{pmatrix} \\
&= (\Delta x, \Delta y) \begin{pmatrix} \sum_{(x_i, y_i) \in W} \left(\frac{\partial f}{\partial x} \Big|_{(x_i, y_i)} \right)^2 & \sum_{(x_i, y_i) \in W} \left(\frac{\partial f}{\partial x} \Big|_{(x_i, y_i)} \right) \left(\frac{\partial f}{\partial y} \Big|_{(x_i, y_i)} \right) \\
\sum_{(x_i, y_i) \in W} \left(\frac{\partial f}{\partial x} \Big|_{(x_i, y_i)} \right) \left(\frac{\partial f}{\partial y} \Big|_{(x_i, y_i)} \right) & \sum_{(x_i, y_i) \in W} \left(\frac{\partial f}{\partial y} \Big|_{(x_i, y_i)} \right)^2 \end{pmatrix} \begin{pmatrix} \Delta x \\ \Delta y \end{pmatrix}
\end{aligned} \tag{4-3}$$

我们令,

$$M = \begin{pmatrix} \sum_{(x_i, y_i) \in W} \left(\frac{\partial f}{\partial x} \Big|_{(x_i, y_i)} \right)^2 & \sum_{(x_i, y_i) \in W} \left(\frac{\partial f}{\partial x} \Big|_{(x_i, y_i)} \right) \left(\frac{\partial f}{\partial y} \Big|_{(x_i, y_i)} \right) \\ \sum_{(x_i, y_i) \in W} \left(\frac{\partial f}{\partial x} \Big|_{(x_i, y_i)} \right) \left(\frac{\partial f}{\partial y} \Big|_{(x_i, y_i)} \right) & \sum_{(x_i, y_i) \in W} \left(\frac{\partial f}{\partial y} \Big|_{(x_i, y_i)} \right)^2 \end{pmatrix} \tag{4-4}$$

, 则 $s_W(\Delta x, \Delta y) = (\Delta x, \Delta y) M \begin{pmatrix} \Delta x \\ \Delta y \end{pmatrix}$ 。如果我们让新旧窗口所覆盖图像区域的像素值的差异

$s_W(\Delta x, \Delta y)$ 为一个常数, 比如 $s_W(\Delta x, \Delta y) = 1$, 即

$$(\Delta x, \Delta y) M \begin{pmatrix} \Delta x \\ \Delta y \end{pmatrix} = 1 \tag{4-5}$$

那么, 方程式 4-5 代表了能够使得新旧窗口所覆盖区域像素值的差异为 1 的窗口移动量 $(\Delta x, \Delta y)$ 所形成的轨迹。对于式 4-5 来说, 矩阵 M 为已知量, 它是由点 \mathbf{x} 处的局部窗口 W 所唯一确定的。可以证明, 按式 4-4 方式所定义的矩阵 M 为半正定矩阵。实际上, 除非是极特殊情况 (比如窗口 W 所覆盖的图像块的像素值全部为相同常数), M 为正定矩阵。当 M 为正定矩阵时, 可以证明方程式 4-5 所描述的 $(\Delta x, \Delta y)$ 的轨迹为一个椭圆 (见附录 A)。显然,

该椭圆的几何属性完全由 M 决定。如图 4-2 (a) 所示, 假设 M 的两个特征值分别为 λ_1 和 λ_2 ,

且 $\lambda_1 \geq \lambda_2$, 则该椭圆的长半轴长度为 $\lambda_2^{-1/2}$, 其短半轴长度为 $\lambda_1^{-1/2}$, 该结论留作习题请读者完成证明。

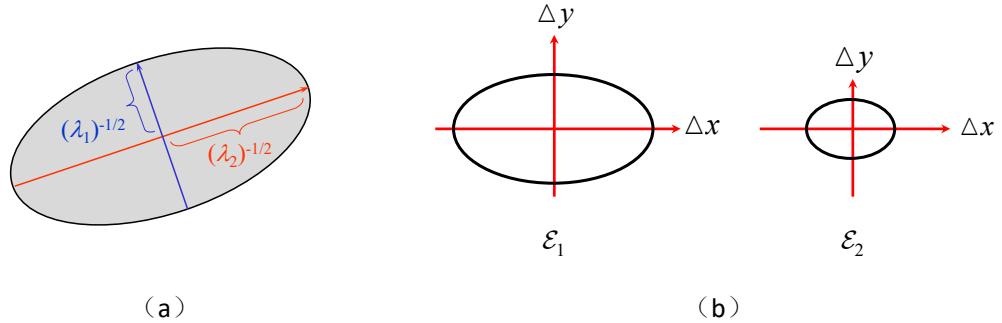


图 4-2: (a) 式 4-5 所确定的椭圆, 该椭圆的长半轴长度为 $\lambda_2^{-1/2}$ 、短半轴长度为 $\lambda_1^{-1/2}$; (b)

假设图像上有两个点 x_1 和 x_2 , 在它们周围取相同大小的窗口, 按照式 4-5 的方式得到两个相应的椭圆 E_1 和 E_2 , 那么较小的椭圆 E_2 所对应的点 x_2 更可能是一个角点。

考虑图像 I 上的两个点 x_1 和 x_2 , 在它们周围取相同大小的窗口, 之后按照式 4-4 的方式分别计算与 x_1 和 x_2 对应的实对称矩阵 M_1 和 M_2 。之后, 按照式 4-5, 我们可以有两个相应的椭圆 E_1 : $(\Delta x, \Delta y) M_1 \begin{pmatrix} \Delta x \\ \Delta y \end{pmatrix} = 1$ 和 E_2 : $(\Delta x, \Delta y) M_2 \begin{pmatrix} \Delta x \\ \Delta y \end{pmatrix} = 1$ 。假设椭圆 E_1 和 E_2 的形态如图 4-2(b)

所示, 那么相应地, x_1 和 x_2 中的哪一个更可能是一个角点呢? 答案是与较小的椭圆 E_2 所对应的点 x_2 更可能是一个角点。这是因为, 小的椭圆意味着只要对原始覆盖窗口施加一个很小的移动量就可以使窗口覆盖区域的像素值变化为 1, 而大的椭圆意味着要对原始覆盖窗口施加一个相对较大的移动量才可以使窗口覆盖区域的像素值变化为 1。实际上, 更准确地说, 不但要小而且要“接近于圆”的椭圆所对应的被考察点才更有可能是一个角点; “接近于圆”意味着无论沿哪个方向对覆盖该点的窗口施加相同幅度的移动量都会使得新旧窗口所覆盖的区域的像素值产生较为一致的变化, 这才符合我们在 4.1.1 节中对角点特性的定性分析。注意到, 小的椭圆意味着式 4-5 中的 M 的特征值会比较大, 而“接近于圆”则意味着 M 的两个特征值要差不多大, 这就说明我们可以根据 M 的特征值的情况来对角点进行判定: 当 M 的两个特征值都很大而且差不多大时, 它所刻画的点 x 更可能是角点。类似地, 对于 M 中其他特征值情况我们也可以得到相应判断: 当 λ_1 和 λ_2 都很小时, 点 x 更可能位于图像平滑区域上; 当 λ_1 和 λ_2 其中一个很大、另一个很小时, 点 x 可能位于图像边缘上。我们把 x 点所属类型与 λ_1 、 λ_2 之间的关系总结在了图 4-3 中。

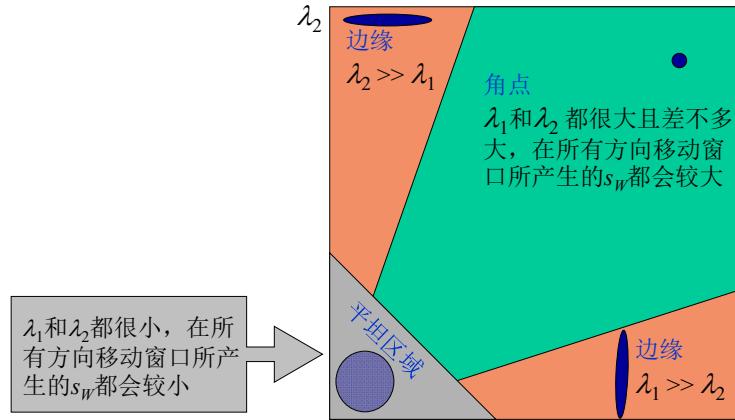


图 4-3：图像上一点 x 所在区域属性与 x 所对应的 M 矩阵的特征值 λ_1 和 λ_2 之间的关系。

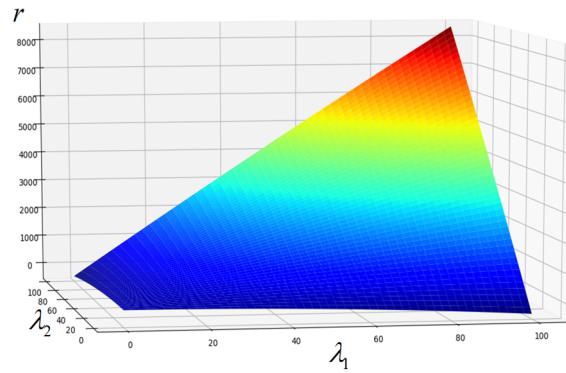


图 4-4：角点程度数值 r 与矩阵 M 的两个特征值 λ_1 和 λ_2 之间的关系，可以看到只有当 λ_1 和 λ_2 同时都很大时， r 才会很大。

在编程实现中，如果真的要对 M 进行特征值分解的话，角点检测操作的效率会很低，因为矩阵特征值分解的计算代价较高。幸运的是，Harris 和 Stephens 给出了一个计算点 x 处角点程度（cornerness）的经验公式，避免了对 M 进行显式的特征值分解。利用该公式，点 x 处的角点程度值 $r(x)$ 被表达为，

$$r(x) = \det(M(x)) - k(\text{trace}(M(x)))^2 \quad (4-6)$$

其中， $M(x)$ 表示按照式 4-4 的方式计算点 x 处的 M 矩阵， $\det(M(x))$ 表示计算矩阵 $M(x)$ 的行列式， $\text{trace}(M(x))$ 表示计算矩阵 $M(x)$ 的迹， k 为一个事先设定的超参数，一般设置为 0.04 到 0.06 之间。 r 的值越大，说明该点处是角点的可能性就越高。需要注意到，在式 4-6 中，虽然在计算 $r(x)$ 的过程中并没有显式计算 $M(x)$ 的特征值，但 $r(x)$ 的数值实质上是依赖于 $M(x)$ 的特征值的，这是因为 $\det(M(x))$ 与 $\text{trace}(M(x))$ 都完全决定于 $M(x)$ 的特征值。若 $M(x)$ 的两个特征值分别为 λ_1 和 λ_2 ，则 $\det(M(x)) = \lambda_1 \lambda_2$ ， $\text{trace}(M(x)) = \lambda_1 + \lambda_2$ 。图 4-4 清晰

地展示了 r 与 λ_1 和 λ_2 的关系。可以看出，只有当 λ_1 和 λ_2 同时都很大时， r 才会很大，而这刚好符合我们对角点属性的定性分析。矩阵行列式与迹的计算相比矩阵特征值分解来说，计算复杂度会低很多。

若点 \mathbf{x} 处的 $r(\mathbf{x})$ 大于预先设定的阈值 t ，即 $r(\mathbf{x}) > t$ ，则认为 \mathbf{x} 为一个候选角点。但为了得到图像 f 上合理的稀疏角点集合，我们还需要对候选角点集合进行一步后处理操作，非极大值抑制（non-maximum suppression）。这是因为如果 \mathbf{x} 处的 $r(\mathbf{x})$ 很大，则它的近邻 \mathbf{x}' 处的 $r(\mathbf{x}')$ 通常也会非常大，这就导致单一阈值化操作会认为 \mathbf{x} 附近的“一大片区域”都是角点，这显然与客观物理世界是不相符合的。非极大值抑制这个操作就在一个预设大小的局部范围内，只保留角点程度值最大的候选角点，而把该局部区域内其他的候选角点剔除出角点集合，从而保证最后得到的角点集合是较为稀疏的。

我们再来谈一下窗口 W 的具体形式。由于我们是借助 W 所覆盖的图像区域来分析 W 中心位置 \mathbf{x} 的特性的，可以合理地认为与 \mathbf{x} 距离越近的点对 \mathbf{x} 特性的影响越高，与 \mathbf{x} 离得较远的点，对 \mathbf{x} 特性的影响会小一些。因此，在算法实现中， W 通常被取为高斯窗口。 W 的大小是算法的超参数，需要用户预先设定。

为了计算式 4-4 所定义的矩阵 M ，需要近似计算图像函数 f 的偏导数。由于实际的图像为离散数字图像，我们只能用差分方法来近似计算图像函数的偏导数。对此部分内容不熟悉的读者，可参见附录 B。

为了使读者能够对哈里斯角点检测算法的处理流程能有一个整体上的认识，我们在图 4-5 中总结了该算法的关键处理步骤，并以可视化的形式给出了每一步所得到的处理结果。

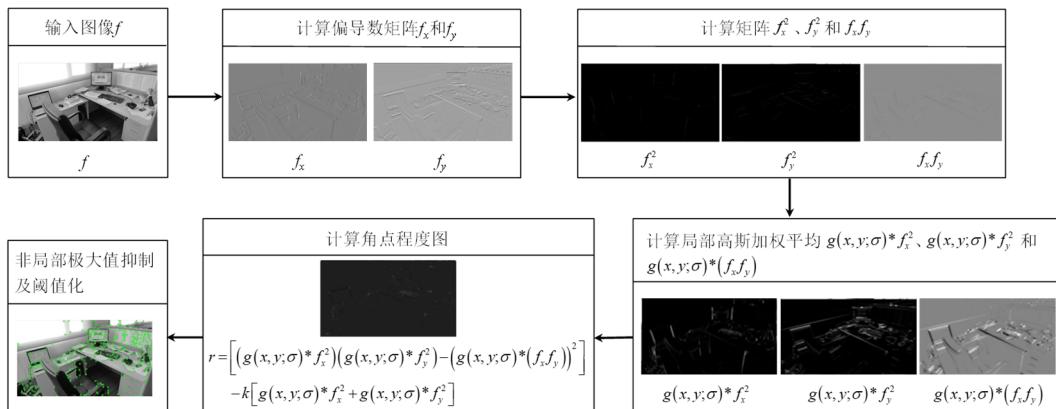


图 4-5：哈里斯角点检测算法处理流程。

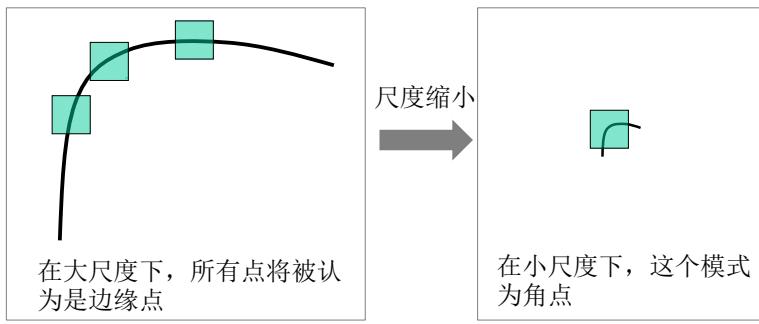


图 4-6: 哈里斯角点检测算法不具有尺度不变性。在这个理想模型中, 当分析窗口大小相同时, 在大尺度下, 所有的点都被认为是边缘点, 而在小尺度下, 该模式将被认为是角点。

接下来我们来分析一下哈里斯角点检测算法对光照变化和几何变换的不变性。图像上某点处的角点属性完全是由该点处的由式 4-4 所定义的矩阵 M 来决定的, 而 M 是由该点邻域中的一阶偏导数所决定的。根据导数的计算规则容易知道, 哈里斯角点检测算法对图像的整体光照变化具有不变性, 即当图像 $f(x,y)$ 变为 $f(x,y)+b$ (其中 b 为常数) 时, 每点处的角点程度值保持不变。但对除此之外的其他类型的光照变化, 哈里斯角点检测算法都不具有不变性。从哈里斯角点检测算法对角点的定义(图 4-1(c)), 不难理解, 从理论上来说该算法具有“旋转不变性”, 也就是说, 在图像发生了旋转变换的前后, 用相同的哈里斯角点检测程序可以(大致)检测出相同的特征点。但该算法不具有“尺度不变性”, 我们可以通过一个简单的理想模型来说明这个问题, 如图 4-6 所示。在图 4-6 中, 在尺度变换前后, 角点检测算法的分析窗口大小一致, 这就会导致在大尺度下所有的点都被认为是边缘点, 而在小尺度下, 该模式将被认为是角点。导致哈里斯角点检测算法不具有尺度不变性的根本原因就在于该算法中使用的分析窗口 W 的大小是预先设定的, 它没有一种自动化的、与尺度大小相适应的分析窗口大小设定机制。

4.1.3 哈里斯角点的特征描述子

特征点实际就是图像上的一个位置。为了后续应用, 比如要匹配不同图像中的特征点, 我们需要对特征点建立特征描述子以表达该特征点。对于一个给定特征点 \mathbf{x} 来说, 它的特征描述子 \mathbf{d} 是一个向量, \mathbf{d} 是基于 \mathbf{x} 的邻域图像信息构造出来的。

假设 \mathbf{x} 为一个哈里斯角点。为构造 \mathbf{x} 的特征向量 \mathbf{d} , 我们需要以 \mathbf{x} 为中心取一个大小为 $s \times s$ 的窗口 W , 然后基于 W 所覆盖的图像区域来构造 \mathbf{d} 。 s 的值是需要用户事先设定的。

最简单的特征描述子称为“块”(block)描述子, 它是直接把 $s \times s$ 的图像块 W 拉成一个列向量并进行单位化(即, 使得该向量的 l_2 -范数为 1), 以这个单位化之后的列向量作为 \mathbf{x} 的特征描述子 \mathbf{d} 。不难理解, “块”描述子不具有旋转不变性, 也不具有尺度不变性。

为了要对描述子进行匹配, 我们首先要知道如何计算两个描述子之间的距离。设 $\mathbf{d}_1 \in \mathbb{R}^{n \times 1}$ 、 $\mathbf{d}_2 \in \mathbb{R}^{n \times 1}$ 为两个块描述子。常用的计算 \mathbf{d}_1 、 \mathbf{d}_2 距离的方式包括平方差之和(Sum of Squared Differences, SSD) 距离、绝对差之和(Sum of Absolute Differences, SAD) 距离与

规范化互相关 (Normalized Cross Correlation, NCC) 距离。SSD 距离定义为,

$$SSD_{dist}(\mathbf{d}_1, \mathbf{d}_2) = \|\mathbf{d}_1 - \mathbf{d}_2\|_2^2 = \sum_{i=1}^n (d_1^i - d_2^i)^2 \quad (4-7)$$

其中, d_1^i 表示向量 \mathbf{d}_1 的第 i 个元素。SAD 距离定义为,

$$SAD_{dist}(\mathbf{d}_1, \mathbf{d}_2) = \sum_{i=1}^n |d_1^i - d_2^i| \quad (4-8)$$

规范化互相关距离定义为,

$$NCC_{dist}(\mathbf{d}_1, \mathbf{d}_2) = 1 - \frac{1}{n} \frac{(\mathbf{d}_1 - \mu(\mathbf{d}_1)) \cdot (\mathbf{d}_2 - \mu(\mathbf{d}_1))}{std(\mathbf{d}_1) std(\mathbf{d}_2)} \quad (4-9)$$

其中, $std(\cdot)$ 返回向量数据的标准差, $\mu(\cdot)$ 返回向量数据的均值, $\mathbf{d}_1 - \mu(\mathbf{d}_1)$ 这个操作指的是

向量 \mathbf{d}_1 中每一个元素都要减去标量 $\mu(\mathbf{d}_1)$ 。实际上在式 4-9 中, $\frac{1}{n} \frac{(\mathbf{d}_1 - \mu(\mathbf{d}_1)) \cdot (\mathbf{d}_2 - \mu(\mathbf{d}_1))}{std(\mathbf{d}_1) std(\mathbf{d}_2)}$

为 \mathbf{d}_1 与 \mathbf{d}_2 的皮尔逊线性相关系数, 其取值范围为 $[-1, 1]$, 因此 $NCC_{dist}(\mathbf{d}_1, \mathbf{d}_2)$ 的取值范围为

$[0, 2]$ 。也有另外一种方式来定义规范化互相关距离,

$$NCC_{dist}(\mathbf{d}_1, \mathbf{d}_2) = \arccos \left(\frac{1}{n} \frac{(\mathbf{d}_1 - \mu(\mathbf{d}_1)) \cdot (\mathbf{d}_2 - \mu(\mathbf{d}_1))}{std(\mathbf{d}_1) std(\mathbf{d}_2)} \right) \quad (4-10)$$

即为 \mathbf{d}_1 与 \mathbf{d}_2 两个向量之间线性相关系数的反余弦值, 因此其取值范围为 $[0, \pi]$ 。

对于一个哈里斯角点, 除了块描述子以外, 我们也可以为其构建更加“高端”的描述子, 比如 SIFT 描述子^[2]、SURF 描述子^[3]、KAZE 描述子^[4]、BRISK 描述子^[5]等, 但这些描述子都不是为哈里斯角点而专门设计的, 它们都配合了特征尺度选择机制。如果要将这些“高端”描述子配合哈里斯角点来使用, 只能假设一张图像上的所有哈里斯角点具有相同的、预先设定的特征尺度。我们将在 4.2 节中结合 SIFT 特征点检测, 详细介绍 SIFT 描述子, 在 4.3 节中结合 SURF 特征点检测, 详细介绍 SURF 描述子。在图 4-7 中, 我们通过一个具体例子展示了基于块描述子匹配的哈里斯角点匹配结果。在这个例子中, 利用块描述子, 大部分的角点对应关系都是正确的, 但也有一些对应关系是错误的, 这也说明块描述子对图像局部特征的刻画能力十分有限, 我们后面将要学习的几个精心设计的描述子的性能要远优于块描述子。

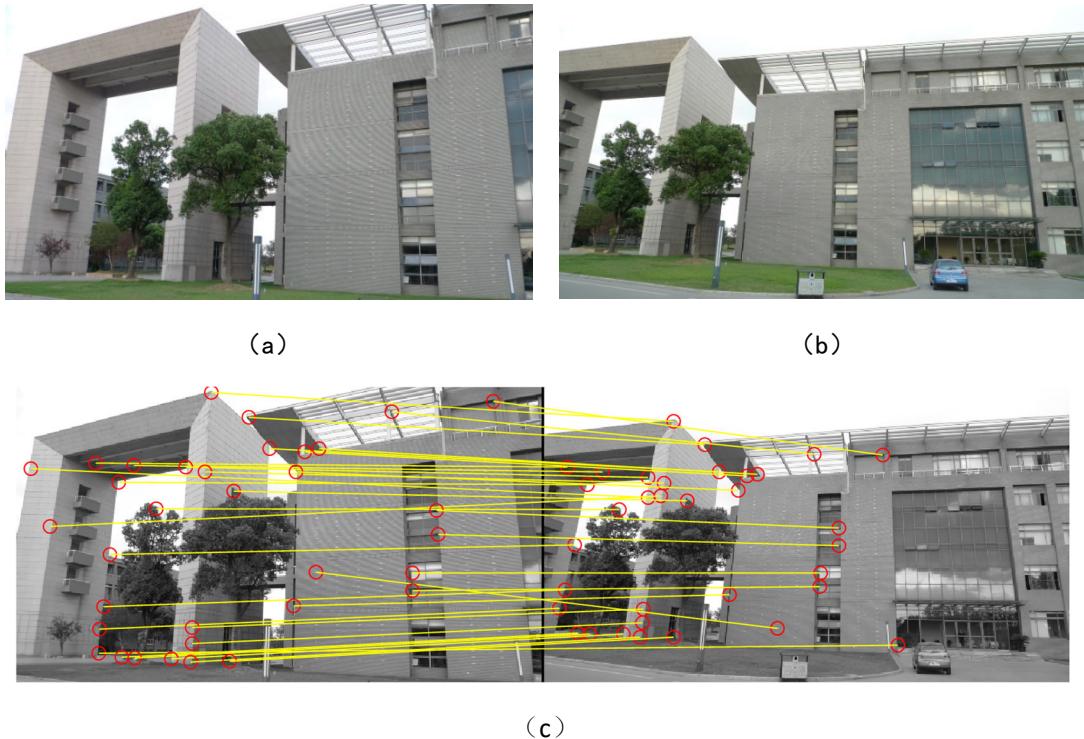


图 4-7: 哈里斯角点以及基于块特征的角点匹配。(a) 和 (b) 是两张输入图像; 对 (a) 和 (b) 分别进行角点检测, 并对每个角点提取块描述子, 之后基于描述子匹配结果建立起两张图像上角点之间的对应关系, 如图 (c) 所示。

4.2 SIFT 特征点及其特征描述子



图 4-8: 大卫·罗维 (David G. Lowe), 加拿大英属哥伦比亚大学计算机科学系教授。他于 1999 年发表 SIFT 算法, 是 SIFT 算法的创始人。他的研究方向主要是计算机视觉, 目标识别, 人类视觉的计算模型。

SIFT 的全称为尺度不变特征变换 (scale-invariant feature transform), 它实际上包含两部分, 尺度不变特征点的检测和尺度不变特征描述子的构建。SIFT 由加拿大英属哥伦比亚大学的 David Lowe 教授(图 4-8)提出, 其最初版本发表在 1999 年的 ICCV(International Conference on Computer Vision, 国际计算机视觉大会) 上^[6], 其完整版本发表在 2004 年的 IJCV

(International Journal of Computer Vision, 国际计算机视觉杂志) 上^[2]。SIFT 可以说是图像特征点检测与匹配领域中的里程碑式的工作, 它对后来许多优秀的同类算法都产生了很大影响。

接下来, 我们将在 4.2.1 节和 4.2.2 节中讲述 SIFT 框架下的特征点检测算法, 在 4.2.3 节中讲述 SIFT 框架下的尺度不变特征描述子的构造方法。

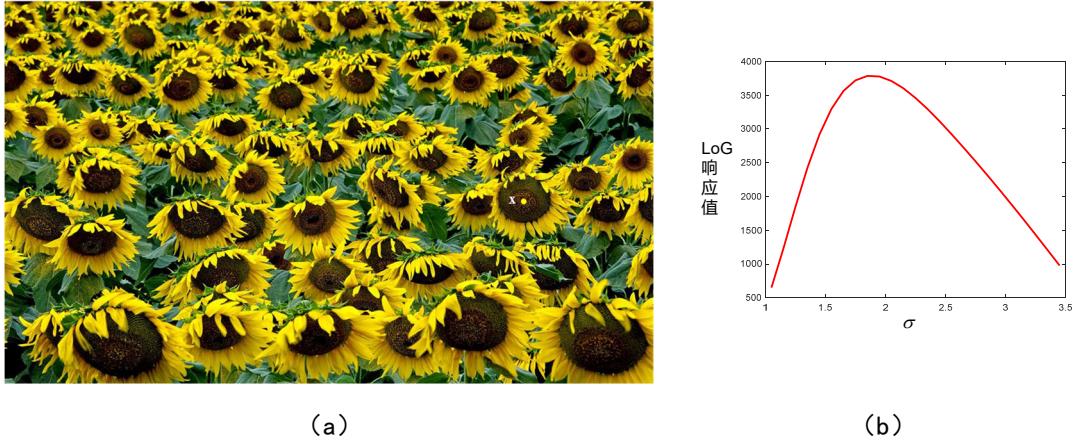


图 4-9: (a) 典型的斑点特征, 向日葵的中心位置是典型的图像斑点特征; (b) 对图像 (a) 用一系列不同尺度的尺度归一化 LoG 算子进行卷积, 点 x 处的响应值随 LoG 算子尺度变化的曲线; 可以看出, 该曲线为单峰曲线, 即它只有一个极大值点。

4.2.1 特征点检测基本思想

我们先来大致描述一下 SIFT 特征点检测算法的基本思想。在 SIFT 框架下, 特征点是一类被称之为“斑点 (blob)”的特殊的点。如图 4-9 中, 向日葵的中心点就是典型的斑点特征点。通过观察我们发现, 刻画一个斑点特征不单单需要知道它的中心位置, 还要知道它的空间大小。为了要检测斑点这种特殊的图像结构, David Lowe 使用了由瑞典学者 Tony Liderberg 首先提出的尺度归一化高斯-拉普拉斯 (scale-normalized Laplacian of Gaussian) 算子^[7], 简称为尺度归一化 LoG 算子。那这个尺度归一化 LoG 算子是什么样子的呢?

二维各向同性的高斯函数 $g(x,y):\mathbb{R}^2 \rightarrow \mathbb{R}$ 为,

$$g(x,y) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{x^2+y^2}{2\sigma^2}\right) \quad (4-11)$$

其参数 σ 称为高斯函数的尺度。函数的拉普拉斯算子为函数二阶偏导数之和, 因此高斯函数 $g(x,y)$ 的拉普拉斯 $\nabla^2 g$ 为,

$$\nabla^2 g = \frac{\partial^2 g}{\partial x^2} + \frac{\partial^2 g}{\partial y^2} = \frac{x^2 + y^2 - 2\sigma^2}{2\pi\sigma^6} e^{-\frac{x^2+y^2}{2\sigma^2}} \quad (4-12)$$

相应地, 尺度归一化 LoG 算子便是在 $\nabla^2 g$ 之前乘上 σ^2 , 即,

$$\sigma^2 \nabla^2 g = \frac{x^2 + y^2 - 2\sigma^2}{2\pi\sigma^4} e^{-\frac{x^2+y^2}{2\sigma^2}} \quad (4-13)$$

图 4-10 (a) 展示了 $\sigma^2 \nabla^2 g$ 算子的空间几何形状；不难理解， $\sigma^2 \nabla^2 g$ 算子非常适合于检测图像中圆盘状的斑点结构。可以看到，尺度归一化 LoG 算子 $\sigma^2 \nabla^2 g$ 有一个控制其尺度大小的参数 σ ，通过改变 σ 的值为 $\sigma_1, \sigma_2, \dots, \sigma_n$ ，可以得到一系列不同尺度的尺度归一化 LoG 算子， $\{\sigma_i^2 \nabla^2 g(\sigma_i)\}_{i=1}^n$ 。

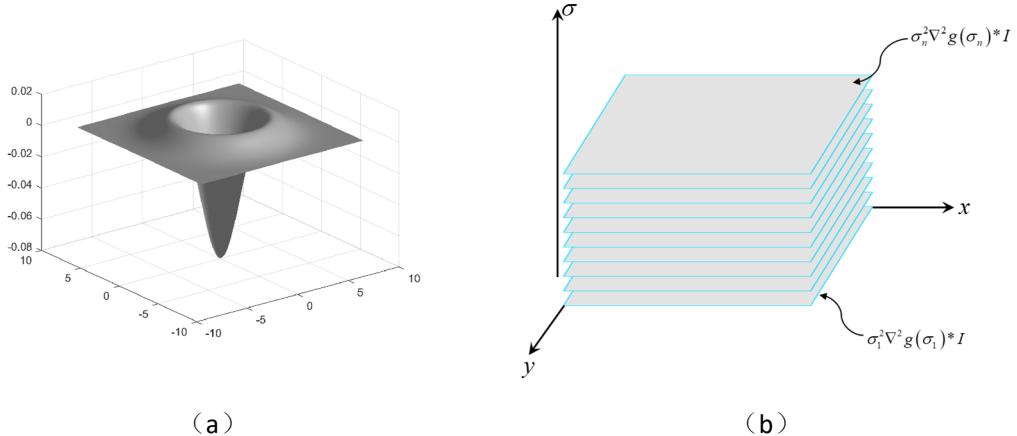


图 4-10：(a) $\sigma^2 \nabla^2 g$ 算子的三维形状，一系列不同尺度的 $\sigma^2 \nabla^2 g$ 算子可以用来检测图像中的斑点结构；(b) 图像 I 的尺度归一化 LoG 尺度空间。

斑点结构检测问题本质上是一种基于模板匹配思想的图像模式检测问题。在“绝对清晰”图像 I 中²，假设 \mathbf{x} 为斑点特征的中心位置。当我们用一系列尺度递增的尺度归一化 LoG 算子 $\{\sigma_i^2 \nabla^2 g(\sigma_i)\}_{i=1}^n$ （其中， $\sigma_{i+1} > \sigma_i$ ）和 I 进行卷积后，就会得到 I 的尺度归一化 LoG 尺度空间，如图 4-10 (b) 所示。在该尺度空间中，与 \mathbf{x} 对应位置处会有一组响应值 $\{r_i\}_{i=1}^n$ 。 r_i 可以看作是关于 σ_i 的函数，而且该函数曲线只有一个峰（谷）值，即斑点结构中心 \mathbf{x} 处关于 σ 的尺度归一化 LoG 响应值曲线只有一个极值点（如图 4-9 (b) 所示）。如果 $\sigma_k^2 \nabla^2 g(\sigma_k)$ 的形状是最接近于该斑点结构的，那么 r_k 将是响应值中的极值点，我们把相应的 σ_k 称作这个斑点结构的**特征尺度**。特征尺度实际上反映了斑点结构的空间大小。另外，如果 \mathbf{x} 是一个斑点结

² 这里假设 I 是“绝对清晰图像”，是为了叙述方便但又保持严谨；在这个条件下，如果极值点出现在了尺度空间的第 k 层，即该层为 $\sigma_k^2 \nabla^2 g(\sigma_k)^* I$ ，我们就说该极值点的特征尺度为 σ_k 。但实际上，“绝对清晰”的图像是不存在的，这个问题我们会在 4.2.2 节中再具体解决。

构的中心位置，它的近邻 \mathbf{x}' 处的关于 $\{\sigma_i^2 \nabla^2 g(\sigma_i)\}_{i=1}^n$ 的响应值的极值会不如 r_k 显著，所以我们可以通过邻域内响应值的比较来得到稀疏的合理的特征点。



(a)



(b)

图 4-11：特征尺度的尺度协变性。在 (a) 中， \mathbf{x} 点为斑点特征点，利用尺度归一化 LoG 算子确定出其特征尺度为 σ_1 ，(a) 中白色圆圈的半径即为 σ_1 。把图像 (a) 放大 2 倍得到图像 (b)，当然，为了方便比较，(b) 实际上只显示了 (a) 图放大两倍结果的一部分。在 (b) 中，与 \mathbf{x} 对应的点为 \mathbf{y} ，显然 \mathbf{y} 是 (b) 中的斑点特征点，且利用尺度归一化 LoG 算子确定出的 \mathbf{y} 的特征尺度为 σ_2 ，(b) 中白色圆圈的半径即为 σ_2 。我们会发现 σ_1 与 σ_2 之间恰好满足关系 $\sigma_2 = 2\sigma_1$ 。

我们再对特征尺度的性质进一步明确一下。通过上述方式确定出的斑点结构的特征尺度具有**尺度协变性**。假设在图像 I_1 中，点 \mathbf{x} 为斑点结构中心，利用尺度归一化 LoG 算子确定出其特征尺度为 σ_1 。把图像 I_1 放大 s 倍得到图像 I_2 ，点 \mathbf{y} 为 \mathbf{x} 的对应点，显然 \mathbf{y} 所在的斑点结构的空间大小应该是 \mathbf{x} 所在的斑点结构空间大小的 s 倍。假设我们利用尺度归一化 LoG 算子确定出了 \mathbf{y} 所在的斑点结构的特征尺度为 σ_2 。我们会发现 σ_2 与 σ_1 之间恰好会满足 $\sigma_2 = s\sigma_1$ 。特征尺度的这种性质便称为**尺度协变性**，在图 4-11 中我们通过一个具体的例子对该性质进行了说明。特征尺度的尺度协变性是非常重要的一个性质，这意味着我们可以**基于特征尺度来构建尺度不变的特征点描述子**。

需要强调一下，我们对以 \mathbf{x} 为中心的斑点结构的特征尺度的确定，是以寻找 \mathbf{x} 位置处 $\{\sigma_i^2 \nabla^2 g(\sigma_i)\}_{i=1}^n$ 响应值的极值点的方式来完成的。“极值点”意味着该极值有可能是极大值，也有可能是极小值，不难理解，到底是极大值还是极小值要取决于斑点结构的特点：如果它是“中部暗、周边亮”的结构，那么上述极值就是极大值，反之如果它是“中间亮、周边暗”的结构，上述极值就是极小值。

斑点结构是通过一组尺度归一化 LoG 算子检测出来的，由于 LoG 算子是各向同性的算子，因此相应的斑点结构检测算法显然具有旋转不变性。另外，斑点特征点是以在尺度归一化 LoG 尺度空间中寻找极值点的方式来确定的，容易理解，这种特征点检测方式会对图像的光照变化具有很强的鲁棒性。

本小节简要描述了 SIFT 框架下特征点检测算法设计的基本思想。在 4.2.2 中，我们将描

述该算法实现的细节。

4.2.2 特征点检测算法实现

1) 用 DoG 对尺度归一化 LoG 进行近似

在具体实现过程中, David Lowe 建议可用高斯差分 (Difference of Gaussians, DoG) 算子来近似代替尺度归一化 LoG 算子 $\sigma^2 \nabla^2 g$, 这会使得尺度不变的特征点检测在实现上更加简洁和高效。顾名思义, DoG 算子是由两个不同尺度的二维高斯函数相减得到的。可以证明, 当 $k \rightarrow 1$ 时, DoG 算子 $DoG(\sigma) \triangleq g(x, y; k\sigma) - g(x, y; \sigma)$ 与 LoG 算子 $\sigma^2 \nabla^2 g$ 之间有如下关系,

$$DoG(\sigma) \approx (k-1)\sigma^2 \nabla^2 g(x, y; \sigma) \quad (4-14)$$

我们来简要证明一下式 4-14。根据高斯函数 $g(x, y; \sigma)$ 的定义式 4-11 可知,

$$\frac{\partial g}{\partial \sigma} = \frac{(x^2 + y^2 - 2\sigma^2)}{2\pi\sigma^5} e^{-\frac{(x^2+y^2)}{2\sigma^2}} = \sigma \nabla^2 g \quad (4-15)$$

而当 $k \rightarrow 1$ 时,

$$\frac{\partial g}{\partial \sigma} \approx \frac{g(x, y; k\sigma) - g(x, y; \sigma)}{k\sigma - \sigma} = \frac{DoG(\sigma)}{(k-1)\sigma} \quad (4-16)$$

结合式 4-16 和式 4-15 可知,

$$DoG(\sigma) = (k-1)\sigma \frac{\partial g}{\partial \sigma} = (k-1)\sigma (\sigma \nabla^2 g(x, y; \sigma)) = (k-1)\sigma^2 \nabla^2 g(x, y; \sigma) \quad (4-17)$$

证毕。虽然从理论上来说, 只有当 $k \rightarrow 1$ 时, 式 4-14 才能成立, 但实践表明, 即使 k 明显比 1 大, DoG 算子检测尺度不变特征点的性能也不会受到明显影响^[2]。

利用 DoG 算子, 在图像尺度归一化 LoG 尺度空间中的特征点检测问题就转换成了在 DoG 尺度空间中的特征点检测问题。根据卷积运算的性质可知, 图像 I 在高斯差分算子 $DoG(\sigma)$ 下的卷积响应输出 $DoG(\sigma)*I$ 就是 $g(x, y; k\sigma)*I - g(x, y; \sigma)*I$ 。因此, I 的 DoG 尺度空间中的尺度层 $DoG(\sigma)*I$ 可以通过 I 的高斯响应差分 $g(x, y; k\sigma)*I - g(x, y; \sigma)*I$ 来得到。

2) 在高斯差分尺度空间中特征点的检测

假设有图像 I 。如图 4-11 所示, D_1 、 D_2 和 D_3 分别为图像 I 高斯差分尺度空间中的相邻三层, 且它们的尺度分别为 σ_1 、 σ_2 和 σ_3 。如果 p 点处的值是 DoG 尺度空间中的一个局部极值, 即 $D_2(p)$ 比 p 在尺度空间中的 26 个近邻的值都大 (或者都小), 点 p 所对应的图像空间中的像素坐标 (x, y) 就被认为是候选特征点, 该特征点的特征尺度为 σ_2 。

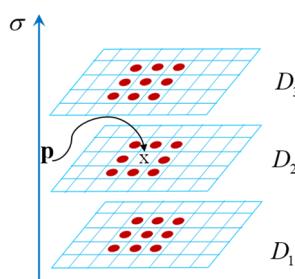


图 4-12: SIFT 中的特征点被定义为 DoG 尺度空间中的局部极值点。 D_1 、 D_2 、 D_3 为 DoG 尺度空间中具有相邻尺度的三层，如果 p 点的值 $D_2(p)$ 比它的 26 个邻居都大或者都小，则 p 点所对应的图像空间中的像素位置就被称为是候选特征点，同时，此特征点的特征尺度就是 σ_2 。

3) 高斯差分尺度空间的构造

给定图像 I ，我们现在的任务是要构建 I 的 DoG 尺度空间。由于 I 的高斯差分被定义为 $g(x, y; k\sigma)*I - g(x, y; \sigma)*I$ ，因此构建 I 的 DoG 尺度空间的前提是要构建 I 的高斯尺度空间。从理论上来说，图像 I 的高斯尺度空间是用一系列具有由小到大变化的标准差的高斯函数同 I 所对应的“清晰场景”进行卷积得到，每一个高斯函数 $g(x, y; \sigma_i)$ 同 I 所对应的“清晰场景”进行卷积之后得到高斯尺度空间中的一层，该层的尺度即为 σ_i 。

然而与 I 所对应的“清晰场景”是无法准确获得的，这是因为在拍摄实际图像时，镜头不可避免地对“清晰场景”进行了低通滤波，即输入图像 I “自带”一个较小的尺度 σ_{init} ，换句话说就是 I 是“清晰场景”与 $g(x, y; \sigma_{init})$ 进行卷积之后的结果。David Lowe 建议 σ_{init} 的值可以设为 0.5。我们对 I 施加标准差为 $\sqrt{\sigma^2 - \sigma_{init}^2}$ 的高斯滤波，得到的结果 $g(x, y; \sqrt{\sigma^2 - \sigma_{init}^2})*I$ 便是高斯尺度空间中尺度为 σ 的图像。这里我们用到了高斯函数卷积运算的一个性质：对图像 I 先后进行尺度为 σ_1 和 σ_2 的两次高斯卷积的结果，等于对原始图像 I 进行尺度为 $\sqrt{\sigma_1^2 + \sigma_2^2}$ 的高斯卷积的结果，即 $g(x, y; \sigma_2)*(g(x, y; \sigma_1)*I) = g(x, y; \sqrt{\sigma_1^2 + \sigma_2^2})*I$ （见附录 C）

为了使 DoG 尺度空间中每层的 DoG 算子 $DoG(\sigma)$ 与其对应的尺度归一化 LoG 算子 $\sigma^2 \nabla^2 g(\sigma)$ 之间保持恒定的倍数关系 $DoG(\sigma) \approx (k-1)\sigma^2 \nabla^2 g(\sigma)$ ，在构建高斯尺度空间时，相邻两层高斯函数的尺度（即标准差）之间，即 σ_{i+1} 与 σ_i 之间，要满足关系 $\sigma_{i+1}/\sigma_i = k$ 。

为了计算效率的提升， I 的高斯尺度空间可以分为不同的组（octave）。第 $o+1$ 组的第 0 层³ 的尺度是第 o 组第 0 层尺度的 2 倍。每一组又可以分为 s 个间隔，这样显然有 $k = 2^{\frac{1}{s}}$ 。由于第 $o+1$ 组的第 0 层的尺度是第 o 组第 0 层尺度的 2 倍，我们可以把第 $o+1$ 组的第 0 层的图像分辨率降为第 o 组图像分辨率的 $\frac{1}{2}$ 而不会损失任何信息⁴，而之后第 $o+1$ 组的其他层都是在该组第 0 层的基础上通过累积高斯卷积得到。每组内各个尺度层的图像分辨率相同。这样，每一组的图像空间分辨率都是上一组的一半。那么，为了进行基于 DoG 尺度空间的特征点检测，图像 I 的高斯尺度空间每组只有 s 层是否是合理的呢？答案是否定的！为了使 I 的高斯尺度空间能够满足特征点检测需求，我们还需要一些精巧的设计。

³ 为了和本节配套学习的 C++ 代码相容，这节对尺度空间中组和层的计数是从 0 开始的。

⁴ 当高斯函数与图像进行卷积时，高斯函数便成为了低通滤波器，其低通范围与其标准差成反比。设第 o 组第 0 层图像为 I_1 ，第 $o+1$ 组第 0 层图像为 I_2 。由于 I_2 的尺度为 I_1 的两倍，则相对于由 I_1 所定义的频率范围来说， I_2 的有效频率范围为 I_1 频率范围的 $1/2$ 。对 I_2 进行“隔二取一”的降采样操作得到 I_3 ，这相当于在频域中只保留了 I_2 频率范围的一半（低频部分），但 I_2 的有效信息恰恰都在低频这一半，因此 I_3 与 I_2 相比并没有信息损失。对此内容的更深入理解，需要读者学习与信号的傅里叶分析有关的内容。

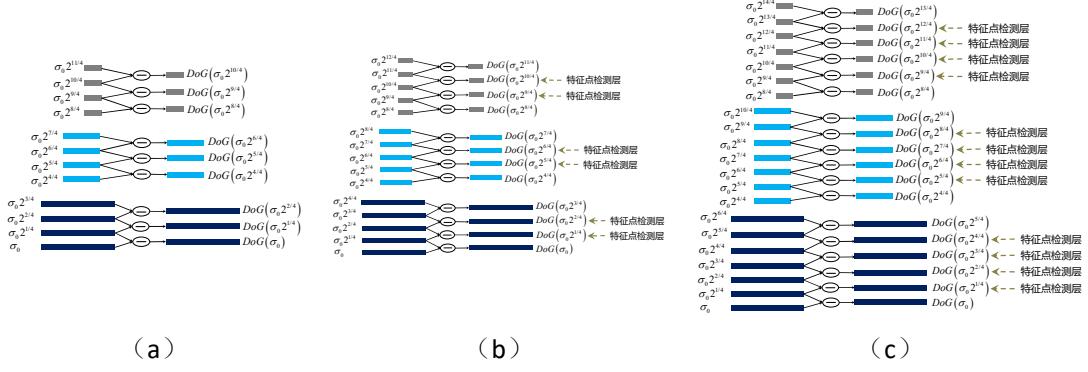


图 4-13：满足 DoG 尺度空间中特征点检测需求的高斯尺度空间的构造。(a) 按此方式构建的高斯尺度空间，会使得相应的 DoG 尺度空间在尺度维度上不是连续等间隔采样的；(b) 按此方式构建的高斯尺度空间，会使得能够进行特征点检测的 DoG 尺度层在尺度维度上不是连续等间隔采样的；(c) 满足 DoG 尺度空间中特征点检测需求的正确的高斯尺度空间构造方式。本图中，在每层侧方都标示出了该层尺度，但所有尺度值都是相对于初始第 0 层图像的分辨率来定义的。

为了便于读者理解，我们假定要构造的高斯尺度空间包括 3 组，每组的尺度分隔为 4，即 $s=4$ ，初始尺度为 σ_0 。对图像高斯尺度空间的尺度维度进行离散采样且要保证相邻两层之间的尺度关系满足 $\sigma_{i+1}/\sigma_i=k$ 的话，高斯尺度空间中的尺度层次设计会如图 4-13 (a) 所示。要注意到，特征点检测是在 DoG 尺度空间中进行的，而图 4-13 (a) 所示的高斯尺度空间尺度采样层所形成的 DoG 尺度空间是不“连续”的，根本原因在于高斯尺度空间的相邻两组之间会有个下采样操作，这使得跨组的两个相邻尺度层之间无法完成相减操作，导致 DoG 尺度空间中缺失了 $DoG(\sigma_0 2^{3/4})$ 和 $DoG(\sigma_0 2^{7/4})$ 这两层。沿着这个思路不难想象，要使形成的 DoG 尺度空间是连续等间隔采样的，我们就需要在高斯尺度空间的每组顶上额外再加上一层，这便形成了如图 4-12 (b) 所示的高斯尺度空间采样层设计方案；该方案所相应形成的 DoG 尺度空间便是连续等间隔采样的了。但这依然不能满足特征点检测的要求，这是因为如前所述，特征点的检测需要在相邻的三个 DoG 层之间才能进行。在图 4-12 (b) 中，我们标记出了能够进行特征点检测操作的层；显然，能够进行特征点检测的层的尺度不是连续等间隔的，缺失了 $DoG(\sigma_0 2^{3/4})$ 、 $DoG(\sigma_0 2^{4/4})$ 、 $DoG(\sigma_0 2^{7/4})$ 、 $DoG(\sigma_0 2^{8/4})$ 等。为了使能够进行特征点检测的 DoG 层在尺度空间上是等间隔连续的，我们需要在高斯尺度空间中每组的顶部继续增加额外的尺度层，如图 4-12 (c) 所示。在图 4-12 (c) 中，最终能够进行特征点检测的 DoG 层在尺度维度上是连续等间隔采样的，这才是满足特征点检测要求的尺度空间尺度层级设计方案。在该方案中，高斯尺度空间尺度层的设计具有如下特性：1) 每组中尺度层的数目为 $s+3$ ；2) $o+1$ 组的第 0 层图像由第 o 组中倒数第 3 层图像进行分辨率减半的下采样操作得到；3) 相邻两个尺度层之间的尺度关系满足 $\sigma_{i+1}/\sigma_i=k=2^{1/s}$ 。

接下来我们讲述在实现高斯尺度空间时会遇到的一些细节问题。

➤ 问题 1：初始尺度 σ_0 设为多少合适？

根据 David Lowe 的建议， σ_0 可以取为 1.6。

➤ 问题 2：假设输入图像为 I ，那么与 σ_0 对应的初始层的图像是什么？是 $g(x, y; \sigma_0)^* I$

吗？

答案是否定的。为了能更加充分地利用图像信息、有效提升检测到的尺度不变特征点的数量，David Lowe 建议要对 I 进行 2 倍上采样，即通过图像插值的办法构造出空间分辨率与 I 的 2 倍的图像 I_{us} 。如前所述，图像 I 也“自带”一个较小的高斯尺度 σ_{init} ，那么，由 I 进行 2 倍上采样得到的 I_{us} 的“自带”尺度就是 $2\sigma_{init}$ 。我们对 I_{us} 施加标准差为 $\sqrt{\sigma_0^2 - (2\sigma_{init})^2}$ 的高斯滤波，得到的结果 $I_0 \triangleq g(x, y; \sqrt{\sigma_0^2 - (2\sigma_{init})^2}) * I_{us}$ 便是尺度为 σ_0 的高斯尺度空间的初始图像。

➤ 问题 3：高斯尺度空间的组数如何确定？

由于在高斯尺度空间中，每组的空间分辨率都为其上一组的 $1/2$ ，因此组数的确定必然和输入图像 I 的空间分辨率有关。一个常用的确定高斯尺度空间组数的经验公式^[8]为，

$$octNum = \frac{\log(\min(w, h))}{\log 2} - 2 \quad (4-18)$$

其中 w 和 h 分别为 I 的宽度和高度。比如，如果输入图像 I 的像素分辨率为 256×256 ，按照式 4-18， I 的高斯尺度空间会有 6 组，它们的空间分辨率分别为 512×512 、 256×256 、 128×128 、 64×64 、 32×32 和 16×16 。

➤ 问题 4：如何计算得到高斯尺度空间中每一层的图像？

基于高斯函数卷积运算的性质，不难理解，高斯尺度空间的构造可以迭代进行：设第 $i+1$ 层的尺度为 σ_{i+1} 、第 i 层的尺度为 σ_i ，我们只需要对第 i 层的图像 I_i 施加尺度为 $\sqrt{\sigma_{i+1}^2 - \sigma_i^2}$ 的高斯卷积，便可以得到第 $i+1$ 层的图像 I_{i+1} ，即 $I_{i+1} = g(x, y; \sqrt{\sigma_{i+1}^2 - \sigma_i^2}) * I_i$ ，这种实现方式会使得在构造高斯尺度空间时所使用的高斯卷积核都比较小。还有一个实现方面的细节：第 $o+1$ 组的基准尺度（该组内第 0 层的尺度）与第 o 组的基准尺度之间正好是 2 倍的关系，但这两个基准尺度所对应的高斯函数的标准差确是一样的（而不是 2 倍的关系），这是因为第 $o+1$ 组的空间分辨率恰好是第 o 组空间分辨率的一半。我们举个具体的例子来说明一下这个问题。如图 4-13 (c) 所示，第 0 组的基准尺度为 σ_0 ，第 1 组的基准尺度为第 0 组基准尺度的两倍，即 $2\sigma_0$ 。在计算第 0 组第 1 层图像时，需要对该组第 0 层图像施加的高斯卷积的标准差为 $\sqrt{(\sigma_0 2^{1/4})^2 - \sigma_0^2}$ 。由于第 1 组的空间分辨率为第 0 组的一半，相对于第 1 组的空间分辨率来说，第 1 组的基准尺度所对应的高斯标准差为 σ_0 ；同样地，相对于第 1 组的空间分辨率来说，第一组第 1 层尺度所对应的高斯标准差为 $\sigma_0 2^{5/4}/2 = \sigma_0 2^{1/4}$ 。因此，在计算第 1 组第 1 层图像时，需要对该组第 0 层图像施加的高斯卷积的标准差也为 $\sqrt{(\sigma_0 2^{1/4})^2 - \sigma_0^2}$ 。

4) 粗略极值点位置的精化以及对低对比度极值点的剔除

我们可把图像 I 在 DoG 尺度空间中的响应看作一个三元函数 $f(x, y, l) : \mathbb{R}^3 \rightarrow \mathbb{R}$ ， (x, y) 为 DoG 尺度空间中某一层图像上的位置， l 为该层在 DoG 尺度空间中的尺度序号（假设 DoG 尺度空间中的尺度层从 0 开始统一编号），该层的尺度为 $\sigma_0 2^{l/s}$ 。在 DoG 尺度空间中，假设

通过比较某点的值与其周围 26 个邻居值大小关系的方式，我们初步确定出点 $\mathbf{x}_0=(x_0, y_0, l_0)^T$ 为一个局部极值点。由于我们构造的尺度空间是离散的， \mathbf{x}_0 的坐标一定都是整数。但在真实连续的 DoG 尺度空间中，点 \mathbf{x}_0 很可能并不是准确的尺度空间局部极值点。我们可以对初始取得的“粗略”极值点位置 \mathbf{x}_0 进行“精化”，以得到更加精准的极值点位置。我们可以通过一个简单的例子来理解一下粗略极值点位置精化的思想。在图 4-14 中，假设连续可微函数 $f(x)$ 的一个真正的局部极值点为 x^* 。我们只有 $f(x)$ 的在离散采样点 x_1, x_2, x_3, \dots 处的数据。通过比较离散采样点处的数据，可知 x_2 可能为 $f(x)$ 一个局部极值点，但实际上它并不是 $f(x)$ 真正精确的极值点。基于离散采样数据，利用位置精化操作，我们希望能找到比 x_2 更接近于真实极值点 x^* 的点。

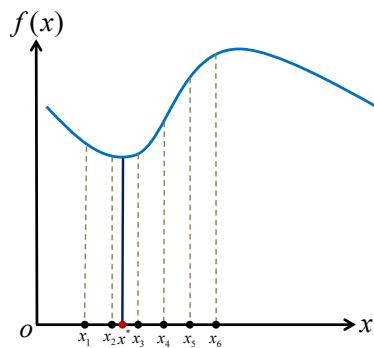


图 4-14：离散极值点的位置精化操作示意图。假设对于连续可微函数 $f(x)$ ，我们只有它在离散采样点 x_1, x_2, x_3, \dots 处的数据。利用精化操作，可以从已有离散采样数据中插值出比 x_2 更接近于真实局部极值点 x^* 的点。

对粗略极值点 \mathbf{x}_0 进行的精化是通过不断尝试迭代来完成的。对 $f(x, y, l)$ 在点 $\mathbf{x}_0=(x_0, y_0, l_0)^T$ 近旁进行二阶泰勒展开，

$$f(\mathbf{x}_0 + \Delta\mathbf{x}) \approx f(\mathbf{x}_0) + (\nabla f(\mathbf{x}_0))^T \Delta\mathbf{x} + \frac{1}{2} (\Delta\mathbf{x})^T (\nabla^2 f(\mathbf{x}_0)) \Delta\mathbf{x} \quad (4-19)$$

其中 $\Delta\mathbf{x} \triangleq (\Delta x, \Delta y, \Delta l)^T$ ， $\nabla f(\mathbf{x}_0)$ 表示在点 \mathbf{x}_0 处函数 $f(\mathbf{x})$ 的梯度， $\nabla^2 f(\mathbf{x}_0)$ 表示在点 \mathbf{x}_0 处函数 $f(\mathbf{x})$ 的海森矩阵。式 4-19 中的 $f(\mathbf{x}_0 + \Delta\mathbf{x})$ 是关于 $\Delta\mathbf{x}$ 的函数，我们现在要求出它的极值点 $\Delta\mathbf{x}^*$ 。

如果 $\Delta\mathbf{x}^* = \mathbf{0}$ ，那说明 \mathbf{x}_0 本身就是极值点，否则真正的极值点可能应该更接近于 $\mathbf{x}_0 + \Delta\mathbf{x}^*$ 。由于式 4-19 是关于 $\Delta\mathbf{x}$ 的二次函数，要找到它的极值点 $\Delta\mathbf{x}^*$ 只需要找到它关于 $\Delta\mathbf{x}$ 的驻点即可，即求方程 $\frac{df(\mathbf{x}_0 + \Delta\mathbf{x})}{d\Delta\mathbf{x}} = \mathbf{0}$ 的解 $\Delta\mathbf{x}^*$ ，容易知道，

$$\Delta\mathbf{x}^* = -(\nabla^2 f(\mathbf{x}_0))^{-1} \nabla f(\mathbf{x}_0) \quad (4-20)$$

当然，我们需要保证 $\nabla^2 f(\mathbf{x}_0)$ 可逆，式 4-20 才会成立。如果 $\mathbf{x}_0 + \Delta\mathbf{x}^*$ 更接近于另外一个整数位置点 \mathbf{x}_1 （此时的 $\Delta\mathbf{x}^*$ 至少有一个维度上的值大于 0.5），说明真正的极值点应该更接近于 \mathbf{x}_1 ，

则我们需要把 \mathbf{x}_0 更新为 \mathbf{x}_1 , $\mathbf{x}_0 := \mathbf{x}_1$, 然后继续按照上述方式在 \mathbf{x}_0 点进行粗略极值点位置的精化; 否则的话, 与 \mathbf{x}_0 对应的精确极值点的位置被估计为,

$$\hat{\mathbf{x}}_0 := \mathbf{x}_0 + \Delta\mathbf{x}^* \quad (4-21)$$

同时更进一步, 我们可以根据式 4-19 估计出 DoG 尺度空间中点 $\hat{\mathbf{x}}_0$ 处的值 $f(\hat{\mathbf{x}}_0)$,

$$\begin{aligned} f(\hat{\mathbf{x}}_0) &= f(\mathbf{x}_0 + \Delta\mathbf{x}^*) \approx f(\mathbf{x}_0) + (\nabla f(\mathbf{x}_0))^T \Delta\mathbf{x}^* + \frac{1}{2} (\Delta\mathbf{x}^*)^T (\nabla^2 f(\mathbf{x}_0)) \Delta\mathbf{x}^* \\ &= f(\mathbf{x}_0) + (\nabla f(\mathbf{x}_0))^T \left(-(\nabla^2 f(\mathbf{x}_0))^{-1} \nabla f(\mathbf{x}_0) \right) + \frac{1}{2} \left(-(\nabla^2 f(\mathbf{x}_0))^{-1} \nabla f(\mathbf{x}_0) \right)^T (\nabla^2 f(\mathbf{x}_0)) \left(-(\nabla^2 f(\mathbf{x}_0))^{-1} \nabla f(\mathbf{x}_0) \right) \\ &= f(\mathbf{x}_0) - \frac{1}{2} (\nabla f(\mathbf{x}_0))^T (\nabla^2 f(\mathbf{x}_0))^{-1} \nabla f(\mathbf{x}_0) \\ &= f(\mathbf{x}_0) + \frac{1}{2} (\nabla f(\mathbf{x}_0))^T \Delta\mathbf{x}^* \end{aligned} \quad (4-22)$$

如果 $|f(\hat{\mathbf{x}}_0)|$ 的值太小, 则说明点 $\hat{\mathbf{x}}_0$ 并不是一个稳定的特征点, 需要被剔除掉。需要格外注意,

式 4-21 和式 4-22 中的 \mathbf{x}_0 很有可能并不是那个初始给定的粗略极值点, 而是最后一步精化迭代步骤中的整数位置点。另外, 粗略极值点位置精化操作只能限制在组内进行, 也就是说如果初始给定的粗略极值点 \mathbf{x}_0 在第 o 组, 那么最终精化后的点 $\hat{\mathbf{x}}_0$ 也需要被限制在第 o 组之内。

算法 4-1 给出了对 DoG 尺度空间中粗略极值点 \mathbf{x}_0 进行位置精化操作的处理流程伪码。

算法 4-1: 粗略极值点 \mathbf{x}_0 的位置精化

```

 $K = 5$ 
 $iter = 0$ 
 $\Delta\mathbf{x}^* = -(\nabla^2 f(\mathbf{x}_0))^{-1} \nabla f(\mathbf{x}_0)$ 
while  $iter < K$ 
    if any dimension of  $\Delta\mathbf{x}^*$  is smaller than 0.5
        break
    end
     $\mathbf{x}_0 := round(\mathbf{x}_0 + \Delta\mathbf{x}^*)$ 
     $\Delta\mathbf{x}^* = -(\nabla^2 f(\mathbf{x}_0))^{-1} \nabla f(\mathbf{x}_0)$ 
     $iter++$ 
end
if  $iter \geq K$  //说明算法没有收敛, 精化操作失败
    return false
end
 $\hat{\mathbf{x}}_0 := \mathbf{x}_0 + \Delta\mathbf{x}^*$ 
 $f(\hat{\mathbf{x}}_0) = f(\mathbf{x}_0) + \frac{1}{2} (\nabla f(\mathbf{x}_0))^T \Delta\mathbf{x}^*$ 

```

5) 对不稳定的边缘响应点的剔除

假设 \mathbf{x}_0 是 DoG 尺度空间中的一个初始给定的粗略极值点, 它被精化之后的位置为 $\hat{\mathbf{x}}_0$ 。

如前所述, 若 $|f(\hat{\mathbf{x}}_0)|$ 的值太小, $\hat{\mathbf{x}}_0$ 将不会被认为是一个有效特征点。若 $|f(\hat{\mathbf{x}}_0)|$ 的值足够大,

我们还需要对 $\hat{\mathbf{x}}_0$ 进行进一步检查, 看看它是否是图像边缘点。对于一个图像边缘点, 我们很

难获得其在空间上的精确位置，因此 David Lowe 建议要尽可能剔除掉位于图像边缘结构上的候选特征点。如果 $\hat{\mathbf{x}}_0$ 确实是图像边缘点，它将被剔除，否则 $\hat{\mathbf{x}}_0$ 最终被确认为一个有效特征点。那么，有什么办法可以判定 $\hat{\mathbf{x}}_0$ 是否位于图像边缘上呢？我们首先说明一点，由于 $\hat{\mathbf{x}}_0$ 是从初始粗略极值点经过迭代精化之后得到的，其坐标很有可能不是整数，这会给边缘点判别算法的实现带来很大困难。一个可行的解决方案是：将 $\hat{\mathbf{x}}_0$ 是否为边缘点的判定问题近似等价为 $\mathbf{x}_r = \text{round}(\hat{\mathbf{x}}_0)$ 是否为边缘点的判定问题。

我们现在的任务是要判定 DoG 尺度空间中的一个极值点⁵ $\mathbf{x}_r = (x_r, y_r, l_r)$ 是否位于图像边缘上，其中 \mathbf{x}_r 的坐标分量皆为整数。我们把 \mathbf{x}_r 所在的 DoG 尺度空间层记为函数 $f(x, y)$ ，则 $(x_r, y_r, f(x_r, y_r))$ 会形成一个三维曲面。如果点 (x_r, y_r) 位于图像边缘上，不难理解，曲面上点 $(x_r, y_r, f(x_r, y_r))$ 处的两个主曲率（曲面的主曲率的定义见附录 D）的绝对值一定会相差很大，依据图像边缘点的这个特性我们便可以对位于边缘上的特征点进行甄别。

令 K_{\max} 表示 $(x_r, y_r, f(x_r, y_r))$ 处的两个主曲率中绝对值较大的一个，令 K_{\min} 表示另一个绝对值较小的主曲率。设 $H \in \mathbb{R}^{2 \times 2}$ 为 $f(x, y)$ 在点 (x_r, y_r) 处的海森矩阵。根据附录 D 可知，

$$\begin{cases} \text{tr}(H) = K_{\max} + K_{\min} \\ \det(H) = K_{\max} K_{\min} \end{cases} \quad (4-23)$$

令 $r = \frac{K_{\max}}{K_{\min}}$ ，根据上面的条件容易知道， $r \geq 1$ 。此时我们有，

$$\frac{(\text{tr}(H))^2}{\det(H)} = \frac{(K_{\max} + K_{\min})^2}{K_{\max} K_{\min}} = \frac{(rK_{\min} + K_{\min})^2}{rK_{\min} \cdot K_{\min}} = \frac{(r+1)^2}{r} \quad (4-24)$$

容易验证，在 $r=1$ 时，式 4-24 会取得最小值；当 $r > 1$ 时，随着 r 的增大， $\frac{(\text{tr}(H))^2}{\det(H)}$ 的值会单调递增。而 r 越大就意味着点 (x_r, y_r) 越像是一个图像边缘点。按照 David Lowe 建议， r 的阈

值可以设置为 10，因此只要 $\frac{(\text{tr}(H))^2}{\det(H)} > \frac{(10+1)^2}{10}$ ，我们便认为 (x_r, y_r) 位于图像边缘结构上，即点 \mathbf{x}_r 是图像边缘点，近似地， $\hat{\mathbf{x}}_0$ 便被认为是图像边缘点。

⁵ “极值点”这个条件意味着下文中的函数 $f(x, y)$ 在 (x_r, y_r) 处的梯度为 0，且其海森矩阵半正定，即其海森矩阵的两个特征值不可能是异号的，证明见附录 H 中的定理 H.1 和 H.2。

6) $\hat{\mathbf{x}}_0$ 的空间位置与特征尺度的最终估计

假设 $\hat{\mathbf{x}}_0 = (\hat{x}_0, \hat{y}_0, \hat{l}_0)^T$ 为 DoG 尺度空间中经过精化后的特征点且已满足判定条件，即该点

处的 DoG 值幅度较大且该点非图像边缘点。需要注意到， $\hat{\mathbf{x}}_0$ 的空间位置 (\hat{x}_0, \hat{y}_0) 是定义在它所在的组上的，而我们需要知道该点在原始输入图像 I 中的位置。设 $\hat{\mathbf{x}}_0$ 所在组的序号为 $octIndex$ ($octIndex$ 的计数从 0 开始，第 0 组的图像空间分辨率为输入图像 I 的 2 倍)，则 $\hat{\mathbf{x}}_0$ 相对于原始输入图像 I 的空间位置 (\hat{x}_0, \hat{y}_0) 为，

$$\begin{cases} \hat{x}'_0 = \hat{x}_0 \cdot 2^{octIndex-1} \\ \hat{y}'_0 = \hat{y}_0 \cdot 2^{octIndex-1} \end{cases} \quad (4-25)$$

假设整个 DoG 尺度空间中的尺度层从序号 0 开始统一编号，如图 4-13 (c) 所示，相对于初始图像 I_0 来说（分辨率为 I 的 2 倍），序号为 l 的 DoG 层的尺度为 $\sigma_0 \cdot 2^{l/s}$ 。类似地， $\hat{\mathbf{x}}_0$ 在 DoG 尺度空间中的层序号为 \hat{l}_0 （注意， \hat{l}_0 为精化操作之后的层序号，它不一定是整数），因此相对于输入图像 I 的分辨率来说，它的特征尺度 $\hat{\sigma}_0$ 可以被估计为，

$$\hat{\sigma}_0 = \sigma_0 \cdot 2^{\frac{\hat{l}_0}{s}} / 2 \quad (4-26)$$

总结下来，对于一个有效的 SIFT 尺度不变特征点来说，它由三部分信息组成（在后续操作中我们会用到这三类信息）：相对于输入图像的空间位置 (\hat{x}_0, \hat{y}_0) ，相对于输入图像空间分辨率的特征尺度 $\hat{\sigma}_0$ ，在 DoG 尺度空间中与它离得最近的整数尺度层的层号，即 $round(\hat{l}_0)$ 。

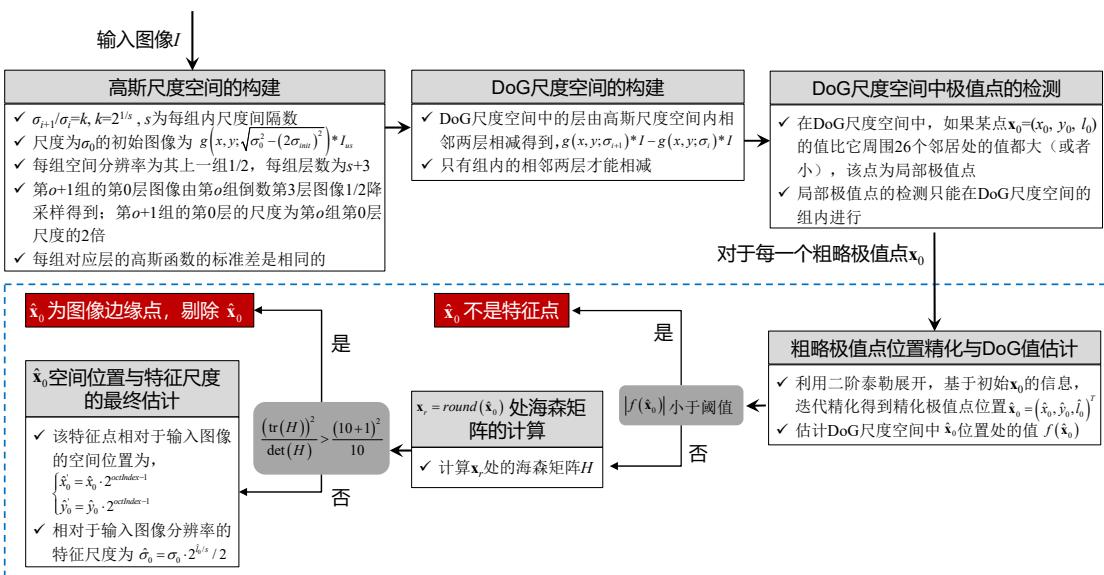


图 4-15：SIFT 框架中尺度不变的特征点检测算法整体流程。

在本小节中，我们详细讲述了 SIFT 框架下尺度不变特征点检测算法的实现细节。为了使读者能从整体上把握该算法的结构和处理流程，我们将该算法的主要步骤总结在了图 4-15 中。在图 4-16 中，我们通过一个具体的例子展示了 SIFT 特征点检测算法的输出结果。在该图中，每个圆圈代表了一个 SIFT 特征点，圆圈的中心为该特征点在原始输入图像空间中的位置，其半径为该特征点的特征尺度。



图 4-16：(a) 和 (b) 分别是图 4-7 中 (a) 和 (b) 两张图像的 SIFT 特征点检测结果。每个圆圈的中心为该特征点在原始图像空间中的位置，圆圈的半径为该特征点的特征尺度。

4.2.3 描述子构造

在 4.2.2 节中，我们学习了 SIFT 框架下的特征点检测算法。假设 I 为输入图像， $\mathbf{x}=(x, y, \sigma, l)$ 为 I 上的一个 SIFT 特征点，其中 (x, y) 为该点在图像 I 上的位置， σ 为该点相对于 I 空间分辨率的特征尺度， l 为 DoG 尺度空间中离该点最近的尺度层的序号 (l 为整数)。基于这些信息，本节的任务是要构建特征点 \mathbf{x} 的尺度不变的特征描述子向量。

1) 用于构建描述子的图像邻域的确定

为了提升计算效率， \mathbf{x} 描述子的构造可在 I 的高斯尺度空间中对应的尺度层上来进行。特征点 \mathbf{x} 在 DoG 尺度空间中的层序号为 l ，与此 DoG 层尺度最接近的高斯尺度空间中的尺度层的序号也为 l ，我们把该高斯尺度空间层记为 g_l 。 \mathbf{x} 的描述子的构建便在 g_l 上进行。

假设 g_l 所在的组序号为 $octIndex$ ，则与图像 I 上 (x, y) 点对应的 g_l 上的点的位置为 $\mathbf{x}_d = (x/2^{octIndex-1}, y/2^{octIndex-1})$ 。为了使构造的特征描述子具有尺度不变性，构造描述子的图像块的大小显然需要具有尺度协变性。我们已经知道 \mathbf{x} 的特征尺度为 σ ，但这个“ σ ”的值是相对于 I 的空间分辨率来定义的。在 g_l 上，与特征尺度 σ 所对应的高斯标准差的值应该为 $\sigma_l = \sigma/2^{octIndex-1}$ 。在后面的步骤中，不论是确定局部主方向还是构建描述子，我们在 g_l 上选定 \mathbf{x}_d 周围的邻域范围时，都是以 σ_l 作为基准。

2) 邻域主方向的确定

为了使构造的描述子具有旋转不变性，我们需要确定出 \mathbf{x}_d 周围局部邻域的主方向 θ ，之后 \mathbf{x}_d 周围用于计算描述子的邻域点都绕 \mathbf{x}_d 旋转 $-\theta$ ，这便实现了方向的归一化。在图 4-17 中，我们通过一个示例来进一步说明一下方向归一化操作的目的。在图 4-17 (a) 和 (b) 中， \mathbf{p}_1 和 \mathbf{p}_2 是两个对应的特征点，它们的邻域内容只是“相差了”一个旋转变换。不难理解，我们希望构造出的 \mathbf{p}_1 和 \mathbf{p}_2 的描述子应该是相同的，也就是说，描述子的构造算法要具有旋转

不变性。为了满足这个要求，在构造描述子之前，先要进行方向归一化。在图 4-17 中，红色箭头标识出了根据特征点局部邻域内点的梯度方向估计出的主方向。之后，要对 \mathbf{p}_1 和 \mathbf{p}_2 的邻域内的点进行旋转，以使得各自的主方向与 X 轴重合。旋转之后的 \mathbf{p}_1 和 \mathbf{p}_2 的邻域分别显示在了图 4-17 (a) 和 (b) 的右下角。可以看出，经过方向归一化操作后， \mathbf{p}_1 和 \mathbf{p}_2 的局部邻域内容完全相同。后续的描述子构建操作将会基于旋转之后的图像邻域来进行。

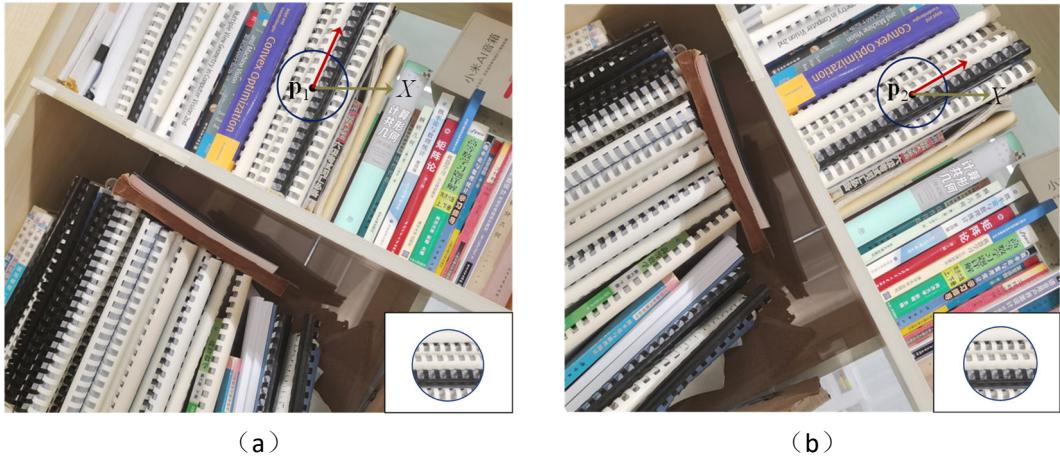


图 4-17：特征点的局部邻域主方向。 \mathbf{p}_1 和 \mathbf{p}_2 是两个对应的特征点，它们的邻域内容只是“相差了”一个旋转变换。红色箭头标识出了根据特征点局部邻域内点的梯度方向估计出的主方向。进行了方向归一化操作之后的 \mathbf{p}_1 和 \mathbf{p}_2 的邻域分别显示在了 (a) 和 (b) 的右下角。

