最 优 化 模 型

III 应 用

刘红英 编

北京航空航天大学数学与系统科学学院

2017 年 3 月

内容简介

本书是Optimization Models(G.C. Calafiore and L. El Ghaoui, Cambridge University Press. Oct. 2014)一书第III部分应用的中文翻译版.

前　言

本文档由北航本科生翻译并录入(详细分工如下)，由刘红英统一校对并编辑排版。编者在此向他们的辛勤劳动表示衷心的感谢。

第十三章的协调组长为董昊，开头至13.4.2节结束由董昊翻译，13.4.3节至结束由王翔翻译。刘晗负责排版、插入图表，并校对公式。

第十四章的协调组长为周仕佶，开头至14.1节结束由周仕佶翻译，14.2节至14.3.1节结束由张林翻译，14.3.2节至14.4节结束由于浩翻译，14.5节至结束由徐亮翻译。

第十五章的协调组长为钞乾，开头至15.1节结束由钞乾翻译，15.2节由钱楚楚翻译，15.3节至结束由阮钰泽翻译，李冠巡负责排版校对。

第十六章的协调组长为沙铜，开头至16.1节结束由乐然翻译，16.2节和16.3节由沙铜翻译，16.4节和16.5节由周子易翻译，16.6节至结束由李博翻译，乐然负责处理和插入图表，沙铜负责排版校对。

由于这些翻译文稿尚未经过特别仔细的雕琢和严格的校对，仅供学习交流参考用。任何意见、建议或其它反馈都可以发送至liuhongying@buaa.edu.cn，在此深表感谢。

刘红英

2017.3 于北京

目 录

第十三章 基于数据的学习 **1**

13.1 监督学习概述 1

13.2 基于多项式模型的最小二乘预测 3

13.2.1 模型 3

13.2.2 性能评价 5

13.2.3 正则化和稀疏性 6

13.3 二分类 8

13.3.1 支持向量机 9

13.3.2 正则化与稀疏性 10

13.3.3 几何解释 10

13.3.4 鲁棒性 14

13.3.5 Logistic回归 15

13.3.6 Fisher判别 16

13.4 一种通用的监督学习问题 18

13.4.1 损失函数 19

13.4.2 惩罚和约束函数 19

13.4.3 Kernel方法 20

13.5 无监督学习 22

13.5.1 主成分分析（Principal component analysis, PCA） . . 23

13.5.2 稀疏PCA 23

13.5.3 非负矩阵分解 26

13.5.4 鲁棒PCA 27

13.5.5 稀疏图的高斯模型 29

13.6 习题 33

第十四章 计算金融 **39**

14.1 单期投资组合最优化 39
14.1.1 均值-方差最优化 40
14.1.2 投资组合约束和交易成本 42
14.1.3 夏普比率(SR)最优化 43
14.1.4 风险价值(Value-at-Risk, VaR)最优化 45
14.2 鲁棒的投资组合最优化 46
14.2.1 鲁棒的均值-方差最优化 47
14.2.2 鲁棒的风险价值(VaR)最优化 48
14.3 多期投资组合最优化 50
14.3.1 投资组合动力学 51
14.3.2 最优的开环策略 53
14.3.3 带仿射价格的闭环分配 54
14.4 稀疏指标跟踪 56
14.5 习题 58

第十五章 控制问题 **67**

15.1 连续及离散时间模型 67
15.1.1 连续时间LTI 系统 67
15.1.2 离散时间LTI系统 69
15.2 基于最优化的控制综合问题 71
15.2.1 状态跟踪的控制命令的综合 71
15.2.2 轨迹追踪控制命令综合 75
15.2.3 模型预测控制 78
15.3 分析和控制器设计中的最优化 80
15.3.1 连续时间的Lyapunov 稳定性分析 80
15.3.2 镇定状态反馈设计 82
15.3.3 离散时间的Lyapunov稳定性和镇定 86
15.4 练习 87

第十六章 工程设计 **91**

16.1 数字滤波器设计 91

16.1.1 线性相位FIR滤波器 92

16.1.2 低通FIR设计规范 93

16.1.3 通过线性规划设计FIR 94

16.1.4 逼近参考匹配的滤波器设计 96

16.2 天线阵列设计 99

16.2.1 天线方向图赋形 101

16.2.2 最小二乘设计法 103

16.2.3 SOCP设计法 105

16.3 数字电路设计 106

16.3.1 电路拓扑 107

16.3.2 设计变量 107

16.3.3 设计目标 108

16.3.4 一个电路设计问题 109

16.4 航天器设计 110

16.5 供应链管理 115

16.5.1 针对区间不确定需求的鲁棒法 116

16.5.2 区间不确定性下的仿射订购策略 117

16.5.3 用于一般的随机不确定性的场景法 121

16.6 练习 125

第十三章　基于数据的学习

如果事实与理论不符，改变那个事实.　　阿尔伯特 爱因斯坦

这章以最优化的视角简单介绍了机器学习中出现的一些典型问题。我们首先研究所谓的监督学习问题，这里的基本问题是用给定的响应数据拟合出某一模型。然后我们研究无监督学习问题，这里也是为数据构建模型，与监督学习问题的区别之处是没有特定的响应。

在这章中，我们用$\boldsymbol{X} = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_m] \in \mathbb{R}^{n\times m}$表示一般的数据矩阵，其中$\boldsymbol{x}_i$是第$i$个数据点，也称其为**样本**(example)。我们称数据点$\boldsymbol{x}$的一个特定维度为一个**特征**(feature)。常见的特征包括：给定单词在字典中出现的频率[1]；决定一个特定特征是否出现的Boolean 变量，例如一个特定演员是否在数据点代表的电影中出演；或者数值量，比如血压，温度，价格等等。

§13.1　监督学习概述

在监督学习中，我们需要为未知的函数 $\boldsymbol{x} \to y(\boldsymbol{x})$ 建立一个模型，这里 $\boldsymbol{x} \in \mathbb{R}^n$ 是输入向量， $y(\boldsymbol{x}) \in \mathbb{R}$ 是对应的输出。假设给了一些观测或样本的集合，即给定一些输入输出点对 $(\boldsymbol{x}_i, y_i), i = 1, ..., m$ ，我们用这些样本来学习一个函数模型， $\boldsymbol{x} \to \hat{y}_{\boldsymbol{w}}(\boldsymbol{x})$ ，其中 $\boldsymbol{w}$ 是模型中的参数向量。一旦得到这个模型，我们就可以预测：如果$\boldsymbol{x}$是一个新的输入点，并且也没有观察到它的输出，则我们预测其输出为 $\hat{y}(\boldsymbol{x}) = \hat{y}_{\boldsymbol{w}}(\boldsymbol{x})$ 。如果对预测的响应是任意实数，则我们可以假设输出是输入的线性函数(线性模型)，从而我们的模型形如 $\hat{y}_{\boldsymbol{w}} = \boldsymbol{w}^\top \boldsymbol{x}$ ，其中 $\boldsymbol{w} \in \mathbb{R}^n$ 包含模型的所有系数。更一般地，我们假设 $\hat{y}_{\boldsymbol{w}}(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x})$ ，其中 $\boldsymbol{x} \to \boldsymbol{\phi}(\boldsymbol{x})$ 是给定的非线性映射；置 $\boldsymbol{\phi}(\boldsymbol{x}) = (1, \boldsymbol{x})$ 可使这种一般性的表述涵盖仿射模型。

例13.1. (**库存管理中的需求量预测**) 一家大型商店需要预测来自特定地理区域的顾客对某种特定商品的需求量。假设需求量的对数是任意的实数，并且需求量的对数线性地依赖于某些输入(特征)：一年中的时刻，商品的类型，之前的销售量等等。这个问题是根据最近的刚刚过去的输入-需求对来预测未来的需求量。

有时候，需要预测二进制(binary)的响应，即对每个输入 $\boldsymbol{x}$ ， $y(\boldsymbol{x}) \in \{-1, 1\}$ ；我们将此称为一个 **标签**(label)。这样，我们可以建立形如 $\hat{y}_{\boldsymbol{w}}(\boldsymbol{x}) =$

[1]参见2.1节中的文本bag-of-words表示.

$\mathrm{sgn}(\boldsymbol{w}^\top \boldsymbol{x}+b)$的符号一线性模型，其中$\boldsymbol{w}\in\mathbb{R}^n, b\in\mathbb{R}$是模型的待定参数. 类似地，我们可以用一种包含给定的非线性映射的更复杂的模型。这里不讨论分类问题中出现的其他响应类型，比如类别响应或者更复杂的对象(比如树)。

例13.2. (**医疗效果的分类**) 二分类问题通常出现在医疗决策中，典型的场景是希望根据临床测量来预测治疗结果。威斯康新诊断乳腺癌数据(Wisconsin Diagnostic Breast Cancer, WDBC)集是公开数据库，该数据库包含了与$m=569$ 名病人有关的数据。对每个病人，我们有一个输入向量$\boldsymbol{x}_i\in\mathbb{R}^{30}$，其由30个来自于FNA乳腺肿块的数字化映像的数值组成，其描述了数字影像所呈现的细胞核的特征，像半径，纹理，周长，面积，光滑度，紧凑性，中央凹陷，对称性和细胞核的分数维数。对于每个输入点，当诊断结果是恶性的时，对应的标签是 $+1$ ；当诊断为良性的时，标签是 -1 。这里的目标是对已知根据进行训练以得到一个分类器，其可以对新来的病人进行诊断(恶性/良性诊断)。如果把新病人的各种临床测量结果表示为$\boldsymbol{x}\in\mathbb{R}^{30}$ ，则我们可以根据这个新的数据 $\boldsymbol{x}$ 和通过学习得到的分类器的预测值对这个病人做出诊断。

例13.3. (**信用卡申请的二分类**) 一家信用卡公司收到上千份的信用卡申请单。每份申请包含了下列申请人信息：年龄，婚姻状况，每年的薪金，未尝还的债务等等。问题：决定是否通过一份申请，即将申请分成两类，通过和不通过。

例13.4. (**分类问题的其他例子**) 很多情况下都会出现二分类问题。在推荐系统中，比如说电影[2]，我们有特定用户的观影偏好信息；分类的目的是确定用户是否会喜欢一部给定的新电影。在垃圾邮件过滤系统中，我们可能会收到已确定是垃圾邮件的电子邮件子集，和另外一些已知是合法的电子邮件子集；分类的目的是确定新收到的一封邮件是垃圾邮件，还是合法邮件。在时间序列的预测中，我们试图基于过去观测到的价格来决定：相对与目前的价格而言，未来的价格(比如一支股票的价格)是增加还是减少。

学习(**或训练**)(learning(or training))问题指找到“最好的”模型的系数向量 $\boldsymbol{w}$ ，使得 $\hat{y}_{\boldsymbol{w}}(\boldsymbol{x}_i)\simeq y_i, i=1,...,m.$ 因此，我们试图以模型变量 $\boldsymbol{w}$ 为决策变量，最小化预测向量 $\hat{\boldsymbol{y}}$ 和观测到的响应 $\boldsymbol{y}$ 的不一致性的某种度量。在实践中，我们希望对未看到的测试点 $\boldsymbol{x}$ 也能做出好的预测。后面会进一步讨论样本外性能(out-of-sample performance)的关键问题。

[2]参见11.4.1.4节.

如何把某种可靠性保证附加到预测输出 $\hat{y}$ 中去是监督学习的另一个关键问题。这里的可靠性保证，换句话说，就是量化的新数据的**错误分类**(misclassification) 概率[3]。这个重要的问题是统计学习理论[4] 的主题，其超出了这本书的范畴，将不再讨论。

我们首先考虑一个详细但是很基础的例子，其表明诸如正则化和样本外性能等基本概念是如何出现的。

§13.2 基于多项式模型的最小二乘预测

考虑如图13.1所示的数据集，这里输入-输出对 $(x_i, y_i), x_i \in \mathbb{R}, i = 1, ..., m$ 。我们的目标是预测与未看到的输入 x 对应的 y 的值。

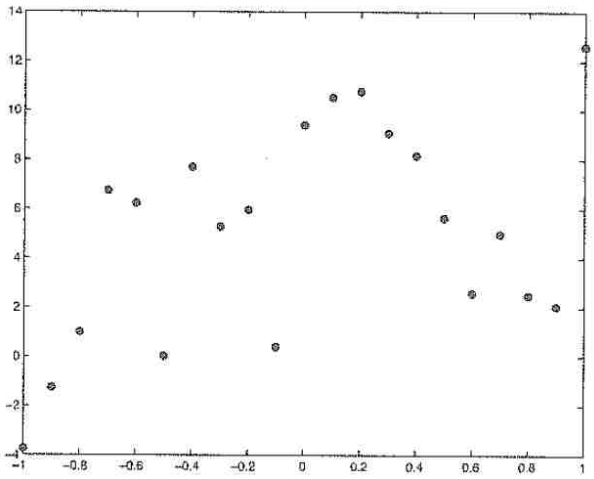

图 13.1: 输入-输出对为 $(x_i, y_i), i = 1, ..., 20$ ，的数据集

§13.2.1 模型

线性模型. 我们先为这些数据建立基本的线性模型，即对输入x，假定输出

$$y(x) = w_1 + xw_2 + e(x) = \boldsymbol{w}^\top \boldsymbol{\phi}(x) + e(x),$$

其中 w_1, w_2 是待定的权重， $\boldsymbol{\phi}(x) \doteq (1, x)$, $e(x)$ 是误差项。为了确定权重向量 $\boldsymbol{w}$ ，我们用最小二乘法，此即导致所谓的**训练误差**(training error)最

[3]在例13.2的医疗背景下，错误分类概率对应于诊断误差.

[4]参加参考书，比如T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical learning, Springer, 2008.

小化问题，即

$$\min_{\boldsymbol{w}} \frac{1}{m}\sum_{i=1}^{m}(y_i - \boldsymbol{\phi}_i^\top \boldsymbol{w})^2,$$

其中 $\boldsymbol{\phi}_i \doteq \boldsymbol{\phi}(x_i) = (1, x_i), i = 1, ..., m.$ 上述问题用最小二乘可以表述为

$$\min_{\boldsymbol{w}} \|\boldsymbol{\Phi}^\top \boldsymbol{w} - \boldsymbol{y}\|_2^2, \tag{13.1}$$

其中向量 $\boldsymbol{y} \doteq (y_1, ..., y_m)$ 是**响应**(response)，矩阵 $\boldsymbol{\Phi} \in \mathbb{R}^{2\times m}$ 是以 $\boldsymbol{\phi}_i (i = 1, .., m)$ 为列向量的数据矩阵，这里的特征包括置为1的常数特征(Φ 的第一行)和置为输入的第二特征（ Φ 的第二行)。这个问题的目标涉及到所谓的**平方损失**(squared loss)函数

$$L(z, y) = \|z - y\|_2^2.$$

通常将这个使用已有数据来拟合一个特定模型的过程称为**训练**(training)阶段，并且称问题(13.1)为训练(或学习)问题。一旦得到问题(13.1)的解$\boldsymbol{w}^*$，我们就可以用向量 $\boldsymbol{w}^*$ 为任意的输入 x 确定输出 $\hat{y}(x)$ ，即与x对应的预测值，即

$$\hat{y}(x) = \boldsymbol{\phi}(x)^\top \boldsymbol{w}^* = w_1^* + w_2^* x.$$

如此选取$\boldsymbol{w}$的合理性可以解释为：$\boldsymbol{w}$ 已经使得已有数据的误差 $\boldsymbol{e} = \boldsymbol{\Phi}^\top \boldsymbol{w} - \boldsymbol{y}$ 最小了；对新的输入 x ，我们希望上面的预测方法是准确的。在实践中，除非能够对数据的生成方式做出一些其他的假定，否则没法保证预测精度。

多项式模型

显然，如果输入的数据 x 接近于线性，我们的预测算法会表现得很好。如果不是线性的，我们可以尝试一种更复杂的模型，比如形如

$$y(x) = w_1 + xw_2 + x^2 w_3 + e(x) = \boldsymbol{w}^\top \boldsymbol{\phi}(x) + e(x)$$

的二次模型. 这里 $\boldsymbol{w} \in \mathbb{R}^3$ ， $\boldsymbol{\phi}(x) = (1, x, x^2)$ 。同上，求解最小二乘问题

$$\min_{\boldsymbol{w}} \|\boldsymbol{\Phi}^\top \boldsymbol{w} - \boldsymbol{y}\|_2^2,$$

可以得到最合适的向量 $\boldsymbol{w}$. 这里 $\boldsymbol{y} = (y_1, ..., y_m)$ ，数据矩阵 $\boldsymbol{\Phi} \in \mathbb{R}^{3\times m}$ 的列向量为 $\boldsymbol{\phi}_i = (1, x_i, x_i^2), i = 1, ..., m$ 。更一般地，我们可以用 k 次多项式模型(见图13.2)，即

$$y(x) = w_1 + xw_2 + x^2 w_3 + ... + x^k w_{k+1} + e(x).$$

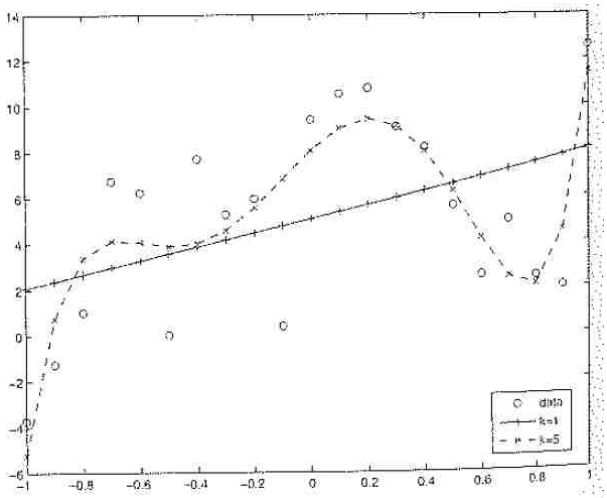

图 13.2: 两个不同的多项式模型，其中一个是线性的，另一个是7次的

§13.2.2 性能评价

我们的预测有多好？如何来比较这里的各种多项式模型？为了评价所得到的预测方法的性能，我们可以用一种称为**留一**(leave-one-out)的方法，即我们留出第j 个数据点 (x_j, y_j)，然后使用其余的数据对 $(x_i, y_i), i = 1, ..., m, i \neq j$，来拟合模型，然后使用留出的第$j$ 个数据点作为“测试点”（或者样本外），即用这个点来评价模型的性能。这样，我们把原始的数据点分成两部分，一部分是训练点的集合，其包含除了除第j个点之外的所有数据点，剩下的是测试点集合，其只包含第j个点。这种留一法可以推广到留出多个数据点的情况；另外一些方式将数据随机的剖分成训练集和测试集，后者通常包含 30% 的数据。这些不同的技术均可纳入**交叉验证**。

我们去除 $\boldsymbol{X}$ 的最后一列和向量 $\boldsymbol{y}$ 的最后一个分量，然后求解最小二乘问题。然后比较测试数据的实际值与对应的预测值，即可得到预测误差 $(\hat{y}(x_j) - y_j)^2$ 这种留一误差与包含所有样本的误差可能有所不同. (13.1)的最优值度量了具体的不同。

我们让留一法中的指标 j 从 1 到 m 来重复上述过程，然后计算测试误差的平均值，即

$$\frac{1}{m}\sum_{j=1}^{m}(\hat{y}(x_j) - y_j)^2.$$

对每个 j，在预测算法中不会被看到对应的数据。因此，希望测试误差

变成零或者特别小都是不现实的。我们画出多项式模型的阶数所对应的平均测试和训练误差，得到图13.2. 我们观察到：随着阶数 k 的增加，训练误差在减少，然而测试误差的一般是先减少，然后保持不变，然后再次增加。称这种现象为**过拟合**(over-fitting)。这种现象表明：高复杂度的模型来拟合训练数据时更好(即训练误差小)，但是一个太复杂的模型(比如一个高次多项式)对未知的数据拟合的不是很好。我们希望有这样一个“甜蜜点”，其对应的测试误差是最小的。

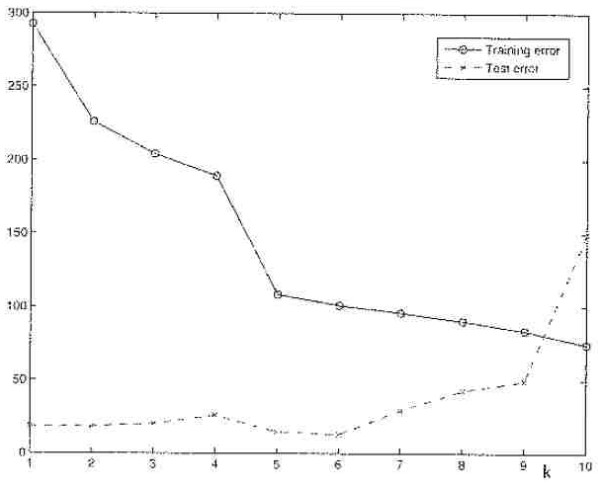

图 13.3: 训练误差和测试误差随多项式模型的阶数的变化

§13.2.3 正则化和稀疏性

正则化(Regularization). 假设我们将多项式模型的阶数置为特定的 k 。对于给定的阶数 k ，所有k次多项式的变化不尽相同；有些变化非常剧烈，这对应着大的导数值。如果没有新输入点的精确值，多项式值的剧烈变化可能会导致严重的分类测试误差。

事实上，一个多项式的“复杂度”不仅依赖于阶数，还依赖于系数的**大小**(size)。更正式的表述[5]是：对多项式 $p_{\boldsymbol{w}}(x) = w_1 + w_2 x + \cdots + w_{k+1} x^k$ ，我们有

$$\forall x \in [-1, 1] : \left| \frac{\boldsymbol{d}}{\boldsymbol{d}x} p_{\boldsymbol{w}}(x) \right| \leq k^{3/2} \|\boldsymbol{w}\|_2. \tag{13.2}$$

这意味着，如果先验地知道数据是有界的(比如 $|x| \leq 1$)，我们极小化系数向量 $\boldsymbol{w}$ 的欧氏范数，由此来控制这个给定阶数的多项式的导数。

[5]参见习题2.8.

这样，我们就可以给损失函数增加一个涉及 $\|\boldsymbol{w}\|_2$ 的惩罚项，从而得到修正的最优化问题(13.1). 比如，我们用

$$\min_{\boldsymbol{w}} \|\boldsymbol{\Phi}^\top \boldsymbol{w} - \boldsymbol{y}\|_2^2 + \lambda \|\boldsymbol{w}\|_2^2$$

代替(13.1)，其中 $\lambda \geq 0$ 是正则化参数。上述问题就是正则化的最小二乘问题. 利用6.7.3 节中的讨论，可以用线性代数的方法给出这个问题的显式解。

选取好的参数 λ 的模式与确定模型的阶数的相同。事实上，为了同时选取两个参数(正则化参数 λ 和模型的阶数 k)，需要在对应的二维空间上进行搜索。如图13.4所表明的，在很多情况下，额外的自由度确实会提高模型的测试误差。

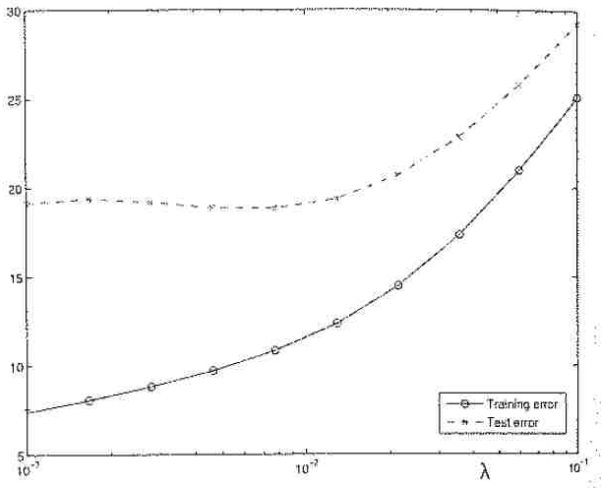

图 13.4: 固定阶数 $k = 10$ 时，训练误差和测试误差随正则化参数 λ 的演变

欧氏范数的平方并不是唯一的用来控制多项式导数的方法。事实上，我们可以用(13.2)的许多变形；比如

$$\forall x \in [-1, 1] : \left| \frac{\boldsymbol{d}}{\boldsymbol{d}x} p_{\boldsymbol{w}}(u) \right| \leq k \|\boldsymbol{w}\|_1,$$

这对应于惩罚问题

$$\min_{\boldsymbol{w}} \|\boldsymbol{\Phi}^\top \boldsymbol{w} - \boldsymbol{y}\|_2^2 + \lambda \|\boldsymbol{w}\|_1. \tag{13.3}$$

上述问题是一个LASSO问题，已经在9.6.2节对其进行了广泛的讨论。

在训练问题中，历史上一直偏好选欧氏范数的平方，因为欧氏范数的平方是因为由此得到问题可以用直接用线性代数的方法求解。与此相反，LASSO问题得不到不能直接用线性代数方法求解。随着凸优化方法的出现，LASSO 和相关的变形变得从计算上可以与正则化最小二乘法相媲美。使用非欧氏范数不仅得到可处理的问题，下面还会看到它同时还提供了一些别的有用的选择。

稀疏性(Sparsity). 在有些情况下，获得稀疏的模型是很理想的，这里稀疏的模型指系数向量 $\boldsymbol{w}$ 有许多零分量的模型。就像在例9.11中讨论的一样，在惩罚项中使用 ℓ_1- 范数能带来最优解向量 $\boldsymbol{w}$ 的稀疏性。例如，解决(13.3)的LASSO问题，其中 $\lambda = 0.1$ ，产生了稀疏的多项式

$$\begin{aligned} p_{\boldsymbol{w}}(x) &= 7.4950 + 10.7504x - 6.9644x^2 - 45.0750x^3 + \\ &\quad 42.9250x^5 - 5.4516x^8 - 0.4539x^{15} + 9.2869x^{20}, \end{aligned}$$

图13.5画出了这个多项式和最小二乘确定的多项式的函数图形及数据点. 由图可以看出，使用 ℓ_1- 范数作为惩罚项让我们找到了一种稀疏的多项式，并且这个多项式在数据所在的区间 $[-1, 1]$ 中的变化也不太剧烈。

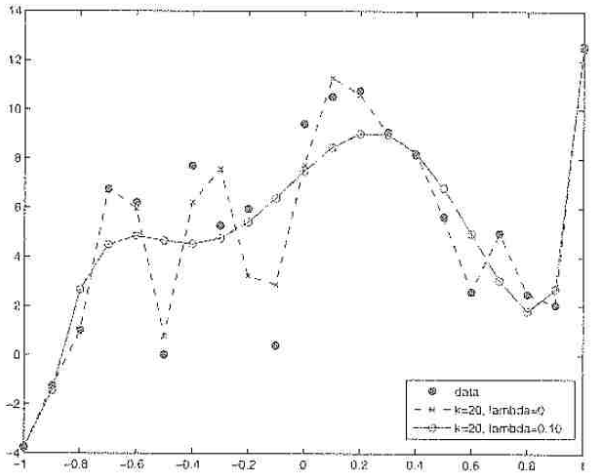

图 13.5: 在LASSO问题(13.3)中取 $k = 20, \lambda = 0.1$ 时所得到的稀疏多项式模型

§13.3 二分类

考虑二分类问题，假设我们有 m 个数据点 $\boldsymbol{x}_i$ 和对应的标签 $y_i \in$

$\{-1,1\}, i=1,...,m$ 。我们的目标是为未知的数据点 $\boldsymbol{x}$ 预测对应的标签 $\hat{y}$ 。

§13.3.1 支持向量机

在13.2节，我们需要预测一个实值变量，我们最初用了仿射函数 $x \to \boldsymbol{w}^\top \boldsymbol{\phi}(x)$ ，其中 $\boldsymbol{\phi}(x)=(1,x)$ 。当待预测的标签 $y \in \{-1,1\}$ ，可能是最简单的修正是形如

$$\hat{y} = \operatorname{sgn}(\boldsymbol{x}^\top \boldsymbol{w} + b)$$

的预测规则. 一开始，我们可能试着去找使得训练误差的平均值取到最小值的 $\boldsymbol{w} \in \mathbb{R}^n, b \in \mathbb{R}$ ，这里的训练误差可以用 0/1 函数正式地表述为

$$E(\alpha) \doteq \begin{cases} 1 & \text{如果} \alpha < 0 \\ 0 & \text{否则.} \end{cases}$$

对给定的输入-输出对 $(\boldsymbol{x}, y)$ ，当且仅当 $y(\boldsymbol{w}^\top x + b) < 0$ 时，这个分类规则 $\boldsymbol{x} \to \operatorname{sgn}(\boldsymbol{w}^\top \boldsymbol{x} + b)$ 才会产生误差。因此，对给定的 $\boldsymbol{w}$ 我们在训练集上观测到的误差的平均值是

$$\frac{1}{m} \sum_{i=1}^{m} E(y_i(\boldsymbol{w}^\top x_i + b)).$$

由此得到了形如

$$\min_{\boldsymbol{w},b} \frac{1}{m} \sum_{i=1}^{m} E(y_i(\boldsymbol{x}_i^\top \boldsymbol{w} + b)) \tag{13.4}$$

的训练问题. 不幸的是，上面的问题不是凸的，从而算法可能收敛于局部极小点，这让人很不满意。

为了克服这个问题，我们观察到上面的函数 E 不大于凸的铰链(hinge)函数 $\alpha \to \max(0, 1-\alpha)$ 。由此得到训练问题的凸近似(实际上是上界)，即

$$\min_{\boldsymbol{w},b} \frac{1}{m} \sum_{i=1}^{m} \max(0, 1 - y_i(\boldsymbol{x}_i^\top \boldsymbol{w} + b)). \tag{13.5}$$

称这个目标函数是**铰链损失**(hinge loss)函数，它是凸函数，也是多面集的；因此可将上述问题重新表述成线性规划。一个可能的线性规划表述是

$$\begin{array}{ll} \underset{\boldsymbol{w},b,e}{\text{minimize}} & \frac{1}{m}\sum_{i=1}^{m} e_i \\ \text{subject to} & e \geq 0, , i = 1, ..., m, \\ & e_i \geq 1 - y_i(\boldsymbol{x}_i^\top \boldsymbol{w} + b), i = 1, ..., m. \end{array} \tag{13.6}$$

求解这个模型，设最优解是点对 $(\boldsymbol{w}^*, b^*)$，则对新数据点$\boldsymbol{x} \in \mathbb{R}^n$的预测标签为 $\hat{y} = \operatorname{sgn}(\boldsymbol{w}^\top \boldsymbol{x} + b)$ 。(13.5)是所谓的支持向量机模型的基本构件。支持向量集的命名源于13.3.3节讨论的几何解释。

§13.3.2 正则化与稀疏性

像13.2节的例子那样，我们对控制模型的复杂度感兴趣。一种方法是控制线性函数 $\boldsymbol{x} \to \boldsymbol{x}^\top \boldsymbol{w} + b$ 在 $\boldsymbol{x}$ 取遍观测到的输入时的变化量。易见该函数的梯度就是 $\boldsymbol{w}$ ，因此控制 $\boldsymbol{w}$ 的大小就能实现我们的目标。

假设这些数据点都包含在一个以原点为圆心，半径为 R 的欧氏球里，因为

$$\max_{\boldsymbol{x},\boldsymbol{x}':\|\boldsymbol{x}\|_2 \le R, \|\boldsymbol{x}'\|_2 \le R} |\boldsymbol{w}^\top (\boldsymbol{x} - \boldsymbol{x}')| \le 2R\|\boldsymbol{w}\|_2,$$

所以我们用 ℓ_2- 范数来正则化，相应的学习问题就变成了

$$\min_{\boldsymbol{w},b} \frac{1}{m} \sum_{i=1}^{m} \max(0, 1 - y_i(\boldsymbol{x}_i^\top \boldsymbol{w} + b)) + \lambda \|\boldsymbol{w}\|_2^2, \tag{13.7}$$

其中 $\lambda \ge 0$ 是正则化参数. 通常用留一法或者类似的方法选取该参数。

假设数据点都包含在大小为 R 的盒子里，即 $\{\boldsymbol{x} : \|\boldsymbol{x}\|_\infty \le R\}$ ，则相应的边界

$$\max_{\boldsymbol{x},\boldsymbol{x}':\|\boldsymbol{x}\|_\infty \le R, \|\boldsymbol{x}'\|_\infty \le R} |\boldsymbol{w}^\top (\boldsymbol{x} - \boldsymbol{x}')| \le 2R\|\boldsymbol{w}\|_1$$

则可将上述问题修正为

$$\min_{\boldsymbol{w},b} \frac{1}{m} \sum_{i=1}^{m} \max(0, 1 - y_i(\boldsymbol{x}_i^\top \boldsymbol{w} + b)) + \lambda \|\boldsymbol{w}\|_1, \tag{13.8}$$

其中 $\lambda \ge 0$ 仍然是正则化参数。ℓ_1- 范数在上述问题中使得最优的向量 $\boldsymbol{w}$ 具有稀疏性。对这样一个向量，标量 $\boldsymbol{w}^\top \boldsymbol{x}$ 仅仅包含了 $\boldsymbol{x}$ 的一部分分量。因此这个技巧能够来辨别一些有助于得到好的预测的关键特征(x 的元素)。

§13.3.3 几何解释

基本的支持向量机模型(13.5)和它的正则化变形有许多几何解释。

线性可分数据(linearly separable data). 假设问题(13.5)的最优值是零，蕴含着训练误差为零是可达的。这表明存在向量 $\boldsymbol{w} \in \mathbb{R}^n$ 和标量 b 使得

$$y_i(\boldsymbol{x}_i^\top \boldsymbol{w} + b) \ge 0, i = 1, ..., m. \tag{13.9}$$

从几何的角度来说，这意味着超平面 $\{x : \boldsymbol{w}^\top \boldsymbol{x} + b = 0\}$ 可以将数据**线性可分的**(linear separable). 这里线性可分指标签 $y_i = 1$ 对应的数据点$\boldsymbol{x}_i$ 在这个超平面的一侧，标签 $y_i = -1$ 对应的数据点$\boldsymbol{x}_i$在另一侧，如图13.6所示。这里决策规则 $\hat{y} = \text{sgn}(\boldsymbol{w}^\top \boldsymbol{x} + b)$ 即对应着辨别新点 $\boldsymbol{x}$ 落在超平面的哪一侧。

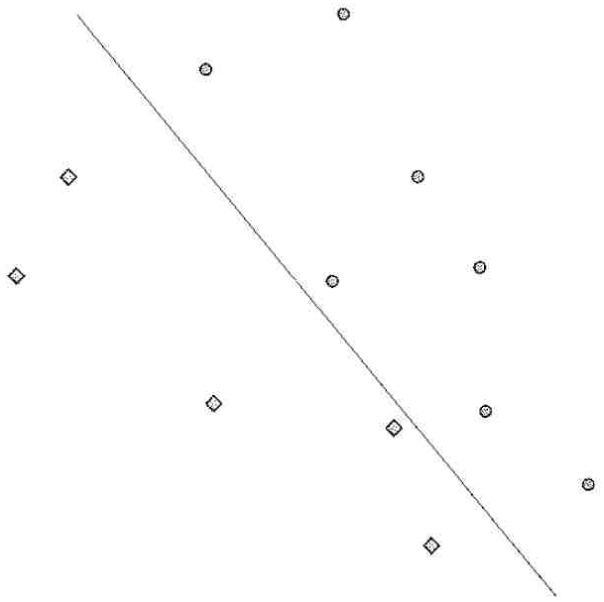

图 13.6: 严格线性可分的数据集

接下来假设是数据是**严格线性可分的**(strict linearly separable)，也就是说，条件(13.9)中不等式严格成立。这种情况出现当且仅当存在正数 $\beta > 0$ 使得

$$y_i(\boldsymbol{x}_i^\top \boldsymbol{w} + b) \geq \beta, i = 1, ..., m.$$

在这种情形下，我们可以给上述不等式两边同时除以 β 以规范化模型参数 $(\boldsymbol{w}, b)$ ，得出严格的线性可分性等价于存在$(\boldsymbol{w}, b)$ 使得

$$y_i(\boldsymbol{x}_i^\top \boldsymbol{w} + b) \geq 1, i = 1, ..., m. \tag{13.10}$$

两个“临界”(critical)超平面 $\mathcal{H}_\pm = \{\boldsymbol{x} : \boldsymbol{w}^\top \boldsymbol{x} + b = \pm 1\}$ 勾画出一个平板(slab). 没有点落在这个厚板内，同时这个厚板也把标签分别为正的和负的数据点分开了，如图13.7所示.