根据 David Lowe 的建议，为了确定主方向，需要在 \mathbf{g}_l 上 \mathbf{x}_d 点周围划定一个 $9\sigma_l \times 9\sigma_l$ 的区域范围，把该图像区域记为 P 。那么，如何找到 P 的主方向呢？我们需要借助于 P 的梯度方向直方图。

对 P 中的每一点 \mathbf{p}_i ，计算出它的梯度模 m_i 和梯度方向 $\alpha_i \in [0, 2\pi]$ 。梯度方向直方图 $oriHist$ 包含 n （在实现中， n 一般取为 36）个小仓（bin），小仓 k ($0 \leq k < n$) 覆盖角度范围 $\left[\frac{2\pi}{n}k, \frac{2\pi}{n}(k+1)\right)$ 。为了构建 $oriHist$ ，需要遍历 P 中所有的点：对于点 \mathbf{p}_i ，若该点处的梯度方向 α_i 满足 $\frac{2\pi}{n}k \leq \alpha_i < \frac{2\pi}{n}(k+1)$ ，则，

$$oriHist[k] := oriHist[k] + \omega_i \cdot m_i \quad (4-27)$$

其中， ω_i 为按照点 \mathbf{p}_i 到 \mathbf{x}_d 距离设定的高斯权重，该高斯函数的标准差为 $1.5\sigma_l$ 。初步得到 $oriHist$ 后，需要对 $oriHist$ 进行 2 次局部平滑以增强其稳定性，局部平滑窗口权重为 $[0.25, 0.5, 0.25]$ 。按此局部平滑策略，在一次平滑操作之后， $oriHist[k]$ 被更新为，

$$oriHist[k] := 0.25 \times oriHist[k-1] + 0.50 \times oriHist[k] + 0.25 \times oriHist[k+1] \quad (4-28)$$

假设 $oriHist$ 的峰值出现在小仓 k (k 当然为整数)。与 4.2.2 节中从粗略极值点位置精化出准确极值点位置所用的思想类似，我们可根据 $oriHist[k-1]$ 、 $oriHist[k]$ 和 $oriHist[k+1]$ 的值，估计出更加准确的峰值位置，所用理论工具还是函数的二阶泰勒展开。不难验证，整数峰值位置 k 所对应的准确的峰值位置 k^* 可被估计为（具体证明过程作为练习，请读者完成），

$$k^* = k + \frac{oriHist[k-1] - oriHist[k+1]}{2(oriHist[k-1] + oriHist[k+1] - 2oriHist[k])} \quad (4-29)$$

则最终与此 $oriHist$ 峰值位置 k^* 所对应的主方向角度可插值为 $\frac{2\pi}{n}k^*$, 它便是特征点 \mathbf{x} 所在局部区域的主方向。

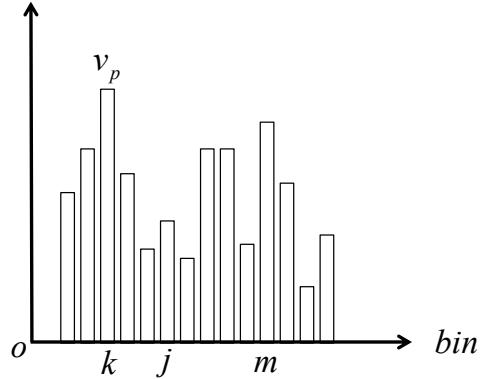


图 4-18: 特征点有两个局部主方向的情况。小仓 k 的值为 v_p , 是全局峰值, 显然小仓 k 可确定一个主方向。小仓 m 处是一个局部峰值, 且其值大于 $v_p \times 80\%$, 我们认为小仓 m 也可确定一个主方向。小仓 j 处虽然也取得了局部峰值, 但它的值小于 $v_p \times 80\%$, 因此它不能确定一个主方向。

在实际情况中, 当 P 的图像结构比较复杂时, 它可能会有多个主方向。为了后续特征匹配操作的稳定性, 我们可以按如下方式处理。设 $oriHist$ 的最高峰值为 v_p 。若小仓 m 的值也是 $oriHist$ 的一个局部极大峰值且 $oriHist[m] > v_p \times 80\%$ (如图 4-18 所示), 我们也会按照同样的方式基于 $oriHist[m-1]$ 、 $oriHist[m]$ 和 $oriHist[m+1]$, 估计出小仓 m 所代表的主方向角度值 o_m 。之后, 我们把特征点 \mathbf{x} 的信息“复制”一份为 \mathbf{x}_c , 并把 o_m 作为 \mathbf{x}_c 的主方向。在后续的所有处理中, \mathbf{x} 和 \mathbf{x}_c 将被当做两个完全独立的特征点。换句话说, 在图像位置 (x, y) 处, 有两个特征点 \mathbf{x} 和 \mathbf{x}_c , 它们的位置相同、特征尺度相同, 唯一的不同之处就是它们的局部主方向不同。当然, 点 \mathbf{x} 处的主方向也许会多于两个, 对每一个主方向都按照上述方式处理即可。

3) 特征描述子的构建

假设按照上述方式确定出特征点 \mathbf{x} 的主方向为 θ , 接下来我们来看看如何根据 \mathbf{x}_d 的邻域信息来构建 \mathbf{x} 的描述子。

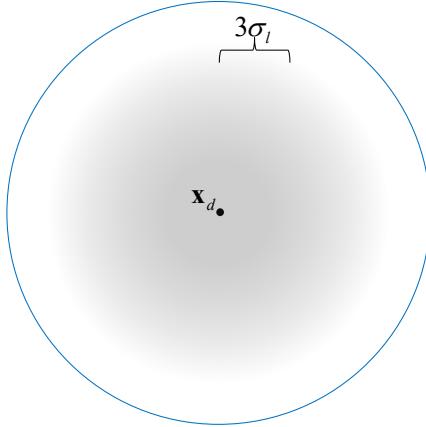


图 4-19: SIFT 描述子的构建。以 \mathbf{x}_d 为圆心，在 \mathbf{x}_d 周围取大小为 $12\sigma_l \times 12\sigma_l$ 的方形区域，将该区域记为 Q 。把 Q 划分为 4×4 共 16 个子区域，从每个子区域构建 8 维梯度方向直方图。最终，将所有子直方图连接起来形成一个 128 维的向量，该向量便是特征点 \mathbf{x} 的 SIFT 尺度不变描述子。在构建直方图时，某点对直方图的贡献要进行基于该点到 \mathbf{x}_d 距离的高斯加权。

在 g_l 上，以 \mathbf{x}_d 为圆心，在 \mathbf{x}_d 周围取半径为 $7.5\sigma_l \cdot \sqrt{2}$ 的区域，将该区域中的每个点绕 \mathbf{x}_d 旋转 $-\theta^6$ ，以使得后续构建出的描述子具有旋转不变性。之后，以 \mathbf{x}_d 为圆心，在 \mathbf{x}_d 周围取大小为 $12\sigma_l \times 12\sigma_l$ 的方形区域，将该区域记为 Q 。如图 4-19 所示，把 Q 划分为 4×4 共 16 个子区域，每个子区域记作 $Q_i (i=1, \dots, 16)$ 。从每个 Q_i 中构建一个 8 维的梯度方向直方图 $hist_i$ ，显然直方图的每个小仓覆盖的角度范围为 $2\pi/8 = \pi/4$ 。最终，将 $\{hist_i\}_{i=1}^{16}$ 连接起来形成一个 128 维的向量 $hist$ ， $hist$ 便是特征点 \mathbf{x} 的 SIFT 尺度不变描述子。在基于 Q 构建描述子 $hist$ 的过程中，David Lowe 给出了一些实现上需要注意的细节和技巧，以使得构建出的描述子更加稳定可靠。比如，在构建 $hist_i$ 的时候，某点对 $hist_i$ 的贡献要进行基于该点到 \mathbf{x}_d 距离的高斯加权，以弱化离 \mathbf{x}_d 较远的点对 \mathbf{x}_d 描述子的影响。另外，每个点的贡献需要按照距离依照比例“线性分配”到与它最近邻的子区域的直方图的相关小仓上去，这样 Q 中每一个点实际上会影响到 8 个小仓的值⁷。但这些细节过于琐碎，本书就不再详加阐述了。感兴趣的读者可参阅与本章配套学习的代码。

得到了 128 维的描述子向量 $hist$ 以后，还要对它进行一些后处理操作。首先，要对 $hist$ 向量进行单位化得到 $hist_n$ ，以使得描述子向量能够达到对光照反射变化的不变性。然后，对 $hist_n$ 中过大的小仓值进行限定，以使得描述子向量能够对一定程度的更加广泛的非线性光照变化具有鲁棒性；David Lowe 建议，可把 $hist_n$ 中值大于 0.2 的小仓的值限定为 0.2。假设上一步骤处理之后的特征描述子向量为 $hist_{nr}$ ，最终，再对最终 $hist_{nr}$ 进行一次单位化操作，得

⁶ 从概念上来理解，我们需要对局部图像绕 \mathbf{x}_d 旋转 $-\theta$ 。但在实际编程实现时，我们并不需要真的对局部图像块进行旋转得到一个新的图像块，而只需要把点的坐标重新计算以确定它落在 16 个子区域中的哪一个当中，并把每点处的梯度方向加上 $-\theta$ 。

⁷ 沿行方向，它会影响行方向相邻 2 个子区域的直方图；沿列方向，它会影响列方向相邻 2 个子区域的直方图；对于每个子区域的直方图，它会影响与它方向最近的 2 个小仓；因此，总共会影响 $2 \times 2 \times 2 = 8$ 个小仓的值。

到最终的单位化特征描述子向量。

SIFT 特征描述子具有很好的对特征点的描述性，同时从理论上来说又具有旋转不变性和尺度不变性；另外，它对环境的光照条件变化也具有很强的鲁棒性。

在图 4-20 中，我们通过一个具体的例子展示了 SIFT 特征点的匹配结果。图 4-20 中的两幅输入图像来自图 4-7 中的 (a) 和 (b)。在对它们进行了 SIFT 特征点检测、描述子构建以及特征点匹配之后，特征点之间的对应关系如图 4-20 所示。

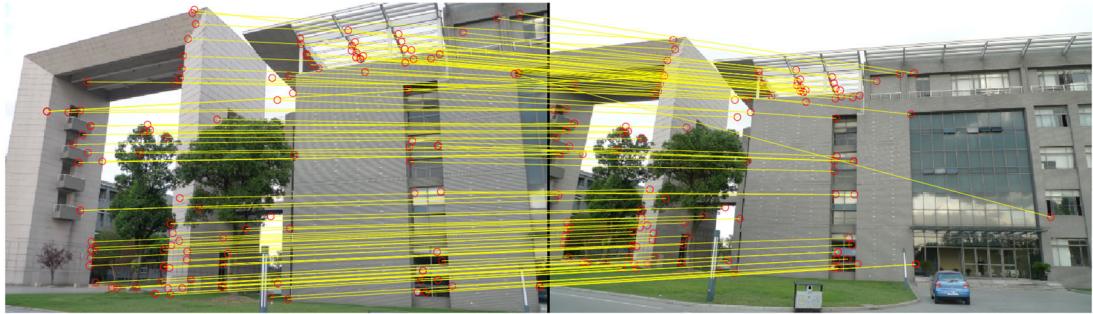


图 4-20：SIFT 特征点以及基于 SIFT 特征描述子的特征点匹配。两幅输入图像为图 4-7 中的 (a) 和 (b)。在对它们进行了 SIFT 特征点检测、描述子构建以及特征点匹配之后，特征点之间的对应关系如本图所示。

4.3 SURF 特征点及其特征描述子

4.4 ORB 特征点及其特征描述子

4.3 特征点匹配

假设有两幅图像 I_1 和 I_2 。 I_1 上的特征点集合为 $\{\mathbf{x}_i\}_{i=1}^m$ ，对应的特征描述子集合为 $\mathcal{P}=\{\mathbf{d}_i\}_{i=1}^m$ 。 I_2 上的特征点集合为 $\{\mathbf{y}_j\}_{j=1}^n$ ，对应的特征描述子集合为 $\mathcal{Q}=\{\mathbf{e}_j\}_{j=1}^n$ ，且 $\mathbf{d}_i (i=1, \dots, n)$ 与 $\mathbf{e}_j (j=1, \dots, m)$ 为同类型的特征描述子。我们现在的目标是要进行特征点匹配，也就是要建立起点集 $\{\mathbf{x}_i\}_{i=1}^m$ 和 $\{\mathbf{y}_j\}_{j=1}^n$ 之间的对应关系，这需要借助于对比它们的特征描述子集合来完成。若点 \mathbf{x}_i 与点 \mathbf{y}_j 的特征描述子 \mathbf{d}_i 与 \mathbf{e}_j 满足以下三个条件，我们则认为 \mathbf{x}_i 与 \mathbf{y}_j 为一

对匹配的特征点：

1) \mathbf{d}_i 与 \mathbf{e}_j 之间的距离要小于某个预设阈值 t_1

\mathbf{d}_i 与 \mathbf{e}_j 之间的距离可以按照本章 4.1.3 节中介绍的特征描述子距离的某种定义方式来计算， t_1 为预设阈值。

2) \mathbf{d}_i 与 \mathbf{e}_j 满足“双向”确认准则

\mathbf{e}_j 是特征描述子集合 \mathcal{Q} 的所有元素中，与 \mathbf{d}_i 距离最小的元素； \mathbf{d}_i 是特征描述子集合 \mathcal{P} 的所有元素中，与 \mathbf{e}_j 距离最小的元素。

3) \mathbf{d}_i 与 \mathbf{e}_j 的匹配无歧义

设 $d_1 = \text{dist}(\mathbf{d}_i, \mathbf{e}_j)$ 。若 \mathbf{e}_k 是集合 \mathcal{Q} 中除了 \mathbf{e}_j 之外，与 \mathbf{d}_i 的距离最近的，且它们之间的距离为 $d_2 = \text{dist}(\mathbf{d}_i, \mathbf{e}_k)$ 。若，

$$d_1 / d_2 < t_2 \quad (4-8)$$

其中 t_2 为预先设定的参数，则认为 \mathbf{d}_i 与 \mathbf{e}_j 的匹配无歧义。容易理解，“匹配无歧义”这条准则想表达的含义是正确匹配时的距离要比错误匹配的距离小很多。这条准则由 David Lowe 提出^[2]。

我们最后再来谈一下给定了 \mathbf{e}_j ，如何能从集合 \mathcal{P} 中找出与 \mathbf{e}_j 距离最近的描述子元素。最简单、最容易理解的方法就是穷举法，即要遍历整个集合 \mathcal{P} ，计算 \mathcal{P} 中每一个元素与 \mathbf{e}_j 的距离，然后从中挑出与 \mathbf{e}_j 最近的集合 \mathcal{P} 中的元素。这种穷举法适合于集合 \mathcal{P} 的规模不大且需要在线动态生成的情况。如果集合 \mathcal{P} 规模较大且可以以离线方式提前构建，那么我们可以用一些精巧的索引结构来存储 \mathcal{P} ，从而可有效提升从 \mathcal{P} 中寻找与 \mathbf{e}_j 最近邻元素的计算效率。在特征点匹配领域，常用的特征描述子索引数据结构有 kd-树^[9]、spill-树^[10]等。由于这部分内容隶属于图像检索领域，本书就不再对这些数据结构的构建和查找操作的细节详加展开了，感兴趣的读者可以参见本章参考文献 11。

4.4 习题

(1) 请证明按照式 4-4 的方式所构造的矩阵 M 必为半正定矩阵。

(2) 对于实际图像来说，除非是极特殊情况，一般情况下式 4-4 中的 M 是正定矩阵。请证

明，如果 $M \in \mathbb{R}^{2 \times 2}$ 为正定矩阵，则满足方程 $(\Delta x, \Delta y) M \begin{pmatrix} \Delta x \\ \Delta y \end{pmatrix} = 1$ 的点 $(\Delta x, \Delta y)$ 所形成的轨

迹为椭圆。更进一步，设 M 的两个特征值为 λ_1, λ_2 且 $\lambda_1 \geq \lambda_2$ ，则上述椭圆的长半轴长度

为 $(\lambda_2)^{-1/2}$ ，其短半轴长度为 $(\lambda_1)^{-1/2}$ 。

- (3) 证明式 4-29。假设 $oriHist$ 为一直方图，其峰值出现在小仓 k (k 当然为整数)。我们可根据 $oriHist[k-1]$ 、 $oriHist[k]$ 和 $oriHist[k+1]$ 的值，估计出更加准确的峰值位置 k^* ，

$$k^* = k + \frac{oriHist[k-1] - oriHist[k+1]}{2(oriHist[k-1] + oriHist[k+1] - 2oriHist[k])}.$$

- (4) 运行并理解与本章配套的 Matlab 哈里斯角点检测程序“`harrisCornerDetector`”。把示例程序中的图像替换成你自己的一张图像，再运行角点检测程序，看看结果如何。尝试改变一下该程序的超参数设置，看看会对角点检测结果产生什么影响。
- (5) 运行并理解与本章配套的 Matlab 哈里斯角点检测与匹配程序“`harrisCornerDescriptorMatching`”。该示例程序实现了哈里斯角点检测、块描述子构造以及基于块描述子的角点匹配，默认设置下运行会生成图 4-7 的角点检测及匹配结果。把示例程序中的图像替换成你自己的两张图像，再运行角点检测与匹配程序，看看结果如何。
- (6) 运行并理解与本章配套的 C++ 程序“`openSIFTVS`”。该程序实现了 SIFT 特征点检测、描述子构建以及描述子匹配等功能。建议读者认真学习此示例程序，它可帮助读者深刻理解 SIFT 算法设计原理。在运行该程序之前，请读者先学习该示例程序附带的文档“`opensift` 编译指南”。该文档可指导读者在 Windows+Visual Studio 的开发环境下（作者所用的具体开发环境为 Win11+VS2017）正确配置和部署运行该程序所需的软件环境。正确部署并运行后，该程序可输出如下图所示的 SIFT 特征点匹配结果，



参考文献

- [1] Harris, C., and M. Stephens, “A combined corner and edge detector,” *Proceedings of the 4th Alvey Vision Conference*, August 1988, pp. 147-151.

- [2] D.G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int'l J. Comput. Vis.*, vol. 60, pp. 91-110, 2004.
- [3] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool, "SURF: Speeded up robust features", *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346–359, 2008.
- [4] P. F. Alcantarilla, A. Bartoli, and A. J. Davison, "KAZE features," *Proc. Euro. Conf. Comput. Vis.*, pp. 214-227, 2012.
- [5] S. Leutenegger, M. Chli, and R. Siegwart, "BRISK: Binary robust invariant scalable keypoints," *Proc. IEEE International Conference on Computer Vision*, 2011.
- [6] David G. Lowe, "Object recognition from local scale-invariant features," *Proc. International Conference on Computer Vision*, pp. 1150–1157, 1999.
- [7] T. Lindeberg, "Scale-space theory: A basic tool for analysing structures at different scales", *J. Applied Statistics*, vol. 21, no. 2, pp. 224-270, 1994.
- [8] OPENSIFT, An open-source SIFT library, <http://robwhess.github.io/opensift/>
- [9] J.L. Bentley, Multidimensional binary search trees used for associative searching, *Communications of the ACM*, 18 (9), 1975.
- [10] T. Liu, A. Moore, A. Gray et al., An investigation of practical approximate nearest neighbor algorithms, in *Proc. NIPS*, 2004.
- [11] 王永明, 王贵锦, 图像局部不变性特征与描述, 国防工业出版社, 2010 年。

第 5 章 线性最小二乘问题

在第 4 章中，通过描述子匹配，我们已经得到了图像 I_1 和 I_2 中特征点对应点对关系集合 $\mathcal{S} = \{\mathbf{x}_i \leftrightarrow \mathbf{x}'_i\}_{i=1}^p$ ，其中 \mathbf{x}_i 是来自 I_1 的特征点， \mathbf{x}'_i 是来自 I_2 的特征点， $\mathbf{x}_i \leftrightarrow \mathbf{x}'_i$ 表示 \mathbf{x}_i 与 \mathbf{x}'_i 是一对对应的特征点， p 为 I_1 和 I_2 中具有对应关系的特征点对的个数。根据全景拼接问题的描述，图像 I_1 和 I_2 的所有对应点可以通过同一个线性几何变换 H 关联起来。在没有任何其他先验知识的情况下，我们把 H 考虑成平面之间最具有普适性的线性变换，射影变换，则 $\forall \mathbf{x}_i \leftrightarrow \mathbf{x}'_i \in \mathcal{S}$ ，

$$c\mathbf{x}'_i = H_{3 \times 3}\mathbf{x}_i \quad (5-1)$$

其中 $c \neq 0$ 是一个与 \mathbf{x}'_i 有关的常数， $H = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix}$ 是一个 3×3 的表达平面上射影变换的 8

自由度矩阵。由于 \mathbf{x}_i 、 \mathbf{x}'_i 都是从图像上检测到的特征点，因此它们都是平面上的正常点（非无穷远点），因此可假定式 5-1 中的 \mathbf{x}_i 、 \mathbf{x}'_i 都是规范化齐次坐标形式（如果不是，则可以先转化为规范化齐次坐标形式），即 $\mathbf{x}_i = (x_i, y_i, 1)^T$ ， $\mathbf{x}'_i = (x'_i, y'_i, 1)^T$ 。则式 5-1 可进一步变为，

$$c \begin{bmatrix} x'_i \\ y'_i \\ 1 \end{bmatrix} = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix} \begin{bmatrix} x_i \\ y_i \\ 1 \end{bmatrix} \quad (5-2)$$

这样，从每一个点对关系 $\mathbf{x}_i \leftrightarrow \mathbf{x}'_i$ 中都可以得到一个形如式 5-2 的关于 H 的等式。那么，点对关系集合 \mathcal{S} 中共有 p 个元素，因此可以得到 p 个形如式 5-2 的等式。接下去的任务就是要从这 p 个等式中解出 H 。根据具体处理方式的不同，这个问题可以建模为齐次线性最小二乘问题或者非齐次线性最小二乘问题。接下去我们就对这两个问题详加阐述。

在本章后面的推导过程中，会遇到函数或自变量的表达中包含矩阵或向量的求导问题，如果读者对这些内容不是很熟悉的话，请参见本书附录 C。

5.1 齐次线性最小二乘问题

5.1.1 问题定义

给定一个点对关系 $\mathbf{x}_i \leftrightarrow \mathbf{x}'_i$ ，我们得到了一个形如式 5-2 的方程。对这个方程左右展开得到，

$$\begin{cases} h_{11}x_i + h_{12}y_i + h_{13} = cx_i \\ h_{21}x_i + h_{22}y_i + h_{23} = cy_i \\ h_{31}x_i + h_{32}y_i + h_{33} = c \end{cases} \quad (5-3)$$

将式 5-3 中第一式和第三式的左右两边相除、第二式和第三式的左右两边相除，得到，

$$\begin{cases} \frac{h_{11}x_i + h_{12}y_i + h_{13}}{h_{31}x_i + h_{32}y_i + h_{33}} = x_i \\ \frac{h_{21}x_i + h_{22}y_i + h_{23}}{h_{31}x_i + h_{32}y_i + h_{33}} = y_i \end{cases} \quad (5-4)$$

把式 5-4 从形式上整理得到，

$$\begin{pmatrix} x_i & y_i & 1 & 0 & 0 & 0 & -x_i x_i' & -y_i x_i' & -x_i' \\ 0 & 0 & 0 & x_i & y_i & 1 & -x_i y_i' & -y_i y_i' & -y_i' \end{pmatrix} \begin{pmatrix} h_{11} \\ h_{12} \\ h_{13} \\ h_{21} \\ h_{22} \\ h_{23} \\ h_{31} \\ h_{32} \\ h_{33} \end{pmatrix} = \mathbf{0} \quad (5-5)$$

从式 5-5 中可以看出，由一对点对关系 $\mathbf{x}_i \leftrightarrow \mathbf{x}'_i$ ，我们可以得到两个方程。如果有 4 对点对

关系 $\{\mathbf{x}_i \leftrightarrow \mathbf{x}'_i\}_{i=1}^4$ 的话，便可以得到 8 个线性方程，写成矩阵形式即为，

$$A_{8 \times 9} \mathbf{h}_{9 \times 1} = \mathbf{0} \quad (5-6)$$

其中， $A_{8 \times 9}$ 是方程组的系数矩阵， $\mathbf{h}_{9 \times 1} = (h_{11}, h_{12}, h_{13}, h_{21}, h_{22}, h_{23}, h_{31}, h_{32}, h_{33})^T$ 。在一般情况下， $rank(A_{8 \times 9}) = 8$ ，则齐次线性方程组 5-6 的解空间中存在 $(9-8)=1$ 个线性无关的解向量^[1]，这个解向量便对应于我们最终需要的射影矩阵 H 。因此，从理论上来说，两个平面间的射影变换关系可以由 4 个有效对应点对唯一确定。我们强调是“有效”对应点对，指的是 $\{\mathbf{x}_i\}_{i=1}^4$ 和 $\{\mathbf{x}'_i\}_{i=1}^4$ 中不能存在三点共线的情况。

但在估计平面间的射影变换时，我们得到的点对关系集合 $\mathcal{S} = \{\mathbf{x}_i \leftrightarrow \mathbf{x}'_i\}_{i=1}^p$ 中的元素数量往往会远多于 4 对，即 $p > 4$ ，这时问题会变成什么形式呢？显然，这时从 p 个点对中，我们可以得到 $2p$ 个线性方程，其矩阵形式为，

$$A_{2p \times 9} \mathbf{h}_{9 \times 1} = \mathbf{0} \quad (5-7)$$

其中， $A_{2p \times 9}$ 是方程组的系数矩阵， $\mathbf{h}_{9 \times 1} = (h_{11}, h_{12}, h_{13}, h_{21}, h_{22}, h_{23}, h_{31}, h_{32}, h_{33})^T$ 。在一般情况下， $rank(A_{2p \times 9}) = 9$ ，因此根据齐次线性方程组解的理论^[1]，方程组 5-7 只有平凡的零解。然而零

解对于我们的问题来说没有意义，我们并不需要零解。我们希望能在最小二乘意义之下找到一个适合于方程组 5-7 的非零解 \mathbf{h}^* 。同时注意到，在我们的问题中，解向量 \mathbf{h}^* 实际上代表了射影变换矩阵，而由射影变换的性质可知，对于任意实数 $k \neq 0$ ， $A\mathbf{h}^*$ 与 \mathbf{h}^* 会表达相同的射影变换。因此，不失一般性，我们可以约束 $\|\mathbf{h}^*\|_2^2 = 1$ 。这样，我们的问题就被建模为，

$$\mathbf{h}^* = \arg \min_{\mathbf{h}} \|A\mathbf{h}\|_2^2, \text{ subject to } \|\mathbf{h}\|_2^2 = 1, A \in \mathbb{R}^{2p \times 9}, \mathbf{h} \in \mathbb{R}^{9 \times 1} \quad (5-8)$$

其中， $p > 4$ ， $\text{rank}(A) = 9$ 。

我们可以以一种更加普适的表达方式来描述形如 5-8 所代表的一类问题，

$$\mathbf{x}^* = \arg \min_{\mathbf{x}} \|A\mathbf{x}\|_2^2, \text{ subject to } \|\mathbf{x}\|_2^2 = 1, A \in \mathbb{R}^{m \times n}, \mathbf{x} \in \mathbb{R}^{n \times 1} \quad (5-9)$$

其中， $\text{rank}(A) = n$ 。我们将在 5.1.2 中讲述如何解式 5-9 所定义的这个优化问题。

5.1.2 问题的求解

式 5-9 的求解问题是一个典型的带有等式约束的求函数最小值点的问题。目标函数 $f(\mathbf{x}) = \|A\mathbf{x}\|_2^2$ 与等式约束函数 $g(\mathbf{x}) = 1 - \|\mathbf{x}\|_2^2$ 关于优化变量 \mathbf{x} 都有连续的一阶偏导数。因此，我们可以用拉格朗日乘子法来找出 $f(\mathbf{x})$ 在等式约束 $g(\mathbf{x}) = 0$ 下所有可能的极值点。

构造拉格朗日函数，

$$L(\mathbf{x}, \lambda) = f(\mathbf{x}) + \lambda g(\mathbf{x}) = \|A\mathbf{x}\|_2^2 + \lambda(1 - \|\mathbf{x}\|_2^2) \quad (5-10)$$

根据拉格朗日乘子法的原理，首先要找出 $L(\mathbf{x}, \lambda)$ 的驻点。设 $(\mathbf{x}_0, \lambda_0)$ 是 $L(\mathbf{x}, \lambda)$ 的一个驻点，则它必须要满足，

$$\begin{cases} \frac{\partial L}{\partial \mathbf{x}} \Big|_{\mathbf{x}=\mathbf{x}_0, \lambda=\lambda_0} = \mathbf{0} \\ \frac{\partial L}{\partial \lambda} \Big|_{\mathbf{x}=\mathbf{x}_0, \lambda=\lambda_0} = 0 \end{cases} \Rightarrow \begin{cases} \frac{\partial (\mathbf{x}^T A^T A \mathbf{x} + \lambda(1 - \mathbf{x}^T \mathbf{x}))}{\partial \mathbf{x}} \Big|_{\mathbf{x}=\mathbf{x}_0, \lambda=\lambda_0} = \mathbf{0} \\ \frac{\partial (\mathbf{x}^T A^T A \mathbf{x} + \lambda(1 - \mathbf{x}^T \mathbf{x}))}{\partial \lambda} \Big|_{\mathbf{x}=\mathbf{x}_0, \lambda=\lambda_0} = 0 \end{cases} \Rightarrow \begin{cases} A^T A \mathbf{x}_0 = \lambda_0 \mathbf{x}_0 \\ \mathbf{x}_0^T \mathbf{x}_0 = 1 \end{cases} \quad (5-11)$$

即 λ_0 是 $A^T A$ 的特征值， \mathbf{x}_0 是 $A^T A$ 对应于特征值 λ_0 的单位特征向量。显然，满足这样条件的 “ $(\mathbf{x}_0, \lambda_0)$ ” 并不是唯一的。设集合 $\mathcal{S} = \{(\mathbf{x}_i, \lambda_i) : A^T A \mathbf{x}_i = \lambda_i \mathbf{x}_i, \mathbf{x}_i^T \mathbf{x}_i = 1\}$ ，则 \mathcal{S} 表示 $L(\mathbf{x}, \lambda)$ 的所有驻点。设集合 $\mathcal{C} = \{\mathbf{x}_i : (\mathbf{x}_i, \lambda_i) \in \mathcal{S}\}$ ，则 \mathcal{C} 表示 $f(\mathbf{x})$ 在等式约束 $g(\mathbf{x}) = 0$ 下所有可能的极值点。接下来，我们要在 \mathcal{C} 中挑选出能使 $f(\mathbf{x})$ 取得最小值（在等式约束 $g(\mathbf{x}) = 0$ 下）的点。

若 $\mathbf{x}_i \in \mathcal{C}$ ，则，

$$f(\mathbf{x}_i) = \|A\mathbf{x}_i\|_2^2 = \mathbf{x}_i^T A^T A \mathbf{x}_i = \mathbf{x}_i^T \lambda_i \mathbf{x}_i = \lambda_i \quad (5-12)$$

则可知 $f(\mathbf{x})$ 的最小值为 $\min\{\lambda_i\}$ ⁸, 即为 $A^T A$ 最小的特征值。而 $f(\mathbf{x})$ 能取到这个最小值 (在等式约束 $g(\mathbf{x})=0$ 下) 的点为 $A^T A$ 的对应于其最小特征值的单位特征向量。

5.2 非齐次线性最小二乘问题

5.2.1 问题定义

我们知道, 表达平面间射影变换的矩阵 H 虽然有 9 个元素, 但它只有 8 个自由度。在 5.1 节的实际处理中, 我们以限定“向量化之后的 H 为 9 维单位向量”的方式 (式 5-8) 把它的自由度限定为 8。本节我们用另外一种思路来限定 H 的自由度。

假设 H 中的元素 $h_{ij} \neq 0$, 则在求解 H 的过程中可以把 h_{ij} 固定为一个常数 $c \neq 0$ 。不失一般性, 假设 $h_{33} \neq 0$, 我们固定 h_{33} 为 $h_{33} = 1$ 。这样, 给定一对对应点对 $\mathbf{x}_i^* = (x_i^*, y_i^*, 1)^T$ 和 $\mathbf{x}_i = (x_i, y_i, 1)^T$, 它们之间的关系可表达为,

$$c \begin{bmatrix} x_i^* \\ y_i^* \\ 1 \end{bmatrix} = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & 1 \end{bmatrix} \begin{bmatrix} x_i \\ y_i \\ 1 \end{bmatrix} \quad (5-13)$$

对式 5-13 左右展开得到,

$$\begin{cases} h_{11}x_i + h_{12}y_i + h_{13} = cx_i^* \\ h_{21}x_i + h_{22}y_i + h_{23} = cy_i^* \\ h_{31}x_i + h_{32}y_i + 1 = c \end{cases} \quad (5-14)$$

将式 5-14 中第一式和第三式的左右两边相除、第二式和第三式的左右两边相除, 得到,

$$\begin{cases} \frac{h_{11}x_i + h_{12}y_i + h_{13}}{h_{31}x_i + h_{32}y_i + 1} = x_i^* \\ \frac{h_{21}x_i + h_{22}y_i + h_{23}}{h_{31}x_i + h_{32}y_i + 1} = y_i^* \end{cases} \quad (5-15)$$

把式 5-15 从形式上整理得到,

⁸ $\{\lambda_i\}$ 表示由 $A^T A$ 的所有特征值构成的集合。

$$\begin{pmatrix} x_i & y_i & 1 & 0 & 0 & 0 & -x_i \vec{x}_i & -y_i \vec{x}_i \\ 0 & 0 & 0 & x_i & y_i & 1 & -x_i \vec{y}_i & -y_i \vec{y}_i \end{pmatrix} \begin{pmatrix} h_{11} \\ h_{12} \\ h_{13} \\ h_{21} \\ h_{22} \\ h_{23} \\ h_{31} \\ h_{32} \end{pmatrix} = \begin{pmatrix} \vec{x}_i \\ \vec{y}_i \end{pmatrix} \quad (5-16)$$

从式 5-16 中可以看出, 由一对点对关系 $\mathbf{x}_i \leftrightarrow \vec{\mathbf{x}}_i$, 我们可以得到两个方程。如果有 4 对点对关系 $\{\mathbf{x}_i \leftrightarrow \vec{\mathbf{x}}_i\}_{i=1}^4$ 的话, 便可以得到 8 个线性方程, 写成矩阵的形式为,

$$A_{8 \times 8} \mathbf{h}_{8 \times 1} = \mathbf{b}_{8 \times 1} \quad (5-17)$$

其中, $A_{8 \times 8}$ 是方程组的系数矩阵, $\mathbf{h}_{8 \times 1} = (h_{11}, h_{12}, h_{13}, h_{21}, h_{22}, h_{23}, h_{31}, h_{32})^T$, $\mathbf{b}_{8 \times 1} = (\vec{x}_1, \vec{y}_1, \vec{x}_2, \vec{y}_2, \vec{x}_3, \vec{y}_3, \vec{x}_4, \vec{y}_4)^T$ 。在一般情况下 ($\{\mathbf{x}_i\}_{i=1}^4$ 中以及 $\{\vec{\mathbf{x}}_i\}_{i=1}^4$ 中都不能有三点共线), $\text{rank}(A_{8 \times 8}) = \text{rank}([A_{8 \times 8}; \mathbf{b}]) = 8$, 则方程组 5-17 有唯一解^[1], 从这个解向量我们就可以相应得到最终的射影矩阵 H 。因此, 从理论上来说, 通过两个平面内 4 个有效对应点对, 我们便可以唯一确定这两个平面间的射影变换关系。

但一般情况下, 我们的点对关系集合 $\mathcal{S} = \{\mathbf{x}_i \leftrightarrow \vec{\mathbf{x}}_i\}_{i=1}^p$ 中的元素数量远多于 4 对, 即 $p > 4$ 。这时从 p 个点对中, 可以得到 $2p$ 个线性方程, 其矩阵形式为,

$$A_{2p \times 8} \mathbf{h}_{8 \times 1} = \mathbf{b}_{2p \times 1} \quad (5-18)$$

其中, $A_{2p \times 8}$ 是方程组的系数矩阵, $\mathbf{h}_{8 \times 1} = (h_{11}, h_{12}, h_{13}, h_{21}, h_{22}, h_{23}, h_{31}, h_{32})^T$, $\mathbf{b}_{2p \times 1}$ 为非零常数向量。在一般情况下, 方程组 5-18 的系数矩阵的秩 $\text{rank}(A_{2p \times 8}) = 8$, 而其增广矩阵的秩 $\text{rank}([A_{2p \times 8}; \mathbf{b}]) = 9$, 因此根据线性方程组解的理论^[1], 方程组 5-18 无解。既然从理论上来说, 方程组 5-18 无解, 我们只能退而求其次, 希望能在最小二乘意义之下找到一个适合于方程组 5-18 的 \mathbf{h}^* 。这样, 我们的问题就被建模为,

$$\mathbf{h}^* = \arg \min_{\mathbf{h}} \|A\mathbf{h} - \mathbf{b}\|_2^2, A \in \mathbb{R}^{2p \times 8}, \mathbf{h} \in \mathbb{R}^{8 \times 1}, \mathbf{b} \in \mathbb{R}^{2p \times 1} \quad (5-19)$$

其中, $p > 4$, $\text{rank}(A) = 8$ 。

我们可以以一种更加普适的表达方式来描述形如 5-19 所代表的一类问题,

$$\mathbf{x}^* = \arg \min_{\mathbf{x}} \|A\mathbf{x} - \mathbf{b}\|_2^2, A \in \mathbb{R}^{m \times n}, \mathbf{x} \in \mathbb{R}^{n \times 1}, \mathbf{b} \neq \mathbf{0} \in \mathbb{R}^{m \times 1} \quad (5-20)$$

其中, $\text{rank}(A) = n$ 。我们将在 5.2.2 中讲述如何解式 5-20 所定义的这个优化问题。

5.2.2 问题的求解

求式 5-20 最优解的问题是一个典型的无约束凸优化问题。我们可以首先来证明该问题的目标函数，

$$f(\mathbf{x}) = \|A\mathbf{x} - \mathbf{b}\|_2^2, A \in \mathbb{R}^{m \times n}, \mathbf{x} \in \mathbb{R}^{n \times 1}, \mathbf{b} \neq \mathbf{0} \in \mathbb{R}^{m \times 1} \quad (5-21)$$

为凸函数，这个证明作为练习请读者来完成。然后，我们来找到目标函数 $f(\mathbf{x})$ 的驻点。如果 \mathbf{x}_s 为 $f(\mathbf{x})$ 的驻点，那么 \mathbf{x}_s 需要满足，

$$\begin{aligned} \nabla f(\mathbf{x})|_{\mathbf{x}=\mathbf{x}_s} &= \frac{d\|A\mathbf{x}-\mathbf{b}\|_2^2}{d\mathbf{x}}|_{\mathbf{x}=\mathbf{x}_s} \\ &= \frac{d((A\mathbf{x}-\mathbf{b})^T(A\mathbf{x}-\mathbf{b}))}{d\mathbf{x}}|_{\mathbf{x}=\mathbf{x}_s} \\ &= \frac{d(\mathbf{x}^T A^T A \mathbf{x} - 2\mathbf{x}^T A^T \mathbf{b} + \mathbf{b}^T \mathbf{b})}{d\mathbf{x}}|_{\mathbf{x}=\mathbf{x}_s} \\ &= 2A^T A \mathbf{x} - 2A^T \mathbf{b}|_{\mathbf{x}=\mathbf{x}_s} = \mathbf{0} \end{aligned} \quad (5-22)$$

因此，

$$\mathbf{x}_s = (A^T A)^{-1} A^T \mathbf{b} \quad (5-23)$$

我们需要注意的是，式 5-23 要成立的话， $A^T A$ 必须要可逆才可以。事实上，在我们这个问题中，由于我们要求 $\text{rank}(A) = n$ ，即 A 是列满秩矩阵，则可以证明 $A^T A$ 一定是可逆的，这个证明作为练习请读者来完成。待优化的目标函数 $f(\mathbf{x})$ 为凸函数，则它的驻点一定也是全局最小值点^[2]。因此，问题 5-20 的最优解就是 $\mathbf{x}^* = \mathbf{x}_s = (A^T A)^{-1} A^T \mathbf{b}$ 。

5.2.3 基于奇异值分解原理的求解方法

本节将介绍非齐次线性最小二乘问题的另外一种解法：基于奇异值分解的方法。该方法同 5.2.2 中介绍的方法相比，有两个优越之处：1) 要用 5.2.2 节中介绍的方法来解非齐次线性最小二乘问题时，问题中的系数矩阵必须是列满秩矩阵，如式 5-20 中， $\text{rank}(A_{m \times n}) = n$ ，而本节介绍的方法并不需要待解问题满足这个附加条件；2) 从计算机算法实现的角度来说，本节中介绍的基于奇异值分解的方法所产生的解会具有更高的数值精度^[3]。如果读者对矩阵奇异值分解的基本内容不太熟悉的话，可参见本书附录 D。

首先再来梳理一下我们要解决的问题。我们想要解如下线性方程组，

$$A_{m \times n} \mathbf{x}_{n \times 1} = \mathbf{b}_{m \times 1}, \mathbf{b} \neq \mathbf{0} \quad (5-24)$$

方程组 5-24 的解会出现的情况无外乎以下三种：1) $\text{rank}(A) = \text{rank}([A; \mathbf{b}]) = n$ ，此时方程组有唯一解，我们要把这个解找出来；2) $\text{rank}(A) = \text{rank}([A; \mathbf{b}]) < n$ ，此时方程组有无穷多组解，我们要找到其中一个来解决我们手里的实际问题；3) $\text{rank}(A) \neq \text{rank}([A; \mathbf{b}])$ ，此时方

程组无解，我们要在最小二乘意义之下找到一个“最适合”该方程组的解。不难看出，对以上三种情况的处理都可以归结为求解如下问题，

$$\mathbf{x}^* = \arg \min_{\mathbf{x}} \|A\mathbf{x} - \mathbf{b}\|_2^2, A \in \mathbb{R}^{m \times n}, \mathbf{x} \in \mathbb{R}^{n \times 1}, \mathbf{b} \neq \mathbf{0} \in \mathbb{R}^{m \times 1} \quad (5-25)$$

需要注意的是，式 5-25 所定义的问题同式 5-20 不同，后者有一个额外的要求 $\text{rank}(A) = n$

而前者没有。下面我们就来看看如何具体来求解问题 5-25。

对 A 进行奇异值分解得到，

$$A = U_{m \times m} \Sigma_{m \times n} V_{n \times n}^T \quad (5-26)$$

其中 U 和 V 为正交矩阵。假设 $\text{rank}(A) = r$ ，则，

$$\Sigma_{m \times n} = \begin{bmatrix} \sigma_1 & & & \\ & \sigma_2 & & \\ & & \ddots & \\ & & & \sigma_r \\ O_{(m-r) \times r} & & & O_{(m-r) \times (n-r)} \end{bmatrix}_{m \times n} \quad (5-27)$$

其中 $\sigma_1, \sigma_2, \dots, \sigma_r > 0$ 为矩阵 A 的奇异值。进一步有，

$$\begin{aligned} A\mathbf{x} - \mathbf{b} &= U\Sigma V^T \mathbf{x} - \mathbf{b} \\ &= U(\Sigma V^T \mathbf{x}) - U(U^T \mathbf{b}) \\ &= U(\Sigma V^T \mathbf{x} - U^T \mathbf{b}) \\ &\triangleq U(\Sigma \mathbf{y}_{n \times 1} - \mathbf{c}_{m \times 1}) \end{aligned} \quad (5-28)$$

其中 $\mathbf{y}_{n \times 1} = V^T \mathbf{x}$, $\mathbf{c}_{m \times 1} = U^T \mathbf{b}$ 。由于 U 是正交矩阵，它可以保持向量长度，因此，

$$\|A\mathbf{x} - \mathbf{b}\|_2 = \|U(\Sigma \mathbf{y} - \mathbf{c})\|_2 = \|\Sigma \mathbf{y}_{n \times 1} - \mathbf{c}_{m \times 1}\|_2 \quad (5-29)$$

我们的最终目的是要找到 $\mathbf{x}^* = \arg \min_{\mathbf{x}} \|A\mathbf{x} - \mathbf{b}\|_2$ 。由于有 5-29 式，我们可以间接地先找出

最优的 $\mathbf{y}^* = \arg \min_{\mathbf{y}} \|\Sigma \mathbf{y}_{n \times 1} - \mathbf{c}_{m \times 1}\|_2$ ，再根据 $\mathbf{y}^* = V^T \mathbf{x}^*$ 解出 \mathbf{x}^* 即可。

由于，

$$\Sigma \mathbf{y}_{n \times 1} = \begin{bmatrix} \sigma_1 & & & \\ & \sigma_2 & & \\ & & \ddots & \\ & & & \sigma_r \\ O_{(m-r) \times r} & & & O_{(m-r) \times (n-r)} \end{bmatrix}_{m \times n} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \sigma_1 y_1 \\ \sigma_2 y_2 \\ \vdots \\ \sigma_r y_r \\ 0 \\ \vdots \\ 0 \end{bmatrix}_{m \times 1} \quad (5-30)$$

所以，

$$\Sigma \mathbf{y}_{n \times 1} - \mathbf{c}_{m \times 1} = \begin{bmatrix} \sigma_1 y_1 - c_1 \\ \sigma_2 y_2 - c_2 \\ \vdots \\ \sigma_r y_r - c_r \\ -c_{r+1} \\ \vdots \\ -c_m \end{bmatrix}_{m \times 1} \quad (5-31)$$

根据式 5-31, 只要让 $y_i = \frac{c_i}{\sigma_i}, 1 \leq i \leq r$ (此时 y_{r+1}, \dots, y_n 可以是任意值), 则 $\|\Sigma \mathbf{y}_{n \times 1} - \mathbf{c}_{m \times 1}\|_2$

便可取到最小长度 $\left(\sum_{i=r+1}^m c_i^2 \right)^{1/2}$ 。满足这个要求的一个“最简单”的 $\mathbf{y}_{n \times 1}^*$ 可以为,

$$\mathbf{y}_{n \times 1}^* = \begin{bmatrix} \frac{1}{\sigma_1} \\ \frac{1}{\sigma_2} \\ \ddots \\ \frac{1}{\sigma_r} \\ O_{(n-r) \times r} \end{bmatrix}_{n \times 1} O_{r \times (m-r)} \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_m \end{bmatrix}_{m \times 1} = \begin{bmatrix} c_1 / \sigma_1 \\ c_2 / \sigma_2 \\ \vdots \\ c_r / \sigma_r \\ 0 \\ \vdots \\ 0 \end{bmatrix}_{n \times 1} \triangleq \Sigma^+ \mathbf{c}_{m \times 1} \quad (5-32)$$

其中, Σ^+ 代表了对矩阵 Σ 做转置并把非零对角元取倒数的操作。需要强调的是, 当 $r < n$ 时, 最优的 \mathbf{y}^* 是不唯一的, 式 5-32 中给出的 \mathbf{y}^* 只是满足条件 $y_i = \frac{c_i}{\sigma_i} (1 \leq i \leq r)$ 的最优 \mathbf{y}^* 中的一个。当然, 如果 $r = n$, 即系数矩阵 $A_{m \times n}$ 是列满秩矩阵, 则最优 \mathbf{y}^* 是唯一的。

有了 \mathbf{y}^* 之后, 可以自然得出 \mathbf{x}^* ,

$$\mathbf{x}^* = V \mathbf{y}^* = V \Sigma^+ \mathbf{c} = V \Sigma^+ U^T \mathbf{b} \quad (5-33)$$

其中, $A^+ \triangleq V \Sigma^+ U^T$ 称为 A 的 Moore-Penrose 广义逆。

5.3 习题

(1) 式 5-12 中出现的 λ_i 有没有可能是负数? 为什么?

(2) 请 证 明 式 5-20 中 的 优 化 问 题 中 , 目 标 函 数 $f(\mathbf{x}) = \|A\mathbf{x} - \mathbf{b}\|_2^2, A \in \mathbb{R}^{m \times n}, \mathbf{x} \in \mathbb{R}^{n \times 1}, \mathbf{b} \neq \mathbf{0} \in \mathbb{R}^{m \times 1}$ 为凸函数。提示: 由于 $f(\mathbf{x})$ 二阶可微, 我

们只需要证明该函数的定义域为凸集并且它的 Hessian 矩阵为半正定矩阵^[2]即可。

(3) 有矩阵 $A \in \mathbb{R}^{m \times n}$ 且 $\text{rank}(A) = n$ ，请证明矩阵 $A^T A$ 必为可逆矩阵。

参考文献

- [1] 李世栋, 乐经良, 冯卫国, 王纪林, 线性代数, 科学出版社, 2000 年。
- [2] Stephen Boyd, Lieven Vandenberghe, Convex Optimization, Cambridge University Press, 2004.
- [3] Gene H. Golub, Charles F. Van Loan, Matrix Computations, John Hopkins Univ Press, Baltimore, 1983.

第 6 章 射影矩阵的鲁棒估计与图像的插值

6.1 随机抽样一致算法

在第 5 章中，我们解决了这样一个问题：假设得到了图像 I_1 和 I_2 中特征点对应点对关系结合 $\mathcal{S} = \{\mathbf{x}_i \leftrightarrow \mathbf{x}'_i\}_{i=1}^p$ ，通过最小二乘法对线性方程组 $\{\mathbf{c}\mathbf{x}'_i = H_{3 \times 3}\mathbf{x}_i\}_{i=1}^p$ 进行求解，便可得到图像 I_1 和 I_2 之间的射影变换矩阵 H 。在这个过程中，我们实际上隐含了一个很强的假设，那就是要假设集合 \mathcal{S} 中的所有点对关系都是正确的，即不存在错误的匹配。但在绝大多数现实情况中，特征点检测算法、描述子构造算法以及特征点匹配策略，都不是完美无缺的，这会导致我们手里的对应点对关系集合 \mathcal{S} 中很可能会存在某些错误的对应关系。在 \mathcal{S} 中存在错误对应点对关系的情况下，若直接将 \mathcal{S} 中的数据不加区别地输入给最小二乘法来解出 H ，那这个 H 很有可能离“正确的 H ”相去甚远。那么，我们是否有一种处理策略，可以在从集合 \mathcal{S} 估计射影变换 H 的过程中，尽可能地摆脱错误对应点对关系的影响？

实际上，我们可以将射影矩阵估计这个问题拓广到一类更加广泛的问题：如何从可能存在外点（outlier）的观测数据集合中鲁棒地拟合出参数模型？下面通过一个简单的具体实例来对该问题及相关的概念进行阐释。

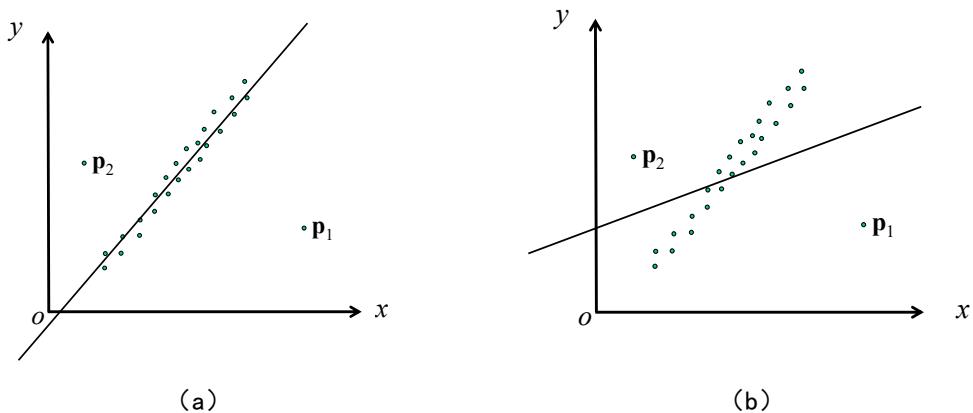


图 6-1：直线拟合。(a) 把 p_1 、 p_2 视作外点，在直线拟合过程中不考虑此两点；(b) 没有进行外点区分，所有的数据点都参与直线拟合。

如图 6-1 所示，假设我们的任务是要从一组平面二维数据点中拟合出一条平面直线。平面直线的方程为 $y=ax+b$ ，进行直线拟合也就是要基于观测数据点确定出模型中待定参数 a 和 b 的值。由于我们已经知道要拟合的数学模型为一条直线，而且大部分观测数据应该是可靠的，因此可以合理地认为大部分观测点应该大致沿着一条直线分布。带着这个先验知识，我们来看一下图 6-1 中的观测数据点。除了 p_1 、 p_2 以外，大部分的观测点都是正常的。唯有 p_1 、 p_2 显得有些不正常，因为它们显然游离在了大部分点所组成的一致集合之外，因此， p_1 、 p_2 便是两个外点。如果在直线拟合过程中，不对外点进行区分，即利用所有的观测数据点来

进行直线拟合，得到的直线就会如图 6-1 (b) 所示，这显然不是我们所期待的正确结果。如果我们能有办法识别并剔除外点 \mathbf{p}_1 、 \mathbf{p}_2 ，而只使用剩余的“内点”集合来进行直线拟合的话，得到的便是图 6-1 (a) 中所示的结果，显然这是符合预期的正确结果。

那么，如何才能在基于观测数据的模型拟合过程中消除掉外点的影响呢？一个常用的用于解决这一类问题的算法框架是随机抽样一致(Random Sample Consensus, RANSAC)算法。该算法最早由美国学者 Martin Fischler 和 Robert Bolles 于 1981 年发表在 ACM 通讯上^[1]。RANSAC 是一种迭代算法，在迭代过程中不断尝试从输入观测数据集合 \mathcal{S} 中寻找更好的一致集。在每一次迭代过程中，基于从观测数据集中随机选取的观测数据点来进行模型拟合，并计算当前模型的一致集。模型的一致集是由观测数据集中的这样一些点组成的：该点带入模型后，根据某种选定的度量函数计算出来的误差值小于预先设定的阈值。算法最终要么会返回由最好的一致集所拟合出来的模型，要么算法失败（即无法从观测数据集中拟合出满足条件的参数模型）。算法 6-1 给出了 RANSAC 算法框架的伪码。

算法 6-1: RANSAC 模型拟合算法

输入:

```

data //观测数据集
n //拟合模型所需要的最少的数据点个数
k //最大允许迭代次数
t //阈值，若数据点带入模型所得误差小于 t，则认为该数据点属于该模型的一致集
d //阈值，若当前模型的一致集中数据点的个数多于 d，则认为该一致集已经足够好

```

输出: $bestFit$ //拟合出来的模型参数，若为空则表明拟合失败

```

iterations = 0
bestFit = null
bestErr = something really large

while iterations < k do
    maybeInliers := n randomly selected values from data
    maybeModel := model parameters fitted to maybeInliers
    alsoInliers := empty set
    for every point in data not in maybeInliers do //计算 maybeModel 的一致集
        if point fits maybeModel with an error smaller than t
            add point to alsoInliers
        end if
    end for
    if the number of elements in alsoInliers is > d then
        // 这意味着我们可能已经找到了一个很好的模型
        // 把该模型从当前一致集中拟合出来
        betterModel := model parameters fitted to all points in maybeInliers and alsoInliers
        thisErr := a measure of how well betterModel fits these points
        if thisErr < bestErr then //完成输出模型及其误差更新
            bestFit := betterModel
            bestErr := thisErr
        end if
    end if
    increment iterations
end while

```

```
return bestFit
```

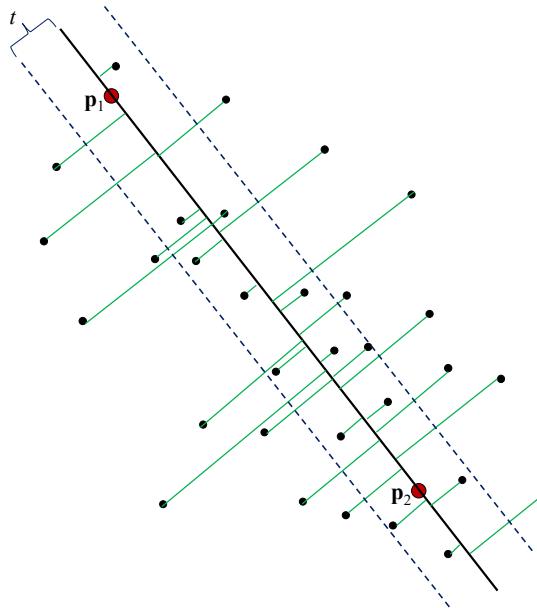


图 6-2: 在用 RANSAC 框架来解决直线拟合问题的一次迭代过程中, maybeInliers、maybeModel 以及 alsoInliers。此次迭代过程中, 随机选择的两点 p_1 、 p_2 构成了 maybeInliers; 由 p_1 、 p_2 所拟合的直线便是 maybeModel; 之后, 计算给定平面点中除了 p_1 、 p_2 之外的每一点到 maybeModel 所代表的直线的距离, 若相应的距离小于 t , 则该点属于 alsoInliers; maybeInliers 和 alsoInliers 集合一起构成了当前拟合出的直线 maybeModel 的一致集。

结合着上面提到的直线拟合这个具体任务, 我们来理解一下算法 6-1 中的关键变量和处理步骤。 $data$ 就是给定的二维数据点集合。由于两个不重合的点可以唯一地确定一条直线, 因此 $n=2$ 。最大迭代次数 k 以及两个阈值 t 和 d 的取值需要根据具体任务的经验知识来确定, 或者需要通过尝试性的试验来确定。如图 6-2 所示, 在某次迭代过程中, 随机选择的两点 p_1 、 p_2 构成了 maybeInliers; 由 p_1 、 p_2 所拟合的直线便是 maybeModel。一个数据点在 maybeModel 下的拟合误差可以用该点到 maybeModel 所确定的直线的距离来表示。这样, 接下来计算数据集中除了 p_1 、 p_2 之外的每一点到 maybeModel 所表示的直线的距离; 若相应的距离小于 t , 则该点属于 alsoInliers; maybeInliers 和 alsoInliers 集合一起构成了当前模型 maybeModel 的一致集。若当前模型的一致集中的元素足够多了 (大于 $d+2$), 则可基于该一致集中的全部数据采用最小二乘法估计出直线模型 betterModel, 并度量该 betterModel 的精度 thisErr; thisErr 可以用当前一致集中所有点到 betterModel 所代表的直线的距离的平均值来表示。

我们来稍微展开讨论一下算法 6-1 中最大迭代次数 k 的确定。对于某些类型的问题, 我们可以事先大致估计出给定观测数据的内点比例 ω 。这样, 一次估计中随机选取的用于估计模型的 n 个点都为内点的概率就为 ω^n 。如果要保证在 k 次迭代过程中, 至少有一次估计模型时所用的所有 n 个数据点都是内点的概率为 p , 那么我们是可以把 k 确定出来的。设事件 A 为“每次随机选取的 n 个用于估计模型的点中至少有一个是外点”, 则事件 A 每次发生的

概率为 $1-\omega^n$ 。同时，在 k 次迭代中，事件 A 是独立的。基于这些条件可知，事件 A 满足了贝努利试验的条件。这样，迭代了 k 次之后，事件 A 发生了 k 次的概率为 $P(k) = C_k^k (1-\omega^n)^k (\omega^n)^0$ 。

同时，根据条件可知， $P(k) = 1 - p$ 。因此有，

$$C_k^k (1-\omega^n)^k (\omega^n)^0 = 1 - p \quad (6-1)$$

则有，

$$k = \frac{\log(1-p)}{\log(1-\omega^n)} \quad (6-2)$$

最后再强调一下，RANSAC 是一个算法框架，它并不是为解决某一个具体问题而设计的，而是用于解决“从可能存在外点的观测数据集中鲁棒地拟合出参数模型”这一类问题的通用框架。回到本章需要解决的问题：从可能包含外点的对应点对关系集合 $\mathcal{S} = \{\mathbf{x}_i \leftrightarrow \mathbf{x}'_i\}_{i=1}^p$ 中估计图像 I_1 和 I_2 之间的射影变换矩阵 H 。如果用 RANSAC 模型拟合算法（算法 6-1）来解决这个具体的从观测数据集合（ \mathcal{S} ）中拟合出参数模型（ H ）的问题的话，算法 6-1 中的每个处理步骤是什么？应该如何来做？这个问题请读者作为练习来完成。

6.2 图像的插值

到目前为止，我们已经可以估计出图像 I_1 与 I_2 之间的射影变换矩阵 H 了，即如果 $\mathbf{x}_i \in I_1$ 与 $\mathbf{x}'_i \in I_2$ 为对应点，则有 $\mathbf{x}'_i = H\mathbf{x}_i$ 。之后，便可以把 I_1 中的每个像素点 \mathbf{x}_i 变换到新的位置 $H\mathbf{x}_i$ ，以对齐 I_1 和 I_2 中的图像内容，进行 I_1 和 I_2 的全景拼接。本节将讨论如何具体实现将 I_1 中的像素点 \mathbf{x}_i 变换到位置 $H\mathbf{x}_i$ 这个操作。

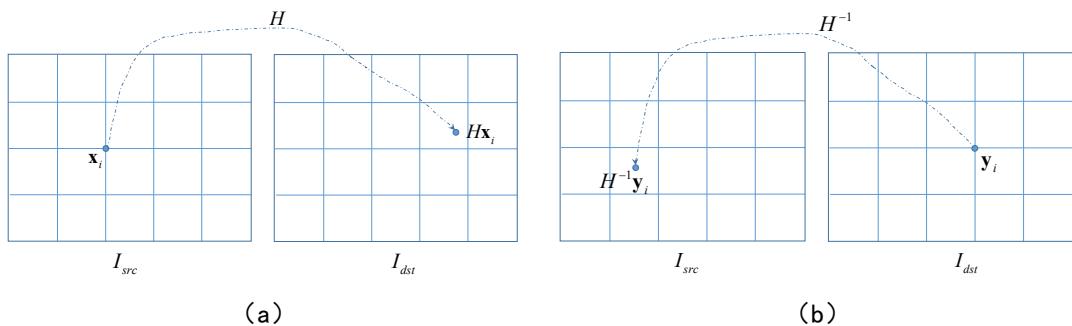


图 6-3：图像坐标变换实现思路示意图。（a）“正向”思路，把源图像坐标映射至目标图像坐标；（b）“逆向”思路，根据目标图像坐标到源图中找对应位置。

假设变换之前的图像为 I_{src} ，变换之后的图像为 I_{dst} 。如图 6-3 (a) 所示，我们先考虑按照“正向”思路来实现从 I_{src} 至 I_{dst} 的变换。对于 I_{src} 中的某一点 \mathbf{x}_i ，它变换之后的位置为 $H\mathbf{x}_i$ ，

因此我们只需要把 I_{dst} 中的对应位置赋值为像素值 $I_{src}(\mathbf{x}_i)$ 即可，即 $I_{dst}(H\mathbf{x}_i)=I_{src}(\mathbf{x}_i)$ 。但实际上，这个“正向”思路是很难实现的，这是因为：由于图像是数字图像的原因，我们只能对图像上整数坐标处的像素值进行存取。 \mathbf{x}_i 为整数坐标，但 $H\mathbf{x}_i$ 几乎不可能也为整数坐标，因此 “ $I_{dst}(H\mathbf{x}_i)=I_{src}(\mathbf{x}_i)$ ” 这个像素赋值操作实际上是不能完成的。

接下来看看按照“逆向”思路是否能实现我们的目的。如图 6-2 (b) 所示，对于 I_{dst} 上的任意一点 \mathbf{y}_i ，我们只要能在 I_{src} 上找到与之对应的点并把那一点的像素值赋值给 $I_{dst}(\mathbf{y}_i)$ 即可。容易知道，在这个过程中，点 \mathbf{y}_i 的坐标为整数，但 I_{src} 上与之对应的位置 $H^1\mathbf{y}_i$ 很大概率上不是整数。我们可以利用 I_{src} 上点 $H^1\mathbf{y}_i$ 周围整数坐标位置处的像素值来“估计”出像素值 $I_{src}(H^1\mathbf{y}_i)$ 。这个根据周围邻域整数坐标位置处的像素值来估计出非整数坐标位置处像素值的过程便称为图像的插值（interpolation）。

图像的插值问题属于典型的数字图像处理问题，常见的解决方法有最近邻插值法、双线性（bilinear）插值法、双三次（bicubic）插值法等。最近邻插值法是最简单的，同时也是计算代价最小的图像插值算法；它直接从 $H^1\mathbf{y}_i$ 的 4 个整数坐标位置处的邻居中挑选一个最近的，然后把这个最近邻居处的像素值作为 $H^1\mathbf{y}_i$ 处的像素值。最近邻插值法的插值效果较差，经常会出现较为明显的锯齿效应或块效应。如果从计算复杂度、易理解性、插值效果三个方面综合考虑的话，双线性差值法是一个非常好的折中选择，因此目前它也是使用的最为广泛的图像插值算法。本节接下来将详细介绍双线性插值法。关于双三次插值法以及更加高级的图像插值算法，读者可参考专门的图像处理书籍^[2]。

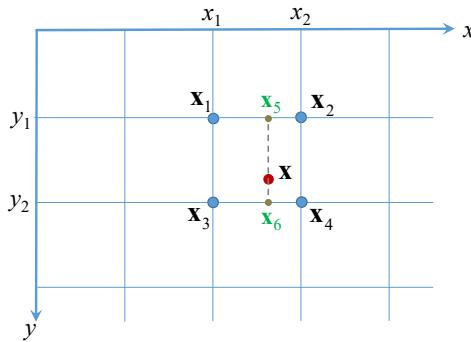


图 6-4：图像的双线性插值算法原理示意图。

如图 6-4 所示，在数字图像 f 上，我们的目标是要估计出非整数坐标位置 $\mathbf{x}=(x, y)$ 处的像素值 $f(\mathbf{x})$ 。首先，要确定出 \mathbf{x} 的 4 个整数位置处的最近邻节点， $\mathbf{x}_1=(x_1, y_1)$ 、 $\mathbf{x}_2=(x_2, y_1)$ 、 $\mathbf{x}_3=(x_1, y_2)$ 和 $\mathbf{x}_4=(x_2, y_2)$ 。所谓“双线性插值”，顾名思义，就是要执行两次线性插值操作。首先，根据 \mathbf{x}_1 、 \mathbf{x}_2 和 \mathbf{x} 沿 x -方向的坐标值线性插值出 $\mathbf{x}_5=(x, y_1)$ 处的像素值 $f(\mathbf{x}_5)$ ； $f(\mathbf{x}_1)$ 与 $f(\mathbf{x}_2)$ 对像素值 $f(\mathbf{x}_5)$ 的贡献线性反比于点 \mathbf{x}_1 、 \mathbf{x}_2 到 \mathbf{x}_5 的距离，

$$f(\mathbf{x}_5) = \frac{x_2 - x}{x_2 - x_1} f(\mathbf{x}_1) + \frac{x - x_1}{x_2 - x_1} f(\mathbf{x}_2) \quad (6-3)$$

同理，也可以从像素值 $f(\mathbf{x}_3)$ 与 $f(\mathbf{x}_4)$ 线性插值出点 $\mathbf{x}_6=(x, y_2)$ 处的像素值 $f(\mathbf{x}_6)$ ，

$$f(\mathbf{x}_6) = \frac{x_2 - x}{x_2 - x_1} f(\mathbf{x}_3) + \frac{x - x_1}{x_2 - x_1} f(\mathbf{x}_4) \quad (6-4)$$

有了点 $\mathbf{x}_5=(x, y_1)$ 和点 $\mathbf{x}_6=(x, y_2)$ 处的像素值 $f(\mathbf{x}_5)$ 和 $f(\mathbf{x}_6)$ 以后，根据 \mathbf{x}_5 、 \mathbf{x}_6 和 \mathbf{x} 沿 y -方向的坐标值再一次通过线性插值便可以得到 \mathbf{x} 处的像素值 $f(\mathbf{x})$ ，

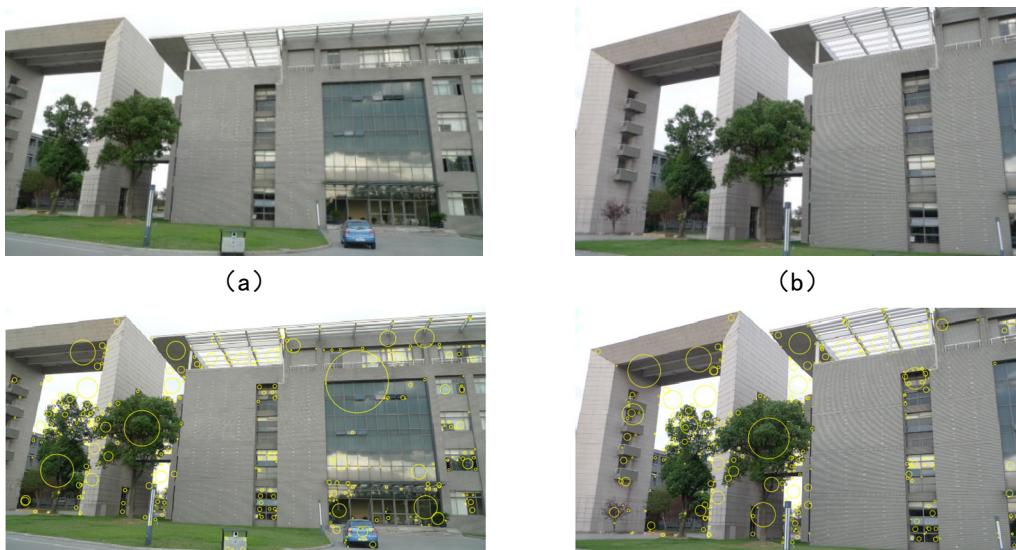
$$f(\mathbf{x}) = \frac{y_2 - y}{y_2 - y_1} f(\mathbf{x}_5) + \frac{y - y_1}{y_2 - y_1} f(\mathbf{x}_6) \quad (6-5)$$

通过联合式 6-3、式 6-4 和式 6-5，同时注意到 $y_2 - y_1 = 1$ 和 $x_2 - x_1 = 1$ ，我们最终便可得到图像的双线性插值公式，

$$f(\mathbf{x}) = (y_2 - y)(x_2 - x)f(\mathbf{x}_1) + (y_2 - y)(x - x_1)f(\mathbf{x}_2) + (y - y_1)(x_2 - x)f(\mathbf{x}_3) + (y - y_1)(x - x_1)f(\mathbf{x}_4) \quad (6-6)$$

利用式 6-6，我们便可以实现对图像的几何变换。

到这里为止，本书第一篇的内容“图像的全景拼接”就全部讲述完毕了。我们以一个具体的图像全景拼接的例子来结束本篇。图 6-5 (a) 和 (b) 是两幅待拼接的图像， I_1 和 I_2 。首先在 I_1 和 I_2 中检测 SIFT 尺度不变特征点，其中 DoG 响应值较强的部分特征点被显示在了图 6-5 (c) 和 (d) 中。在图 6-5 (c) 和 (d) 中，每个圆圈代表了一个 SIFT 特征点，圆圈的中心为特征点的空间位置，圆圈的半径为对应特征点的特征尺度。之后，为每个 SIFT 特征点构建尺度不变特征描述子，并基于特征描述子进行特征点匹配，建立起 I_1 和 I_2 上特征点之间的对应点对关系。图 6-5 (e) 展示了基于特征点匹配建立起来的特征点对应关系；要注意：一般情况下，这些点对关系中会存在对应错误的情况。接下来使用 RANSAC 算法，从可能存在外点（错误匹配关系）的点对关系集合中估计出最优的一致集。图 6-5 (f) 展示了最优一致集中点对关系情况；可以看出，一致集中就不存在错误匹配的点对关系了。基于最优一致集中的数据，使用线性最小二乘法便可以解出 I_1 与 I_2 之间的射影变换矩阵 H 。之后，对 I_1 施加几何变换 H ，便把图像平面 I_1 与 I_2 进行了对齐，在这个过程中需要使用图像插值技术。变换后的 I_1 显示在了图 6-5 (g) 中。最后，把变换后的 I_1 和 I_2 填充到一张图像上，完成全景拼接任务。



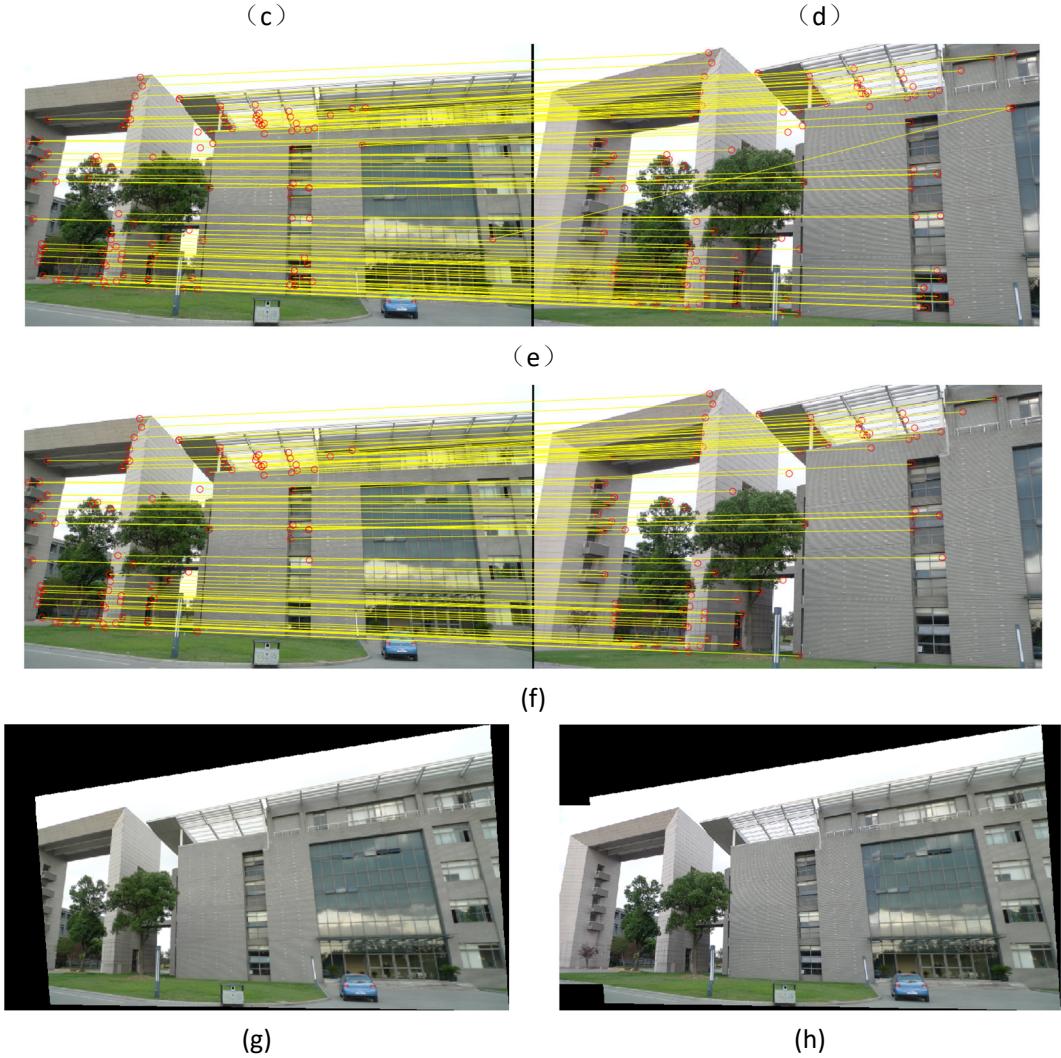


图 6-5：基于特征点匹配思想的图像全景拼接关键步骤处理结果示例。(a) 和 (b) 是待拼接的两幅图像；(c) 和 (d) 为 SIFT 特征点检测结果；(e) 为特征点匹配结果；(f) 利用 RANSAC 算法找到的一致集中的特征点对应关系集合；(g) 对图像 I_1 施加射影变换 H 的结果；(h) I_1 与 I_2 全景拼接的最终结果。

6.3 习题

- (1) 基于 RANSAC 算法框架的图像平面间的射影变换估计。假设得到了图像 I_1 和 I_2 中特征点对应点对关系结合 $\mathcal{S} = \{\mathbf{x}_i \leftrightarrow \mathbf{x}'_i\}_{i=1}^p$ ，但该集合中可能存在外点，即 \mathcal{S} 中的某些对应点对关系有可能是错误的。假设我们用“算法 6-1：RANSAC 模型拟合算法”来解决该问题，请显式地写明算法 6-1 在解决这个具体问题时的处理步骤。
- (2) 运行并理解与本章配套的 Matlab 全景拼接示例程序“PanoramaStitchingSIFTTRANSAC”。该示例程序实现了 SIFT 特征点检测、尺度不变特征描述子构造、特征点匹配、基于 RANSAC 框架的射影变换矩阵估计以及图像的几何变换拼接。把示例程序中的图像替换成你自己的两张图像，再运行该全景拼接程序，看看结果如何。

参考文献

- [1] M.A. Fischler and R.C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography", Communications of the ACM, 24 (6): 381–395, 1981.
- [2] R.C. Gonzalez and R.E. Woods, Digital Image Processing (3rd Edition), Prentice Hall, 2008.

第二篇：单目测量

第 7 章 单目测量问题概述

7.1 问题的定义

在本篇中，读者将要学习到如何给图像中的目标赋予“度量”信息，即要回答图像中的指定目标在实际物理空间中的位置是什么、它的大小是多少等问题。我们可以通过两个例子来更加直观地理解一下本篇中要解决的问题。图像 7-1 (a) 拍摄的是一枚硬币放在桌面上的场景。假设我们使用某种目标分割算法在该图像上分割出了硬币，我们能否进一步知道该硬币的真实直径是多少毫米？图像 7-1 (b) 是由安装在机器人上的相机所采集到的图像。假设我们用目标检测算法在 7-1 (b) 上检测并框出了行人及减速带目标，那么能否知道这些目标在实际物理空间中距离机器人有多远？不难想象，如果没有附加其他额外信息的话，以上描述的两个任务都是不可能的。那么，我们需要提前知道些什么信息才能完成这两个任务呢？

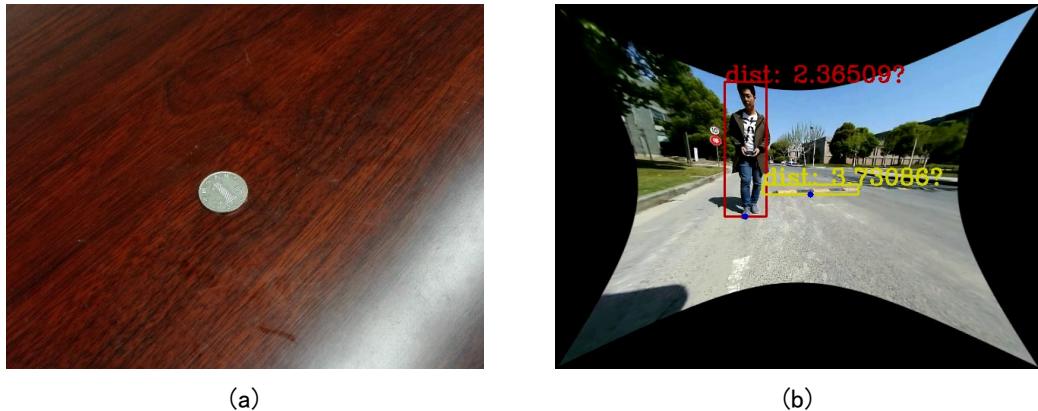


图 7-1：单目测量问题举例。(a) 桌面之上放置一枚硬币，如何从给定图像中提取出硬币的物理直径信息？(b) 该图像由安装在机器人上的相机拍摄获得，如何如同该图所示的那样计算出兴趣目标（行人与减速带）到机器人的距离？

我们把在单张图像上对目标的大小或位置进行测量的问题称为单目测量问题。这类问题需要满足一个前提假设：**待测量的目标要位于一个物理平面之上，图像平面与该物理平面之间满足线性几何变换的关系。**比如，在图 7-1 (a) 中，待测量直径的硬币是位于桌面上的；在图 7-1 (b) 中，我们要假设行人与减速带目标包围框的下边缘是位于路面上的。解决这类问题的关键在于：要事先通过离线标定，计算出图像平面与实际物理平面（如图 7-1 (a) 中的桌面、图 7-1 (b) 中的路面）之间的线性几何变换 H ；当有了 H 以后，我们便可以把图像上的任意点映射到实际物理平面之上，这样便可以得到图像平面上目标物体的实际几何信息。

7.2 方案流程

我们在第一篇中已经学习了如何求解两个（图像）平面之间的线性几何变换。但在那时我们额外附加了一些假设条件，假设两个（图像）平面的所有对应点之间确实是可以经由同一个线性几何变换联系起来的。但在实际情况下，对于一般的相机而言，由于镜头存在畸变，我们并不能保证它所拍摄的物理平面与图像平面之间一定满足线性几何变换的关系。因此，对于单目测量问题，我们首先需要对所拍摄的图像进行去畸变处理。去畸变处理的本质目的是使相机的成像过程严格满足针孔（pin-hole）相机成像模型，从而使得物理空间中的平面与成像平面之间满足线性几何变换关系。

为了对图像进行去畸变处理，我们需要知道包括畸变系数在内的相机所有内参数。**相机内参数**指的是在参数化相机成像模型中与相机自身有关的参数，这些参数的取值仅与相机自身的物理属性有关，与相机所处的外在空间位置无关。对于给定的一个相机，需要对它进行“内参标定”才能获得它的内参数值。有了相机内参之后，我们便可以对该相机所采集的图像进行去畸变处理（实际上，图 7-1 (b) 就是一张经过了去畸变处理之后的图像）。之后，便可以通过离线外参数标定，确定出相机的成像平面（去畸变之后）与目标物体所处平面（如图 7-1 (a) 中的桌面、图 7-1 (b) 中的路面）间的线性几何变换 H 。值得强调的是，对图像进行畸变去除并不是相机内参标定的唯一用途。当构建双目或多目立体视觉系统、基于视觉的三维重建系统、基于视觉的空间定位系统时，我们都必须要知道系统中每个相机的内外参数，唯有如此，我们才能从图像信息中得到三维物理空间中的度量信息。

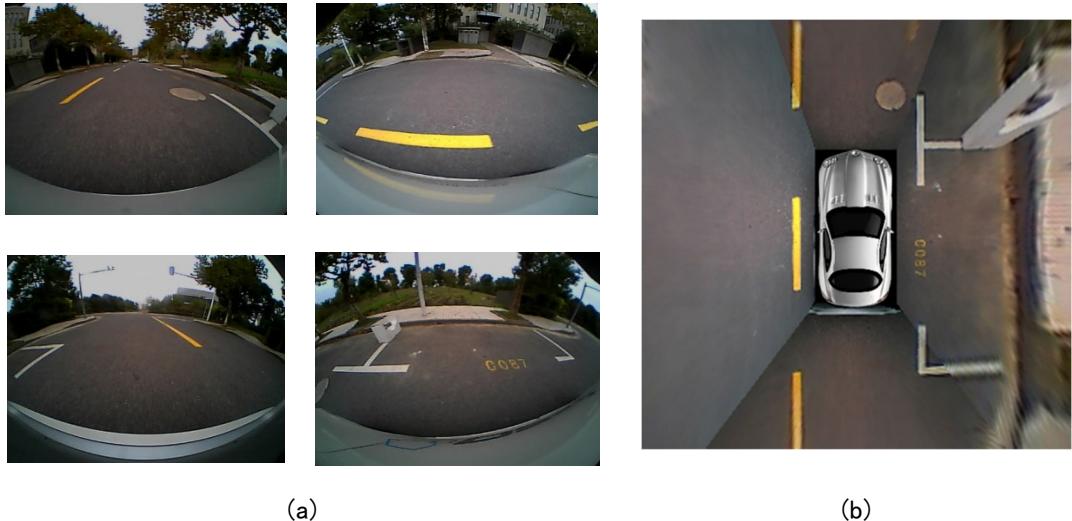


图 7-2：车载环视鸟瞰视图。(a) 四幅由安装在车身四周的鱼眼相机拍摄的图像。(b) 由 (a) 中的四幅图像生成的环视鸟瞰视图，该视图与路面之间的几何变换关系为相似变换。

借助于图像平面与物理平面间的线性几何变换矩阵 H ，我们还可以生成出该物理平面的鸟瞰视图。从几何上来说，鸟瞰视图与它所代表的物理平面之间是相似变换关系。如果待分析的目标是比较“扁”的、位于平面上的目标，比如图 7-1 (a) 中的硬币、图 7-1 (b) 中的

减速带等，在鸟瞰视图中对它们进行观察、检测和测量等操作，会更加方便和直观。图 7-2 中给出了一个环视鸟瞰视图的示例。环视鸟瞰视图经常用于辅助驾驶任务，比如泊车位的检测与定位^[1]等。**7-2 (a)** 是四张由安装在车身四周的鱼眼相机拍摄的图像。经过图像去畸变、外参标定等操作之后，我们可以从**7-2 (a)** 的四幅图像中拼合成**7-2 (b)** 所示的环视鸟瞰视图。显然，在鸟瞰视图之下，对路面上的平面目标（比如车道线、泊车位等）进行检测和测量会更加容易进行。

在接下来的第 8 至 11 章中，将详细阐述单目测量所需的理论和技术知识。

为了学习相机的内参标定，读者需具备一些初步的射影几何方面的基础知识。鉴于大部分计算机领域的初学者可能都不曾系统学习过这方面的内容，我们在第 8 章中会介绍射影几何的基本内容。

从数学角度来看，相机的内参标定问题最终会归结为解一个非线性最小二乘的优化问题。我们会在第 9 章中介绍非线性最小二乘问题及其解法。

第 8 章和第 9 章的内容都是为了要解决相机标定问题而需要事先学习的预备知识。第 10 章首先会介绍针孔相机成像模型，然后会系统讲解对相机内参进行标定的一个最常用的方法—张正友平面标定法^[2]。

最后，在第 11 章中，我们将学习如何生成平面的鸟瞰视图。

参考文献

- [1] L. Zhang, J. Huang, X. Li, and L. Xiong, “Vision-based parking-slot detection: A DCNN-based approach and a large-scale benchmark dataset,” *IEEE Trans. Image Processing*, vol. 27, no. 11, pp. 5350-5364, 2018.
- [2] Z. Zhang, “A flexible new technique for camera calibration,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 22, no. 11, pp. 1330-1334, 2000.

第9章 非线性最小二乘问题

从数学形式上来说，相机参数标定问题最终会转化为一个非线性最小二乘问题，该问题是一类具有特殊结构的无约束优化问题。工程实践中的很多问题都可以被建模为非线性最小二乘问题，因此这类问题求解方法的应用范畴绝不仅限于相机参数标定这个具体任务。本章将先介绍无约束优化问题的相关基本概念和基本方法，之后再具体介绍非线性最小二乘这个特殊的无约束优化问题的求解方法。

9.1 无约束优化问题基础

9.1.1 问题定义与基本概念

考虑这样一个问题： $f(\mathbf{x}) : \mathbb{R}^n \rightarrow \mathbb{R}$ 为一个连续可微函数，其在定义域内的取值有界，我们的目标是要找到它的最小值点。对于这个问题，我们在高等数学中学过的做法是：先要找到 $f(\mathbf{x})$ 的所有驻点，然后通过比较 $f(\mathbf{x})$ 在所有驻点处和可能的定义域端点处的值来找出 $f(\mathbf{x})$ 的最小值点。在这个过程中，为了要找到驻点，我们需要解关于 \mathbf{x} 的方程 $\nabla f(\mathbf{x}) = \mathbf{0}$ ，这个方程只有在极特殊的情况下才存在闭式解，而在绝大多数情况下我们无法找到该方程的闭式解。因而实际上，对于大多数情况来说，由于 $f(\mathbf{x})$ 的形式较为复杂，我们并不能找到 $f(\mathbf{x})$ 的所有驻点，因此在没有其他额外条件的情况下，想找到 $f(\mathbf{x})$ 的全局最小值点是非常困难的。一般来说，我们只能退而求其次，通过迭代优化的办法找到 $f(\mathbf{x})$ 的局部极小值点。如果对迭代的初始点选择恰当的话，局部极值点（虽然它不是问题的全局最优解）对于实际工程问题来说也是足够的。本章要解决的问题就限定为：从一个初始点 \mathbf{x}_0 开始，经过迭代优化，寻找非线性函数 $f(\mathbf{x})$ 的局部极小值点（局部极小值点的定义见附录 E）。

对于一般的非线性优化问题，求解方法都是迭代进行的：从初始迭代点 \mathbf{x}_0 开始，算法在每一次迭代之后都产生一个新的迭代点 $\mathbf{x}_1, \mathbf{x}_2, \dots$ ；我们希望这个过程可以在有限次内完成并最终收敛于函数 $f(\mathbf{x})$ 的一个极小值点 \mathbf{x}^* 。在这个过程中，算法需要有某种度量准则，来确保迭代是沿着使函数值不断减小的方向行进的，即要保证，

$$f(\mathbf{x}_{k+1}) < f(\mathbf{x}_k) \quad (9-1)$$

这样一个准则可以使得迭代过程最终不会收敛于一个局部极大值点，同时也降低了它收敛到一个鞍点（关于鞍点的定义与讨论见附录 E）的可能性^[1]。迭代算法的每一步，从本质上来说，就是要确定迭代更新向量：考虑从 \mathbf{x}_k 开始的一次迭代，我们就是要确定出更新向量 \mathbf{h} ，然后根据 \mathbf{h} 得到下一个迭代点 $\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{h}$ 。从本质上来说，不同迭代优化算法之间的不同之处就在于它们在每次迭代中计算更新向量的方式不同。在下一节，我们将学习一个非常直观和常用的迭代优化算法框架，阻尼法（damped method）。

在过渡到下一节讲解具体的迭代更新算法之前，还有一个宏观层面的问题我们需要明确一下。如果函数 $f(\mathbf{x})$ 有很多个局部极小值点，迭代算法最终会收敛到哪个局部极小值点会和初始迭代点 \mathbf{x}_0 的选择有很大关系，但这并不意味着算法最终收敛到的局部极小值点一定是距离 \mathbf{x}_0 最近的那个局部极小值点^[1]。

9.1.2 阻尼法

现在要解决的问题是，如何确定函数 $f(\mathbf{x}): \mathbb{R}^n \rightarrow \mathbb{R}$ 在点 \mathbf{x} 处的更新向量 \mathbf{h}_{dm} 以得到下一个迭代点 $\mathbf{x} + \mathbf{h}_{dm}$ 。一般来说， $f(\mathbf{x})$ 的形式较为复杂，导致我们不太容易确定合理的更新向量 \mathbf{h}_{dm} 。一个直观的想法便是，可以用一个简单的函数 $l(\mathbf{h}): \mathbb{R}^n \rightarrow \mathbb{R}$ 来近似代替 f 在 \mathbf{x} 附近的形式 $f(\mathbf{x} + \mathbf{h})$ 。兼顾考虑到函数的复杂程度以及对 f 的逼近能力， l 的一个常用的合理选择就是二次函数（quadratic function）形式。同时， $l(\mathbf{h})$ 还要满足 $l(\mathbf{0}) = f(\mathbf{x})$ 。因此， $l(\mathbf{h})$ 需要被构造为，

$$l(\mathbf{h}) = f(\mathbf{x}) + \mathbf{h}^T \mathbf{c} + \frac{1}{2} \mathbf{h}^T B \mathbf{h} \quad (9-2)$$

其中， $\mathbf{c} \in \mathbb{R}^n$ ， $B \in \mathbb{R}^{n \times n}$ 为实对称矩阵。显然， \mathbf{c} 和 B 需要根据函数 f 在 \mathbf{x} 点处的信息来构造，不同的方法会采用不同的构造方式，我们稍后会见到具体的 \mathbf{c} 和 B ，在这里我们姑且认为 \mathbf{c} 和 B 已经构造好了。当 $\|\mathbf{h}\|$ 足够小时， $l(\mathbf{h})$ 可以作为 $f(\mathbf{x} + \mathbf{h})$ 的很好的近似。我们的目标是要找到 \mathbf{h}_{dm} 以使得 $f(\mathbf{x} + \mathbf{h}_{dm})$ 尽可能地小，而我们又假定 $l(\mathbf{h})$ 可以用来近似 $f(\mathbf{x} + \mathbf{h})$ ，因此借助于 $l(\mathbf{h})$ ， \mathbf{h}_{dm} 可以被合理地估计为，

$$\mathbf{h}_{dm} = \arg \min_{\mathbf{h}} l(\mathbf{h}) \quad (9-3)$$

同时我们要注意到，只有当 $\|\mathbf{h}\|$ 很小时， $l(\mathbf{h})$ 才可以很好地近似 $f(\mathbf{x} + \mathbf{h})$ 。式 9-3 仅仅是以最小化 $l(\mathbf{h})$ 为目标，得到的 $\|\mathbf{h}_{dm}\|$ 可能会很大；这就导致尽管 $l(\mathbf{h}_{dm})$ 可能会很小，但 $f(\mathbf{x} + \mathbf{h}_{dm})$ 可能并不一定小，因为这时 $l(\mathbf{h}_{dm})$ 并不一定能很好地近似 $f(\mathbf{x} + \mathbf{h}_{dm})$ 。综合这些分析，不难理解，我们需要在式 9-3 的基础上对大的 $\|\mathbf{h}\|$ 进行适当“惩罚”，从而使得到的 $\|\mathbf{h}_{dm}\|$ 不至于过大。这样，最终设计的迭代更新向量 \mathbf{h}_{dm} 的求解方式就变成了，

$$\mathbf{h}_{dm} = \arg \min_{\mathbf{h}} \left\{ l(\mathbf{h}) + \frac{1}{2} \mu \mathbf{h}^T \mathbf{h} \right\} \quad (9-4)$$

其中， $\mu > 0$ 称为阻尼系数， $\frac{1}{2} \mu \mathbf{h}^T \mathbf{h}$ 称为阻尼项。以式 9-4 所表示的方式来求解更新向量的迭代优化框架便称为“阻尼法（damped method）”（本节中出现的迭代更新向量记为了 \mathbf{h}_{dm} ，其下标 dm 就是 damped 的缩写）。不难理解，阻尼项的目的就是为了要对大的更新步长进行“惩罚”，从而使迭代更新保持“稳步前进”，而不会朝着一个方向一步“走的太远”。

要找式 9-4 中目标函数的最小值点 \mathbf{h}_{dm} , 就需要计算目标函数 $l(\mathbf{h}) + \frac{1}{2}\mu\mathbf{h}^T\mathbf{h}$ 的驻点。该目标函数的形式比较简单, 我们可以容易求出其驻点为 $-(B + \mu I)^{-1}\mathbf{c}$, 其中 I 为 n 阶单位矩阵。容易知道, 目标函数 $l(\mathbf{h}) + \frac{1}{2}\mu\mathbf{h}^T\mathbf{h}$ 的海森矩阵为 $B + \mu I$, 而且这个矩阵和 \mathbf{h} 无关, 即在驻点 $-(B + \mu I)^{-1}\mathbf{c}$ 处的海森矩阵也是 $B + \mu I$ 。我们假定 μ 是合适的以使得 $B + \mu I$ 为正定矩阵, 那么根据定理 E.3 可知, 驻点 $-(B + \mu I)^{-1}\mathbf{c}$ 必为函数 $l(\mathbf{h}) + \frac{1}{2}\mu\mathbf{h}^T\mathbf{h}$ 的局部极小值点。而实际上如果 $B + \mu I$ 为正定矩阵, 则 $l(\mathbf{h}) + \frac{1}{2}\mu\mathbf{h}^T\mathbf{h}$ 必为凸函数, 那么它的局部极小值点 $-(B + \mu I)^{-1}\mathbf{c}$ 也即是它的全局最小值点^[2]。因此, 在 $B + \mu I$ 是正定矩阵的条件下, 式 9-4 的解为,

$$\mathbf{h}_{dm} = -(B + \mu I)^{-1}\mathbf{c} \quad (9-5)$$

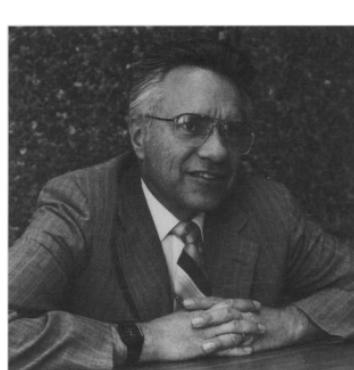


图 9-1: 唐纳德·马夸尔特^[5] (Donald W. Marquardt, 1929 年 3 月 13 日-1997 年 7 月 5 日), 美国统计学家, 因独立提出 Levenberg - Marquardt 非线性最小二乘解法而闻名。他于 1950 年在哥伦比亚大学获得物理学和数学学士学位, 并于 1956 年在特拉华大学获得数学和统计学硕士学位。1953 年, 他加入杜邦, 在那里工作了 39 年, 创立并管理了杜邦质量管理与技术中心。1975 年, 他当选为美国统计协会会士, 1986 年他获得了休哈特奖章。

当迭代算法以上述方式在当前迭代点 \mathbf{x} 处计算出了更新向量 \mathbf{h}_{dm} 之后, 并不会贸然接受 \mathbf{h}_{dm} , 因为毕竟这个更新向量仅是根据 $f(\mathbf{x}+\mathbf{h})$ 的“替身” $l(\mathbf{h})$ 得到的, 它对于 f 来说是否真正合适还需要判断一下: 只有当 $f(\mathbf{x}+\mathbf{h}_{dm}) < f(\mathbf{x})$ 时, 算法才会接受 \mathbf{h}_{dm} 并前进到下一迭代点 $\mathbf{x}+\mathbf{h}_{dm}$; 否则, 不会进行迭代点的更新。**不论当前这一步计算出来的 \mathbf{h}_{dm} 是否被接受, 都需要调整 μ 值来为下一次计算更新向量做准备。**那么如何进行 μ 值的调整呢? 我们可以先来定性的分析一下。如果当前这一步计算出来的 \mathbf{h}_{dm} “不够理想”, 那在下一次计算更新向量的时候就不能过于相信模型 $l(\mathbf{h})$ 了, 就需要对大的更新步长“加大惩罚力度”, 即增加 μ 值; 反之, 我们可以适当调低 μ 值。而什么叫做“不够理想呢”? 这指的是从当前点前进了 \mathbf{h}_{dm} 之后, 模型 l 的值下降了很多 (即 $l(\mathbf{0}) - l(\mathbf{h}_{dm})$ 比较大), 而真正的目标函数 f 的值下降的却很有限 (即 $f(\mathbf{x}) - f(\mathbf{x}+\mathbf{h}_{dm})$ 比较小)。受这个想法启发, 我们可以定义一个量来定量刻画 \mathbf{h}_{dm} 的理想程度,

这个量称为增益比 (gain ratio) ρ , 按如下定义,

$$\rho = \frac{f(\mathbf{x}) - f(\mathbf{x} + \mathbf{h}_{dn})}{l(\mathbf{0}) - l(\mathbf{h}_{dn})} \quad (9-6)$$

需要注意的是, $l(\mathbf{0}) - l(\mathbf{h}_{dn})$ 这部分一定是正的。不难理解, ρ 越小, 说明 \mathbf{h}_{dn} 越差; ρ 越大, 说明 \mathbf{h}_{dn} 越好。这样, 我们就可以根据当前的 ρ 值来动态调整下一步迭代时所用的 μ 值。在文献中, 有两种常用的 μ 值调整算法: 一个是由美国统计学家马夸尔特 (Donald W. Marquardt, 图 9-1) 于 1963 年提出来的^[3], 一个是由丹麦数学家尼尔森 (Hans Bruun Nielsen) 于 1999 年提出来的^[4], 它们分别被总结在了算法 9-1 和算法 9-2 中。

算法 9-1: 马夸尔特阻尼系数 μ 调整算法

```

if  $\rho < 0.25$ 
     $\mu := \mu \times 2$ 
elseif  $\rho > 0.75$ 
     $\mu := \mu \times \frac{1}{3}$ 
end

```

算法 9-2: 尼尔森阻尼系数 μ 调整算法

```

if  $\rho > 0$ 
     $\mu := \mu \times \max \left\{ \frac{1}{3}, 1 - (2\rho - 1)^3 \right\}; v := 2$ 
else
     $\mu := \mu \times v; v := 2 \times v$ 
end

```

在前面介绍模型 $l(\mathbf{h})$ 时, 我们并没有具体说明 l 中的 \mathbf{c} 和 B 是如何根据目标函数 f 在 \mathbf{x} 处的信息来构造的。实际上, 可以有不同的 \mathbf{c} 和 B 的构造方式。这里我们介绍一个具体的例子。如果 \mathbf{c} 取为 f 在 \mathbf{x} 处的梯度、 B 取为 f 在 \mathbf{x} 处的海森矩阵, 即 $\mathbf{c} = \nabla f(\mathbf{x})$ 、 $B = \nabla^2 f(\mathbf{x})$, 则此时由式 9-5 所确定的更新向量的具体形式为,

$$\mathbf{h}_{dn} = -(\nabla^2 f(\mathbf{x}) + \mu I)^{-1} \nabla f(\mathbf{x}) \quad (9-7)$$

通过式 9-7 来计算更新向量的这个阻尼法的具体形式称为“阻尼牛顿法 (damped Newton method)”^[1] (式 9-7 中的更新向量记为了 \mathbf{h}_{dn} , 其下标 dn 就是 damped Newton 的缩写)。

以上我们讲述了在阻尼法迭代优化框架之下, 一次迭代更新所需要完成的两个核心步骤: 更新向量的计算以及阻尼系数的调整。我们还有最后一个问题没有讲清楚, 那就是该迭代算法什么时候终止, 或者说算法终止迭代的条件是什么。不难理解, 如果当前点已经是驻点了或者当前这一步迭代得到的更新向量已经足够小 ($\|\mathbf{h}_{dn}\|$ 足够小), 那么算法就可以停止迭代了。我们将在 9.2.3 节中以列文伯格-马夸尔特法作为一个具体的例子给出阻尼法迭代优化框架的完整算法伪码。

9.2 非线性最小二乘问题及其解法

9.2.1 问题定义与基本概念

在第 5 章中，我们学习了线性最小二乘问题及其解法。在线性最小二乘问题中，我们的目标是求目标函数 $f(\mathbf{x}) = \frac{1}{2} \|A\mathbf{x} - \mathbf{b}\|_2^2$, $A \in \mathbb{R}^{m \times n}$, $\mathbf{x} \in \mathbb{R}^{n \times 1}$, $\mathbf{b} \in \mathbb{R}^{m \times 1}$ (该函数在式 5-25 中给出) 的最

小值点。我们可以对这个目标函数 $f(\mathbf{x})$ 做一下形式上的改变：把 A 按行来表示， $A = \begin{bmatrix} \mathbf{a}_1^T \\ \mathbf{a}_2^T \\ \vdots \\ \mathbf{a}_m^T \end{bmatrix}$

其中 \mathbf{a}_i^T 代表 A 的第 i 行， $\mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix}$ ，其中 b_i 是 \mathbf{b} 的第 i 个元素。令 $f_i(\mathbf{x}) = \mathbf{a}_i^T \mathbf{x} - b_i$ ($i=1,\dots,m$)，

$\mathbf{f}(\mathbf{x}) = (f_1(\mathbf{x}), f_2(\mathbf{x}), \dots, f_m(\mathbf{x}))^T$ ，则有，

$$f(\mathbf{x}) = \frac{1}{2} \sum_{i=1}^m f_i^2(\mathbf{x}) = \frac{1}{2} \|\mathbf{f}(\mathbf{x})\|_2^2 = \frac{1}{2} \mathbf{f}^T(\mathbf{x}) \mathbf{f}(\mathbf{x}) \quad (9-8)$$

在线性最小二乘问题中， $f_i(\mathbf{x})$ 中与优化变量 \mathbf{x} 有关的部分为线性函数，因此该问题才被称之为线性最小二乘问题。而如果 $f_i(\mathbf{x})$ 中关于 \mathbf{x} 的部分不能写为关于 \mathbf{x} 的线性函数，那么以式 9-8 为目标函数的最小二乘问题便称为**非线性最小二乘问题**。线性最小二乘问题有闭式解，我们在第 5 章中已经学习过了。而非线性最小二乘问题一般没有闭式解，只能使用本节所讲述的迭代优化算法求解。本节接下来将讲述非线性最小二乘问题的解法，该问题的目标函数由式 9-8 给出。

若对矢量函数 $\mathbf{f}(\mathbf{x}) : \mathbb{R}^n \rightarrow \mathbb{R}^m$ 进行一阶泰勒展开，可以得到，

$$\begin{aligned} \mathbf{f}(\mathbf{x} + \mathbf{h}) &= \begin{bmatrix} f_1(\mathbf{x} + \mathbf{h}) \\ f_2(\mathbf{x} + \mathbf{h}) \\ \vdots \\ f_m(\mathbf{x} + \mathbf{h}) \end{bmatrix} \\ &= \begin{bmatrix} f_1(\mathbf{x}) + (\nabla f_1(\mathbf{x}))^T \mathbf{h} \\ f_2(\mathbf{x}) + (\nabla f_2(\mathbf{x}))^T \mathbf{h} \\ \vdots \\ f_m(\mathbf{x}) + (\nabla f_m(\mathbf{x}))^T \mathbf{h} \end{bmatrix} = \begin{bmatrix} f_1(\mathbf{x}) \\ f_2(\mathbf{x}) \\ \vdots \\ f_m(\mathbf{x}) \end{bmatrix} + \begin{bmatrix} (\nabla f_1(\mathbf{x}))^T \\ (\nabla f_2(\mathbf{x}))^T \\ \vdots \\ (\nabla f_m(\mathbf{x}))^T \end{bmatrix} \mathbf{h} \\ &= \mathbf{f}(\mathbf{x}) + J(\mathbf{x})\mathbf{h} \end{aligned} \quad (9-9)$$

其中, $J(\mathbf{x}) \triangleq \begin{bmatrix} (\nabla f_1(\mathbf{x}))^T \\ (\nabla f_2(\mathbf{x}))^T \\ \vdots \\ (\nabla f_m(\mathbf{x}))^T \end{bmatrix} \in \mathbb{R}^{m \times n}$ 称为矢量函数 $\mathbf{f}(\mathbf{x})$ 的雅可比矩阵 (Jacobian matrix), 显然它

是由 $\{f_i(\mathbf{x})\}_{i=1}^m$ 的一阶偏导数所组成的一个矩阵; 矩阵 $J(\mathbf{x})$ 有 m 行, 它的第 i 行正是函数 $f_i(\mathbf{x})$ 的梯度向量的转置。

我们再来看一看式 9-8 所定义的目标函数 $f(\mathbf{x})$ 的梯度。先来看看 $\frac{\partial f(\mathbf{x})}{\partial x_j}, j = 1, 2, \dots, n$ 的形式,

$$\begin{aligned} \frac{\partial f(\mathbf{x})}{\partial x_j} &= \frac{1}{2} \frac{\partial [f_1^2(\mathbf{x}) + f_2^2(\mathbf{x}) + \dots + f_m^2(\mathbf{x})]}{\partial x_j} \\ &= f_1(\mathbf{x}) \frac{\partial f_1(\mathbf{x})}{\partial x_j} + f_2(\mathbf{x}) \frac{\partial f_2(\mathbf{x})}{\partial x_j} + \dots + f_m(\mathbf{x}) \frac{\partial f_m(\mathbf{x})}{\partial x_j} \\ &= \sum_{i=1}^m \left[f_i(\mathbf{x}) \frac{\partial f_i(\mathbf{x})}{\partial x_j} \right] \end{aligned} \quad (9-10)$$

基于式 9-10, 可知 $f(\mathbf{x})$ 的梯度 $\nabla f(\mathbf{x})$ 为,

$$\begin{aligned} \nabla f(\mathbf{x}) &= \begin{bmatrix} \frac{\partial f(\mathbf{x})}{\partial x_1} \\ \frac{\partial f(\mathbf{x})}{\partial x_2} \\ \vdots \\ \frac{\partial f(\mathbf{x})}{\partial x_n} \end{bmatrix} = \begin{bmatrix} f_1(\mathbf{x}) \frac{\partial f_1}{\partial x_1} + f_2(\mathbf{x}) \frac{\partial f_2}{\partial x_1} + \dots + f_m(\mathbf{x}) \frac{\partial f_m}{\partial x_1} \\ f_1(\mathbf{x}) \frac{\partial f_1}{\partial x_2} + f_2(\mathbf{x}) \frac{\partial f_2}{\partial x_2} + \dots + f_m(\mathbf{x}) \frac{\partial f_m}{\partial x_2} \\ \vdots \\ f_1(\mathbf{x}) \frac{\partial f_1}{\partial x_n} + f_2(\mathbf{x}) \frac{\partial f_2}{\partial x_n} + \dots + f_m(\mathbf{x}) \frac{\partial f_m}{\partial x_n} \end{bmatrix} \\ &= \begin{bmatrix} \frac{\partial f_1(\mathbf{x})}{\partial x_1} & \frac{\partial f_2(\mathbf{x})}{\partial x_1} & \dots & \frac{\partial f_m(\mathbf{x})}{\partial x_1} \\ \frac{\partial f_1(\mathbf{x})}{\partial x_2} & \frac{\partial f_2(\mathbf{x})}{\partial x_2} & \dots & \frac{\partial f_m(\mathbf{x})}{\partial x_2} \\ \vdots & & & \\ \frac{\partial f_1(\mathbf{x})}{\partial x_n} & \frac{\partial f_2(\mathbf{x})}{\partial x_n} & \dots & \frac{\partial f_m(\mathbf{x})}{\partial x_n} \end{bmatrix}_{n \times m} \begin{bmatrix} f_1(\mathbf{x}) \\ f_2(\mathbf{x}) \\ \vdots \\ f_m(\mathbf{x}) \end{bmatrix} \\ &= (J(\mathbf{x}))^T \mathbf{f}(\mathbf{x}) \end{aligned} \quad (9-11)$$

9.2.2 高斯牛顿法

我们现在要解决的问题是: 对于式 9-8 所给出的目标函数 $f(\mathbf{x})$, 在当前迭代点 \mathbf{x} 处如何得到行进至下一个迭代点的更新向量。我们先来讲述解决这一问题的最直接的方法, 高斯牛顿法 (Gauss-Newton method)。该方法是基于对 $\mathbf{f}(\mathbf{x})$ 在 \mathbf{x} 这点附近的一阶泰勒展开来构造的。

根据式 9-9 可知, 当 $\|\mathbf{h}\|$ 很小时, $\mathbf{f}(\mathbf{x}+\mathbf{h}) \approx \mathbf{f}(\mathbf{x}) + J(\mathbf{x})\mathbf{h}$ 。这样, 式 9-8 中的目标函数 $f(\mathbf{x})$ 在点 $\mathbf{x}+\mathbf{h}$ 处的值近似为,

$$\begin{aligned} f(\mathbf{x}+\mathbf{h}) &= \frac{1}{2} \mathbf{f}^T(\mathbf{x}+\mathbf{h}) \mathbf{f}(\mathbf{x}+\mathbf{h}) \\ &\approx \frac{1}{2} (\mathbf{f}(\mathbf{x}) + J(\mathbf{x})\mathbf{h})^T (\mathbf{f}(\mathbf{x}) + J(\mathbf{x})\mathbf{h}) \\ &= \frac{1}{2} \mathbf{f}^T(\mathbf{x}) \mathbf{f}(\mathbf{x}) + \mathbf{h}^T (J^T(\mathbf{x}) \mathbf{f}(\mathbf{x})) + \frac{1}{2} \mathbf{h}^T J^T(\mathbf{x}) J(\mathbf{x}) \mathbf{h} \\ &= f(\mathbf{x}) + \mathbf{h}^T (J^T(\mathbf{x}) \mathbf{f}(\mathbf{x})) + \frac{1}{2} \mathbf{h}^T J^T(\mathbf{x}) J(\mathbf{x}) \mathbf{h} \end{aligned} \quad (9-12)$$

记,

$$l(\mathbf{h}) = f(\mathbf{x}) + \mathbf{h}^T (J^T(\mathbf{x}) \mathbf{f}(\mathbf{x})) + \frac{1}{2} \mathbf{h}^T J^T(\mathbf{x}) J(\mathbf{x}) \mathbf{h} \quad (9-13)$$

则 $f(\mathbf{x}+\mathbf{h}) \approx l(\mathbf{h})$ 。我们的目标是要找到更新向量 \mathbf{h}_{gn} 使得 $f(\mathbf{x}+\mathbf{h}_{gn})$ 尽可能地小, 这个目标可以借助于求解 $f(\mathbf{x}+\mathbf{h})$ 的“替身” $l(\mathbf{h})$ 的最小值点来实现, 即

$$\mathbf{h}_{gn} = \arg \min_{\mathbf{h}} l(\mathbf{h}) \quad (9-14)$$

容易知道, 函数 $l(\mathbf{h})$ 的驻点为 $-(J^T(\mathbf{x}) J(\mathbf{x}))^{-1} J^T(\mathbf{x}) \mathbf{f}(\mathbf{x})$ 。这里我们需要附加一个额外条件: 雅可比矩阵 $J(\mathbf{x})$ 是列满秩的, 即 $\text{rank}(J(\mathbf{x})) = n$; 这样的话: $J^T(\mathbf{x}) J(\mathbf{x})$ 必为正定矩阵, 因而 $l(\mathbf{h})$ 的驻点 $-(J^T(\mathbf{x}) J(\mathbf{x}))^{-1} J^T(\mathbf{x}) \mathbf{f}(\mathbf{x})$ 也是 $l(\mathbf{h})$ 的全局最小值点(这个结论留作练习请读者完成证明), 即,

$$\mathbf{h}_{gn} = -(J^T(\mathbf{x}) J(\mathbf{x}))^{-1} J^T(\mathbf{x}) \mathbf{f}(\mathbf{x}) \quad (9-15)$$

式 9-15 便是在点 \mathbf{x} 处由高斯牛顿法所确定的更新向量的计算方式(式 9-15 中的更新向量记为了 \mathbf{h}_{gn} , 其下标 gn 就是 Gauss-Newton 的缩写)。

在 9.1.2 节中, 我们说二次函数 $l(\mathbf{h})$ (式 9-2)可以用来作为目标函数 f 在 $\mathbf{x}+\mathbf{h}$ 处的“替身”。那时我们说函数 $l(\mathbf{h})$ 中的 \mathbf{c} 和 B 会有不同的具体构造方式。对照高斯牛顿法中使用的具体的 $l(\mathbf{h})$ (式 9-13), 不难发现, 在高斯牛顿法这个解决非线性最小二乘问题的具体方法中, $\mathbf{c} = J^T(\mathbf{x}) \mathbf{f}(\mathbf{x})$, $B = J^T(\mathbf{x}) J(\mathbf{x})$ 。

我们最后再从阻尼法的视角来审视一下高斯牛顿法。阻尼法计算更新向量时所用的目标函数为式 9-4, 而高斯牛顿法计算更新向量所用的目标函数为式 9-14, 对照之下, 不难发现, 高斯牛顿法可以看作是阻尼系数恒为 0 的用于解非线性最小二乘这个具体问题的阻尼法, 是完全没有启用阻尼项的阻尼法。按照我们在介绍阻尼法时所做的分析, 像高斯牛顿法这种直接以“替身”函数 $l(\mathbf{h})$ 的最小值点来作为更新向量的方式并不是很合理, 我们需要引入适当的阻尼项以惩罚由 $l(\mathbf{h})$ 所诱导出的“大的”更新向量。直观地, 我们可以在高斯牛顿法计算更新向量的优化目标函数(式 9-14)的基础之上引入阻尼项, 这便是下一节要介绍的列文伯格-马夸尔特法(Levenberg-Marquardt method)。

9.2.3 列文伯格-马夸尔特法

列文伯格-马夸尔特法所要解决的问题与 9.2.2 节中所讲述的高斯牛顿法是相同的。它与高斯牛顿法相比，唯一的不同之处在于：在计算目标函数 f （式 9-8）在 \mathbf{x} 点处的更新向量时引入了阻尼项，即更新向量 \mathbf{h}_{lm} 通过下式来确定，

$$\mathbf{h}_{lm} = \arg \min_{\mathbf{h}} \left\{ l(\mathbf{h}) + \frac{1}{2} \mu \mathbf{h}^T \mathbf{h} \right\} \quad (9-16)$$

其中 $l(\mathbf{h})$ 由式 9-13 给出， $\mu > 0$ 为阻尼系数。可以证明，式 9-16 中的目标函数 $l(\mathbf{h}) + \frac{1}{2} \mu \mathbf{h}^T \mathbf{h}$

为凸函数，则其驻点 $-(J^T(\mathbf{x})J(\mathbf{x}) + \mu I)^{-1} J^T(\mathbf{x})\mathbf{f}(\mathbf{x})$ 便为其最小值点，因此，

$$\mathbf{h}_{lm} = -(J^T(\mathbf{x})J(\mathbf{x}) + \mu I)^{-1} J^T(\mathbf{x})\mathbf{f}(\mathbf{x}) \quad (9-17)$$

式 9-17 便是在点 \mathbf{x} 处由列文伯格-马夸尔特法所确定的更新向量的计算方式（式 9-17 中的更新向量记为了 \mathbf{h}_{lm} ，其下标 lm 就是 Levenberg-Marquardt 的缩写）。

列文伯格-马夸尔特法是解决非线性优化问题的阻尼法通用框架在非线性最小二乘这个具体问题上的“实例化”体现，对照式 9-5 和式 9-17 也可以清楚地体会到这一点：只要把通用阻尼法中的 \mathbf{c} 和 B 分别取为 $\mathbf{c}=J^T(\mathbf{x})\mathbf{f}(\mathbf{x})$ 、 $B=J^T(\mathbf{x})J(\mathbf{x})$ ，便得到了列文伯格-马夸尔特这个“实例化”方法。

9.2.2 节中讲述的高斯牛顿法和本节讲述的列文伯格-马夸尔特法都是解决非线性最小二乘问题的具体算法，它们用来计算更新向量的方式分别为式 9-15 和式 9-17。通过对照式 9-15 和式 9-17，可以发现，列文伯格-马夸尔特法要比高斯牛顿法更加稳健，这是因为在高斯牛顿法中，迭代的每一步都要以 $J(\mathbf{x})$ 为列满秩矩阵为前提条件，否则 $J^T(\mathbf{x})J(\mathbf{x})$ 不可逆，迭代将无法进行下去；而列文伯格-马夸尔特法并没有这个要求，不管 $J(\mathbf{x})$ 是否为列满秩矩阵， $J^T(\mathbf{x})J(\mathbf{x}) + \mu I$ 都为正定矩阵，迭代一定可以进行下去。式 9-17 同式 9-15 相比，它额外引入了阻尼项，因此列文伯格-马夸尔特法也被称为“阻尼高斯牛顿法”^[1]。

在 9.1.2 节介绍阻尼法时，我们并没有详细讨论如何设定阻尼系数 μ 的初值、如何设置迭代终止条件等细节问题，这里我们以列文伯格-马夸尔特法这个“实例化”阻尼法为对象，讨论一下这些细节问题。

如何得到 μ 的初始设定值 μ_0 。 μ_0 与初始迭代点 \mathbf{x}_0 处的雅可比矩阵 $J(\mathbf{x}_0)$ 有关，它被设定为，

$$\mu_0 = \tau \cdot \max_i \left\{ [J^T(\mathbf{x}_0)J(\mathbf{x}_0)]_{ii} \right\}, i=1,2,\dots,n \quad (9-18)$$

其中， $[J^T(\mathbf{x}_0)J(\mathbf{x}_0)]_{ii}$ 表示方阵 $J^T(\mathbf{x}_0)J(\mathbf{x}_0)$ 第 i 行 i 列的对角元； τ 是一个需要预先设定的超参数，算法对 τ 的设置并不敏感，但作为一般准则，如果我们确信 \mathbf{x}_0 已经非常接近我们要找的局部极小值点时， τ 可以设置的小一些，比如 $\tau=10^{-6}$ ，否则 τ 需要设置的相对大一些，比如设置为 $\tau=10^{-3}$ 甚至是 $\tau=1$ 。在算法迭代过程中， μ 值可根据算法 9-1 或算法 9-2 进行动态调整。

如何设置迭代终止条件。如果目标函数 f 在当前迭代点 \mathbf{x} 的梯度 $\nabla f(\mathbf{x})=J^T(\mathbf{x})\mathbf{f}(\mathbf{x})$ 已经

为 $\mathbf{0}$ 的话, 说明 \mathbf{x} 已经是 f 的极小值点 (我们不考虑极特殊的鞍点情况) 了, 迭代就需要终止。转化为程序实现, 我们只需要判断以下条件是否满足,

$$\|\nabla f(\mathbf{x})\|_{\infty} \leq \varepsilon_1 \quad (9-19)$$

其中, ε_1 为预先设定的一个很小的正数, $\|\cdot\|_{\infty}$ 为矢量的无穷范数。另外, 如果当前这一步迭代的更新向量已经非常小了, 说明迭代已经“走不动了”, 迭代也需要终止了。转化为程序实现, 这个终止条件被表达为,

$$\|\mathbf{h}_{new}\|_2 \leq \varepsilon_2 (\|\mathbf{x}\|_2 + \varepsilon_2) \quad (9-20)$$

其中, ε_2 为预先设定的一个很小的正数; 当 $\|\mathbf{x}\|_2$ 相对较大时, 更新向量长短的判定 (大致上) 是通过与一个相对量 $\varepsilon_2 \|\mathbf{x}\|_2$ 进行比较而得出的, 而当 $\|\mathbf{x}\|_2$ 非常小时, 更新向量长短的判定 (大致上) 是通过与一个绝对量 ε_2^2 进行比较而得出的。除了以上两个终止条件之外, 为了避免无限循环, 还需要设置一个最大迭代次数限制: 当迭代次数超过 k_{max} 时, 迭代停止。

算法 9-3 给出了完整的列文伯格-马夸尔特法的程序伪码。

算法 9-3: 列文伯格-马夸尔特法

```

begin
     $k := 0; v := 2; \mathbf{x} = \mathbf{x}_0$ 
     $A := J^T(\mathbf{x})J(\mathbf{x}); \mathbf{g} := J^T(\mathbf{x})\mathbf{f}(\mathbf{x})$ 
     $found := (\|\mathbf{g}\|_{\infty} \leq \varepsilon_1); \mu = \tau \cdot \max_i \{A_{ii}\}$ 
    while (not  $found$ ) and ( $k < k_{max}$ )
         $k := k + 1; \mathbf{h}_{lm} := -(A + \mu I)^{-1} \mathbf{g}$ 
        if  $\|\mathbf{h}_{lm}\|_2 \leq \varepsilon_2 (\|\mathbf{x}\|_2 + \varepsilon_2)$ 
             $found := \text{true}$ 
        else
             $\mathbf{x}_{new} := \mathbf{x} + \mathbf{h}_{lm}$ 
             $\rho := (f(\mathbf{x}) - f(\mathbf{x}_{new})) / (l(\mathbf{0}) - l(\mathbf{h}_{lm}))$ 
            if  $\rho > 0$ 
                 $\mathbf{x} := \mathbf{x}_{new}$ 
                 $A := J^T(\mathbf{x})J(\mathbf{x}); \mathbf{g} := J^T(\mathbf{x})\mathbf{f}(\mathbf{x})$ 
                 $found := \|\mathbf{g}\|_{\infty} \leq \varepsilon_1$ 
                 $\mu := \mu \times \max \left\{ \frac{1}{3}, 1 - (2\rho - 1)^3 \right\}; v := 2$ 
            else
                 $\mu := \mu \times v; v := 2 \times v$ 
            end if
        end if
    end while
    return  $\mathbf{x}$ 
end
```

列文伯格-马夸尔特法可以说是目前使用的最为广泛的用于解非线性最小二乘问题的方法。该方法最早于 1944 年由美国数学家肯尼斯·列文伯格 (Kenneth Levenberg) 提出^[6], 后

来唐纳德·马夸尔特于 1963 年又独立提出了该方法^[3]，因此该方法后来被命名为列文伯格-马夸尔特法，以纪念这两位数学家。

最后，我们再来强调一点，不论是高斯牛顿法还是列文伯格-马夸尔特法，从它们计算在点 \mathbf{x} 处的更新向量时所用的方式（式 9-15 和式 9-17）可以看出，我们需要知道的信息包括 $\mathbf{f}(\mathbf{x})$ 和 \mathbf{f} 在点 \mathbf{x} 处的雅可比矩阵 $J(\mathbf{x})$ 。当我们已经有了最小二乘问题的目标函数以后，在点 \mathbf{x} 处的 $\mathbf{f}(\mathbf{x})$ 的值当然是容易知道的，因而具体实现迭代算法的关键就在于我们要得到 $J(\mathbf{x})$ 的表达式。在 10.3.3 中，我们将以相机内参标定这个具体任务为载体，介绍如何推导得到非线性最小二乘问题中的雅可比矩阵 $J(\mathbf{x})$ 。

9.3 习题

- (1) 请证明：在式 9-14 中，如果雅可比矩阵 $J(\mathbf{x})$ 为列满秩矩阵，即 $\text{rank}(J(\mathbf{x}))=n$ ，则 $J^T(\mathbf{x})J(\mathbf{x})$ 必为正定矩阵，同时， $I(\mathbf{h})$ 的驻点 $-\left(J^T(\mathbf{x})J(\mathbf{x})\right)^{-1}J^T(\mathbf{x})\mathbf{f}(\mathbf{x})$ 也是 $I(\mathbf{h})$ 的全局最小值点。
- (2) 请证明：式 9-16 中的目标函数为凸函数。

参考文献

- [1] K. Madsen, H.B. Nielsen, and O. Tingleff, *Methods for Non-linear Least Squares Problems*, Technical University of Denmark, 2004.
- [2] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.
- [3] D. Marquardt, “An algorithm for least squares estimation on nonlinear parameters,” *SIAM J. Appl. Math.*, vol. 11, pp. 431-441, 1963.
- [4] H.B. Nielsen, “Damping parameter in Marquardt’s method”, Technical Report, Technical University of Denmark, 1999.
- [5] Donald W. Marquardt, https://en.wikipedia.org/wiki/Donald_Marquardt
- [6] K. Levenberg, “A method for the solution of certain problems in least squares,” *Quart. Appl. Math.* 2, pp 164–168, 1944.

第 10 章 相机成像模型与内参标定

本章将首先介绍针孔相机成像模型。进而会详细讲解如何对给定相机进行内参标定，以得到该相机成像模型中的内参数值。

10.1 不考虑镜头畸变的成像模型

为了便于分析，在计算机视觉领域，最常使用的相机成像模型为针孔（pin-hole）相机模型。在该模型下，如果不考虑镜头畸变的话，世界坐标系中的一点 \mathbf{p}_w 到其在图像上的像点 $\mathbf{u}=(u, v)^T$ 的成像流程可以被图 10-1 来表示。接下来我们将详细建模这个成像流程。

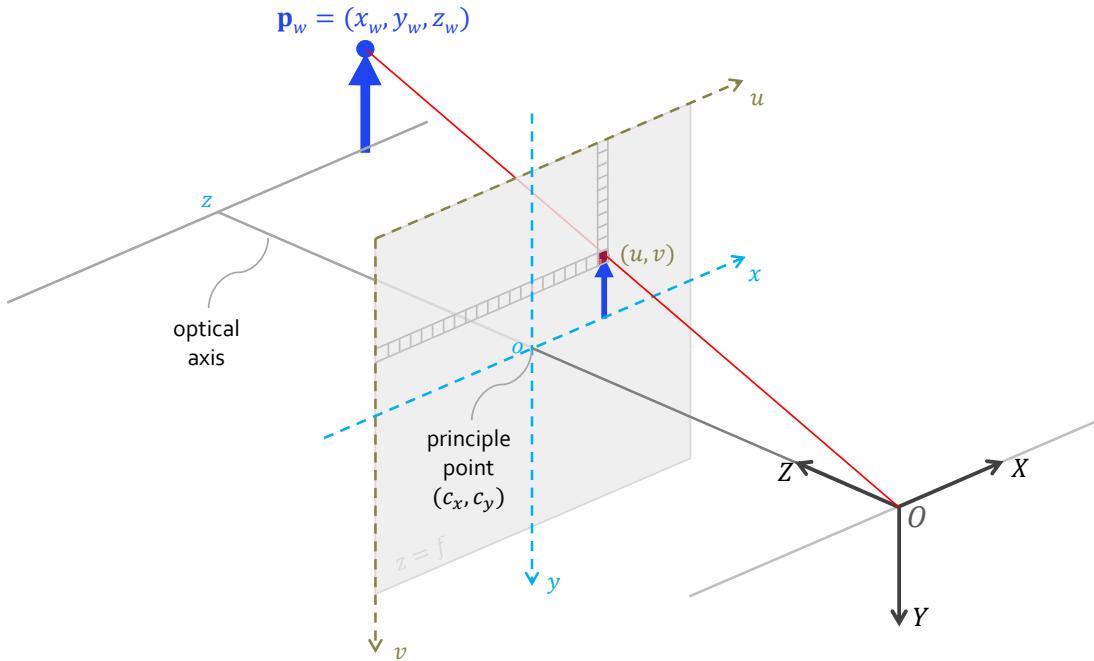


图 10-1：针孔相机模型。

设三维世界坐标系中的 \mathbf{p}_w 点坐标为 $\mathbf{p}_w=(x_w, y_w, z_w, 1)^T$ (齐次坐标表示)。相机自身也建立了一个三维坐标系，称为**相机坐标系**，这个坐标系以相机的光心 O 为坐标原点，其三个正交坐标轴被表示为 X 、 Y 和 Z 。由于相机坐标系和世界坐标系之间的关系可以通过旋转和平移来刻画，因此，点 \mathbf{p}_w 在相机坐标系下的坐标 \mathbf{p}_c 可被表达为，

$$\mathbf{p}_c = [R_{3 \times 3} \ \mathbf{t}_{3 \times 1}] \mathbf{p}_w = [R_{3 \times 3} \ \mathbf{t}_{3 \times 1}] \begin{bmatrix} x_w \\ y_w \\ z_w \\ 1 \end{bmatrix} \triangleq \begin{bmatrix} x_c \\ y_c \\ z_c \end{bmatrix} \quad (10-1)$$

其中 R 为正交矩阵且 $\det(R)=1$ ， \mathbf{p}_c 为非齐次坐标表示。

相机的成像平面为一个到光心 O 的距离为 f 、垂直于 Z 轴且在 Z 轴正向的一个平面。在这个平面上，我们定义**成像平面坐标系**，这是一个二维平面坐标系。该坐标系的原点 o 为 Z 轴与该平面的交点，其 x -轴与 X 轴方向相同，其 y -轴与 Y 轴方向相同。显然 o 即是相机光心 O 在成像平面上的像。根据相似三角形的知识容易知道， \mathbf{p}_c 在成像平面坐标系下的投影点的坐标 $(x, y)^T$ 为，

$$\begin{cases} x = f \frac{x_c}{z_c} \\ y = f \frac{y_c}{z_c} \end{cases} \quad (10-2)$$

其齐次坐标表达为，

$$\begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \frac{1}{z_c} \begin{bmatrix} f & 0 & 0 \\ 0 & f & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_c \\ y_c \\ z_c \end{bmatrix} \quad (10-3)$$

为了后续的推导，我们还需要引入一个**归一化成像平面坐标系**，这个坐标系的定义和构建方式与成像平面坐标系类似，唯一的一点区别是归一化成像平面到光心 O 的距离为单位“1”，即 $f=1$ ，这也是为什么该平面称为“归一化”成像平面。这个单位“1”是个无量纲的数值。借助于式 10-3，容易知道，令 $f=1$ ，便可以得到点 \mathbf{p}_c 在归一化成像平面上投影点的齐次坐标 $(x_n, y_n, 1)^T$ ，

$$\begin{bmatrix} x_n \\ y_n \\ 1 \end{bmatrix} = \frac{1}{z_c} \begin{bmatrix} x_c \\ y_c \\ z_c \end{bmatrix} \quad (10-4)$$

可以看出，点 $\mathbf{p}_c=(x_c, y_c, z_c)^T$ 在归一化成像平面坐标系下的投影坐标只与 x_c 、 y_c 和 z_c 三者之间的比值有关，与它们的绝对数值大小以及单位都无关。

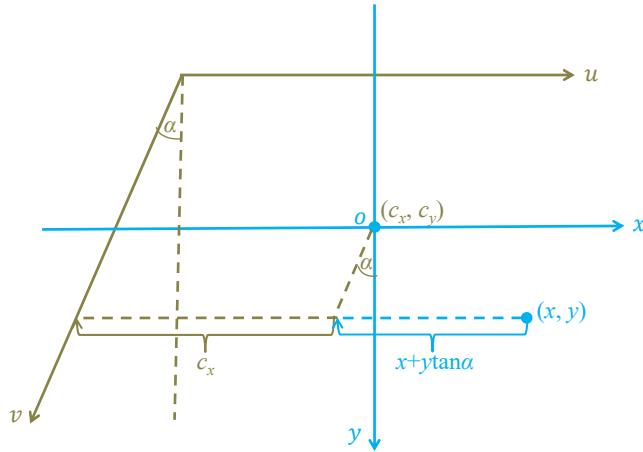


图 10-2：成像平面坐标系与像素坐标系关系示意图。图中，绿色部分代表的是像素坐标系下的信息，蓝色部分代表的是成像平面坐标系下的信息。

最后，我们对成像平面进行“像素化”，得到成像平面上的点在**像素坐标系**之下的坐标。如图 10-2 所示，像素坐标系($u-v$)和成像平面坐标系($x-y$)都在同一平面上，且成像平面坐标

系的原点 o 在像素坐标系下的像素坐标为 $(c_x, c_y)^T$, 这个坐标称为主点 (principal point) 坐标, 主点坐标实际上就是相机光心 O 在最终图像上的成像位置。像素坐标系的 u 轴与 x 轴平行。而由于感光器件制造工艺可能不完美的原因, v 轴与 u 轴可能并不一定是严格垂直的, 因此, 在成像模型中我们需要建模 u 与 v 之间的不垂直特性, 我们把 v 轴与 y 轴方向的夹角记为 α 。设成像器件上一个像素的物理宽度为 dx 、高度为 dy 。从图 10-2 所示的关系图中, 容易知道成像平面坐标系下的一点 $(x, y)^T$ 在像素坐标系下的坐标 $(u, v)^T$ 为,

$$\begin{cases} u = c_x + \frac{x + y \tan \alpha}{dx} \\ v = c_y + \frac{y}{dy} \end{cases} \quad (10-5)$$

写成矩阵乘法的形式为,

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} \frac{1}{dx} & \frac{\tan \alpha}{dx} & c_x \\ 0 & \frac{1}{dy} & c_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \quad (10-6)$$

通过联立等式 10-1、10-3 和 10-6, 我们便可以得到从世界坐标系下的一点 $\mathbf{p}_w = (x_w, y_w, z_w, 1)^T$ 到它最终在成像平面像素坐标系下的位置 $(u, v, 1)^T$ 的完整映射过程,

$$\begin{aligned} \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} &= \begin{bmatrix} \frac{1}{dx} & \frac{\tan \alpha}{dx} & c_x \\ 0 & \frac{1}{dy} & c_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \begin{bmatrix} \frac{1}{dx} & \frac{\tan \alpha}{dx} & c_x \\ 0 & \frac{1}{dy} & c_y \\ 0 & 0 & 1 \end{bmatrix} \frac{1}{z_c} \begin{bmatrix} f & 0 & 0 \\ 0 & f & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_c \\ y_c \\ z_c \end{bmatrix} \\ &= \frac{1}{z_c} \begin{bmatrix} \frac{1}{dx} & \frac{\tan \alpha}{dx} & c_x \\ 0 & \frac{1}{dy} & c_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} f & 0 & 0 \\ 0 & f & 0 \\ 0 & 0 & 1 \end{bmatrix} [R_{3 \times 3} \ t_{3 \times 1}] \begin{bmatrix} x_w \\ y_w \\ z_w \\ 1 \end{bmatrix} = \frac{1}{z_c} \begin{bmatrix} \frac{f}{dx} & \frac{f \tan \alpha}{dx} & c_x \\ 0 & \frac{f}{dy} & c_y \\ 0 & 0 & 1 \end{bmatrix} [R_{3 \times 3} \ t_{3 \times 1}] \begin{bmatrix} x_w \\ y_w \\ z_w \\ 1 \end{bmatrix} \\ &\triangleq \frac{1}{z_c} \begin{bmatrix} f_x & s & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} [R_{3 \times 3} \ t_{3 \times 1}] \begin{bmatrix} x_w \\ y_w \\ z_w \\ 1 \end{bmatrix} \end{aligned} \quad (10-7)$$

在式 10-7 最后一步中, 我们令 $f_x = f/dx$ 、 $f_y = f/dy$ 和 $s = f_x \tan \alpha$ 。 f_x 、 f_y 称为相机在 x 方向和 y 方向的焦距, s 刻画了成像器件两个坐标轴的不垂直性, 称为扭曲参数 (skew parameter)。容易

知道, f_x 、 f_y 和 s 都只与相机的物理属性有关, 都是相机的内参数, 所以矩阵 $K = \begin{bmatrix} f_x & s & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix}$ 称

为相机的内参矩阵。式 10-7 中的 R 和 t 都与相机相对于世界坐标系的位姿有关, 因此它们是相机的外参数。

我们还需要明确一下成像建模过程中各个变量的单位的问题。 \mathbf{p}_w 点的坐标使用的是物理长度单位，比如毫米、厘米。假定在成像模型中的物理长度所采用的单位都为毫米，则 \mathbf{t} 、 \mathbf{p}_c 、 $(x, y)^T$ 、 f 的单位都为毫米； dx 、 dy 的单位为毫米/像素； f_x 、 f_y 和 s 的单位都为像素，主点坐标 $(c_x, c_y)^T$ 的单位为像素。内参矩阵 K 中的参数都以像素为单位。

当读者用一些现有的工具包来执行相机标定任务时，可能会发现有些工具包（比如 Matlab）所用的成像模型考虑了扭曲参数 s ；但也有些工具包（比如 OpenCV）并不考虑这个参数，即认为 $s=0$ 。对于绝大多数现代相机而言，把扭曲参数 s 简单地认为取值为 0 也是合理的，这是因为现代相机的制作工艺精度较高，可以认为它们的成像器件几乎是完美的长方形，这就意味着式 10-7 中的 $\alpha=0$ ，因此相应地 $s=0$ 。为了使论述的问题尽可能简洁，本书后面再讨论相机模型时，将不再考虑扭曲参数 s ，即认为 $s=0$ 。因此，本书中的内参矩阵 K 为，

$$K = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \quad (10-8)$$

我们可以用矢量形式来表达点的坐标，从而成像流程式 10-7 可以被简洁地表达为，

$$\mathbf{u} = \frac{1}{z_c} K [R_{3 \times 3} \ \mathbf{t}_{3 \times 1}] \mathbf{p}_w \quad (10-9)$$

其中， \mathbf{p}_w 为三维世界坐标系中一点的归一化齐次坐标表示， \mathbf{u} 为点 \mathbf{p}_w 在最终像素坐标系下的归一化齐次像素坐标。

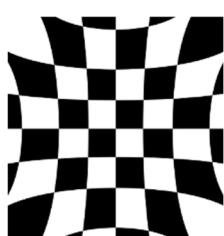
根据式 10-4 可知 $\frac{1}{z_c} [R_{3 \times 3} \ \mathbf{t}_{3 \times 1}] \mathbf{p}_w = (x_n, y_n, 1)^T$ ，结合式 10-9 我们会有，

$$\mathbf{u} = K \begin{pmatrix} x_n \\ y_n \\ 1 \end{pmatrix} \quad (10-10)$$

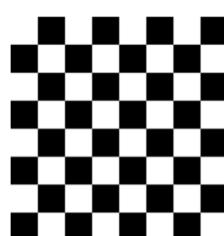
上式表达了归一化成像平面坐标系下的点 $(x_n, y_n, 1)^T$ 与其在像素坐标系下的对应点 $\mathbf{u}=(u, v, 1)^T$ 之间的关系。

10.2 考虑镜头畸变的成像模型

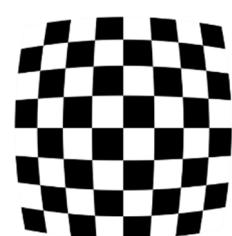
式 10-9 所描述的成像模型没有考虑相机镜头的畸变，这是因为理想的针孔相机模型中是没有镜头的。为了更加准确地建模真实相机的成像流程，我们需要在式 10-9 的基础上额外考虑镜头所引起的径向畸变（radial distortion）和切向畸变（tangential distortion）^[1, 2]。



(a)



(b)



(c)

图 10-3：镜头的径向畸变示意图。（b）镜头没有发生畸变时所成图像；（a）镜头发生了“枕型”畸变，成像点到光轴的距离会变大；（c）镜头发生了“桶形”畸变，成像点到光轴的距离会变小。

当光线在透镜边缘的弯曲程度大于在透镜光学中心的弯曲程度时，就会发生径向畸变。镜头越小，畸变程度越大。镜头的径向畸变又分为两种：枕型畸变（pincushion distortion）和桶形畸变（barrel distortion），又分别称为正径向畸变（positive radial distortion）和负径向畸变（negative radial distortion）。我们可以通过图 10-3 来直观地理解一下这两类径向畸变。10-3（b）中的“图像”是在镜头不存在畸变时，得到的理想图像；10-3（a）中，镜头发生了“枕型”畸变，成像点到光轴的距离会变大；10-3（c）中，镜头发生了“桶形”畸变，成像点到光轴的距离会变小。我们平时见到的广角鱼眼镜头就利用了“桶形”畸变的特性，这种镜头通过减小成像点到成像中心的距离，有效增大了成像的视场范围。

对镜头径向畸变现象的建模是在归一化成像平面坐标系下进行的。假设在没有镜头径向畸变的情况下，归一化成像平面坐标系下的一点为 $(x_n, y_n)^T$ ；当发生了径向畸变之后，这一点被映射到了 $(x_{dr}, y_{dr})^T$ ，则 $(x_{dr}, y_{dr})^T$ 与 $(x_n, y_n)^T$ 之间的关系可被表达为，

$$\begin{cases} x_{dr} = x_n(1+k_1r^2 + k_2r^4 + k_3r^6) \\ y_{dr} = y_n(1+k_1r^2 + k_2r^4 + k_3r^6) \end{cases} \quad (10-11)$$

其中， $r^2 = x_n^2 + y_n^2$ ， k_1 、 k_2 和 k_3 为待定参数。

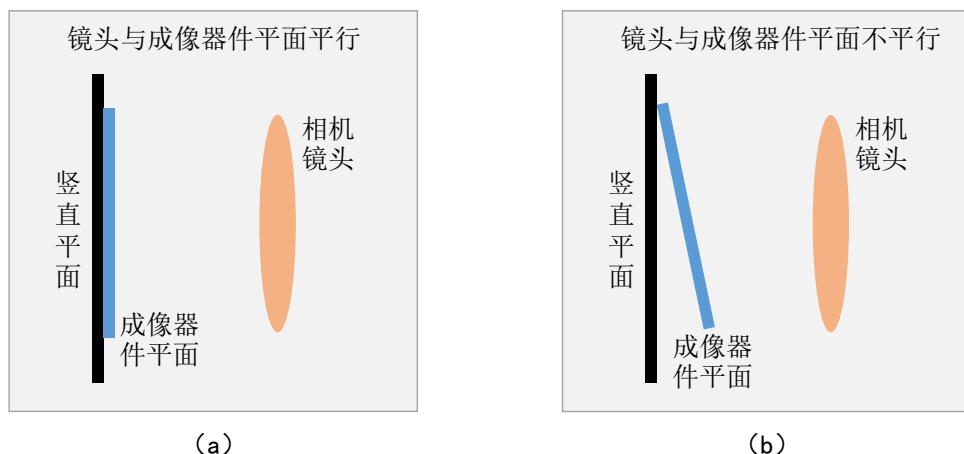


图 10-4：镜头的切向畸变示意图。（a）当相机镜头与成像器件平面严格平行时，所成图像不会发生切向畸变；（b）当相机镜头与成像器件平面不平行时，所成图像会发生切向畸变。

另外一种镜头畸变称为切向畸变。如图 10-4 所示，当镜头平面与成像器件平面不是严格平行时，就会产生切向畸变。同径向畸变一样，镜头的切向畸变也是在归一化成像坐标系下来进行建模的。假设在没有镜头切向畸变的时候，归一化成像平面坐标系下的一点为 $(x_n, y_n)^T$ ；当发生了切向畸变之后，这一点被映射到了 $(x_{dt}, y_{dt})^T$ ，则 $(x_{dt}, y_{dt})^T$ 与 $(x_n, y_n)^T$ 之间的关系可被表达为，

$$\begin{cases} x_{dt} = x_n + (2\rho_1 x_n y_n + \rho_2 (r^2 + 2x_n^2)) \\ y_{dt} = y_n + (2\rho_2 x_n y_n + \rho_1 (r^2 + 2y_n^2)) \end{cases} \quad (10-12)$$

其中, $r^2 = x_n^2 + y_n^2$, ρ_1 、 ρ_2 是和切向畸变相关的两个参数。

我们当然可以在成像模型中同时考虑镜头的径向畸变和切向畸变。假设在没有镜头畸变时, 归一化成像平面坐标系下的一点为 $(x_n, y_n)^T$; 当发生了径向与切向畸变之后, 这一点被映射到了 $(x_d, y_d)^T$, 则 $(x_d, y_d)^T$ 与 $(x_n, y_n)^T$ 之间的关系可被表达为,

$$\begin{cases} x_d = x_n (1 + k_1 r^2 + k_2 r^4) + 2\rho_1 x_n y_n + \rho_2 (r^2 + 2x_n^2) + x_n k_3 r^6 \\ y_d = y_n (1 + k_1 r^2 + k_2 r^4) + 2\rho_2 x_n y_n + \rho_1 (r^2 + 2y_n^2) + y_n k_3 r^6 \end{cases} \quad (10-13)$$

其中, k_1 、 k_2 、 ρ_1 、 ρ_2 、 k_3 称为相机的畸变参数, 它们当然也是相机的内参数。

本书后续部分讨论中所使用的相机镜头畸变模型由式 10-13 给出, 该模型可以很好地对绝大多数普通相机镜头的畸变情况进行建模。当然, 在研究领域, 还有一些更加复杂的镜头畸变模型, 比如薄棱镜模型、倾斜模型、鱼眼模型等, 但为了保持叙述的简洁性, 本书就不再具体讨论它们了。

相机的镜头畸变是在归一化成像平面坐标系之下被进行建模的。根据式 10-10 中的结论, 归一化成像平面坐标系下的点乘上内参矩阵 K 就得到了最终的像素坐标系下点的坐标。

我们在成像模型 10-9 的基础上, 引入一个畸变算子 \mathcal{D} 来表示在归一化成像平面坐标系之下发生的镜头畸变。这样, 考虑了镜头畸变的完整的相机成像模型为,

$$\mathbf{u} = K \cdot \mathcal{D} \left\{ \frac{1}{z_c} [R \ t]_{3 \times 4} \mathbf{p}_w \right\} \quad (10-14)$$

其中, 畸变算子 \mathcal{D} 表示在归一化成像平面坐标系下把无畸变的投影点进行由式 10-13 所示的畸变映射。

10.3 相机内参标定

10.3.1 相机内参标定算法的基本流程

在 10.2 节中, 我们对相机成像过程的完整流程进行了建模。在 10.3 中, 我们将解决这样一个问题: 对于一个给定的相机, 如何能知道刻画它的相机模型的内参数值, 即成像模型式 10-14 中的待定参数 $\{f_x, f_y, c_x, c_y, k_1, k_2, \rho_1, \rho_2, k_3\}$ 具体的值。求解相机内参数的过程称为**相机的内参标定**, 这往往是用相机来执行空间测量任务前的必要操作。那么应该如何来解决这个问题呢?