最大间隔分类(Maximum margin classifier). 在之前考虑的严格线性可分数据中，存在无限多的 $(\boldsymbol{w}, b)$ 满足条件(13.10)，即有无穷多个平

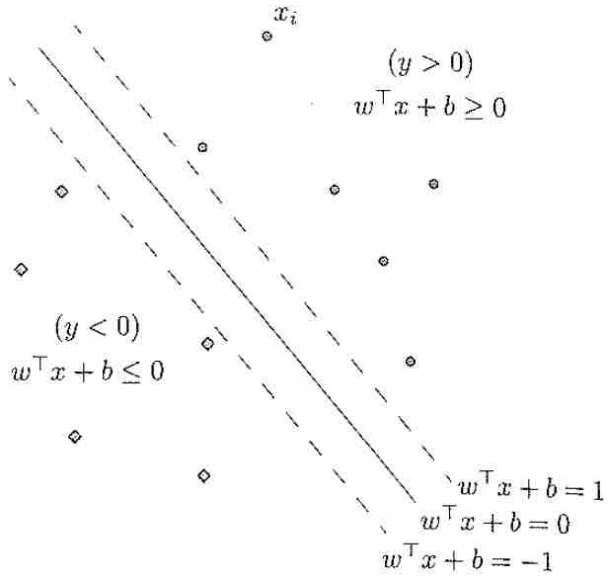

图 13.7: 严格线性可分数据集和一个分离平板

板可以分开数据集。解决这种解不惟一问题的合理作法是寻找**分离间隔**(separation margin)最大的分离平板，这里分离间隔指两个“临界”超平面 $\mathcal{H}_\pm = \{\boldsymbol{x} : \boldsymbol{w}^\top \boldsymbol{x} + b = \pm 1\}$ 之间的距离。最大化间隔的超平面无疑会具有好的性质：对于一个非常接近正(或者负)标签数据的测试点 $\boldsymbol{x}$ 而言，它将以更大的概率落在决策超平面相应的一侧。

易于验证两个临界超平面 $\mathcal{H}_+, \mathcal{H}_-$ 之间的距离是 $2/\|\boldsymbol{w}\|_2$. 在满足标签约束的前提下，为了最大化间隔，我们就需要最小化 $\|\boldsymbol{w}\|_2$ 。由此即得最优化问题

$$\begin{array}{ll} \underset{\boldsymbol{w},b}{\text{minimize}} & \|\boldsymbol{w}\|_2 \\ \text{subject to} & y_i(\boldsymbol{x}_i^\top \boldsymbol{w} + b) \geq 1, \quad i = 1, ..., m. \end{array} \tag{13.11}$$

一般称此为**最大间隔**SVM(Maximum margin SVM)，见图13.8。**支持向量**(support vector)的名字源于这样的事实：在实践中，只有很少一些点落在临界边界 $\mathcal{H}_\pm$ 上。粗略地说，如果数据点是对某种连续分布的随机抽取得到的，则超过n个特征点落在临界超平面上的概率为零。称这些落在临界超平面上的特定点为**支持向量**(support vectors)。

不可分数据(Non-separable data.) 当数据不可分(图13.9)，我们将在严格可分的条件(13.10)下引入“松弛变量”，这些松弛变量表示对违返约

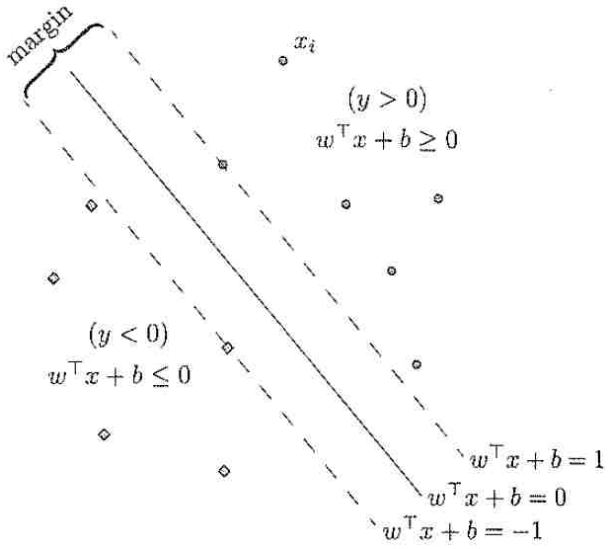

图 13.8: 最大间隔分类器

束的惩罚，这时的变量为 $(\boldsymbol{w}, b)$ 和误差向量 $\boldsymbol{e}$ ，与之对应的条件是

$$\boldsymbol{e} \geq \mathbf{0}, \quad y_i(\boldsymbol{x}_i^\top \boldsymbol{w} + b) \geq 1 - e_i, \quad i = 1, ..., m.$$

理想的作法是最小化 $\boldsymbol{e}$ 中非零元素的**个数**(number)，也就是向量$\boldsymbol{e}$的基数。这将涉及非凸问题。然而，我们可以近似这个问题，用 ℓ_1- 范数代替基数函数。因为 $\boldsymbol{e} \geq 0, \ell_1-$范数的减少对应着 $\boldsymbol{e}$ 的元素之和的减少。通过这种方式，我们精确地得到了(13.6)中表述的基本SVM问题。

软间隔分类(Classification with soft margin.) 在不可分的情况下，我们需要找到一种间隔(两个临界超平面间的距离)和训练集上的分类误差之间的折衷。这就导致了形如(13.7)的问题. 用松弛变量，我们可将(13.7)重新表述成

$$\begin{array}{ll} \underset{\boldsymbol{w},b,\boldsymbol{e}}{\text{minimize}} & \frac{1}{m}\sum_{i=1}^m e_i + \lambda\|\boldsymbol{w}\|_2^2 \\ \text{subject to} & \boldsymbol{e} \geq \mathbf{0}, e_i \geq 1 - y_i(\boldsymbol{x}_i^\top \boldsymbol{w} + b), \quad i = 1, ..., m. \end{array} \tag{13.12}$$

这里的目标函数代表着分类间隔(我们希望大一些)和所有违反项的和(我们希望小一些)之间的折衷。针对这些原因，称这种方法是软间隔分类。

由于当初的构造方式，问题(13.12)总是可行的。如果 $(\boldsymbol{w}^*, b^*, \boldsymbol{e}^*)$ 是一个最优解，并且 $\boldsymbol{e}^* = 0$ ，则 $(\boldsymbol{w}^*, b^*)$ 也是最大化间隔问题(13.11)的最

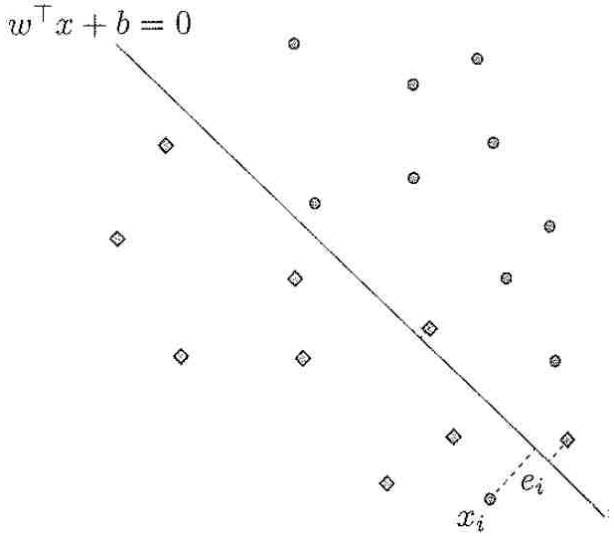

图 13.9: 不可分数据集的近似分离.

优解。如果 $\boldsymbol{e}^* \geq \mathbf{0}, \boldsymbol{e}^* \neq \mathbf{0}$ ，那么最大化间隔问题是不可行的，即数据是不可分的。在这种情况下，对 e_i^* 非零的所有 i ，问题(13.12) 中对应的约束在最优解处必须以等式成立，否则我们能够减小这些非最优的 e_i^* . 如图13.10所示，这些 e_i^* 表示数据点 $\boldsymbol{x}_i$ 到超平面 $y_i(\boldsymbol{w}^\top \boldsymbol{x} + b) = 1$ 的规范化距离. 与 $e_i^* \in (0, 1)$ 相应的点 $\boldsymbol{x}_i$ 落在间隔带内，但是仍然在决策边界正确的一侧；与 $e_i^* > 1$ 对应的点落在决策边界错误的一侧，即被错误分类。

§13.3.4 鲁棒性

下面，我们从鲁棒性的角度解释惩罚式变形(13.7)和(13.8)。假设对每个 $i = 1, ..., m$ ，仅知道数据点 $\boldsymbol{x}_i$ 落在以 $\hat{x}_i$ 为中心，以 ρ 为半径的给定的球 S_i 内。我们首先来寻找数据点能被分离的条件. 这里的分离指不管数据点在球上的精确位置，即要求数据点位于给定球内的任何位置时，都要求是可分的. 然后我们尝试最大化鲁棒性水平，即球的半径。

对给定的模型的参数向量 $(\boldsymbol{w}, b)$ ，条件

$$y_i(\boldsymbol{x}_i^\top \boldsymbol{w} + b) \geq 0, \ i = 1, ..., m, \ \forall \boldsymbol{x}_i \in S_i$$

成立当且仅当

$$y_i(\hat{\boldsymbol{x}}_i^\top \boldsymbol{w} + b) \geq \rho \|\boldsymbol{w}\|_2, i = 1, ..., m.$$

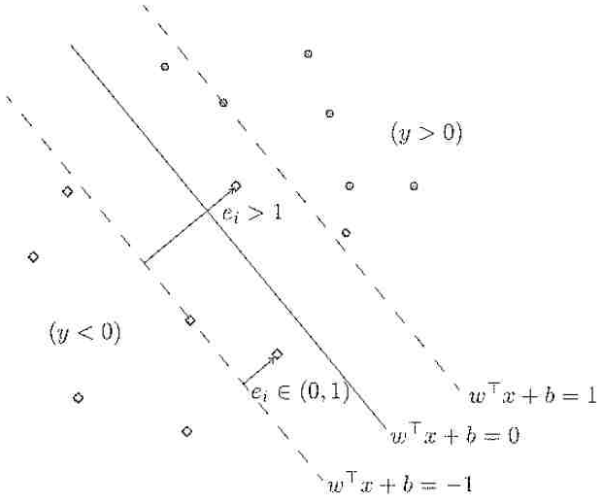

图 13.10: 软间隔分类.

由于上面的条件关于 $(\boldsymbol{w},b)$ 是齐次的，我们总是可以假设 $\rho\|\boldsymbol{w}\|_2=1$。然后，最大化 ρ 就相当于最小化 $\|\boldsymbol{w}\|_2$。用 x_i 代替中心 $\hat{x}_i$，我们就得到了问题(13.11)。几何解释是我们分离了球而不是点。

类似的解释也可适用于 ℓ_1-范数，此时假设数据未知但是有界，即落在超立方体 $\{\boldsymbol{x}:\|\boldsymbol{x}-\hat{\boldsymbol{x}}\|_\infty\le\rho\}$ 里。

§13.3.5 Logistic回归

Logistic回归是支持向量机的一种变形，即用光滑凸的对数函数 $\alpha\to\ln(1+\mathrm{e}^{-\alpha})$ 代替SVM中的Hinge函数 $\alpha\to\max(0,1-\alpha)$ 。

模型. 与支持向量机的学习问题(13.5)相对照的Logistic回归是

$$\min_{\boldsymbol{w},b}\sum_{i=1}^{m}\ln(1+\mathrm{e}^{-y_i(\boldsymbol{x}_i^\top\boldsymbol{w}+b)}). \tag{13.13}$$

当然，在实践中通常使用惩罚变形，即在目标中加上包含$\boldsymbol{w}$的范数的惩罚项。因为目标中的log-sum-exp是凸的[6]. Logistic回归模型的一个优点是可以将它们自然地推广到多分类问题，此时标签的取值不是二进制的，而是取有限个值。我们这里不讨论这种推广。

概率解释. SVMs中采用Logistic回归的主要优点是与它对应的有一个概率模型。这个模型允许为新数据点附加上属于正负类别的概率值.

[6]参见习题 2.14

这是它与SVMs的不同之处，因为SVMs仅为新点的类别提供了“是或否”的答案。Logistic模型的解释如下. 假定一个点$\boldsymbol{x}$的标签为 $y \in \{-1, 1\}$ 的概率形如

$$\pi_{\boldsymbol{w},b}(\boldsymbol{x}, y) = \frac{1}{1 + \mathrm{e}^{-y(\boldsymbol{w}^\top \boldsymbol{x} + b)}}.$$

几何上可以清楚地解释这个概率值. 如果一个点$\boldsymbol{x}$远离由 $\boldsymbol{w}^\top \boldsymbol{x} + b = 0$ 定义的超平面，且这个点所在区域对应于正标签，则 $\boldsymbol{w}^\top \boldsymbol{x} + b$ 的数值就会很大，这时上面定义的概率就接近于 1。反过来，如果 $\boldsymbol{w}^\top \boldsymbol{x} + b$ 是负数，且绝对值很大，这个点 $\boldsymbol{x}$ 就落在超平面 $\boldsymbol{w}^\top \boldsymbol{x} + b = 0$ 的另一侧，且如上定义的概率值就接近于 0。

我们采用极大似然法来拟合这个模型。这里的似然是上述模型为每个数据点 $\boldsymbol{x}_i$ 得到观测标签 y_i 所赋予的概率，即

$$\prod_{i:y_i=+1} \pi_{\boldsymbol{w},b}(\boldsymbol{x}_i, 1) \prod_{i:y_i=-1} \pi_{\boldsymbol{w},b}(\boldsymbol{x}_i, -1).$$

取负对数，并关于参数 $(\boldsymbol{w}, b)$ 来极小化，即得到问题(13.13)。

一旦观测到新数据点 $\boldsymbol{x}$，我们计算概率 $\pi_{\boldsymbol{w},b}(\boldsymbol{x}, \pm 1)$，并以 $\pi_{\boldsymbol{w},b}(\boldsymbol{x}, 1)$ 和 $\pi_{\boldsymbol{w},b}(\boldsymbol{x}, -1)$ 中最大者对应的y作为$\boldsymbol{x}$的预测标签. 这类似于SVM中置 $\hat{y} = \mathrm{sgn}(\boldsymbol{w}^\top \boldsymbol{x} + b)$，但Logistic回归有了这种额外的优点，使得我们可以在我们的预测中附加上概率水平。

§13.3.6 Fisher判别

目前解决二分类问题的另一种方法是Fisher**判别**(Fisher discrimination)。像我们将看到的那样，用线性代数的方法即可解决这种方法中所涉及的计算问题. 费舍判别是要在数据空间找到一个方向 $\boldsymbol{u} \in \mathbb{R}^n$，并且使得分别属于正标签和负标签的两类数据各自在方向 $\boldsymbol{u}$ 上的投影分得越开越好. 这里需要度量每类投影数据之间的距离相对于各自数据类的变化性。

为了便于表述，我们分别定义与正负类别相对应的两个矩阵

$$\boldsymbol{A}_+ = [\boldsymbol{x}_i]_{i:y_i=+1} \in \mathbb{R}^{n \times m_+},\ \boldsymbol{A}_- = [\boldsymbol{x}_i]_{i:y_i=-1} \in \mathbb{R}^{n \times m_-},$$

其中 $m_\pm = \mathrm{card}\{i : y_i = \pm 1\}$ 表示每个类别的大小。类似地，我们用 $\bar{\boldsymbol{a}}_+, \bar{\boldsymbol{a}}_-$ 表示两个类别的质心(centroid)，即

$$\bar{\boldsymbol{a}}_\pm = \frac{1}{m_\pm} \boldsymbol{A}_\pm \mathbf{1}.$$

最后，令

$$\tilde{\boldsymbol{A}}_{\pm} = \boldsymbol{A}_{\pm}\left(\boldsymbol{I}_{\pm} - \frac{1}{m_{\pm}}\mathbf{1}\mathbf{1}^{\top}\right).$$

表示中心化的数据矩阵.

这里，两个类别的质心在方向 $\boldsymbol{u}$ 上的距离的平方可以表示为$(\boldsymbol{u}^{\top}(\bar{a}_{+}-\bar{\boldsymbol{a}}_{-}))^2$，我们把这个量作为分离距离的一种度量，并用数据的均方差来规范化。数据集 $\boldsymbol{A}=[\boldsymbol{a}_1,...,\boldsymbol{a}_p]$ 关于平均值 $\bar{\boldsymbol{a}}$ 在方向 $\boldsymbol{u}$ 上的均方差即

$$\begin{aligned} s^2 &= \frac{1}{p}\sum_{i=1}^{p}(\boldsymbol{u}^{\top}(\boldsymbol{a}_i-\bar{\boldsymbol{a}}))^2 = \frac{1}{p}\sum_{i=1}^{p}\boldsymbol{u}^{\top}\tilde{\boldsymbol{a}}_i\tilde{\boldsymbol{a}}_i^{\top}\boldsymbol{u} \\ &= \frac{1}{p}\boldsymbol{u}^{\top}\left(\sum_{i=1}^{p}\tilde{\boldsymbol{a}}_i\tilde{\boldsymbol{a}}_i^{\top}\right)\boldsymbol{u} \\ &= \frac{1}{p}\boldsymbol{u}^{\top}\tilde{\boldsymbol{A}}\tilde{\boldsymbol{A}}^{\top}\boldsymbol{u}. \end{aligned}$$

将两个类别的质心沿方向 $\boldsymbol{u}$ 的距离的平方进行规范化，得判别准则

$$f_0(\boldsymbol{u}) = \frac{\boldsymbol{u}^{\top}(\bar{\boldsymbol{a}}_{+}-\bar{\boldsymbol{a}}_{-})(\bar{\boldsymbol{a}}_{+}-\bar{\boldsymbol{a}}_{-})^{\top}\boldsymbol{u}}{\boldsymbol{u}^{\top}(\frac{1}{m_{+}}\tilde{\boldsymbol{A}}_{+}\tilde{\boldsymbol{A}}_{+}^{\top}+\frac{1}{m_{-}}\tilde{\boldsymbol{A}}_{-}\tilde{\boldsymbol{A}}_{-}^{\top})\boldsymbol{u}} = \frac{(\boldsymbol{u}^{\top}\boldsymbol{c})^2}{\boldsymbol{u}^{\top}\boldsymbol{M}\boldsymbol{u}},$$

其中 $\boldsymbol{c}=\bar{\boldsymbol{a}}_{+}-\bar{\boldsymbol{a}}_{-}$，$\boldsymbol{M}=\frac{1}{m_{+}}\tilde{\boldsymbol{A}}_{+}\tilde{\boldsymbol{A}}_{+}^{\top}+\frac{1}{m_{-}}\tilde{\boldsymbol{A}}_{-}\tilde{\boldsymbol{A}}_{-}^{\top}$.