假设我们知道一组空间三维点的世界坐标 $\{\mathbf{p}_{wi}\}_{i=1}^n$ ($\mathbf{p}_{wi} \in \mathbb{R}^{4 \times 1}$ 为空间三维点 i 的齐次坐标), 并且知道与它们对应的相机图像上点集的像素坐标 $\{\mathbf{u}_i\}_{i=1}^n$ ($\mathbf{u}_i \in \mathbb{R}^{3 \times 1}$ 为空间点 i 在图像上二维

投影像素点的齐次坐标)。根据成像模型式 10-14, 我们可以得到一组关于相机未知参数(包括内参数与外参数)的方程。通过解这个方程, 便可得到相机模型中的待定参数值。所有现有的相机标定方案都是基于这一基本思想来设计的。



图 10-5: 张正友, 博士, 男, 汉族, 浙江温岭市人, 1965 年 8 月 1 日出生。先后毕业于浙江大学、法国南锡(Nancy)大学、法国巴黎第十一大学, ACM Fellow, IEEE Fellow。是世界知名的人工智能和机器人科学家, 在多个领域都有开创性的贡献。2013 年, 因为“张氏标定法”, 张正友获得了 IEEE Helmholtz 时间考验奖, 目前这一相机标定法在全世界被普遍采用。2021 年 1 月 8 日张正友受聘腾讯历史上最高专业职级——17 级研究员/杰出科学家。

在众多的相机内参标定方案中, 目前使用最为广泛的便是张正友(图 10-5)平面标定法^[3], 这是由于该方法具有操作简便、标定精度高的优点。本节后面的内容将详细介绍该内参标定方案的具体细节。

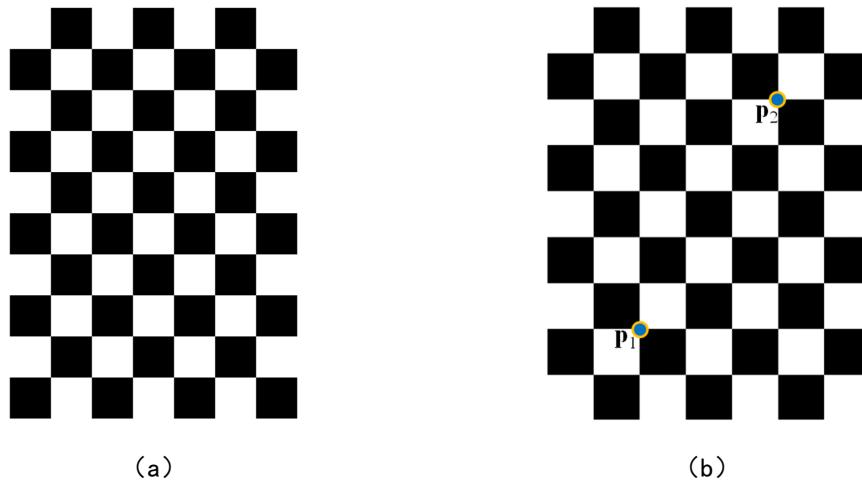


图 10-6: (a) 平面标定法中常用的棋盘格标定板, 以黑白格为单位, 其行数与列数的奇偶性必须不同; (b) 此棋盘格标定板有 9 行 7 列, 这样的话我们无法对两个交叉点 p_1 与 p_2 进行区分, p_2 可能是与 p_1 不一样的一个点, 也可能是在标定板旋转了 180 度之后的结果, 即 p_2 就是 p_1 。

首先要制作一张平面标定板, 其上要印有易于检测的周期性图像模式。最常使用的标定

板图像模式是黑白棋盘格模式，它由周期性排布的等边长的黑白正方形块组成，如图 10-6 (a) 所示。棋盘格标定板上的交叉点便是我们将要用于相机标定的三维特征点。为了使标定板上交叉点的坐标不具有歧义性，标定板上行与列的黑白块数目的奇偶性不能相同。即，如果它有奇数列（以黑白块为单位），它就需要有偶数行；反之，如果它有偶数列，它就需要有奇数行。否则，标定板上交叉点的坐标会出现歧义性。图 10-6 (b) 展示了棋盘格标定板的行数与列数都是奇数的情况下，可以看到，在这种情况下，我们无法对两个交叉点 \mathbf{p}_1 与 \mathbf{p}_2 进行区分。

在标定板平面之上选定好坐标原点，同时规定好 X 轴和 Y 轴方向，之后再按照右手法则确定出垂直于标定板平面的 Z 轴，这样我们便建立了一个基于标定板平面的三维世界坐标系，如图 10-7 所示。显然，为了便于操作，原点要选在某一交叉点上， X 轴与 Y 轴要沿着黑白格的边的方向。在这个三维世界坐标系下，如果我们预先测量好了每个黑白块的边长，那么可以容易得到标定板上所有交叉点的世界坐标 $\{\mathbf{p}_j\}_{j=1}^n$ ，其中 $\mathbf{p}_j = (x_j, y_j, 0, 1)^T$ 为交叉点的三维齐次坐标， n 为标定板上交叉点的个数（图 10-7 中所示的标定板有 54 个交叉点）。由于交叉点 $\mathbf{p}_j (j=1, \dots, n)$ 位于标定板平面之上，因此 \mathbf{p}_j 坐标的 Z 值为 0。

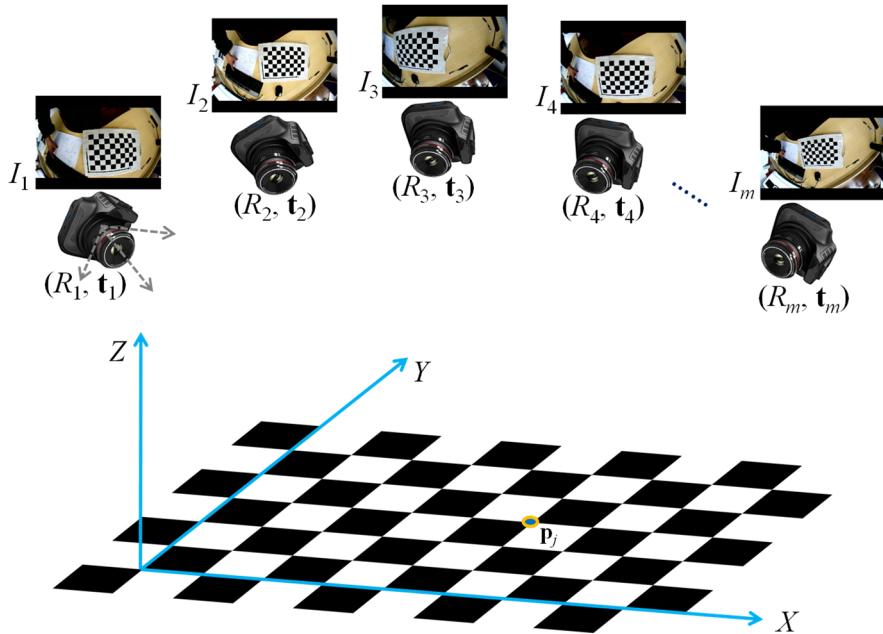


图 10-7：相机内参标定操作示意图。改变相机的位置，并在每个位置 i 处拍摄下标定板的一张图像 I_i ，总计拍摄 m 张图像。 (R_i, t_i) 代表在位置 i 处相机在由标定板平面所确立的世界坐标系下的位姿。

准备好标定板以后，将待标定相机放置在不同的合适位置，并在每个位置处拍摄一张标定板照片，如图 10-7 所示。假设总共在 m 个不同位置处拍摄了 m 张标定板图像 $\{I_i\}_{i=1}^m$ 。在位置 i 处，相机相对于由标定板所建立的世界坐标系的位姿记为 (R_i, t_i) ，即当相机在位置 i 处

时，世界坐标系下的点 \mathbf{p}_w （归一化齐次坐标形式）在相机坐标系下的坐标为 $[R_i \ \mathbf{t}_i] \mathbf{p}_w$ 。对图像 $I_i (i=1,..,m)$ ，我们用特征点检测算法在其上检测出图像空间中的交叉点；通过简单的后处理后，可以建立起标定板平面上的物理交叉点与 I_i 上图像空间中的交叉点之间的对应关系 $\{\mathbf{p}_j \leftrightarrow \mathbf{u}_{ij}\}_{j=1}^n$ ，即标定板平面上的物理交叉点 \mathbf{p}_j 在图像 I_i 中的像为 \mathbf{u}_{ij} 。

考虑标定板上的交叉点 \mathbf{p}_j ，如果给定相机内参（内参矩阵 K 以及镜头畸变系数）和相机在位置 i 处的位姿 (R_i, \mathbf{t}_i) ，根据成像模型式 10-14，它在 I_i 上的投影点应该为

$K \cdot \mathcal{D} \left\{ \frac{1}{z_{cij}} [R_i \ \mathbf{t}_i]_{3 \times 4} \mathbf{p}_j \right\}$ ；而我们观测到的它在 I_i 上的实际投影点为 \mathbf{u}_{ij} 。我们定义，

$$\frac{1}{2} \left\| K \cdot \mathcal{D} \left\{ \frac{1}{z_{cij}} [R_i \ \mathbf{t}_i]_{3 \times 4} \mathbf{p}_j \right\} - \mathbf{u}_{ij} \right\|_2^2 \quad (10-15)$$

为 \mathbf{p}_j 在 I_i 上的重投影误差 (reprojection error)，即重投影误差表征的是在某组参数下根据理论模型计算出来的理论投影位置与实际观测到的投影位置之间的误差。式 10-15 中的 $\frac{1}{2}$ 只是为了后续便于求导而加上去的。更进一步，我们可以得到标定板上全体交叉点 $\{\mathbf{p}_j\}_{j=1}^n$ 在全部标定板图像 $\{I_i\}_{i=1}^m$ 上的重投影误差之和，

$$e = \sum_{i=1}^m \sum_{j=1}^n \frac{1}{2} \left\| K \cdot \mathcal{D} \left\{ \frac{1}{z_{cij}} [R_i \ \mathbf{t}_i] \mathbf{p}_j \right\} - \mathbf{u}_{ij} \right\|_2^2 \quad (10-16)$$

容易理解， e 实际上是相机模型中待定参数集合 Θ 的函数，集合 Θ 包含了相机的内参数以及相机在各个位置上相对于世界坐标系的外参数（位姿），即 $\Theta = \{f_x, f_y, c_x, c_y, k_1, k_2, \rho_1, \rho_2, k_3, \{R_i\}_{i=1}^m, \{\mathbf{t}_i\}_{i=1}^m\}$ 。我们进一步把 e 明确写为 Θ 的函数，即，

$$e(\Theta) = \sum_{i=1}^m \sum_{j=1}^n \frac{1}{2} \left\| K \cdot \mathcal{D} \left\{ \frac{1}{z_{cij}} [R_i \ \mathbf{t}_i] \mathbf{p}_j \right\} - \mathbf{u}_{ij} \right\|_2^2 \quad (10-17)$$

显然，相机模型参数越趋近于正确，相应得到的重投影误差之和 e 就会越小。因此，我们可以把相机参数的求解问题转换为求函数 $e(\Theta)$ 的最小值点的问题，即要求解如下最优化问题，

$$\Theta^* = \arg \min_{\Theta} \sum_{i=1}^m \sum_{j=1}^n \frac{1}{2} \left\| K \cdot \mathcal{D} \left\{ \frac{1}{z_{cij}} [R_i \ \mathbf{t}_i] \mathbf{p}_j \right\} - \mathbf{u}_{ij} \right\|_2^2 \quad (10-18)$$

式 10-18 的最优解 Θ^* 便是待标定相机的内外参数集合。

细心的读者可能会注意到，一对标定板交叉点与其像点的对应关系 $\mathbf{p}_j \leftrightarrow \mathbf{u}_{ij}$ 可以提供两条独立约束，那么如果标定板上的交叉点足够多（也就是 n 足够大），能否只需要拍摄一张标定板图像就能够执行相机标定任务了？答案是否定的^[4]！我们先来考虑一下镜头畸变。镜头畸变只是建模了归一化平面内部的几何变换关系，它有 5 个自由度 ($k_1, k_2, \rho_1, \rho_2, k_3$)。因此，在相机模型中其他参数已知的情况下，只需要一张标定板图像上的 3 个点对关系便可以

唯一地把与镜头畸变有关的 5 个参数确定下来。如果我们不考虑与镜头畸变有关的 5 个参数，拍摄一张标定板图像时，相机模型中待定参数的个数是 10，包括 4 个内参和 6 个外参（三维空间中的相机位姿有 6 个自由度），而此时标定板平面和图像平面之间满足射影变换关系。根据 3.1.5 节中的知识可知，平面间的射影变换的自由度为 8，并且可以通过 4 个有效对应点对关系唯一确定（见 5.1.1 节），即即使有再多的点对关系，它们能提供的约束信息从本质上来说与 4 个有效点对是一样多的，并不会提供更多的约束。因此，如果只有一张标定板图像的话，即使它上面有很多个交叉点，但它对相机模型能够提供的独立约束的个数也就是 8 个，当然，基于这些约束信息并不能唯一解出此时的 10 个待定相机模型参数。假设标定板上交叉点的个数不少于 4 个，如果有两张标定板图像的话，便可提供 16 个约束，这时的待定参数正好也为 16 个（4 个内参和代表两组相机位姿的 12 外参），因此便可以唯一地解出所有的待定相机参数。这样我们便知，为了要执行相机内参标定，至少需要拍摄两张标定板图像。但在实际操作中，由于要考虑到噪声的存在，还要考虑到解的稳定性，一般往往需要拍摄 10~20 张标定板图像。

10.3.2 三维空间旋转的轴角表达

在具体介绍如何对相机模型参数进行求解之前，我们还有一个关键问题需要解决一下，那就是三维欧氏空间中旋转的表示问题。在目标函数式 10-18 中，拍摄第 i 张标定板图像时，相机坐标系相对于由标定板平面所建立的世界坐标系来说的位姿被表达为 (R_i, \mathbf{t}_i) ，其中 R_i 是用来刻画旋转的特殊正交阵（见 3.3 节中有关三维空间中线性几何变换的介绍）。这种以旋转矩阵来表达三维空间旋转的方法，会给迭代优化算法的实现带来很大困难，因为这会迫使我们需要引入额外的约束来确保与旋转有关的这 9 个优化变量确实能够组成一个能有效表达三维空间旋转的特殊正交矩阵，导致该问题会变为比较困难的有约束优化问题。我们知道，三维欧氏空间中的旋转有 3 个自由度，那么是否存在一种三维空间旋转的三维向量表示，而且这个表示向量中的三个数是相互独立的而不是耦合在一起的？如果存在这样的旋转表示方式的话，那么以三维空间旋转为优化变量的优化问题就变成了相对简单的无约束优化问题。幸运的是，这种表达方式是存在的，它就是三维欧氏空间旋转的轴角（axis-angle）表达。

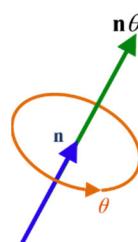


图 10-8：三维空间旋转的轴角表示，单位向量 \mathbf{n} 为旋转轴， $\theta > 0$ 为按右手法则绕轴 \mathbf{n} 旋转的角度，该旋转的轴角表示便为 $\mathbf{n}\theta$ 。

不难想象，任何一个三维欧氏空间中的旋转都可以表达成绕某一个旋转轴 $\mathbf{n} \in \mathbb{R}^{3 \times 1}$

($\|\mathbf{n}\|_2 = 1$) 按右手法则逆时针旋转 θ ($\theta > 0$, 以弧度为单位) 的形式, 只要 \mathbf{n} 和 θ 确定了, 它们所代表的旋转会被唯一确定 (图 10-8)。我们可以用三维矢量 $\mathbf{d}=\theta\mathbf{n}$ 来表达由 \mathbf{n} 和 θ 所

确定的旋转, 这样 $\mathbf{n}=\frac{\mathbf{d}}{\|\mathbf{d}\|_2}$, $\theta=\|\mathbf{d}\|_2$; 这种三维空间旋转的表达方式称为轴角 (axis-angle)。

既然旋转矩阵和轴角都可用来表达三维欧氏空间中的旋转, 那么这两种表达方式之间是否可以进行相互转化呢? 答案是肯定的! 如果一个三维旋转的轴角表达为 $\theta\mathbf{n}$, $\mathbf{n}=(n_1, n_2, n_3)^T$, 且 $\|\mathbf{n}\|_2 = 1$, 那么表达该旋转的旋转矩阵为,

$$R=\cos\theta \cdot I + (1-\cos\theta)\mathbf{n}\mathbf{n}^T + \sin\theta \cdot \hat{\mathbf{n}} \quad (10-19)$$

其中, $I \in \mathbb{R}^{3 \times 3}$ 为单位矩阵, $\hat{\mathbf{n}} = \begin{bmatrix} 0 & -n_3 & n_2 \\ n_3 & 0 & n_1 \\ -n_2 & n_1 & 0 \end{bmatrix}$ 。式 10-19 给出了从轴角表示到旋转矩阵

表示的转化方式, 该公式称为罗德里格斯公式 (Rodrigues formula), 最早由法国数学家奥林德·罗德里格斯 (Olinde Rodrigues) 提出; 也有文献认为该公式的提出应该归功于莱昂哈德·欧拉 (Leonhard Euler) [5]。罗德里格斯公式的具体证明见本书附录 F。

基于式 10-19, 我们也可以从一个给定的旋转矩阵 R 得到与之对应的轴角 $\theta\mathbf{n}$ 。由式 10-19 可知,

$$R = \begin{bmatrix} \cos\theta & & \\ & \cos\theta & \\ & & \cos\theta \end{bmatrix} + (1-\cos\theta) \begin{bmatrix} n_1^2 & n_1n_2 & n_1n_3 \\ n_1n_2 & n_2^2 & n_2n_3 \\ n_1n_3 & n_2n_3 & n_3^2 \end{bmatrix} + \sin\theta \begin{bmatrix} 0 & -n_3 & n_2 \\ n_3 & 0 & n_1 \\ -n_2 & n_1 & 0 \end{bmatrix} \quad (10-20)$$

对上式两端进行求迹操作, 得到,

$$\text{tr}(R) = 3\cos\theta + (1-\cos\theta)(n_1^2 + n_2^2 + n_3^2) = 1 + 2\cos\theta \quad (10-21)$$

其中, $\text{tr}(R)$ 表示矩阵 R 的迹。由式 10-21 可知,

$$\theta = \arccos\left(\frac{\text{tr}(R)-1}{2}\right) \quad (10-22)$$

然后需要确定旋转轴 \mathbf{n} 。旋转轴 \mathbf{n} 应该是在施加旋转变换 R 之后保持不动的向量, 因此它要满足,

$$R\mathbf{n} = \mathbf{n} \quad (10-23)$$

则 \mathbf{n} 为矩阵 R 对应于特征值 1 的单位特征向量。因此, 要计算 \mathbf{n} , 只需要对 R 进行特征值分解, 与特征值 1 相对应的单位特征向量就是 \mathbf{n} 。这其中还有三个细节问题值得我们考虑一下。矩阵 R 一定会有一个特征值是 1 吗? 答案是肯定的, 证明请读者作为练习自行完成。第 2 个问题是, 1 这个特征值所对应的特征子空间 (几何重数) 的维度会不会大于 1? 如果可能存在这种情况的话, 旋转轴 \mathbf{n} 就会有无穷多种可行情况, 会不会给我们确定旋转轴带来困扰? 答案是: 确实存在特征值 1 所对应的特征子空间的维度大于 1 的情况, 也就是说这时候满足条件式 10-23 的 \mathbf{n} 有无穷多中可行情况, 但幸运的是, 这并不会给我们确定 \mathbf{n} 带来困扰。可以证明, 如果 R 有重根 1 的话, 它的三个特征值必然全部为 1; 而这样的 R 所对应的

$$\theta = \arccos\left(\frac{\text{tr}(R)-1}{2}\right) = \arccos\left(\frac{(1+1+1)-1}{2}\right) = 0, \text{ 也就是说这种情况根本就没有旋转, 轴角表达就是 } \mathbf{0}.$$

第 3 个问题是, 与特征值 1 对应的 R 的单位特征向量不是唯一的, 如果 \mathbf{x} 是满足条件的向量, 那么 $-\mathbf{x}$ 一定也满足条件, 那么 \mathbf{n} 到底应该是 \mathbf{x} 呢还是 $-\mathbf{x}$ 呢? 这需要结合着 θ 的取值来确定。根据式 10-22 可知, θ 的取值范围为 $[0, \pi]$ 。假设我们要表达的真实旋转

是绕 \mathbf{x} 轴旋转 $\frac{3}{2}\pi$, 但根据式 10-22, 计算出来的 θ 值为 $\frac{1}{2}\pi$, 则我们需要把旋转轴选为 $-\mathbf{x}$;

也就是说, 我们需要通过检验 $\theta\mathbf{x}$ 与 $-\theta\mathbf{x}$ 哪一个能通过式 10-19 得到给定的 R 来决定旋转轴的选择。

由于轴角表达中的三个分量是相互独立的, 因此在以旋转为优化变量的问题中轴角这种表达方式更适合用来表达三维空间中的旋转。在问题式 10-18 中, 假设与 R_i 对应的轴角为 $\mathbf{d}_i \in \mathbb{R}^{3 \times 1}$, 可以把 R_i 用 \mathbf{d}_i 的映射 $\mathcal{R}(\mathbf{d}_i): \mathbb{R}^{3 \times 1} \rightarrow \mathbb{R}^{3 \times 3}$ 来表示, $\mathcal{R}(\mathbf{d}_i)$ 表示把轴角 \mathbf{d}_i 映射到与其对应的旋转矩阵 (该映射由式 10-19 所确定)。这样, 式 10-18 所描述的优化问题可改写为,

$$\Theta^* = \arg \min_{\Theta} \sum_{i=1}^m \sum_{j=1}^n \frac{1}{2} \left\| K \cdot \mathcal{D} \left\{ \frac{1}{z_{cij}} [\mathcal{R}(\mathbf{d}_i) \mathbf{t}_i] \mathbf{p}_j \right\} - \mathbf{u}_{ij} \right\|_2^2 \quad (10-24)$$

其中, 待优化的参数集合 $\Theta = \{f_x, f_y, c_x, c_y, k_1, k_2, \rho_1, \rho_2, k_3, \{\mathbf{d}_i\}_{i=1}^m, \{\mathbf{t}_i\}_{i=1}^m\}$ 。我们将在 10.3.3 节中讲解如何对参数集合 Θ 进行合理的初始化, 在 10.3.4 节中讲解迭代优化求解式 10-24 的具体细节。

10.3.3 相机成像模型参数的初始估计

接下来考虑如何解式 10-24 这个优化问题。根据第 9 章中的知识, 我们知道, 式 10-24 这个问题是一个非线性最小二乘问题, 同时它也是一个非凸优化问题。对于这样一个问题来说, 实际上很难找到它的全局最优解。但对于大多数实际工程问题来说, 合适的局部最优解 (目标函数在此处取得局部极小值) 往往也是足够用的。我们需要用第 9 章中介绍的方法来迭代求解式 10-24 这个问题, 即从优化变量的一个初始值开始, 按照下降方向不断迭代, 直至最终收敛于目标函数的一个局部极小值点。不难理解, 对于这样的迭代优化算法来说, 优化变量初始值的选取会对最终得到的局部极小值点有很大影响。如图 10-9 所示, $f(x)$ 是一个非凸连续函数。在定义域内, 它的全局最优解应该为 x^* ; 它还有几个局部极小值点, 比如 x_l^1 、

x_l^2 等。如果我们把优化变量的初始值选在了 x_1 处, 经迭代优化后, 迭代算法很可能会收敛于局部极小值点 x_l^1 ; 类似地, 如果我们把优化变量的初始值选在了 x_2 处, 经迭代优化后, 迭代算法很可能会收敛于局部极小值点 x_l^2 。显然, 即使 x_l^2 并不是全局最优解, 但它明显要

比 x_l^1 好。因此，对于迭代优化算法来说，对优化变量初始值的合理估计是非常重要的，而这个步骤往往需要基于要解决的实际问题的领域知识来进行。针对相机标定这个特定问题，本节将详细介绍如何对待优化变量集合 Θ 进行合理的初始值估计。

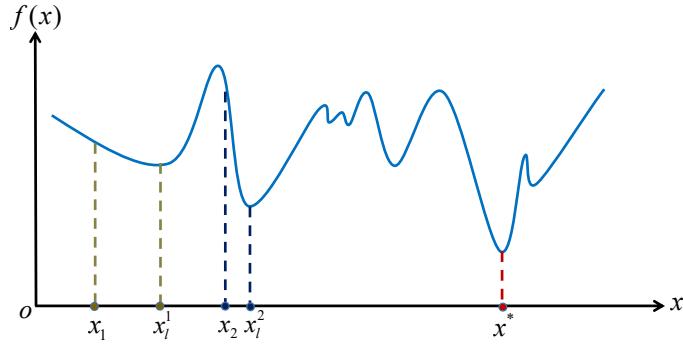


图 10-9：针对非凸问题的迭代优化方法的局部极小值点与优化变量初始值选取的关系。在该示意图中，定义域内的全局最优解是 x^* ；如果优化变量的初始值选在了 x_1 点，迭代优化算法很有可能会收敛于局部极小值点 x_l^1 ；如果优化变量的初始值选在了 x_2 点，迭代优化算法很有可能会收敛于局部极小值点 x_l^2 。

首先来考虑与镜头畸变有关的 5 个参数 $(k_1, k_2, \rho_1, \rho_2, k_3)$ 。如果关于镜头畸变我们并没有任何先验知识，可以把这 5 个参数都初始化为 0。也就是说，在参数初始化阶段，我们姑且简单地认为相机镜头不存在畸变。这样，在这个阶段所用的相机成像模型便是式 10-9。

再来考虑主点坐标 $(c_x, c_y)^T$ 。从相机成像流程示意图图 10-1 中可以看出，主点坐标是相机光心在成像平面上所成的像的像素坐标，基本上大致处于最终图像的中心位置。因此， c_x 和 c_y 可以被合理地分别初始化为所成图像宽和高的一半，即，

$$c_x = \frac{width}{2}, c_y = \frac{height}{2} \quad (10-25)$$

其中， $width$ 和 $height$ 分别为相机所拍摄图像的宽和高（单位是像素）。

对另外两个内参 (f_x, f_y) 的初始化稍微复杂了一些，需要用到消失点（vanishing point）的概念和性质，我们需要循序渐进地铺垫一些定义和命题。

定义 10.1 消失点。在针孔相机成像模型下，物理平面上一条直线的无穷远点在成像平面上所成的像称为这条直线所对应的消失点。

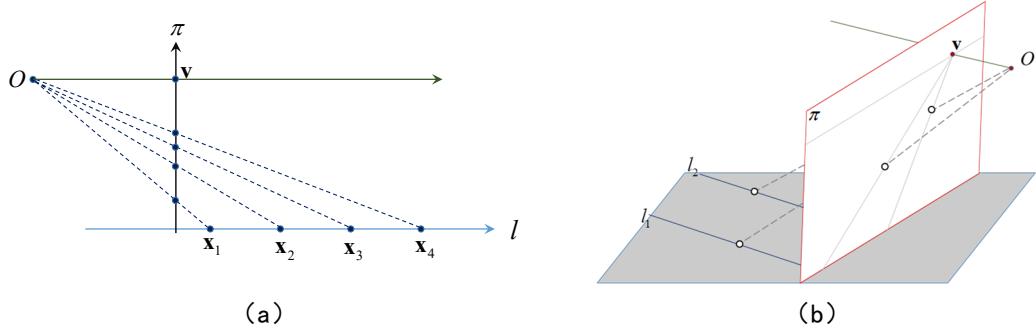


图 10-10: 消失点。(a) 物理平面上一条直线 l 的无穷远点在成像平面 π 上的像为 v , v 便称为 l 在成像平面 π 上的消失点; (b) 平面上两条平行直线 l_1 和 l_2 在成像平面 π 上的像会汇聚于同一个消失点 v 。

我们可以通过图 10-10 (a) 来理解一下消失点的定义。在图 10-10 (a) 中, O 为相机光心, π 是其成像平面 (图中显示的是该成像平面的截面)。物理平面上有一条直线 l , 我们现在来看看 l 在 π 上的像。考虑在 l 上取一些等间距的点, $x_1, x_2, x_3, x_4, \dots$ 无穷远点。这些点在投影平面 π 上的投影点之间的间距会越来越小。不难想象, l 的无穷远点的像最终会收敛于 v , v 是经过 O 且与 l 平行的直线与 π 的交点, 称为 l 的消失点。

由消失点的定义不难知道, 如图 10-10 (b) 所示, 欧氏平面上的两条平行线 l_1 和 l_2 , 它们在成像平面 π 上的像会汇聚在同一消失点 v 。从射影平面的视角来看, 两条平行线 l_1 和 l_2 会相交于无穷远点, 即它们具有相同的无穷远点, 而消失点是无穷远点的像, 因而 l_1 和 l_2 在成像平面上的像会汇聚于相同的消失点。更进一步, 对于平面上的一组平行线, 它们在成像平面上会具有相同的消失点; 对于平面上方向不同的两条直线, 它们的消失点也不同。而且, 我们有如下命题成立,

命题 10.1 设相机光心为 O , v 为成像平面上一点, 则直线 OV 平行于空间中以 v 为消失点的直线。

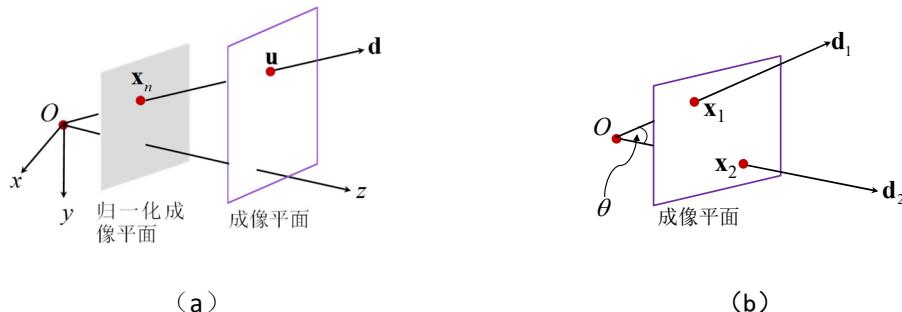


图 10-11: (a) 在相机坐标系下, 连接光心 O 和成像平面上像素坐标系下一点 u 的射线的方向为 $d = K^{-1}u$; (b) O 为相机光心, x_1, x_2 为成像平面上像素坐标系下的两点, 在相机坐标系下,

$$\text{射线 } \overrightarrow{Ox_1} \text{ 与 } \overrightarrow{Ox_2} \text{ 的夹角为 } \theta = \arccos \frac{\mathbf{x}_1^T (K^{-T} K^{-1}) \mathbf{x}_2}{\sqrt{\mathbf{x}_1^T (K^{-T} K^{-1}) \mathbf{x}_1} \sqrt{\mathbf{x}_2^T (K^{-T} K^{-1}) \mathbf{x}_2}}.$$

接下来我们要看看给定图像上一点，如何来表达连接相机光心和这点的射线的方向。有如下命题，

命题 10.2 设相机光心为 O , \mathbf{u} 为成像平面像素坐标系下一点的齐次坐标表示，则在相机坐标系下，射线 $\overrightarrow{O\mathbf{u}}$ 的方向 \mathbf{d} 可表示为 $\mathbf{d} = K^{-1}\mathbf{u}$ ，其中 K 为相机内参矩阵。

证明：

设像素 \mathbf{u} 的归一化齐次坐标表达为 \mathbf{u}' 。根据式 10-10 可知， \mathbf{u}' 与归一化成像平面上的对应点 $\mathbf{x}_n = (x_n, y_n, 1)^T$ （注意：这是归一化成像平面上二维点的归一化齐次坐标）之间的关系为， $\mathbf{x}_n = K^{-1}\mathbf{u}'$ 。需要注意的是，如图 10-11 (a) 所示， $(x_n, y_n, 1)^T$ 恰好也是 \mathbf{x}_n 这个点在相机坐标系下的三维空间坐标。同时，根据相机成像模型知道， O 、 \mathbf{x}_n 与 \mathbf{u} 共线。因此， \mathbf{d} 可以表达为 $\mathbf{d} = \overrightarrow{O\mathbf{u}} = \overrightarrow{O\mathbf{x}_n} = \mathbf{x}_n = K^{-1}\mathbf{u}'$ 。显然， $\forall c \neq 0$ ， $cK^{-1}\mathbf{u}' = K^{-1}(c\mathbf{u}')$ 都可以用来表示方向 \mathbf{d} ，而 $c\mathbf{u}'$ 就是 \mathbf{u}' 的普通齐次坐标表示，即为 \mathbf{u} 。因此， $\mathbf{d} = K^{-1}\mathbf{u}$ 。

命题 10.3 设相机光心为 O , \mathbf{x}_1 、 \mathbf{x}_2 为成像平面像素坐标系下两点的齐次坐标，则在相机坐标系下，射线 $\overrightarrow{O\mathbf{x}_1}$ 与 $\overrightarrow{O\mathbf{x}_2}$ 的夹角 θ 为，

$$\theta = \arccos \frac{\mathbf{x}_1^T (K^{-T} K^{-1}) \mathbf{x}_2}{\sqrt{\mathbf{x}_1^T (K^{-T} K^{-1}) \mathbf{x}_1} \sqrt{\mathbf{x}_2^T (K^{-T} K^{-1}) \mathbf{x}_2}} \quad (10-26)$$

其中， K 为相机内参矩阵。

证明：

根据图 10-11(b)，设 $\overrightarrow{O\mathbf{x}_1}$ 的方向为 \mathbf{d}_1 、 $\overrightarrow{O\mathbf{x}_2}$ 的方向为 \mathbf{d}_2 。根据命题 10.2 可知， $\mathbf{d}_1 = K^{-1}\mathbf{x}_1$ ， $\mathbf{d}_2 = K^{-1}\mathbf{x}_2$ 。因此有，

$$\cos \theta = \frac{\mathbf{d}_1 \cdot \mathbf{d}_2}{\|\mathbf{d}_1\| \|\mathbf{d}_2\|} = \frac{(K^{-1}\mathbf{x}_1)^T K^{-1}\mathbf{x}_2}{\sqrt{(K^{-1}\mathbf{x}_1)^T (K^{-1}\mathbf{x}_1)} \sqrt{(K^{-1}\mathbf{x}_2)^T (K^{-1}\mathbf{x}_2)}} = \frac{\mathbf{x}_1^T (K^{-T} K^{-1}) \mathbf{x}_2}{\sqrt{\mathbf{x}_1^T (K^{-T} K^{-1}) \mathbf{x}_1} \sqrt{\mathbf{x}_2^T (K^{-T} K^{-1}) \mathbf{x}_2}},$$

$$\text{则有, } \theta = \arccos \frac{\mathbf{x}_1^T (K^{-T} K^{-1}) \mathbf{x}_2}{\sqrt{\mathbf{x}_1^T (K^{-T} K^{-1}) \mathbf{x}_1} \sqrt{\mathbf{x}_2^T (K^{-T} K^{-1}) \mathbf{x}_2}}.$$

命题 10.4 设 l_1 和 l_2 是同一物理平面上的两条直线，它们在成像平面像素坐标系下的消失点分别为 \mathbf{v}_1 和 \mathbf{v}_2 ， O 为相机光心。 θ 为射线 $\overrightarrow{O\mathbf{v}_1}$ 与 $\overrightarrow{O\mathbf{v}_2}$ 之间的夹角。则直线 l_1 和 l_2 之间的两个夹角分别为 θ 和 $\pi - \theta$ 。

证明：

由于 \mathbf{v}_1 和 \mathbf{v}_2 是空间直线 l_1 和 l_2 在成像平面上的消失点，根据命题 10.1 可知， $l_1 \parallel O\mathbf{v}_1$ ， $l_2 \parallel O\mathbf{v}_2$ 。而向量 $\overrightarrow{O\mathbf{v}_1}$ 与 $\overrightarrow{O\mathbf{v}_2}$ 之间的夹角为 θ ，因此显然 l_1 和 l_2 之间的两个夹角分别为 θ 和 $\pi - \theta$ 。

命题 10.5 设 l_1 和 l_2 是同一物理平面上两条相互垂直的直线，它们在成像平面像素坐标系下的消失点分别为 \mathbf{v}_1 和 \mathbf{v}_2 , O 为相机光心，则有 $\mathbf{v}_1^T(K^{-T}K^{-1})\mathbf{v}_2=0$ ，其中 K 为相机内参矩阵。

证明：

由于 \mathbf{v}_1 和 \mathbf{v}_2 是空间直线 l_1 和 l_2 在成像平面上的消失点，根据命题 10.1 可知， $l_1 \parallel O\mathbf{v}_1$, $l_2 \parallel O\mathbf{v}_2$ 。又由于 $l_1 \perp l_2$ ，则有 $O\mathbf{v}_1 \perp O\mathbf{v}_2$ ，即矢量 $O\mathbf{v}_1$ 、 $O\mathbf{v}_2$ 之间的夹角 $\theta=\frac{\pi}{2}$ 。又由命题 10.3 证明过程可知， $\cos\theta=\frac{\mathbf{v}_1^T(K^{-T}K^{-1})\mathbf{v}_2}{\sqrt{\mathbf{v}_1^T(K^{-T}K^{-1})\mathbf{v}_1}\sqrt{\mathbf{v}_2^T(K^{-T}K^{-1})\mathbf{v}_2}}=\cos\frac{\pi}{2}=0$ 。因此有， $\mathbf{v}_1^T(K^{-T}K^{-1})\mathbf{v}_2=0$ 。

对相机内参 (f_x, f_y) 的初始化估计就是利用了标定板平面上相互垂直直线的消失点的性质来进行的。由于在本节相机模型参数初始化操作中，我们假定相机镜头无畸变，因此标定板平面和成像平面之间满足射影变换关系，即标定板平面上一点可以通过射影变换矩阵 $H_i \in \mathbb{R}^{3 \times 3}$ 变换到标定板图像 I_i 中的对应点。对于 H_i ，我们可以预先根据标定板平面上交叉点与 I_i 上图像空间中的交叉点之间的对应关系 $\{\mathbf{p}_j \leftrightarrow \mathbf{u}_{ij}\}_{j=1}^n$ ，使用线性最小二乘法将其解算出来。需要强调的一点是，这里的 \mathbf{p}_j 是标定板上第 j 个交叉点在标定板二维平面坐标系下的二维齐次坐标，其形式为 $\mathbf{p}_j=(x_j, y_j, 1)^T$ 。从对应点对关系集合来估计两个平面间的射影变换矩阵的具体方法可参见第 5 章中的相关内容。

考虑标定板平面坐标系下的 4 条特殊直线， $l_1: Y=0$ 、 $l_2: X=0$ 、 $l_3: Y=X$ 和 $l_4: Y=-X$ 。容易知道， $l_1 \perp l_2$ 、 $l_3 \perp l_4$ 。根据第 8 章知识可知，如果把标定板平面看作射影平面的话，可以求得 l_1 、 l_2 、 l_3 和 l_4 的无穷远点 $\mathbf{p}_{\infty 1}$ 、 $\mathbf{p}_{\infty 2}$ 、 $\mathbf{p}_{\infty 3}$ 和 $\mathbf{p}_{\infty 4}$ 坐标分别为，

$$\mathbf{p}_{\infty 1} = (1, 0, 0)^T, \quad \mathbf{p}_{\infty 2} = (0, 1, 0)^T, \quad \mathbf{p}_{\infty 3} = (1, 1, 0)^T, \quad \mathbf{p}_{\infty 4} = (1, -1, 0)^T \quad (10-27)$$

把 H_i 按列展开表示为 $H_i = [\mathbf{h}_{i1} \ \mathbf{h}_{i2} \ \mathbf{h}_{i3}]$ 。这样，标定板平面上的点 $\mathbf{p}_{\infty 1}$ 、 $\mathbf{p}_{\infty 2}$ 、 $\mathbf{p}_{\infty 3}$ 和 $\mathbf{p}_{\infty 4}$ 在 I_i 中的像 \mathbf{v}_{i1} 、 \mathbf{v}_{i2} 、 \mathbf{v}_{i3} 和 \mathbf{v}_{i4} 分别为，

$$\begin{aligned}
\mathbf{v}_{i1} &= [\mathbf{h}_{i1} \ \mathbf{h}_{i2} \ \mathbf{h}_{i3}] \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} = \mathbf{h}_{i1} \\
\mathbf{v}_{i2} &= [\mathbf{h}_{i1} \ \mathbf{h}_{i2} \ \mathbf{h}_{i3}] \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} = \mathbf{h}_{i2} \\
\mathbf{v}_{i3} &= [\mathbf{h}_{i1} \ \mathbf{h}_{i2} \ \mathbf{h}_{i3}] \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} = \mathbf{h}_{i1} + \mathbf{h}_{i2} \\
\mathbf{v}_{i4} &= [\mathbf{h}_{i1} \ \mathbf{h}_{i2} \ \mathbf{h}_{i3}] \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix} = \mathbf{h}_{i1} - \mathbf{h}_{i2}
\end{aligned} \tag{10-28}$$

根据消失点的定义可知, \mathbf{v}_{i1} 、 \mathbf{v}_{i2} 、 \mathbf{v}_{i3} 和 \mathbf{v}_{i4} 实际上正是直线 l_1 、 l_2 、 l_3 和 l_4 在图像 I_i 上的消失点。更进一步, 由于 $l_1 \perp l_2$ 、 $l_3 \perp l_4$, 根据命题 10.5 有,

$$\begin{cases} \mathbf{v}_{i1}^T (K^{-T} K^{-1}) \mathbf{v}_{i2} = 0 \\ \mathbf{v}_{i3}^T (K^{-T} K^{-1}) \mathbf{v}_{i4} = 0 \end{cases} \tag{10-29}$$

我们再来审视一下内参矩阵 $K = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix}$ 。令 $P = \begin{bmatrix} 1 & 0 & c_x \\ 0 & 1 & c_y \\ 0 & 0 & 1 \end{bmatrix}$ 、 $Q = \begin{bmatrix} f_x & 0 & 0 \\ 0 & f_y & 0 \\ 0 & 0 & 1 \end{bmatrix}$, 则显然有 $K = PQ$ 。

这样有,

$$K^{-T} K^{-1} = (PQ)^{-T} (PQ)^{-1} = P^{-T} (Q^{-T} Q^{-1}) P^{-1} \tag{10-30}$$

将式 10-30 带入式 10-29 得到,

$$\begin{cases} (P^{-1} \mathbf{v}_{i1})^T (Q^{-T} Q^{-1}) (P^{-1} \mathbf{v}_{i2}) = 0 \\ (P^{-1} \mathbf{v}_{i3})^T (Q^{-T} Q^{-1}) (P^{-1} \mathbf{v}_{i4}) = 0 \end{cases} \tag{10-31}$$

在式 10-31 中, $P^{-1} \mathbf{v}_{i1}$ 、 $P^{-1} \mathbf{v}_{i2}$ 、 $P^{-1} \mathbf{v}_{i3}$ 和 $P^{-1} \mathbf{v}_{i4}$ 实际上都为已知量。为了简化表述, 我们令

$$\begin{pmatrix} a_{i1} \\ b_{i1} \\ c_{i1} \end{pmatrix} \triangleq P^{-1} \mathbf{v}_{i1}、\begin{pmatrix} a_{i2} \\ b_{i2} \\ c_{i2} \end{pmatrix} \triangleq P^{-1} \mathbf{v}_{i2}、\begin{pmatrix} a_{i3} \\ b_{i3} \\ c_{i3} \end{pmatrix} \triangleq P^{-1} \mathbf{v}_{i3} \text{ 和 } \begin{pmatrix} a_{i4} \\ b_{i4} \\ c_{i4} \end{pmatrix} \triangleq P^{-1} \mathbf{v}_{i4}。 \text{ 同时, } Q^{-T} Q^{-1} = \begin{bmatrix} \frac{1}{f_x^2} & 0 & 0 \\ 0 & \frac{1}{f_y^2} & 0 \\ 0 & 0 & 1 \end{bmatrix}。 \text{ 这}$$

样, 式 10-31 可变形为,

$$\begin{bmatrix} a_{i1}a_{i2} & b_{i1}b_{i2} \\ a_{i3}a_{i4} & b_{i3}b_{i4} \end{bmatrix} \begin{bmatrix} \frac{1}{f_x^2} \\ \frac{1}{f_y^2} \end{bmatrix} = \begin{bmatrix} -c_{i1}c_{i2} \\ -c_{i3}c_{i4} \end{bmatrix} \tag{10-32}$$

式 10-32 实际上是由 2 个关于未知数 $\left(\frac{1}{f_x^2}, \frac{1}{f_y^2}\right)^T$ 的线性方程所组成的线性方程组，且这个方

程组是由标定板平面和它的图像 I_i 所确定的。由于我们一共拍摄了 m 张标定板图像 $\{I_i\}_{i=1}^m$ ，

所以相应地会得到 m 个形如式 10-32 的关于 $\left(\frac{1}{f_x^2}, \frac{1}{f_y^2}\right)^T$ 的线性方程组。我们把它们联立在一

起便得到了由标定板平面和它的图像集合 $\{I_i\}_{i=1}^m$ 所确定的关于 $\left(\frac{1}{f_x^2}, \frac{1}{f_y^2}\right)^T$ 的线性方程组，

$$\begin{bmatrix} a_{11}a_{12} & b_{11}b_{12} \\ a_{13}a_{14} & b_{13}b_{14} \\ a_{21}a_{22} & b_{21}b_{22} \\ a_{23}a_{24} & b_{23}b_{24} \\ \vdots & \\ a_{m1}a_{m2} & b_{m1}b_{m2} \\ a_{m3}a_{m4} & b_{m3}b_{m4} \end{bmatrix}_{2m \times 2} \begin{bmatrix} \frac{1}{f_x^2} \\ \frac{1}{f_y^2} \end{bmatrix} = \begin{bmatrix} -c_{11}c_{12} \\ -c_{13}c_{14} \\ -c_{21}c_{22} \\ -c_{23}c_{24} \\ \vdots \\ -c_{m1}c_{m2} \\ -c_{m3}c_{m4} \end{bmatrix}_{2m \times 2} \quad (10-33)$$

显然，式 10-33 中 $\left(\frac{1}{f_x^2}, \frac{1}{f_y^2}\right)^T$ 的求解问题是一个非齐次线性最小二乘问题，可以用 5.2 节中

所讲述的方法来解决该问题。当从式 10-33 中求解出 $\left(\frac{1}{f_x^2}, \frac{1}{f_y^2}\right)^T$ 之后，我们便相应地得到了

f_x 和 f_y 。这样到目前为止，在问题式 10-24 中，待优化参数集合 Θ 中的相机内参数 $\{f_x, f_y, c_x, c_y, k_1, k_2, \rho_1, \rho_2, k_3\}$ 就都已经初始化好了。接下来，我们将考虑如何对 Θ 中的外参数

$\{\mathbf{d}_i\}_{i=1}^m$ 、 $\{\mathbf{t}_i\}_{i=1}^m$ 进行合理初始化。

对于标定板图像 I ，假设我们已经得到了标定板平面上交叉点与 I 上图像空间中交叉点的对应关系集合 $\{\mathbf{p}_j \leftrightarrow \mathbf{u}_j\}_{j=1}^n$ ，其中 \mathbf{p}_j 为标定板上的交叉点，其在标定板平面坐标系下的齐次坐标可表示为 $(x_j, y_j, 1)^T$ ，相应地，其在标定板平面所定义出来的三维世界坐标系下的坐标为 $(x_j, y_j, 0, 1)^T$ ， \mathbf{u}_j 为 I 上与 \mathbf{p}_j 对应的点。设与 \mathbf{p}_j (及 \mathbf{u}_j) 对应的相机归一化成像平面上的点为 \mathbf{x}_{nj} ，根据式 10-10 有 $\mathbf{x}_{nj} = K^{-1} \mathbf{u}_j$ 。由于在参数初始化过程中我们假定相机镜头不存在畸变，因此标定板平面和归一化成像平面之间满足射影变换关系，相应的射影变换矩阵 P 可以通过对称关系集合 $\{\mathbf{p}_j \leftrightarrow \mathbf{x}_{nj}\}_{j=1}^n$ (注意：这里的 $\mathbf{p}_j = (x_j, y_j, 1)^T$ 为标定板上第 j 个交叉点在标定板平面坐标系下的二维齐次坐标) 解算出来。根据平面间射影变换的定义可知，对于标定板平面上任意的交叉点 $(x_j, y_j, 1)^T$ 有，

$$c_j \mathbf{x}_{nj} = P \begin{pmatrix} x_j \\ y_j \\ 1 \end{pmatrix} \quad (10-34)$$

其中, c_j 为与点 \mathbf{x}_{nj} 相关的一个常数。另一方面, 根据相机成像模型式 10-9 可知,

$$z_q K^{-1} \mathbf{u}_j = [R \mathbf{t}] (x_j, y_j, 0, 1)^T, \text{ 即}$$

$$z_{cj} \mathbf{x}_{nj} = [R \mathbf{t}] \begin{pmatrix} x_j \\ y_j \\ 0 \\ 1 \end{pmatrix} = [\mathbf{r}_1 \mathbf{r}_2 \mathbf{r}_3 \mathbf{t}] \begin{pmatrix} x_j \\ y_j \\ 0 \\ 1 \end{pmatrix} = [\mathbf{r}_1 \mathbf{r}_2 \mathbf{t}] \begin{pmatrix} x_j \\ y_j \\ 1 \end{pmatrix} \quad (10-35)$$

其中, \mathbf{r}_1 、 \mathbf{r}_2 、 \mathbf{r}_3 分别为矩阵 R 的第 1、2、3 列。需要注意的是, 由于坐标的齐次性, 式 10-34 的左端 $c_j \mathbf{x}_{nj}$ 和式 10-35 的左端 $z_{cj} \mathbf{x}_{nj}$ 实际上表达的是归一化成像平面上的同一个点。这样,

对比式 10-34 和式 10-35, 我们发现矩阵 P 和矩阵 $[\mathbf{r}_1 \mathbf{r}_2 \mathbf{t}]$ 把标定板平面上的点 $(x_j, y_j, 1)^T$ 映射到了归一化成像平面上的同一个点。因此, P 和 $[\mathbf{r}_1 \mathbf{r}_2 \mathbf{t}]$ 实际上表达了相同的平面间的射影变换。根据 3.1.5 中关于射影变换矩阵的知识可知, P 和 $[\mathbf{r}_1 \mathbf{r}_2 \mathbf{t}]$ 之间只相差了一个倍数关系, 即存在数 λ 使得,

$$P = \lambda [\mathbf{r}_1 \mathbf{r}_2 \mathbf{t}] \quad (10-36)$$

把 P 按列展开, 表达为形式 $[\mathbf{p}_1 \mathbf{p}_2 \mathbf{p}_3]$, 结合式 10-36 有,

$$\mathbf{r}_1 = \frac{1}{\lambda} \mathbf{p}_1, \mathbf{r}_2 = \frac{1}{\lambda} \mathbf{p}_2, \mathbf{t} = \frac{1}{\lambda} \mathbf{p}_3 \quad (10-37)$$

由于 R 为正交矩阵, 因此 $\|\mathbf{r}_1\|_2 = \|\mathbf{r}_2\|_2 = 1$, 结合式 10-37, 可得 $|\lambda| = \|\mathbf{p}_1\|_2 = \|\mathbf{p}_2\|_2$ 。在编程实现时, λ 一般可被取为 $\lambda = (\|\mathbf{p}_1\|_2 + \|\mathbf{p}_2\|_2) / 2$ 。当 λ 的值确定了以后, 根据式 10-37, \mathbf{r}_1 、 \mathbf{r}_2 和 \mathbf{t} 可以相应地被确定下来。由于 R 为正交矩阵且 $\det(R) = 1$, 可以推出 $\mathbf{r}_3 = \mathbf{r}_1 \times \mathbf{r}_2$, 具体证明过程给读者留作练习。最后, 我们再根据式 10-22 和 10-23, 把 R 转换为对应的轴角 \mathbf{d} 。这样, 与标定板图像 I 所对应的相机外参 \mathbf{d} 和 \mathbf{t} 就被初始化完毕了。由于我们给标定板拍摄了 m 张图像 $\{I_i\}_{i=1}^m$, 与每一张图像相关联的相机外参都不相同, 因此上述外参初始化过程需要针对每一张标定板图像都要进行一次以得到与之相关联的相机外参, 最终便可得到 Θ 中的全部相机外参数 $\{\mathbf{d}_i\}_{i=1}^m$ 、 $\{\mathbf{t}_i\}_{i=1}^m$ 的初始值。

10.3.4 相机成像模型参数的迭代优化

根据第 9 章中的知识可知, 式 10-24 所定义的问题为一个非线性最小二乘问题。我们在 10.3.3 中给该问题中的待优化变量 Θ 进行了合理的初始化。接下来就可以用第 9 章中介绍的高斯牛顿法或者列文伯格-马夸尔特法来迭代求解这个问题。

令,

$$\mathbf{f}_{ij}(\Theta) = K \cdot \mathcal{D} \left\{ \frac{1}{z_{cij}} [\mathcal{R}(\mathbf{d}_i) \mathbf{t}_i] \mathbf{p}_j \right\} - \mathbf{u}_{ij} \quad (10-38)$$

$$\mathbf{f}(\Theta) = (\mathbf{f}_{11}(\Theta) \mathbf{f}_{12}(\Theta) \cdots \mathbf{f}_{mn}(\Theta))^T \quad (10-39)$$

则式 10-24 所定义的优化问题可被表达为,

$$\Theta^* = \arg \min_{\Theta} \left(\frac{1}{2} \mathbf{f}^T(\Theta) \mathbf{f}(\Theta) \right) \quad (10-40)$$

式 10-40 便是标准化的非线性最小二乘问题的表达方式了。根据第 9 章中的内容我们知道,无论是用高斯牛顿法还是用列文伯格-马夸尔特法来求解问题 10-40, 关键都在于要推导出 $\mathbf{f}(\Theta)$ 的雅可比矩阵 $J(\Theta)$ 的表达形式。

在式 10-38 中, 令 $\mathbf{u}_{ij} = K \cdot \mathcal{D} \left\{ \frac{1}{z_{cij}} [\mathcal{R}(\mathbf{d}_i) \mathbf{t}_i] \mathbf{p}_j \right\}$, 则 \mathbf{u}_{ij} 表示的是根据成像模型计算出的

标定板上的交叉点 \mathbf{p}_j (其表达是在由标定板平面所定义的世界坐标系下) 在标定板图像 I_i 上的投影点。 $\mathbf{f}_{ij}(\Theta)$ 中与优化变量 Θ 有关的部分仅为 \mathbf{u}_{ij} , 因此 $\mathbf{f}(\Theta)$ 的雅可比矩阵 $J(\Theta)$ 为,

$$J(\Theta) = \begin{pmatrix} \frac{d\mathbf{f}_{11}(\Theta)}{d\Theta^T} \\ \frac{d\mathbf{f}_{12}(\Theta)}{d\Theta^T} \\ \vdots \\ \frac{d\mathbf{f}_{ln}(\Theta)}{d\Theta^T} \\ \frac{d\mathbf{f}_{21}(\Theta)}{d\Theta^T} \\ \vdots \\ \frac{d\mathbf{f}_{mn}(\Theta)}{d\Theta^T} \end{pmatrix}_{(2mn) \times (9+6m)} = \begin{pmatrix} \frac{d\mathbf{u}_{11}}{d\Theta^T} \\ \frac{d\mathbf{u}_{12}}{d\Theta^T} \\ \vdots \\ \frac{d\mathbf{u}_{ln}}{d\Theta^T} \\ \frac{d\mathbf{u}_{21}}{d\Theta^T} \\ \vdots \\ \frac{d\mathbf{u}_{mn}}{d\Theta^T} \end{pmatrix}_{(2mn) \times (9+6m)} \quad (10-41)$$

要得出 $J(\Theta)$ 的表达式, 关键在于要得到 $\frac{d\mathbf{u}_{ij}}{d\Theta^T}$ 的表达式。具体来说, 我们要得到 \mathbf{u}_{ij} 关于相机

内参的偏导数 $\frac{\partial \mathbf{u}_{ij}}{\partial f_x}$ 、 $\frac{\partial \mathbf{u}_{ij}}{\partial f_y}$ 、 $\frac{\partial \mathbf{u}_{ij}}{\partial c_x}$ 、 $\frac{\partial \mathbf{u}_{ij}}{\partial c_y}$ 、 $\frac{\partial \mathbf{u}_{ij}}{\partial k_1}$ 、 $\frac{\partial \mathbf{u}_{ij}}{\partial k_2}$ 、 $\frac{\partial \mathbf{u}_{ij}}{\partial \rho_1}$ 、 $\frac{\partial \mathbf{u}_{ij}}{\partial \rho_2}$ 和 $\frac{\partial \mathbf{u}_{ij}}{\partial k_3}$; 也要得到

\mathbf{u}_{ij} 关于相机外参的偏导数 $\frac{\partial \mathbf{u}_{ij}}{\partial \mathbf{d}_k}$ ($k = 1, \dots, m$) 和 $\frac{\partial \mathbf{u}_{ij}}{\partial \mathbf{t}_k}$ ($k = 1, \dots, m$), 但我们要注意到图像 I_i 和相机位姿 $\forall k \neq i, (\mathbf{d}_k, \mathbf{t}_k)$ 是没有关系的, 因此 $\forall k \neq i$, 有 $\frac{\partial \mathbf{u}_{ij}}{\partial \mathbf{d}_k} = \mathbf{0}$ 和 $\frac{\partial \mathbf{u}_{ij}}{\partial \mathbf{t}_k} = \mathbf{0}$, 因此实际上我们只需

要推导 $\frac{\partial \mathbf{u}_{ij}}{\partial \mathbf{d}_i}$ 和 $\frac{\partial \mathbf{u}_{ij}}{\partial \mathbf{t}_i}$ 的表达式。

设 \mathbf{u}_{ij} 的非齐次坐标为 $\mathbf{u}_{ij} = (u, v)^T$; 与 \mathbf{u}_{ij} 对应的世界坐标系下的三维空间点为 \mathbf{p}_j , 我们把

它的坐标记为 $\mathbf{p}_j = (x, y, z, 1)^T$ (齐次坐标); \mathbf{p}_j 在相机 i 坐标系下的坐标记为 $\mathbf{p}_c = (x_c, y_c, z_c)^T$ (非齐次坐标); \mathbf{p}_j 在相机 i 归一化成像平面上的投影记为 $\mathbf{p}_n = (x_n, y_n)^T$, 它在相机 i 归一化成像平面上经过镜头畸变建模之后的投影记为 $\mathbf{p}_d = (x_d, y_d)^T$ 。记 $\mathbf{d}_i = (d_1, d_2, d_3)^T$, $\mathbf{t}_i = (t_1, t_2, t_3)^T$ 。需要明确一点, \mathbf{p}_j 是已知量而且它在迭代优化过程中始终保持不变, 其他坐标值 \mathbf{p}_c 、 \mathbf{p}_n 、 \mathbf{p}_d 和 \mathbf{u}_y^i 可以以当前迭代点 Θ 为成像模型参数值, 根据相机成像模型式 10-14, 通过投影 \mathbf{p}_j 来计算得到, 因此, 对于当前迭代点 Θ 来说, Θ 、 \mathbf{p}_j 、 \mathbf{p}_c 、 \mathbf{p}_n 、 \mathbf{p}_d 和 \mathbf{u}_y^i 实际上都是已知量。

根据式 10-14 可知, 归一化成像平面上的点 \mathbf{p}_d 与其在成像平面像素坐标系下的投影点 \mathbf{u}_y^i 之间只相差了一个内参矩阵 K , 即 $\mathbf{u}_y^i = K\mathbf{p}_d$, 展开之后即为,

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} x_d \\ y_d \\ 1 \end{bmatrix} \quad (10-42)$$

根据式 10-42, 可得,

$$\frac{\partial \mathbf{u}_y^i}{\partial f_x} = \begin{bmatrix} \frac{\partial u}{\partial f_x} \\ \frac{\partial v}{\partial f_x} \\ \frac{\partial 1}{\partial f_x} \end{bmatrix} = \begin{bmatrix} x_d \\ 0 \\ 0 \end{bmatrix}, \frac{\partial \mathbf{u}_y^i}{\partial f_y} = \begin{bmatrix} \frac{\partial u}{\partial f_y} \\ \frac{\partial v}{\partial f_y} \\ \frac{\partial 1}{\partial f_y} \end{bmatrix} = \begin{bmatrix} 0 \\ y_d \\ 0 \end{bmatrix}, \frac{\partial \mathbf{u}_y^i}{\partial c_x} = \begin{bmatrix} \frac{\partial u}{\partial c_x} \\ \frac{\partial v}{\partial c_x} \\ \frac{\partial 1}{\partial c_x} \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \frac{\partial \mathbf{u}_y^i}{\partial c_y} = \begin{bmatrix} \frac{\partial u}{\partial c_y} \\ \frac{\partial v}{\partial c_y} \\ \frac{\partial 1}{\partial c_y} \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \quad (10-43)$$

同时, 从式 10-42 中, 我们还可得到,

$$\frac{\partial \mathbf{u}_y^i}{\partial \mathbf{p}_d^T} = \begin{bmatrix} \frac{\partial u}{\partial x_d} \frac{\partial u}{\partial y_d} \\ \frac{\partial v}{\partial x_d} \frac{\partial v}{\partial y_d} \\ \frac{\partial 1}{\partial x_d} \frac{\partial 1}{\partial y_d} \end{bmatrix} = \begin{bmatrix} f_x & 0 \\ 0 & f_y \end{bmatrix} \quad (10-44)$$

我们把与镜头畸变建模有关的内参数组合为一个向量 $\mathbf{k} \triangleq (k_1, k_2, \rho_1, \rho_2, k_3)^T$ 。根据式 10-13, 可得,

$$\frac{\partial \mathbf{p}_d}{\partial \mathbf{k}^T} = \begin{bmatrix} \frac{\partial x_d}{\partial k_1} \frac{\partial x_d}{\partial k_2} \frac{\partial x_d}{\partial \rho_1} \frac{\partial x_d}{\partial \rho_2} \frac{\partial x_d}{\partial k_3} \\ \frac{\partial y_d}{\partial k_1} \frac{\partial y_d}{\partial k_2} \frac{\partial y_d}{\partial \rho_1} \frac{\partial y_d}{\partial \rho_2} \frac{\partial y_d}{\partial k_3} \end{bmatrix} = \begin{bmatrix} x_n r^2 & x_n r^4 & 2x_n y_n & r^2 + 2x_n^2 & x_n r^6 \\ y_n r^2 & y_n r^4 & r^2 + 2y_n^2 & 2x_n y_n & y_n r^6 \end{bmatrix} \quad (10-45)$$

其中, $r^2 = x_n^2 + y_n^2$ 。结合式 10-44 和 10-45, 根据链式求导法则得到,

$$\frac{\partial \mathbf{u}_y^i}{\partial \mathbf{k}^T} = \frac{\partial \mathbf{u}_y^i}{\partial \mathbf{p}_d^T} \cdot \frac{\partial \mathbf{p}_d}{\partial \mathbf{k}^T} = \begin{bmatrix} f_x x_n r^2 & f_x x_n r^4 & 2f_x x_n y_n & f_x (r^2 + 2x_n^2) & f_x x_n r^6 \\ f_y y_n r^2 & f_y y_n r^4 & f_y (r^2 + 2y_n^2) & 2f_y x_n y_n & f_y y_n r^6 \end{bmatrix} \quad (10-46)$$

至此为止, 我们已经得到了 \mathbf{u}_{ij}^+ 关于相机所有内参的偏导数形式, 即 $\frac{\partial \mathbf{u}_{ij}^+}{\partial f_x}$ 、 $\frac{\partial \mathbf{u}_{ij}^+}{\partial f_y}$ 、 $\frac{\partial \mathbf{u}_{ij}^+}{\partial c_x}$ 、 $\frac{\partial \mathbf{u}_{ij}^+}{\partial c_y}$

和 $\frac{\partial \mathbf{u}_{ij}^+}{\partial \mathbf{k}^T}$ 。接下来需要确定 \mathbf{u}_{ij}^+ 关于相机外参 $(\mathbf{d}_i, \mathbf{t}_i)$ 的偏导数形式。

根据式 10-13, 可得,

$$\begin{aligned} \frac{\partial \mathbf{p}_d}{\partial \mathbf{p}_n^T} &= \begin{bmatrix} \frac{\partial x_d}{\partial x_n} & \frac{\partial x_d}{\partial y_n} \\ \frac{\partial x_n}{\partial x_n} & \frac{\partial y_n}{\partial y_n} \\ \frac{\partial y_d}{\partial x_n} & \frac{\partial y_d}{\partial y_n} \\ \frac{\partial x_n}{\partial x_n} & \frac{\partial y_n}{\partial y_n} \end{bmatrix} \\ &= \begin{bmatrix} 1+k_1r^2+k_2r^4+k_3r^6+2x_n^2(k_1+2k_2r^2+3k_3r^4)+2\rho_1y_n+6\rho_2x_n & 2x_ny_n(k_1+2k_2r^2+3k_3r^4)+2(\rho_1x_n+\rho_2y_n) \\ 2x_ny_n(k_1+2k_2r^2+3k_3r^4)+2(\rho_1x_n+\rho_2y_n) & 1+k_1r^2+k_2r^4+k_3r^6+2y_n^2(k_1+2k_2r^2+3k_3r^4)+2\rho_2x_n+6\rho_1y_n \end{bmatrix} \end{aligned} \quad (10-47)$$

根据式 10-4 可得,

$$\frac{\partial \mathbf{p}_n}{\partial \mathbf{p}_c^T} = \begin{bmatrix} \frac{\partial x_n}{\partial x_c} & \frac{\partial x_n}{\partial y_c} & \frac{\partial x_n}{\partial z_c} \\ \frac{\partial y_n}{\partial x_c} & \frac{\partial y_n}{\partial y_c} & \frac{\partial y_n}{\partial z_c} \\ \frac{\partial z_n}{\partial x_c} & \frac{\partial z_n}{\partial y_c} & \frac{\partial z_n}{\partial z_c} \end{bmatrix} = \begin{bmatrix} \frac{1}{z_c} & 0 & -\frac{x_c}{z_c^2} \\ 0 & \frac{1}{z_c} & -\frac{y_c}{z_c^2} \end{bmatrix} \quad (10-48)$$

设与轴角 \mathbf{d}_i 对应的旋转矩阵为 $R = \begin{bmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{bmatrix}$, 把矩阵 R 的元素按行排列形成列向量

$\mathbf{r} = (r_{11}, r_{12}, r_{13}, r_{21}, r_{22}, r_{23}, r_{31}, r_{32}, r_{33})^T$ 。式 10-1 表达了世界坐标系下的一点与相机坐标系下点的关系, 根据式 10-1, 我们可知 \mathbf{p}_c 与 \mathbf{p}_j 之间满足下式关系,

$$\begin{bmatrix} x_c \\ y_c \\ z_c \end{bmatrix} = \begin{bmatrix} r_{11}x + r_{12}y + r_{13}z + t_1 \\ r_{21}x + r_{22}y + r_{23}z + t_2 \\ r_{31}x + r_{32}y + r_{33}z + t_3 \end{bmatrix} \quad (10-49)$$

由式 10-49 可得,

$$\frac{\partial \mathbf{p}_c}{\partial \mathbf{r}^T} = \begin{bmatrix} \frac{\partial x_c}{\partial r_{11}} & \frac{\partial x_c}{\partial r_{12}} & \frac{\partial x_c}{\partial r_{13}} & \frac{\partial x_c}{\partial r_{21}} & \frac{\partial x_c}{\partial r_{22}} & \frac{\partial x_c}{\partial r_{23}} & \frac{\partial x_c}{\partial r_{31}} & \frac{\partial x_c}{\partial r_{32}} & \frac{\partial x_c}{\partial r_{33}} \\ \frac{\partial y_c}{\partial r_{11}} & \frac{\partial y_c}{\partial r_{12}} & \frac{\partial y_c}{\partial r_{13}} & \frac{\partial y_c}{\partial r_{21}} & \frac{\partial y_c}{\partial r_{22}} & \frac{\partial y_c}{\partial r_{23}} & \frac{\partial y_c}{\partial r_{31}} & \frac{\partial y_c}{\partial r_{32}} & \frac{\partial y_c}{\partial r_{33}} \\ \frac{\partial z_c}{\partial r_{11}} & \frac{\partial z_c}{\partial r_{12}} & \frac{\partial z_c}{\partial r_{13}} & \frac{\partial z_c}{\partial r_{21}} & \frac{\partial z_c}{\partial r_{22}} & \frac{\partial z_c}{\partial r_{23}} & \frac{\partial z_c}{\partial r_{31}} & \frac{\partial z_c}{\partial r_{32}} & \frac{\partial z_c}{\partial r_{33}} \end{bmatrix} = \begin{bmatrix} x & y & z & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & x & y & z & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & x & y & z \end{bmatrix} \quad (10-50)$$

$$\frac{d\mathbf{p}_c}{d\mathbf{t}_i^T} = \begin{bmatrix} \frac{\partial x_c}{\partial t_1} & \frac{\partial x_c}{\partial t_2} & \frac{\partial x_c}{\partial t_3} \\ \frac{\partial y_c}{\partial t_1} & \frac{\partial y_c}{\partial t_2} & \frac{\partial y_c}{\partial t_3} \\ \frac{\partial z_c}{\partial t_1} & \frac{\partial z_c}{\partial t_2} & \frac{\partial z_c}{\partial t_3} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad (10-51)$$

我们把轴角 $\mathbf{d}_i = (d_1, d_2, d_3)^T$ 显示地表示为旋转轴与旋转角乘积的形式, $\mathbf{d}_i = \theta \mathbf{n}$, 其中

$\theta = \|\mathbf{d}_i\|_2$, $\mathbf{n} = (n_1, n_2, n_3)^T = \frac{\mathbf{d}_i}{\|\mathbf{d}_i\|_2}$, 记 $\alpha = \sin \theta$ 、 $\beta = \cos \theta$ 、 $\gamma = 1 - \cos \theta$, 由式 10-20 可得,

$$\begin{bmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{bmatrix} = \beta \begin{bmatrix} 1 & & \\ & 1 & \\ & & 1 \end{bmatrix} + \gamma \begin{bmatrix} n_1^2 & n_1 n_2 & n_1 n_3 \\ n_1 n_2 & n_2^2 & n_2 n_3 \\ n_1 n_3 & n_2 n_3 & n_3^2 \end{bmatrix} + \alpha \begin{bmatrix} 0 & -n_3 & n_2 \\ n_3 & 0 & n_1 \\ -n_2 & n_1 & 0 \end{bmatrix} \quad (10-52)$$

基于以上信息, 我们可推导出 \mathbf{r} 与 \mathbf{d}_i 的导数关系,

$$\frac{\partial \mathbf{r}}{\partial \mathbf{d}_i^T} = \begin{bmatrix} \frac{2\gamma n_1(1-n_1^2)}{\theta} + \alpha n_1(n_1^2-1) & -\frac{2\gamma n_1^2 n_2}{\theta} + \alpha n_2(n_1^2-1) & -\frac{2\gamma n_1^2 n_3}{\theta} + \alpha n_3(n_1^2-1) \\ n_1(\alpha n_1 n_2 - \beta n_3) + \frac{\gamma n_2(1-2n_1^2) + \alpha n_1 n_3}{\theta} & n_2(\alpha n_1 n_2 - \beta n_3) + \frac{\gamma n_1(1-2n_2^2) + \alpha n_2 n_3}{\theta} & n_3(\alpha n_1 n_2 - \beta n_3) + \frac{\alpha(n_3^2-1) - 2\gamma n_1 n_2 n_3}{\theta} \\ n_1(\alpha n_1 n_3 + \beta n_2) + \frac{\gamma n_3(1-2n_1^2) - \alpha n_1 n_2}{\theta} & n_2(\alpha n_1 n_3 + \beta n_2) + \frac{\alpha(1-n_2^2) - 2\gamma n_1 n_2 n_3}{\theta} & n_3(\alpha n_1 n_3 + \beta n_2) + \frac{\gamma n_1(1-2n_3^2) - \alpha n_2 n_3}{\theta} \\ n_1(\alpha n_1 n_2 + \beta n_3) + \frac{\gamma n_2(1-2n_1^2) - \alpha n_1 n_3}{\theta} & n_2(\alpha n_1 n_2 + \beta n_3) + \frac{\gamma n_1(1-2n_2^2) - \alpha n_2 n_3}{\theta} & n_3(\alpha n_1 n_2 + \beta n_3) + \frac{\alpha(1-n_3^2) - 2\gamma n_1 n_2 n_3}{\theta} \\ \frac{-2\gamma n_1 n_2^2}{\theta} + \alpha n_1(n_2^2-1) & \frac{2\gamma n_2(1-n_2^2)}{\theta} + \alpha n_2(n_2^2-1) & \frac{-2\gamma n_2 n_3^2}{\theta} + \alpha n_3(n_2^2-1) \\ n_1(\alpha n_2 n_3 - \beta n_1) - \frac{\alpha(1-n_1^2) + 2\gamma n_1 n_2 n_3}{\theta} & n_2(\alpha n_2 n_3 - \beta n_1) + \frac{\gamma n_3(1-2n_2^2) + \alpha n_1 n_2}{\theta} & n_3(\alpha n_2 n_3 - \beta n_1) + \frac{\alpha n_1 n_3 + \gamma n_2(1-2n_3^2)}{\theta} \\ n_1(\alpha n_1 n_3 - \beta n_2) + \frac{\alpha n_1 n_2 + \gamma n_3(1-2n_1^2)}{\theta} & n_2(\alpha n_1 n_3 - \beta n_2) - \frac{\alpha(1-n_2^2) + 2\gamma n_1 n_2 n_3}{\theta} & n_3(\alpha n_1 n_3 - \beta n_2) + \frac{\alpha n_2 n_3 + \gamma n_1(1-2n_3^2)}{\theta} \\ n_1(\alpha n_2 n_3 + \beta n_1) + \frac{\alpha(1-n_1^2) - 2\gamma n_1 n_2 n_3}{\theta} & n_2(\alpha n_2 n_3 + \beta n_1) + \frac{\gamma n_3(1-2n_2^2) - \alpha n_1 n_2}{\theta} & n_3(\alpha n_2 n_3 + \beta n_1) + \frac{\gamma n_2(1-2n_3^2) - \alpha n_1 n_3}{\theta} \\ \frac{-2\gamma n_1 n_3^2}{\theta} + \alpha n_1(n_3^2-1) & \frac{-2\gamma n_2 n_3^2}{\theta} + \alpha n_2(n_3^2-1) & \frac{2\gamma n_3(1-n_3^2)}{\theta} + \alpha n_3(n_3^2-1) \end{bmatrix} \quad (10-53)$$

作为练习, 请读者完成式 10-53 的推导。结合式 10-44、10-47、10-48、10-50 和 10-53, 根据链式求导法则得到,

$$\frac{\partial \dot{\mathbf{u}}_{ij}^i}{\partial \mathbf{d}_i^T} = \frac{\partial \dot{\mathbf{u}}_{ij}^i}{\partial \mathbf{p}_d^T} \cdot \frac{\partial \mathbf{p}_d}{\partial \mathbf{p}_n^T} \cdot \frac{\partial \mathbf{p}_n}{\partial \mathbf{p}_c^T} \cdot \frac{\partial \mathbf{p}_c}{\partial \mathbf{r}^T} \cdot \frac{\partial \mathbf{r}}{\partial \mathbf{d}_i^T} \quad (10-54)$$

结合式 10-44、10-47、10-48 和 10-51, 根据链式求导法则得到,

$$\frac{\partial \dot{\mathbf{u}}_{ij}^i}{\partial \mathbf{t}_i^T} = \frac{\partial \dot{\mathbf{u}}_{ij}^i}{\partial \mathbf{p}_d^T} \cdot \frac{\partial \mathbf{p}_d}{\partial \mathbf{p}_n^T} \cdot \frac{\partial \mathbf{p}_n}{\partial \mathbf{p}_c^T} \cdot \frac{\partial \mathbf{p}_c}{\partial \mathbf{t}_i^T} \quad (10-55)$$

到这里为止, 计算 $\frac{d\dot{\mathbf{u}}_{ij}^i}{d\Theta^T}$ 所需的所有必要的表达形式我们都已经得到了, 继而可以确定 $\mathbf{f}(\Theta)$

的雅可比矩阵 $J(\Theta)$, 然后就可以利用第 9 章中所介绍的高斯牛顿法或者列文伯格-马夸尔特法来对相机模型参数集合 Θ 进行迭代优化了, 最终便可得到相机参数的标定结果 Θ^* 。

10.4 镜头畸变去除

当有了相机模型的参数以后, 便可以基于图像观测来对物理空间进行测量。一般来说, 为了便于建模和分析, 往往先要对获取的图像进行镜头畸变去除。在进行了镜头畸变去除以后, 便可以使用理想的针孔相机成像模型 (式 10-9) 来建模成像流程了, 而无需再考虑镜头畸变这件事儿了。在相机内参数已知的情况下, 图像的镜头畸变去除是很容易执行的。

假设原始拍摄的带有镜头畸变的图像为 I_d , 去畸变之后的图像记为 I 。对于 I 上一点 \mathbf{u} ,

我们需要计算出 I_d 上与之对应的点 \mathbf{u}_d 。与 \mathbf{u} 对应的归一化成像坐标系下的点为 $K^{-1}\mathbf{u}$, 该点经镜头畸变映射至归一化成像坐标系下的点 $\mathcal{D}(K^{-1}\mathbf{u})$ 。点 $\mathcal{D}(K^{-1}\mathbf{u})$ 在成像平面像素坐标系下的投影为 $K(\mathcal{D}(K^{-1}\mathbf{u}))$, 即 $\mathbf{u}_d=K(\mathcal{D}(K^{-1}\mathbf{u}))$ 。之后, 我们便可把 $I_d(\mathbf{u}_d)$ 的像素值赋值给 $I(\mathbf{u})$ 。当然, 在实际编程实现的时候, 由于 \mathbf{u} 为整数, 它在 I_d 上的对应点坐标 $\mathbf{u}_d=K(\mathcal{D}(K^{-1}\mathbf{u}))$ 几乎不可能也为整数, 因此 $I_d(\mathbf{u}_d)$ 的像素值的获取需要通过对 \mathbf{u}_d 邻域整数位置点的像素值的插值来得到。关于图像插值的内容, 我们已经在第 6 章中介绍过了。

10.5 习题

- (1) 矩阵 $R \in \mathbb{R}^{3 \times 3}$ 为正交矩阵且 $\det(R)=1$ 。若把 R 按列展开, 写成形式 $R = [\mathbf{r}_1 \ \mathbf{r}_2 \ \mathbf{r}_3]$, 其中 \mathbf{r}_1 、 \mathbf{r}_2 、 \mathbf{r}_3 分别为 R 的第 1、2、3 列, 请证明 $\mathbf{r}_3 = \mathbf{r}_1 \times \mathbf{r}_2$ 。
- (2) 矩阵 $R \in \mathbb{R}^{3 \times 3}$ 为正交矩阵且 $\det(R)=1$ 。请证明 1 必然是 R 的特征值。
- (3) 请推导式 10-47 和式 10-53。

参考文献

- [1] D.C. Brown, "Close-range camera calibration," *Photogrammetric Engineering* 37 (1971): 855–866.
- [2] J.G. Fryer and D. C. Brown, "Lens distortion for close-range photogrammetry," *Photogrammetric Engineering and Remote Sensing* 52 (1986): 51–58.
- [3] Z. Zhang, "A flexible new technique for camera calibration," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 22, no. 11, pp. 1330–1334, 2000.
- [4] A. Kaehler and G. Bradski, *Learning OpenCV 3*, O'Reilly Media, Inc., 2016.
- [5] Cheng, Hui; Gupta, K. C. (March 1989). "An Historical Note on Finite Rotations". *Journal of Applied Mechanics. American Society of Mechanical Engineers.* 56 (1): 139–145. Retrieved 2022-04-11.

第 11 章 鸟瞰视图

在第 7 章中提到，为了便于使用视觉技术来对平面上的目标进行检测或测量，我们可以生成物理平面的鸟瞰视图。鸟瞰视图又称为**逆透视投影**，这是因为当拍摄物理平面信息时，在针孔相机模型下，图像平面是物理平面通过透视投影产生的。逆透视投影便是将图像平面信息反投影至它所对应的物理平面上，得到物理平面的“像素化”表示。鸟瞰视图被广泛应用在辅助驾驶中的环视系统和各类工业流水线中的工件属性测量系统中。我们将在这一章学习如何从物理平面的图像中构造出该平面的鸟瞰视图。

11.1 基本流程

如果读者已经掌握了相机内参标定技术、图像镜头畸变去除技术以及平面间射影变换估计技术的话，会很容易理解和掌握鸟瞰视图生成技术，因为后者正是对前面几项技术的综合应用。

假定对于某个物理平面，我们拍摄了它的图像 I_D ，现在的任务是要从 I_D 中生成该平面的鸟瞰视图 I_B 。完成这个任务的关键在于要建立起从鸟瞰视图 I_B 坐标系下的点到原始图像 I_D 坐标系下的点的映射查找表 $T_{B \rightarrow D}$ ，即对于给定的 I_B 上的一点 \mathbf{x}_B ，通过查询查找表 $T_{B \rightarrow D}$ ，我们可以得到 I_D 上与之对应的点 \mathbf{x}_D 。这样便可以把 $I_B(\mathbf{x}_B)$ 赋值为 $I_D(\mathbf{x}_D)$ ，从而生成出鸟瞰视图 I_B 。

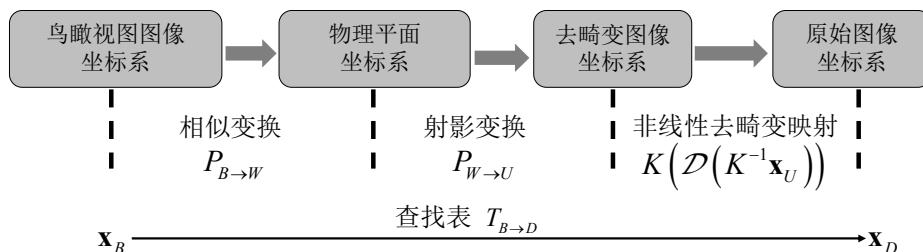


图 11-1：鸟瞰视图坐标系下的一点 \mathbf{x}_B 与原始图像坐标系下对应点 \mathbf{x}_D 之间的映射关系。

如图 11-1 所示，从概念上来说，查找表 $T_{B \rightarrow D}$ 的构建需要借助几个坐标系，鸟瞰视图图像坐标系、物理平面坐标系、去畸变图像坐标系和原始图像坐标系^[1]。只要我们建立起了从鸟瞰视图图像坐标系到物理平面坐标系的映射关系、从物理平面坐标系到去畸变图像坐标系的映射关系以及从去畸变图像坐标系到原始图像坐标系的映射关系，我们便可以生成查找表 $T_{B \rightarrow D}$ 。需要强调一下， $T_{B \rightarrow D}$ 的构建是离线完成的，它和图像的内容无关，只与相机相对于物理平面的位姿有关；因此，只要相机相对于物理平面的位姿保持不变， $T_{B \rightarrow D}$ 在构建完毕

之后就不需要再重新构建了。

11.2 鸟瞰视图坐标系到物理平面坐标系的映射

鸟瞰视图图像坐标系与物理平面坐标系之间的映射关系 $P_{B \rightarrow W}$ 实际上是一个相似变换。

一般情况下,为了测量和表示的方便,鸟瞰视图图像坐标系的两个坐标轴与物理平面坐标系的两个坐标轴分别平行,这样只要预先确定好鸟瞰图像的像素分辨率、它所覆盖的物理平面范围以及鸟瞰视图中心点在物理平面坐标系下的位置便可以确定出 $P_{B \rightarrow W}$ 。鸟瞰视图图像坐标系的单位为像素,物理平面坐标系的单位为物理长度单位,比如米或者毫米。

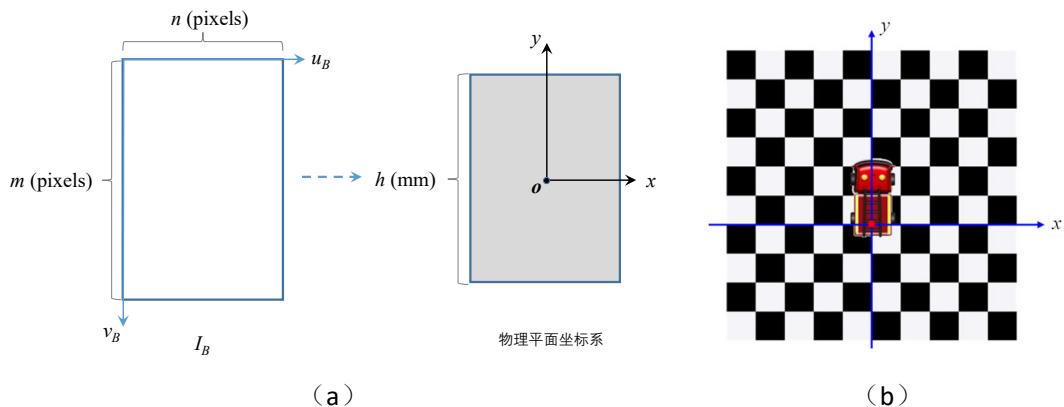


图 11-2: (a) 鸟瞰视图图像坐标系与物理平面坐标系之间的映射关系示意图; (b) 基于棋盘格标定场所建立的物理平面坐标系, 该标定场由边长为 1 米的黑白方格组成。

如图 11-2 所示,假设生成的鸟瞰视图图像的分辨率为 $m \times n$ (单位为像素),所覆盖的物理平面的长度为 h (单位为物理长度单位,比如毫米),图像的中心对应于物理平面坐标系的原点 \mathbf{o} 。设 $\mathbf{x}_B = (x_B, y_B)^T$ 为鸟瞰视图图像 I_B 上一点,与之对应的物理平面坐标系下的一点为 $\mathbf{x}_W = (x_W, y_W)^T$,那么容易知道, \mathbf{x}_B 与 \mathbf{x}_W 之间的关系为,

$$\begin{pmatrix} x_W \\ y_W \\ 1 \end{pmatrix} = \begin{bmatrix} \frac{h}{m} & 0 & -\frac{hn}{2m} \\ 0 & -\frac{h}{m} & \frac{h}{2} \\ 0 & 0 & 1 \end{bmatrix} \begin{pmatrix} x_B \\ y_B \\ 1 \end{pmatrix} \quad (11-1)$$

即, $P_{B \rightarrow W}$ 为,

$$P_{B \rightarrow W} = \begin{bmatrix} \frac{h}{m} & 0 & -\frac{hn}{2m} \\ 0 & -\frac{h}{m} & \frac{h}{2} \\ 0 & 0 & 1 \end{bmatrix} \quad (11-2)$$

当然，式 11-2 是在鸟瞰视图中心对应于物理平面坐标系原点 \mathbf{o} 的情况下得到的结果，如果鸟瞰视图中心对应于物理平面坐标系中的其他位置，则 $P_{B \rightarrow W}$ 的具体形式也需要随之改变。

需要强调一点，在下一步建立物理平面坐标系与去畸变图像坐标系的映射关系时，需要借助于棋盘格标定场。为了方便起见，物理平面坐标系的建立往往也是借助于标定场的，即它的坐标原点会选在某个棋盘格交叉点上、它的两个坐标轴要沿着标定场的边的方向，如图 11-2 (b) 所示。

11.3 物理平面坐标系到去畸变图像坐标系的映射

假设物理平面的去畸变图像为 I_U ，则物理平面与 I_U 之间满足射影变换关系。根据第 5 章中的知识可知，要估计两个平面之间的射影变换，必须要知道这两个平面之间的对应点对集合且集合中的元素个数不少于 4 个。为此，我们需要在物理平面上铺设棋盘格标定场，标定场中每个交叉点在物理平面坐标系下的坐标都可以提前测得。如图 11-3 所示，在图像 I_U 中，我们可以选择一些在视场内可见的棋盘格图像交叉点并标注出它们在 I_U 上的像素坐标 $\{\mathbf{x}_U^i\}_{i=1}^p (p > 4)$ ；与 \mathbf{x}_U^i 对应的物理平面坐标系下的棋盘格交叉点 \mathbf{x}_W^i 是已知的。假设物理平面与去畸变图像 I_U 之间的射影变换矩阵为 $P_{W \rightarrow U}$ ，则 $\mathbf{x}_U^i = P_{W \rightarrow U} \mathbf{x}_W^i$ 。这样，基于点对关系集合

$\{\mathbf{x}_U^i \leftrightarrow \mathbf{x}_W^i\}_{i=1}^p$ ，使用最小二乘法（参见第 5 章）便可以解出 $P_{W \rightarrow U}$ 。

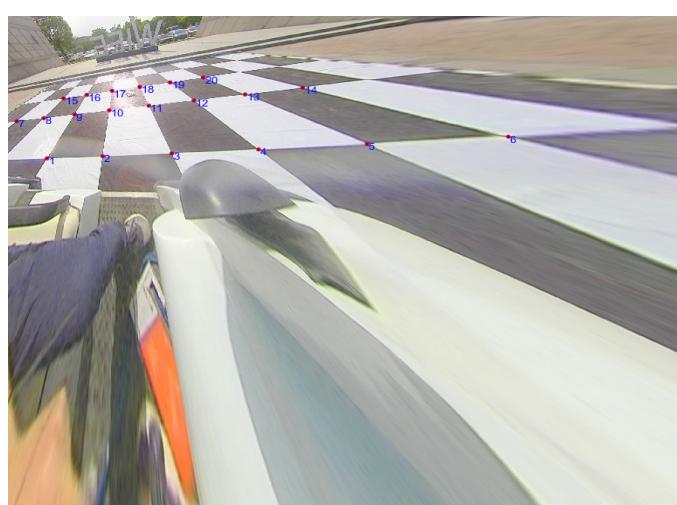


图 11-3：在去畸变图像上选取可视范围内的棋盘格图像交叉点并标注它们的像素坐标。

11.4 去畸变图像坐标系到原始图像坐标系的映射

去畸变图像坐标系到原始图像坐标系的映射实际上就是图像镜头畸变去除的过程。根据 10.4 节的内容可知，设 \mathbf{x}_U （二维齐次坐标）为去畸变图像 I_U 上的一点，它所对应的带有镜头畸变的原始图像 I_D 上的一点为 $\mathbf{x}_D = K \mathcal{D}(K^{-1} \mathbf{x}_U)$ ，其中 K 为相机的内参矩阵， $\mathcal{D}(\cdot)$ 为相机镜头畸变算子（式 10-14）。

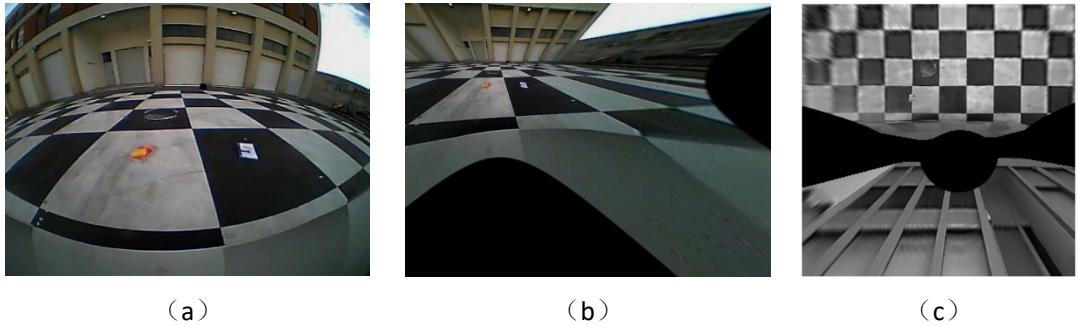


图 11-4：(a) 原始标定场图像；(b) 去除镜头畸变图像；(c) 鸟瞰视图图像。

这样，假设我们对相机进行了内参标定得到了内参矩阵 K 和镜头畸变算子 $\mathcal{D}(\cdot)$ ，通过外参标定得到了矩阵 $P_{B \rightarrow W}$ 和矩阵 $P_{W \rightarrow U}$ ，对于鸟瞰视图图像中的一点 \mathbf{x}_B ，便可以通过下式得到在原始输入图像 I_D 中的对应点 \mathbf{x}_D ，

$$\mathbf{x}_D = K \left(\mathcal{D} \left(K^{-1} \left(P_{W \rightarrow U} P_{B \rightarrow W} \mathbf{x}_B \right) \right) \right) \quad (11-3)$$

通过式 11-3，我们便可以建立起从鸟瞰视图坐标系下的点到原始图像坐标系下的点的映射查找表 $T_{B \rightarrow D}$ ，进而便可以生成鸟瞰视图。图 11-4 通过一个实例展示了鸟瞰视图图像的生成过程，(a) 为原始标定场图像，(b) 为对 (a) 进行了镜头畸变去除之后的图像，(c) 为最终得到的鸟瞰视图图像。

参考文献

- [1] L. Zhang, X. Li, J. Huang, Y. Shen and D. Wang, "Vision-based parking-slot detection: A benchmark and a learning-based approach," *Symmetry*, vol. 2018, no. 10, pp. 64:1-18, 2018.

第三篇：目标检测

第 12 章 目标检测问题概述

在第 7 章中提到，为了便于使用视觉技术来对平面上的目标进行检测或测量，我们可以生成物理平面的鸟瞰视图。鸟瞰视图又称为**逆透视投影**，这是因为当拍摄物理平面信息时，在针孔相机模型下，图像平面是物理平面通过透视投影产生的。逆透视投影便是将图像平面信息反投影至它所对应的物理平面上，得到物理平面的“像素化”表示。鸟瞰视图被广泛应用在辅助驾驶中的环视系统和各类工业流水线中的工件属性测量系统中。我们将在这一章学习如何从物理平面的图像中构造出该平面的鸟瞰视图。

参考文献

- [2] L. Zhang, X. Li, J. Huang, Y. Shen and D. Wang, “Vision-based parking-slot detection: A benchmark and a learning-based approach,” *Symmetry*, vol. 2018, no. 10, pp. 64:1-18, 2018.

第 13 章 凸优化基础

13.1 凸优化问题

13.1.1 凸集与仿射集

定义 16.1 凸集 (Convex set)。如果一个集合 \mathcal{C} 是凸集，当且仅当对于 \mathcal{C} 中任意的两个元素 $\mathbf{x}_1 \in \mathcal{C}$ 和 $\mathbf{x}_2 \in \mathcal{C}$ ，对于任意的实数 $\theta \in [0,1]$ 有，

$$\theta\mathbf{x}_1 + (1-\theta)\mathbf{x}_2 \in \mathcal{C}$$

$\theta\mathbf{x}_1 + (1-\theta)\mathbf{x}_2$ 也称为元素 \mathbf{x}_1 、 \mathbf{x}_2 的凸组合。从凸集的定义可以看出，如果一个集合 \mathcal{C} 是凸集的话， \mathcal{C} 中任意两个元素 \mathbf{x}_1 、 \mathbf{x}_2 连接所形成的“线段”上的点也一定属于集合 \mathcal{C} 。图 13-1 给出了二维空间上的几个典型的凸集和非凸集的示例。

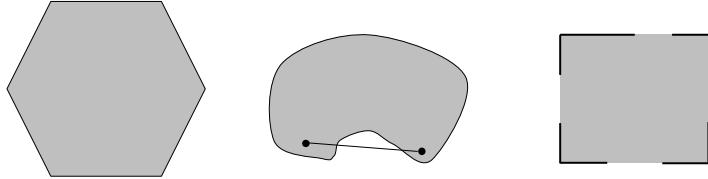


图 13-1：几个典型的二维空间上的凸集与非凸集的示例：左图是一个六边形且包含边界，它是一个凸集；中间的一个不是凸集，因为如图所示的连接两点的线段有一部分处于集合之外；右边的四边形包含了一部分边界，有一部分边界没有被包含在集合中，按照凸集的定义容易知道，它不是凸集。

命题 13.1 集合的交运算保持凸性。也就是说，如果集合 \mathcal{C}_1 和 \mathcal{C}_2 为两个凸集，那么它们的交集 $\mathcal{C}_1 \cap \mathcal{C}_2$ 也是凸集。

定义 13.2 仿射集 (Affine set)。如果一个集合 \mathcal{C} 是仿射集，当且仅当对于 \mathcal{C} 中任意的两个元素 $\mathbf{x}_1 \in \mathcal{C}$ 和 $\mathbf{x}_2 \in \mathcal{C}$ ，对于任意的实数 θ ，有

$$\theta\mathbf{x}_1 + (1-\theta)\mathbf{x}_2 \in \mathcal{C}$$

$\theta\mathbf{x}_1 + (1-\theta)\mathbf{x}_2$ 也称为元素 \mathbf{x}_1 、 \mathbf{x}_2 的仿射组合。从仿射集的定义可以看出，如果一个集合 \mathcal{C} 是仿射集的话，经过 \mathcal{C} 中任意两个元素 \mathbf{x}_1 、 \mathbf{x}_2 的“直线”上的点也一定属于集合 \mathcal{C} 。对照仿射集和凸集的定义不难看出，如果一个集合是仿射集，它也必为凸集。

定义 13.3 仿射包 (Affine hull)。由集合 \mathcal{C} 中元素所有可能的仿射组合所形成的集合称为集合 \mathcal{C} 的仿射包，记作 $\text{aff}\mathcal{C}$:

$$\text{aff}\mathcal{C} = \{\theta \mathbf{x}_1 + (1-\theta) \mathbf{x}_2 \mid \mathbf{x}_1, \mathbf{x}_2 \in \mathcal{C}\}$$

其中， θ 为任意实数。

由仿射包的定义不难看出， $\text{aff}\mathcal{C}$ 一定是仿射集，且是包含集合 \mathcal{C} 的最小仿射集；如果 \mathcal{C} 本身已经是仿射集，那么 $\text{aff}\mathcal{C}=\mathcal{C}$ 。

在 \mathbb{R}^n 空间中，基于集合 \mathcal{C} 的仿射包 $\text{aff}\mathcal{C}$ ，我们可以定义集合 \mathcal{C} 的相对内部 (relative interior) 和相对边界 (relative boundary) 这两个概念。作为铺垫，我们首先回顾一下集合的内部 (interior)、闭包 (closure)、边界 (boundary) 这几个概念。集合 $\mathcal{C} \subseteq \mathbb{R}^n$ 的内部 $\text{int}\mathcal{C}$ 被定义为，

$$\text{int}\mathcal{C} = \{\mathbf{x} \in \mathcal{C} \mid \exists \varepsilon > 0, B(\mathbf{x}, \varepsilon) \subseteq \mathcal{C}\}$$

即，如果 \mathbf{x} 为 $\text{int}\mathcal{C}$ 中的一点，这意味着 \mathbf{x} 属于集合 \mathcal{C} ，并且一定存在以 \mathbf{x} 为中心的某个 \mathbb{R}^n 空间中的“闭球” $B(\mathbf{x}, \varepsilon) = \{\mathbf{y} \mid \|\mathbf{y} - \mathbf{x}\| \leq \varepsilon\}$ ($\|\cdot\|$ 可以是任意范数)，该球全部被包含在 \mathcal{C} 中。

我们说一个集合 \mathcal{C} 为开集 (open)，如果 $\text{int}\mathcal{C}=\mathcal{C}$ ，即集合 \mathcal{C} 中的每一个点都是它的内部点；集合 $\mathcal{C} \subseteq \mathbb{R}^n$ 为闭集 (closed)，如果它的补集 (complement) $\mathbb{R}^n \setminus \mathcal{C} = \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{x} \notin \mathcal{C}\}$ 为开集。

集合 \mathcal{C} 的闭包被定义为 $\text{cl}\mathcal{C} = \mathbb{R}^n \setminus \text{int}\{\mathbb{R}^n \setminus \mathcal{C}\}$ ；也可以从另外一个角度来理解闭包，

$$\text{cl}\mathcal{C} = \{\mathbf{x} \in \mathbb{R}^n \mid \forall \varepsilon > 0, \exists \mathbf{y} \in \mathcal{C}, \|\mathbf{y} - \mathbf{x}\| \leq \varepsilon\}$$

集合 \mathcal{C} 的边界被定义为 $\text{bd}\mathcal{C} = \text{cl}\mathcal{C} \setminus \text{int}\mathcal{C}$ ；边界上一点 \mathbf{x} 满足如下性质： $\forall \varepsilon > 0, \exists \mathbf{y} \in \mathcal{C}$, $\exists \mathbf{z} \notin \mathcal{C}$ ，使得 $\|\mathbf{y} - \mathbf{x}\| \leq \varepsilon$, $\|\mathbf{z} - \mathbf{x}\| \leq \varepsilon$ ，即同时存在离 \mathbf{x} 无限近的 \mathcal{C} 中的点和离 \mathbf{x} 无限近的 $\mathbb{R}^n \setminus \mathcal{C}$ 中的点。根据边界的定义可知，如果集合 \mathcal{C} 为闭集，则它包含其边界，即 $\text{bd}\mathcal{C} \subseteq \mathcal{C}$ ；如果集合 \mathcal{C} 为开集，则它不包含边界中的任何点，即 $\text{bd}\mathcal{C} \cap \mathcal{C} = \emptyset$ 。

需要注意，上面阐述的集合 \mathcal{C} 的内部、边界等概念是相对于 \mathcal{C} 中元素所在的空间 \mathbb{R}^n 来说的。为了后续论述需要，我们还需要引入集合 \mathcal{C} 相对于它的仿射包 $\text{aff}\mathcal{C}$ 的“相对内部” $\text{relint}\mathcal{C}$ 这个概念：

定义 13.4 相对内部 (relative interior)。有集合 $\mathcal{C} \subseteq \mathbb{R}^n$ ，其相对内部 $\text{relint}\mathcal{C}$ 被定义为，

$$\text{relint}\mathcal{C} = \{\mathbf{x} \in \mathcal{C} \mid \exists \varepsilon > 0, \text{such that } (B(\mathbf{x}, \varepsilon) \cap \text{aff}\mathcal{C}) \subseteq \mathcal{C}\} \quad (13-1)$$

其中, $B(\mathbf{x}, \varepsilon) = \{\mathbf{y} \mid \|\mathbf{y} - \mathbf{x}\| \leq \varepsilon\}$ 是以 \mathbf{x} 为中心、以 ε 为半径的 \mathbb{R}^n 空间中的闭球 ($\|\cdot\|$ 可以是任意范数)。基于 $\text{relint}\mathcal{C}$, 我们还可以定义集合 \mathcal{C} 相对它的仿射包 $\text{aff}\mathcal{C}$ 来说的“相对边界”,

$\text{cl}\mathcal{C} \setminus \text{relint}\mathcal{C}$ 。容易理解, 如果 $\text{aff}\mathcal{C} = \mathbb{R}^n$, 则有 $\text{relint}\mathcal{C} = \text{int}\mathcal{C}$ 和 $\text{cl}\mathcal{C} \setminus \text{relint}\mathcal{C} = \text{bd}\mathcal{C}$ 。

我们通过一个具体的例子来理解一下内部、边界、相对内部、相对边界这几个概念。

例 13.1: 考虑 \mathbb{R}^3 空间中 (x_1, x_2) -平面上的一个正方形内的点所形成的集合 \mathcal{C} (图 13-2 左图),

$$\mathcal{C} = \{\mathbf{x} \in \mathbb{R}^3 \mid -1 \leq x_1 \leq 1, -1 \leq x_2 \leq 1, x_3 = 0\}$$

其中, $\mathbf{x} = (x_1, x_2, x_3)^T$ 。集合 \mathcal{C} 的仿射包为整个 (x_1, x_2) -平面, 即 $\text{aff}\mathcal{C} = \{\mathbf{x} \in \mathbb{R}^3 \mid x_3 = 0\}$ 。集合 \mathcal{C} 的内部 $\text{int}\mathcal{C}$ (在 \mathbb{R}^3 中) 为空集, 但它的相对内部 $\text{relint}\mathcal{C}$ (图 13-2 中间图) 为,

$$\text{relint}\mathcal{C} = \{\mathbf{x} \in \mathbb{R}^3 \mid -1 < x_1 < 1, -1 < x_2 < 1, x_3 = 0\}$$

集合 \mathcal{C} 的边界 $\text{bd}\mathcal{C}$ (在 \mathbb{R}^3 中) 为其自身; 它的相对边界 $\text{cl}\mathcal{C} \setminus \text{relint}\mathcal{C}$ 为这个正方形的“边框”(图 13-2 右图),

$$\text{cl}\mathcal{C} \setminus \text{relint}\mathcal{C} = \{\mathbf{x} \in \mathbb{R}^3 \mid \max\{|x_1|, |x_2|\} = 1, x_3 = 0\}$$

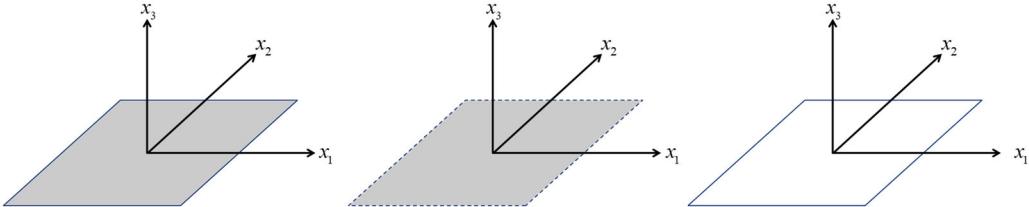


图 13-2: 左图是例 13.1 中所描述的集合 \mathcal{C} , 它是 \mathbb{R}^3 空间中的一个正方形; 中间图示意了集合 \mathcal{C} 的相对内部 $\text{relint}\mathcal{C}$; 右图示意了集合 \mathcal{C} 的相对边界 $\text{cl}\mathcal{C} \setminus \text{relint}\mathcal{C}$ 。

从这个例子可以看出, 相对内部(相对边界)这个概念是用来描述嵌在高维空间(比如, \mathbb{R}^3 空间)中的低维空间(比如, \mathbb{R}^3 中的某个平面)上的某个集合相对于低维空间的“内外”关系的。

13.1.2 凸函数

定义 13.5 仿射函数(Affine function)。如果函数 $f: \mathbb{R}^n \rightarrow \mathbb{R}^m$ 是一个线性函数和一个常量之和, 也就是说, 它具有如下形式,

$$f(\mathbf{x}) = A\mathbf{x} + \mathbf{b}, A \in \mathbb{R}^{m \times n}, \mathbf{x} \in \mathbb{R}^{n \times 1}, \mathbf{b} \in \mathbb{R}^{m \times 1} \quad (13-2)$$

则 $f(\mathbf{x})$ 被称为仿射函数。

定义 13.6 凸函数 (Convex function)。 函数 $f(\mathbf{x}): \mathbb{R}^n \rightarrow \mathbb{R}$, 如果它的定义域 $\text{dom} f$ 是凸集, 且对于定义域中任意的两点 \mathbf{x} 和 \mathbf{y} , 对于任意的 $\theta \in [0,1]$ 都有,

$$\theta f(\mathbf{x}) + (1-\theta)f(\mathbf{y}) \geq f(\theta\mathbf{x} + (1-\theta)\mathbf{y}) \quad (13-3)$$

则称 $f(\mathbf{x})$ 为凸函数 (如图 13-3 (a) 所示)。

从凸函数的定义可以看出, 如果一个函数是凸函数, 当且仅当其定义域内两点凸组合的函数值要小于等于两点函数值的凸组合, 当然, 还要满足定义域是凸集的前提条件。如果式 13-3 中的大于等于号是严格大于号, 那么函数 $f(\mathbf{x})$ 便称为**严格凸函数**(strictly convex function)。如果 $-f(\mathbf{x})$ 为凸函数, 则称 $f(\mathbf{x})$ 为**凹函数**(Concave function); 相应地, 如果 $-f(\mathbf{x})$ 为严格凸函数, 则称 $f(\mathbf{x})$ 为**严格凹函数**。

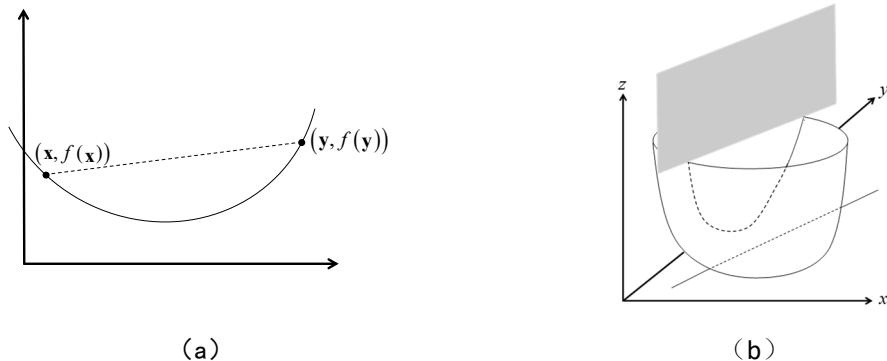


图 13-3: (a) 凸函数的图形。凸函数图形上任意两点的连线所形成的弦 (chord) 都在函数图形之上; (b) 一个函数为凸函数, 当且仅当该函数被限定在其定义域内的任意一条线上时, 所形成的一元函数也为凸函数。

命题 13.2 仿射函数既是凸函数也是凹函数。

证明:

仿射函数的定义为式 13-2。其定义域为 \mathbb{R}^n , 显然为凸集。另外, $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n, \forall \theta \in [0,1]$, 有,

$$\begin{aligned} \theta f(\mathbf{x}) + (1-\theta)f(\mathbf{y}) &= \theta(A\mathbf{x} + \mathbf{b}) + (1-\theta)(A\mathbf{y} + \mathbf{b}) \\ &= A\theta\mathbf{x} + A(1-\theta)\mathbf{y} + \theta\mathbf{b} + (1-\theta)\mathbf{b} \\ &= A(\theta\mathbf{x} + (1-\theta)\mathbf{y}) + \mathbf{b} \\ &= f(\theta\mathbf{x} + (1-\theta)\mathbf{y}) \end{aligned}$$

则可知仿射函数 $f(\mathbf{x})$ 为凸函数。类似地, 也可证明 $-f(\mathbf{x})$ 也为凸函数, 则 $f(\mathbf{x})$ 为凹函数。这样综合起来, 仿射函数既是凸函数也是凹函数。

命题 13.3 一个函数为凸函数，当且仅当若该函数被限定在其定义域内的任意一条线上时所形成的一元函数也为凸函数。

也就是说，函数 $f(\mathbf{x})$ 为凸函数当且仅当对于其定义域内的任意两点 \mathbf{x} 和 \mathbf{v} ，函数 $g(t) = f(\mathbf{x} + t\mathbf{v})$ 为凸函数，其中 t 要使得 $\mathbf{x} + t\mathbf{v}$ 依然在定义域内，如图 13-3 (b) 所示。我们可以通过一个比喻来帮助读者理解一下该性质。平放在地面上的一口锅就是一个凸函数。这时，用任意一个垂直于地面的平面去截这口锅，所得到的交线就是一个一元函数了，该一元函数的定义域就是平面与地面相交所得到的直线，显然这样得到的一元函数也是凸函数。凸函数的这个性质是一个很有用的性质：若要检查某函数 $f(\mathbf{x})$ 是否为凸函数，可以把该函数限定在一维直线上，再检查所得到的一维函数是否为凸函数；若以这种方式得到的任意一维函数均为凸函数，则 $f(\mathbf{x})$ 便为凸函数。

定理 13.1 凸函数的一阶判定条件。设函数 $f(\mathbf{x}): \mathbb{R}^n \rightarrow \mathbb{R}$ 一阶可微，也就是说函数在其定义域内的每一点处的梯度 $\nabla f(\mathbf{x})$ 都存在，则函数 $f(\mathbf{x})$ 是凸函数的充要条件是：它的定义域是凸集并且对于定义域内的任意两点 \mathbf{x} 和 \mathbf{y} ，下式成立，

$$f(\mathbf{y}) \geq f(\mathbf{x}) + (\nabla f(\mathbf{x}))^\top (\mathbf{y} - \mathbf{x}) \quad (13-4)$$

证明：

(1) 我们先来证明 $n=1$ 的情况，即需要证明：可微函数 $f(x): \mathbb{R} \rightarrow \mathbb{R}$ 为凸函数的充要条件是 $\text{dom}f$ 为凸集，且 $f(y) \geq f(x) + f'(x)(y-x)$ ， $\forall x, y \in \text{dom}f$ 都成立，其中 $\text{dom}f$ 为函数 f 的定义域。

必要性。由于 $f(x)$ 为凸函数，根据凸函数的定义，其定义域 $\text{dom}f$ 为凸集。根据凸集的性质， $\forall x, y \in \text{dom}f$ ， $\forall t \in (0, 1]$ ，有 $x + t(y-x) \in \text{dom}f$ 。由于 $f(x)$ 为凸函数，根据凸函数定义有，

$$f(x + t(y-x)) = f(ty + (1-t)x) \leq tf(y) + (1-t)f(x)$$

上式两边同时除以 t 得到，

$$f(y) \geq \frac{f(x + t(y-x)) - f(x)}{t} + f(x)$$

上式两边关于 $t \rightarrow 0$ 取极限得到，

$$\begin{aligned} f(y) &\geq f(x) + \lim_{t \rightarrow 0} \frac{f(x + t(y-x)) - f(x)}{t} \\ &= f(x) + \frac{f(x) + t(y-x)f'(x) - f(x)}{t} \\ &= f(x) + f'(x)(y-x) \end{aligned}$$

充分性。已知 $\text{dom}f$ 为凸集，且 $f(y) \geq f(x) + f'(x)(y-x)$ ， $\forall x, y \in \text{dom}f$ 都成立，要证

明 $f(x)$ 为凸函数。 $\forall x, y \in \text{dom}f$, $\theta \in [0,1]$, 令 $z = \theta x + (1-\theta)y$, 由于 $\text{dom}f$ 为凸集, 则 $z \in \text{dom}f$ 。

根据已知条件,

$$f(x) \geq f(z) + f'(z)(x-z), \quad f(y) \geq f(z) + f'(z)(y-z)$$

上面两个不等式的两边分别乘以 θ 和 $1-\theta$ 得到,

$$\theta f(x) \geq \theta f(z) + \theta f'(z)(x-z), \quad (1-\theta)f(y) \geq (1-\theta)f(z) + (1-\theta)f'(z)(y-z)$$

上述两个不等式左右两边分别相加得到,

$$\begin{aligned} \theta f(x) + (1-\theta)f(y) &\geq \theta f(z) + \theta f'(z)(x-z) + (1-\theta)f(z) + (1-\theta)f'(z)(y-z) \\ &= \theta f(z) + \theta x f'(z) - \theta z f'(z) + f(z) - \theta f(z) + y f'(z) - z f'(z) - \theta y f'(z) + \theta z f'(z) \\ &= f(z) + [\theta x + (1-\theta)y - z] f'(z) \\ &= f(z) + [z - z] f'(z) \\ &= f(z) = f(\theta x + (1-\theta)y) \end{aligned}$$

因此函数 $f(x)$ 为凸函数。

(2) 我们再来证明一般情况, 也就是函数形式为 $f(\mathbf{x}): \mathbb{R}^n \rightarrow \mathbb{R}$ 时的情况。

必要性。也就是要证明如果 $f(\mathbf{x})$ 为凸函数时, $\text{dom}f$ 为凸集且 $\forall \mathbf{x}, \mathbf{y} \in \text{dom}f$,

$f(\mathbf{y}) \geq f(\mathbf{x}) + (\nabla f(\mathbf{x}))^\top (\mathbf{y} - \mathbf{x})$ 成立。根据凸函数定义, 若 $f(\mathbf{x})$ 为凸函数, 其定义域 $\text{dom}f$ 必然为凸集。 $\forall \mathbf{x}, \mathbf{y} \in \text{dom}f$, 设 $f(\mathbf{x})$ 被限定在直线 $t\mathbf{y} + (1-t)\mathbf{x}$ (t 的取值要使得 $t\mathbf{y} + (1-t)\mathbf{x} \in \text{dom}f$)

之 上 所 形 成 的 一 元 函 数 为 $g(t) = f(t\mathbf{y} + (1-t)\mathbf{x})$ 。记 $\mathbf{u} = t\mathbf{y} + (1-t)\mathbf{x}$, 则

$$g'(t) = \frac{dg}{d\mathbf{u}^\top} \frac{d\mathbf{u}}{dt} = (\nabla f(t\mathbf{y} + (1-t)\mathbf{x}))^\top (\mathbf{y} - \mathbf{x}), \text{ 因此有 } g'(0) = (\nabla f(\mathbf{x}))^\top (\mathbf{y} - \mathbf{x})。 \text{ 根据命题 13.3 我们}$$

知道, $g(t)$ 为凸函数, 因此, 根据上面我们已经证明的 $n=1$ 时的情况可知,

$$g(1) \geq g(0) + g'(0)(1-0), \text{ 即 } f(\mathbf{y}) \geq f(\mathbf{x}) + (\nabla f(\mathbf{x}))^\top (\mathbf{y} - \mathbf{x})。$$

充分性。也就是要证明: 如果 $\text{dom}f$ 为凸集且 $\forall \mathbf{x}, \mathbf{y} \in \text{dom}f$, $f(\mathbf{y}) \geq f(\mathbf{x}) + (\nabla f(\mathbf{x}))^\top (\mathbf{y} - \mathbf{x})$ 都成立, 那么 $f(\mathbf{x})$ 必为凸函数。 $\forall \mathbf{x}, \mathbf{y} \in \text{dom}f$, 任取两个数 t 和 s , 它们要满足

$t\mathbf{y} + (1-t)\mathbf{x} \in \text{dom}f$, $s\mathbf{y} + (1-s)\mathbf{x} \in \text{dom}f$ 。那么, 根据已知条件有,

$$\begin{aligned} f(t\mathbf{y} + (1-t)\mathbf{x}) &\geq f(s\mathbf{y} + (1-s)\mathbf{x}) + (\nabla f(s\mathbf{y} + (1-s)\mathbf{x}))^\top (t\mathbf{y} + (1-t)\mathbf{x} - s\mathbf{y} - (1-s)\mathbf{x}) \\ &= f(s\mathbf{y} + (1-s)\mathbf{x}) + (\nabla f(s\mathbf{y} + (1-s)\mathbf{x}))^\top (\mathbf{y} - \mathbf{x})(t-s) \end{aligned}$$

上式也就是 $g(t) \geq g(s) + g'(s)(t-s)$, 且 t 的取值集合 $\text{dom}g$ 为凸集。根据已经证明的 $n=1$ 时

的情况可知， $g(t)$ 为凸函数。在上面证明过程中， \mathbf{x} 、 \mathbf{y} 和 t 是任意取的，这就意味着把 $f(\mathbf{x})$ 限制在它定义域内任意一条线 $t\mathbf{y} + (1-t)\mathbf{x}$ 上，所得到的一元函数 $g(t) = f(t\mathbf{y} + (1-t)\mathbf{x})$ 都是凸函数，根据命题 13.3 可知， $f(\mathbf{x})$ 必为凸函数。

图 13-4(a) 以可视化的方式展示了定理 13.1 所描述的凸函数一阶判定条件的几何含义。我们也可以用一种更加生活化的方式来阐述这个定理的思想：如图 13-4(b) 所示，你有一口锅（即使是平底锅也没有问题），在锅曲面上任取一点，在该点放一根与锅曲面相切的棍子，那么这根棍子的全部一定都在锅的下面。

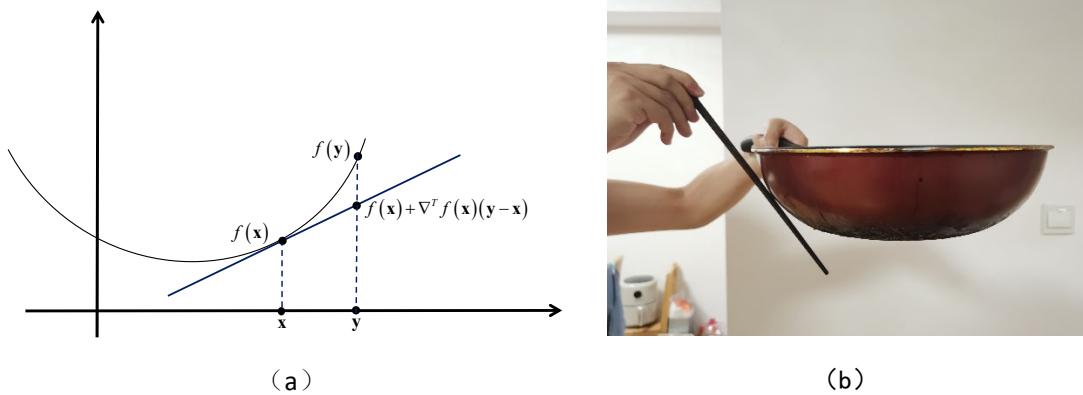


图 13-4: (a) 凸函数的一阶判定条件几何示意图；(b) 生活中，锅就是个典型的“凸函数”，相应地，凸函数一阶判定条件的意思就是：在锅曲面上任取一点，在该点放一根与锅曲面相切的棍子，那么这根棍子的全部一定都在锅的下面。

类似地，我们可以证明如下关于严格凸函数的判定命题：

命题 13.4 严格凸函数的一阶判定条件。设函数 $f(\mathbf{x}): \mathbb{R}^n \rightarrow \mathbb{R}$ 一阶可微，则函数 $f(\mathbf{x})$ 是严格凸函数的充要条件是： $\text{dom } f$ 是凸集并且 $\forall \mathbf{x}, \mathbf{y} \in \text{dom } f$ 且 $\mathbf{x} \neq \mathbf{y}$ 下式成立，

$$f(\mathbf{y}) > f(\mathbf{x}) + (\nabla f(\mathbf{x}))^T (\mathbf{y} - \mathbf{x}) \quad (13-5)$$

从定理 13.1，我们可以得到凸函数的一个非常重要的性质：

命题 13.5 可微凸函数的驻点就是该函数的全局最小值点。设凸函数 $f(\mathbf{x}): \mathbb{R}^n \rightarrow \mathbb{R}$ 一阶可微，在点 $\mathbf{x}^* \in \text{dom } f$ 处，若 $\nabla_{\mathbf{x}=\mathbf{x}^*} f(\mathbf{x}) = \mathbf{0}$ ，则 \mathbf{x}^* 是 $f(\mathbf{x})$ 的全局最小值点。

证明：

由于 $f(\mathbf{x})$ 为可微凸函数，根据定理 13.1 可知， $\forall \mathbf{x}, \mathbf{y} \in \text{dom } f$ 有，

$$f(\mathbf{y}) \geq f(\mathbf{x}) + (\nabla f(\mathbf{x}))^T (\mathbf{y} - \mathbf{x})$$

在点 $\mathbf{x} = \mathbf{x}^*$ 处则有， $f(\mathbf{y}) \geq f(\mathbf{x}^*) + (\nabla_{\mathbf{x}=\mathbf{x}^*} f(\mathbf{x}))^T (\mathbf{y} - \mathbf{x}^*)$ 。因为 $\nabla_{\mathbf{x}=\mathbf{x}^*} f(\mathbf{x}) = \mathbf{0}$ ，则有

$f(\mathbf{y}) \geq f(\mathbf{x}^*)$, 因此 \mathbf{x}^* 是 $f(\mathbf{x})$ 的全局最小值点。

定理 13.2 凸函数的二阶判定条件。如果函数 $f(\mathbf{x}): \mathbb{R}^n \rightarrow \mathbb{R}$ 二阶可微, 则函数 $f(\mathbf{x})$ 是凸函数的充要条件是: 它的定义域是凸集并且它的海森矩阵 $\nabla^2 f(\mathbf{x})$ 为半正定矩阵。

证明:

必要性。我们需要证明: 若 $f(\mathbf{x})$ 是凸函数, 则它的定义域是凸集并且它的海森矩阵 $\nabla^2 f(\mathbf{x})$ 为半正定矩阵。由于 $f(\mathbf{x})$ 是凸函数, 根据凸函数定义, $\text{dom}f$ 必然为凸集。由于 $f(\mathbf{x})$ 二阶可微, $\forall \mathbf{x} \in \text{dom}f$, 我们可在 \mathbf{x} 处进行二阶泰勒展开, 有,

$$f(\mathbf{x}+\mathbf{h}) = f(\mathbf{x}) + \mathbf{h}^T \nabla f(\mathbf{x}) + \frac{1}{2} \mathbf{h}^T \nabla^2 f(\mathbf{x}) \mathbf{h} + O(\|\mathbf{h}\|^2)$$

其中, $\mathbf{h} \neq \mathbf{0}$ 。由于已知 $f(\mathbf{x})$ 为凸函数, 根据定理 13.1 可知, $f(\mathbf{x}+\mathbf{h}) \geq f(\mathbf{x}) + \nabla^T f(\mathbf{x}) \mathbf{h}$ 。结合上式有 $\frac{1}{2} \mathbf{h}^T \nabla^2 f(\mathbf{x}) \mathbf{h} \geq 0$, 因此 $\nabla^2 f(\mathbf{x})$ 为半正定矩阵。

充分性。我们需要证明: 若 $f(\mathbf{x})$ 的定义域是凸集并且它的海森矩阵 $\nabla^2 f(\mathbf{x})$ 为半正定矩阵, 则 $f(\mathbf{x})$ 必为凸函数。 $\forall \mathbf{x}, \mathbf{y} \in \text{dom}f$, 根据泰勒展开有,

$$f(\mathbf{y}) = f(\mathbf{x}) + \nabla^T f(\mathbf{x})(\mathbf{y}-\mathbf{x}) + \frac{1}{2} (\mathbf{y}-\mathbf{x})^T \nabla^2 f(\mathbf{x}+t(\mathbf{y}-\mathbf{x})) (\mathbf{y}-\mathbf{x})$$

其中, $t \in [0,1]$ 是存在的某个数能使上式成立。由于 $\text{dom}f$ 为凸集, 根据凸集的性质可知 $\mathbf{x}+t(\mathbf{y}-\mathbf{x}) \in \text{dom}f$ 。根据已知条件, $f(\mathbf{x})$ 在其定义域内任意一点处的海森矩阵都为半正定矩阵, 则 $\nabla^2 f(\mathbf{x}+t(\mathbf{y}-\mathbf{x}))$ 为半正定矩阵, 则 $\frac{1}{2} (\mathbf{y}-\mathbf{x})^T \nabla^2 f(\mathbf{x}+t(\mathbf{y}-\mathbf{x})) (\mathbf{y}-\mathbf{x}) \geq 0$ 。再结合上述泰勒展开结果可知, $f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla^T f(\mathbf{x})(\mathbf{y}-\mathbf{x})$ 。根据定理 13.1 可知, $f(\mathbf{x})$ 为凸函数。

命题 13.6 严格凸函数的二阶判定条件。函数 $f(\mathbf{x}): \mathbb{R}^n \rightarrow \mathbb{R}$ 二阶可微, 若其定义域为凸集且其海森矩阵 $\nabla^2 f(\mathbf{x})$ 为正定矩阵, 则函数 $f(\mathbf{x})$ 为严格凸函数。

证明:

$\forall \mathbf{x}, \mathbf{y} \in \text{dom}f$, 根据泰勒展开有,

$$f(\mathbf{y}) = f(\mathbf{x}) + \nabla^T f(\mathbf{x})(\mathbf{y}-\mathbf{x}) + \frac{1}{2} (\mathbf{y}-\mathbf{x})^T \nabla^2 f(\mathbf{x}+t(\mathbf{y}-\mathbf{x})) (\mathbf{y}-\mathbf{x})$$

其中, $t \in [0,1]$ 是存在的某个数能使上式成立。由于 $\text{dom}f$ 为凸集, 根据凸集的性质可知

$\mathbf{x} + t(\mathbf{y} - \mathbf{x}) \in \text{dom}f$ 。根据已知条件, $f(\mathbf{x})$ 在其定义域内任意一点处的海森矩阵都为正定矩阵,

则 $\nabla^2 f(\mathbf{x} + t(\mathbf{y} - \mathbf{x}))$ 为正定矩阵, 则 $\frac{1}{2}(\mathbf{y} - \mathbf{x})^T \nabla^2 f(\mathbf{x} + t(\mathbf{y} - \mathbf{x}))(\mathbf{y} - \mathbf{x}) > 0$ 。结合泰勒展开结果

则有, $f(\mathbf{y}) > f(\mathbf{x}) + \nabla^T f(\mathbf{x})(\mathbf{y} - \mathbf{x})$ 。根据命题 13.4 可知, $f(\mathbf{x})$ 为严格凸函数。

需要格外注意的是, 命题 13.6 阐述的是判定一个函数为严格凸函数的充分条件, 但该条件并不是严格凸函数判定的必要条件。比如, $f(x) = x^4$ 为严格凸函数, 但其在 $x=0$ 处的海森矩阵 (即二阶导数 $f''(0)=0$) 并不是正定的 (对于只有一个元素的矩阵来说, 矩阵正定和该唯一元素大于零等价)。

命题 13.7 两个凸函数逐点求最大, 得到的函数依然是凸函数。即, 若 $f_1(\mathbf{x}), f_2(\mathbf{x})$ 为两个凸函数, 则 $f(\mathbf{x}) = \max\{f_1(\mathbf{x}), f_2(\mathbf{x})\}$ 也为凸函数。

证明:

设 $f_1(\mathbf{x}), f_2(\mathbf{x})$ 的定义域分别为 $\text{dom}f_1, \text{dom}f_2$, 因为这两个函数均为凸函数, 根据凸函数的定义 (定义 13.6) 可知, $\text{dom}f_1, \text{dom}f_2$ 均为凸集。 $f(\mathbf{x})$ 的定义域 $\text{dom}f$ 显然为 $f_1(\mathbf{x}), f_2(\mathbf{x})$ 定义域的交集, 即 $\text{dom}f = \text{dom}f_1 \cap \text{dom}f_2$ 。根据命题 13.1, 凸集的交集依然是凸集, 因此 $\text{dom}f$ 为凸集。另外, $\forall \mathbf{x}, \mathbf{y} \in \text{dom}f, \forall \theta \in [0, 1]$ 有,

$$\begin{aligned} f(\theta \mathbf{x} + (1-\theta)\mathbf{y}) &= \max\{f_1(\theta \mathbf{x} + (1-\theta)\mathbf{y}), f_2(\theta \mathbf{x} + (1-\theta)\mathbf{y})\} \\ &\leq \max\{\theta f_1(\mathbf{x}) + (1-\theta)f_1(\mathbf{y}), \theta f_2(\mathbf{x}) + (1-\theta)f_2(\mathbf{y})\} \\ &\leq \theta \max\{f_1(\mathbf{x}), f_2(\mathbf{x})\} + (1-\theta) \max\{f_1(\mathbf{y}), f_2(\mathbf{y})\} \\ &= \theta f(\mathbf{x}) + (1-\theta)f(\mathbf{y}) \end{aligned}$$

综合以上信息可知, $f(\mathbf{x})$ 为凸函数。

命题 13.8 两个凹函数逐点求最小, 得到的函数依然是凹函数。即, 若 $f_1(\mathbf{x}), f_2(\mathbf{x})$ 为两个凹函数, 则 $f(\mathbf{x}) = \min\{f_1(\mathbf{x}), f_2(\mathbf{x})\}$ 也为凹函数。

该命题的证明作为练习, 请读者自行完成 (提示, 可直接利用命题 13.7 的结论)。

命题 13.9 凸函数的非负组合依然是凸函数。假设 $f_1(\mathbf{x}), f_2(\mathbf{x}), \dots, f_m(\mathbf{x})$ 都是凸函数, 则,

$$f(\mathbf{x}) = \omega_1 f_1(\mathbf{x}) + \omega_2 f_2(\mathbf{x}) + \dots + \omega_m f_m(\mathbf{x}), \omega_i \geq 0, i=1, \dots, m \text{ 为凸函数。}$$

证明:

$\text{dom}f = \bigcap_{i=1}^m \text{dom}f_i$, 由于 $\text{dom}f_i$ 是凸集, 则 $\text{dom}f$ 是一系列凸集的交集, 根据命题 13.1,

$\text{dom}f$ 为凸集。另外, $\forall \mathbf{x}, \mathbf{y} \in \text{dom}f, \forall \theta \in [0, 1]$ 有,

$$\begin{aligned} f(\theta \mathbf{x} + (1-\theta)\mathbf{y}) &= \sum_{i=1}^m \omega_i f_i(\theta \mathbf{x} + (1-\theta)\mathbf{y}) \\ &\leq \sum_{i=1}^m \omega_i (\theta f_i(\mathbf{x}) + (1-\theta)f_i(\mathbf{y})) = \theta \sum_{i=1}^m \omega_i f_i(\mathbf{x}) + (1-\theta) \sum_{i=1}^m \omega_i f_i(\mathbf{y}) \\ &= \theta f(\mathbf{x}) + (1-\theta)f(\mathbf{y}) \end{aligned}$$

综上, $f(\mathbf{x})$ 为凸函数。

定义 13.7 二次函数 (Quadratic function)。具有如下形式的函数 $f(\mathbf{x}): \mathbb{R}^n \rightarrow \mathbb{R}$ 被称为二次函数,

$$f(\mathbf{x}) = \frac{1}{2} \mathbf{x}^T P \mathbf{x} + \mathbf{q}^T \mathbf{x} + r \quad (13-6)$$

其中, $P \in \mathbb{R}^{n \times n}$ 为实对称矩阵, $\mathbf{q} \in \mathbb{R}^{n \times 1}$, r 为实数。

对于式 13-6 中的二次函数 $f(\mathbf{x})$, 容易知道它是二次可微的, 且其定义域显然为凸集,

同时其海森矩阵为 $\nabla^2 f(\mathbf{x}) = P$ ⁹。根据定理 13.2 可知, 该二次函数为凸函数的充要条件为 P 为半正定矩阵; 根据命题 13.6 可知, 该二次函数为严格凸函数的充分条件为 P 为正定矩阵。

13.1.3 优化问题

定义 13.8 优化问题 (Optimization problem)。一般我们用如下形式来表示优化问题,

$$\begin{aligned} \mathbf{x}^* &= \arg \min_{\mathbf{x}} f_0(\mathbf{x}) \\ \text{subject to } &f_i(\mathbf{x}) \leq 0, i = 1, \dots, m \\ &h_i(\mathbf{x}) = 0, i = 1, \dots, p \end{aligned} \quad (13-7)$$

上式所表达的含义是, 我们想在所有满足条件 $f_i(\mathbf{x}) \leq 0, i = 1, \dots, m$ 和 $h_i(\mathbf{x}) = 0, i = 1, \dots, p$ 的 \mathbf{x} 中找到能够使函数 $f_0(\mathbf{x})$ 取得最小值的点 \mathbf{x}^* 。其中, $\mathbf{x} \in \mathbb{R}^n$ 称为优化变量 (optimization variable),

$f_0: \mathbb{R}^n \rightarrow \mathbb{R}$ 称为目标函数 (objective function), $f_i(\mathbf{x}) \leq 0, i = 1, \dots, m$ 称为不等式约束,

$h_i(\mathbf{x}) = 0, i = 1, \dots, p$ 称为等式约束; 如果 $m=p=0$, 即该问题没有任何约束, 则该问题称为无约束优化问题。

在式 13-7 中, 我们把目标函数与约束函数都能够有定义的优化变量取值集合称为优化问题的定义域 (domain), 记为 $\mathcal{D} = \bigcap_{i=0}^m \text{dom} f_i \cap \bigcap_{i=1}^p \text{dom} h_i$, 即 \mathcal{D} 是目标函数与所有约束函数定义域的交集。如果 $\mathbf{x} \in \mathcal{D}$ 且 \mathbf{x} 满足所有约束, 即 $f_i(\mathbf{x}) \leq 0, i = 1, \dots, m$, $h_i(\mathbf{x}) = 0, i = 1, \dots, p$, 则称 \mathbf{x} 为该问题的一个可行点 (feasible point)。当优化问题式 13-7 至少存在一个可行点时, 称该问题是可行的 (feasible), 否则称该问题是不可行的 (infeasible)。由所有可行点构成的集合称为该问题的可行集 (feasible set)。显然, 一个优化问题如果存在最优解的话, 最优解一定在该问题的可行集中取得。需要注意的是, 我们说一个优化问题是可行的, 这并不意味着该问题一定存在最优解。比如如下这个优化问题,

$$x^* = \arg \min_{\mathbf{x}} 2x, \text{ subject to } x \leq 0$$

⁹ 如果读者不熟悉向量与矩阵形式的求导操作, 请仔细阅读本书附录 F。

它的可行集显然是 $x \leq 0$ ，但在该可行集上目标函数 $2x$ 的取值没有下界 (unbounded below)，因此该问题的最优解不存在。

最优值 (optimal value)。最优值被定义为在约束条件下目标函数能取得的最小值，

$$v^* = \min \{f_0(\mathbf{x}) \mid f_i(\mathbf{x}) \leq 0, i = 1, \dots, m, h_i(\mathbf{x}) = 0, i = 1, \dots, p\} \quad (13-8)$$

显然，如果优化问题式 13-7 存在最优解 \mathbf{x}^* ，则最优值 $v^* = f_0(\mathbf{x}^*)$ ；如果问题式 13-7 是不可行的，我们定义 $v^* = +\infty$ ；如果问题式 13-7 的目标函数 $f_0(\mathbf{x})$ 在约束条件下的取值没有下界 (unbounded below)，我们定义 $v^* = -\infty$ 。

13.1.4 凸优化问题

定义 13.9 凸优化问题 (Convex optimization problem)。我们把如下形式的优化问题称为凸优化问题，

$$\begin{aligned} \mathbf{x}^* &= \arg \min_{\mathbf{x}} f_0(\mathbf{x}) \\ \text{subject to } &f_i(\mathbf{x}) \leq 0, i = 1, \dots, m \\ &\mathbf{a}_i^T \mathbf{x} - b_i = 0, i = 1, \dots, p \end{aligned} \quad (13-9)$$

其中， $f_i(\mathbf{x})$ 是凸函数。

显然，凸优化问题是优化问题的“子集”。通过对比一般优化问题的定义（定义 13.5）和凸优化问题的定义（定义 13.6），我们可以看出凸优化问题在一般优化问题的基础上还有三个额外要求：(1) 目标函数必须是凸函数；(2) 不等式约束函数是凸函数；(3) 等式约束函数必须是仿射函数 $h_i(\mathbf{x}) = \mathbf{a}_i^T \mathbf{x} - b_i$ 。凸优化问题的可行集一定是凸集¹⁰。

作为凸优化问题的一个典型代表，我们来介绍一下凸二次规划问题：

定义 13.10 凸二次规划问题 (Convex quadratic program problem)。我们把如下形式的凸优化问题称为凸二次规划问题，

$$\begin{aligned} \mathbf{x}^* &= \arg \min_{\mathbf{x}} \frac{1}{2} \mathbf{x}^T P \mathbf{x} + \mathbf{q}^T \mathbf{x} + r \\ \text{subject to } &G \mathbf{x} \leq \mathbf{h}, G \in \mathbb{R}^{m \times n} \\ &A \mathbf{x} = \mathbf{b}, A \in \mathbb{R}^{p \times n} \end{aligned} \quad (13-10)$$

其中， P 为半正定矩阵。

容易验证，凸二次规划问题当然是凸优化问题：1) 它的目标函数为二次函数，由于 P 为半正定矩阵，根据定理 13.2 可知该目标函数为凸函数；2) 不等式约束函数均为仿射函数，根据命题 13.2，它们当然也都是凸函数；(3) 等式约束函数均为仿射函数。

¹⁰ 如果要证明凸优化问题的可行集一定为凸集，还需要引入次水平集 (sub-level set) 的概念，这个概念与本书的其他内容无关，为了叙述简洁，我们就不再引入这个概念了。读者只需要要知道“凸优化问题的可行集一定为凸集”这个结论就可以了。

13.2 对偶

13.2.1 对偶函数

定义 13.11 拉格朗日函数 (Lagrangian)。如下函数 $l(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta}) : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p \rightarrow \mathbb{R}$ 称为定义 13.8 所描述的优化问题的拉格朗日函数,

$$l(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = f_0(\mathbf{x}) + \sum_{i=1}^m \alpha_i f_i(\mathbf{x}) + \sum_{i=1}^p \beta_i h_i(\mathbf{x}) \quad (13-11)$$

其中, $\boldsymbol{\alpha} = \{\alpha_i\}_{i=1}^m$, $\boldsymbol{\beta} = \{\beta_i\}_{i=1}^p$, 它们被称作对偶变量 (dual variables) 或拉格朗日乘子向量 (Lagrange multiplier vectors); 函数 $l(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta})$ 的定义域为 $\text{dom } l = \mathcal{D} \times \mathbb{R}^m \times \mathbb{R}^p$ 。

定义 13.12 拉格朗日对偶函数 (Lagrange dual function), 也简称为对偶函数 (dual function)。如下函数 $g(\boldsymbol{\alpha}, \boldsymbol{\beta}) : \mathbb{R}^m \times \mathbb{R}^p \rightarrow \mathbb{R}$ 称为拉格朗日函数式 13-11 的对偶函数,

$$g(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \min_{\mathbf{x} \in \mathcal{D}} l(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \min_{\mathbf{x} \in \mathcal{D}} \left(f_0(\mathbf{x}) + \sum_{i=1}^m \alpha_i f_i(\mathbf{x}) + \sum_{i=1}^p \beta_i h_i(\mathbf{x}) \right) \quad (13-12)$$

也就是说, 对偶函数 $g(\boldsymbol{\alpha}, \boldsymbol{\beta})$ 是关于对偶变量 $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ 的函数, 其值是在固定 $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ 时, 拉格朗日函数 $l(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta})$ 在变化 \mathbf{x} 时所能取得的最小值。若在 $(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)$ 处, $l(\mathbf{x}, \boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)$ 关于 \mathbf{x} 没有下界, 则定义 $g(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0) = -\infty$ 。

命题 13.10 式 13-12 所定义的拉格朗日对偶函数为凹函数。

证明:

式 13-12 所定义的对偶函数 $g(\boldsymbol{\alpha}, \boldsymbol{\beta})$ 可变形为如下形式,

$$g(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \min_{\mathbf{x} \in \mathcal{D}} \left(f_1(\mathbf{x}) f_2(\mathbf{x}) \cdots f_m(\mathbf{x}) h_1(\mathbf{x}) h_2(\mathbf{x}) \cdots h_p(\mathbf{x}) \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_m \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} + f_0(\mathbf{x}) \right)$$

显然, $g(\boldsymbol{\alpha}, \boldsymbol{\beta})$ 为一系列关于 $\begin{pmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{pmatrix}$ 的仿射函数逐点求最小得到的; 而根据命题 13.2, 仿射函数为凹函数。这也就是说, 函数 $g(\boldsymbol{\alpha}, \boldsymbol{\beta})$ 为一系列关于 $\begin{pmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{pmatrix}$ 的凹函数逐点求最小而得到的, 根据

命题 13.8 可知, $g(\boldsymbol{\alpha}, \boldsymbol{\beta})$ 为凹函数。

命题 13.11 式 13-12 所定义的拉格朗日对偶函数的值是优化问题式 13-7 的最优值 v^* 的

下界。即 $\forall \alpha \geq \mathbf{0}$, $\forall \beta$, 有 $g(\alpha, \beta) \leq v^*$ 。

证明：

先考虑优化问题式 13-7 为可行的情况。假设 $\tilde{\mathbf{x}}$ 是优化问题式 13-7 的一个可行点，根据可行点的定义可知，

$$f_i(\tilde{\mathbf{x}}) \leq 0, i = 1, 2, \dots, m, \quad h_i(\tilde{\mathbf{x}}) = 0, i = 1, 2, \dots, p$$

则有，

$$\sum_{i=1}^m \alpha_i f_i(\tilde{\mathbf{x}}) + \sum_{i=1}^m \beta_i h_i(\tilde{\mathbf{x}}) \leq 0,$$

因此，

$$l(\tilde{\mathbf{x}}, \alpha, \beta) = f_0(\tilde{\mathbf{x}}) + \sum_{i=1}^m \alpha_i f_i(\tilde{\mathbf{x}}) + \sum_{i=1}^m \beta_i h_i(\tilde{\mathbf{x}}) \leq f_0(\tilde{\mathbf{x}}),$$

因此，

$$g(\alpha, \beta) = \min_{\mathbf{x} \in \mathcal{D}} l(\mathbf{x}, \alpha, \beta) \leq l(\tilde{\mathbf{x}}, \alpha, \beta) \leq f_0(\tilde{\mathbf{x}})$$

由于上式对任意的可行点 $\tilde{\mathbf{x}}$ 都成立，而优化问题式 13-7 的最优值 v^* 当然也是在某个可行点处取得的，因此有 $g(\alpha, \beta) \leq v^*$ 。

再先考虑优化问题式 13-7 为不可行的情况。若优化问题式 13-7 不可行，则其最优值 $v^* = +\infty$ ，则显然有 $g(\alpha, \beta) \leq v^*$ 。

需要注意的是，命题 13.11 对 α 的取值是有要求的，要求 $\alpha \geq \mathbf{0}$ ；只有满足 $\alpha \geq \mathbf{0}$ 这个条件，上述推导才能成立。

为了帮助读者能够深刻理解拉格朗日对偶函数以及命题 13.11 的涵义，我们举两个简单的例子。

例 13.2：考虑如下优化问题，

$$\mathbf{x}^* = \arg \min_{\mathbf{x}} \mathbf{x}^T \mathbf{x}, \text{ subject to } A \mathbf{x} = \mathbf{b}, A \in \mathbb{R}^{p \times n}$$

这个优化问题没有不等式约束，只有 p 个等式约束。它的拉格朗日函数为，

$$l(\mathbf{x}, \beta) = \mathbf{x}^T \mathbf{x} + \beta^T (A \mathbf{x} - \mathbf{b})$$

其中 $\beta \in \mathbb{R}^p$, l 的定义域为 $\text{dom } l = \mathbb{R}^n \times \mathbb{R}^p$ 。相应地，拉格朗日对偶函数为 $g(\beta) = \min_{\mathbf{x} \in \mathbb{R}^n} l(\mathbf{x}, \beta)$ 。

考虑 $l(\mathbf{x}, \beta)$ ，对于固定的 β ， $l(\mathbf{x}, \beta)$ 为关于 \mathbf{x} 的凸二次函数。因此，我们可以找到使 $l(\mathbf{x}, \beta)$ 最小化的 \mathbf{x} ，即解 $\nabla_{\mathbf{x}} l(\mathbf{x}, \beta) = \mathbf{0}$ ，得到 $\mathbf{x} = -\frac{1}{2} A^T \beta$ 。因此，对偶函数 $g(\beta)$ 为，

$$g(\beta) = l\left(-\frac{1}{2} A^T \beta, \beta\right) = -\frac{1}{4} \beta^T A A^T \beta - \beta^T \mathbf{b}$$

该函数显然为一个凹二次函数。由命题 13.11 可知， $\forall \beta \in \mathbb{R}^p$ 有，

$$g(\beta) = -\frac{1}{4} \beta^T A A^T \beta - \beta^T \mathbf{b} \leq \min_{\mathbf{x}} (\mathbf{x}^T \mathbf{x} \mid A \mathbf{x} = \mathbf{b})$$

例 13.3: 考虑如下优化问题,

$$\begin{aligned} \mathbf{x}^* &= \arg \min_{\mathbf{x}} \mathbf{c}^T \mathbf{x} \\ \text{subject to } \mathbf{x} &\geq \mathbf{0} \\ A \mathbf{x} &= \mathbf{b}, A \in \mathbb{R}^{p \times n} \end{aligned}$$

其中, $\mathbf{x} \in \mathbb{R}^n$ 。

首先, 把该优化问题写成标准优化问题的形式,

$$\begin{aligned} \mathbf{x}^* &= \arg \min_{\mathbf{x}} \mathbf{c}^T \mathbf{x} \\ \text{subject to } -\mathbf{x} &\leq \mathbf{0} \\ A \mathbf{x} - \mathbf{b} &= \mathbf{0}, A \in \mathbb{R}^{p \times n} \end{aligned}$$

它的拉格朗日函数为,

$$l(\mathbf{x}, \alpha, \beta) = \mathbf{c}^T \mathbf{x} - \alpha^T \mathbf{x} + \beta^T (A \mathbf{x} - \mathbf{b})$$

其中 $\alpha \in \mathbb{R}^n$, $\beta \in \mathbb{R}^p$, l 的定义域为 $\text{dom } l = \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^p$ 。相应地, 拉格朗日对偶函数为,

$$g(\alpha, \beta) = \min_{\mathbf{x} \in \mathbb{R}^n} l(\mathbf{x}, \alpha, \beta) = -\beta^T \mathbf{b} + \min_{\mathbf{x} \in \mathbb{R}^n} (\mathbf{c} + A^T \beta - \alpha)^T \mathbf{x}$$

上式关于 \mathbf{x} 的部分为关于 \mathbf{x} 的线性函数, 只要 $\mathbf{c} + A^T \beta - \alpha \neq \mathbf{0}$, 那么 $(\mathbf{c} + A^T \beta - \alpha)^T \mathbf{x}$ 必然是没有下界的, 因此可知,

$$g(\alpha, \beta) = \begin{cases} -\beta^T \mathbf{b}, & \text{if } \mathbf{c} + A^T \beta - \alpha = \mathbf{0} \\ -\infty, & \text{otherwise} \end{cases}$$

因此, 当 $\alpha \geq \mathbf{0}$ 且 $\mathbf{c} + A^T \beta - \alpha = \mathbf{0}$ 时, 对偶函数 $g(\alpha, \beta)$ 给出了原始优化问题最优值的一个非平凡下界 $-\beta^T \mathbf{b}$ 。

13.2.2 对偶问题

拉格朗日对偶函数 (式 13-12) 给出了优化问题 (式 13-7) 最优值 v^* 的下界。这些下界的取值是决定于 α 和 β 的。实际上, 这些由对偶函数所确定的优化问题最优值的下界中, 最大的那个下界是“最有意义的”。为什么这么说呢? 我们可以拿生活中的一个例子作为类比。假如你想在上海买一套房子, 地段、楼层、户型、面积等因素构成了一系列约束条件。你当然很想知道买这样一套房子最少得准备多少钱 (最优值)。为了解除心中的疑惑, 你找了 10 个朋友 (对偶函数), 他们每个人都是靠谱的、值得信赖的, 也就是说, 他们对你买房子最少花多少钱这件事所给出的估计 (最优值的下界) 都是正确的。A 说你最少得花 10 块, B 说你最少要花 100 块, C 说你最少要花 1 万块……。这 10 个朋友当中 E 说的最高, 他说你最少得准备 500 万。那么, 这 10 个朋友所提供的信息当中, 谁的信息对你来说是最有价值

的呢？显然是 E。因此，接下去我们要回答一个自然引出的问题：由对偶函数 $g(\alpha, \beta)$ 所确定出的最优（最大）下界是什么？即要求解如下拉格朗日对偶问题：

定义 13.13 拉格朗日对偶问题 (Lagrange dual problem)。我们称如下优化问题为定义 13.8 所定义的优化问题的拉格朗日对偶问题，

$$\begin{aligned} \alpha^*, \beta^* &= \arg \max_{\alpha, \beta} g(\alpha, \beta) \\ \text{subject to } \alpha &\geq 0 \end{aligned} \quad (13-13)$$

其中， $g(\alpha, \beta)$ 为由式 13-12 所定义的拉格朗日对偶函数。问题式 13-13 的最优解 (α^*, β^*) 被称作是对偶最优的 (dual optimal)，或最优的拉格朗日乘子 (optimal Lagrange multipliers)。

与对偶问题相对应的由定义 13.8 所定义的优化问题也被称为对偶问题的原问题 (primal problem)。如果 $\alpha \geq 0$ 且 $g(\alpha, \beta) > -\infty$ ，则 (α, β) 被称作是对偶可行 (dual feasible) 的，即此时 (α, β) 是优化问题式 13-13 的一个可行点。只有 (α, β) 是对偶可行的， $g(\alpha, \beta)$ 所定义的值才是原问题的一个非平凡的下界。

需要注意的是，不论原优化问题 (定义 13.8) 是否为凸优化问题，式 13-13 所描述的拉格朗日对偶问题一定是一个凸优化问题：1) 它的目标是要最大化一个凹函数，这相当于是要最小化一个凸函数，因此如果写成标准优化问题形式的话，其目标函数为凸函数；2) 其不等式约束函数均为仿射函数，当然也都是凸函数。

命题 13.12 优化问题的弱对偶性 (weak duality)。设式 13-13 所定义的对偶问题的最优值为 d^* ，即 $d^* = \max_{\alpha, \beta} g(\alpha, \beta)$, subject to $\alpha \geq 0$ ，则 d^* 是原问题最优值 v^* 由对偶函数 $g(\alpha, \beta)$ 所确定的最大的下界，因此必有 $d^* \leq v^*$ 。这个性质被称作优化问题的弱对偶性。

13.2.3 强对偶性与斯莱特条件

如前所述，由定义 13.8 所定义的一般化的优化问题具有弱对偶性，即原问题的最优值 v^* 和其对偶问题的最优值 d^* 之间满足关系，

$$d^* \leq v^* \quad (13-14)$$

我们现在想知道在什么样的条件下式 13-14 中的等号可以成立呢？若式 13-14 中的等号成立，即一个优化问题 (由定义 13.8 所定义) 的最优值和它的对偶问题的最优值相等，我们称该优化问题具有强对偶性 (strong duality)。显然，我们需要为一般化的优化问题附加一些额外的限制条件，才能使得到的优化问题具有强对偶性。一个广泛使用的强对偶判定条件便是斯莱特 (Slater) 条件^[2]：

定理 13.3 斯莱特条件 (Slater condition)。如果一个优化问题为由定义 13.9 所定义的凸优化问题，即该优化问题具有如下形式，

$$\begin{aligned} \mathbf{x}^* &= \arg \min_{\mathbf{x}} f_0(\mathbf{x}) \\ \text{subject to } f_i(\mathbf{x}) &\leq 0, i = 1, \dots, m \\ A\mathbf{x} &= \mathbf{b}, A \in \mathbb{R}^{p \times n} \end{aligned} \quad (13-15)$$

其中 $f_i(\mathbf{x}), i = 0, \dots, m$ 为凸函数。若至少存在一点 $\mathbf{x} \in \text{relint}\mathcal{D}$ (相对内部的定义见式 13-1) 使得,

$$f_i(\mathbf{x}) < 0, i = 1, \dots, m, A\mathbf{x} = \mathbf{b} \quad (13-16)$$

则该优化问题必具有强对偶性。该定理的证明需要较多细节且与本书的主线内容关系不大, 这里就不给出了, 感兴趣的读者可以参见^[1]。

斯莱特条件要求要至少存在一个 $\mathbf{x} \in \text{relint}\mathcal{D}$ 严格满足所有不等式约束 (小于号严格成立), 但如果该不等式约束函数为仿射函数, 这个条件可以放松一下: 只要满足该不等式约束即可 (可以是小于等于号), 而不一定是严格满足。这样便有了针对仿射型不等式约束的修正斯莱特条件:

定理 13.4 修正斯莱特条件 (refined Slater condition)。如果凸优化问题式 13-15 的前 k 个不等式约束函数 f_1, f_2, \dots, f_k 为仿射函数, 若至少存在一点 $\mathbf{x} \in \text{relint}\mathcal{D}$ 使得,

$$f_i(\mathbf{x}) \leq 0, i = 1, \dots, k, f_i(\mathbf{x}) < 0, i = k+1, \dots, m, A\mathbf{x} = \mathbf{b} \quad (13-17)$$

则该优化问题具有强对偶性。

根据定理 13.4, 我们自然会有如下命题成立:

命题 13.13 约束函数全为仿射函数的凸优化问题的斯莱特条件。如果一个凸优化问题的约束函数 (包括不等式约束和等式约束) 都是仿射函数而且其目标函数的定义域 $\text{dom}f_0$ 为开集, 那么只要该问题是可行的 (至少存在一个可行点), 它必满足 (修正) 斯莱特条件, 即它具有强对偶性。

证明:

根据已知条件, 该凸优化问题的不等式约束函数 f_1, f_2, \dots, f_m 都为仿射函数, 等式约束函数为仿射函数 $A\mathbf{x} = \mathbf{b}$ 。仿射函数的定义域均为 \mathbb{R}^n 。这样, 该优化问题的定义域 \mathcal{D} 为 $\text{dom}f_0$ 与所有约束函数定义域的交集, 则 $\mathcal{D} = \text{dom}f_0 \cap \mathbb{R}^n \cap \mathbb{R}^n \cdots \cap \mathbb{R}^n = \text{dom}f_0$, 而 $\text{dom}f_0$ 又是开集, 则 \mathcal{D} 为开集, 则 $\text{relint}\mathcal{D} = \mathcal{D}$ 。若该问题可行, 则说明至少有一点 $\mathbf{x} \in \mathcal{D} = \text{relint}\mathcal{D}$ 满足所有约束条件, 即,

$$f_i(\mathbf{x}) \leq 0, i = 1, \dots, m, A\mathbf{x} = \mathbf{b}$$

根据定理 13.4 可知, 该凸优化问题满足修正斯莱特条件, 因此它具有强对偶性。

命题 13.14 考虑形如式 13-10 所定义的凸二次规划问题, 如果该问题可行, 则其必然具有强对偶性。这个结论可由命题 13.13 直接得出。

需要注意: 斯莱特条件是一个优化问题具有强对偶性的充分条件, 但它不是必要条件。

我们通过一个具体例子来进一步理解一下斯莱特条件。考虑例 13.2, 原问题为,

$$\mathbf{x}^* = \arg \min_{\mathbf{x}} \mathbf{x}^T \mathbf{x}, \text{subject to } A\mathbf{x} = \mathbf{b}, A \in \mathbb{R}^{p \times n}$$

它的对偶问题为,

$$\underset{\beta}{\text{maximize}} -\frac{1}{4} \beta^T A A^T \beta - \beta^T b$$

原问题为凸优化问题, 没有不等式约束, 目标函数的定义域为 \mathbb{R}^n , 其为开集, 那么根据命题 13.13, 只需要原问题是可行的它便是强对偶的。对于这个问题来说, 原问题是可行的等价于 $Ax = b, A \in \mathbb{R}^{p \times n}$ 有解, 即只需要 $\text{rank}(A) = \text{rank}([A \ b])$ 即可。

13.2.4 强弱对偶性的“最大-最小”刻画

式 13-11 给出了原问题式 13-7 的拉格朗日函数 $l(x, \alpha, \beta) : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p \rightarrow \mathbb{R}$, 由该函数的定义可知,

$$\max_{\alpha \geq 0, \beta} l(x, \alpha, \beta) = \begin{cases} f_0(x), & \text{if } f_i(x) \leq 0, i = 1, \dots, m, h_i(x) = 0, i = 1, \dots, p \\ +\infty, & \text{otherwise} \end{cases}$$