接下来，我们需要以$\boldsymbol{u}$为变量来极大化判别准则$f_0(\boldsymbol{u})$，这实质上是极大化两个二次型的比值，即

$$\max_{\boldsymbol{u}\neq\mathbf{0}}\frac{(\boldsymbol{u}^{\top}\boldsymbol{c})^2}{\boldsymbol{u}^{\top}\boldsymbol{M}\boldsymbol{u}}.$$

由于问题的目标函数具有齐次性，我们总可以重新调整 $\boldsymbol{u}$ 使得 $\boldsymbol{u}^{\top}\boldsymbol{c}=1$，于是得到最小化问题

$$\begin{aligned} \underset{\boldsymbol{u}}{\text{minimize}} \quad & \boldsymbol{u}^{\top}\boldsymbol{M}\boldsymbol{u} \\ \text{subject to} \quad & \boldsymbol{u}^{\top}\boldsymbol{c}=1. \end{aligned}$$

这里对称矩阵 $\boldsymbol{M}$ 是半正定的，因此上面的问题是凸二次规划。简单起见，我们假设 $\boldsymbol{M}\succ\mathbf{0}$。为了求解上述的凸优化问题，我们用拉格朗日乘子法，即由Lagrange函数

$$L(\boldsymbol{u},\mu) = \frac{1}{2}\boldsymbol{u}^{\top}\boldsymbol{M}\boldsymbol{u}+\mu(1-\boldsymbol{u}^{\top}\boldsymbol{c}),$$

关于变量 $(\boldsymbol{u},\mu)$ 的稳定点条件

$$0=\nabla_{\boldsymbol{u}}L(\boldsymbol{u},\mu)=\boldsymbol{M}\boldsymbol{u}-\mu\boldsymbol{c},\ \boldsymbol{u}^{\top}\boldsymbol{c}=1.$$

来表征解[7]。因为 $\boldsymbol{M}$ 是正定的，我们得到 $\boldsymbol{u} = \mu \boldsymbol{M}^{-1}\boldsymbol{c}$ ，将其带入约束 $\boldsymbol{u}^\top \boldsymbol{c} = 1$ ，我们得到

$$\boldsymbol{u} = \frac{1}{\boldsymbol{c}^\top \boldsymbol{M}^{-1}\boldsymbol{c}} \boldsymbol{M}^{-1}\boldsymbol{c}.$$

图13.11的左半边展示了 $\mathbb{R}^2$ 中的两组数据云，箭头代表着最优的Fisher判别方向 $\boldsymbol{u}$ 。图13.11的右半边表明：数据沿着方向$\boldsymbol{u}$投影后的直方图被很好地分开了。

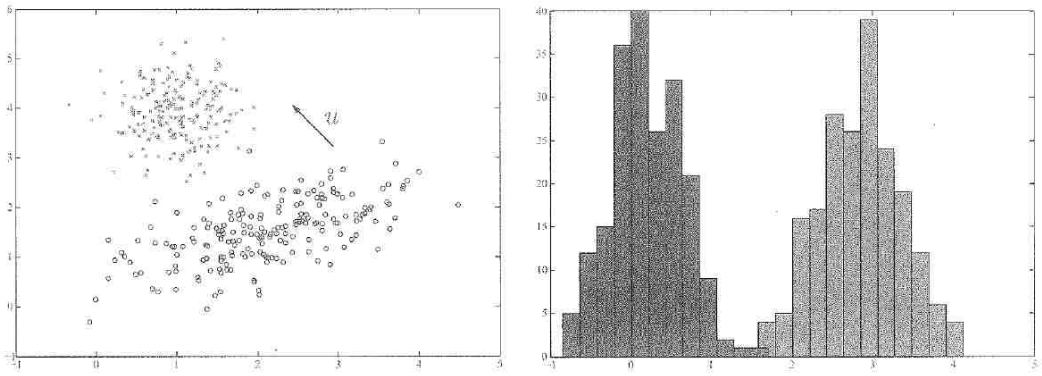

图 13.11: Fisher数据判别

§13.4 一种通用的监督学习问题

在这一节，我们考虑一种形如

$$\min_{\boldsymbol{w}} L(\boldsymbol{\Phi}^\top \boldsymbol{w}, \boldsymbol{y}) + \lambda p(\boldsymbol{w}) \tag{13.14}$$

的通用监督学习模型，其中 $\boldsymbol{w} \in \mathbb{R}^n$ 包含了模型的参数； p 是惩罚函数(通常是一种范数或者范数的平方)； λ 是正则化参数；函数 L 代表**损失**(loss)函数，用来度量预测值和响应 y 之间的差异。在上述情况下，矩阵 $\boldsymbol{\Phi}$ 包含着某种变换之后的数据点；准确来说，每列 $\boldsymbol{\phi}_i$ 形如 $\boldsymbol{\phi}_i = \boldsymbol{\phi}(\boldsymbol{x}_i)$ ，其中 $\boldsymbol{\phi}$ 是某种由用户选定的映射(可能是非线性的)[8]。

这种通用的问题类还远不能涵盖所有的监督学习的方法，但它涵盖了最小二乘回归，SVM，Logistic回归和更多的方法。

与这类学习问题相关的预测规则取决于任务和模型。在回归中，输出的预测值是一个实数，并且预测规则形如 $\hat{y} = \boldsymbol{\phi}(\boldsymbol{x})^\top \boldsymbol{w}$ 。在二分类中，

[7]参见8.5节.

[8]参见13.2节中关于映射$\boldsymbol{\phi}$的说明

我们使用符号线性函数 $\hat{y}=\operatorname{sgn}(\boldsymbol{\phi}(\boldsymbol{x})^{\top}\boldsymbol{w})$ 。我们用符号 $\hat{y}=S(\phi(\boldsymbol{x})^{\top}\boldsymbol{w})$ 来表示所有这些情况，其中 S 是 1 或者是符号函数。

§13.4.1 损失函数

损失函数 L 的形式是基于任务的(回归或者分类)，也就是说，是基于待预测的输出值的属性，还有其他因素，比如对这些数据的先验假设。通常假设损失函数可以分解成求和的形式，即

$$L(z,y)=\sum_{i=1}^{m} l(z_i,y_i),$$

其中 l 是某(通常是凸的)函数。表13.1给出了标准选择。

表 13.1: 标准的损失函数

损失函数	$l(z,y)$	任务
欧氏范数的平方	$\|z-y\|_2^2$	回归
ℓ_1-范数	$\|z-y\|_1$	回归
ℓ_∞-范数	$\|z-y\|_\infty$	回归
Hinge函数	$\max(0,1-yz)$	二分类
Logistic函数	$\ln(1+\mathrm{e}^{-yz})$	分类

选择特定的损失函数需要基于任务和对数据的假设，此外实践中还应考虑可用的软件(用其求解学习问题(13.14))等。

§13.4.2 惩罚和约束函数

惩罚函数 L 有好多种，具体的选择取决于任务和其它假设。历史上偏好选 ℓ_2-范数的平方是因为对应的最优化问题(13.14) 继承了诸如光滑性(当损失函数自身是光滑的时)等好的性质。

选取一个惩罚(Choosing a penalty). 如果我们认为最优的向量$\boldsymbol{w}$是稀疏的，即只有很少一部分特征主导了对输出的预测，则可用 ℓ_1- 范数惩罚来提高稀疏性。如果向量 $\boldsymbol{w}$ 由分块形式 $\boldsymbol{w}^{(i)},i=1,...,p$ ，如果我们认为许多块都应该是零向量，则可以使用复合范数

$$\sum_{i=1}^{p}\|\boldsymbol{w}^{(i)}\|_2.$$

事实上，上述函数是由范数组成的向量 $(\|\boldsymbol{w}^{(1)}\|_2, ..., \|\boldsymbol{w}^{(p)}\|_2)$ 的 ℓ_1-范数。我们希望提高这个向量的稀疏性。

实践中很重要的一种考量是：设计的惩罚应该使得学习问题(13.14)有惟一解。为此，通常会添加 ℓ_2- 范数的平方的小的倍数，这能确保目标函数是强凸的。

约束版(Constrained Versions). 注意到惩罚形式(13.14)涵盖了形如

$$\min_{\boldsymbol{w}}\{L(\boldsymbol{\Phi}^\top \boldsymbol{w}, \boldsymbol{y}) : \boldsymbol{w} \in C\}$$

的约束式变形，这里 C 是一个(通常是凸的)集合，比如欧氏球。如果恰当地选取 p 为

$$p(\boldsymbol{w}) = \begin{cases} 1 & 若\boldsymbol{w} \in C \\ 0 & 否则 \end{cases}$$

就可将上述这个约束形式的学习问题表述成惩罚形式的问题. 你也许好奇使用给定的惩罚，这个惩罚的平方，及与之相对应的约束版在实际中的差异[9]。为了说明这个问题，我们考虑LASSO 和两种变形，即

$$\begin{aligned} p^*_{\text{lasso}} &\doteq \min_{\boldsymbol{w}} \|\boldsymbol{\Phi}^\top \boldsymbol{w} - \boldsymbol{y}\|_2^2 + \lambda\|\boldsymbol{w}\|_1, \\ p^*_{\text{sqrt}} &\doteq \min_{\boldsymbol{w}} \|\boldsymbol{\Phi}^\top \boldsymbol{w} - \boldsymbol{y}\|_2 + \mu\|\boldsymbol{w}\|_1, \\ p^*_{\text{constr}} &\doteq \min_{\boldsymbol{w}}\{\|\boldsymbol{\Phi}^\top \boldsymbol{w} - \boldsymbol{y}\|_2 : \|\boldsymbol{w}\|_1 \le \alpha\}, \end{aligned}$$

其中 $\lambda \ge 0, \mu \ge 0, \alpha \ge 0$ 是正则化参数。对一个固定的三元对 (λ, μ, α)，与上述问题相对应的解通常是不同的。然而，当参数 λ, μ, α 遍历所有非负实数时，则上述每个最优化问题的解所生成的路径对于同一个实际问题(即数据$\boldsymbol{\Phi}$ 和$\boldsymbol{y}$)而言是相同的。

§13.4.3 Kernel方法

倘若学习问题是欧氏范数的平方的惩罚式

$$\min_{\boldsymbol{w}} L(\boldsymbol{\Phi}^\top \boldsymbol{w}, \boldsymbol{y}) + \lambda\|\boldsymbol{w}\|_2^2 \tag{13.15}$$

时，Kernel方法使得非线性法则的背景下也能够进行快速计算。我们发现第一项涉及损失，且仅在 $\boldsymbol{\Phi}^\top \boldsymbol{w}$ 中依赖于 $\boldsymbol{w}$ 。由线性代数基本定理(见3.2.4节)，可将任一向量 $\boldsymbol{w} \in \mathbb{R}^n$ 分解为 $\boldsymbol{w} = \boldsymbol{\Phi}\boldsymbol{v} + \boldsymbol{r}$，其中 $\boldsymbol{\Phi}^\top \boldsymbol{r} = \boldsymbol{0}$ 。从而可

[9]参见习题13.5.

将 $\boldsymbol{w}$ 的欧氏范数的平方分解成 $\|\boldsymbol{\Phi v}\|_2^2+\|\boldsymbol{r}\|_2^2$ 。将$\boldsymbol{w}$和$\|\boldsymbol{w}\|_2^2$的这种分解代入学习问题(13.15)，得

$$\min_{\boldsymbol{v},\boldsymbol{r}} L(\boldsymbol{\Phi}^\top\boldsymbol{\Phi v},\boldsymbol{y})+\lambda(\|\boldsymbol{\Phi v}\|_2^2+\|\boldsymbol{r}\|_2^2).$$

显然最优的 $\boldsymbol{r}$ 是零向量。从几何上看，这意味着最优的 $\boldsymbol{w}$ 属于数据 $\boldsymbol{\Phi}$ 生成的空间，即存在 $\boldsymbol{v}$ 使得最优解 $\boldsymbol{w}=\boldsymbol{\Phi v}$ 。此时，学习问题(13.15)就变成了

$$\min_{\boldsymbol{v}} L(\boldsymbol{Kv},\boldsymbol{y})+\lambda\boldsymbol{v}^\top\boldsymbol{Kv}, \tag{13.16}$$

我们称这里的 $\boldsymbol{K}=\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\in\mathbb{R}^{m\times m}$ 为**核矩阵**(kernel matrix)。因此，上述问题仅依赖于标量积 $K_{ij}=\boldsymbol{\phi}(\boldsymbol{x}_i)^\top\boldsymbol{\phi}(\boldsymbol{x}_j)$ ，其中 $1\le i,j\le m$.

预测规则是相似的：如果 $\boldsymbol{x}$ 是一个新数据点，则预测值

$$\hat{y}=S(\boldsymbol{\phi}(\boldsymbol{x})^\top\boldsymbol{w})=S(\boldsymbol{\phi}(\boldsymbol{x})^\top\boldsymbol{\Phi v}).$$

现在的预测值也只涉及变换后的新点 $\boldsymbol{\phi}(\boldsymbol{x})$ 和变换后的数据 $\boldsymbol{\phi}(\boldsymbol{x}_i)$ 的标量积。

这意味着所有的计算取决于我们是否能快速求出任意一对变换后的点 $\boldsymbol{\phi}(\boldsymbol{x}),\boldsymbol{\phi}(\boldsymbol{x}')$ 的标量积。考虑 $n=2$ 时二维问题的例子，我们寻找形如

$$\hat{y}=w_1+\sqrt{2}w_2x_1+\sqrt{2}w_3x_2+w_4x_1^2+\sqrt{2}w_5x_1x_2+w_6x_2^2.$$

的**二次**分类规则。这个规则对应于非线性映射 $\boldsymbol{x}\to\boldsymbol{w}^\top\boldsymbol{\phi}(\boldsymbol{x})$ ，其中

$$\boldsymbol{\phi}(\boldsymbol{x})=(1,\sqrt{2}x_1,\sqrt{2}x_2,x_1^2,\sqrt{2}x_1x_2,x_2^2).$$

我们发现对任意一对点 $\boldsymbol{x},\boldsymbol{x}'\in\mathbb{R}^n$，

$$\boldsymbol{\phi}(\boldsymbol{x})^\top\boldsymbol{\phi}(\boldsymbol{x}')=(1+\boldsymbol{x}^\top\boldsymbol{x}')^2.$$

一般来说，对于形式是 $\boldsymbol{x}$ 的 d 阶多项式的分类规则，我们可以作映射 $\boldsymbol{\phi}(\boldsymbol{x})\in\mathbb{R}^p$ ，其中[10] $p=O(n^d)$ ，使得任意一对点 $\boldsymbol{x},\boldsymbol{x}'\in\mathbb{R}^n$有

$$\boldsymbol{\phi}(\boldsymbol{x})^\top\boldsymbol{\phi}(\boldsymbol{x}')=(1+\boldsymbol{x}^\top\boldsymbol{x}')^d.$$

计算上述标量积**先验上**(a priori)需要 $O(p)=O(n^d)$ ，这是指数(关于特征数目 n)时间的。上述公式可以将计算复杂度降到 $O(n)$ ，这样的计算复杂度是非常低的。

[10]这里 $O(q)$ 表示是 q 的某线性函数.

表 13.2: 常用的有效核函数

名称	$K(\boldsymbol{x}, \boldsymbol{x}')$	参数
线性	$\boldsymbol{x}^\top \boldsymbol{x}'$	无
多项式	$(1 + \boldsymbol{x}^\top \boldsymbol{x}')^d$	阶数 d
高斯	$\mathrm{e}^{\frac{1}{2\sigma^2}\|\boldsymbol{x}-\boldsymbol{x}'\|_2^2}$	宽度 σ

为了形成优化问题(13.16)以及给出预测值，我们需要作 $O(m^2)$ 个标量积。因为求解QP(13.16) 的复杂度为 $O(m^3)$ ，kernel方法的总复杂度为 $O(m^3 + nm^2)$, 这与多项式的阶数 d 无关。概括起来，即我们求解多项式模型的且具有欧氏范数惩罚的学习问题所需的计算量与解决同一问题的线性模型所需的计算量一样多。

Kernel的思想也可用于非多项式模型。相对于多项式多项模型，这里不必显式地定义非线性映射，即可以绕过非线性映射 ϕ 的显式定义，只需确定核函数

$$K(\boldsymbol{x}, \boldsymbol{x}') \doteq \boldsymbol{\phi}(\boldsymbol{x})^\top \boldsymbol{\phi}(\boldsymbol{x}')$$

的值即可。当然，并不是每个函数 $(\boldsymbol{x}, \boldsymbol{x}') \to K(\boldsymbol{x}, \boldsymbol{x}')$ 都能定义一个有效的核. 对形如 $\boldsymbol{x} \to \boldsymbol{\phi}(\boldsymbol{x})$ 的映射，需要满足半正定条件。在实践中，对核函数的取值附加一个条件将会使计算效率变得很低。表13.2 中给出了通常选取的有效核。

例13.5. (**乳腺癌的分类**) 回到关于乳腺癌的例13.2(基于威斯康星州的数据集)。为了使数据的软间隔分离的图像可视化，我们先将数据降维，即利用PCA方法(参见5.3.2 节)把数据从30维降到2 维。需要注意的是，在实际中这不一定是一个好的想法，因为将原空间中可分离的数据集投影到低维空间后，所得投影数据不一定是可分离的。然而，在这里我们让数据的维数急剧减少只是为了展示数据的二维图像。

对于投影后的二维数据点，我们求解 $\lambda = 1$ 的问题(13.16)以便训练出一个高斯核($\sigma = 1$)的SVM，结果如图13.12所示，其中 + 代表良性数据点，* 代表恶性数据点，圆圈标记的是支持向量，实线表示分离面。

§13.5 无监督学习

在无监督学习中，仅有数据点 $\boldsymbol{x}_i \in \mathbb{R}^n, i = 1, ..., m$，没有指定的标签

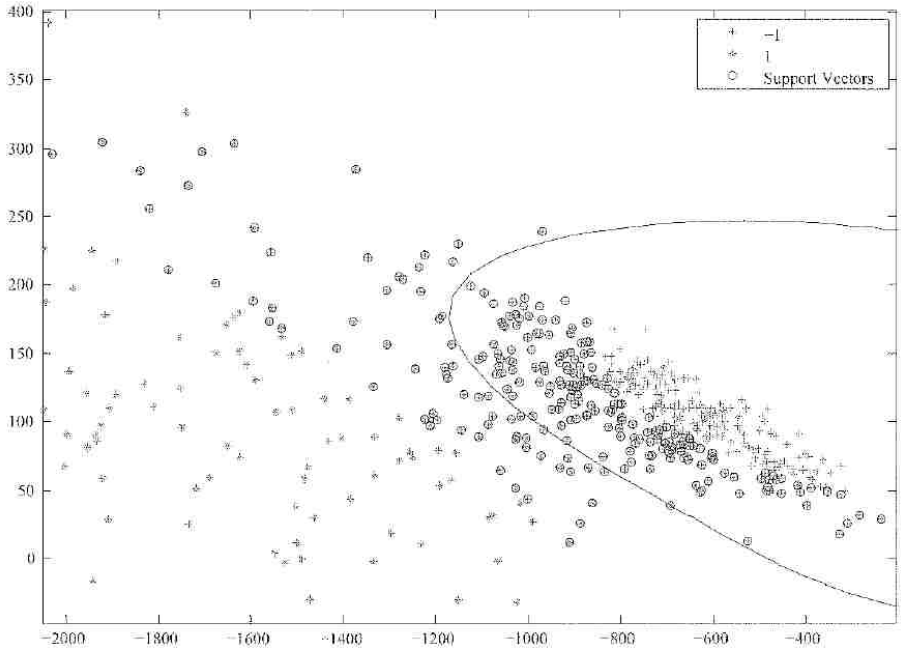

图 13.12: 乳腺癌数据集的二维投影数据的SVM分类

或反馈。我们的任务是从数据中得到一些信息或者结构。

§13.5.1 主成分分析（Principal component analysis, PCA）

在主成分分析中，我们的目标是寻找数据集中最重要的(或就含有信息而言)方向，即沿着该方向数据的变化最大。已经在5.3.2节中详细描述了PCA。数值上，PCA 问题的计算归结为计算数据矩阵的SVD。下面我们将描述基本PCA问题的一些变形。