也就是说, 在条件 $f_i(x) \leq 0, i = 1, \dots, m, h_i(x) = 0, i = 1, \dots, p$ 下, $\max_{\alpha \geq 0, \beta} l(x, \alpha, \beta) = f_0(x)$, 而该条件实际上是原问题式 13-7 的可行条件。因此不难理解, 原问题式 13-7 的最优值 v^* 可表达为,

$$v^* = \min_{x \in \mathcal{D}} \max_{\alpha \geq 0, \beta} l(x, \alpha, \beta)$$

另一方面, 结合对偶问题的定义(式 13-13)以及拉格朗日对偶函数的定义(式 13-12), 可知对偶问题的最优值 d^* 可表达为,

$$d^* = \max_{\alpha \geq 0, \beta} \min_{x \in \mathcal{D}} l(x, \alpha, \beta)$$

根据优化问题的弱对偶性质(命题 13.12)可知 $d^* \leq v^*$, 即以下不等式成立,

$$\max_{\alpha \geq 0, \beta} \min_{x \in \mathcal{D}} l(x, \alpha, \beta) \leq \min_{x \in \mathcal{D}} \max_{\alpha \geq 0, \beta} l(x, \alpha, \beta) \quad (13-18)$$

更进一步, 如果原问题具有强对偶性, 则意味着原问题的最优值 v^* 和对偶问题的最优值 d^* 相等, 则有,

$$\max_{\alpha \geq 0, \beta} \min_{x \in \mathcal{D}} l(x, \alpha, \beta) = \min_{x \in \mathcal{D}} \max_{\alpha \geq 0, \beta} l(x, \alpha, \beta) \quad (13-19)$$

也就是说, 如果原问题具有强对偶性, 对其拉格朗日函数在优化变量 x 上求最小以及在对偶变量 $(\alpha \geq 0, \beta)$ 上求最大的两个操作可以交换顺序。

13.2.5 KKT 最优条件

假设 x 为原问题的一个可行点, (α, β) 为对偶问题的对偶可行点, 我们把此时原问题目标函数值 $f_0(x)$ 与对偶问题目标函数值 $g(\alpha, \beta)$ 之差,

$$f_0(\mathbf{x}) - g(\boldsymbol{\alpha}, \boldsymbol{\beta})$$

称为对偶间隔 (duality gap)。

由原问题与对偶问题的性质容易知道：

命题 13.15 对于给定的一对可行优化变量 \mathbf{x}_0 与对偶可行变量 $(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)$ ，若相应的对偶间隔为 0，即 $f_0(\mathbf{x}_0) = g(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)$ ，则 \mathbf{x}_0 是原问题的最优解， $(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)$ 是对偶问题的最优解，并且原问题是强对偶的。该结论的证明作为练习，请读者自行完成。

设原问题是强对偶的。假设 \mathbf{x}^* 是原问题的最优解， $(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$ 是对偶问题的最优解，则有，

$$\begin{aligned} f_0(\mathbf{x}^*) &= g(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*) \\ &= \min_{\mathbf{x}} \left(f_0(\mathbf{x}) + \sum_{i=1}^m \alpha_i^* f_i(\mathbf{x}) + \sum_{i=1}^p \beta_i^* h_i(\mathbf{x}) \right) \\ &\leq f_0(\mathbf{x}^*) + \sum_{i=1}^m \alpha_i^* f_i(\mathbf{x}^*) + \sum_{i=1}^p \beta_i^* h_i(\mathbf{x}^*) \\ &\leq f_0(\mathbf{x}^*) \end{aligned} \tag{13-20}$$

第一行等号成立是因为原问题具有强对偶性，则原问题的最优值 $f_0(\mathbf{x}^*)$ 等于对偶问题的最优值 $g(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$ ；第二行等号成立，根据的是对偶函数的定义；第三行小于等于号成立，是因为当遍历所有 \mathbf{x} 之后拉格朗日函数的最小值当然要小于等于该函数在点 \mathbf{x}^* 处的值；最后一行等号成立，是因为 $\alpha_i^* \geq 0$, $f_i(\mathbf{x}^*) \leq 0, i=1, \dots, m$, $h_i(\mathbf{x}^*) = 0, i=1, \dots, p$ 。由于上述一系列推导链条的两端都是 $f_0(\mathbf{x}^*)$ ，因此两个小于等于号实际上都是等号！继而我们可以得到一些有用的结论：

- 1) 由于第 3 行的小于等于号实际上是等号，因此 \mathbf{x}^* 就是拉格朗日函数 $l(\mathbf{x}; \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$ 的最小值点；
- 2) 由于第 4 行的不等号实际上也为等号，因此， $\sum_{i=1}^m \alpha_i^* f_i(\mathbf{x}^*)$ 必为 0；又由于已知 $\alpha_i^* f_i(\mathbf{x}^*) \leq 0, i=1, \dots, m$ ，因此必有 $\alpha_i^* f_i(\mathbf{x}^*) = 0, i=1, \dots, m$ 。

基于上述分析，我们可引出优化领域中的一个重要结论：KKT 条件 (Karush-Kuhn-Tucker conditions)。在描述 KKT 条件时，首先要假定优化问题 (定义 13.8) 中的目标函数和所有约束函数都是可微的，即要假设 $f_0, f_1, \dots, f_m, h_1, \dots, h_p$ 都是可微的。

定理 13.5 针对一般优化问题的 KKT 条件。 设原优化问题是强对偶的。假设 \mathbf{x}^* 是原问题的最优解， $(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$ 是对偶问题的最优解，由式 13-20 所得到的结论 1) 我们知道， \mathbf{x}^* 是就是拉格朗日函数 $l(\mathbf{x}; \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$ 的最小值点，由于 $l(\mathbf{x}; \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$ 是可微的，因此它在 \mathbf{x}^* 处的梯度

$\nabla_{\mathbf{x}=\mathbf{x}^*} l(\mathbf{x}; \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$ 为 $\mathbf{0}$, 即, $\nabla_{\mathbf{x}=\mathbf{x}^*} f_0(\mathbf{x}) + \sum_{i=1}^m \alpha_i^* \nabla_{\mathbf{x}=\mathbf{x}^*} f_i(\mathbf{x}) + \sum_{i=1}^p \beta_i^* \nabla_{\mathbf{x}=\mathbf{x}^*} h_i(\mathbf{x}) = \mathbf{0}$ 。综合在一起我们有,

$$\begin{aligned} f_i(\mathbf{x}^*) &\leq 0, i = 1, \dots, m \\ h_i(\mathbf{x}^*) &= 0, i = 1, \dots, p \\ \alpha_i^* &\geq 0, i = 1, \dots, m \\ \alpha_i^* f_i(\mathbf{x}^*) &= 0, i = 1, \dots, m \end{aligned} \tag{13-21}$$

$$\nabla_{\mathbf{x}=\mathbf{x}^*} f_0(\mathbf{x}) + \sum_{i=1}^m \alpha_i^* \nabla_{\mathbf{x}=\mathbf{x}^*} f_i(\mathbf{x}) + \sum_{i=1}^p \beta_i^* \nabla_{\mathbf{x}=\mathbf{x}^*} h_i(\mathbf{x}) = \mathbf{0}$$

式 13-21 便称为 KKT 条件。总结下来, 对于任意的由定义 13.8 所定义的优化问题, 如果它的目标函数和所有约束函数都是可微的, 并且该优化问题是强对偶的, 那么任意一对原问题和对偶问题的最优解 \mathbf{x}^* 、 $(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$ 都满足式 13-21 所列的 KKT 条件。

定理 13.6 针对凸优化问题的 KKT 条件。 假设一个凸优化问题的目标函数和约束函数都可微, 且该问题满足斯莱特条件。在这种情况下, KKT 条件是一对原问题可行点 \mathbf{x}_0 与对偶问题对偶可行点 $(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)$ 分别为原问题最优与对偶最优的充要条件。

证明:

必要性, 即要证明: 如果 \mathbf{x}_0 和 $(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)$ 分别为原问题和对偶问题的最优解, 则它们满足 KKT 条件。由于已知原问题满足斯莱特条件, 则可知原问题具有强对偶性; 又因为 \mathbf{x}_0 和 $(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)$ 分别为原问题和对偶问题的最优解且原问题所有约束函数都可微, 根据定理 13.5, \mathbf{x}_0 和 $(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)$ 满足 KKT 条件。

充分性, 即要证明: 如果 \mathbf{x}_0 和 $(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)$ 满足 KKT 条件, 则它们必然分别是原问题和对偶问题的最优解。由于 \mathbf{x}_0 和 $(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)$ 满足 KKT 条件, 则有,

$$\begin{aligned} f_i(\mathbf{x}_0) &\leq 0, i = 1, \dots, m \\ h_i(\mathbf{x}_0) &= 0, i = 1, \dots, p \\ \alpha_{0i} &\geq 0, i = 1, \dots, m \\ \alpha_{0i} f_i(\mathbf{x}_0) &= 0, i = 1, \dots, m \\ \nabla_{\mathbf{x}=\mathbf{x}_0} f_0(\mathbf{x}) + \sum_{i=1}^m \alpha_{0i} \nabla_{\mathbf{x}=\mathbf{x}_0} f_i(\mathbf{x}) + \sum_{i=1}^p \beta_{0i} \nabla_{\mathbf{x}=\mathbf{x}_0} h_i(\mathbf{x}) &= \mathbf{0} \end{aligned}$$

由于原问题为凸优化问题, 则可知 f_0, f_1, \dots, f_m 都是凸函数且 h_1, \dots, h_p 都为仿射函数(见凸优化问题定义 13.9)。又从 KKT 条件第 3 条知道, $\alpha_{0i} \geq 0$, 则可知拉格朗日函数,

$$l(\mathbf{x}, \boldsymbol{\alpha}_0, \boldsymbol{\beta}_0) = f_0(\mathbf{x}) + \sum_{i=1}^m \alpha_{0i} f_i(\mathbf{x}) + \sum_{i=1}^p \beta_{0i} h_i(\mathbf{x})$$

为关于 \mathbf{x} 的凸函数(证明留作练习请读者完成); 同时, KKT 条件的最后一条表明了函数

$l(\mathbf{x}; \boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)$ 在 \mathbf{x}_0 处的梯度为 $\mathbf{0}$, 因此必有 \mathbf{x}_0 为 $l(\mathbf{x}; \boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)$ 的最小值点。这样我们便有,

$$g(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0) = \min_{\mathbf{x}} l(\mathbf{x}, \boldsymbol{\alpha}_0, \boldsymbol{\beta}_0) = l(\mathbf{x}_0, \boldsymbol{\alpha}_0, \boldsymbol{\beta}_0) = f_0(\mathbf{x}_0) + \sum_{i=1}^m \alpha_{0i} f_i(\mathbf{x}_0) + \sum_{i=1}^p \beta_{0i} h_i(\mathbf{x}_0) = f_0(\mathbf{x}_0)$$

其中, 最后一个等式应用了 $h_i(\mathbf{x}_0) = 0, \alpha_{0i} f_i(\mathbf{x}_0) = 0$ 这两个已知条件(KKT 条件中的第 2、4 条)。

这说明在 \mathbf{x}_0 和 $(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)$ 处, 对偶间隔为 0, 再根据命题 13.15 可知, \mathbf{x}_0 和 $(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)$ 分别是原问题的最优解和对偶问题的最优解。

KKT 条件在优化领域占有重要地位。在一些少量的特殊情况下, 可以解析求解出 KKT 条件, 从而得到优化问题的最优解。更加广泛地, 很多解决凸优化问题的算法可以被理解为求解 KKT 条件的方法。KKT 条件最初以哈罗德·库恩 (Harold W. Kuhn) 和阿尔伯特·塔克 (Albert W. Tucker) 的名字命名, 他们于 1951 年提出了这些条件^[3]。后来, 学者们发现威廉·卡鲁什 (William Karush) 在他 1939 年的硕士论文中已经阐述了这些条件^[4]。最终, 这项数学成果根据他们三个人姓氏的首字母被命名为 KKT (Karush-Kuhn-Tucker) 条件。

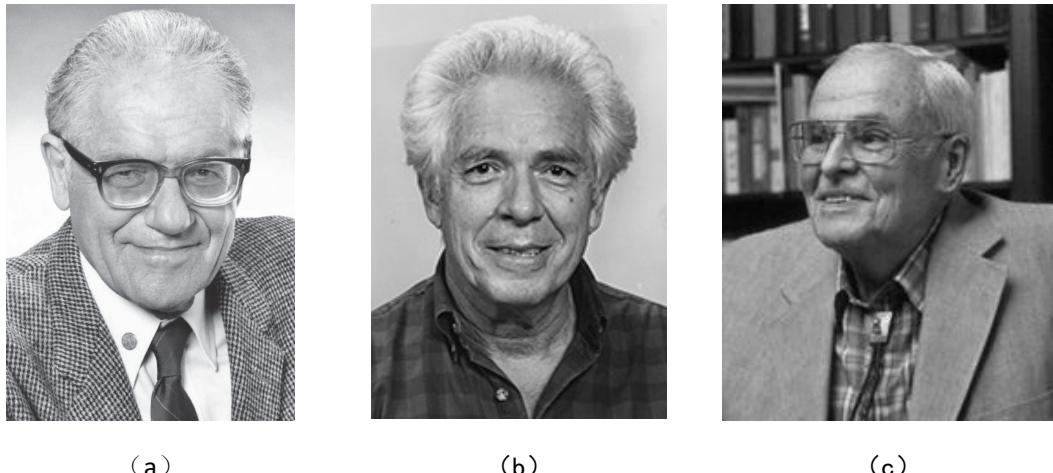


图 13-5: KKT 条件的三个提出者。(a) 威廉·卡鲁什 (William Karush, 1917 年 3 月 1 日至 1997 年 2 月 22 日), 美国加州州立大学北岭分校的数学教授, 以对 KKT 条件的贡献而闻名;在其硕士论文中, 他首次提出了不等式约束问题最优解的必要条件。(b) 哈罗德·库恩 (Harold W. Kuhn, 1925 年 7 月 29 日至 2014 年 7 月 2 日), 研究博弈论的美国数学家, 普林斯顿大学前数学名誉教授, 他与 David Gale 和 Albert William Tucker 一起获得了 1980 年冯诺依曼理论奖; 他与约翰·福布斯·纳什 (John Forbes Nash) 是多年朋友和同事, 他也是纳什获得诺贝尔奖委员会关注的关键人物 (纳什于 1994 年获诺贝尔经济学奖); 他在 2001 年拍摄的改编自纳什生活的电影《美丽心灵》中担任数学顾问。(c) 阿尔伯特·威廉·塔克 (Albert William Tucker, 1905 年 11 月 28 日至 1995 年 1 月 25 日), 加拿大数学家, 在拓扑学、博弈论和非线性规划方面做出了重要贡献。塔克出生于加拿大安大略省的奥沙瓦, 1928 年在多伦多大学获得学士学位, 1929 年在同一所大学获得硕士学位; 1932 年, 他在普林斯顿大学获得博士学位; 1932 年至 1933 年, 他先后在剑桥大学、哈佛大学和芝加哥大学担任研究员; 他于 1933 年回

到普林斯顿大学，直到 1974 年退休。

例 13.4: 考虑如下等式约束的凸二次规划问题。

$$\begin{aligned}\mathbf{x}^* &= \arg \min_{\mathbf{x}} \frac{1}{2} \mathbf{x}^T P \mathbf{x} + \mathbf{q}^T \mathbf{x} + r \\ \text{subject to } A \mathbf{x} &= \mathbf{b}, A \in \mathbb{R}^{p \times n}\end{aligned}$$

其中， P 为半正定矩阵。

该问题的拉格朗日函数为， $l(\mathbf{x}, \boldsymbol{\beta}) = \frac{1}{2} \mathbf{x}^T P \mathbf{x} + \mathbf{q}^T \mathbf{x} + r + \boldsymbol{\beta}^T (A \mathbf{x} - \mathbf{b})$ 。设原问题的最优解为 \mathbf{x}^* ，对偶问题的最优解为 $\boldsymbol{\beta}^*$ 。KKT 条件的最后一条在这个具体的问题中为，

$$\nabla_{\mathbf{x}=\mathbf{x}^*} l(\mathbf{x}, \boldsymbol{\beta}^*) = P \mathbf{x}^* + \mathbf{q} + A^T \boldsymbol{\beta}^* = \mathbf{0}$$

KKT 条件的第二条在这个具体问题中为 $A \mathbf{x}^* = \mathbf{b}$ 。这两个条件可组合表达为，

$$\begin{bmatrix} P & A^T \\ A & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{x}^* \\ \boldsymbol{\beta}^* \end{bmatrix} = \begin{bmatrix} -\mathbf{q} \\ \mathbf{b} \end{bmatrix}$$

通过解该线性方程组便可得到原问题以及对偶问题的最优解。

例 13.5: 考虑如下优化问题，

$$\begin{aligned}x_1, x_2 &= \arg \min_{x_1, x_2} x_1^2 + x_2^2 \\ \text{subject to } x_2 &\leq b \\ x_1 + x_2 &= 1\end{aligned}$$

解：

容易验证，该优化问题是一个凸优化问题且满足斯莱特条件。该优化问题的拉格朗日函数为， $l(x_1, x_2, \alpha, \beta) = x_1^2 + x_2^2 + \alpha(x_2 - b) + \beta(1 - x_1 - x_2)$ 。只要我们能够找到一对满足 KKT 条件的原问题的可行点 (x_1^*, x_2^*) 和对偶问题的对偶可行点 (α^*, β^*) ，那么 (x_1^*, x_2^*) 必为原问题的最优解。那我们接下来只需从 KKT 条件方程组中，把 (x_1^*, x_2^*) 和 (α^*, β^*) 求解出来，即解方程组，

$$\begin{cases} x_2^* - b \leq 0 \\ 1 - x_1^* - x_2^* = 0 \\ \alpha^* \geq 0 \\ \alpha^* (x_2^* - b) = 0 \\ \frac{\partial l}{\partial x_1} \Big|_{x_1=x_1^*} = 0, \frac{\partial l}{\partial x_2} \Big|_{x_2=x_2^*} = 0 \end{cases}$$

由最后一条梯度为 $\mathbf{0}$ 的约束可得出，

$$\begin{cases} 2x_1^* - \beta^* = 0 \\ 2x_2^* + \alpha^* - \beta^* = 0 \end{cases} \Rightarrow \begin{cases} x_1^* = \frac{\beta^*}{2} \\ x_2^* = \frac{\beta^* - \alpha^*}{2} \end{cases}, \text{再带入 KKT 条件第 2 条，得到 } \beta^* = \frac{2 + \alpha^*}{2}, \text{因此有}$$

$$\begin{cases} x_1^* = \frac{2+\alpha^*}{4} \\ x_2^* = \frac{2-\alpha^*}{4} \end{cases} \text{。再考虑不等式约束 } x_2^* \leq b \text{，则有 } \frac{2-\alpha^*}{4} \leq b \text{，即 } \alpha^* \geq 2 - 4b \text{。下面对 } b \text{ 分类讨论}$$

(如图 13-6 所示):

1) 若 $b > 1/2$ ，只需取 $\alpha^* = 0$ 即可满足所有约束，这时的极值点出现在可行集相对内部，为

$$(x_1^* = 1/2, x_2^* = 1/2) \text{ (图 13-6 (a));}$$

2) 若 $b = 1/2$ ，只需取 $\alpha^* = 0$ 即可满足所有约束，这时的极值点出现在可行集相对边界，为

$$(x_1^* = 1/2, x_2^* = 1/2) \text{ (图 13-6 (b));}$$

3) 若 $b < 1/2$ ， α^* 必然要大于 0，此时根据 KKT 条件第 4 条，必有 $x_2^* = b$ ，则相应地 $x_1^* = 1-b$ ，

此时的最优解也出现在可行集相对边界上 (图 13-6 (c))。

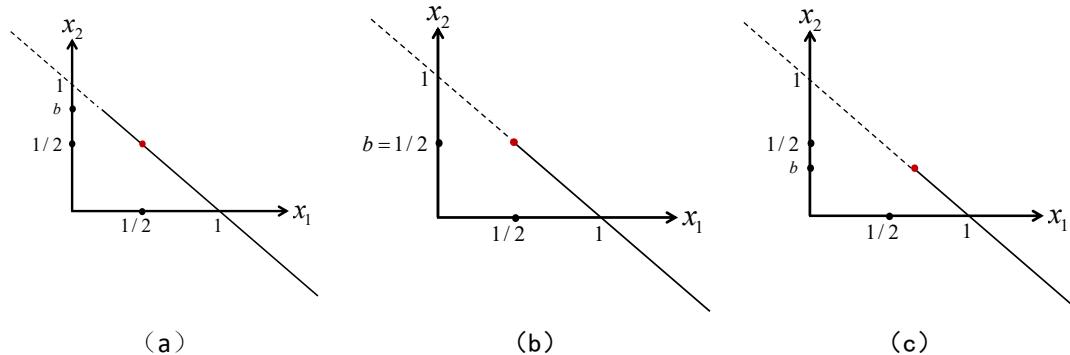


图 13-6: 例 13.5 最优解的 3 种情况。图中，实线表示可行集，红色点表示最优解。在情况 (a) 中， $b > 1/2$ ，最优解出现在可行集相对内部；在情况 (b) ($b = 1/2$) 和情况 (c) ($b < 1/2$) 中，最优解出现在可行集相对边界上。

13.2.6 利用对偶问题来求解原问题

在上一小节中，从式 13-20 得到的结论 1) 我们知道，如果原问题具有强对偶性且对偶问题的最优解为 (α^*, β^*) ，则原问题的最优解 \mathbf{x}^* 一定是拉格朗日函数 $l(\mathbf{x}; \alpha^*, \beta^*)$ 的最小值点。

利用这个结论，有时候我们可以通过对偶问题的最优解来计算得到原问题的最优解。

具体来说，假设原问题具有强对偶性且对偶问题的最优解为 (α^*, β^*) 。假设函数 $l(\mathbf{x}; \alpha^*, \beta^*)$ 的最小值点，

$$\mathbf{x}^* = \arg \min_{\mathbf{x}} f_0(\mathbf{x}) + \sum_{i=1}^m \alpha_i^* f_i(\mathbf{x}) + \sum_{i=1}^p \beta_i^* h_i(\mathbf{x})$$

是唯一的。那么，只要 \mathbf{x}^* 对于原问题来说是可行的，它必是原问题的最优解；如果 \mathbf{x}^* 对于原问题来说是不可行的，则原问题没有最优解。

利用这个性质，当对偶问题较容易解的时候，我们可以先求出对偶问题的最优解，再利用上述过程求出原问题的最优解。

13.3 总结

本章的内容从凸集的概念开始，到 KKT 条件结束，简要介绍了凸优化领域的基础概念和理论，这些材料可为读者进一步深入学习数学优化理论和算法打下基础。由于本章的概念和结论较多，为了方便初学者厘清它们之间的逻辑推理关系、知晓每个概念或结论是如何支撑其他概念或结论的，图 13-7 给出了本章基本概念和结论之间的支撑关系，图中的箭头表示支撑关系，比如“仿射函数 → 凸优化问题”，表示的含义是“凸优化问题”这个概念在定义的时候需要“仿射函数”这个概念来支撑。

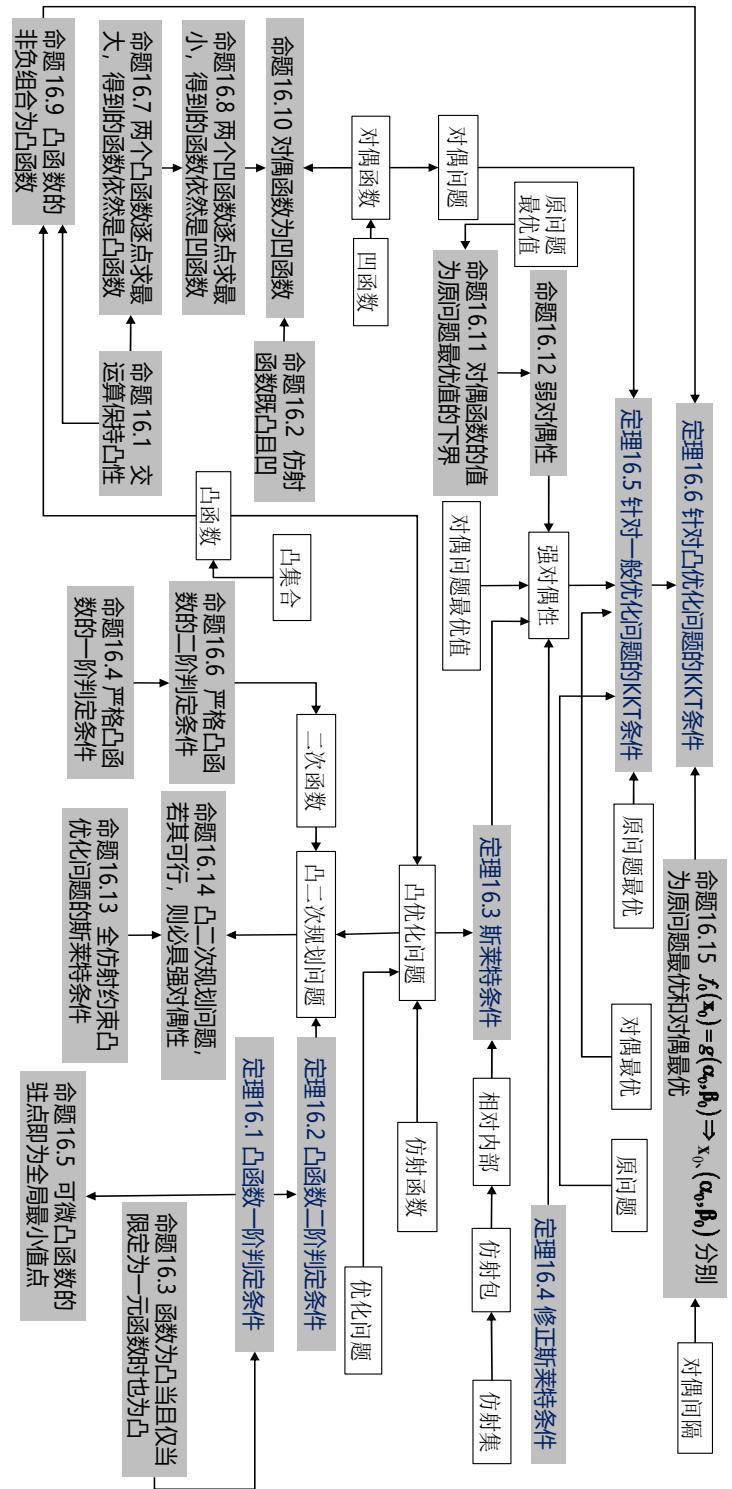


图 13-7：本章基本概念和结论之间的支撑关系。

13.4 习题

- (4) 请证明命题 13.8, 即, 若 $f_1(\mathbf{x}), f_2(\mathbf{x})$ 为两个凹函数, 则 $f(\mathbf{x})=\min\{f_1(\mathbf{x}), f_2(\mathbf{x})\}$ 也为凹函数。
- (5) 考虑由式 13-7 所定义的原优化问题, 以及由式 13-13 所定义的对偶问题。若 \mathbf{x}_0 是

原问题的一个可行点, (α_0, β_0) 是对偶问题的一个对偶可行点, 若相应的对偶间隔为 0,

即 $f_0(\mathbf{x}_0) = g(\alpha_0, \beta_0)$, 请证明: \mathbf{x}_0 必为原问题的最优解, (α_0, β_0) 必为对偶问题的最优解, 并且原问题是强对偶的。

(6) 考虑由式 13-9 所定义的凸优化问题, 其拉格朗日函数为,

$$l(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = f_0(\mathbf{x}) + \sum_{i=1}^m \alpha_i f_i(\mathbf{x}) + \sum_{i=1}^p \beta_i h_i(\mathbf{x})$$

若 $\boldsymbol{\alpha} \geq \mathbf{0}$, 请证明 $l(\mathbf{x}; \boldsymbol{\alpha}, \boldsymbol{\beta})$ 为关于 \mathbf{x} 的凸函数。

参考文献

- [1] S. Boyd and L. Vandenberghe, Convex Optimization, Cambridge University Press, 2004.
- [2] Morton Slater. Lagrange Multipliers Revisited, Cowles Commission Discussion Paper No. 403 (Report), 1950.
- [3] H. W. Kuhn, A. W. Tucker, "Nonlinear programming," Proc. 2nd Berkeley Symposium. Berkeley: University of California Press. pp. 481–492, 1951.
- [4] W. Karush. Minima of Functions of Several Variables with Inequalities as Side Constraints (M.Sc. thesis), Dept. of Mathematics, Univ. of Chicago, Chicago, Illinois, 1939.

第 14 章 支持向量机与基于支持向量机的目标检测

14.1 线性分类问题

考虑这样一个二类分类的机器学习问题：给定训练集 $\mathcal{D} = \{(\mathbf{x}_i, y_i) : \mathbf{x}_i \in \mathbb{R}^d, y_i \in \{+1, -1\}\}_{i=1}^n$ ，要从中学习出一个二类分类模型 h 出来。其中， (\mathbf{x}_i, y_i) 为第 i 个训练样本， \mathbf{x}_i 为一个 d 维特征向量， y_i 为样本 i 的标签。为了后续讨论方便，我们姑且假定训练集 \mathcal{D} 是线性可分的¹¹，即 \mathcal{D} 中的正负样本可以用一个超平面完美区分开来，那什么又是超平面呢？

定义 14.1 超平面^[1]。欧氏空间 \mathbb{R}^d 中的一个超平面是由满足如下条件的点 \mathbf{x} 组成的集合，

$$\mathbf{w}^T \mathbf{x} + b = 0 \quad (14-1)$$

其中， $\mathbf{w} \neq \mathbf{0} \in \mathbb{R}^d$ 为超平面的法向量；如果以 \mathbf{w} 为正向的话，从原点出发到超平面的带符号的距离为 $\frac{-b}{\|\mathbf{w}\|}$ 。

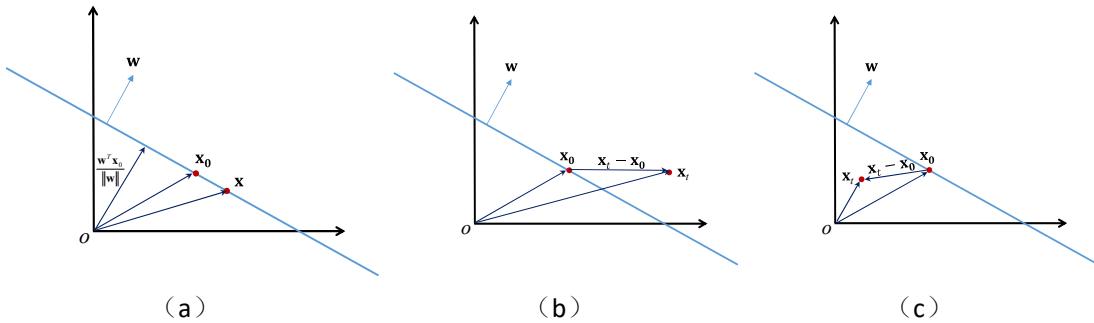


图 14-1：(a) 超平面方程的几何解释： $\mathbf{w}^T \mathbf{x} + b = 0$ 定义了一个超平面，其法向量为 \mathbf{w} ，若该平面过一已知点 \mathbf{x}_0 ，则 $b = -\mathbf{w}^T \mathbf{x}_0$ ；以 \mathbf{w} 为正向，从原点出发到超平面的有向距离为 $\frac{-b}{\|\mathbf{w}\|}$ ；(b) 位于超平面正侧 (\mathbf{w} 指向的一侧) 的点 \mathbf{x}_t 满足 $\mathbf{w}^T \mathbf{x}_t + b > 0$ ；(c) 位于超平面负侧的点 \mathbf{x}_t 满足 $\mathbf{w}^T \mathbf{x}_t + b < 0$ 。

容易知道，若 $d=1$ ，则超平面就是数轴上一个孤立的点；若 $d=2$ ，超平面表现为二维平

¹¹ 而在大多数实际情况中，分类问题既不是二类分类问题，训练集也不是线性可分的。请读者稍安勿躁，我们在后续章节中会逐步增加问题的复杂度以适配实际情况。

面中的一条直线；若 $d=3$ ，超平面则为三维欧氏空间中的一个平面。如图 14-1 (a) 所示，我们以二维空间为例，来解释一下超平面方程中参数的几何意义。设超平面的法向量为 \mathbf{w} ，且该超平面过一已知点 \mathbf{x}_0 ，那么该超平面上的任意点 \mathbf{x} 满足方程，

$$\mathbf{w}^T(\mathbf{x} - \mathbf{x}_0) = 0 \quad (14-2)$$

对照定义式 14-1，可知 $b = -\mathbf{w}^T \mathbf{x}_0$ 。根据图 14-1 (a)，由原点 O 出发到超平面的有向距离为

$$\mathbf{x}_0 \cdot \frac{\mathbf{w}}{\|\mathbf{w}\|} = \frac{\mathbf{w}^T \mathbf{x}_0}{\|\mathbf{w}\|} = \frac{-b}{\|\mathbf{w}\|}。这些关于超平面方程的几何解释拓广到高维欧氏空间中也是成立的。$$

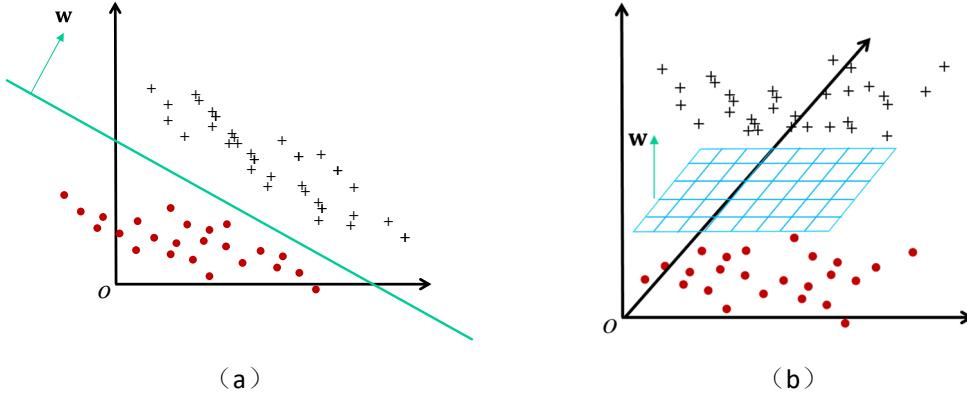


图 14-2：线性可分数据示例。(a) 二维情况下的线性可分数据，正负样本分别位于由直线所形成的分类面的两侧；(b) 三维情况下的线性可分数据，正负样本分别位于由平面所形成的分类面的两侧。

如果样本集 \mathcal{D} 是线性可分的，指的是在 d 维空间中存在超平面 $\mathbf{w}^T \mathbf{x} + b = 0$ ，使得 \mathcal{D} 中的所有正样本都在超平面的正侧（ \mathbf{w} 所指向的一侧），所有负样本都在超平面的负侧。图 14-2 (a) 和 (b) 分别给出了二维和三维情况下线性可分数据的示例；图中，“加号”表示正样本，“圆点”表示负样本。线性分类学习算法就是要基于训练集 \mathcal{D} ，在线性模型集合 \mathcal{H} 中，找到能把正负样本完全分开的超平面 $\mathbf{w}^T \mathbf{x} + b = 0$ 。

假如基于 \mathcal{D} 我们利用某种方法已经找到了满足条件的超平面 $\mathbf{w}^T \mathbf{x} + b = 0$ ，那么在测试阶段，对于一个新来的测试样本 \mathbf{x}_t ，如何利用该超平面来判断 \mathbf{x}_t 到底是正样本还是负样本呢？如图 14-1 (b) 所示，如果点 \mathbf{x}_t 位于超平面 $\mathbf{w}^T \mathbf{x} + b = 0$ 正侧，则矢量 $\mathbf{x}_t - \mathbf{x}_0$ 与 \mathbf{w} 夹角为锐角，因此 $\mathbf{w}^T(\mathbf{x}_t - \mathbf{x}_0) > 0$ ，即 $\mathbf{w}^T \mathbf{x}_t + b > 0$ ，此结论反之也成立，即如果 $\mathbf{w}^T \mathbf{x}_t + b > 0$ ，那么 \mathbf{x}_t 就是正样本。类似地如图 14-1 (c)，如果点 \mathbf{x}_t 位于超平面 $\mathbf{w}^T \mathbf{x} + b = 0$ 负侧，则矢量 $\mathbf{x}_t - \mathbf{x}_0$ 与 \mathbf{w} 夹角为钝角，相应地 $\mathbf{w}^T(\mathbf{x}_t - \mathbf{x}_0) < 0$ ，即 $\mathbf{w}^T \mathbf{x}_t + b < 0$ ，此结论反之也成立，即如果 $\mathbf{w}^T \mathbf{x}_t + b < 0$ ，那么 \mathbf{x}_t 就是负样本。总结下来，我们可以根据超平面 $\mathbf{w}^T \mathbf{x} + b = 0$ 来诱导出一个分类模型 $h_{\mathbf{w}, b}(\mathbf{x})$ ，

$$h_{\mathbf{w}, b}(\mathbf{x}) = \begin{cases} +1 & \text{if } \mathbf{w}^T \mathbf{x} + b \geq 0 \\ -1 & \text{if } \mathbf{w}^T \mathbf{x} + b < 0 \end{cases} \quad (14-3)$$

将来再来了任何一个特征向量 \mathbf{x} , 我们都可以按照式 14-3 所确定的分类模型来判断 \mathbf{x} 是正样本还是负样本。式 14-3 便是一个线性分类模型。我们还剩下一个非常棘手的问题没有解决: 基于训练集 \mathcal{D} , 如何能找到参数 \mathbf{w} 和 b , 使得超平面 $\mathbf{w}^T \mathbf{x} + b = 0$ 能把 \mathcal{D} 中的正负样本完全分开 (所有正样本都在超平面的正侧, 所有负样本都在超平面的负侧)? 下一节我们将学习解决这个问题的一个算法, 感知器算法。

14.2 感知器算法

感知器 (Perceptron) 可以说是最简单的线性分类器, 该算法由美国学者弗兰克·罗森布拉特 (Frank Rosenblatt, 图 14-3) 于 1958 年提出^[2]。在后续章节中可以看到, 感知器实际上可以看作是神经网络的基本组成单元。



(a)



(b)

图 14-3: (a) 弗兰克·罗森布拉特 (Frank Rosenblatt, 1928 年 7 月 11 日–1971 年 7 月 11 日) 是美国心理学家, 在人工智能领域享有盛誉。1971 年, 43 岁生日那天, 他在切萨皮克湾 (Chesapeake Bay) 驾驶一艘名为 Clearwater 的单桅帆船时溺水身亡。由于他最早提出了感知器学习算法, 而感知器模型现在被认为是神经网络的基础构件, 因此也有文献认为他是“深度学习之父”^[3]; (b) 为了纪念 Frank Rosenblatt, IEEE (电气电子工程师学会) 从 2004 年开始设立了 IEEE Frank Rosenblatt Award 奖, 用以表彰对生物和语言驱动的计算范式和系统做出杰出贡献的学者。

感知器学习算法要解决的问题是: 从线性可分的训练集 $\mathcal{D} = \{(\mathbf{x}_i, y_i) : \mathbf{x}_i \in \mathbb{R}^d, y_i \in \{+1, -1\}\}_{i=1}^n$ 中确定出 \mathbf{w} 和 b , 使得对于 \mathcal{D} 中的每一个样本 i ,

$h_{\mathbf{w}, b}(\mathbf{x}_i) = \text{sign}(\mathbf{w}^T \mathbf{x}_i + b) = y_i$ 都能成立。为了后续论述方便, 我们记 $\hat{\mathbf{x}}_i = \begin{pmatrix} \mathbf{x}_i \\ 1 \end{pmatrix}$ 、 $\hat{\mathbf{w}} = \begin{pmatrix} \mathbf{w} \\ b \end{pmatrix}$, 这样显

然有 $\mathbf{w}^T \mathbf{x}_i + b = \hat{\mathbf{w}}^T \hat{\mathbf{x}}_i$ 。相应地, 从 \mathcal{D} 中确定超平面参数 \mathbf{w} 和 b 的问题也就转换成了从 \mathcal{D} 中确

定 $\hat{\mathbf{w}}$ 的问题。算法 14-1 给出了感知器算法的伪码。

算法 14-1: 感知器算法

输入:

$$\text{训练集 } \mathcal{D} = \left\{ (\mathbf{x}_i, y_i) : \mathbf{x}_i \in \mathbb{R}^d, y_i \in \{+1, -1\} \right\}_{i=1}^n$$

输出:

超平面参数 $\hat{\mathbf{w}}$

随机初始化 $\hat{\mathbf{w}}$

$$\text{misclassified_examples} := \{(\mathbf{x}_i, y_i) \in \mathcal{D} : h_{\hat{\mathbf{w}}}(\mathbf{x}_i) \neq y_i\}$$

while misclassified_examples 非空

从 misclassified_examples 中随机选取一个样本 (\mathbf{x}_m, y_m)

//注意: y_m 是这个错分样本的真实类标

$$\hat{\mathbf{w}} := \hat{\mathbf{w}} + \hat{\mathbf{x}}_m y_m$$

$$\text{misclassified_examples} := \{(\mathbf{x}_i, y_i) \in \mathcal{D} : h_{\hat{\mathbf{w}}}(\mathbf{x}_i) \neq y_i\}$$

end

返回最终得到的 $\hat{\mathbf{w}}$

从算法 14-1 可以看出, 感知器学习算法首先会初始化一个超平面, 之后会进入迭代阶段。在每次迭代中, 算法首先用当前超平面去对训练集 \mathcal{D} 进行分类测试, 进而从被错误分类的样本集合中随机挑选一个出来, 并用该挑选出来的错分样本信息对超平面参数 $\hat{\mathbf{w}}$ 进行更新。迭代过程持续进行, 直至在当前所得的超平面下, 集合 \mathcal{D} 中没有被错分的样本为止。对于该算法的几个细节我们有必要进一步解读一下。

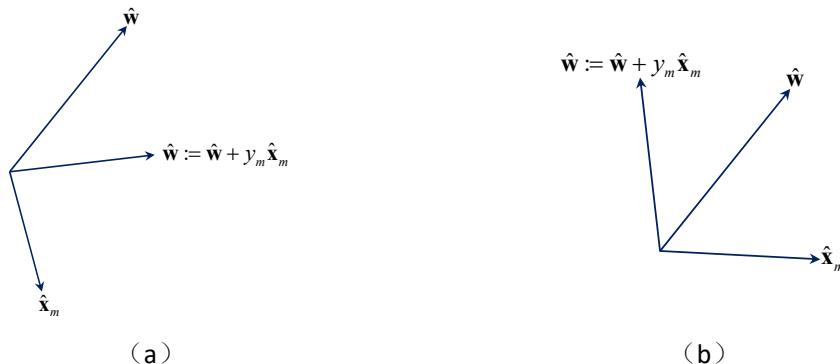


图 14-4: (a) 原始超平面 $\hat{\mathbf{w}}$ 将 $\hat{\mathbf{x}}_m$ 错分为了负样本, 更新之后的超平面 $\hat{\mathbf{w}} := \hat{\mathbf{w}} + \hat{\mathbf{x}}_m y_m$ 具有更高的可能性可将 $\hat{\mathbf{x}}_m$ 正确分类为正样本; (b) 原始超平面 $\hat{\mathbf{w}}$ 将 $\hat{\mathbf{x}}_m$ 错分为了正样本, 更新之后的超平面 $\hat{\mathbf{w}} := \hat{\mathbf{w}} + \hat{\mathbf{x}}_m y_m$ 具有更高的可能性可将 $\hat{\mathbf{x}}_m$ 正确分类为负样本。

在每次迭代过程中, 感知器算法根据随机选取的一个错分样本 (\mathbf{x}_m, y_m) 的信息来对超平面参数 $\hat{\mathbf{w}}$ 进行更新, 更新准则为 $\hat{\mathbf{w}} := \hat{\mathbf{w}} + \hat{\mathbf{x}}_m y_m$ 。那么, 该准则为什么是合理的? 不难理解, 如果一个更新准则是合理的, 那么更新之后的超平面应该有很大概率可以将之前被错分的样本 (\mathbf{x}_m, y_m) 分对。那么我们就来检查一下, 更新准则 $\hat{\mathbf{w}} := \hat{\mathbf{w}} + \hat{\mathbf{x}}_m y_m$ 是否具有该特性。如图 14-4

(a) 所示, 由于 $\hat{\mathbf{w}}^T \hat{\mathbf{x}}_m < 0$, 因此原始超平面 $\hat{\mathbf{w}}$ 将 \mathbf{x}_m 错分为了负样本, 相应地必有 $y_m=1$ 。在这种情况下, 更新之后的向量 $\hat{\mathbf{w}} + \hat{\mathbf{x}}_m y_m = \hat{\mathbf{w}} + \hat{\mathbf{x}}_m$ 与 $\hat{\mathbf{x}}_m$ 之间的夹角会减小, 因此超平面 $\hat{\mathbf{w}} + \hat{\mathbf{x}}_m y_m$ 与原超平面 $\hat{\mathbf{w}}$ 相比, 其有更高的可能性将 \mathbf{x}_m 正确分类为正样本。类似地, 如图 14-4 (b) 所示, 由于 $\hat{\mathbf{w}}^T \hat{\mathbf{x}}_m > 0$, 因此原始超平面 $\hat{\mathbf{w}}$ 将 \mathbf{x}_m 错分为了正样本, 相应地必有 $y_m=-1$ 。在这种情况下, 更新之后的向量 $\hat{\mathbf{w}} + \hat{\mathbf{x}}_m y_m = \hat{\mathbf{w}} - \hat{\mathbf{x}}_m$ 与 $\hat{\mathbf{x}}_m$ 之间的夹角会增大, 因此超平面 $\hat{\mathbf{w}} + \hat{\mathbf{x}}_m y_m$ 与原超平面 $\hat{\mathbf{w}}$ 相比, 其有更高的可能性将 \mathbf{x}_m 正确分类为负样本。因此, 可以看出, 超平面参数更新准则 $\hat{\mathbf{w}} := \hat{\mathbf{w}} + \hat{\mathbf{x}}_m y_m$ 是合理的。但需要注意的是, 这并不意味着经过一次超平面参数更新之后, 新的超平面就一定会把之前相应的错分样本 \mathbf{x}_m 分类正确了, 往往需要经过几次迭代之后才能达到这个目标。

另外, 有时我们根据某个错分样本 \mathbf{x}_1 对超平面进行了更新之后, 原本已经被正确分类的样本 \mathbf{x}_2 在新的超平面之下反而会被错误分类了。有效处理这个问题的一个简单办法就是只有当更新之后的超平面会降低错分样本数目的时候, 我们才接受此次更新。

从算法 14-1 中可以看出, 感知器算法的迭代停止条件是训练集中不再有被错分的样本了。那么问题来了, 我们是否能够保证: 按照算法 14-1 中的超平面参数更新准则, 经过有限次迭代以后, 训练集中的所有样本一定都会被正确分类? 幸运的是答案是肯定的。1962 年, 美国学者 Novikoff 证明了如下结论: 如果一个集合 \mathcal{D} 中的正负样本是线性可分的, 那么按照算法 14-1 所述的超平面参数更新准则, 经过有限次迭代更新之后, 最终得到的超平面一定能够将 \mathcal{D} 中的正负样本完全分开^[4]。Novikoff 定理的证明也可以参见中文教科书《统计学习方法》^[5]。但如果训练集 \mathcal{D} 本身并不是线性可分的话, 感知器学习算法不会收敛, 迭代结果会发生震荡。

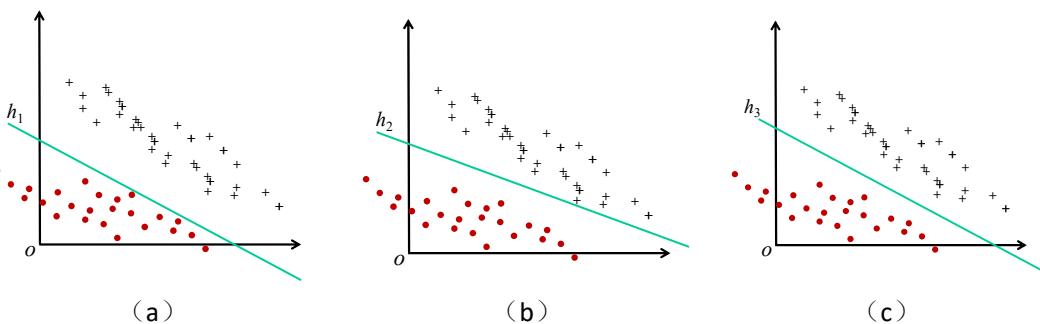


图 14-5: 运行感知器学习算法 3 次, 从同一组训练数据中会得到 3 个不同的分类超平面。通过直观观察不难理解, 超平面 h_3 比 h_1 和 h_2 更好, 因为对于 h_1 和 h_2 来说, 某些样本到它们的距离非常小 (远小于训练样本集到 h_3 的最小距离), 说明它们潜在的泛化能力不如 h_3 。

如果训练集 \mathcal{D} 是线性可分的, 那么感知器算法一定可以从中学习出一个超平面, 该超平

面可以把 \mathcal{D} 中的正负样本完全分开。但要注意的是，由于超平面是随机初始化的，并且在每次更新超平面参数时，所使用的错分样本也是随机选取的，因此若多次运行感知器算法，可能会得到多个不同的分类超平面。如图 14-5 所示，在某个具体的线性可分训练集上，运行感知器算法 3 次，我们得到了 3 个分类超平面 h_1 、 h_2 和 h_3 。初看起来这似乎没有什么潜在的问题，因为这些超平面虽然不同，但它们都可以将训练集中的正负样本完全区分开。那是否便可以认为这些超平面是“同样好的”呢？很遗憾的是答案是否定的！不要忘记，机器学习的任务是要从训练集中学习出一个模型，我们希望这个模型能在将来未见的测试数据上具有较好的泛化能力。虽然图 14-5 中的超平面 h_1 、 h_2 和 h_3 都能将训练样本正确分类（即它们的训练分类误差都为 0），但需要注意，对于 h_1 、 h_2 来说，某些样本离它们非常近，一旦对这些样本增加一些小的扰动，相应的超平面一定会产生分类错误，也就是说图 14-5 中的分类超平面 h_1 和 h_2 的泛化能力较差。相比之下，考虑图 14-5 (c) 中的超平面 h_3 ，相对来说，所有样本到 h_3 的距离都很大，这就意味着 h_3 抗扰动能力很强，即它会具有较好的泛化能力。那么，我们如何才能从线性可分的数据集中学习出类似 h_3 这样的较优的分类超平面呢？我们在下一节中将解决这个问题。

14.3 线性可分支持向量机

在上一节最后我们提到，对于线性可分的训练集，感知器算法并不能返回唯一的最优分类超平面。在这一节，我们将学习线性可分支持向量机（Support Vector Machine, SVM）学习算法，它可以从线性可分的训练集合 \mathcal{D} 中学习出唯一的最优的（类似于图 14-5 (c) 中的 h_3 ）分类超平面。我们将首先定义出什么是最优分类超平面；之后，再逐步把从集合 \mathcal{D} 中学习出最优分类超平面的问题建模为一个典型的凸优化问题；最后，我们将详细介绍如何求解这样一个凸优化问题。



图 14-6：弗拉基米尔·瓦普尼克 (Vladimir N. Vapnik, 1936 年 12 月 6 日-)，统计学家，因提出了支持向量机、VC 理论等而著名。他出生于前苏联。1958 年，他在撒马尔罕（现属乌兹别克斯坦）的乌兹别克国立大学完成了硕士学业。1964 年，他于莫斯科控制科学学院获得博士学位。毕业后，他一直在该校工作直到 1990 年，在此期间，他成为了该校计算机科学与研

究系的系主任。1990 年底，弗拉基米尔·瓦普尼克移居美国，加入了位于新泽西州霍姆德的 At&T 贝尔实验室的自适应系统研究部门。1995 年，他被伦敦大学聘为计算机与统计科学专业的教授。现在，他工作于新泽西州普林斯顿的 NEC 实验室。他同时是哥伦比亚大学的特聘教授。2006 年，他成为美国国家工程院院士。

线性可分支持向量机算法于二十世纪六十年代由前苏联学者弗拉基米尔·瓦普尼克 (Vladimir N. Vapnik) 等人提出，其相关工作被总结在了其俄文版专著之中^[6]。线性可分支持向量机也称为线性硬间隔支持向量机 (linear hard-margin SVM)。“硬间隔”的意思就是要假定训练数据集本身是线性可分的，这样就一定会存在分隔超平面可以将训练集中的全部样本完全正确分类。

14.3.1 线性可分支持向量机的问题建模

我们在 14.1 节中已经介绍了分隔超平面的概念。为了方便建模 SVM 学习问题，我们需要重新梳理一下这个概念。对于数据集 \mathcal{D} ，如果超平面 $\mathbf{w}^T \mathbf{x} + b = 0$ 是其分隔超平面，那么必有，

$$\forall (\mathbf{x}_i, y_i) \in \mathcal{D}, y_i(\mathbf{w}^T \mathbf{x}_i + b) > 0 \quad (14-4)$$

但分类信号 $y_i(\mathbf{w}^T \mathbf{x}_i + b)$ 本身的大小是没有意义的，因为通过缩放 (\mathbf{w}, b) ，我们可以得到任意大小的 $y_i(\mathbf{w}^T \mathbf{x}_i + b)$ 。这是因为 (\mathbf{w}, b) 和 $(\mathbf{w}/\rho, b/\rho), \forall \rho > 0$ 表达的是完全相同的超平面，但通过改变 ρ ，我们却可以随意改变样本 i 的分类信号的大小。我们考虑按如下方式取 ρ ，

$$\rho = \min_{i=1,\dots,n} y_i(\mathbf{w}^T \mathbf{x}_i + b) \quad (14-5)$$

并对超平面参数 (\mathbf{w}, b) 进行缩放，得到其新的表达 $(\mathbf{w}/\rho, b/\rho)$ 。这样便有，

$$\min_{i=1,\dots,n} y_i\left(\frac{\mathbf{w}^T}{\rho} \mathbf{x}_i + \frac{b}{\rho}\right) = \frac{1}{\rho} \min_{i=1,\dots,n} y_i(\mathbf{w}^T \mathbf{x}_i + b) = \frac{\rho}{\rho} = 1 \quad (14-6)$$

上述分析表明，对于线性可分的数据集 \mathcal{D} 来说，对于它的任意一个分隔超平面，都可以通过选择合适的参数 (\mathbf{w}, b) 来表达该分隔超平面，使得 $\forall (\mathbf{x}_i, y_i) \in \mathcal{D}, y_i(\mathbf{w}^T \mathbf{x}_i + b) \geq 1$ ，并且 \mathcal{D} 中至少有一个样本会使得上述不等式中的等号严格成立。受此启发，我们给出如下在 SVM 学习领域所使用的分隔超平面定义：

定义 14.2 分隔超平面 (Separating hyperplane)。 给定训练集 $\mathcal{D} = \{(\mathbf{x}_i, y_i) : \mathbf{x}_i \in \mathbb{R}^d, y_i \in \{+1, -1\}\}_{i=1}^n$ ，超平面 h 是 \mathcal{D} 的分隔超平面当且仅当它可以被满足如下条件的参数 (\mathbf{w}, b) 所表达，

$$\min_{i=1,\dots,n} y_i(\mathbf{w}^T \mathbf{x}_i + b) = 1 \quad (14-7)$$

观察图 14-5 中的三个分类超平面 h_1 、 h_2 和 h_3 ，我们为什么会感觉 h_3 要比 h_1 和 h_2 更好

呢？那是因为样本到 h_3 的最小距离要比样本到 h_1 (h_2) 的最小距离要大，这就使得 h_3 对样本的扰动具有更高的鲁棒性。基于这个直观观察，我们就有了比较两个分类超平面优劣的基本准则：

设 h_1 和 h_2 是两个可将线性可分训练集 \mathcal{D} 正确分类的超平面， \mathcal{D} 中样本到 h_1 的最小距离为 m_1 ， \mathcal{D} 中样本到 h_2 的最小距离为 m_2 。若 $m_1 > m_2$ ，则超平面 h_1 优于超平面 h_2 。

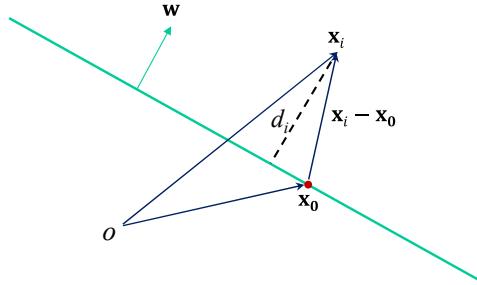


图 14-7：点 \mathbf{x}_i 到超平面 $\mathbf{w}^T \mathbf{x} + b = 0$ 的欧氏距离计算示意图。

那如何计算样本 \mathbf{x}_i 到超平面 $\mathbf{w}^T \mathbf{x} + b = 0$ 的“距离”呢？我们先来看看如何计算 \mathbf{x}_i 到超平面 $\mathbf{w}^T \mathbf{x} + b = 0$ 的欧氏距离。如图 14-7 所示，点 \mathbf{x}_i 到超平面 $\mathbf{w}^T \mathbf{x} + b = 0$ 的欧氏距离记为 d_i ， d_i 可通过如下方式计算得出，

$$d_i = \left| \frac{\mathbf{w}}{\|\mathbf{w}\|} \cdot (\mathbf{x}_i - \mathbf{x}_0) \right| = \frac{|\mathbf{w} \cdot (\mathbf{x}_i - \mathbf{x}_0)|}{\|\mathbf{w}\|} = \frac{|\mathbf{w}^T \mathbf{x}_i - \mathbf{w}^T \mathbf{x}_0|}{\|\mathbf{w}\|} = \frac{|\mathbf{w}^T \mathbf{x}_i + b|}{\|\mathbf{w}\|} \quad (14-8)$$

需要注意到，式 14-8 计算的是点 \mathbf{x}_i 到超平面 $\mathbf{w}^T \mathbf{x} + b = 0$ 的欧氏距离，这个距离始终是非负数，它并不能反映出 \mathbf{x}_i 是否能被该超平面所正确分类。由于我们寻找最优分类超平面的准则是“最大化样本到超平面的最小距离”，我们希望“距离”的计算方式能够体现分类的正确性与否：当 \mathbf{x}_i 被正确分类时，相应的 \mathbf{x}_i 到超平面的“距离”要为非负数；当 \mathbf{x}_i 被错误分类时，相应的 \mathbf{x}_i 到超平面的距离要小于 0；这样，寻找最优分隔超平面的准则便会优先选择能对训练集进行完全正确分类的超平面。通过观察不难发现，我们只需要对式 14-8 稍加修改，便可得到满足期望的计算样本 (\mathbf{x}_i, y_i) 到超平面 $\mathbf{w}^T \mathbf{x} + b = 0$ “距离”的计算方式，

$$\gamma_i = \frac{y_i (\mathbf{w}^T \mathbf{x}_i + b)}{\|\mathbf{w}\|} \quad (14-9)$$

当样本 \mathbf{x}_i 被正确分类时，式 14-9 计算的就是样本点到超平面的欧氏距离；当 \mathbf{x}_i 被错误分类时，式 14-9 计算的就是样本点到超平面欧氏距离的相反数，为负数。在文献中，按照式 14-9 的方式计算的样本 (\mathbf{x}_i, y_i) 到超平面 $\mathbf{w}^T \mathbf{x} + b = 0$ 的“距离”有个名称，称为样本 (\mathbf{x}_i, y_i) （在超平面 $\mathbf{w}^T \mathbf{x} + b = 0$ 下）的几何间隔（geometric margin）。基于样本点的几何间隔定义，我们可以进一步给出超平面的几何间隔的定义：

定义 14.3 超平面的几何间隔。给定训练集 $\mathcal{D} = \{(\mathbf{x}_i, y_i) : \mathbf{x}_i \in \mathbb{R}^d, y_i \in \{+1, -1\}_{i=1}^n\}$ ，超平面

$\mathbf{w}^T \mathbf{x} + b = 0$ 的几何间隔 γ 为 \mathcal{D} 中所有样本在该超平面下几何间隔的最小值，即，

$$\gamma = \min_{i=1,2,\dots,n} \gamma_i = \min_{i=1,2,\dots,n} \frac{y_i(\mathbf{w}^T \mathbf{x}_i + b)}{\|\mathbf{w}\|} = \frac{1}{\|\mathbf{w}\|} \min_{i=1,2,\dots,n} y_i(\mathbf{w}^T \mathbf{x}_i + b) \quad (14-10)$$

对于线性可分训练集 \mathcal{D} ，我们要寻找的它的最优分隔超平面便是在 \mathcal{D} 上具有最大几何间隔的那个超平面。该问题可被建模为如下优化问题，

$$\begin{aligned} \mathbf{w}^*, b^* &= \arg \max_{\mathbf{w}, b} \gamma = \arg \max_{\mathbf{w}, b} \left(\frac{1}{\|\mathbf{w}\|} \min_{i=1,2,\dots,n} y_i(\mathbf{w}^T \mathbf{x}_i + b) \right) = \arg \max_{\mathbf{w}, b} \frac{1}{\|\mathbf{w}\|} \\ \text{subject to } &\min_{i=1,2,\dots,n} y_i(\mathbf{w}^T \mathbf{x}_i + b) = 1 \end{aligned} \quad (14-11)$$

其中的约束条件表示该超平面首先一定要是数据集 \mathcal{D} 的一个分隔超平面。问题式 14-11 可继续变形为如下同解问题，

$$\begin{aligned} \mathbf{w}^*, b^* &= \arg \min_{\mathbf{w}, b} \frac{1}{2} \mathbf{w}^T \mathbf{w} \\ \text{subject to } &-y_i(\mathbf{w}^T \mathbf{x}_i + b) + 1 \leq 0, i = 1, \dots, n \end{aligned} \quad (14-12)$$

式 14-12 就是最终得到的线性可分 SVM 的数学模型。对该问题进行求解，得到的参数 (\mathbf{w}^*, b^*) 便定义了能将线性可分训练集 \mathcal{D} 完全正确分类且具有最大几何间隔的最优分隔超平面。当有了最优分隔超平面 (\mathbf{w}^*, b^*) 之后，若样本 $(\mathbf{x}_k, y_k) \in \mathcal{D}$ 满足 $y_k(\mathbf{w}^T \mathbf{x}_k + b) = 1$ ，则称该样本为分隔超平面 (\mathbf{w}^*, b^*) 的支持向量 (Support vector)，这也就是为什么这类寻找最优分隔超平面的方法称为支持向量机。

若记 $\mathbf{u} = \begin{pmatrix} b \\ \mathbf{w} \end{pmatrix}$ ，则有 $\mathbf{w}^T \mathbf{w} = \mathbf{u}^T \begin{bmatrix} 0 & \mathbf{0}_{1 \times d} \\ \mathbf{0}_{d \times 1} & I_{d \times d} \end{bmatrix} \mathbf{u}$ ， $-y_i(\mathbf{w}^T \mathbf{x}_i + b) + 1 = (-y_i - y_i \mathbf{x}_i^T) \mathbf{u} + 1$ 。相应地，

式 14-12 可变为如下同解问题，

$$\begin{aligned} \mathbf{u}^* &= \arg \min_{\mathbf{u}} \frac{1}{2} \mathbf{u}^T \begin{bmatrix} 0 & \mathbf{0}_{1 \times d} \\ \mathbf{0}_{d \times 1} & I_{d \times d} \end{bmatrix} \mathbf{u} \\ \text{subject to } &(-y_i - y_i \mathbf{x}_i^T) \mathbf{u} + 1 \leq 0, i = 1, \dots, n \end{aligned} \quad (14-13)$$

可验证其中的矩阵 $\begin{bmatrix} 0 & \mathbf{0}_{1 \times d} \\ \mathbf{0}_{d \times 1} & I_{d \times d} \end{bmatrix}$ 为半正定矩阵。对照式 13-10 凸二次规划问题的定义，容易验证，上述问题便是标准的凸二次规划问题。因此，原则上来说，我们可以用通用的求解标准凸二次规划问题的方法来求解 SVM 问题。但通用算法并没有考虑 SVM 问题的特点，不能很好地处理在大规模数据集 (d 很大， n 很大) 上的 SVM 训练问题。因此，在机器学习领域，学者们提出了专门针对 SVM 学习问题的求解算法，这将在下节中进行介绍。

14.3.2 线性可分支持向量机问题的求解

在 14.3.1 节中，我们对线性可分 SVM 学习问题进行了数学建模，得到了由式 14-12 所表达的一个优化问题。在这一节中，我们将详细讲述如何对该问题进行求解。

我们要求解的优化问题为，

$$\begin{aligned} \mathbf{w}^*, b^* &= \arg \min_{\mathbf{w}, b} \frac{1}{2} \mathbf{w}^T \mathbf{w} \\ \text{subject to } &-y_i(\mathbf{w}^T \mathbf{x}_i + b) + 1 \leq 0, i = 1, \dots, n \end{aligned} \quad (14-14)$$

其拉格朗日函数（见定义 13.11）为，

$$\begin{aligned} l(\mathbf{w}, b, \alpha) &= \frac{1}{2} \mathbf{w}^T \mathbf{w} + \sum_{i=1}^n \alpha_i (-y_i(\mathbf{w}^T \mathbf{x}_i + b) + 1) \\ &= \frac{1}{2} \mathbf{w}^T \mathbf{w} - \sum_{i=1}^n \alpha_i y_i (\mathbf{w}^T \mathbf{x}_i + b) + \sum_{i=1}^n \alpha_i \end{aligned} \quad (14-15)$$

其中， $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_n)^T$ 。如果把式 14-14 所描述的优化问题看作是原问题的话，其对偶问题（见定义 13.13）为，

$$\begin{aligned} \alpha^* &= \arg \max_{\alpha} \left\{ \min_{\mathbf{w}, b} l(\mathbf{w}, b, \alpha) \right\} \\ \text{subject to } &\alpha \geq 0 \end{aligned} \quad (14-16)$$

我们在 14.3.1 节中已经提到，问题式 14-14 为一个凸二次规划问题，又由于已经假定训练集 \mathcal{D} 是线性可分的，因此该问题必然存在可行点，根据命题 13.14 可知，该问题满足斯莱特条件，即具有强对偶性。另外注意到，该问题的目标函数和所有约束函数都可微。根据定理 13.6 可知，只要我们能找到一对原问题的可行点和对偶问题的可行点， (\mathbf{w}^*, b^*) 和 α^* ，使得它们满足 KKT 条件，那么 (\mathbf{w}^*, b^*) 和 α^* 必然分别是原问题式 14-14 和其对偶问题式 14-16 的最优解。因此，我们求解问题 14-14 的思路就是，先找到其对偶问题式 14-16 的最优解 α^* 的表达式，再根据 KKT 条件所形成的等式约束关系找到 (\mathbf{w}^*, b^*) ，这样最终得到的 (\mathbf{w}^*, b^*) 和 α^* 当然就会满足 KKT 条件，相应地 (\mathbf{w}^*, b^*) 就是我们要找的原问题的最优解。

我们先来求解对偶问题式 14-16。这个问题包含了两个部分，里层是关于原问题优化变量 (\mathbf{w}, b) 的最小化问题，外层是关于对偶变量 α 的最大化问题，我们需要“从里到外”顺次解决。

(1) 求解 $\min_{\mathbf{w}, b} l(\mathbf{w}, b, \alpha)$

这个最小化问题的目标函数为 $l_1(\mathbf{w}, b) = \frac{1}{2} \mathbf{w}^T \mathbf{w} - \sum_{i=1}^n \alpha_i y_i (\mathbf{w}^T \mathbf{x}_i + b) + \sum_{i=1}^n \alpha_i$ ， l_1 关于 (\mathbf{w}, b) 为一个凸函数与一组仿射函数的求和，容易验证它为可微凸函数。根据命题 13.5，该

函数的驻点便是它的全局最小值点。因此，我们只需要找到 l_1 关于 (\mathbf{w}, b) 的驻点，便可得出

它的最小值。找 l_1 的驻点便是要求解方程组，

$$\begin{cases} \frac{\partial l_1}{\partial \mathbf{w}} = \mathbf{0} \\ \frac{\partial l_1}{\partial b} = 0 \end{cases} \Rightarrow \begin{cases} \mathbf{w} - \sum_{i=1}^n \alpha_i y_i \mathbf{x}_i = \mathbf{0} \\ -\sum_{i=1}^n \alpha_i y_i = 0 \end{cases} \Rightarrow \begin{cases} \mathbf{w} = \sum_{i=1}^n \alpha_i y_i \mathbf{x}_i \\ \sum_{i=1}^n \alpha_i y_i = 0 \end{cases}$$

因此有，

$$\mathbf{w} = \sum_{i=1}^n \alpha_i y_i \mathbf{x}_i \quad (14-17)$$

$$\sum_{i=1}^n \alpha_i y_i = 0 \quad (14-18)$$

因此， l_1 的最小值，即 $\min_{\mathbf{w}, b} l(\mathbf{w}, b, \alpha)$ ，为

$$\begin{aligned} & \frac{1}{2} \left(\sum_{i=1}^n \alpha_i y_i \mathbf{x}_i \right) \cdot \left(\sum_{j=1}^n \alpha_j y_j \mathbf{x}_j \right) - \sum_{i=1}^n \alpha_i y_i \left(\left(\sum_{j=1}^n \alpha_j y_j \mathbf{x}_j \right) \cdot \mathbf{x}_i + b \right) + \sum_{i=1}^n \alpha_i \\ &= \frac{1}{2} \left(\sum_{i=1}^n \alpha_i y_i \mathbf{x}_i \right) \cdot \left(\sum_{j=1}^n \alpha_j y_j \mathbf{x}_j \right) - \sum_{i=1}^n \alpha_i y_i \mathbf{x}_i \cdot \left(\sum_{j=1}^n \alpha_j y_j \mathbf{x}_j \right) - b \sum_{i=1}^n \alpha_i y_i + \sum_{i=1}^n \alpha_i \\ &= -\frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j \mathbf{x}_i \cdot \mathbf{x}_j + \sum_{i=1}^n \alpha_i \end{aligned}$$

即，

$$\min_{\mathbf{w}, b} l(\mathbf{w}, b, \alpha) = -\frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j \mathbf{x}_i \cdot \mathbf{x}_j + \sum_{i=1}^n \alpha_i$$

(2) 求解

$$\begin{aligned} \alpha^* &= \arg \max_{\alpha} \left\{ -\frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j \mathbf{x}_i \cdot \mathbf{x}_j + \sum_{i=1}^n \alpha_i \right\}, \\ &\text{subject to } \alpha \geq \mathbf{0} \\ &\sum_{i=1}^n \alpha_i y_i = 0 \end{aligned}$$

首先把该问题转换成标准优化问题的形式，

$$\begin{aligned} \alpha^* &= \arg \min_{\alpha} \left\{ \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j \mathbf{x}_i \cdot \mathbf{x}_j - \sum_{i=1}^n \alpha_i \right\}, \\ &\text{subject to } -\alpha \leq \mathbf{0} \end{aligned} \quad (14-19)$$

$$\sum_{i=1}^n \alpha_i y_i = 0$$

上述问题可进一步转换成如下形式，

$$\alpha^* = \arg \min_{\alpha} \left\{ \frac{1}{2} \alpha^T Q \alpha - \mathbf{1}_{n \times 1}^T \alpha \right\},$$

subject to $\begin{bmatrix} \mathbf{y}^T \\ -\mathbf{y}^T \\ -I_{n \times n} \end{bmatrix} \alpha \leq \mathbf{0}_{(n+2) \times 1}$

(14-20)

其中, 若令 $X = \begin{bmatrix} -y_1 \mathbf{x}_1^T & - \\ -y_2 \mathbf{x}_2^T & - \\ \vdots & \\ -y_n \mathbf{x}_n^T & - \end{bmatrix}$, 则 $Q = XX^T = \begin{bmatrix} y_1 y_1 \mathbf{x}_1^T \mathbf{x}_1 & y_1 y_2 \mathbf{x}_1^T \mathbf{x}_2 & \cdots & y_1 y_n \mathbf{x}_1^T \mathbf{x}_n \\ y_2 y_1 \mathbf{x}_2^T \mathbf{x}_1 & y_2 y_2 \mathbf{x}_2^T \mathbf{x}_2 & \cdots & y_2 y_n \mathbf{x}_2^T \mathbf{x}_n \\ \vdots & & & \\ y_n y_1 \mathbf{x}_n^T \mathbf{x}_1 & y_n y_2 \mathbf{x}_n^T \mathbf{x}_2 & \cdots & y_n y_n \mathbf{x}_n^T \mathbf{x}_n \end{bmatrix}$, 根据附录命

题 G.4 可知, Q 为半正定矩阵。对照凸二次规划问题的定义 (定义 13.10) 可知, 问题式 14-20 是一个标准的凸二次规划问题。读者会发现, 我们本来要解决的线性 SVM 学习问题式 14-14 就是一个凸二次规划问题, 经过了一系列推导之后, 我们把该问题转换成了式 14-20, 但问题式 14-20 依旧是一个凸二次规划问题, 那我们岂不是陷入了一个死循环? 直接用通用的求解凸二次规划问题的算法包求解问题 14-14 不就行了吗? 实际上, 式 14-20 这个凸二次规划问题要比式 14-14 那个凸二次规划问题容易求解, 因为它的优化变量里面只有拉格朗日乘子 α 。基于这个特点, 研究人员已经设计出了针对式 14-20 这个特殊凸二次规划问题的快速高效求解算法, 比如序列最小最优化算法 (sequential minimal optimization, SMO) [7] 及其变种。SMO 算法的本质思想是把一个大规模的复杂优化问题转化成能有解析解的小规模简单的优化问题。SMO 算法具有较多琐碎的实现细节且与本书的主线内容关联度较弱, 因此我们就不再具体介绍该算法了。感兴趣的读者可以参见【5】。当然, 如果问题的规模不是很大的话, 我们确实可以直接用解凸二次规划问题的标准程序¹²来解问题式 14-20。

当有了对偶问题式 14-16 的最优解 α^* 之后, 如前所述, 我们便可以根据 KKT 条件中的等式约束关系找到 (\mathbf{w}^*, b^*) , 使得 (\mathbf{w}^*, b^*) 与 α^* 满足 KKT 条件, 这样 (\mathbf{w}^*, b^*) 便是我们最终要找的原问题的最优解。这个过程可以表达成如下命题的形式:

命题 14.1 设 $\alpha^* = (\alpha_1^*, \alpha_2^*, \dots, \alpha_n^*)$ 是对偶问题式 14-16 的最优解, 则存在下标 j , 使得

$\alpha_j^* > 0$, 并可按照如下方式求得原问题式 14-14 的最优解 (\mathbf{w}^*, b^*) :

$$\mathbf{w}^* = \sum_{i=1}^n \alpha_i^* y_i \mathbf{x}_i$$
(14-21)

$$b^* = y_j - \sum_{i=1}^n \alpha_i^* y_i (\mathbf{x}_i \cdot \mathbf{x}_j)$$
(14-22)

证明:

根据定理 13.6 可知, 在已知对偶问题的最优解为 α^* 的情况下, 如果能找到原问题的可行点 (\mathbf{w}^*, b^*) , 使得 (\mathbf{w}^*, b^*) 和 α^* 满足 KKT 条件, 那么 (\mathbf{w}^*, b^*) 必为原问题的最优解。根据 KKT

¹² 比如 Matlab 中提供的库函数 quadprog。

条件，列出方程组：

$$\begin{aligned}\nabla_{\mathbf{w}=\mathbf{w}^*} l(\mathbf{w}, b, \boldsymbol{\alpha}^*) &= \mathbf{0} \\ \nabla_{b=b^*} l(\mathbf{w}, b, \boldsymbol{\alpha}^*) &= 0\end{aligned}\tag{14-23}$$

$$\begin{aligned}\alpha_i^* (-y_i (\mathbf{w}^* \cdot \mathbf{x}_i + b^*) + 1) &= 0, i = 1, \dots, n \\ -y_i (\mathbf{w}^* \cdot \mathbf{x}_i + b^*) + 1 &\leq 0, i = 1, \dots, n \\ \alpha_i^* &\geq 0, i = 1, \dots, n\end{aligned}\tag{14-24}$$

由式 14-23 可知， $\mathbf{w}^* = \sum_{i=1}^n \alpha_i^* y_i \mathbf{x}_i$ 。

至少存在一个下标 j , $\alpha_j^* > 0$ 。可以用反证法：如果不存在这样的下标 j , 则 $\boldsymbol{\alpha}^* = \mathbf{0}$, 则

有 $\mathbf{w}^* = \sum_{i=1}^n \alpha_i^* y_i \mathbf{x}_i = \mathbf{0}$ 。但 $\mathbf{w}^* = \mathbf{0}$ 显然不是原问题式 14-14 的最优解，产生矛盾，因此必有