§13.5.2 稀疏PCA

PCA方法在许多领域中得到广泛应用，且已发展出了多种变形。最近发展了一种称为稀疏PCA的变形，其提升了所产生的主方向的可解释性，具有更大的使用价值。在问题(5.18)中添加一个约束用以限制决策变量中非零元素的个数，可得到稀疏PCA，即

$$\begin{array}{ll}\underset{\boldsymbol{z}\in\mathbb{R}^n}{\text{maximize}} & \boldsymbol{z}^\top(\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\top)\boldsymbol{z} \\ \text{subject to} & \|\boldsymbol{z}\|_2 = 1, \mathrm{card}(\boldsymbol{z}) \le k,\end{array} \tag{13.16}$$

其中 $k \leq n$ 是由用户指定的基数的上界。如同在注记5.3中讨论的那样，原始PCA问题是寻找方向 $\boldsymbol{z}$ ，其能“解释”数据中的最大变化量；在实践中，这个方向 $\boldsymbol{z}$ 通常不是稀疏的。相比之下，由稀疏PCA 给出的方向如定义所述是稀疏的，比如 $k \ll n$，而这种情况正是我们在实践中所感兴趣的。如果数据的每一维(数据矩阵的行)对应于一个可解释的量，如价格、温度、体积等，那么方向越稀疏，表明它只涉及很少的(即 k 个)基本量，从而它的可解释性就越强。

当数据的维数较小时，稀疏PCA问题是易于求解的：只需考虑将数据矩阵 $\tilde{\boldsymbol{X}}$ 移走$n-k$维(行)，然后对剩下的行所组成的子矩阵求解一般的PCA 问题。对于维数较大的情况，会出现行向量的组合爆炸问题，即生成非常多的可能的子矩阵，从而使得这种穷举法失效。在12.5.4节中描述了一个有效的解决稀疏PCA的方法，13.7节的关于文本分析的问题使用了这种求解方法。

例13.6. (**市场数据的稀疏**PCA) 回到例5.3，我们发现由矩阵 $\boldsymbol{U}$ 的第一列

$$\boldsymbol{u}_1 = [-0.4143 \quad -0.4671 \quad -0.4075 \quad -0.5199 \quad -0.0019 \quad -0.4169]^\top$$

给出的方向对应的方差最大，但是 $\boldsymbol{u}_1$ 是一个稠密向量，其中只有一个分量(第五个)接近于零. 与 $\boldsymbol{u}_1$ 对应的最大奇异值 $\sigma_1 = 1.0765$ ，方差比

$$\eta_1 = \frac{\sigma_1^2}{\sigma_1^2 + \cdots + \sigma_6^2} = 67.766\%,$$

这意味着大约 68% 的方差包含在一个由 $\boldsymbol{u}_1$ 作为权重的特定指标中，该指标几乎是所有指标的非平凡组合。

现在求解 $k = 2$ 的稀疏PCA问题(13.16)，由此可得到一个基数小的方向，沿着该方向方差也较大。我们可以通过穷尽原始六维空间的所有二维组合来搜索，即我们要解决原始矩阵 $\tilde{\boldsymbol{X}}$ 任意保留两行并去除其余行后所得到的所有可能的矩阵 $\tilde{\boldsymbol{X}}_2$ 的PCA问题。我们发现与行指标 4 和 5 (称为指标PAC和BOT)对应的PCA的可解释方差在所有可能的情况中是最大的。这时对应的数据矩阵的最大奇异值 $\tilde{\sigma}_1 = 0.6826$ ，这与原始数据矩阵的最大奇异值 $\sigma_1 = 1.0765$ 是可比的。依据解释方差，这两个指标相对于总方差的比值为

$$\frac{\tilde{\sigma}_1^2}{\sigma_1^2 + \cdots + \sigma_6^2} = 0.2725\%.$$

因此，仅这两个指标就能捕获大约 28% 的总市场方差。

表 13.3: 与NYT数据的前5个稀疏主成分的非零元素对应的词语

第一PC	第二PC	第三PC	第四PC	第五PC
million	point	official	president	school
percent	play	government	Campaign	program
business	team	United_states	bush	children
company	season	u_s	administration	student
market	game	attack		
companies				

例13.7. (**主题发现中的稀疏**PCA) 当要从大量文档中发现主题时，可以应用稀疏PCA。UCI机器学习数据库中的**纽约时报**(The New York Times, NYT)文本集包含了300,000篇文章，和一个包含了102,660 个独一无二的单词的字典，这使得文件大小达到1GB。这个给定的数据没有分类标签，且由于版权原因，文件名和文档级别元数据(如文章段落信息)均未知。文本通过“词袋”(bag-of-words)法转换为数值形式。

表13.3列出了排名前五的稀疏主成分的非零系数对应的非零特征(词语)。每个成分代表了一个确定的话题。的确，**纽约时报**的第一主成分是关于商业的，第二个是关于体育的，第三个是关于美国的，第四个是关于政治的，第五个是关于教育的。即使未提供元数据，稀疏PCA 仍然能清楚地识别，并完美地与**纽约时报**在其网站上对文章进行分类时所采用的主题相对应。

稀疏PCA的思想也可以推广到具有向量基数约束的矩阵的低秩逼近问题。例如，带基数约束的秩一逼近问题形如

$$\begin{array}{ll}\underset{\boldsymbol{p}\in\mathbb{R}^n,\boldsymbol{q}\in\mathbb{R}^m}{\text{minimize}} & \|\boldsymbol{X}-\boldsymbol{p}\boldsymbol{q}^\top\|_{\mathrm{F}} \\ \text{subject to} & \mathrm{card}(\boldsymbol{p})\le k,\ \mathrm{card}(\boldsymbol{q})\le h,\end{array}$$

其中 $\boldsymbol{X}\in\mathbb{R}^{n\times m}$ ，且 $k\le n, h\le m$ 是用于限制向量 $\boldsymbol{p},\boldsymbol{q}$ 中非零元数的数目的参数，由用户选定。当 n,m 较小时，类似于上述稀疏PCA问题那样，可用用SVD解决这里的问题；否则，求解该问题将变得非常困难，但已经给出了基于幂迭代的有效算法，详见12.5.4 节。

§13.5.3 非负矩阵分解

给定非负的 $n \times m$ 阶数据矩阵 $\boldsymbol{X}$ ，**非负矩阵分解**(Non-Negative Matrix Factorization, NNMF)指的是用非负低秩成分矩阵 $\boldsymbol{P}, \boldsymbol{Q}$ 的乘积 $\boldsymbol{P}\boldsymbol{Q}^\top$ 来逼近数据矩阵 $\boldsymbol{X}$，这里 $\boldsymbol{P} \in \mathbb{R}^{n\times k}, \boldsymbol{Q} \in \mathbb{R}^{m\times k}$表示低秩非负矩阵，即它们的列数 $k \ll \min(n, m)$ 。因此，这是在5.3.2.4节中讨论过的低秩矩阵分解问题的简单改进。

为说明NNMF，我们先看数据矩阵 $\boldsymbol{X}$ 的普通秩一逼近，即

$$\min_{\boldsymbol{p}\in\mathbb{R}^n, \boldsymbol{q}\in\mathbb{R}^m} \|\boldsymbol{X} - \boldsymbol{p}\boldsymbol{q}^\top\|_{\mathrm{F}}.$$

如果存在 $\boldsymbol{p}, \boldsymbol{q}$ 使得问题的目标值很小，即有 $\boldsymbol{X} \approx \boldsymbol{p}\boldsymbol{q}^\top$ ，我们就可以将 $\boldsymbol{p}$ 解释为典型数据点，$\boldsymbol{q}$ 解释为典型特征。如果 $\boldsymbol{X}$ 非负，就可能存在一个秩一逼近 $\boldsymbol{X} \approx \boldsymbol{p}\boldsymbol{q}^\top$ ，但是这里 p, q 的某些分量是负的。对于这种情况，很难解释结果，比如当 $\boldsymbol{p}$ 存在负分量时，我们就不能将 $\boldsymbol{p}$ 解释成典型数据点。

NNMF方法可应用于非负数据矩阵。如果目标秩 $k = 1$ ，则NNMF问题即

$$\begin{array}{ll} \underset{\boldsymbol{p}\in\mathbb{R}^n, \boldsymbol{q}\in\mathbb{R}^m}{\text{minimize}} & \|\boldsymbol{X} - \boldsymbol{p}\boldsymbol{q}^\top\|_{\mathrm{F}} \\ \text{subject to} & \boldsymbol{p} \geq \boldsymbol{0},\ \boldsymbol{q} \geq \boldsymbol{0}. \end{array}$$

该结果的解释为：如果 $\boldsymbol{p}, \boldsymbol{q}$ 为非负向量，且 $\boldsymbol{X} \approx \boldsymbol{p}\boldsymbol{q}^\top$ ，则 $\boldsymbol{X}$ 的每列均与向量 $\boldsymbol{q}$ 成正比例，向量 $\boldsymbol{p}$ 的分量分别对应着这些比例系数。因此每个数据点都沿着单个“轮廓” $\boldsymbol{q}$ 运动，数据点的不同之处是与每个数据点对应一个特定的非负比例因子。更一般的，NNMF 问题可表述为

$$\begin{array}{ll} \underset{\boldsymbol{P}\in\mathbb{R}^{n\times k}, \boldsymbol{Q}\in\mathbb{R}^{m\times k}}{\text{minimize}} & \|\tilde{\boldsymbol{X}} - \boldsymbol{P}\boldsymbol{Q}^\top\|_{\mathrm{F}} \\ \text{subject to} & \boldsymbol{P} \geq \boldsymbol{0}, \boldsymbol{Q} \geq \boldsymbol{0}. \end{array}$$

这里的解释是数据点是 $\boldsymbol{Q}$ 的 k 个列向量确定的基本轮廓的线性组合。该问题是非凸的，难以精确求解。经常用到的方法是一种基于坐标块下降[11]的方法，即关于 $\boldsymbol{P}, \boldsymbol{Q}$ 交替地进行极小化。每个子问题(关于 $\boldsymbol{P}$ 或者 $\boldsymbol{Q}$ 的极小化)都是凸问题，事实上是凸二次规划。

例13.8. (**用于文本文档主题发现的**NNMF) 利用“词袋”方法(见例2.1)，我们可以将包含m个文本文档的集合用矩阵 $\boldsymbol{X}$ 来表示，其中 x_{ij} 表示术语 i 在文件 j 中出现的次数。对于这种矩阵，自然会想到用NNMF。秩

[11]参见12.5.3 节

一非负逼近 $\boldsymbol{X} \approx \boldsymbol{p}\boldsymbol{q}^\top$，其中 $\boldsymbol{p} \geq \boldsymbol{0}, \boldsymbol{q} \geq \boldsymbol{0}$ 表示每个文档 i （ $\boldsymbol{X}$ 的列)可用 $p_i\boldsymbol{q}$ 近似，即秩一逼近下所有文档均相似于一个文档。为了将 $p_i\boldsymbol{q}$ 解释为一个文档，这里 $\boldsymbol{p}, \boldsymbol{q}$ 的非负性是至关重要的。秩 k 阶逼近表示所有文档是 k 个基本文档的“混合”。

例13.9. (**学生得分矩阵的**NNMF) 考虑关于四门考试和五位学生的二进制得分矩阵

$$\boldsymbol{X} = \begin{bmatrix} 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 \end{bmatrix},$$

其中当学生 j 通过考试 i 时，x_{ij} 为1，否则为 0。我们可以将其写作 $\boldsymbol{X} = \boldsymbol{P}\boldsymbol{Q}^\top$，其中 $\boldsymbol{P}, \boldsymbol{Q}$ 均只有三列

$$\boldsymbol{P} = \begin{bmatrix} 0 & 1/2 & 0 \\ 1/2 & 0 & 0 \\ 0 & 1/2 & 1 \\ 1/2 & 0 & 0 \end{bmatrix}, \quad \boldsymbol{Q} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 2 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \\ 0 & 2 & 2 \end{bmatrix}.$$

我们可将 $\boldsymbol{P}, \boldsymbol{Q}$ 的三列解释为“成功通过考试所需要的三项‘技能’”。矩阵 $\boldsymbol{P}$ 代表着通过每门考试需要何种技能；矩阵 $\boldsymbol{Q}$ 表明每个学生各掌握了什么技能。以第一门考试和第一个学生为例，我们从 $\boldsymbol{P}$ 中可以看到通过考试 1 需要技能 2，但从 $\boldsymbol{Q}$ 中看到学生 1 仅掌握技能 1，故学生 1 不能通过考试 1。事实上，学生 1 只能通过考试 3，因为其余考试都要求技能 2 或 3.

---

§13.5.4 鲁棒PCA

低秩矩阵逼近和与之密切相关的PCA均可解释为将给定的数据矩阵 $\boldsymbol{X}$ 表示为矩阵之和 $\boldsymbol{X} = \boldsymbol{Y} + \boldsymbol{Z}$，其中 $\boldsymbol{Y}$ 为低秩矩阵，$\boldsymbol{Z}$ 较小。

鲁棒PCA是PCA的另一种变形[12]，我们将借此尝试将给定的 $n \times m$

[12]参见E.J. Candès, X. Li, Y. Ma, and J. Wright, Robust principal component analysis?, J. ACM, 2011.

矩阵 $\boldsymbol{X}$ 分解为低秩成分和**稀疏**成分之和，即形如

$$\begin{aligned} \underset{\boldsymbol{Y},\boldsymbol{Z}\in\mathbb{R}^{n\times m}}{\text{minimize}} \quad & \text{rank}(\boldsymbol{Y})+\lambda\text{card}(\boldsymbol{Z}) \\ \text{subject to} \quad & \boldsymbol{X}=\boldsymbol{Y}+\boldsymbol{Z}, \end{aligned}$$

其中 $\text{card}(\boldsymbol{Z})$ 是 $n\times m$ 矩阵 $\boldsymbol{Z}$ 中非零元的个数，λ 是一个参数，其允许我们在 $\boldsymbol{X}$ 的成分 $\boldsymbol{Y}$ 的低秩性和另一成分 $\boldsymbol{Z}$ 的稀疏性之间进行折衷。当认为测量数据具有低秩结构，但是存在着由稀疏成分表现的“spike”（或outlier）时，会出现这种问题。

鲁棒PCA问题通常很难求解。在11.4.1.4节中提到，我们可以用矩阵的核范数作为秩的凸近似，并用 ℓ_1 -范数控制基数，由此得到上述鲁棒PCA问题的一个凸逼近，即

$$\begin{aligned} \underset{\boldsymbol{Y},\boldsymbol{Z}\in\mathbb{R}^{n\times m}}{\text{minimize}} \quad & \|\boldsymbol{Y}\|_*+\lambda\|\boldsymbol{Z}\|_1 \\ \text{subject to} \quad & \boldsymbol{X}=\boldsymbol{Y}+\boldsymbol{Z}, \end{aligned}$$

这里 $\|\boldsymbol{Z}\|_1$ 是矩阵 $\boldsymbol{Z}$ 中元素的绝对值之和，$\|\boldsymbol{Y}\|_*$ 是矩阵$\boldsymbol{Y}$的核范数(即$\boldsymbol{Y}$的奇异值之和)。上述问题是一个SDP，若问题规模适当，则是可解的；更大规模的实例需要用专门的算法来求解。

例13.10. (**视频中的背景建模**) 对视频数据而言，由于图像间的相关性，自然地可用低秩模型来对其进行建模。视频监测中最基本的算法功能之一是为场景中的背景变化建立好的模型。这个任务因为前景中出现的物体而变得复杂：比如在繁忙的背景中，每个图像中都可能出现一些异常现象。背景模型必须有足够强的适应性，以便适应背景中的变化，比如变化的光线。在这种情况下，很自然地就可用低秩近似来建模背景变化。前景中的物体(如汽车和行人)通常只占据图像像素的一小部分，因此可处理成稀疏误差。

我们考虑原始视频中的五幅图像，如图13.13第一行所示，其中包含一个过路人。每幅图的分辨率为 176×144 像素。我们首先将其分成三个频段(RGB)。对于每个频段，我们将每张图像排放成矩阵 $\mathbb{R}^{25344\times 5}$ 中的一列，然后通过鲁棒PCA法将 $\boldsymbol{X}$ 分解为一个低秩成分 $\boldsymbol{Y}$ 和稀疏成分 $\boldsymbol{Z}$ 。之后我们将三个频段重新结合到一起，形成两个图像集，其中一个为低秩成分，另一个为稀疏成分。如图13.13所示，低秩成分正确恢复了背景，同时稀疏成分正确地显示了正在移动的人。

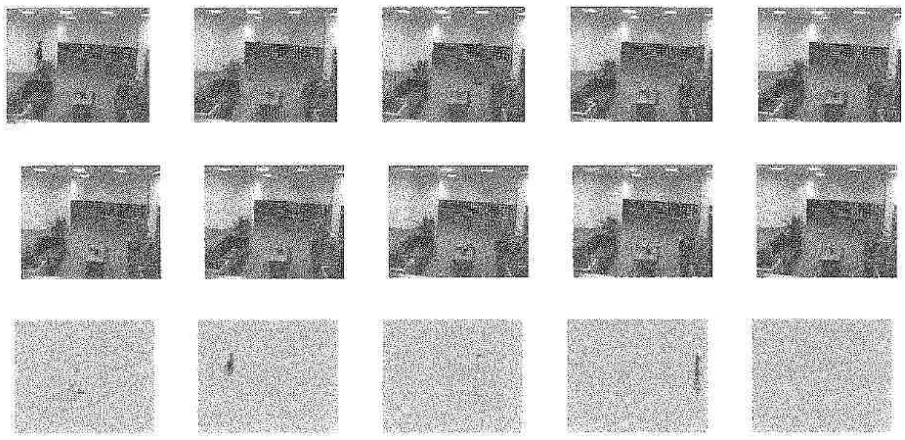

图 13.13: 视频帧集合：顶部为原始集合，中部为只显示背景的低秩成分，底部为只显示前景的稀疏成分

§13.5.5 稀疏图的高斯模型

稀疏图的高斯建模的目的是发现数据集中隐藏的图结构。基本的假设是数据由多维高斯分布生成，而我们的任务是由数据得到该分布的参数，这些参数能让模型显示出大量的条件独立性。为了让这些术语更清楚，我们将给出一些背景知识。

用高斯分布拟合数据. 设取值为$\mathbb{R}^n$中的向量的随机向量$\boldsymbol{\xi}$服从高斯分布，即它的概率密度函数形如

$$p(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^n \det(\boldsymbol{S})}} \mathrm{e}^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{S}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})},$$