某个下标 j , 使得 $\alpha_j^* > 0$ 。对于这样的下标 j , 由式 14-24 可知,

$$-y_j (\mathbf{w}^* \cdot \mathbf{x}_j + b^*) + 1 = 0$$

将 $\mathbf{w}^* = \sum_{i=1}^n \alpha_i^* y_i \mathbf{x}_i$ 带入上式并注意到 $y_j^2 = 1$, 我们便有 $b^* = y_j - \sum_{i=1}^n \alpha_i^* y_i (\mathbf{x}_i \cdot \mathbf{x}_j)$ 。

有了 (\mathbf{w}^*, b^*) 之后，我们便得到了线性可分数据集 \mathcal{D} 的最优分隔超平面；相应地，也可得到最终的基于该超平面的分类决策函数，

$$h_{\mathbf{w}^*, b^*}(\mathbf{x}) = \text{sign}\left(\sum_{i=1}^n \alpha_i^* y_i (\mathbf{x} \cdot \mathbf{x}_i) + b^*\right)$$

最后再提醒读者注意一点，如式 14-21 和 14-22 在计算 \mathbf{w}^* 和 b^* 时，其中的求和是对所有样本进行求和。但实际上， $\boldsymbol{\alpha}^*$ 中的大部分分量都为零，显然 $\boldsymbol{\alpha}^*$ 中只有不为零的那些元素才会在计算 \mathbf{w}^* 和 b^* 时真正发挥作用。更进一步，若 $\alpha_i^* \neq 0$ ，则根据 KKT 条件式 14-24，必有

$y_i (\mathbf{w}^* \cdot \mathbf{x}_i + b^*) = 1$ ，即样本特征向量 \mathbf{x}_i 是最优分类超平面的支持向量，那显而易见， \mathbf{w}^* 和 b^* 的计算值实际上只与支持向量集合有关。

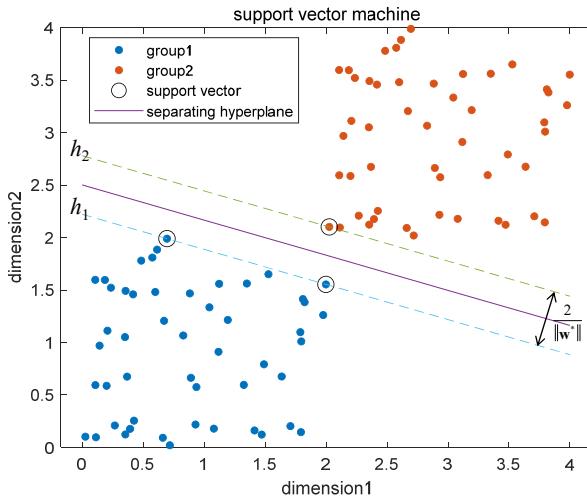


图 14-8：从线性可分数据中学习出最优分隔超平面，实线代表最优分隔超平面，圈出来的样本点为最优分隔超平面的支持向量。

最后，以一个可视化的例子来结束本节。如图 14-8 所示，有一组线性可分的二维数据，我们用本节介绍的线性可分 SVM 模型从该数据集中学习出了最优分隔超平面（图中的实线）。图中圈出的样本点即为最优分隔超平面的支持向量，它们满足条件 $y_i(\mathbf{w}^* \cdot \mathbf{x}_i + b^*) = 1$ 。对于正样本支持向量 ($y_i=1$) 来说，它所在的超平面为 $h_1: \mathbf{w}^* \cdot \mathbf{x} + b^* = 1$ ；对于负样本支持向量 ($y_i=-1$) 来说，它所在的超平面为 $h_2: \mathbf{w}^* \cdot \mathbf{x} + b^* = -1$ 。显然， h_1 与 h_2 平行，并且没有样本点在 h_1 与 h_2 之间。 h_1 和 h_2 界定出了一条带状区域，最优分隔超平面位于它们中央。考虑某个支持向量 (\mathbf{x}_i, y_i) ，根据式 14-8，它到最优分隔超平面 $\mathbf{w}^* \cdot \mathbf{x} + b^* = 0$ 的欧氏距离为 $\frac{|\mathbf{w}^* \cdot \mathbf{x}_i + b^*|}{\|\mathbf{w}^*\|} = \frac{1}{\|\mathbf{w}^*\|}$ ，因此可知超平面 h_1 与超平面 h_2 之间的距离为 $\frac{2}{\|\mathbf{w}^*\|}$ 。实际上，最优分隔超平面仅由支持向量集合确定；对于非支持向量来说，它们可任意移动，只要不进入由 h_1 与 h_2 所界定的“带状区域”，都不会引起最优分隔超平面的改变。

14.4 软间隔与线性支持向量机

14.4.1 问题建模

如果数据集是线性可分的，那么用 14.3 节中讲述的线性可分 SVM 就可将此类分类问题完美解决。但遗憾的是，现实世界中的（原始采集的）数据集很少是线性可分的。当数据集不是线性可分的时候，我们便不能用线性可分 SVM 来解决这类数据的分类问题了。这是因为式 14-12 中的不等式约束不会全都满足，即问题式 14-12 的可行集为空（问题无解）。那要对线性可分 SVM 进行怎样的扩展才能使它可以解决线性不可分问题呢？本节将介绍解决这个问题的一种思路，软间隔支持向量机。

软间隔支持向量机解决从这样的数据集中学习出最优分隔超平面的问题：训练集 $\mathcal{T} = \{(\mathbf{x}_i, y_i) : |\mathbf{x}_i \in \mathbb{R}^d, y_i \in \{+1, -1\}\}_{i=1}^n$ 中存在一些外点 (outlier)，这导致 \mathcal{T} 不是线性可分的；但若将这些外点剔除之后，剩下的大部分的样本点所组成的集合是线性可分的。

数据集 \mathcal{T} 线性不可分，这意味着 \mathcal{T} 中的某些样本点 (\mathbf{x}_i, y_i) 不能严格满足 $y_i(\mathbf{w}^T \mathbf{x}_i + b) \geq 1$ 这个约束条件。我们可对这个约束条件放松一下，引入松弛变量 $\xi_i \geq 0$ ，将约束条件修改为， $y_i(\mathbf{w}^T \mathbf{x}_i + b) \geq 1 - \xi_i$ 。不难发现，对一给定超平面 (\mathbf{w}, b) ，对于任意 $(\mathbf{x}_i, y_i) \in \mathcal{T}$ ，一定存在某个 $\xi_i \geq 0$ ，使得条件 $y_i(\mathbf{w}^T \mathbf{x}_i + b) \geq 1 - \xi_i$ 满足。我们当然希望 ξ_i 不能太大，这就需要在目标函数中对大的 ξ_i 进行“惩罚”，因此可将问题式 14-12 中的目标函

数修改为 $\frac{1}{2} \mathbf{w}^T \mathbf{w} + C \sum_{i=1}^n \xi_i$ ，其中 $C > 0$ 称为惩罚参数，一般需要根据问题性质由用户设定。

这样，针对线性不可分数据集的线性 SVM 学习问题可被建模为如下形式，

$$\begin{aligned} \mathbf{w}^*, b^* &= \arg \min_{\mathbf{w}, b, \xi} \frac{1}{2} \mathbf{w}^T \mathbf{w} + C \sum_{i=1}^n \xi_i \\ \text{subject to } &y_i(\mathbf{w}^T \mathbf{x}_i + b) \geq 1 - \xi_i, i = 1, \dots, n \\ &\xi_i \geq 0, i = 1, \dots, n \end{aligned} \tag{14-25}$$

我们在 14.3 节中提到，能够解决线性可分数据集分类问题的线性支持向量机称为硬间隔支持向量机，相应地，本节介绍的能够处理线性不可分数据集分类问题的线性支持向量机称为软间隔 (soft margin) 支持向量机。同硬间隔 SVM 相比，软间隔 SVM 在寻找最优分隔超平面时并不是试图找一个“不会分错”的超平面，而是在找一个“犯错最少”的超平面。不难看出，软间隔线性支持向量机包含硬间隔线性支持向量机，它既可以处理线性不可分的情况，当然也能处理线性可分的情况。后面我们就把软间隔线性支持向量机简称为线性支持向量机。

14.4.2 问题求解

接下来我们考虑如何求解问题式 14-25。实际上，经过变形之后，该问题也是一个凸二次规划问题。记 $\mathbf{u} = \begin{pmatrix} b \\ \mathbf{w} \\ \xi \end{pmatrix}$ ，其中 $\xi \triangleq (\xi_1, \dots, \xi_n)^T$ ，记，

$$P = \begin{bmatrix} 0 & \mathbf{0}_{1 \times d} & \mathbf{0}_{1 \times n} \\ \mathbf{0}_{d \times 1} & I_{d \times d} & \mathbf{0}_{d \times n} \\ \mathbf{0}_{n \times 1} & \mathbf{0}_{n \times d} & \begin{bmatrix} \frac{2C}{\xi_1} \\ \ddots \\ \frac{2C}{\xi_n} \end{bmatrix}_{n \times n} \end{bmatrix}_{(1+d+n) \times (1+d+n)}$$

注意矩阵 P 的右下角的 $n \times n$ 的对角子矩阵，只有当 $\xi_i > 0$ 时，对应位置处的元素才是 $\frac{2C}{\xi_i}$ ，

如果 $\xi_i = 0$ ，直接把对应位置置为 0 即可。容易知道，矩阵 P 为半正定矩阵。同时我们有，

$$\begin{aligned} \mathbf{u}^T P \mathbf{u} &= (b \ \mathbf{w}^T \ \boldsymbol{\xi}^T) \begin{bmatrix} 0 & \mathbf{0}_{1 \times d} & \mathbf{0}_{1 \times n} \\ \mathbf{0}_{d \times 1} & I_{d \times d} & \mathbf{0}_{d \times n} \\ \mathbf{0}_{n \times 1} & \mathbf{0}_{n \times d} & \begin{bmatrix} \frac{2C}{\xi_1} \\ \ddots \\ \frac{2C}{\xi_n} \end{bmatrix}_{n \times n} \end{bmatrix}_{(1+d+n) \times (1+d+n)} \begin{pmatrix} b \\ \mathbf{w} \\ \boldsymbol{\xi} \end{pmatrix} \\ &= \mathbf{w}^T \mathbf{w} + 2C \sum_{i=1}^n \xi_i \end{aligned} \quad (14-26)$$

问题式 14-25 中的不等式约束 $y_i(\mathbf{w}^T \mathbf{x}_i + b) \geq 1 - \xi_i$ 可变形为，

$$\left(-y_i \ -y_i \mathbf{x}_i^T \ (0, 0, \dots, -1_{(i)}, 0, \dots, 0)_{1 \times n} \right) \begin{pmatrix} b \\ \mathbf{w} \\ \boldsymbol{\xi} \end{pmatrix} + 1 \leq 0 \quad (14-27)$$

其中， $(0, 0, \dots, -1_{(i)}, 0, \dots, 0)_{1 \times n}$ 表示这是一个 $1 \times n$ 的行向量，-1 出现在位置 i 处。问题式 14-25 中

的不等式约束 $\xi_i \geq 0$ 可变形为，

$$\left(0 \ \mathbf{0}_{1 \times d} \ (0, 0, \dots, -1_{(i)}, 0, \dots, 0)_{1 \times n} \right) \begin{pmatrix} b \\ \mathbf{w} \\ \boldsymbol{\xi} \end{pmatrix} \leq 0 \quad (14-28)$$

结合式 14-26、式 14-27 和式 14-28，问题式 14-25 可变形为如下问题，

$$\begin{aligned} \mathbf{u}^* &= \arg \min_{\mathbf{u}} \frac{1}{2} \mathbf{u}^T P \mathbf{u} \\ \text{subject to } & \left(-y_i - y_i \mathbf{x}_i^T \left(0, 0, \dots, -1_{(i)}, 0, \dots, 0 \right)_{1 \times n} \right) \mathbf{u} + 1 \leq 0, i = 1, \dots, n \\ & \left(0 \ \mathbf{0}_{1 \times d} \left(0, 0, \dots, -1_{(i)}, 0, \dots, 0 \right)_{1 \times n} \right) \mathbf{u} \leq 0, i = 1, \dots, n \end{aligned} \quad (14-29)$$

对照凸二次规划问题的定义（定义 13.10）可知，式 14-29 所描述的问题正是一个凸二次规划问题，即问题式 14-25 是一个凸二次规划问题。与 14.3.2 节中线性可分 SVM 问题求解的分析过程类似，问题式 14-25 也具有强对偶性。因此，我们也是同样的求解思路：找出对偶问题的最优解，然后根据 KKT 条件的等式约束，找出原问题的最优解。

与优化问题式 14-25 所对应的拉格朗日函数为，

$$l((\mathbf{w}, b, \xi), \alpha, \mu) = \frac{1}{2} \mathbf{w}^T \mathbf{w} + C \sum_{i=1}^n \xi_i + \sum_{i=1}^n \alpha_i (-y_i (\mathbf{w}^T \mathbf{x}_i + b) - \xi_i + 1) + \sum_{i=1}^n \mu_i (-\xi_i) \quad (14-30)$$

其中， $\alpha = (\alpha_1, \dots, \alpha_n)^T$, $\alpha_i \geq 0$, $\mu = (\mu_1, \dots, \mu_n)^T$, $\mu_i \geq 0$ 。如果把问题式 14-25 看作原问题的话，

其对偶问题为，

$$\begin{aligned} \alpha^*, \mu^* &= \arg \max_{\alpha, \mu} \left\{ \min_{(\mathbf{w}, b, \xi)} l((\mathbf{w}, b, \xi), \alpha, \mu) \right\} \\ \text{subject to } & \alpha \geq \mathbf{0} \\ & \mu \geq \mathbf{0} \end{aligned} \quad (14-31)$$

对偶问题式 14-31 的求解包含了两个部分，里层是关于原问题优化变量 (\mathbf{w}, b, ξ) 的最小化问题，外层是关于对偶变量 (α, μ) 的最大化问题，我们需要“从里到外”顺次解决。

(1) 求解 $\min_{(\mathbf{w}, b, \xi)} l((\mathbf{w}, b, \xi), \alpha, \mu)$

这个最小化问题的目标函数为，

$$l_1(\mathbf{w}, b, \xi) = \frac{1}{2} \mathbf{w}^T \mathbf{w} + C \sum_{i=1}^n \xi_i + \sum_{i=1}^n \alpha_i (-y_i (\mathbf{w}^T \mathbf{x}_i + b) - \xi_i + 1) + \sum_{i=1}^n \mu_i (-\xi_i)$$

l_1 关于 (\mathbf{w}, b, ξ) 为一个凸函数与一组仿射函数的组合，容易验证它为可微凸函数。根据命题 13.5，该函数的驻点便是它的全局最小值点。因此，我们只需要找到 l_1 关于 (\mathbf{w}, b, ξ) 的驻点，便可得出 l_1 的最小值。找 l_1 的驻点便是要求解方程组，

$$\begin{cases} \frac{\partial l_1}{\partial \mathbf{w}} = \mathbf{0} \\ \frac{\partial l_1}{\partial b} = 0 \\ \frac{\partial l_1}{\partial \xi_i} = 0 \end{cases} \Rightarrow \begin{cases} \mathbf{w} - \sum_{i=1}^n \alpha_i y_i \mathbf{x}_i = \mathbf{0} \\ -\sum_{i=1}^n \alpha_i y_i = 0 \\ C - \alpha_i - \mu_i = 0 \end{cases} \Rightarrow \begin{cases} \mathbf{w} = \sum_{i=1}^n \alpha_i y_i \mathbf{x}_i \\ \sum_{i=1}^n \alpha_i y_i = 0 \\ C - \alpha_i - \mu_i = 0 \end{cases} \quad (14-32)$$

$$(14-33)$$

$$(14-34)$$

因此， l_1 的最小值，即 $\min_{(\mathbf{w}, b, \xi)} l((\mathbf{w}, b, \xi), \alpha, \mu)$ ，为

$$\begin{aligned}
& \frac{1}{2} \left(\sum_{i=1}^n \alpha_i y_i \mathbf{x}_i \right) \cdot \left(\sum_{j=1}^n \alpha_j y_j \mathbf{x}_j \right) + C \sum_{i=1}^n \xi_i - \left(\sum_{i=1}^n \alpha_i y_i \mathbf{x}_i \right) \cdot \left(\sum_{j=1}^n \alpha_j y_j \mathbf{x}_j \right) - b \sum_{j=1}^n \alpha_j y_j - \sum_{i=1}^n \alpha_i \xi_i + \sum_{i=1}^n \alpha_i - \sum_{i=1}^n \mu_i \xi_i \\
& = -\frac{1}{2} \left(\sum_{i=1}^n \alpha_i y_i \mathbf{x}_i \right) \cdot \left(\sum_{j=1}^n \alpha_j y_j \mathbf{x}_j \right) + \sum_{i=1}^n \alpha_i + \sum_{i=1}^n (C - \alpha_i - \mu_i) \xi_i \\
& = -\frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j \mathbf{x}_i \cdot \mathbf{x}_j + \sum_{i=1}^n \alpha_i
\end{aligned}$$

即，

$$\min_{(\mathbf{w}, b, \boldsymbol{\xi})} l((\mathbf{w}, b, \boldsymbol{\xi}), \boldsymbol{\alpha}, \boldsymbol{\mu}) = -\frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j \mathbf{x}_i \cdot \mathbf{x}_j + \sum_{i=1}^n \alpha_i$$

(2) 求解

$$\begin{aligned}
\boldsymbol{\alpha}^*, \boldsymbol{\mu}^* &= \arg \max_{\boldsymbol{\alpha}, \boldsymbol{\mu}} \left\{ -\frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j \mathbf{x}_i \cdot \mathbf{x}_j + \sum_{i=1}^n \alpha_i \right\} \\
\text{subject to } \alpha_i &\geq 0, i = 1, \dots, n \\
\mu_i &\geq 0, i = 1, \dots, n \\
\sum_{i=1}^n \alpha_i y_i &= 0 \\
C - \alpha_i - \mu_i &= 0, i = 1, \dots, n
\end{aligned}$$

根据 $C - \alpha_i - \mu_i = 0$ 这个等式约束，上述问题可进一步精简为，

$$\begin{aligned}
\boldsymbol{\alpha}^* &= \arg \min_{\boldsymbol{\alpha}} \left\{ \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j \mathbf{x}_i \cdot \mathbf{x}_j - \sum_{i=1}^n \alpha_i \right\} \\
\text{subject to } C &\geq \alpha_i \geq 0, i = 1, \dots, n \\
\sum_{i=1}^n \alpha_i y_i &= 0
\end{aligned} \tag{14-35}$$

读者会发现，上述问题与问题式 14-19 几乎是一模一样的，唯一的区别就是对 $\boldsymbol{\alpha}$ 的约束从象限约束变成了“盒子”约束。同样的，该问题可以用 SMO 算法求解；当问题规模不大时，也可以直接用解决凸二次规划问题的通用算法求解。

当有了对偶问题式 14-31 的最优解 $(\boldsymbol{\alpha}^*, \boldsymbol{\mu}^*)$ 之后，如前所述，我们便可以根据 KKT 条件中的等式约束关系找到 $(\mathbf{w}^*, b^*, \boldsymbol{\xi}^*)$ ，使得 $(\mathbf{w}^*, b^*, \boldsymbol{\xi}^*)$ 与 $(\boldsymbol{\alpha}^*, \boldsymbol{\mu}^*)$ 满足 KKT 条件，这样 (\mathbf{w}^*, b^*) 便是我们最终要找的最优分类超平面。这个过程可以表达成如下形式的命题：

命题 14.2 设 $\boldsymbol{\alpha}^* = (\alpha_1^*, \alpha_2^*, \dots, \alpha_n^*)$ 是对偶问题式 14-31 的最优解，若存在下标 j ，使得

$0 < \alpha_j^* < C$ ，则可按照如下方式求得原问题式 14-25 的最优解 (\mathbf{w}^*, b^*) ：

$$\mathbf{w}^* = \sum_{i=1}^n \alpha_i^* y_i \mathbf{x}_i \tag{14-36}$$

$$b^* = y_j - \sum_{i=1}^n \alpha_i^* y_i (\mathbf{x}_i \cdot \mathbf{x}_j) \tag{14-37}$$

证明：

根据定理 13.6 可知，在已知对偶问题的最优解为 (α^*, μ^*) 的情况下，如果能找到原问题的可行点 (w^*, b^*, ξ^*) ，使得 (w^*, b^*, ξ^*) 和 (α^*, μ^*) 满足 KKT 条件，那么 (w^*, b^*, ξ^*) 必为原问题的最优解。根据 KKT 条件，列出方程组：

$$\nabla_{w=w^*} l((w, b, \xi), \alpha^*, \mu^*) = w^* - \sum_{i=1}^n \alpha_i^* y_i \mathbf{x}_i = \mathbf{0} \quad (14-38)$$

$$\nabla_{b=b^*} l((w, b, \xi), \alpha^*, \mu^*) = -\sum_{i=1}^n \alpha_i^* y_i = 0 \quad (14-39)$$

$$\nabla_{\xi_i=\xi_i^*} l((w, b, \xi), \alpha^*, \mu^*) = C - \alpha_i^* - \mu_i^* = 0 \quad (14-39)$$

$$\alpha_i^* (-y_i (w^* \cdot \mathbf{x}_i + b^*) + 1 - \xi_i^*) = 0, i = 1, \dots, n \quad (14-40)$$

$$\mu_i^* \xi_i^* = 0, i = 1, \dots, n \quad (14-41)$$

$$-y_i (w^* \cdot \mathbf{x}_i + b^*) + 1 - \xi_i \leq 0, i = 1, \dots, n$$

$$-\xi_i^* \leq 0, i = 1, \dots, n$$

$$\alpha_i^* \geq 0, i = 1, \dots, n$$

$$\mu_i^* \geq 0, i = 1, \dots, n$$

由式 14-38 可知， $w^* = \sum_{i=1}^n \alpha_i^* y_i \mathbf{x}_i$ 。若存在下标 j 使得 $0 < \alpha_j^* < C$ ，由式 14-39 可知 $\mu_j^* > 0$ ，

又由式 14-41 可知 $\xi_j^* = 0$ ；此时由式 14-40 可知， $-y_j (w^* \cdot \mathbf{x}_j + b^*) + 1 - \xi_j^* = 0$ ，则有，

$$b^* = y_j - \sum_{i=1}^n \alpha_i^* y_i (\mathbf{x}_i \cdot \mathbf{x}_j)$$

最终，最优分类超平面 $w^* \cdot \mathbf{x} + b^* = 0$ 表示为，

$$\sum_{i=1}^n \alpha_i^* y_i (\mathbf{x} \cdot \mathbf{x}_i) + b^* = 0$$

相应地，基于该超平面的分类决策函数为，

$$h_{w^*, b^*}(\mathbf{x}) = \text{sign}\left(\sum_{i=1}^n \alpha_i^* y_i (\mathbf{x} \cdot \mathbf{x}_i) + b^*\right).$$

我们给出线性支持向量机学习算法伪码：

算法 14-2：线性支持向量机学习算法

输入：

训练集 $\mathcal{T} = \{(\mathbf{x}_i, y_i) : \mathbf{x}_i \in \mathbb{R}^d, y_i \in \{+1, -1\}\}_{i=1}^n$

输出：

分类决策函数

(1) 选取惩罚参数 $C > 0$ ，构造并求解凸二次规划问题：

$$\alpha^* = \arg \min_{\alpha} \left\{ \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j) - \sum_{i=1}^n \alpha_i \right\}$$

subject to $C \geq \alpha_i \geq 0, i = 1, \dots, n$

$$\sum_{i=1}^n \alpha_i y_i = 0$$

求得的最优解为 $\alpha^* = (\alpha_1^*, \alpha_2^*, \dots, \alpha_n^*)^T$ 。

(2) 计算 $\mathbf{w}^* = \sum_{i=1}^n \alpha_i^* y_i \mathbf{x}_i$

(3) 从 α^* 中选择一个分量 α_j^* 符合条件 $0 < \alpha_j^* < C$, 计算

$$b^* = y_j - \sum_{i=1}^n \alpha_i^* y_i (\mathbf{x}_i \cdot \mathbf{x}_j)$$

(4) 构造决策函数: $h(\mathbf{x}) = \text{sign} \left(\sum_{i=1}^n \alpha_i^* y_i (\mathbf{x} \cdot \mathbf{x}_i) + b^* \right)$

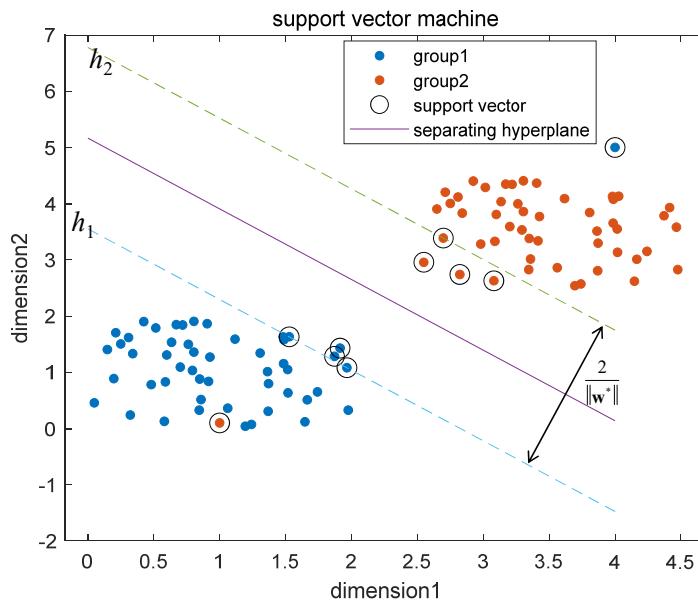


图 14-9：从线性不可分数据中利用软间隔线性支持向量机算法学习出最优分隔超平面，实线代表最优分隔超平面，圈出来的样本点为最优分隔超平面的支持向量。

我们以一个可视化的例子来结束本节。如图 14-9 所示，有一组线性不可分的二维数据，如果把少量样本剔除后，剩下的大部分数据点实际上是线性可分的，这种情况就比较适合用软间隔线性支持向量机来处理。用本节介绍的（软间隔）线性 SVM 模型从该数据集中学习出的最优分隔超平面如图 14-9 中的实线所示。图中圈出的样本点即为最优分隔超平面的支持向量，它们满足条件 $y_i (\mathbf{w}^* \cdot \mathbf{x}_i + b^*) = 1 - \xi_i$ 。

14.5 非线性支持向量机与核函数

对解线性分类问题，线性分类支持向量机是一种非常有效的方法。但是，有时分类问题是非线性的，这时可以使用非线性支持向量机。本节讲述非线性支持向量机，其主要特点是利用核技巧（kernel trick），为此要先介绍核技巧。

14.5.1 核函数与核技巧

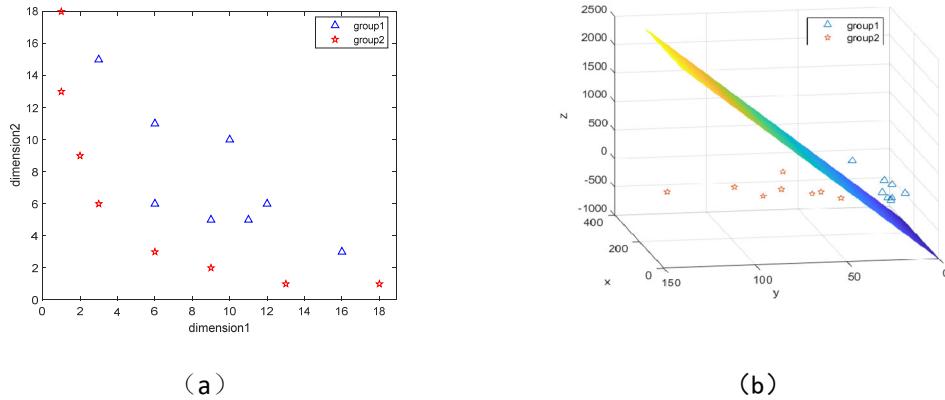


图 14-10：(a) 一组二维数据点，在二维空间中它们不是线性可分的；(b) 将 (a) 中的二维数据点，经过多项式映射变换至三维空间，在三维空间中这些数据点是线性可分的。

我们来看一个具体的例子。假设你有一组如图 14-10 (a) 所示的二维数据。你想用之前咱们介绍过的线性支持向量机来对它们进行分类，不难想象，不会得到很好的结果，我们找不到一条直线可以合理的将这两类数据分开。但需要强调的是，在二维空间中不能将这些数据分开，并不意味着在更高维的空间中不能将它们分开。可以尝试如下方案。对原始二维数据进行一个变换，把它们映射到三维空间，比如可用如下的多项式映射， $\phi: \mathbb{R}^2 \rightarrow \mathbb{R}^3$ ，

$$\phi(x_1, x_2) = (x_1^2, \sqrt{2}x_1x_2, x_2^2)$$

利用该映射，在图 14-10 (a) 中的数据会被映射至如图 14-10 (b) 所示的三维空间中。我们惊喜地发现，在三维空间中，这些数据点是线性可分的！

通过这个例子，我们可以得出利用支持向量机来解决非线性分类问题的一个大致思路：

- 1) 首先，把原始训练数据通过某种映射函数 ϕ ，从低维特征空间映射至高维特征空间 \mathcal{H} ；
- 2) 在 \mathcal{H} 中用映射后的训练数据训练出线性支持向量机；
- 3) 在测试阶段，对于一个待分类样本 t （在低维空间中表达），先用映射函数 ϕ 把它映射至 \mathcal{H} 为 $\phi(t)$ ，之后，再用在 2) 中训练好的线性支持向量机对 $\phi(t)$ 进行分类。

敏锐的读者会注意到，在上述方案中有一个关键问题：对于某个给定的数据集，如何选择合适的映射函数来完成从低维特征空间到高维特征空间的变换？不幸的是，这个问题没有

固定的正确答案。对于某个给定的具体问题，使用者往往需要根据经验进行一定的“试错”，才能确定出合适的映射函数。

在线性支持向量机模型的求解过程中，最终的核心问题会归结为要解一个对偶问题式 14-35。为方便阅读，我们把该问题写在了下方，

$$\begin{aligned}\alpha^* &= \arg \min_{\alpha} \left\{ \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j \mathbf{x}_i \cdot \mathbf{x}_j - \sum_{i=1}^n \alpha_i \right\} \\ \text{subject to } C &\geq \alpha_i \geq 0, i = 1, \dots, n \\ \sum_{i=1}^n \alpha_i y_i &= 0\end{aligned}$$

在这个优化问题中，与输入数据特征向量有关的项只有 $\mathbf{x}_i \cdot \mathbf{x}_j$ 。如果我们将训练数据从原始

输入特征空间经过映射 ϕ 映射到了高维特征空间 \mathcal{H} 中之后，数据特征就会相应地变换为 $\phi(\mathbf{x}_i)$ ($\phi(\mathbf{x}_j)$)。因此，在映射后的高维特征空间 \mathcal{H} 中进行线性支持向量机的学习的核心问题就会相应地变为，

$$\begin{aligned}\alpha^* &= \arg \min_{\alpha} \left\{ \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j) - \sum_{i=1}^n \alpha_i \right\} \\ \text{subject to } C &\geq \alpha_i \geq 0, i = 1, \dots, n \\ \sum_{i=1}^n \alpha_i y_i &= 0\end{aligned}\tag{14-42}$$

相应的分类决策函数会变为，

$$h_{w^*, b^*}(\mathbf{x}) = \text{sign} \left(\sum_{i=1}^n \alpha_i^* y_i \phi(\mathbf{x}) \cdot \phi(\mathbf{x}_i) + b^* \right)\tag{14-43}$$

上述“将线性不可分数据经过一个映射变换至高维特征空间，再在高维特征空间中训练线性支持向量机”的思路清晰直观，易于理解。然而在处理实际问题时，这种方式存在效率不高的缺点，这主要是因为我们需要把每一个训练数据都要映射至高维空间。观察式 14-42 和式 14-43 发现，不论是在支持向量机的训练阶段还是在样本分类预测阶段，我们希望计算的实际上不是映射之后的特征向量，而是映射之后的特征向量之间的内积。那么是否存在一种函数 $K(\cdot): \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ ，当它以 \mathbf{x}_i 和 \mathbf{x}_j 为输入时，其输出值 $K(\mathbf{x}_i, \mathbf{x}_j)$ 恰好是 \mathbf{x}_i 和 \mathbf{x}_j 映射至高维空间中之后的表示 $\phi(\mathbf{x}_i)$ 和 $\phi(\mathbf{x}_j)$ 的内积，即 $K(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j)$ ？如果这样的函数 K 存在的话，我们就可以通过计算 $K(\mathbf{x}_i, \mathbf{x}_j)$ 来代替 $\phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j)$ ，从而省去了计算特征 \mathbf{x}_i (\mathbf{x}_j) 高维映射的操作。满足我们要求的函数 K 便称为核函数：

定义 14.4 核函数。设 \mathcal{X} 是原始输入特征空间， \mathcal{H} 为高维特征空间，如果存在一个从 \mathcal{X} 到 \mathcal{H} 的映射， $\phi(\mathbf{x}): \mathcal{X} \rightarrow \mathcal{H}$ ，使得对所有 $\mathbf{x}, \mathbf{z} \in \mathcal{X}$ ，函数 $K(\mathbf{x}, \mathbf{z})$ 满足 $K(\mathbf{x}, \mathbf{z}) = \phi(\mathbf{x}) \cdot \phi(\mathbf{z})$ ，则称 $K(\mathbf{x}, \mathbf{z})$ 为核函数， $\phi(\mathbf{x})$ 为映射函数，其中 $\phi(\mathbf{x}) \cdot \phi(\mathbf{z})$ 表示向量 $\phi(\mathbf{x})$ 和 $\phi(\mathbf{z})$ 的内积。

核技巧的想法就是，在学习和预测中只定义核函数 K ，而不显式地定义映射函数 ϕ 。需要注意的是，对于给定的核函数 K ，特征空间 \mathcal{H} 和映射函数 ϕ 的取法并不唯一。下面通过一个例子来说明一下核函数和映射函数之间的关系。

例 14.1：假设原始输入特征空间为 \mathbb{R}^2 ，核函数是 $K(\mathbf{x}, \mathbf{z}) = (\mathbf{x} \cdot \mathbf{z})^2$ ，请找出相关的高维特征空间 \mathcal{H} 和映射函数 $\phi(\mathbf{x}) : \mathbb{R}^2 \rightarrow \mathcal{H}$ 。

解：

取高维特征空间 $\mathcal{H} = \mathbb{R}^3$ ，由于 $(\mathbf{x} \cdot \mathbf{z})^2 = x_1^2 z_1^2 + 2x_1 x_2 z_1 z_2 + x_2^2 z_2^2$ ，可取映射函数为，

$$\phi(\mathbf{x}) = (x_1^2, \sqrt{2}x_1 x_2, x_2^2)^T$$

容易验证此时有 $K(\mathbf{x}, \mathbf{z}) = (\mathbf{x} \cdot \mathbf{z})^2 = \phi(\mathbf{x}) \cdot \phi(\mathbf{z})$ 。

仍取高维特征空间 $\mathcal{H} = \mathbb{R}^3$ 。也可取映射函数为， $\phi(\mathbf{x}) = \frac{1}{\sqrt{2}}(x_1^2 - x_2^2, 2x_1 x_2, x_1^2 + x_2^2)$ 。此时同样会有 $K(\mathbf{x}, \mathbf{z}) = (\mathbf{x} \cdot \mathbf{z})^2 = \phi(\mathbf{x}) \cdot \phi(\mathbf{z})$ 。

还可取高维特征空间 $\mathcal{H} = \mathbb{R}^4$ ，以及映射函数 $\phi(\mathbf{x}) = (x_1^2, x_1 x_2, x_1 x_2, x_2^2)$ 。

下面介绍几个常用的核函数。

(1) 多项式核函数 (polynomial kernel function)

$$K(\mathbf{x}, \mathbf{z}) = (\mathbf{x} \cdot \mathbf{z} + 1)^p$$

相应的分类决策函数的形式为，

$$h_{w^*, b^*}(\mathbf{x}) = \text{sign}\left(\sum_{i=1}^n \alpha_i^* y_i (\mathbf{x}_i \cdot \mathbf{x} + 1)^p + b^*\right)$$

(2) 高斯核函数 (Gaussian kernel function)

$$K(\mathbf{x}, \mathbf{z}) = \exp\left(-\frac{\|\mathbf{x} - \mathbf{z}\|_2^2}{2\sigma^2}\right)$$

该核函数也称为径向基函数 (radial basis function, RBF)。相应的分类决策函数的形式为，

$$h_{w^*, b^*}(\mathbf{x}) = \text{sign}\left(\sum_{i=1}^n \alpha_i^* y_i \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}_i\|_2^2}{2\sigma^2}\right) + b^*\right)$$

借助于核函数，在高维特征空间 \mathcal{H} 中进行线性支持向量机学习的核心问题式 14-42 就会相应地变成，

$$\begin{aligned}
\alpha^* &= \arg \min_{\alpha} \left\{ \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) - \sum_{i=1}^n \alpha_i \right\} \\
\text{subject to } C &\geq \alpha_i \geq 0, i = 1, \dots, n \\
\sum_{i=1}^n \alpha_i y_i &= 0
\end{aligned} \tag{14-44}$$

相应地，决策分类函数就会变成，

$$h_{w^*, b^*}(\mathbf{x}) = \text{sign} \left(\sum_{i=1}^n \alpha_i^* y_i K(\mathbf{x}, \mathbf{x}_i) + b^* \right) \tag{14-45}$$

14.5.2 非线性支持向量机

式 14-44 给出了在映射后的高维特征空间 \mathcal{H} 中的线性支持向量机的学习模型。当映射函数是非线性函数时，学习到的含有核函数的支持向量机是非线性模型，称为**非线性支持向量机**。非线性支持向量机的学习是隐式地在特征空间中进行的，不需要显式地定义特征空间和映射函数。这样的技巧称为核技巧，它是巧妙地利用线性分类学习方法与核函数来解决非线性分类问题的技术。在实际应用中，往往需要依赖领域知识来选择核函数，核函数选择的有效性需要通过实验来验证。

下面我们给出非线性支持向量机学习算法伪码：

算法 14-3：非线性支持向量机学习算法

输入：

$$\text{训练集 } \mathcal{T} = \{(\mathbf{x}_i, y_i) : \mathbf{x}_i \in \mathbb{R}^d, y_i \in \{+1, -1\}\}_{i=1}^n$$

输出：

分类决策函数

(1) 选取合适的核函数 K 和适当的参数 C ，构造并求解优化问题：

$$\begin{aligned}
\alpha^* &= \arg \min_{\alpha} \left\{ \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) - \sum_{i=1}^n \alpha_i \right\} \\
\text{subject to } C &\geq \alpha_i \geq 0, i = 1, \dots, n \\
\sum_{i=1}^n \alpha_i y_i &= 0
\end{aligned}$$

求得的最优解为 $\alpha^* = (\alpha_1^*, \alpha_2^*, \dots, \alpha_n^*)^T$ 。

(2) 从 α^* 中选择一个分量 $0 < \alpha_j^* < C$ ，计算

$$b^* = y_j - \sum_{i=1}^n \alpha_i^* y_i K(\mathbf{x}_i, \mathbf{x}_j)$$

(3) 构造决策函数： $h(\mathbf{x}) = \text{sign} \left(\sum_{i=1}^n \alpha_i^* y_i K(\mathbf{x}, \mathbf{x}_i) + b^* \right)$

最后，我们通过一个具体的例子来感受一下非线性支持向量机在解决非线性分类问题上的能力。如图 14-11 (a) 所示，有一个由二维数据点组成的包含了两个类别的数据集。现在要找一个分类面将此两类数据点完全分开，这显然是一个非线性分类问题。线性支持向量机不能解决该问题，即我们不可能找到一个超平面（在该具体问题中为一条直线）可将图 14-

11 (a) 中的两类数据点完全分开。可使用本节介绍的非线性支持向量机来解决该问题。具体来说，我们采用高斯核函数，基于给定数据集训练出非线性 SVM 模型。若用该模型对二维平面上的数据点进行分类测试，会得到如图 14-11 (b) 所示的实线所代表的分类决策面

(若点 \mathbf{x} 在这条实线上，它满足 $\sum_{i=1}^n \alpha_i^* y_i K(\mathbf{x}, \mathbf{x}_i) + b^* = 0$)：其内侧数据点为负类，其外侧数据点为正类。我们看到，在二维空间中，该非线性分类决策面可将给定数据完全正确分类。

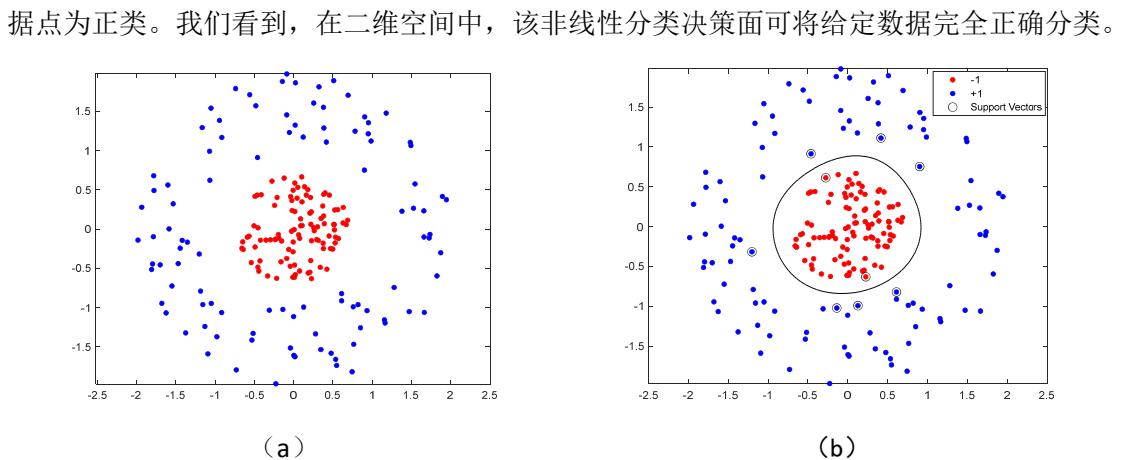


图 14-11：(a) 由二维数据点组成的包含了两个类别的数据集；(b) 实线代表了基于 (a) 中数据训练出的非线性 SVM 分类决策面，圈出的样本点为该分类面的支持向量。

14.6 针对多类分类问题的支持向量机

本章前面介绍的支持向量机分类模型解决的问题都是二类分类问题。而在实际应用中，我们遇到的问题绝大多数都属于多类分类问题。那如何利用二类分类模型来完成多类分类任务呢？幸运的是，针对二类分类问题的支持向量机经简单扩展便可用于解决多类分类问题。我们将介绍两种最常用的扩展方式，“一对多(one-against-all)”的方式和“一对一(one-against-one)”的方式。

“一对多”的方式

假设要解决的问题是一个 K 类分类问题。在“一对多”的方式下，我们需要训练 K 个二类分类器。在训练第 k 个二类分类器 h_k （参数为 \mathbf{w}_k 和 b_k ）的时候，将属于第 k 个类的样本看作正样本，将其余所有类的样本看作负样本。如图 14-12 (a) 所示，要解决的是一个四分类问题，四个类别的数据分别用黑色方块、蓝色五星、红色三角和绿色圆点来代表。为了要解决该四分类问题，我们需要为每一个类别训练一个二类分类器。比如，在训练针对“黑色方块”二类分类器的时候，“黑色方块”会作为正样本，其他所有数据均作为负样本。

在测试阶段，来了一个待分类的测试样本 \mathbf{t} ，我们需要计算 \mathbf{t} 在每一个二类分类器下的响应值（在支持向量机中，如果分类器的响应值为正，则测试样本为正样本，分类器的响应值为负，则测试样本为负样本）。如果 \mathbf{t} 在分类器 h_c 下的响应值 $\mathbf{w}_c^T \mathbf{t} + b_c$ 最大，即

$$\mathbf{w}_c^T \mathbf{t} + b_c = \max_{j=1, \dots, K} \{\mathbf{w}_j^T \mathbf{t} + b_j\}，则 \mathbf{t} 的类别就被判定为 c。在这种多类分类策略下，图 14-12 (a)$$

中所示的四分类问题的分类面如图 14-12 (b) 所示。

尽管“一对多”的扩展方式有其明显的不足之处，比如各个分类器的分类响应值可能不具有相同的尺度、在训练每个二类分类器的时候训练样本是不均衡的等等，但由于该方式容易理解、易于实现且在实际任务中性能表现良好，作为一种将二类分类模型扩展到多类分类模型的策略，该方式在实际应用中被广泛使用^[8]。

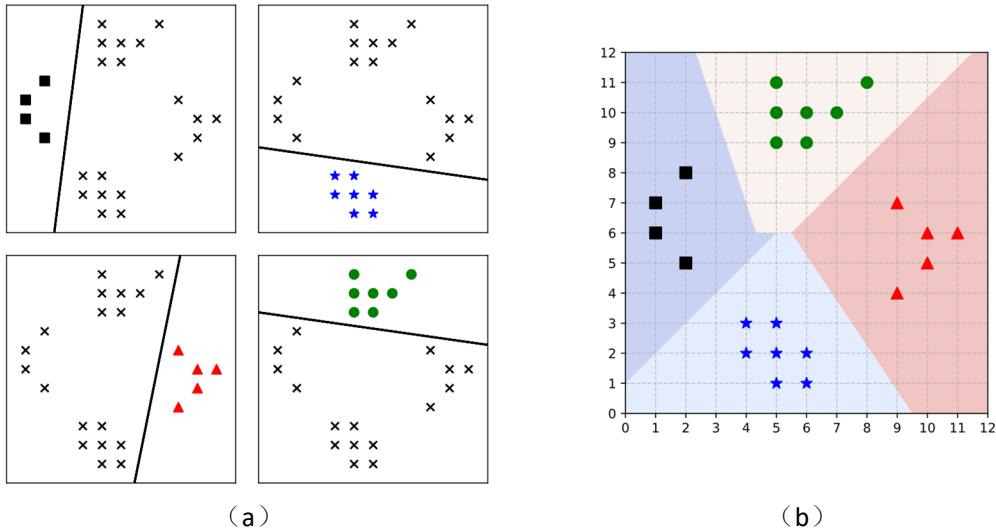


图 14-12：一个四分类问题。(a) “一对多”的多类分类方式会为每一个类别训练一个二类分类器；(b) 按照一对多的方式来解决该四分类问题时所形成的分类面。

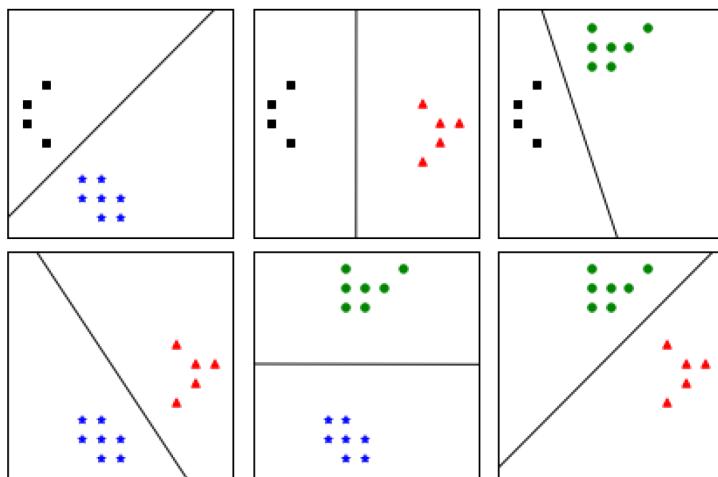


图 14-13：与图 14-12 中一样的四分类问题。若用“一对一”的多类分类方式，我们需要为每两个类别训练一个二类分类器；在这个问题中，共有 4 个类别，因此需要训练 6 个二类分类器。

“一对一”的方式

“一对多”的分类方式是区分一个类和其他所有的类，而“一对一”的分类方式是区分一个类和另一个类。因此，对一个 K 类分类问题来说，在“一对一”的多类分类方式下，我

们需要事先训练好 $\frac{K(K-1)}{2}$ 个二类分类器，即对于每两个类来说，都要训练一个二类分类器，如图 14-13 所示。

在测试阶段，来了一个待分类的测试样本 t ，我们用之前训练好的 $\frac{K(K-1)}{2}$ 个二类分类器逐个对 t 进行分类预测，当然就会得到 $\frac{K(K-1)}{2}$ 个分类预测结果，然后使用“投票”的方法得到 t 的最终分类预测结果，即得票最多的类别就作为 t 的最终类别。

14.7 习题

- (7) 运行并理解与本章配套的 Matlab 示例程序“hard-margin SVM”。基于仿真数据，该程序示范了如何从线性可分的数据集中，利用硬间隔支持向量机模型学习出最优分类超平面。该程序可生成类似于图 14-8 的可视化结果。
- (8) 运行并理解与本章配套的 Matlab 示例程序“soft-margin SVM”。基于仿真数据，该程序示范了软间隔支持向量机的工作方式。该程序可生成类似于图 14-9 的可视化结果。
- (9) 运行并理解与本章配套的 Matlab 示例程序“rbf-kernel SVM”。基于仿真数据，该程序示范了基于核技巧的非线性支持向量机的工作方式。该程序可生成类似于图 14-11 的可视化结果。

参考文献

- [1] S. Boyd and L. Vandenberghe, Convex Optimization, Cambridge University Press, 2004.
- [2] F. Rosenblatt, The perceptron: A probabilistic model for information storage and organization in the brain, Psychological Review, vol. 65, no. 6, pp. 386-408, 1958.
- [3] C. Tappert, Who is the father of deep learning? Proc. IEEE International Conference on Computational Science and Computational Intelligence (CSCI), pp. 343-348, 2019.
- [4] A. Novikoff, On convergence proofs on perceptrons, Symposium on the Mathematical Theory of Automata, Polytechnic Institute of Brooklyn, pp. 615-622, 1962.
- [5] 李航，统计学习方法（第二版），清华大学出版社，2019 年。
- [6] V.N. Vapnik and A.Y. Chervonenkis, Theory of pattern recognition: Statistical problems of learning[俄文版], Nauka, Mosco, 1974.
- [7] J. Platt, Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines, Technical Report, Microsoft, Apr. 1998.
- [8] C.M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.

第 15 章 深度神经网络及基于深度神经网络的目标检测

在第 7 章中提到，为了便于使用视觉技术来对平面上的目标进行检测或测量，我们可以生成物理平面的鸟瞰视图。鸟瞰视图又称为**逆透视投影**，这是因为当拍摄物理平面信息时，在针孔相机模型下，图像平面是物理平面通过透视投影产生的。逆透视投影便是将图像平面信息反投影至它所对应的物理平面上，得到物理平面的“像素化”表示。鸟瞰视图被广泛应用在辅助驾驶中的环视系统和各类工业流水线中的工件属性测量系统中。我们将在这一章学习如何从物理平面的图像中构造出该平面的鸟瞰视图。

参考文献

- [1] L. Zhang, X. Li, J. Huang, Y. Shen and D. Wang, “Vision-based parking-slot detection: A benchmark and a learning-based approach,” *Symmetry*, vol. 2018, no. 10, pp. 64:1-18, 2018.

第四篇：三维立体视觉

第 16 章 三维重建问题概述

在第 7 章中提到，为了便于使用视觉技术来对平面上的目标进行检测或测量，我们可以生成物理平面的鸟瞰视图。鸟瞰视图又称为**逆透视投影**，这是因为当拍摄物理平面信息时，在针孔相机模型下，图像平面是物理平面通过透视投影产生的。逆透视投影便是将图像平面信息反投影至它所对应的物理平面上，得到物理平面的“像素化”表示。鸟瞰视图被广泛应用在辅助驾驶中的环视系统和各类工业流水线中的工件属性测量系统中。我们将在这一章学习如何从物理平面的图像中构造出该平面的鸟瞰视图。

参考文献

- [2] L. Zhang, X. Li, J. Huang, Y. Shen and D. Wang, “Vision-based parking-slot detection: A benchmark and a learning-based approach,” *Symmetry*, vol. 2018, no. 10, pp. 64:1-18, 2018.

第 17 章 运动恢复结构

在第 7 章中提到，为了便于使用视觉技术来对平面上的目标进行检测或测量，我们可以生成物理平面的鸟瞰视图。鸟瞰视图又称为**逆透视投影**，这是因为当拍摄物理平面信息时，在针孔相机模型下，图像平面是物理平面通过透视投影产生的。逆透视投影便是将图像平面信息反投影至它所对应的物理平面上，得到物理平面的“像素化”表示。鸟瞰视图被广泛应用在辅助驾驶中的环视系统和各类工业流水线中的工件属性测量系统中。我们将在这一章学习如何从物理平面的图像中构造出该平面的鸟瞰视图。

参考文献

- [3] L. Zhang, X. Li, J. Huang, Y. Shen and D. Wang, “Vision-based parking-slot detection: A benchmark and a learning-based approach,” *Symmetry*, vol. 2018, no. 10, pp. 64:1-18, 2018.

第 18 章 神经辐射场

2020 年，美国加州大学伯克利分校的学者 Mildenhall 等人提出了神经辐射场（Neural Radiance Field）的概念^[1]。该工作的核心思想是用神经网络来对三维视觉场景进行隐式表达。最早提出的时候，该方法是用来解决场景的新视角合成问题的。很快，它被推广到更多的应用领域，比如三维重建以及实时建图与定位等。

本章将首先介绍什么是场景的辐射场表达以及如何基于场景的辐射场来生成场景的渲染图片。之后介绍基于神经网络的场景辐射场的隐式表达及其训练方法。

18.1 基于辐射场的体渲染

18.1.1 连续型形式

体渲染（volume rendering）属于计算机图形学研究范畴。这种渲染方式将景物所在的物理空间考虑成辐射场（radiance fields）：空间中的每一点都具有不透明度（opacity）和与观察方向有关的颜色（view-dependent color）两个属性。渲染操作的目的是生成场景在某一虚拟相机视角下的照片。虚拟相机的视角由它在世界坐标系下的位姿来表达。渲染操作的本质便是要确定出虚拟相机成像平面上每一点的颜色。

如图 18-1 所示，考虑虚拟相机成像平面上像素点 \mathbf{p} ，我们来看看如何确定 \mathbf{p} 的颜色值。连接相机光心 \mathbf{o} 和 \mathbf{p} 便得到了一条从相机光心 \mathbf{o} 出发经过 \mathbf{p} 的光线。不难理解，像素点 \mathbf{p} 的颜色将完全取决于光线 $\overrightarrow{\mathbf{op}}$ 在传播过程中所遇到的景物。设 $\overrightarrow{\mathbf{op}}$ 的方向为单位向量 \mathbf{d} ，则光线 $\overrightarrow{\mathbf{op}}$ 可表达为 $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}, t \in [t_n, t_f]$ ，其中 t_n, t_f 分别表示近远减裁面，即在渲染计算过程中，我们只考虑有限长度的光线。依据体渲染原理，像素点 \mathbf{p} 处的颜色被计算为，

$$\hat{\mathbf{u}}(\mathbf{p}) = \int_{t_n}^{t_f} T(t) \sigma(\mathbf{r}(t)) \mathbf{c}(\mathbf{r}(t), \mathbf{d}) dt \quad (18-1)$$

其中， $\sigma(\mathbf{x}), \mathbf{x} \in \mathbb{R}^3$ 表示空间位置 \mathbf{x} 处的不透明度， $\mathbf{c}(\mathbf{x}, \mathbf{d})$ 表示空间位置 \mathbf{x} 处在观察方向为 \mathbf{d} 时所观察到的颜色， $T(t) = \exp\left(-\int_{t_n}^t \sigma(\mathbf{r}(s)) ds\right)$ 表示沿着光线 $\overrightarrow{\mathbf{op}}$ 从 $\mathbf{r}(t_n)$ 到 $\mathbf{r}(t)$ 的累积透明度。

渲染公式式 18-1 说明相机成像平面上 \mathbf{p} 点的颜色是通过累积光线 $\overrightarrow{\mathbf{op}}$ 一路行来遇到的所有点的贡献得来的。具体来说，场景中点 $\mathbf{r}(t)$ 处的贡献为 $T(t) \sigma(\mathbf{r}(t)) \mathbf{c}(\mathbf{r}(t), \mathbf{d})$ ，即该点的贡献取决于该点自身的与观察方向有关的颜色 $\mathbf{c}(\mathbf{r}(t), \mathbf{d})$ 、该点的不透明度 $\sigma(\mathbf{r}(t))$ （显然，越透明，该点的贡献越少）以及该点处的累积透明度 $T(t)$ （显然，累积透明度越高，该点贡献度越大）。因此，公式 18-1 所表达的体渲染流程符合我们的直观认知。

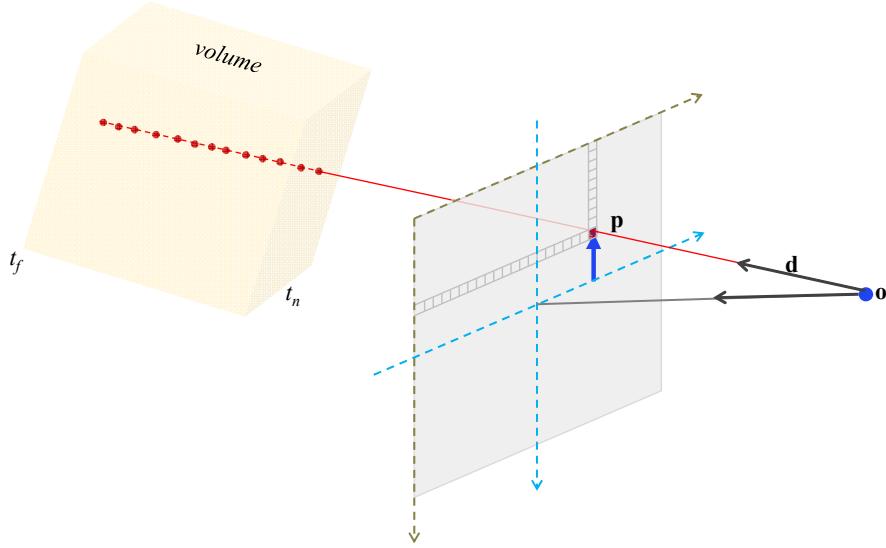


图 18-1：体渲染原理：为确定相机成像平面上一点 p 的颜色值，首先需要确定出光线 \overrightarrow{op} ，之后根据该光线在渲染体内遇到的“景物”信息，按照渲染公式（式 18-1）计算出 p 点的颜色值。

18.1.2 离散型形式

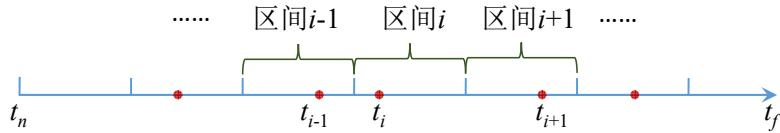


图 18-2：离散化体渲染计算原理。

不难理解，如果用式 18-1 来计算 p 点处的颜色值，我们需要在光线 \overrightarrow{op} 传播的连续路径上进行积分。这种计算方式难以转换为计算机程序。为了方便计算机编程，我们需要给出式 18-1 的离散表达形式。将式 18-1 的连续积分形式转换为离散求和形式的基本思路就是将积分区间 $[t_n, t_f]$ 划分为有限个小区间，然后在每个小区间内采样一个代表点，并近似认为该区间内其他点的属性都与该代表点相同，最后将式 18-1 的积分近似为有限区间上的求和。图 18-2 示意了式 18-1 离散化的主要过程。

将积分区间 $[t_n, t_f]$ 划分为 N 个小区间，那么第 i 个小区间便是 $\left[t_n + \frac{i-1}{N}(t_f - t_n), t_n + \frac{i}{N}(t_f - t_n) \right]$, $i = 1, 2, \dots, N$ 。然后，在每个小区间内按照均匀分布随机采样出一个计算点。在区间 i 内的采样点记为 t_i ，其值服从如下均匀分布，

$$t_i \sim U\left[t_n + \frac{i-1}{N}(t_f - t_n), t_n + \frac{i}{N}(t_f - t_n)\right]$$

这样，式 18-1 的离散化近似为，

$$\hat{\mathbf{u}}(\mathbf{p}) \approx \sum_{i=1}^{N-1} \hat{\mathbf{u}}_i(\mathbf{p}) = \sum_{i=1}^{N-1} \int_{t_i}^{t_{i+1}} T(t) \sigma(\mathbf{r}(t)) \mathbf{c}(\mathbf{r}(t), \mathbf{d}) dt \quad (18-2)$$

上式求和中的每一项 $\hat{\mathbf{u}}_i(\mathbf{p})$ 表达的是光线在 $\mathbf{r}(t_i)$ 到 $\mathbf{r}(t_{i+1})$ 之间所遇到的“景物”对最终 \mathbf{p} 点颜色的贡献,

$$\hat{\mathbf{u}}_i(\mathbf{p}) = \int_{t_i}^{t_{i+1}} T(t) \sigma(\mathbf{r}(t)) \mathbf{c}(\mathbf{r}(t), \mathbf{d}) dt \approx \int_{t_i}^{t_{i+1}} \exp\left(-\int_{t_n}^t \sigma(s) ds\right) \sigma_i \mathbf{c}_i dt \quad (18-3)$$

在上式中, 我们将区间 $[t_i, t_{i+1}]$ 内所有对应的点的不透明度都近似为常数 $\sigma_i = \sigma(\mathbf{r}(t_i))$ 、颜色都近似为常数 $\mathbf{c}_i = \mathbf{c}(\mathbf{r}(t_i), \mathbf{d})$, 但 $T(t)$ 的值在小区间 $[t_i, t_{i+1}]$ 内随 t 的变化而变化, 不能视为常数。式 18-3 可进一步化为,

$$\begin{aligned} \hat{\mathbf{u}}_i(\mathbf{p}) &\approx \sigma_i \mathbf{c}_i \int_{t_i}^{t_{i+1}} \exp\left(-\int_{t_n}^t \sigma(s) ds\right) dt \\ &= \sigma_i \mathbf{c}_i \int_{t_i}^{t_{i+1}} \exp\left(-\int_{t_n}^{t_i} \sigma(s) ds\right) \exp\left(-\int_{t_i}^t \sigma(s) ds\right) dt \\ &= \sigma_i \mathbf{c}_i T_i \int_{t_i}^{t_{i+1}} \exp\left(-\int_{t_i}^t \sigma(s) ds\right) dt \end{aligned} \quad (18-4)$$

在上式推导的最后一步中, 我们把点 $\mathbf{r}(t_i)$ 处的累积透明度记为了 T_i , 即 $T_i = \exp\left(-\int_{t_n}^{t_i} \sigma(s) ds\right)$,

T_i 与积分变量 t 无关, 因此可以拿到积分 $\int_{t_i}^{t_{i+1}}$ 之外。式 18-4 中, 最后一步中的积分部分为,

$$\begin{aligned} \int_{t_i}^{t_{i+1}} \exp\left(-\int_{t_i}^t \sigma(s) ds\right) dt &\approx \int_{t_i}^{t_{i+1}} \exp(-\sigma_i(t - t_i)) dt \\ &= \frac{\exp(-\sigma_i(t - t_i))}{-\sigma_i} \Big|_{t_i}^{t_{i+1}} = \frac{1}{\sigma_i} [1 - \exp(-\sigma_i(t_{i+1} - t_i))] \end{aligned}$$

因此,

$$\hat{\mathbf{u}}_i(\mathbf{p}) \approx \sigma_i \mathbf{c}_i T_i \int_{t_i}^{t_{i+1}} \exp\left(-\int_{t_i}^t \sigma(s) ds\right) dt \approx \sigma_i \mathbf{c}_i T_i \frac{1}{\sigma_i} [1 - \exp(-\sigma_i(t_{i+1} - t_i))] = \mathbf{c}_i T_i [1 - \exp(-\sigma_i(t_{i+1} - t_i))]$$

记采样点间隔为 $\delta_i \triangleq (t_{i+1} - t_i)$, 则有,

$$\hat{\mathbf{u}}(\mathbf{p}) \approx \sum_{i=1}^{N-1} \hat{\mathbf{u}}_i(\mathbf{p}) = \sum_{i=1}^{N-1} \mathbf{c}_i T_i [1 - \exp(-\sigma_i \delta_i)] \quad (18-5)$$

最后, 我们还需要对 T_i 的计算进行一下离散化:

$$\begin{aligned} T_i &= \exp\left(-\int_{t_n}^{t_i} \sigma(s) ds\right) \\ &\approx \exp\left(-[\sigma_1(t_2 - t_1) + \sigma_2(t_3 - t_2) + \dots + \sigma_{i-1}(t_i - t_{i-1})]\right) \\ &= \exp\left(-\sum_{j=1}^{i-1} \sigma_j (t_{j+1} - t_j)\right) \\ &= \exp\left(-\sum_{j=1}^{i-1} \sigma_j \delta_j\right) \end{aligned} \quad (18-6)$$

式 18-5 和 18-6 一起构成了连续型体渲染模型式 18-1 的离散化表达。

18.2 辐射场的隐式表达及其学习

18.2.1 辐射场的隐式表达

在 18.1 节中提到，给定了场景的辐射场以后，我们便可以使用体渲染技术渲染出任意虚拟相机视角下的场景照片。那么场景的辐射场是如何来表示的呢？从 18.1 节中的内容可知，对于给定的辐射场来说，在确定了一个空间位置 \mathbf{x} 以及观察方向 \mathbf{d} 之后，我们希望知道的信息是辐射场中 \mathbf{x} 处的不透明度 σ 以及该点处与观察方向 \mathbf{d} 有关的颜色值 \mathbf{c} 。因此，场景的辐射场可以被自然地表达为一个映射，

$$F_{\Theta}(\mathbf{x}, \mathbf{d}) \rightarrow (\mathbf{c}, \sigma) \quad (18-7)$$

其中， F_{Θ} 为映射模型， Θ 为该模型的参数集合； \mathbf{x} 表示辐射场中的某个三维空间位置， \mathbf{d} 为表示观察方向的三维单位向量， σ 为辐射场中点 \mathbf{x} 处的不透明度， $\mathbf{c}=(r, g, b)$ 为 \mathbf{x} 处与观察方向 \mathbf{d} 有关的颜色向量。对于某个给定场景来说，当其映射模型 F_{Θ} 被确定之后，这个场景相应的辐射场的表达也就确定了。

一个很自然的想法便是用神经网络来隐式地表达映射模型 F_{Θ} ，从 5 维的输入向量 (\mathbf{x}, \mathbf{d}) 中直接回归出 4 维输出向量 (\mathbf{c}, σ) 。然而，Mildenhall 等^[1]在实验中发现，上述的简单处理方式并不能很好地刻画场景的高频细节。为了解决该问题，他们提出了位置数据与方向数据升维编码的概念，用升维编码后的位置与方向数据（再加上原始的位置与方向数据）作为神经网络 F_{Θ} 的输入，目的是为了能让神经网络更好地表达场景中的高频信息。具体来说，升维编码过程可表达为函数 $\gamma(p): \mathbb{R} \rightarrow \mathbb{R}^{2L}$ ，

$$\gamma(p) = (\sin(2^0 \pi p), \cos(2^0 \pi p), \dots, \sin(2^{L-1} \pi p), \cos(2^{L-1} \pi p)) \quad (18-8)$$

举个具体例子来说，对于位置数据部分，取 $L=10$ ，则升维编码后的位置数据的维数为 60 ($2 \times 10 \times 3$)，再加上原始的位置数据自身，则与位置有关的输入数据的维度就是 63 维；对于方向数据¹³部分，取 $L=4$ ，则升维编码后的方向数据的维数为 24 ($2 \times 4 \times 3$)，再加上原始的方向数据自身，则与观察方向有关的输入数据的维度就是 27 维。

确定好了输入与输出以后，神经网络 F_{Θ} 应该怎样设计呢？由于辐射场中某点处的不透明度只与其位置有关，因此我们希望在回归不透明度时只引入与位置有关的输入；而另一方面，辐射场中某点处的颜色值既与该点的空间位置有关，又与相机的观察方向有关，因此在回归某点处的颜色信息时，需要同时使用位置数据和观察方向数据。基于这些考虑，Mildenhall 等^[1]设计了具有如下结构的神经网络 F_{Θ} 来隐式表达场景的辐射场：从整体上来说，

¹³ 三维空间中表示方向的单位向量实际上只有 2 个自由度，但为了方便起见，Mildenhall 等^[1]把它显式表示为了一个三维向量。

F_Θ 是一个多层全连接网络，它先用 8 个具有 256 个输出节点的全连接层来处理升维编码后的位置数据（63 维），最后一层回归出不透明度值和一个 256 维的特征向量 \mathbf{f} ，这其中还使用了一次跳跃连接，把输入向量直接级联到第 5 个全连接层的输出向量；然后把 \mathbf{f} 和升维编码后的观察方向数据（27 维）级联在一起，形成既与位置信息有关又与观察方向信息有关的向量 \mathbf{f}' ，之后再用一个具有 128 个输出节点的全连接层来处理 \mathbf{f}' ，并最终回归出代表颜色信息的三维向量。图 18-3 给出了 F_Θ 完整的网络结构图。



图 18-3: Mildenhall 等^[1]提出的用于隐式表达辐射场的神经网络结构。

18.2.2 神经辐射场的学习

■ 单一网络表示

在 18.2.1 节中讲到，场景的辐射场可被隐式地表达为一个神经网络 F_Θ ，当 F_Θ 给定以后，便可以从给定的位置与观察方向向量 (\mathbf{x}, \mathbf{d}) “查找出” 相应位置处的辐射场信息（不透明度以及与观察方向有关的颜色）。然而，对于给定的场景，其辐射场表达网络 F_Θ 是如何得到的呢？

对于某一场景，为了要训练出表达该场景辐射场的神经网络 F_Θ ，我们首先要有一组拍摄自该场景的、相机内外参数已知的照片 \mathcal{P} 。基于照片集合 \mathcal{P} ，便可以学习出 F_Θ ，具体思路如下。假设 I 是 \mathcal{P} 中的一张照片， \mathbf{p} 为 I 上一点，其颜色值为 $\mathbf{u}(\mathbf{p})$ 。另一方面，我们也可以根据场景的辐射场 F_Θ ，按照式 18-5 和式 18-6 所述的体渲染方式计算出 \mathbf{p} 点的颜色值 $\hat{\mathbf{u}}(\mathbf{p})$ 。

显然，场景的辐射场 F_Θ 越准确，渲染模型计算出来的像素值 $\hat{\mathbf{u}}(\mathbf{p})$ 就越接近于该点的颜色真

值 $\mathbf{u}(\mathbf{p})$ 。因此，很自然地，我们可以用 $\hat{\mathbf{u}}(\mathbf{p})$ 与 $\mathbf{u}(\mathbf{p})$ 之间的差异来构造训练 F_Θ 的损失函数

$$l(\Theta),$$

$$l(\Theta) = \sum_{\mathbf{p} \in \Omega} \left\| \hat{\mathbf{u}}(\mathbf{p}) - \mathbf{u}(\mathbf{p}) \right\|_2^2 \quad (18-9)$$

其中， Ω 为每次训练迭代中从 \mathcal{P} 中随机选取出的图像像素位置集合。

还有一处细节需要详述一下：对于给定的像素位置 \mathbf{p} （齐次坐标表示），要想使用体渲染技术计算出该点的颜色值，我们需要确定出与 \mathbf{p} 对应的辐射场中的光线方程。假设相机的内参矩阵为 K ，拍摄像素 \mathbf{p} 所属的图像时的相机位姿矩阵为 $T_{WC} \in \mathbb{R}^{3 \times 4}$ （即，相机坐标系的一点左乘 T_{WC} 之后，就得到了该点在世界坐标系下的坐标）。这样的话，相机光心在世界坐标系的坐标为 $\mathbf{o}_w = T_{WC}(0 0 0 1)^T$ 。与 \mathbf{p} 对应的归一化成像平面上的点的齐次坐标为 $K^{-1}\mathbf{p}$ ，则

该点在相机坐标系下的齐次坐标为 $\begin{pmatrix} K^{-1}\mathbf{p} \\ 1 \end{pmatrix}$ ，更进一步，其在世界坐标系下的坐标为

$\mathbf{p}_w = T_{WC} \begin{pmatrix} K^{-1}\mathbf{p} \\ 1 \end{pmatrix}$ 。这样，我们便得到了与 \mathbf{p} 对应的辐射场中的光线 $\overrightarrow{\mathbf{o}_w \mathbf{p}_w}$ 。有了光线 $\overrightarrow{\mathbf{o}_w \mathbf{p}_w}$ 之后，我们便可以按照式 18-5 和式 18-6 所述的体渲染方式计算出 \mathbf{p} 点的颜色值 $\hat{\mathbf{u}}(\mathbf{p})$ 。

■ 双网络表示

参考文献

- [1] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, “NeRF: Representing scenes as neural radiance fields for view synthesis,” *Proc. ECCV*, pp. 405–421, 2020.