其中参数 $\boldsymbol{\mu} \in \mathbb{R}^n$ ，$\boldsymbol{S}$ 是一个 $n \times n$ 的对称正定矩阵。这些参数都有自然的解释：$\boldsymbol{\mu}$ 是分布的平均值(期望)，$\boldsymbol{S}$ 是协方差矩阵。事实上进行积分，有

$$\mathbb{E}(\boldsymbol{\xi}) = \int \boldsymbol{x} p(\boldsymbol{x}) d\boldsymbol{x} = \boldsymbol{\mu},$$

$$\mathbb{E}\left((\boldsymbol{\xi}-\boldsymbol{\mu})(\boldsymbol{\xi}-\boldsymbol{\mu})^\top\right) = \int (\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^\top p(\boldsymbol{x}) d\boldsymbol{x} = \boldsymbol{S}.$$

给定数据集 $\boldsymbol{X} = [\boldsymbol{x}_1, ..., \boldsymbol{x}_m]$ ，定义**样本均值**(sample mean)

$$\bar{\boldsymbol{\mu}} \doteq \frac{1}{m} \sum_{i=1}^{m} \boldsymbol{x}_i$$

和**样本协方差阵**(sample covariance matrix)

$$\bar{\boldsymbol{S}} \doteq \frac{1}{m}\sum_{i=1}(\boldsymbol{x}_i-\bar{\boldsymbol{\mu}})(\boldsymbol{x}_i-\bar{\boldsymbol{\mu}})^{\top}.$$

下面在给定数据集 $\boldsymbol{X}$ 后，我们考虑用极大似然的思想来拟合高斯分布，即寻找参数 $\boldsymbol{\mu},\boldsymbol{S}$ 的估计值 $\hat{\boldsymbol{\mu}},\hat{\boldsymbol{S}}$ ，这些值使得似然函数取极大值. 为此，需要极大化密度 $p(\boldsymbol{x}_i), i=1,...,m$, 的乘积，对这个乘积取对数，得问题

$$\begin{aligned}\underset{\boldsymbol{\mu},\boldsymbol{S}}{\text{maximize}} \quad & -\tfrac{1}{2}\textstyle\sum_{i=1}^{m}(\boldsymbol{x}_i-\boldsymbol{\mu})^{\top}\boldsymbol{S}^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu})+\log\det(\boldsymbol{S})\\ \text{subject to} \quad & \boldsymbol{S}\succeq\boldsymbol{0}.\end{aligned}$$

记该问题的解为 $\boldsymbol{\mu}^*$和 $\boldsymbol{S}^*$，易验证 $\boldsymbol{\mu}^*=\bar{\boldsymbol{\mu}}$，即样本均值 $\bar{\boldsymbol{\mu}}$ 是总体均值 $\boldsymbol{\mu}$ 的极大似然估计. 此外，还有

$$\boldsymbol{S}^*=\arg\max_{\boldsymbol{S}\succ\boldsymbol{0}} \ \operatorname{trace}(\boldsymbol{S}^{-1}\bar{\boldsymbol{S}})-\log\det(\boldsymbol{S}), \tag{13.17}$$

其中 $\bar{\boldsymbol{S}}$ 是样本协方差阵. 问题(13.17)本身的表述方式不是凸优化，但通过变量替换 $\boldsymbol{S}\to\boldsymbol{P}=\boldsymbol{S}^{-1}$ ，可将其表述成关于 $\boldsymbol{P}$ 的问题，即

$$\max_{\boldsymbol{P}\succ\boldsymbol{0}} \ \operatorname{trace}(\boldsymbol{P}\bar{\boldsymbol{S}})+\log\det(\boldsymbol{P}). \tag{13.18}$$

该问题关于矩阵 $\boldsymbol{P}$ 是凸的。假定样本协方差矩阵 $\bar{\boldsymbol{S}}$ 正定，考虑该问题的最优性条件即可得到该问题的解，即得到 $\hat{\boldsymbol{P}}=\bar{\boldsymbol{S}}^{-1}$ 。即我们得到总体协方差阵的极大似然估计就是样本协方差阵，即 $\hat{\boldsymbol{S}}=\bar{\boldsymbol{S}}$。综上所述，在该情况下，极大似然方法得到的总体期望的估计是样本平均值，总体协方差阵的估计是样本协方差阵。

条件独立(conditional independence). 假设协方差阵 $\boldsymbol{S}$ 是可逆的. 称逆矩阵 $\boldsymbol{P}=\boldsymbol{S}^{-1}$ 是精度(precision)矩阵. 就**条件独立**(conditional independence)而言，精度矩阵的元素有一个很好的解释。如果固定除ξ_k和ξ_l之外的其余随机变量时，随机变量 ξ_k 和 ξ_l 独立，我们称这对随机变量 (ξ_k,ξ_l) 条件独立。这意味着对应的(条件)概率密度可被分解为两个标量分布的乘积，这两个标量分布分别以 x_k 和 x_l 为变量。现在假设存在随机变量对 (ξ_k,ξ_l) 使得精度矩阵中对应的 $P_{kl}=0$ 。上面给定的密度函数 p 满足

$$-2\ln p(\boldsymbol{x})=(\boldsymbol{x}-\boldsymbol{\mu})^{\top}\boldsymbol{S}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})=\sum_{i,j}P_{ij}(x_i-\mu_i)(x_j-\mu_j).$$

显然，当 $P_{kl}=0$ ，上述二次函数中没有一项会同时包含 x_k 和 x_l 。因此，当固定 $\boldsymbol{x}$ 中除 x_k 和 x_l 之外的其余变量后，我们可将上式写作

$$-2\ln p(\boldsymbol{x})=p_k(x_k)+p_l(x_l).$$

意即概率密度函数 p 可分解成两个函数的乘积，其中每个因子仅由 x_k 和 x_l 分别确定。

例13.11. 假设 $\mathbb{R}^3$ 中的高斯分布的均值为 $\mathbf{0}$，它的 3×3 精度矩阵 $\boldsymbol{P}$ 中有 $P_{12}=0$ 。让我们验证 x_1, x_2 是条件独立的。我们固定变量 x_3 ，观察到

$$\begin{aligned} -2\ln p(\boldsymbol{x}) &= P_{11}x_1^2 + P_{22}x_2^2 + 2P_{12}x_1x_2 + 2P_{13}x_1x_3 + 2P_{23}x_2x_3 + P_{33}x_3^2 \\ &= (P_{11}x_1^2 + 2P_{13}x_1x_3) + (P_{22}x_2^2 + 2P_{23}x_2x_3) + \text{constant}. \end{aligned}$$

如上所述，当 x_3 固定时，密度函数可写为两项乘积，每项分别只与 x_1, x_2 有关。

在仅基于观测数据来试图理解随机变量的关系时，条件独立是一个自然的概念。在实际操作中，众所周知的协方差的概念往往会失去作用。对于现实世界中的大部分数据集而言，所有的变量都是相关的，因此协方差阵是稠密的，这样很难从协方差阵看出有什么明显的结构。与此相反，许多变量对是条件独立的，这表现为对应的精度矩阵是稀疏的，该事实与其逆矩阵(协方差阵)稠密刚好相反。

上述事实促使我们去发现条件独立的结构，例如不同随机变量 $\xi_1, ..., \xi_n$ 间的图，在该图中任意一对条件独立的变量对 ξ_i, ξ_j 间没有边相连接。图越稀疏，就越易读。

下面的例子可以说明这种思想，即在揭示协方差矩阵无法呈现的结构时，精度矩阵是如何起作用的。

例13.12. (**条件独立的一个例子**) 假设我们在大量人群中观察三个随机变量 ξ_1, ξ_2, ξ_3 ，分别对应鞋码，是否白发，以及年龄。假设我们经计算得到了协方差阵的一个很好的估计。不出所料，这个协方差矩阵是稠密的，这意味着所有变量都是相关的。特别的，当年龄增长，鞋码也增加，白发出现的概率也会增加。

然而精度矩阵满足 $P_{12}=0$ ，这表明当年龄给定时，鞋码和是否有白发实际上是相互独立的。这说得通：当我们将相似年纪的人分为一个群体时，我们会发现对某一个年龄群体，鞋码和是否有白发间并没有统计联系。

稀疏的精度矩阵. 为了在极大似然估计的模型中增大条件独立性，我们修正凸问题(13.18)，即给其的目标函数增加一项关于 $\boldsymbol{P}$ 的 ℓ_1 -范数的惩罚，得到变形问题

$$\max_{\boldsymbol{P}\succ\boldsymbol{0}} \ \operatorname{trace}(\boldsymbol{P}\bar{\boldsymbol{S}}) + \log\det(\boldsymbol{P}) - \lambda\|\boldsymbol{P}\|_1, \tag{13.19}$$

其中 λ 是参数，它允许我们对模型的“拟合度”和精度矩阵的稀疏性进行折衷。与原问题(13.18)不同的是，上述惩罚问题没有解析解。但是它是凸的，存在算法能有效地求解它。

例13.13. (**利率数据图**) 在图13.14中，我们呈现了从协方差矩阵的逆中得到的利率(采样超过一年多)的相关结构。每个节点表示一个特定的利率期限(interest rate maturity)，如果协方差矩阵的逆中对应的系数非零，两个节点间有边相连，即它们条件相关。我们将 $\lambda = 0$ (此时就是样本协方差矩阵的逆)和 $\lambda = 1$ 时问题(13.19)的解相比较。在稀疏解中利率清晰地呈现出按期限聚类的现象。

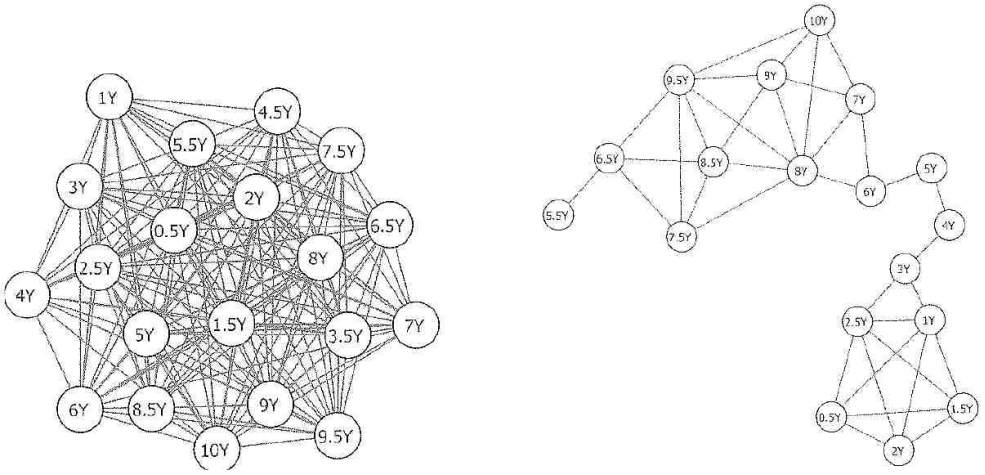

图 13.14: 利率数据的稀疏图模型

例13.14. (**议员的图**)(a graph of Senators) 回到例2.15中的参议院投票数据。图13.15显示了由数据得到的稀疏图，其对应于 $\lambda = 0.1$ 时问题(13.19)的

解。每当最优的稀疏协方差阵的逆 $\boldsymbol{P}$ 中对应的元素为零，图13.15中与此元素对应的两个节点(参议员)之间就没有边相连。

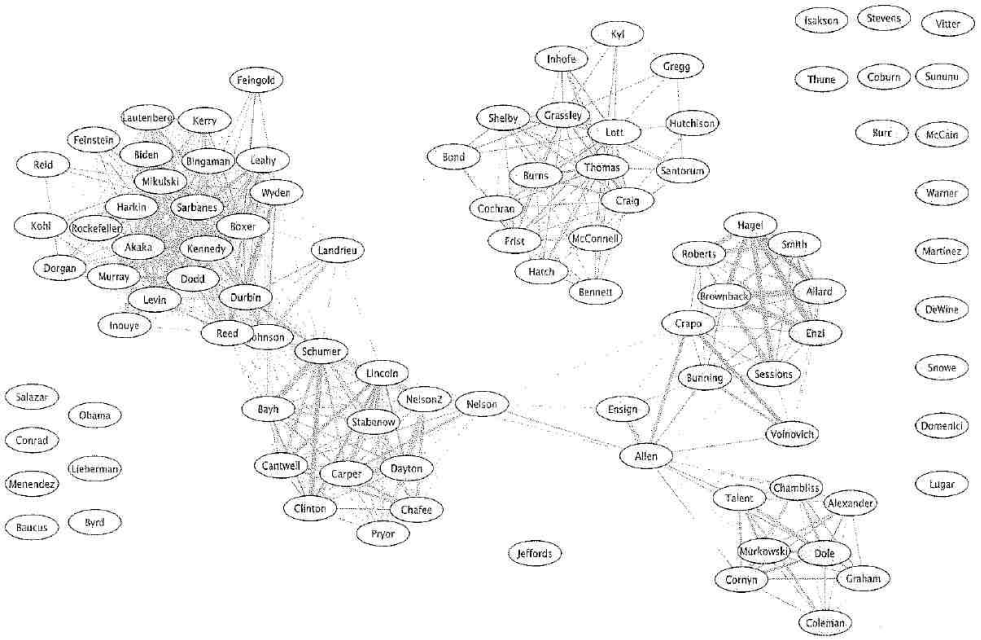

图 13.15: 参议院投票数据的稀疏图模型

§13.6 习题

习题13.1. (文本分析中的SVD) 考虑对应于大量新闻文章所构成集合的数据集，即假设给定 $n \times m$ 的术语文档(term-by-documents)矩阵 $\boldsymbol{X}$ 。确切地说，X 的元素 (i, j) 表示单词 i 在文档 j 中出现的次数。我们想用二维图像对数据集进行可视化。请详述你将如何完成以下事情(用 $\boldsymbol{X}$ 的适当中心化版本的SVD详述你的步骤)。

(a) 在单词空间内，将不同的新闻源作为点，画出方差最大的点。

(b) 在新闻源空间内，将不同的单词作为点，画出方差最大的点。

习题13.2. (**学习一个因素的模型**) 给定一个数据矩阵 $\boldsymbol{X} = [\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(m)}]$ ，其中 $\boldsymbol{x}^{(i)} \in \mathbb{R}^n, i = 1, ..., m$ 。假设已经将数据进行中心化处理，即 $\boldsymbol{x}^{(1)} +$

$\cdots + \boldsymbol{x}^{(m)} = \mathbf{0}$. 一种对协方差矩阵的(经验)估计[13]是

$$\boldsymbol{\Sigma} = \frac{1}{m}\sum_{i=1}^{m} \boldsymbol{x}^{(i)}\boldsymbol{x}^{(i)\top}.$$

在实践中，我们通常会发现上述对协方差矩阵的估计有噪声。去除噪声的一种办法是将协方差矩阵近似为 $\boldsymbol{\Sigma} \approx \lambda \boldsymbol{I} + \boldsymbol{F}\boldsymbol{F}^{\top}$ ，其中 $\boldsymbol{F}$ 是一个包含所谓“因素载荷”的 $n \times k$ 矩阵，$k \ll n$ 是因素的个数，$\lambda \geq 0$ 为“异质噪声”的方差。对应于这种近似方法的随机模型为

$$\boldsymbol{x} = \boldsymbol{F}\boldsymbol{f} + \sigma \boldsymbol{e},$$

其中 $\boldsymbol{x}$ 是中心化的观测(随机)向量；$\boldsymbol{f}$ 和 $\boldsymbol{e}$ 均是随机向量，二者的均值均为零，协方差矩阵均是单位矩阵；$\sigma = \sqrt{\lambda}$ 为异质噪声成分 $\sigma \boldsymbol{e}$ 的标准差。该随机模型可解释为观测是 k 个因素的线性组合(k 较小)和噪声部分，这个噪声对每个维度独立地产生影响。

为了用 $\boldsymbol{F}, \lambda$ 拟合数据，我们需要求解

$$\min_{\boldsymbol{F}, \lambda \geq 0} \|\boldsymbol{\Sigma} - \lambda \boldsymbol{I} - \boldsymbol{F}\boldsymbol{F}^{\top}\|_{\mathrm{F}}. \tag{13.20}$$

(a) 假设 λ 已知且小于 λ_k (经验协方差矩阵 σ 中第 k 大的特征值)。将最优解 $\boldsymbol{F}$ 表示为 λ 的函数，记作 $\boldsymbol{F}(\lambda)$ 。意即：你需要在固定 λ 的情况下解出 $\boldsymbol{F}$ 。

(b) 定义误差 $E(\lambda) = \|\boldsymbol{\Sigma} - \lambda \boldsymbol{I} - \boldsymbol{F}(\lambda)\boldsymbol{F}(\lambda)^{\top}\|_{\mathrm{F}}$，其中 $\boldsymbol{F}(\lambda)$ 是在(a)中得到的矩阵。证明

$$E(\lambda)^2 = \sum_{i=k+1}^{p} (\lambda_i - \lambda)^2.$$

确定极小化误差的最优解 λ 的显式表示，然后总结处理估计问题(13.20)的解决方案。

(c) 假设我们想估计数据空间中关于特定方向的风险(由方差度量)。回忆例4.2，当给定单位范数的 $n-$ 维向量 $\boldsymbol{w}$ 时，沿方向 $\boldsymbol{w}$ 的方差为 $\boldsymbol{w}^{\top}\boldsymbol{\Sigma}\boldsymbol{w}$ 。证明与使用 $\boldsymbol{\Sigma}$ 相比较，使用对 $\boldsymbol{\Sigma}$ 的 k 阶近似是定向风险的一个偏低估计。基于上述因素模型的近似将会怎样？请进行讨论。

习题13.3. (时间序列的运动预测) 我们有一个历史数据集，其包含时间序列 $r(1), ..., r(T)$ 的值。我们的目标是预测时间序列将会上升还是下降。

[13]参见例4.2.