附录

A. 圆锥曲线^[1]

在笛卡尔平面坐标系下，圆锥曲线的一般方程表示为，

$$Ax^2 + Bxy + Cy^2 + Dx + Ey + F = 0 \quad (\text{A-1})$$

其中所有系数都为实数且 A 、 B 和 C 不能同时为零。同时，我们也很容易得出圆锥曲线的等价矩阵表达形式，

$$(xy) \begin{bmatrix} A & \frac{B}{2} \\ \frac{B}{2} & C \end{bmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + (DE) \begin{pmatrix} x \\ y \end{pmatrix} + F = 0 \quad (\text{A-2})$$

或者是，

$$(xy1) \begin{bmatrix} A & \frac{B}{2} & \frac{D}{2} \\ \frac{B}{2} & C & \frac{E}{2} \\ \frac{D}{2} & \frac{E}{2} & F \end{bmatrix} \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} = 0 \quad (\text{A-3})$$

圆锥曲线的具体类型可以根据判别式 $B^2 - 4AC$ 来决定：

- 1) 如果 $B^2 - 4AC < 0$ ，该圆锥曲线为椭圆；在此条件下，如果更进一步有 $A = C$ 并且 $B = 0$ ，则圆锥曲线表示圆；
- 2) 如果 $B^2 - 4AC = 0$ ，该圆锥曲线为抛物线；
- 3) 如果 $B^2 - 4AC > 0$ ，该圆锥曲线为双曲线。

B. 数字图像导数的近似计算

在做理论分析时，我们经常会把图像 $f(x, y) : \mathbb{R}^2 \rightarrow \mathbb{R}$ 考虑为连续函数，且有很多情况需要计算 f 的各阶偏导数。但在编程实现时，实际的图像为离散数字图像，为此我们需要

有一套计算数字图像近似导数的机制。假设 (x, y) 为图像 f 上的整数位置点，我们的任务是

要近似计算 f 在点 (x, y) 处的一阶与二阶偏导数 $\frac{\partial f}{\partial x}$ 、 $\frac{\partial f}{\partial y}$ 、 $\frac{\partial^2 f}{\partial x^2}$ 、 $\frac{\partial^2 f}{\partial y^2}$ 和 $\frac{\partial^2 f}{\partial x \partial y}$ 。

在推导图像导数近似计算表达式的过程中，我们暂时要假设图像函数 $f(x, y)$ 为连续函数且具有二阶偏导数。函数 $f(x, y)$ 在点 (x, y) 近旁的二阶泰勒展开为，

$$f(x+h, y+k) \approx f(x, y) + h \frac{\partial f}{\partial x} + k \frac{\partial f}{\partial y} + \frac{1}{2} h^2 \frac{\partial^2 f}{\partial x^2} + h k \frac{\partial^2 f}{\partial x \partial y} + \frac{1}{2} k^2 \frac{\partial^2 f}{\partial y^2} \quad (\text{B-1})$$

其中， h 和 k 为小量。对于数字图像来说， h 和 k 为整数。取 $h=1$ 、 $k=0$ ，我们有，

$$f(x+1, y) \approx f(x, y) + \frac{\partial f}{\partial x} + \frac{1}{2} \frac{\partial^2 f}{\partial x^2} \quad (\text{B-2})$$

取 $h=-1$ 、 $k=0$ ，我们有，

$$f(x-1, y) \approx f(x, y) - \frac{\partial f}{\partial x} + \frac{1}{2} \frac{\partial^2 f}{\partial x^2} \quad (\text{B-3})$$

式 B-2 与式 B-3 两端相减并适当变形得到，

$$\frac{\partial f}{\partial x} \approx \frac{f(x+1, y) - f(x-1, y)}{2} \quad (\text{B-4})$$

采样类似的方式可以得到，

$$\frac{\partial f}{\partial y} \approx \frac{f(x, y+1) - f(x, y-1)}{2} \quad (\text{B-5})$$

式 B-2 与式 B-3 两端相加并稍加变形得到，

$$\frac{\partial^2 f}{\partial x^2} = f(x+1, y) + f(x-1, y) - 2f(x, y) \quad (\text{B-6})$$

取 $h=0$ 、 $k=1$ ，我们有，

$$f(x, y+1) \approx f(x, y) + \frac{\partial f}{\partial y} + \frac{1}{2} \frac{\partial^2 f}{\partial y^2} \quad (\text{B-7})$$

取 $h=0$ 、 $k=-1$ ，我们有，

$$f(x, y-1) \approx f(x, y) - \frac{\partial f}{\partial y} + \frac{1}{2} \frac{\partial^2 f}{\partial y^2} \quad (\text{B-8})$$

式 B-8 与式 B-9 两端相加并稍加变形得到，

$$\frac{\partial^2 f}{\partial y^2} = f(x, y+1) + f(x, y-1) - 2f(x, y) \quad (\text{B-9})$$

取 $h=1$ 、 $k=1$ ，我们有，

$$f(x+1, y+1) \approx f(x, y) + \frac{\partial f}{\partial x} + \frac{\partial f}{\partial y} + \frac{1}{2} \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial x \partial y} + \frac{1}{2} \frac{\partial^2 f}{\partial y^2} \quad (\text{B-10})$$

取 $h=-1$ 、 $k=-1$, 我们有,

$$f(x-1, y-1) \approx f(x, y) - \frac{\partial f}{\partial x} - \frac{\partial f}{\partial y} + \frac{1}{2} \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial x \partial y} + \frac{1}{2} \frac{\partial^2 f}{\partial y^2} \quad (\text{B-11})$$

取 $h=1$ 、 $k=-1$, 我们有,

$$f(x+1, y-1) \approx f(x, y) + \frac{\partial f}{\partial x} - \frac{\partial f}{\partial y} + \frac{1}{2} \frac{\partial^2 f}{\partial x^2} - \frac{\partial^2 f}{\partial x \partial y} + \frac{1}{2} \frac{\partial^2 f}{\partial y^2} \quad (\text{B-12})$$

取 $h=-1$ 、 $k=1$, 我们有,

$$f(x-1, y+1) \approx f(x, y) - \frac{\partial f}{\partial x} + \frac{\partial f}{\partial y} + \frac{1}{2} \frac{\partial^2 f}{\partial x^2} - \frac{\partial^2 f}{\partial x \partial y} + \frac{1}{2} \frac{\partial^2 f}{\partial y^2} \quad (\text{B-13})$$

式 B-10 和式 B-12 两端相减得到,

$$f(x+1, y+1) - f(x+1, y-1) \approx 2 \frac{\partial f}{\partial y} + 2 \frac{\partial^2 f}{\partial x \partial y} \quad (\text{B-14})$$

式 B-11 和式 B-13 两端相减得到,

$$f(x-1, y-1) - f(x-1, y+1) \approx -2 \frac{\partial f}{\partial y} + 2 \frac{\partial^2 f}{\partial x \partial y} \quad (\text{B-15})$$

式 B-14 和式 B-15 两端相加得到,

$$\frac{\partial^2 f}{\partial x \partial y} = \frac{f(x+1, y+1) + f(x-1, y-1) - f(x+1, y-1) - f(x-1, y+1)}{4} \quad (\text{B-16})$$

上面我们推导了二元离散函数的一阶与二阶偏导数的近似计算方式, 实际上这些结果可以直接推广到三元离散函数的情况。比如在 4.2.2 节中, 我们把 DoG 尺度空间看作三元函数 $f(x, y, l)$, 其中 x 、 y 为空间位置、 l 为尺度层的序号, 因此 x 、 y 和 l 都为整数。我们可以用与本节完全类似的方式来自近似计算函数 $f(x, y, l)$ 在 (x, y, l) 处的梯度向量和海森矩阵。

C. 高斯函数的卷积及其傅里叶变换

设 $g(x, y) : \mathbb{R}^2 \rightarrow \mathbb{R}$ 为二维高斯形状的函数, 则其傅里叶变换 $\mathcal{G}(u, v)$ (u, v 为傅里叶频率域中的角频率坐标) 也为高斯形状函数。具体来说, 若

$$g(x, y) = e^{-\pi(a^2 x^2 + b^2 y^2)} \quad (\text{C-1})$$

其中 a 、 b 为非零常数, 则 $g(x, y)$ 的傅里叶变换 $\mathcal{G}(u, v)$ 为^[2],

$$\mathcal{G}(u, v) = \frac{1}{|ab|} e^{-\frac{1}{4\pi} \left(\frac{u^2}{a^2} + \frac{v^2}{b^2} \right)} \quad (\text{C-2})$$

根据式 C-2, 容易证明, 若 $g_1(x,y;\sigma_1) = \frac{1}{2\pi\sigma_1^2} e^{-\frac{(x^2+y^2)}{2\sigma_1^2}}$, $g_2(x,y;\sigma_2) = \frac{1}{2\pi\sigma_2^2} e^{-\frac{(x^2+y^2)}{2\sigma_2^2}}$, 则 g_1

与 g_2 的卷积 g_3 为,

$$g_3(x,y;\sqrt{\sigma_1^2 + \sigma_2^2}) = \frac{1}{2\pi(\sigma_1^2 + \sigma_2^2)} e^{-\frac{x^2+y^2}{2(\sigma_1^2 + \sigma_2^2)}} \quad (C-3)$$

即 g_3 也为高斯函数, 其标准差为 $\sqrt{\sigma_1^2 + \sigma_2^2}$ 。式 C-3 的简要证明如下。

令式 C-1 与 C-2 中的 $a = \frac{1}{\sqrt{2\pi}\sigma}$, $b = \frac{1}{\sqrt{2\pi}\sigma}$, 则我们有傅里叶变换对,

$$e^{-\frac{x^2+y^2}{2\sigma^2}} \leftrightarrow 2\pi\sigma^2 e^{-\frac{(u^2+v^2)\sigma^2}{2}} \quad (C-4)$$

由式 C-4 可知, $g_1(x,y;\sigma_1)$ 与 $g_2(x,y;\sigma_2)$ 的傅里叶变换分别为 $\mathcal{G}_1(u,v) = e^{-\frac{(u^2+v^2)\sigma_1^2}{2}}$ 和 $\mathcal{G}_2(u,v) = e^{-\frac{(u^2+v^2)\sigma_2^2}{2}}$ 。则 g_1 与 g_2 的卷积为,

$$\begin{aligned} g_1 * g_2 &= \mathcal{F}^{-1}(\mathcal{G}_1(u,v) \cdot \mathcal{G}_2(u,v)) \\ &= \mathcal{F}^{-1}\left(e^{-\frac{(u^2+v^2)\sigma_1^2}{2}} e^{-\frac{(u^2+v^2)\sigma_2^2}{2}}\right) \\ &= \mathcal{F}^{-1}\left(e^{-\frac{(u^2+v^2)(\sigma_1^2 + \sigma_2^2)}{2}}\right) \\ &= \frac{1}{2\pi(\sigma_1^2 + \sigma_2^2)} e^{-\frac{x^2+y^2}{2(\sigma_1^2 + \sigma_2^2)}} \end{aligned} \quad (C-5)$$

其中 $\mathcal{F}^{-1}(\cdot)$ 表示傅里叶反变换, $\mathcal{G}_1(u,v) \cdot \mathcal{G}_2(u,v)$ 表示 $\mathcal{G}_1(u,v)$ 与 $\mathcal{G}_2(u,v)$ 在傅里叶频率域中频率坐标 (u,v) 处的普通复数乘法。

D. 主曲率与海森矩阵

我们知道, 曲率是度量曲线局部弯曲程度的几何量。对于曲面来说, 我们也可以定义曲面上某一点 s 的曲率。设曲面在 s 点的法线为 z 轴, 过 z 轴可以有无限多个剖切平面, 每个剖切平面与曲面相交, 其交线为一条平面曲线, 每条平面曲线在 s 点都有一个曲率。设这些曲率中的最大值为 κ_{\max} 、最小值为 κ_{\min} , 这两个曲率称为曲面在 s 点的主曲率 (principal curvatures)。基于两个主曲率 κ_{\max} 和 κ_{\min} , 可以定义平均曲率 $\kappa_a = (\kappa_1 + \kappa_2)/2$ 和高斯曲率

$$\kappa_G = \kappa_1 \kappa_2.$$

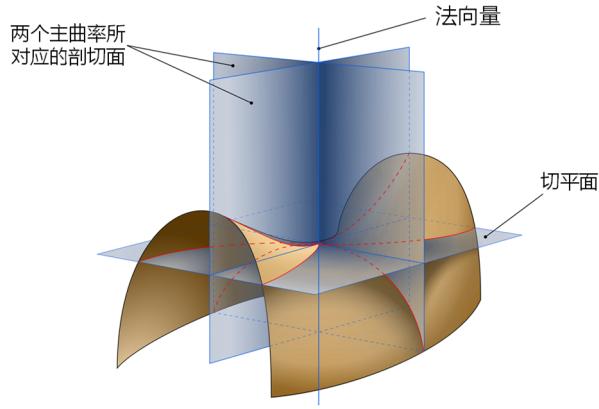


图 D-1：曲面上一点的两个主曲率。

我们考虑一种特殊的由二元函数所确定的三维曲面。设 $f(\mathbf{x}) : \mathbb{R}^2 \rightarrow \mathbb{R}, \mathbf{x} \in \mathbb{R}^2$ 为二元连续函数且具有二阶偏导数，则由该函数可确定出一个三维欧氏空间中的曲面 $(\mathbf{x}, f(\mathbf{x}))$ 。根据微分几何的知识^[3]，该曲面上一点 $(\mathbf{x}, f(\mathbf{x}))$ 处的平均曲率为，

$$\kappa_a(\mathbf{x}, f(\mathbf{x})) = \frac{\kappa_{\max}(\mathbf{x}, f(\mathbf{x})) + \kappa_{\min}(\mathbf{x}, f(\mathbf{x}))}{2} = \frac{(1+f_x^2)f_{yy} + (1+f_y^2)f_{xx} - 2f_x f_y f_{xy}}{2(1+f_x^2 + f_y^2)^{3/2}} \quad (\text{D-1})$$

高斯曲率为，

$$\kappa_G(\mathbf{x}, f(\mathbf{x})) = \kappa_{\max}(\mathbf{x}, f(\mathbf{x})) \cdot \kappa_{\min}(\mathbf{x}, f(\mathbf{x})) = \frac{f_{xx} f_{yy} - f_{xy}^2}{(1+f_x^2 + f_y^2)^2} \quad (\text{D-2})$$

若 \mathbf{x}_0 为 f 的驻点，则 $f_x|_{\mathbf{x}=\mathbf{x}_0} = 0, f_y|_{\mathbf{x}=\mathbf{x}_0} = 0$ 。根据式 D-1 和式 D-2，此时曲面上点 $(\mathbf{x}_0, f(\mathbf{x}_0))$ 处的主曲率 $\kappa_{\max}(\mathbf{x}_0, f(\mathbf{x}_0))$ 和 $\kappa_{\min}(\mathbf{x}_0, f(\mathbf{x}_0))$ 满足，

$$\begin{aligned} \kappa_{\max}(\mathbf{x}_0, f(\mathbf{x}_0)) + \kappa_{\min}(\mathbf{x}_0, f(\mathbf{x}_0)) &= (f_{xx} + f_{yy})|_{\mathbf{x}=\mathbf{x}_0} \\ \kappa_{\max}(\mathbf{x}_0, f(\mathbf{x}_0)) \cdot \kappa_{\min}(\mathbf{x}_0, f(\mathbf{x}_0)) &= (f_{xx} f_{yy} - f_{xy}^2)|_{\mathbf{x}=\mathbf{x}_0} \end{aligned} \quad (\text{D-3})$$

而同时我们知道函数 $f(\mathbf{x})$ 在点 \mathbf{x}_0 处的海森矩阵为，

$$H_0 = \begin{bmatrix} f_{xx} & f_{xy} \\ f_{xy} & f_{yy} \end{bmatrix}_{\mathbf{x}=\mathbf{x}_0} \quad (\text{D-4})$$

则显然，点 $(\mathbf{x}_0, f(\mathbf{x}_0))$ 处的两个主曲率之和便是 H_0 的迹，之积便是 H_0 的行列式，也就是说此时的两个主曲率实际上就是矩阵 H_0 的两个特征值。

E. 拉格朗日乘子法^[3]

拉格朗日乘子法解决的问题是：如何找到在等式约束下函数所有可能的极值点的问题。我们的目标是要找到函数 $f(\mathbf{x}) : \mathbb{R}^n \rightarrow \mathbb{R}$ 在 K 个等式约束 $\{g_k(\mathbf{x}) = 0\}_{k=1}^K$ 之下的极值点，其中 $f(\mathbf{x})$ 和 $\{g_k(\mathbf{x})\}_{k=1}^K$ 都存在连续的一阶偏导数。构造拉格朗日函数，

$$L(\mathbf{x}, \lambda_1, \lambda_2, \dots, \lambda_K) = f(\mathbf{x}) + \sum_{k=1}^K \lambda_k g_k(\mathbf{x}) \quad (\text{E-1})$$

如果 \mathbf{x}^* 是原问题的一个极值点，则必存在 $\lambda_1^*, \lambda_2^*, \dots, \lambda_K^*$ 使得 $(\mathbf{x}^*, \lambda_1^*, \lambda_2^*, \dots, \lambda_K^*)$ 为函数 $L(\mathbf{x}, \lambda_1, \lambda_2, \dots, \lambda_K)$ 的驻点，即 \mathbf{x}^* 是原问题的一个极值点的必要条件是：存在 $\lambda_1^*, \lambda_2^*, \dots, \lambda_K^*$ ，使得 $(\mathbf{x}^*, \lambda_1^*, \lambda_2^*, \dots, \lambda_K^*)$ 为函数 $L(\mathbf{x}, \lambda_1, \lambda_2, \dots, \lambda_K)$ 的驻点。但这个条件不是充分条件，也就是说，即使 $(\mathbf{x}^*, \lambda_1^*, \lambda_2^*, \dots, \lambda_K^*)$ 为函数 $L(\mathbf{x}, \lambda_1, \lambda_2, \dots, \lambda_K)$ 的驻点，但 \mathbf{x}^* 不一定是原问题的极值点， \mathbf{x}^* 到底是不是原问题的极值点还要根据原问题的具体特点进行分析。

这样，只要我们找到了拉格朗日函数的所有驻点，便可以找出原问题所有可能的极值点。具体来说，可以计算出 $L(\mathbf{x}, \lambda_1, \lambda_2, \dots, \lambda_K)$ 的驻点，即解如下方程组，

$$\begin{cases} \frac{\partial L}{\partial \mathbf{x}} = \mathbf{0} \\ \frac{\partial L}{\partial \lambda_1} = 0 \\ \vdots \\ \frac{\partial L}{\partial \lambda_K} = 0 \end{cases} \quad (\text{E-2})$$

假设 $(\mathbf{x}^*, \lambda_1^*, \lambda_2^*, \dots, \lambda_K^*)$ 满足方程组 E-2，即 $(\mathbf{x}^*, \lambda_1^*, \lambda_2^*, \dots, \lambda_K^*)$ 是 $L(\mathbf{x}, \lambda_1, \lambda_2, \dots, \lambda_K)$ 的驻点，则 \mathbf{x}^* 便是原问题一个可能的极值点。

F. 函数或自变量形式为矩阵或向量时的求导运算

F.1 向量和矩阵函数对标量变量求导

定义 F. 1 向量函数对标量自变量的导数。

设有 n 维向量函数 $\mathbf{f}(t) = (f_1(t), f_2(t), \dots, f_n(t))^T$ ，其中 $f_i(t) (i = 1, \dots, n)$ 为关于 t 的

可微函数，则 $\mathbf{f}(t)$ 对 t 的导数定义为： $\frac{d\mathbf{f}(t)}{dt} = \left[\frac{df_1(t)}{dt}, \frac{df_2(t)}{dt}, \dots, \frac{df_n(t)}{dt} \right]^T$

定义 F. 2 矩阵函数对标量自变量的导数。

设 有 $m \times n$ 维 矩 阵 函 数 $F(t) = \begin{bmatrix} f_{11}(t) & f_{12}(t) & \dots & f_{1n}(t) \\ f_{21}(t) & f_{22}(t) & \dots & f_{2n}(t) \\ \vdots & & & \\ f_{m1}(t) & f_{m2}(t) & \dots & f_{mn}(t) \end{bmatrix}_{m \times n}$ ， 其 中

$f_{ij}(t) (i=1, \dots, m, j=1, \dots, n)$ 为关于 t 的可微函数，则 $F(t)$ 关于自变量 t 的导数定义为，

$$\frac{dF(t)}{dt} = \begin{bmatrix} \frac{df_{11}(t)}{dt} \frac{df_{12}(t)}{dt}, \dots, \frac{df_{1n}(t)}{dt} \\ \frac{df_{21}(t)}{dt} \frac{df_{22}(t)}{dt}, \dots, \frac{df_{2n}(t)}{dt} \\ \vdots \\ \frac{df_{m1}(t)}{dt} \frac{df_{m2}(t)}{dt}, \dots, \frac{df_{mn}(t)}{dt} \end{bmatrix}_{m \times n}$$

F.2. 标量函数对矩阵变量求导

定义 F. 3 函数对矩阵自变量的导数。

设函数 $f(X) : \mathbb{R}^{m \times n} \rightarrow \mathbb{R}$ 把一个 $m \times n$ 的矩阵 X 映射成一个实数。定义函数 $f(X)$ 对矩

阵型自变量 $X = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ \vdots & & & \\ x_{m1} & x_{m2} & \dots & x_{mn} \end{bmatrix}$ 的导数为，

$$\frac{df(X)}{dX} = \begin{bmatrix} \frac{\partial f}{\partial x_{11}} \frac{\partial f}{\partial x_{12}}, \dots, \frac{\partial f}{\partial x_{1n}} \\ \vdots \\ \frac{\partial f}{\partial x_{m1}} \frac{\partial f}{\partial x_{m2}}, \dots, \frac{\partial f}{\partial x_{mn}} \end{bmatrix}$$

例 F.1：

假设 $A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$ ， 函数 $f(A) = \frac{3}{2}A_{11} + 5A_{12}^2 + A_{21}A_{22}$ ， 则

$$\frac{df(A)}{dA} = \begin{bmatrix} \frac{3}{2} & 10A_{12} \\ A_{22} & A_{21} \end{bmatrix}$$

F.3. 标量函数对向量变量求导

定义 F. 4 函数对向量自变量的导数。

设有函数 $f(\mathbf{x}): \mathbb{R}^n \rightarrow \mathbb{R}$ ，它是以列向量 $\mathbf{x} = (x_1, x_2, \dots, x_n)^T$ 为自变量的标量函数。

则函数 $f(\mathbf{x})$ 对列向量型自变量 \mathbf{x} 的导数为,

$$\frac{df(\mathbf{x})}{d\mathbf{x}} = \left[\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \dots, \frac{\partial f}{\partial x_n} \right]^T$$

上述定义形式就是多元函数 $f(\mathbf{x})$ 的梯度。可见，标量函数关于列向量型自变量的导数是一个列向量。类似地，我们也可以定义标量函数对行向量型自变量的导数，

$$\frac{df(\mathbf{x})}{d\mathbf{x}^T} = \left[\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \dots, \frac{\partial f}{\partial x_n} \right]$$

F.4. 向量函数对向量变量求导

定义 F. 5 向量函数对向量自变量的导数。

设 $\mathbf{x} \in \mathbb{R}^n$ ， $\mathbf{y}(\mathbf{x}) = [y_1(\mathbf{x}) y_2(\mathbf{x}) \cdots y_m(\mathbf{x})]^T \in \mathbb{R}^m$ ，其中

$y_i(\mathbf{x}): \mathbb{R}^n \rightarrow \mathbb{R} (i=1, 2, \dots, m)$ 是以 \mathbf{x} 为变量的 n 元可微函数。由于 $\mathbf{y}(\mathbf{x})$ 是列向量，可以定

义 $\mathbf{y}(\mathbf{x})$ 对于行向量 \mathbf{x}^T 的导数，

$$\frac{d\mathbf{y}(\mathbf{x})}{d\mathbf{x}^T} = \left[\begin{array}{c} \frac{\partial y_1(\mathbf{x})}{\partial x_1}, \frac{\partial y_1(\mathbf{x})}{\partial x_2}, \dots, \frac{\partial y_1(\mathbf{x})}{\partial x_n} \\ \frac{\partial y_2(\mathbf{x})}{\partial x_1}, \frac{\partial y_2(\mathbf{x})}{\partial x_2}, \dots, \frac{\partial y_2(\mathbf{x})}{\partial x_n} \\ \vdots \\ \frac{\partial y_m(\mathbf{x})}{\partial x_1}, \frac{\partial y_m(\mathbf{x})}{\partial x_2}, \dots, \frac{\partial y_m(\mathbf{x})}{\partial x_n} \end{array} \right]_{m \times n}$$

类似地，也可以定义 $\mathbf{y}^T(\mathbf{x})$ 对于列向量型自变量 \mathbf{x} 的导数，

$$\frac{d\mathbf{y}^T(\mathbf{x})}{d\mathbf{x}} = \begin{bmatrix} \frac{\partial y_1(\mathbf{x})}{\partial x_1}, \frac{\partial y_2(\mathbf{x})}{\partial x_1}, \dots, \frac{\partial y_m(\mathbf{x})}{\partial x_1} \\ \frac{\partial y_1(\mathbf{x})}{\partial x_2}, \frac{\partial y_2(\mathbf{x})}{\partial x_2}, \dots, \frac{\partial y_m(\mathbf{x})}{\partial x_2} \\ \vdots \\ \frac{\partial y_1(\mathbf{x})}{\partial x_n}, \frac{\partial y_2(\mathbf{x})}{\partial x_n}, \dots, \frac{\partial y_m(\mathbf{x})}{\partial x_n} \end{bmatrix}_{n \times m}$$

例 F.2: 设有列向量函数 $\mathbf{y}(\mathbf{x}) = \begin{bmatrix} y_1(\mathbf{x}) \\ y_2(\mathbf{x}) \end{bmatrix}$, 其中 $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$, $y_1(\mathbf{x}) = x_1^2 - x_2$, $y_2(\mathbf{x}) = x_3^2 + 3x_2$,

则行向量函数 $\mathbf{y}^T(\mathbf{x})$ 对列向量 \mathbf{x} 的导数为,

$$\frac{d\mathbf{y}^T(\mathbf{x})}{d\mathbf{x}} = \begin{bmatrix} \frac{\partial y_1(\mathbf{x})}{\partial x_1} & \frac{\partial y_2(\mathbf{x})}{\partial x_1} \\ \frac{\partial y_1(\mathbf{x})}{\partial x_2} & \frac{\partial y_2(\mathbf{x})}{\partial x_2} \\ \frac{\partial y_1(\mathbf{x})}{\partial x_3} & \frac{\partial y_2(\mathbf{x})}{\partial x_3} \end{bmatrix} = \begin{bmatrix} 2x_1 & 0 \\ -1 & 3 \\ 0 & 2x_3 \end{bmatrix}$$

F.5. 常用结论

1) 如 果 $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{y}(\mathbf{x}) = [y_1(\mathbf{x}) \ y_2(\mathbf{x}) \cdots y_m(\mathbf{x})]^T \in \mathbb{R}^m$, 其 中

$y_i(\mathbf{x}): \mathbb{R}^n \rightarrow \mathbb{R} (i=1, 2, \dots, m)$ 是以 \mathbf{x} 为变量的 n 元可微函数, 则有,

$$\frac{d\mathbf{y}^T(\mathbf{x})}{d\mathbf{x}} = \left(\frac{d\mathbf{y}(\mathbf{x})}{d\mathbf{x}^T} \right)^T.$$

证明:

可以直接由向量函数对向量型自变量的导数的定义得到。

2) $\mathbf{x}, \mathbf{a} \in \mathbb{R}^n$, \mathbf{x} 为自变量, \mathbf{a} 为常量, 则 $\frac{d(\mathbf{a}^T \mathbf{x})}{d\mathbf{x}} = \frac{d(\mathbf{x}^T \mathbf{a})}{d\mathbf{x}} = \mathbf{a}$ 。

证明:

$f(\mathbf{x}) = \mathbf{a}^T \mathbf{x}$ 为一个标量函数。设 $\mathbf{a} = (a_1, a_2, \dots, a_n)^T$, $\mathbf{x} = (x_1, x_2, \dots, x_n)^T$, 则

$f(\mathbf{x}) = a_1 x_1 + a_2 x_2 + \dots + a_n x_n$, 则有,

$$\frac{d(\mathbf{a}^T \mathbf{x})}{d\mathbf{x}} = \left(\frac{\partial(\mathbf{a}^T \mathbf{x})}{\partial x_1}, \frac{\partial(\mathbf{a}^T \mathbf{x})}{\partial x_2}, \dots, \frac{\partial(\mathbf{a}^T \mathbf{x})}{\partial x_n} \right)^T = (a_1, a_2, \dots, a_n)^T = \mathbf{a}$$

又由于 $\mathbf{a}^T \mathbf{x} = \mathbf{x}^T \mathbf{a}$, 因此 $\frac{d(\mathbf{x}^T \mathbf{a})}{d\mathbf{x}} = \frac{d(\mathbf{a}^T \mathbf{x})}{d\mathbf{x}} = \mathbf{a}$

3) $A \in \mathbb{R}^{m \times n}$ 为常数矩阵, $\mathbf{x} \in \mathbb{R}^n$ 为自变量, 则 $\frac{d(A\mathbf{x})}{d\mathbf{x}^T} = A$ 。

证明:

$$\text{设 } A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & & & \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}, \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \text{令 } \mathbf{y}(\mathbf{x}) = A\mathbf{x}, \text{则,}$$

$$\mathbf{y}(\mathbf{x}) = A\mathbf{x} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & & & \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n \\ \vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n \end{bmatrix} \triangleq \begin{bmatrix} y_1(\mathbf{x}) \\ y_2(\mathbf{x}) \\ \vdots \\ y_m(\mathbf{x}) \end{bmatrix}$$

$\mathbf{y}(\mathbf{x})$ 是 m 维列向量, 现在要求它对 n 维行向量 \mathbf{x}^T 的导数,

$$\frac{d\mathbf{y}(\mathbf{x})}{d\mathbf{x}^T} = \begin{bmatrix} \frac{\partial y_1(\mathbf{x})}{\partial x_1} \frac{\partial y_1(\mathbf{x})}{\partial x_2} \cdots \frac{\partial y_1(\mathbf{x})}{\partial x_n} \\ \frac{\partial y_2(\mathbf{x})}{\partial x_1} \frac{\partial y_2(\mathbf{x})}{\partial x_2} \cdots \frac{\partial y_2(\mathbf{x})}{\partial x_n} \\ \vdots \\ \frac{\partial y_m(\mathbf{x})}{\partial x_1} \frac{\partial y_m(\mathbf{x})}{\partial x_2} \cdots \frac{\partial y_m(\mathbf{x})}{\partial x_n} \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & & & \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} = A$$

4) $A \in \mathbb{R}^{m \times n}$ 为常数矩阵, $\mathbf{x} \in \mathbb{R}^n$ 为自变量, 则 $\frac{d(\mathbf{x}^T A^T)}{d\mathbf{x}} = A^T$ 。

证明:

$$\frac{d(\mathbf{x}^T A^T)}{d\mathbf{x}} = \frac{d(A\mathbf{x})^T}{d\mathbf{x}}, \text{根据本节结论 1) 和 3) 有,}$$

$$\frac{d(A\mathbf{x})^T}{d\mathbf{x}} = \left(\frac{d(A\mathbf{x})}{d\mathbf{x}^T} \right)^T = A^T$$

5) $A \in \mathbb{R}^{n \times n}$ 为常数矩阵, $\mathbf{x} \in \mathbb{R}^n$ 为自变量, 则 $\frac{d(\mathbf{x}^T A \mathbf{x})}{d\mathbf{x}} = (A + A^T)\mathbf{x}$ 。

证明:

$\mathbf{x}^T A \mathbf{x}$ 是关于列向量 \mathbf{x} 的标量函数。不失一般性，我们假定 A 是 3 阶方阵 $A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$ ，

\mathbf{x} 是 3 维列向量 $\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$ ，则有，

$$\mathbf{x}^T A \mathbf{x} = (x_1, x_2, x_3) \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$$

$$= x_1^2 a_{11} + x_1 x_2 a_{21} + x_1 x_3 a_{31} + x_1 x_2 a_{12} + x_2^2 a_{22} + x_2 x_3 a_{32} + x_1 x_3 a_{13} + x_2 x_3 a_{23} + x_3^2 a_{33}$$

则，

$$\begin{aligned} \frac{d(\mathbf{x}^T A \mathbf{x})}{d\mathbf{x}} &= \begin{bmatrix} \frac{\partial(\mathbf{x}^T A \mathbf{x})}{\partial x_1} \\ \frac{\partial(\mathbf{x}^T A \mathbf{x})}{\partial x_2} \\ \frac{\partial(\mathbf{x}^T A \mathbf{x})}{\partial x_3} \end{bmatrix} \\ &= \begin{bmatrix} 2a_{11}x_1 + (a_{12} + a_{21})x_2 + (a_{13} + a_{31})x_3 \\ (a_{21} + a_{12})x_1 + 2a_{22}x_2 + (a_{23} + a_{32})x_3 \\ (a_{31} + a_{13})x_1 + (a_{32} + a_{23})x_2 + 2a_{33}x_3 \end{bmatrix} \\ &= \begin{bmatrix} 2a_{11} & (a_{12} + a_{21}) & (a_{13} + a_{31}) \\ (a_{21} + a_{12}) & 2a_{22} & (a_{23} + a_{32}) \\ (a_{31} + a_{13}) & (a_{32} + a_{23}) & 2a_{33} \end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \\ &= \left(\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} + \begin{bmatrix} a_{11} & a_{21} & a_{31} \\ a_{12} & a_{22} & a_{32} \\ a_{13} & a_{23} & a_{33} \end{bmatrix} \right) \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \\ &= (A + A^T) \mathbf{x} \end{aligned}$$

6) $X \in \mathbb{R}^{m \times n}$ 为自变量矩阵， $\mathbf{a} \in \mathbb{R}^{m \times 1}$ 、 $\mathbf{b} \in \mathbb{R}^{n \times 1}$ 为常数列向量，则 $\frac{d(\mathbf{a}^T X \mathbf{b})}{dX} = \mathbf{a} \mathbf{b}^T$ 。

证明：

设 $\mathbf{a} = (a_1, a_2, \dots, a_m)^T$, $\mathbf{b} = (b_1, b_2, \dots, b_n)^T$, $X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & & & \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix}$

则,

$$\begin{aligned} \mathbf{a}^T X \mathbf{b} &= (a_1, a_2, \dots, a_m) \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & & & \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix} \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix} \\ &= \left(\sum_{i=1}^m a_i x_{i1}, \sum_{i=1}^m a_i x_{i2}, \dots, \sum_{i=1}^m a_i x_{in} \right) \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix} \\ &= b_1 \sum_{i=1}^m a_i x_{i1} + b_2 \sum_{i=1}^m a_i x_{i2} + \dots + b_n \sum_{i=1}^m a_i x_{in} \end{aligned}$$

则,

$$\begin{aligned} \frac{d(\mathbf{a}^T X \mathbf{b})}{dX} &= \begin{bmatrix} \frac{\partial(\mathbf{a}^T X \mathbf{b})}{\partial x_{11}} \frac{\partial(\mathbf{a}^T X \mathbf{b})}{\partial x_{12}} \cdots \frac{\partial(\mathbf{a}^T X \mathbf{b})}{\partial x_{1n}} \\ \frac{\partial(\mathbf{a}^T X \mathbf{b})}{\partial x_{21}} \frac{\partial(\mathbf{a}^T X \mathbf{b})}{\partial x_{22}} \cdots \frac{\partial(\mathbf{a}^T X \mathbf{b})}{\partial x_{2n}} \\ \vdots \\ \frac{\partial(\mathbf{a}^T X \mathbf{b})}{\partial x_{m1}} \frac{\partial(\mathbf{a}^T X \mathbf{b})}{\partial x_{m2}} \cdots \frac{\partial(\mathbf{a}^T X \mathbf{b})}{\partial x_{mn}} \end{bmatrix} \\ &= \begin{bmatrix} b_1 a_1 & b_2 a_1 \cdots b_n a_1 \\ b_1 a_2 & b_2 a_2 \cdots b_n a_2 \\ \vdots & \\ b_1 a_m & b_2 a_m \cdots b_n a_m \end{bmatrix} \\ &= \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{pmatrix} (b_1 b_2 \cdots b_n) \\ &= \mathbf{a} \mathbf{b}^T \end{aligned}$$

7) $X \in \mathbb{R}^{n \times m}$ 为自变量矩阵, $\mathbf{a} \in \mathbb{R}^{m \times 1}$ 、 $\mathbf{b} \in \mathbb{R}^{n \times 1}$ 为常数列向量, 则 $\frac{d\mathbf{a}^T X^T \mathbf{b}}{dX} = \mathbf{b} \mathbf{a}^T$ 。

证明:

$$\frac{d\mathbf{a}^T X^T \mathbf{b}}{dX} = \frac{d(\mathbf{b}^T X \mathbf{a})^T}{dX} = \frac{d(\mathbf{b}^T X \mathbf{a})}{dX} = \mathbf{b} \mathbf{a}^T$$

8) $\mathbf{X} \in \mathbb{R}^n$ 为自变量, 则 $\frac{d\mathbf{x}^T \mathbf{x}}{d\mathbf{x}} = 2\mathbf{x}$ 。

证明:

设 $\mathbf{x} = (x_1, x_2, \dots, x_n)^T$, 则 $\mathbf{x}^T \mathbf{x} = x_1^2 + x_2^2 + \dots + x_n^2$, 则,

$$\frac{d(\mathbf{x}^T \mathbf{x})}{d\mathbf{x}} = \begin{bmatrix} \frac{\partial(\mathbf{x}^T \mathbf{x})}{\partial x_1} \\ \frac{\partial(\mathbf{x}^T \mathbf{x})}{\partial x_2} \\ \vdots \\ \frac{\partial(\mathbf{x}^T \mathbf{x})}{\partial x_n} \end{bmatrix} = \begin{bmatrix} 2x_1 \\ 2x_2 \\ \vdots \\ 2x_n \end{bmatrix} = 2\mathbf{x}$$

9) $X \in \mathbb{R}^{m \times n}$ 为自变量矩阵, $B \in \mathbb{R}^{n \times m}$ 为常数矩阵, 则 $\frac{d(\text{tr}(XB))}{dX} = B^T$, 其中 $\text{tr}(\cdot)$ 表

示计算矩阵的迹。

证明:

设 $X = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ \vdots & & & \\ x_{m1} & x_{m2} & \dots & x_{mn} \end{bmatrix}$, $B = \begin{bmatrix} b_{11} & b_{12} & \dots & b_{1m} \\ b_{21} & b_{22} & \dots & b_{2m} \\ \vdots & & & \\ b_{n1} & b_{n2} & \dots & b_{nm} \end{bmatrix}$, 则有,

$$XB = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ \vdots & & & \\ x_{m1} & x_{m2} & \dots & x_{mn} \end{bmatrix} \begin{bmatrix} b_{11} & b_{12} & \dots & b_{1m} \\ b_{21} & b_{22} & \dots & b_{2m} \\ \vdots & & & \\ b_{n1} & b_{n2} & \dots & b_{nm} \end{bmatrix}, \text{ 则,}$$

$$\text{tr}(XB) = \sum_{j=1}^n x_{1j} b_{j1} + \sum_{j=1}^n x_{2j} b_{j2} + \dots + \sum_{j=1}^n x_{mj} b_{jm}, \text{ 则,}$$

$$\frac{d(\text{tr}(XB))}{dX} = \begin{bmatrix} \frac{\partial(\text{tr}(XB))}{\partial x_{11}} \frac{\partial(\text{tr}(XB))}{\partial x_{12}}, \dots, \frac{\partial(\text{tr}(XB))}{\partial x_{1n}} \\ \frac{\partial(\text{tr}(XB))}{\partial x_{21}} \frac{\partial(\text{tr}(XB))}{\partial x_{22}}, \dots, \frac{\partial(\text{tr}(XB))}{\partial x_{2n}} \\ \vdots \\ \frac{\partial(\text{tr}(XB))}{\partial x_{m1}} \frac{\partial(\text{tr}(XB))}{\partial x_{m2}}, \dots, \frac{\partial(\text{tr}(XB))}{\partial x_{mn}} \end{bmatrix} = \begin{bmatrix} b_{11} & b_{21}, \dots, b_{n1} \\ b_{12} & b_{22}, \dots, b_{n2} \\ \vdots \\ b_{1m} & b_{2m}, \dots, b_{nm} \end{bmatrix} = B^T$$

10) $X \in \mathbb{R}^{n \times n}$ 为自变量矩阵, $f(X): \mathbb{R}^{m \times n} \rightarrow \mathbb{R}$ 为关于 X 中元素可微的函数, 则

$$\frac{df(X)}{dX^T} = \left(\frac{df(X)}{dX} \right)^T.$$

证明:

根据标量函数对矩阵型自变量导数的定义容易验证。

11) $X \in \mathbb{R}^{n \times n}$ 为非奇异 n 阶方阵, $|X|$ 表示 X 的行列式, 则 $\frac{d|X|}{dX} = |X| (X^{-1})^T$ 。

证明:

设 $X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & & & \\ x_{n1} & x_{n2} & \cdots & x_{nn} \end{bmatrix}$, 它的伴随矩阵为 $X^* = \begin{bmatrix} x_{11}^* & x_{12}^* & \cdots & x_{1n}^* \\ x_{21}^* & x_{22}^* & \cdots & x_{2n}^* \\ \vdots & & & \\ x_{n1}^* & x_{n2}^* & \cdots & x_{nn}^* \end{bmatrix}$ 。由伴随矩阵的性质^[3]可知

$X^* = |X| X^{-1}$ 。另外, 我们知道, $|X| = \sum_{j=1}^n x_{ij} x_{ji}^*$, 而且 X 中第 i 行的任何元素与 X^* 中第 i

列的任何元素都没有关系。则,

$$\frac{d|X|}{dX} = \begin{bmatrix} \frac{\partial|X|}{\partial x_{11}} & \frac{\partial|X|}{\partial x_{12}} & \cdots & \frac{\partial|X|}{\partial x_{1n}} \\ \frac{\partial|X|}{\partial x_{21}} & \frac{\partial|X|}{\partial x_{22}} & \cdots & \frac{\partial|X|}{\partial x_{2n}} \\ \vdots & & & \\ \frac{\partial|X|}{\partial x_{n1}} & \frac{\partial|X|}{\partial x_{n2}} & \cdots & \frac{\partial|X|}{\partial x_{nn}} \end{bmatrix} = \begin{bmatrix} x_{11}^* & x_{21}^* & \cdots & x_{n1}^* \\ x_{12}^* & x_{22}^* & \cdots & x_{n2}^* \\ \vdots & & & \\ x_{1n}^* & x_{2n}^* & \cdots & x_{nn}^* \end{bmatrix} = (X^*)^T = (|X| X^{-1})^T = |X| X^{-T}$$

G. 奇异值分解

G.1. 奇异值分解定理

定理 G. 1 奇异值分解定理

设有矩阵 $A_{m \times n}$, 其秩为 $\text{rank}(A) = r$, 则其可以分解为如下形式,

$$A_{m \times n} = U_{m \times m} \sum_{m \times n} V_{n \times n}^T \quad (\text{G-1})$$

式 G-1 称为矩阵 A 的奇异值分解。其中, $U = [\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_m]$, $V = [\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n]$ 为正交矩阵,

分别称为 A 的左奇异矩阵和右奇异矩阵； $\Sigma_{m \times n} = \begin{bmatrix} \sum_r & O_{r \times (n-r)} \\ O_{(m-r) \times r} & O_{(m-r) \times (n-r)} \end{bmatrix}_{m \times n}$ ，

$\Sigma_r = \text{diag}(\sigma_1, \sigma_2, \dots, \sigma_r)$ 为对角矩阵， $\sigma_1, \sigma_2, \dots, \sigma_r > 0$ 称为矩阵 A 的奇异值。

需要注意的是，矩阵 A 的奇异值必为正数，而且奇异值集合是唯一的，但 U 和 V 并不唯一。如果将 $\sigma_1, \sigma_2, \dots, \sigma_r$ 按照从大小顺序排列，则 $\Sigma_{m \times n}$ 便是被 A 唯一确定的。

G.2. 奇异值分解的经济型 (economy-sized) 表达形式

设有矩阵 $A_{m \times n}$ ，其奇异值分解形式如式 G-1 所示，则

$$\begin{aligned} A_{m \times n} &= U_{m \times m} \Sigma_{m \times n} V_{n \times n}^T \\ &= [\mathbf{u}_1, \dots, \mathbf{u}_r | \mathbf{u}_{r+1}, \dots, \mathbf{u}_m] \begin{bmatrix} \sigma_1 & & & \\ & \sigma_2 & & \\ & & \ddots & \\ & & & \sigma_r \end{bmatrix} O_{r \times (n-r)} \begin{bmatrix} \mathbf{v}_1^T \\ \mathbf{v}_2^T \\ \vdots \\ \mathbf{v}_r^T \\ \hline \mathbf{v}_{r+1}^T \\ \vdots \\ \mathbf{v}_n^T \end{bmatrix} \\ &= [\mathbf{u}_1, \dots, \mathbf{u}_r] \begin{bmatrix} \sigma_1 & & & \\ & \sigma_2 & & \\ & & \ddots & \\ & & & \sigma_r \end{bmatrix} \begin{bmatrix} \mathbf{v}_1^T \\ \mathbf{v}_2^T \\ \vdots \\ \mathbf{v}_r^T \end{bmatrix} \end{aligned} \quad (\text{G-2})$$

式 G-2 这种 SVD 表达形式称为 SVD 的经济型表达形式。

G.3. 奇异值分解的矩阵和表达形式

矩阵乘法的外积 (outer product) 表达。如果 X 是 $m \times k$ 矩阵， \mathbf{x}_i 是它的列； Y 是 $k \times n$ 矩阵，它的行为 \mathbf{y}_i^T ，则

$$XY = \sum_{i=1}^k \mathbf{x}_i \mathbf{y}_i^T \quad (\text{G-3})$$

并且容易证明，每一个子矩阵 $\mathbf{x}_i \mathbf{y}_i^T$ 的秩都为 1。

假设有矩阵 $A_{m \times n}$ ，其经济型奇异值分解形式如式 G-2 所示，则通过矩阵的外积分解，

我们可以得到 $A_{m \times n}$ 的奇异值分解的矩阵和形式。

$$\text{令 } X = [\mathbf{u}_1, \dots, \mathbf{u}_r] \begin{bmatrix} \sigma_1 & & \\ & \sigma_2 & \\ & & \ddots \\ & & & \sigma_r \end{bmatrix} = [\sigma_1 \mathbf{u}_1, \sigma_2 \mathbf{u}_2, \dots, \sigma_r \mathbf{u}_r], \text{ 令 } Y = \begin{bmatrix} \mathbf{v}_1^T \\ \mathbf{v}_2^T \\ \vdots \\ \mathbf{v}_r^T \end{bmatrix} \text{。由 G-3 可知,}$$

$$A = XY = [\sigma_1 \mathbf{u}_1, \sigma_2 \mathbf{u}_2, \dots, \sigma_r \mathbf{u}_r] \begin{bmatrix} \mathbf{v}_1^T \\ \mathbf{v}_2^T \\ \vdots \\ \mathbf{v}_r^T \end{bmatrix} = \sum_{i=1}^r \sigma_i \mathbf{u}_i \mathbf{v}_i^T \quad (\text{G-4})$$

式 G-4 称为矩阵 A 的奇异值分解的矩阵和的形式。

G.4. 奇异值分解与特征值分解之间的联系

奇异值分解与特征值分解之间有着密切的联系，具体内容总结在了命题 G.6~G.8 之中。为了得到这 3 个命题，首先需要铺垫一些预备命题 G.1~G.5。

命题 G. 1: 如果 $r(A_{m \times n}) = r$ ，则 $r(A^T A) = r(A A^T) = r$ 。

证明：只需要证明 $A^T A \mathbf{x} = \mathbf{0}$ 与 $A \mathbf{x} = \mathbf{0}$ 同解即可。

如果 \mathbf{x} 为 $A \mathbf{x} = \mathbf{0}$ 的解，即 $A \mathbf{x} = \mathbf{0}$ ，显然必有 $A^T A \mathbf{x} = \mathbf{0}$ ，也即 \mathbf{x} 为 $A^T A \mathbf{x} = \mathbf{0}$ 的解。如果 \mathbf{x} 为 $A^T A \mathbf{x} = \mathbf{0}$ 的解，则有 $\mathbf{x}^T A^T A \mathbf{x} = 0$ ，也即 $(A \mathbf{x})^T A \mathbf{x} = 0$ ，则向量 $A \mathbf{x}$ 的长度为 0；而向量 $A \mathbf{x}$ 的长度为零的充要条件是 $A \mathbf{x} = \mathbf{0}$ ，即 \mathbf{x} 为 $A \mathbf{x} = \mathbf{0}$ 的解。这样就证明了 $A^T A \mathbf{x} = \mathbf{0}$ 与 $A \mathbf{x} = \mathbf{0}$ 同解，因此， $r(A^T A) = r(A) = r$ 。同理可以证明 $r(A A^T) = r(A) = r$ 。

命题 G. 2: 设 A 为 $m \times n$ 矩阵，则 $A^T A$ 与 $A A^T$ 有相同的非零特征值。

证明：设 λ 为 $A^T A$ 的特征值，则有，

$$A^T A \mathbf{a} = \lambda \mathbf{a} \quad (\text{G-5})$$

其中 \mathbf{a} 为矩阵 $A^T A$ 对应于特征值 λ 的特征向量。将式 G-5 左右同时乘以 A ，得到，

$$A A^T (A \mathbf{a}) = \lambda (A \mathbf{a}) \quad (\text{G-6})$$

则可知 λ 也为矩阵 $A A^T$ 的特征值，对应的特征向量为 $A \mathbf{a}$ 。

命题 G. 3: 如果 A 为 n 阶实对称矩阵且 $\text{rank}(A) = r$ ，则 A 必有且仅有 r 个不为零的特征值。

证明：由于 A 为实对称矩阵，则它必可以（正交）相似于对角矩阵^[3]，

$$A = P \Sigma P^{-1} \quad (\text{G-7})$$

其中， $\Sigma = \begin{bmatrix} \lambda_1 & & \\ & \lambda_2 & \\ & & \ddots \\ & & & \lambda_n \end{bmatrix}$ ， $\lambda_i (i=1, \dots, n)$ 为矩阵 A 的特征值， P 、 P^{-1} 均为正交矩阵。由于

P 、 P^{-1} 均为可逆矩阵，因此它们不会改变矩阵 Σ 的秩，即 $\text{rank}(P \Sigma P^{-1}) = \text{rank}(\Sigma)$ 。

因此， $\text{rank}(\Sigma) = \text{rank}(P \Sigma P^{-1}) = \text{rank}(A) = r$ 。显然，由于 Σ 为对角矩阵，若 $\text{rank}(\Sigma) = r$ ，则 Σ 必有且仅有 r 个不为零的对角元，即 A 有且仅有 r 个不为零的特征值。

命题 G. 4: A 为 $m \times n$ 实矩阵，则 $A^T A$ 与 AA^T 均为半正定矩阵。

证明： $\forall \mathbf{x} \in \mathbb{R}^{n \times 1} \neq \mathbf{0}$ ， $0 \leq (A\mathbf{x})^T (A\mathbf{x}) = \mathbf{x}^T A^T A \mathbf{x}$ ，则 $A^T A$ 为半正定矩阵。 $\forall \mathbf{x} \in \mathbb{R}^{m \times 1} \neq \mathbf{0}$ ，

$0 \leq (A^T \mathbf{x})^T (A^T \mathbf{x}) = \mathbf{x}^T A A^T \mathbf{x}$ ，则 AA^T 为半正定矩阵。

命题 G. 5: A 为 $m \times n$ 实矩阵，且 $\text{rank}(A) = r$ ，则 $(A^T A)_{n \times n}$ 和 $(AA^T)_{m \times m}$ 有且仅有 r 个大于零的特征值，且其他特征值均为零。

证明：由于 $\text{rank}(A) = r$ ，则根据命题 G.1 可知， $\text{rank}(A^T A) = r$ 。容易知道， $A^T A$ 为实对称矩阵，则 $A^T A$ 为秩为 r 的实对称矩阵。由命题 G.3 可知， $A^T A$ 必有且仅有 r 个不为零的特征值（其余特征值均为零）。由命题 G.4 可知， $A^T A$ 是个半正定矩阵，因此它的特征值全部为非负数。这样，由于已经知道了 $A^T A$ 有且仅有 r 个不为零的特征值，则这 r 个不为零的特征值必然都大于零。又由命题 G.2 可知， AA^T 与 $A^T A$ 有相同的非零特征值，因此， AA^T 也是有且仅有 r 个大于零的特征值，且其他特征值均为零。

证毕。

命题 G. 6: 有实矩阵 $A_{m \times n}$ ， $\text{rank}(A_{m \times n}) = r$ ，则它的 r 个奇异值 $\{\sigma_i\}_{i=1}^r$ 是 $(A^T A)_{n \times n}$ 的对应的特征值 $\{\lambda_i\}_{i=1}^r$ 的平方根，即 $\sigma_i = \sqrt{\lambda_i} (i=1, \dots, r)$ ； A 的奇异值分解的右奇异矩阵 V 就是 $(A^T A)_{n \times n}$ 进行正交特征值分解时产生的正交矩阵。

证明：

由命题 G.5 可知， $(A^T A)_{n \times n}$ 有且仅有 r 个正的特征值，且其他特征值均为零。由于 $(A^T A)_{n \times n}$ 是实对称矩阵，它必然可以进行正交特征值分解，

$$A^T A = V \begin{bmatrix} \begin{bmatrix} \lambda_1 & & \\ & \lambda_2 & \\ & & \ddots \\ & & & \lambda_r \end{bmatrix} & O_{r \times (n-r)} \\ O_{(n-r) \times r} & O_{(n-r) \times (n-r)} \end{bmatrix} V^T \quad (\text{G-8})$$

其中, V 为正交矩阵, $\lambda_1, \lambda_2, \dots, \lambda_r$ 为 $A^T A$ 的 r 个正特征值, 且我们要求 $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_r > 0$ 。

需要注意的是, 满足条件的正交矩阵 V 并不是唯一的。

我们再从奇异值分解的结果看看 $A^T A$ 是什么样子的。如果 A 的奇异值分解形式为式 G-1, 且我们要求对角矩阵中的奇异值按照由大到小的顺序排列, 即 $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r > 0$, 则,

$$\begin{aligned} A^T A &= (U \Sigma V^T)^T U \Sigma V^T \\ &= V \Sigma^T U^T U \Sigma V^T \\ &= V (\Sigma^T \Sigma)_{n \times n} V^T \\ &= V \begin{bmatrix} \begin{bmatrix} \sigma_1^2 & & \\ & \sigma_2^2 & \\ & & \ddots \\ & & & \sigma_r^2 \end{bmatrix} & O_{r \times (n-r)} \\ O_{(n-r) \times r} & O_{(n-r) \times (n-r)} \end{bmatrix} V^T \end{aligned} \quad (\text{G-9})$$

显然, 式 G-9 也是矩阵 $(A^T A)_{n \times n}$ 的某个具体正交特征值分解形式。由于当对角元按序排列时, 矩阵相似对角化之后的对角矩阵具有唯一性, 因此必有,

$$\begin{bmatrix} \begin{bmatrix} \lambda_1 & & \\ & \lambda_2 & \\ & & \ddots \\ & & & \lambda_r \end{bmatrix} & O_{r \times (n-r)} \\ O_{(n-r) \times r} & O_{(n-r) \times (n-r)} \end{bmatrix} = \begin{bmatrix} \begin{bmatrix} \sigma_1^2 & & \\ & \sigma_2^2 & \\ & & \ddots \\ & & & \sigma_r^2 \end{bmatrix} & O_{r \times (n-r)} \\ O_{(n-r) \times r} & O_{(n-r) \times (n-r)} \end{bmatrix} \quad (\text{G-10})$$

即 $\sigma_i = \sqrt{\lambda_i}, 1 \leq i \leq r$ 。同时也可以知道, V 和 V' 可以互换, 因此, V 也是 $A^T A$ 正交特征值分解所产生的正交矩阵。

证毕。

基于与上面完全类似的推导过程, 也可以得出矩阵 $A_{m \times n}$ 的奇异值分解与 $(AA^T)_{m \times m}$ 的特征值分解的关系。

命题 G.7: $(AA^T)_{m \times m}$ 实际上与 $(A^T A)_{n \times n}$ 具有相同的非零特征值 (由命题 G.2 得到),

$\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_r > 0$, 则 A 的 r 个奇异值 $\{\sigma_i\}_{i=1}^r$ 也是 $(AA^T)_{m \times m}$ 的对应的特征值 $\{\lambda_i\}_{i=1}^r$ 的平方根, 即 $\sigma_i = \sqrt{\lambda_i}$ 。另外, $(AA^T)_{m \times m} = (U\Sigma V)(U\Sigma V)^T = U\Sigma VV^T\Sigma^T U^T = U\Sigma\Sigma^T U^T$ 。

因此, A 的左奇异矩阵 U 也就是 $(AA^T)_{m \times m}$ 进行正交特征值分解产生的正交矩阵 (虽然不具有唯一性)。

命题 G.8: 如果矩阵 $A_{n \times n}$ 是半正定矩阵, 则它的正交特征值分解与它的奇异值分解是相同的。也就是说, 如果某种分解形式是 A 的一个正交特征值分解, 则它也是 A 的一个奇异值分解。

证明:

先证明半正定矩阵 A 的奇异值就是 A 的特征值。假设 $\text{rank}(A)=r$, 由于 A 是半正定矩阵 (当然也是实对称矩阵), 结合命题 G.3 和 G.5 可知, A 有且仅有 r 个正特征值 $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_r > 0$, 其它特征值均为零。因此, A 可正交相似对角分解为,

$$A = U \begin{bmatrix} \begin{bmatrix} \lambda_1 & & \\ & \lambda_2 & \\ & & \ddots & \\ & & & \lambda_r \end{bmatrix} & O \\ O & O \end{bmatrix} U^T, \text{ 其中 } U \text{ 为正交矩阵。则,}$$

$$AA^T = U \begin{bmatrix} \begin{bmatrix} \lambda_1^2 & & \\ & \lambda_2^2 & \\ & & \ddots & \\ & & & \lambda_r^2 \end{bmatrix} & O \\ O & O \end{bmatrix} U^T \quad (\text{G-11})$$

根据式 G-11, 由命题 G.6 可知, A 的奇异值就是 $\sqrt{\lambda_1^2}, \dots, \sqrt{\lambda_r^2}$, 即为 $\lambda_1, \dots, \lambda_r$ 。

再来证明 A 的左奇异矩阵就是 U 。这可以由命题 G.7 直接得到。

最后证明 A 的右奇异矩阵也是 U 。由于 A 是实对称矩阵, 因此,

$$A^T A = AA^T = U \begin{bmatrix} \begin{bmatrix} \lambda_1^2 & & \\ & \lambda_2^2 & \\ & & \ddots & \\ & & & \lambda_r^2 \end{bmatrix} & O \\ O & O \end{bmatrix} U^T \quad (\text{G-12})$$

由式 G-12, 再根据命题 G.6 可知, A 的右奇异矩阵就是 $A^T A$ 进行正交特征值分解所产生的

正交矩阵，即为 U 。

证毕。

H. 函数的极值点、驻点和鞍点

函数的极值点、驻点和鞍点是一些在函数的定义域空间具有特殊意义的点，在分析函数性质的时候会经常遇到它们。

定义 H. 1 局部极小值点 (local minimizer)。函数 $f(\mathbf{x}): \mathbb{R}^n \rightarrow \mathbb{R}$ 为连续函数，如果存在 $\delta > 0$ ，使得 $\forall \mathbf{x}, \|\mathbf{x} - \mathbf{x}^*\| < \delta$ 有 $f(\mathbf{x}) \geq f(\mathbf{x}^*)$ 成立，则 \mathbf{x}^* 称为 $f(\mathbf{x})$ 的一个局部极小值点^[4]。

与定义 H.1 类似，我们也可以定义局部极大值点。

定义 H. 2 驻点 (stationary point)。函数 $f(\mathbf{x}): \mathbb{R}^n \rightarrow \mathbb{R}$ 连续可微，若在点 \mathbf{x}_s 处有 $\nabla f(\mathbf{x})|_{\mathbf{x}=\mathbf{x}_s} = \mathbf{0}$ ，则称 \mathbf{x}_s 为 $f(\mathbf{x})$ 的一个驻点。

下面我们给出一个点为函数极小值点的必要条件：

定理 H. 1 函数 $f(\mathbf{x}): \mathbb{R}^n \rightarrow \mathbb{R}$ 连续且有二阶偏导数，如果点 \mathbf{x}^* 为函数 $f(\mathbf{x})$ 的极小值点，则必然有 $\nabla f(\mathbf{x}^*) = \mathbf{0}$ 且 $\nabla^2 f(\mathbf{x}^*)$ 为半正定矩阵，即 \mathbf{x}^* 为函数 $f(\mathbf{x})$ 的驻点且 $f(\mathbf{x})$ 在该点处的海森矩阵为半正定矩阵。

证明：

对于可微函数，它的极值点必然也是驻点，这个结论在大多数高等数学教材中都会有，我们就不再给出证明了。我们只证明后半部分，即极小值点处的海森矩阵为半正定矩阵。把 $f(\mathbf{x})$ 在 \mathbf{x}^* 近旁进行泰勒展开得到，

$$f(\mathbf{x}^* + \mathbf{h}) = f(\mathbf{x}^*) + \mathbf{h}^T \nabla f(\mathbf{x}^*) + \frac{1}{2} \mathbf{h}^T \nabla^2 f(\mathbf{x}^*) \mathbf{h} + O(\|\mathbf{h}\|^2) \quad (\text{H-1})$$

其中， $\nabla f(\mathbf{x}^*)$ 表示 $f(\mathbf{x})$ 在点 \mathbf{x}^* 处的梯度、 $\nabla^2 f(\mathbf{x}^*)$ 表示 $f(\mathbf{x})$ 在点 \mathbf{x}^* 处的海森矩阵。由于 \mathbf{x}^* 为函数 $f(\mathbf{x})$ 的驻点，因此 $\mathbf{h}^T \nabla f(\mathbf{x}^*) = 0$ 。又因为 \mathbf{x}^* 为函数 $f(\mathbf{x})$ 的极小值点，因此当 $\|\mathbf{h}\|$ 足够小时，必有 $f(\mathbf{x}^* + \mathbf{h}) - f(\mathbf{x}^*) \geq 0$ 。这样我们便有， $\frac{1}{2} \mathbf{h}^T \nabla^2 f(\mathbf{x}^*) \mathbf{h} + O(\|\mathbf{h}\|^2) \geq 0$ 。由于 $O(\|\mathbf{h}\|^2)$ 是 $\|\mathbf{h}\|^2$ 的高阶无穷小，若 $\frac{1}{2} \mathbf{h}^T \nabla^2 f(\mathbf{x}^*) \mathbf{h} < 0$ ，则必有 $\frac{1}{2} \mathbf{h}^T \nabla^2 f(\mathbf{x}^*) \mathbf{h} + O(\|\mathbf{h}\|^2) < 0$ ，与 $\frac{1}{2} \mathbf{h}^T \nabla^2 f(\mathbf{x}^*) \mathbf{h} + O(\|\mathbf{h}\|^2) \geq 0$ 矛盾。因此必有 $\frac{1}{2} \mathbf{h}^T \nabla^2 f(\mathbf{x}^*) \mathbf{h} \geq 0$ ，即 $\nabla^2 f(\mathbf{x}^*)$ 为半正定矩阵。

类似地，我们也可以得到一个点为函数极大值点的必要条件：

定理 H. 2 函数 $f(\mathbf{x}): \mathbb{R}^n \rightarrow \mathbb{R}$ 连续且有二阶偏导数，如果点 \mathbf{x}^* 为函数 $f(\mathbf{x})$ 的极大值点，

则必然有 $\nabla f(\mathbf{x}^*) = \mathbf{0}$ 且 $\nabla^2 f(\mathbf{x}^*)$ 为半负定矩阵，即 \mathbf{x}^* 为函数 $f(\mathbf{x})$ 的驻点且 $f(\mathbf{x})$ 在该点处的海森矩阵为半负定矩阵。

接下来，我们给出一个点为函数极小值点的一个充分条件：

定理 H.3 函数 $f(\mathbf{x}): \mathbb{R}^n \rightarrow \mathbb{R}$ 连续且有二阶偏导数，若点 \mathbf{x}^* 为 $f(\mathbf{x})$ 的驻点且 $f(\mathbf{x})$ 在 \mathbf{x}^* 处的海森矩阵为正定矩阵，则 \mathbf{x}^* 为 $f(\mathbf{x})$ 的一个局部极小值点。

证明：

$f(\mathbf{x})$ 的泰勒展开形式为，

$$f(\mathbf{x}^* + \mathbf{h}) = f(\mathbf{x}^*) + \mathbf{h}^T \nabla f(\mathbf{x}^*) + \mathbf{h}^T \nabla^2 f(\mathbf{x}^*) \mathbf{h} + O(\|\mathbf{h}\|^2) \quad (\text{H-2})$$

若 \mathbf{x}^* 为驻点，则 $\mathbf{h}^T \nabla f(\mathbf{x}^*) = 0$ ，因此有，

$$f(\mathbf{x}^* + \mathbf{h}) = f(\mathbf{x}^*) + \mathbf{h}^T \nabla^2 f(\mathbf{x}^*) \mathbf{h} + O(\|\mathbf{h}\|^2) \quad (\text{H-3})$$

由于 $\nabla^2 f(\mathbf{x}^*)$ 为正定矩阵，则它的特征值一定全部大于某个数 $\delta > 0$ 。因此有，

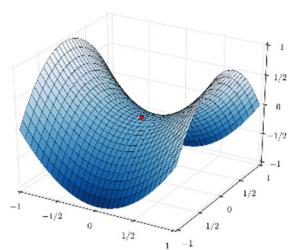
$\mathbf{h}^T \nabla^2 f(\mathbf{x}^*) \mathbf{h} > \delta \|\mathbf{h}\|^2$ 。这样，当 $\|\mathbf{h}\|$ 足够小的时候， $\mathbf{h}^T \nabla^2 f(\mathbf{x}^*) \mathbf{h} + O(\|\mathbf{h}\|^2)$ 的符号完全由 $\mathbf{h}^T \nabla^2 f(\mathbf{x}^*) \mathbf{h}$ 决定，则 $\mathbf{h}^T \nabla^2 f(\mathbf{x}^*) \mathbf{h} + O(\|\mathbf{h}\|^2) > 0$ 。因此 $f(\mathbf{x}^* + \mathbf{h}) > f(\mathbf{x}^*)$ ，即 \mathbf{x}^* 为 $f(\mathbf{x})$ 的一个局部极小值点。

类似地，我们也可以给出一个点为函数极大值点的一个充分条件：

定理 H.4 函数 $f(\mathbf{x}): \mathbb{R}^n \rightarrow \mathbb{R}$ 连续且有二阶偏导数，若点 \mathbf{x}^* 为 $f(\mathbf{x})$ 的驻点且 $f(\mathbf{x})$ 在 \mathbf{x}^* 处的海森矩阵为负定矩阵，则 \mathbf{x}^* 为 $f(\mathbf{x})$ 的一个局部极大值点。

结合定理 H.1~4 可以知道，点 \mathbf{x}^* 为 $f(\mathbf{x})$ 的驻点只是该点为 $f(\mathbf{x})$ 局部极值点的必要条件，也就是说，若点 \mathbf{x}^* 为 $f(\mathbf{x})$ 的驻点，则该点可能是 $f(\mathbf{x})$ 的极大值点，也可能是 $f(\mathbf{x})$ 的极小值点，也可能它根本就不是一个极值点。不是极值点的驻点有一个独特的名字，称为鞍点 (saddle point)。下面我们给出鞍点的正式定义：

定义 H.3 鞍点 (saddle point)。函数 $f(\mathbf{x}): \mathbb{R}^n \rightarrow \mathbb{R}$ 连续可微，若 \mathbf{x}_s 为 $f(\mathbf{x})$ 的驻点且该点并不是函数的局部极值点，则 \mathbf{x}_s 称为 $f(\mathbf{x})$ 的鞍点。



(a)



(b)

图 H-1: (a) 图中红色的点为鞍点所对应的函数曲面上的位置; (b) 典型的马鞍。

如果 $f(\mathbf{x})$ 是二元函数, $f(\mathbf{x})$ 的一个典型鞍点附近的函数曲面在外形上具有这样一个特点: 函数曲面在一个方向上是向上弯曲的, 而在另一个方向上是向下弯曲的, 看上去像一个马鞍, 如图 H-1 所示, 这便是鞍点这个名字的由来。但要注意, 并不是所有鞍点附近的函数曲面都长得像马鞍一样。比如, 考虑函数 $f(x,y)=x^2+y^3$, 点 $(0, 0)$ 为该函数的鞍点, 但函数 $f(x,y)$ 的曲面在这个点附近的形状并不是马鞍形状。我们给出判定鞍点的一个充分条件:

定理 H.5 鞍点判定的一个充分条件。函数 $f(\mathbf{x}): \mathbb{R}^n \rightarrow \mathbb{R}$ 连续且有二阶偏导数, 若 \mathbf{x}_s 为 $f(\mathbf{x})$ 的驻点且 $f(\mathbf{x})$ 在 \mathbf{x}_s 处的海森矩阵 $\nabla^2 f(\mathbf{x}_s)$ 为不定矩阵 (indefinite matrix) (即 $\nabla^2 f(\mathbf{x}_s)$ 同时具有正负特征值), 则 \mathbf{x}_s 为 $f(\mathbf{x})$ 的鞍点。

证明:

我们可以用反证法。我们假设: \mathbf{x}_s 为 $f(\mathbf{x})$ 的驻点且 $f(\mathbf{x})$ 在 \mathbf{x}_s 处的海森矩阵为不定矩阵, 但 \mathbf{x}_s 不是 $f(\mathbf{x})$ 的鞍点。如果能证明该假设不成立, 则点 \mathbf{x}_s 必为 $f(\mathbf{x})$ 的鞍点。

在 \mathbf{x}_s 为 $f(\mathbf{x})$ 驻点的前提下, 由于假设 \mathbf{x}_s 不是 $f(\mathbf{x})$ 的鞍点, 那么它要么是 $f(\mathbf{x})$ 的局部极小值点, 要么是 $f(\mathbf{x})$ 的局部极大值点。若 \mathbf{x}_s 是 $f(\mathbf{x})$ 的局部极小值点, 根据定理 H.1, $\nabla^2 f(\mathbf{x}_s)$ 必为半正定矩阵, 这与 $\nabla^2 f(\mathbf{x}_s)$ 为不定矩阵矛盾, 因此 \mathbf{x}_s 不是 $f(\mathbf{x})$ 的局部极小值点。若 \mathbf{x}_s 是 $f(\mathbf{x})$ 的局部极大值点, 根据定理 H.2, $\nabla^2 f(\mathbf{x}_s)$ 必为半负定矩阵, 这与 $\nabla^2 f(\mathbf{x}_s)$ 为不定矩阵矛盾, 因此 \mathbf{x}_s 也不是 $f(\mathbf{x})$ 的局部极大值点。这样一来, 我们便知假设不成立, 即满足条件的点 \mathbf{x}_s 必为 $f(\mathbf{x})$ 的鞍点。

需要强调的是, 定理 H.5 是鞍点判定的一个充分条件, 但不是必要条件。比如, 函数

$$f(x) = \begin{cases} \frac{1}{2}x^2, & x \geq 0 \\ -\frac{1}{2}x^2, & x < 0 \end{cases}, \text{ 容易知道 } x=0 \text{ 是 } f(x) \text{ 的一个驻点也是鞍点, 但该函数在 } x=0 \text{ 处不存在} \\ \text{二阶导数, 也就不能定义它在 } x=0 \text{ 处的海森矩阵。再比如二元函数 } f(x,y)=x^4-y^4, \text{ 容易知} \\ \text{道 } (0,0) \text{ 是该函数的驻点也是鞍点, 但函数在 } (0,0) \text{ 处的海森矩阵为 } \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}, \text{ 这不是一个不定矩} \end{math>$$

二阶导数, 也就不能定义它在 $x=0$ 处的海森矩阵。再比如二元函数 $f(x,y)=x^4-y^4$, 容易知

道 $(0,0)$ 是该函数的驻点也是鞍点, 但函数在 $(0,0)$ 处的海森矩阵为 $\begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$, 这不是一个不定矩

阵，它可以被看作是一个半正定矩阵，也可以被看作是一个半负定矩阵。

结合定理 H.3、H.4 和 H.5，我们可以总结出来在函数驻点处，函数性质与海森矩阵的关系：

命题 H.1 函数 $f(\mathbf{x}): \mathbb{R}^n \rightarrow \mathbb{R}$ 连续且有二阶偏导数， \mathbf{x}^* 为 $f(\mathbf{x})$ 的驻点，

- (1) 若 $\nabla^2 f(\mathbf{x}^*)$ 为正定矩阵，则 \mathbf{x}^* 为 $f(\mathbf{x})$ 的极小值点；
- (2) 若 $\nabla^2 f(\mathbf{x}^*)$ 为负定矩阵，则 \mathbf{x}^* 为 $f(\mathbf{x})$ 的极大值点；
- (3) 若 $\nabla^2 f(\mathbf{x}^*)$ 为不定矩阵（同时具有正负特征值），则 \mathbf{x}^* 为 $f(\mathbf{x})$ 的鞍点；
- (4) 若仅能确定 $\nabla^2 f(\mathbf{x}_s)$ 为半正（负）定矩阵，则关于 \mathbf{x}_s 点没有进一步结论。

I. 罗德里格斯公式

一个在三维欧氏空间中的旋转，其旋转轴为 \mathbf{n} ，绕 \mathbf{n} 逆时针旋转的角度为 θ ($\theta > 0$)，则该旋转的轴角表达为 $\mathbf{d} = \theta \mathbf{n}$ ，其中 $\theta = \|\mathbf{d}\|_2$ ， $\mathbf{n} = \frac{\mathbf{d}}{\theta}$ 。若 R 为表达该旋转的旋转矩阵，则 R 为，

$$R = \cos \theta \cdot I + (1 - \cos \theta) \mathbf{n} \mathbf{n}^T + \sin \theta \cdot \hat{\mathbf{n}} \quad (\text{I-1})$$

其中， $I \in \mathbb{R}^{3 \times 3}$ 为单位矩阵， $\hat{\mathbf{n}} = \begin{bmatrix} 0 & -n_3 & n_2 \\ n_3 & 0 & n_1 \\ -n_2 & n_1 & 0 \end{bmatrix}$ 。

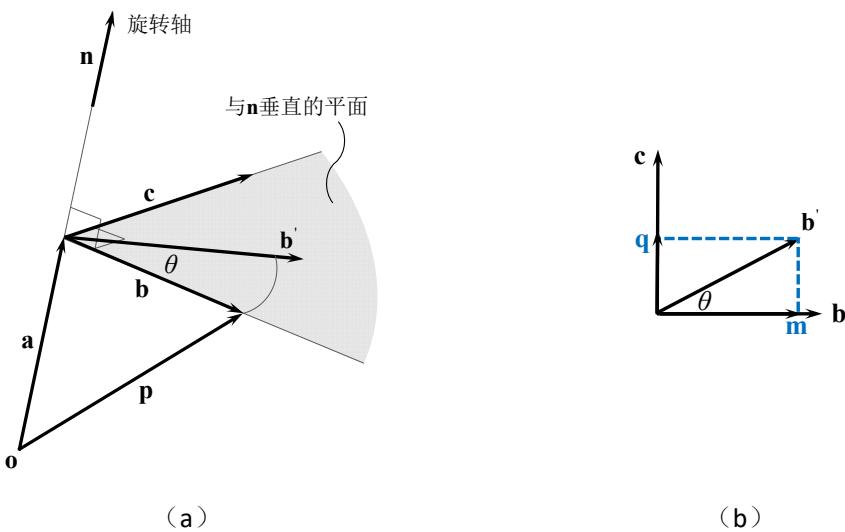


图 I-1：(a) 向量 \mathbf{p} 绕轴旋转示意图；(b) \mathbf{b} 在 \mathbf{b} 和 \mathbf{c} 上的投影分解。

式 I-1 便称为罗德里格斯公式，它给出了旋转的轴角表达到旋转矩阵表达的转换方法。

下面我们来证明一下这个公式。

如图 I-1 (a) 所示, 在三维空间中, 考虑有一个向量 \mathbf{p} , 它绕着轴 \mathbf{d} 逆时针旋转 θ , 我们来看看旋转之后的向量 \mathbf{p}_{rot} 是什么样的。设 \mathbf{d} 轴所对应的单位向量为 \mathbf{n} 。可以把 \mathbf{p} 分解为两部分, 在 \mathbf{n} 上的投影向量 \mathbf{a} 和垂直于 \mathbf{n} 的部分 \mathbf{b} , 显然 $\mathbf{b} = \mathbf{p} - \mathbf{a}$, 这样有

$$\mathbf{a} = \mathbf{n}\mathbf{n}^T\mathbf{p} \quad (I-2)$$

$$\mathbf{b} = \mathbf{p} - \mathbf{a} = \mathbf{p} - \mathbf{n}\mathbf{n}^T\mathbf{p} \quad (I-3)$$

将 \mathbf{p} 绕轴 \mathbf{n} 旋转时, \mathbf{a} 这部分是不会动的, \mathbf{b} 会被转到 \mathbf{b}' , 最终的旋转之后的结果向量 \mathbf{p}_{rot} 与 \mathbf{a} 和 \mathbf{b}' 的关系为,

$$\mathbf{p}_{rot} = \mathbf{a} + \mathbf{b}' \quad (I-4)$$

考虑矢量 $\mathbf{c} = \mathbf{n} \times \mathbf{p}$, 容易知道 $\|\mathbf{c}\| = \|\mathbf{b}\|$, 因此 $\|\mathbf{c}\| = \|\mathbf{b}\| = \|\mathbf{b}'\|$ 。如图 I-1 (b) 所示, \mathbf{b}' 在 \mathbf{b} 上的投影向量为 $\mathbf{m} = \frac{\mathbf{b}}{\|\mathbf{b}\|} \|\mathbf{b}'\| \cos \theta = \mathbf{b} \cos \theta$, \mathbf{b}' 在 \mathbf{c} 上的投影向量为 $\mathbf{q} = \frac{\mathbf{c}}{\|\mathbf{c}\|} \|\mathbf{b}'\| \sin \theta = \mathbf{c} \cos \theta$ 。显然 $\mathbf{m} + \mathbf{q} = \mathbf{b}'$, 因此,

$$\mathbf{b}' = \mathbf{b} \cos \theta + \mathbf{c} \sin \theta \quad (I-5)$$

结合式 I-2、I-4 和 I-5 可得,

$$\begin{aligned} \mathbf{p}_{rot} &= \mathbf{a} + \mathbf{b}' \\ &= \mathbf{n}\mathbf{n}^T\mathbf{p} + \mathbf{b} \cos \theta + \mathbf{c} \sin \theta \\ &= \mathbf{n}\mathbf{n}^T\mathbf{p} + (\mathbf{p} - \mathbf{n}\mathbf{n}^T\mathbf{p}) \cos \theta + \mathbf{n} \times \mathbf{p} \sin \theta \\ &= (\cos \theta \cdot I + (1 - \cos \theta)\mathbf{n}\mathbf{n}^T + \sin \theta \mathbf{n}^\wedge) \mathbf{p} \end{aligned} \quad (I-6)$$

由式 I-6 可以看出, 若以旋转矩阵的形式来表达轴角 $\theta \mathbf{n}$ 所刻画的三维空间旋转时, 该矩阵为,

$$R = \cos \theta \cdot I + (1 - \cos \theta)\mathbf{n}\mathbf{n}^T + \sin \theta \mathbf{n}^\wedge \quad (I-7)$$

参考文献

- [1] Conic section, https://en.wikipedia.org/wiki/Conic_section#CITEREFProtterMorrey1970
- [2] Fourier transform, https://en.wikipedia.org/wiki/Fourier_transform
- [3] A. Gray, Modern Differential Geometry of Curves and Surfaces with Mathematica, 2nd ed. Boca Raton, FL: CRC Press, 1997.
- [4] 同济大学数学系, 高等数学 (第六版), 高等教育出版社, 2007 年。
- [5] 李世栋, 乐经良, 冯卫国, 王纪林, 线性代数, 科学出版社, 2000 年。
- [6] K. Madsen, H.B. Nielsen, and O. Tingleff, Methods for Non-linear Least Squares Problems, Technical University of Denmark, 2004