基本想法是基于使用 n 个过往数据值(这里的 n 是固定的)的自回归模型输出的符号进行预测，即对 t 时刻 $r(t+1)-r(t)$ 的数值的符号预测为

$$\hat{y}_{\boldsymbol{w},b}(t)=\operatorname{sgn}(w_1 r(t)+\cdots+w_n r(t-n+1)+b),$$

其中 $\boldsymbol{w}\in\mathbb{R}^n$ 是我们的分类器的系数，b 是偏移项，$n\ll T$ 决定了我们将用多久远的数据来做预测。

(a) 第一步，我们将求解

$$\min_{\boldsymbol{w},b}\sum_{t=n}^{T-1}(\hat{y}_{\boldsymbol{w},b}(t)-y(t))^2.$$

其中 $y(t)=\operatorname{sgn}(r(t+1)-r(t))$。换句话说，我们尝试在最小二乘意义下将训练集的分类器作出的预测与观测到的事实进行匹配。上述问题是凸优化吗？如果不是，为什么？

(b) 如果要使用凸优化，请说明如何表述问题及训练分类器。确保精确地定义了学习的步骤、产生的优化问题中的变量，并说明如何求解优化问题及如何利用最优解来作出预测。

习题13.4. (PCA的一种变形) 回到例11.2中PCA的变形。使用你选定的(也可以是合成的)数据集来去比较经典PCA和这里所研究的变形，尤其研究二者对于异常数据的敏感度。确保建立尽可能严谨的评价体系。讨论你的结果。

习题13.5. (平方惩罚与非平方惩罚) 考虑问题

$$\begin{aligned}P(\lambda)\ :\ p(\lambda)\ &\doteq\ \min_{\boldsymbol{x}} f(\boldsymbol{x})+\lambda\|\boldsymbol{x}\|,\\ Q(\mu)\ :\ q(\mu)\ &\doteq\ \min_{\boldsymbol{x}} f(\boldsymbol{x})+\frac{1}{2}\mu\|\boldsymbol{x}\|^2,\end{aligned}$$

其中 f 为凸函数，$\|\cdot\|$ 为任一向量范数，参数 $\lambda>0,\mu>0$ 。假设对于给定的参数，相应的问题均有唯一解。

通常，对固定的 λ 和 μ ，上述两个问题的解是不重合的。这个练习将说明：我们可以遍历第一个问题的所有解，得到第二个问题的解集，反之亦然。

(a) 证明： p,q 均为凹函数，且定义为 $\tilde{q}(\mu)=q(1/\mu)$ 的函数在 $\mathbb{R}_+$ 上是凸的。

(b) 证明

$$p(\lambda) = \min_{\mu>0}\ q(\mu) + \frac{\lambda^2}{2\mu}, \quad q(\mu) = \max_{\lambda>0}\ p(\lambda) - \frac{\lambda^2}{2\mu}.$$

对第二个式子，你可假定 f 的定义域的内部非空。

(c) 从第一部分推出解的路径重合，即对于每个 $\lambda > 0$，如果我们已经求解了第一个问题，则对任一 $\mu > 0$，我们如上得到的最优点都将是第二个问题的最优点；反之亦然。不妨将 $P(\lambda)(Q(\mu))$ 的(惟一)解记作 $\boldsymbol{x}^*(\lambda)(\boldsymbol{z}^*(\mu))$。

(d) 关于第三个函数

$$r(\kappa)\ :\ r(\kappa) = \min\{f(\boldsymbol{x}) :\ \|\boldsymbol{x}\| \le \kappa\}$$

陈述并证明相似的结论.

(e) 如果去掉惟一解的假设，结果会怎样？

习题13.6. (基数惩罚最小二乘)考虑问题

$$\phi(\lambda) \doteq \min_{\boldsymbol{w}} \|\boldsymbol{X}^\top \boldsymbol{w} - \boldsymbol{y}\|_2^2 + \rho^2 \|\boldsymbol{w}\|\|_2^2 + \lambda \mathrm{card}(\boldsymbol{w}),$$

其中 $\boldsymbol{X} \in \mathbb{R}^{n\times m}, \boldsymbol{y} \in \mathbb{R}^m, \rho > 0$ 为正则化参数，$\lambda \ge 0$ 允许我们控制解向量的基数(非零元素的个数)。解向量的基数越小，对应的结果的可解释性越强。上述问题通常很难求解。在本练习中，我们将 $\boldsymbol{X}$ 的第 i 行记作 $\boldsymbol{a}_i^\top, i = 1, ..., n$，其对应于一个特定的“特征”(即变量 $\boldsymbol{w}$ 的维数)。

(a) 首先假设没有基数惩罚，即 $\lambda = 0$。证明

$$\phi(0) = \boldsymbol{y}^\top \left(\boldsymbol{I} + \frac{1}{\rho^2} \sum_{i=1}^n \boldsymbol{a}_i \boldsymbol{a}_i^\top \right)^{-1} \boldsymbol{y}.$$

(b) 现在考虑 $\lambda > 0$ 的情况。证明

$$\phi(\lambda) = \min_{\boldsymbol{u}\in\{0,1\}^n} \boldsymbol{y}^\top \left(\boldsymbol{I} + \frac{1}{\rho^2} \sum_{i=1}^m u_i \boldsymbol{a}_i \boldsymbol{a}_i^\top \right)^{-1} + \lambda \sum_{i=1}^n u_i.$$

(c) 将约束 $\boldsymbol{u} \in \{0,1\}^n$ 替换为区间约束 $\boldsymbol{u} \in [0,1]^n$，即可得到问题的一个自然的松弛。证明产生的下界 $\phi(\lambda) \ge \underline{\phi}(\lambda)$ 是凸问题

$$\underline{\phi}(\lambda)\ =\ \max_{\boldsymbol{v}}\ 2\boldsymbol{y}^\top \boldsymbol{v} - \boldsymbol{v}^\top \boldsymbol{v} - \sum_{i=1}^n \left(\frac{(\boldsymbol{a}_i^\top \boldsymbol{v})^2}{\rho^2} - \lambda \right)_+$$

的最优值。如何从上述问题的解 v^* 来恢复次优的稀疏模式？

(d) 将上述问题表示为一个SOCP。

(e) 写出该SOCP的对偶问题,并证明它可化简为

$$\underline{\phi}(\lambda) = \|\boldsymbol{X}^\top \boldsymbol{w} - \boldsymbol{y}\|_2^2 + 2\lambda \sum_{i=1}^{n} B\left(\frac{\rho x_i}{\sqrt{\lambda}}\right),$$

其中B是(凸的)**Hüber函数的反函数**(inverse Hüber function)：对于 $\xi \in \mathbb{R}$，

$$B(\xi) \doteq \frac{1}{2} \min_{0 \le z \le 1} \left(z + \frac{\xi^2}{z}\right) = \begin{cases} |\xi|, & \text{如果} |\xi| \le 1, \\ \frac{\xi^2+1}{2}, & \text{否则} \end{cases}$$

再次说明如何由上述问题的解 $\boldsymbol{w}^*$ 来恢复次优的稀疏模式?

(f) 一个处理基数惩罚的经典方法是将基数替换为 ℓ_1-范数，从而得到松弛的凸优化。将上述方法与 ℓ_1-范数松弛方法进行比较. 请对此进行讨论。

第十四章　计算金融

最优化在计算金融中扮演着越来越重要的角色. 因为H.Markowitz的开创性的工作[1]，金融领域的从业人员和研究者一直在致力于为策略性投资决策寻找合适的模型的研究。在本章中，我们给出了基于凸优化模型的金融应用领域的一个简单的综述. 在这些金融应用案例中，能效求解问题的重要工具是凸优化模型。

§14.1　单期投资组合最优化

已经在例2.6、例4.3和例8.19中介绍了一个基本的金融投资组合模型。在那个问题的背景下，向量$\boldsymbol{r} \in \mathbb{R}^n$ 包含了n种资产在固定周期Δ 内的随机收益率，$\boldsymbol{x} \in \mathbb{R}^n$描述了一个投资者给$n$种资产中的每一种所分配的美元价值量。在这一节中，我们假设投资者具有一个初始的投资组合$\boldsymbol{x}(0) \in \mathbb{R}^n$(所以这个投资者所拥有的初始总资产$w(0) = \sum_{i=1}^n x_i(0) = \mathbf{1}^\top \boldsymbol{x}(0)$)，他希望进行一些市场交易，从而更新了他的投资组合。交易量向量记为$\boldsymbol{u} \in \mathbb{R}^n$(如果给第$i$种资产增加投资，则$u_i > 0$，反之小于0)，从而更新投资组合后，投资组合变成

$$\boldsymbol{x} = \boldsymbol{x}(0) + \boldsymbol{u}.$$

我们进一步假设投资组合是**自我筹资的**(self financing)，意思是，总资产是守恒的(没有新的资金进入或者输出)，这个可以用**预算守恒**(budget conservation) 条件$\mathbf{1}^\top \boldsymbol{x} = \mathbf{1}^\top \boldsymbol{x}(0)$表示，即

$$\mathbf{1}^\top \boldsymbol{u} = 0.$$

我们不失一般性地进一步假设初始资产是一个单位，即$(w(0) = 1)$。有时，投资者的某种资产量可以是负的(即允许$\boldsymbol{x}$的某些元素是负的)。在金融操作中称这种情况为**卖空**(short-selling)，这种情况在实践中意味着从银行或者经纪人处借来资产，然后加以变卖。如果经纪人不允许该特征(或者我们不想去使用该操作)，我们给这个问题施加“无卖空”的限制条件$\boldsymbol{x} \geq \mathbf{0}$，即

$$\boldsymbol{x}(0) + \boldsymbol{u} \geq 0 \qquad \text{无卖空.}$$

[1]由于他将金融折衷决策问题表述成一个凸优化问题的贡献，他于1990年和M.Miller及W.Sharpe获得了诺贝尔经济学奖.

§14.1.1 均值-方差最优化

在一种以诺贝尔经济学奖获得者H.Markowitz命名的经典方法中，人们假设收益向量的期望$\hat{\boldsymbol{r}} = \mathbb{E}\{\boldsymbol{r}\}$和收益的协方差矩阵

$$\boldsymbol{\Sigma} = \mathbb{E}\{(\boldsymbol{r} - \hat{\boldsymbol{r}})(r - \hat{\boldsymbol{r}})^\top\}$$

是已知的，从而期末投资组合的收益就可以用随机变量

$$\rho(\boldsymbol{x}) = \boldsymbol{r}^\top \boldsymbol{x}$$

来描述，它的期望

$$\mathbb{E}\{\rho(\boldsymbol{x})\} = \hat{\boldsymbol{r}}^\top \boldsymbol{x},$$

方差

$$\mathrm{var}\{\rho(\boldsymbol{x})\} = \boldsymbol{x}^\top \boldsymbol{\Sigma} \boldsymbol{x}.$$

在Markowitz方法中，我们假设投资者偏好那些收益期望高的投资组合，同时他是规避**风险**(risk)的，这里的风险用投资组合的方差来度量(金融背景中一般叫做波动性)。为了得到最优的投资组合，或者以容许风险的上限为约束来极大化投资组合收益的期望，或者以收益的下限为约束来极小化风险。用μ表示想得到的期望收益的最小值，$\bar{\sigma}^2$是能接受的风险的最大值，则可将这两个问题表述为

$$\begin{aligned} \rho(\bar{\sigma}) = \quad & \underset{\boldsymbol{u}}{\text{maximize}} \quad && \hat{\boldsymbol{r}}^\top \boldsymbol{x} \\ & \text{subject to} && \boldsymbol{x}^\top \boldsymbol{\Sigma} \boldsymbol{x} \le \bar{\sigma}^2 \\ & && \boldsymbol{x} \in \mathcal{X} \end{aligned}$$

和

$$\begin{aligned} \bar{\sigma}^2(\mu) = \quad & \underset{\boldsymbol{u}}{\text{minimize}} \quad && \boldsymbol{x}^\top \boldsymbol{\Sigma} \boldsymbol{x} \\ & \text{subject to} && \hat{\boldsymbol{r}}^\top \boldsymbol{x} \ge \mu \\ & && \boldsymbol{x} \in \mathcal{X}, \end{aligned}$$

其中

$$\mathcal{X} = \{\boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{x} = \boldsymbol{x}(0) + \boldsymbol{u}, \mathbf{1}^\top \boldsymbol{x} = w(0), \mathrm{ns} \cdot \boldsymbol{x} \ge \mathbf{0}\}$$

这里的ns是一个标识，当不允许卖空时$\mathrm{ns} = 1$，否则，$\mathrm{ns} = 0$。因为$\boldsymbol{\Sigma} \succeq \mathbf{0}$，这两种表述得到的都是凸规划。在实践中，通常对一个递增的关于μ的序列求解第二种表述；或者对一个递减的关于$\bar{\sigma}$的序列求解第一种表述，由此得到风险－收益平面上的由最优值点组成的曲线，称该曲线是**有效前沿**(efficient frontier)。当投资组合$\boldsymbol{x}$使得$(\boldsymbol{x}^\top \boldsymbol{\Sigma} \boldsymbol{x}, \hat{\boldsymbol{r}}^\top \boldsymbol{x})$在这条曲线上时，称$\boldsymbol{x}$是有效的投资组合.

例14.1. (**交易所交易基金的分配**) 作为一个数值的例子，考虑关于分配$n = 7$时的投资组合分配：

1. SPDR道琼斯工业交易所交易基金，DIA；
2. iShare道琼斯交通交易所交易基金，IYT；
3. iShare道琼斯公共事业交易所交易基金，IDU；
4. 第一信托纳斯达克100前科技基金，QQXT；
5. SPDR欧元区斯托克50交易所交易基金，FEZ；
6. iShares巴克莱银行20+Yr Treas。债券交易所交易基金，TLT；
7. iSharesiBoxx USD高犹太人公司债券交易所交易基金，HYG。

我们使用2013年4月16号之前所观察到的300组每日收益数据，得到了收益向量的希望和这些资产的协方差矩阵的估计值[2]分别为

$$\hat{r} = \begin{bmatrix} 5.996 & 4.584 & 6.202 & 7.374 & 3.397 & 1.667 & 3.798 \end{bmatrix} \times 10^{-4},$$

$$\Sigma = \begin{bmatrix} 0.5177 & 0.596 & 0.2712 & 0.5516 & 0.9104 & -0.3859 & 0.2032 \\ 0.596 & 1.22 & 0.3602 & 0.7671 & 1.095 & -0.4363 & 0.2469 \\ 0.2712 & 0.3602 & 0.3602 & 0.2866 & 0.4754 & -0.1721 & 0.1048 \\ 0.5516 & 0.7671 & 0.2866 & 0.8499 & 1.073 & -0.4363 & 0.2303 \\ 0.9104 & 1.095 & 0.4754 & 1.073 & 2.563 & -0.8142 & 0.4063 \\ -0.3859 & -0.4363 & -0.1721 & -0.4363 & -0.8142 & 0.7479 & -0.1681 \\ 0.2032 & 0.2469 & 0.1048 & 0.2303 & 0.4063 & -0.1681 & 0.1478 \end{bmatrix} \times 10^{-4}.$$

我们在μ所在的区间$[3.54, 7.37] \times 10^{-4}$上以等间隔取$N = 20$个点，对于每个点对应的$\mu$值，我们求解ns = 1(即无卖空)时的问题(14.1)，由此我们得到了有效前沿的离散点近似，如图14.1 所示。比如，有效前沿上与数字13对应的投资组合为

$$\boldsymbol{x} = [0.0145 \quad 0.0014 \quad 54.79 \quad 31.68 \quad 0.0008 \quad 13.51 \quad 0.0067] \times 0.01.$$

这个投资组合中含54.8% 的IDU，31.7% 的QQXT，13.5%的TLT；对应的每日收益为0.0596%，标准差为0.494%.

[2]估计收益的期望和协方差矩阵是证券组合中非常精细的一步. 这里我们利用向后回溯的这300天的数据，将经验均值和经验协方差阵分别作为$\hat{\boldsymbol{r}}$ 和$\boldsymbol{\Sigma}$的估计值.

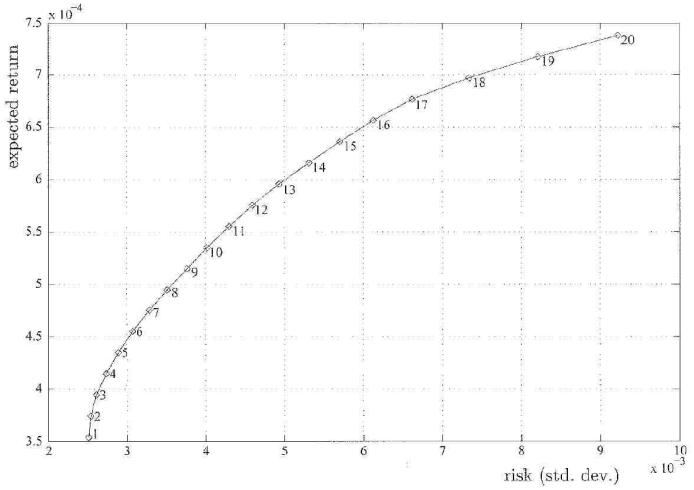

图 14.1: 有效前沿

§14.1.2 投资组合约束和交易成本

下面讨论关于投资组合的进一步的可能提出的设计约束.

§14.1.2.1 板块限界

我们可以给$\boldsymbol{x}$的某些元素施加上界，以限制投资仅仅局限于单个的资产。我们也可以限制在指定板块(比如说对应的前k个资产)中的投资总量不要超出总投资的α倍。给问题添加约束

$$\sum_{i=1}^{k} x_i \leq \alpha \mathbf{1}^\top \boldsymbol{x}$$

可以对此问题进行建模.

§14.1.2.2 多样性

我们还可以施加某些多样性约束。比如，我们可以施加对任何$k(k < n)$种资产的投资均不超过投资总量的η倍，即

$$s_k(\boldsymbol{x}) \doteq \sum_{i=1}^{k} x_{[i]} \leq \eta \mathbf{1}^\top \boldsymbol{x},$$

其中$x_{[i]}$指$\boldsymbol{x}$的第i大分量，所以$s_k(\boldsymbol{x})$是$\boldsymbol{x}$中前k大元素的总和，参见例9.10. 利用(9.16)中的表示，多样性约束可以表示为：存在标量t和$n-$维向量$\boldsymbol{s}$使得

$$kt + \mathbf{1}^\top \boldsymbol{s} \leq \eta \mathbf{1}^\top \boldsymbol{x},\ \boldsymbol{s} \geq \mathbf{0}, \boldsymbol{s} \geq \boldsymbol{x} - t\mathbf{1}.$$

§14.1.2.3 交易成本

我们在最优化模型中也可以考虑交易成本。特别的，我们易于包含形如

$$\phi(\boldsymbol{u}) = c\|\boldsymbol{u}\|_1$$

的正比例成本结构，其中$c \geq 0$是单位交易成本。如果存在交易成本，那么预算平衡条件就必须考虑到因交易而带来的开支：$\mathbf{1}^\top \boldsymbol{x}(0) - \phi(\boldsymbol{u}) = \mathbf{1}^\top \boldsymbol{x}$，即

$$\mathbf{1}^\top \boldsymbol{u} + \phi(\boldsymbol{u}) = 0$$

这个等式约束不是凸的。然而，如果$\hat{\boldsymbol{r}} \geq 0$，并且我们正在考虑问题(14.1)，则可以将这个方程等价地松弛成不等式(参见例8.19)，因此得到

$$\begin{array}{rll} \rho(\bar{\sigma}) = & \underset{\boldsymbol{u}}{\text{maximize}} & \hat{\boldsymbol{r}}^\top \boldsymbol{x} \\ & \text{subject to} & \boldsymbol{x}^\top \boldsymbol{\Sigma} \boldsymbol{x} \leq \bar{\sigma}^2 \\ & & \mathbf{1}^\top \boldsymbol{u} + c\|\boldsymbol{u}\|_1 \leq 0, \\ & & \boldsymbol{x} = \boldsymbol{x}(0) + \boldsymbol{u}, \\ & & \mathrm{ns} \cdot \boldsymbol{x} \geq \mathbf{0}. \end{array} \tag{14.1}$$

这是一个凸优化.

§14.1.3 夏普比率(SR)最优化

夏普比(Sharpe ratio, SR)[3]是回报对风险一种度量，它定量地表示了每单位风险的(无风险利率之外)收益的期望值所对应的数量。投资组合的SR定义为

$$\mathrm{SR}(\boldsymbol{x}) = \frac{\mathbb{E}\{\rho\} - r_f}{\sqrt{\mathrm{var}\{\rho\}}} = \frac{\hat{\boldsymbol{r}}^\top \boldsymbol{x} - r_f}{\sqrt{\boldsymbol{x}^\top \boldsymbol{\Sigma} \boldsymbol{x}}}$$

其中$r_f \geq 0$是**无风险**(risk-free)资产(例如把钱放在银行储蓄所获得的收益是0风险的)的收益，同时也和上面那样假设$w(0) = \mathbf{1}^\top \boldsymbol{x}(0) = 1$。就收益/风险角度来说，SR越高，表明这个投资越好。假设我们的投资组合没有考虑无风险资产，且在$\hat{\boldsymbol{r}}^\top \boldsymbol{x} > r_f, \boldsymbol{\Sigma} \succ \mathbf{0}$的条件下，从几何上来讲，极大化SR最大的投资组合对应着有效前沿与过$(0, r_f)$点的直线(资金分配线，capital allocation line, CAL)的切点，如图14.2 所示。

[3]以William Sharpe命名，Nobel laurerate(1990)

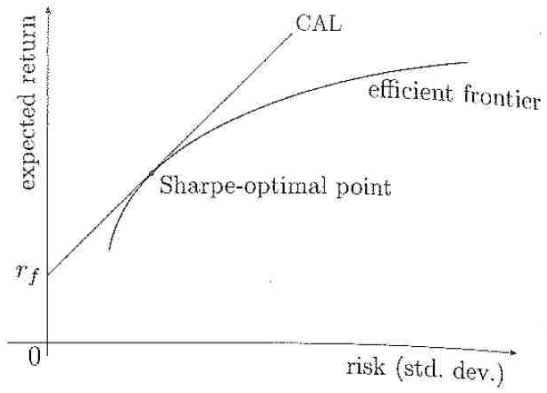

图 14.2: 资金分配线和Sharp最优点.

在如上的这些假设下，寻求夏普最优的投资组合可通过求解如下问题，即

$$\begin{array}{ll}\underset{\boldsymbol{u}}{\text{maximize}} & \frac{\hat{\boldsymbol{r}}^\top \mathcal{X} - r_f}{\sqrt{\boldsymbol{x}^\top \boldsymbol{\Sigma} \boldsymbol{x}}} \\ \text{subject to} & \hat{\boldsymbol{r}}^\top \boldsymbol{x} > r_f \\ & \boldsymbol{x} \in \mathcal{X}.\end{array} \tag{14.2}$$

此种形式的问题是非凸的。然而，我们可以进行如下操作：给SR($\boldsymbol{x}$)的分子和分母都乘上一个松弛变量$\gamma > 0$，并令

$$\tilde{\boldsymbol{x}} \doteq \gamma \boldsymbol{x} = \gamma \boldsymbol{x}(0) + \tilde{\boldsymbol{u}}; \quad \tilde{\boldsymbol{u}} \doteq \gamma \boldsymbol{u}.$$

然后，因为最大化SR($\boldsymbol{x}$)等价于最小化1/SR($\boldsymbol{x}$)是等价的(因为分子和分母都是正的)，我们就可以把问题重新写作

$$\begin{array}{ll}\underset{\tilde{\boldsymbol{u}}, \gamma > 0}{\text{minimize}} & \frac{\sqrt{\tilde{\boldsymbol{x}} \boldsymbol{\Sigma} \tilde{\boldsymbol{x}}}}{\hat{\boldsymbol{r}}^\top \tilde{\boldsymbol{x}} - \gamma r_f} \\ \text{subject to} & \hat{\boldsymbol{r}}^\top \tilde{\boldsymbol{x}} > \gamma r_f, \\ & \mathbf{1}^\top \tilde{\boldsymbol{u}} = 0, \\ & \tilde{\boldsymbol{x}} = \gamma \boldsymbol{x}(0) + \tilde{\boldsymbol{u}}, \\ & \mathrm{ns} \cdot \tilde{\boldsymbol{x}} \geq 0.\end{array}$$

注意到这个问题关于变量$(\tilde{\boldsymbol{u}}, \gamma)$是其次的，即如果$(\tilde{\boldsymbol{u}}, \gamma)$是一个解，那么$\alpha(\tilde{\boldsymbol{u}}, \gamma)$也是一个解（$\alpha > 0$）。我们通过规范化目标函数的分母为 1 来解决齐次

性，即我们施加$\hat{\boldsymbol{r}}^\top\tilde{\boldsymbol{x}}-\gamma r_f=1$. 使用这种规范化，上述问题就变成了

$$\begin{array}{ll}\underset{\tilde{\boldsymbol{u}},\gamma>0}{\text{minimize}} & \sqrt{\tilde{\boldsymbol{x}}\boldsymbol{\Sigma}\boldsymbol{x}}\\ \text{subject to} & \hat{\boldsymbol{r}}^\top\tilde{\boldsymbol{x}}-\gamma r_f=1,\\ & \mathbf{1}^\top\tilde{\boldsymbol{u}}=0,\\ & \tilde{\boldsymbol{x}}=\gamma\boldsymbol{x}(0)+\tilde{\boldsymbol{u}},\\ & \mathrm{ns}\cdot\tilde{\boldsymbol{x}}\geq 0.\end{array}$$

因为$\boldsymbol{\Sigma}\succ 0$，我们可以分解为$\boldsymbol{\Sigma}=\boldsymbol{\Sigma}^{1/2}\boldsymbol{\Sigma}^{1/2}$(参见§4.4.4)，因此$\sqrt{\tilde{\boldsymbol{x}}^\top\boldsymbol{\Sigma}\tilde{\boldsymbol{x}}}=\|\boldsymbol{\Sigma}^{1/2}\tilde{\boldsymbol{x}}\|_2$. 这样，SR的最优化问题就可以转化成形如

$$\begin{array}{ll}\underset{\tilde{\boldsymbol{u}},\gamma>0}{\text{minimize}} & t\\ \text{subject to} & \|\boldsymbol{\Sigma}^{1/2}\tilde{\boldsymbol{x}}\|_2\leq t,\\ & \hat{\boldsymbol{r}}^\top\tilde{\boldsymbol{x}}-\gamma r_f=1,\\ & \mathbf{1}^\top\tilde{\boldsymbol{u}}=0,\\ & \tilde{\boldsymbol{x}}=\gamma\boldsymbol{x}(0)+\tilde{\boldsymbol{u}},\\ & \mathrm{ns}\cdot\tilde{\boldsymbol{x}}\geq 0\end{array}\tag{14.3}$$

的SOCP问题. 一旦由这个问题解出$\tilde{\boldsymbol{u}},\gamma$，我们就可以由$\boldsymbol{u}=\gamma^{-1}\tilde{\boldsymbol{u}}$恢复出初始变量$\boldsymbol{u}$. 作为一个数值例子，在例14.1提供的数据之上求解问题(14.3)，这里假设无风险利率$r_f=1.5\times 10^{-4}$，则我们得到夏普最优的投资组合

$$\boldsymbol{x}=[2.71\quad 0.00\quad 30.95\quad 14.08\quad 0.00\quad 23.82\quad 28.44]^\top\times 0.01,$$

对应的收益的期望和标准差分别是 $0.0460*0.01$ 和$0.3122*0.01$. 这个投资组合在收益/风险有效前沿上，是CAL 直线与有效前沿的切点，如图14.3所示.

§14.1.4 风险价值(Value-at-Risk, VaR)最优化

假设收益向量$\boldsymbol{r}$服从正态分布，且已知均值$\hat{\boldsymbol{r}}$和协方差阵$\boldsymbol{\Sigma}$，所谓的投资组合$\boldsymbol{x}$的水平为$\alpha\in(0,1)$的**风险价值**(Value-at-Risk, VaR)定义为(参见例10.4)

$$\begin{aligned}\mathrm{VaR}_\alpha(\boldsymbol{x})=&\ -\sup\{\zeta:\mathrm{Prob}\{\rho(\boldsymbol{x})\leq\zeta\}\leq\alpha\}\\ =&\ -\inf\{\zeta:\mathrm{Prob}\{\rho(\boldsymbol{x})\leq\zeta\}\geq\alpha\}\\ =&\ -F_\rho^{-1}(\alpha)\end{aligned}$$

其中$F_{\rho^{-1}(\alpha)}$表示随机变量$\rho(\boldsymbol{x})$的累积分布函数的反函数，这里$\rho(\boldsymbol{x})$表示投资组合的收益，服从正态分布. 与经典的方差度量相比，投资组合的VaR更

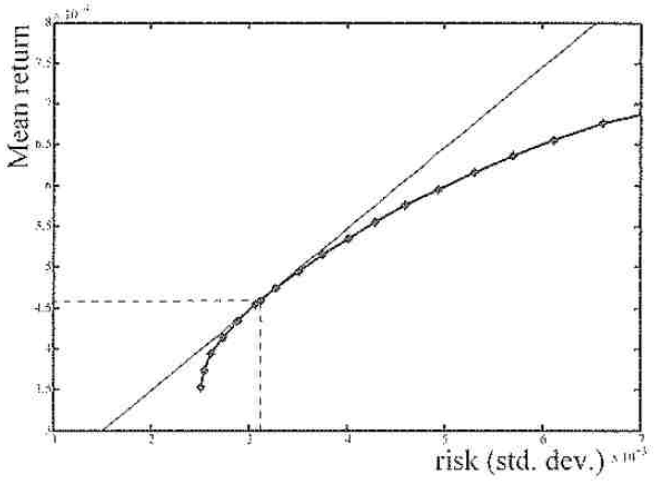

图 14.3: 当$r_f = 1.5 \times 10^{-4}$时Sharp最优的投资组合

受青睐，是一个被广泛采用的风险度量。VaR_α的意思是投资导致高于VaR_α的**损失**(loss)[4]的概率不大于α. 等价地说，投资者的收益高于$-\mathrm{VaR}_\alpha$的概率不低于$1-\alpha$ 。在例10.4中我们已经看到了，对于给定的$\alpha \in (0, 0.5)$，

$$\mathrm{VaR}_\alpha \le \gamma \Leftrightarrow \phi^{-1}(1-\alpha)\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{x}\|_2 \le \gamma + \hat{\boldsymbol{r}}^\top \boldsymbol{x}, \tag{14.4}$$

其中Φ是标准正态累积分布函数。在保证关于VaR_α的上界γ的限制下，最大化收益的期望即可得到最优的投资组合. 求解形如

$$\begin{array}{rll} \rho(\gamma) = & \text{maximize}_{\boldsymbol{x}} & \hat{\boldsymbol{r}}^\top \boldsymbol{x} \\ & \text{subject to} & \Phi^{-1}(1-\alpha)\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{x}\|_2 \le \hat{\boldsymbol{r}}^\top \boldsymbol{x} + \gamma, \\ & & \boldsymbol{x} \in \mathcal{X}. \end{array} \tag{14.5}$$

的SOCP可以得到这个最优的投资组合。

§14.2 鲁棒的投资组合最优化

前面所说的投资组合优化模型是建立在随机收益向量的矩(例如：期望和协方差矩阵)或整个分布是精确已知的假设条件下。然而，在实践中，这些量是不知道的，可能需要结合历史数据和专家知识去估计。因此，由于估计误差和/或在描述市场收益的随机过程性质的错误**先验**(priori)假设，将导致这些估计量具有很高的不确定性。反过来，由使用“标称”(nominal)数

[4]我们的**损失**(loss)ℓ 的意思是收益的相反数，即$\ell = -\rho$. 这样，用我们的记号，VaR_α是一个典型的正值，描述了投资者(以概率$1-\alpha$) 所面对的最大的损失.

据的最优化问题所得到的投资组合分配决策远远偏离实践中的最优策略，即最优的投资组合对输入数据非常敏感。克服这些问题的一种途径是一定程度上考虑这些数据存在的不确定性的**先验**(priori)信息，并寻求“鲁棒”(rubust)的投资组合，其对于不确定性而言，在最坏情况下也表现良好。

§14.2.1 鲁棒的均值-方差最优化

用鲁棒的均值方差投资组合优化方法中，我们假设$\hat{\boldsymbol{r}}$和$\boldsymbol{\Sigma}$是不确定的，我们通过为这些参数确定隶属合集来建模这种不确定性，即我们假设$(\hat{\boldsymbol{r}}, \boldsymbol{\Sigma})$属于某一给定的有界不确定集$\mathcal{U}$。对于一个给定的投资组合$\boldsymbol{x}$，最坏情况的投资组合的方差

$$\sigma^2_{\mathrm{WC}} = \sup_{(\hat{\boldsymbol{r}},\boldsymbol{\Sigma})\in\mathcal{U}} \quad \boldsymbol{x}^\top \boldsymbol{\Sigma} \boldsymbol{x}.$$

最小方差投资组合设计问题(14.1)的鲁棒化版本是

$$\begin{array}{rll} \bar{\sigma}^2_{\mathrm{WC}}(\mu) = & \underset{\boldsymbol{u}}{\text{minimize}} & \sup_{(\hat{\boldsymbol{r}},\boldsymbol{\Sigma})\in\mathcal{U}} \boldsymbol{x}^\top \boldsymbol{\Sigma} \boldsymbol{x} \\ & \text{subject to} & \inf_{(\hat{\boldsymbol{r}},\boldsymbol{\Sigma})\in\mathcal{U}} \hat{\boldsymbol{r}}^\top \boldsymbol{x} \geq \mu, \\ & & \boldsymbol{x} \in \mathcal{X}. \end{array} \tag{14.6}$$

只有在不确定集$\mathcal{U}$足够“简单”时，才能高效准确地求解这个问题。比如$\mathcal{U} = \{(\hat{\boldsymbol{r}}, \boldsymbol{\Sigma}) : \hat{\boldsymbol{r}} \in \mathcal{U}_r, \boldsymbol{\Sigma} \in \mathcal{U}_{\boldsymbol{\Sigma}}\}$，其中$\mathcal{U}_r$和$\mathcal{U}_{\boldsymbol{\Sigma}}$是区间集，即

$$\mathcal{U}_r = \{\hat{\boldsymbol{r}} : r_{\min} \leq \hat{\boldsymbol{r}} \leq r_{\max}\},$$

$$\mathcal{U}_{\boldsymbol{\Sigma}} = \{\boldsymbol{\Sigma} : \boldsymbol{\Sigma}_{\min} \preceq \boldsymbol{\Sigma} \preceq \boldsymbol{\Sigma}_{\max}, \boldsymbol{\Sigma} \succeq \mathbf{0}\}.$$

如果进一步假定$\boldsymbol{x} \geq \mathbf{0}$ 和$\boldsymbol{\Sigma}_{\max} \succeq \mathbf{0}$，则问题(14.6)等价于

$$\begin{array}{rll} \bar{\sigma}^2_{\mathrm{WC}}(\mu) = & \underset{\boldsymbol{u}}{\text{minimize}} & \boldsymbol{x}^\top \boldsymbol{\Sigma}_{\max} \boldsymbol{x} \\ & \text{subject to} & \hat{\boldsymbol{r}}_{\min}^\top \boldsymbol{x} \geq \mu, \\ & & \mathbf{1}^\top \boldsymbol{u} = 0, \\ & & \boldsymbol{x} = \boldsymbol{x}(0) + \boldsymbol{u}, \\ & & \boldsymbol{x} \geq \mathbf{0}. \end{array} \tag{14.7}$$

如果没有施加约束$\boldsymbol{x} \geq \mathbf{0}$，可以用一个**特定的鞍点**(ad-hoc saddle-point)算法来高效求解问题(14.6)。该算法超出本节内容，不予详述[5].

[5]感兴趣的读者可以参考Tütüncü和Koenig 的文章: Robust asset allocation, Annals of Operations Research, 2004.

情景法. 对于一般的不确定集$\mathcal{U}_r,\mathcal{U}_{\boldsymbol{\Sigma}}$及与之相对应的问题，我们无法找到一个有效的方式来表述鲁棒性，如§10.3.4 描述的那样，我们可以用情景法来逼近鲁棒性。在当前的背景下，我们应该假设定义在$\mathcal{U}_r,\mathcal{U}_{\boldsymbol{\Sigma}}$ 上的概率分布，并收集了N组独立同分布样本$(\hat{\boldsymbol{r}}^{(i)},\boldsymbol{\Sigma}^{(i)}), i=1,...,N$，然后我们求解情景问题

$$\begin{array}{ll}\underset{\boldsymbol{x},t}{\text{minimize}} & t \\ \text{subject to} & \boldsymbol{x}^\top\boldsymbol{\Sigma}^{(i)}\boldsymbol{x}\le t, i=1,...,N, \\ & \hat{\boldsymbol{r}}^{(i)T}\boldsymbol{x}\ge\mu, \quad i=1,...,N, \\ & \boldsymbol{x}\in\mathcal{X}.\end{array} \tag{14.8}$$

如果这个情景问题中的N与式(10.30)所描述的相匹配，则求解这个问题可以得到一个具有理想水平的概率鲁棒的证卷组合。这种方法的优点是适用于一般的不确定性结构，并且从类别上来说不会增加求解问题的难度，即“鲁棒化”问题(14.8)仍然是一个凸优化问题。

§14.2.2 鲁棒的风险价值(VaR)最优化

我们在§14.1.4已经看到，当假设收益向量$\boldsymbol{r}$服从均值和协方差矩阵分别为$\hat{\boldsymbol{r}}$和$\boldsymbol{\Sigma}$的正态分布时，式(14.4)给出了投资组合$\boldsymbol{x}$的水平为$\alpha\in(0,0.5)$的风险价值。然而，在实践中，我们可能会遇到两种不确定性来源，其妨碍我们直接使用这个公式。第一种是实际收益的分布不是精确的正态分布(分布不确定性)；第二种是参数$\hat{\boldsymbol{r}},\boldsymbol{\Sigma}$ 的值可能不准确(参数不确定性)。接下来我们讨论在这两种不确定性下如何处理VaR投资组合最优化问题。

§14.2.2.1 分布不确定性下的鲁棒VaR

首先，我们假设收益分布的参数$\hat{\boldsymbol{r}},\boldsymbol{\Sigma}$是精确已知的，但分布本身未知。然后我们考虑一类均值为$\hat{\boldsymbol{r}}$，协方差阵为$\boldsymbol{\Sigma}$ 的概率分布$\mathcal{P}$。最坏的情况下，损失大于某水平ζ的概率是

$$\text{Prob}_{\text{WC}}\{-\rho(\boldsymbol{x})\ge\zeta\}\doteq\sup_{\mathcal{P}}\text{Prob}\{-\rho(\boldsymbol{x})\ge\zeta\}, \tag{14.9}$$

其中上确界是关于$\mathcal{P}$中的所有分布来计算的。最坏情况的$\alpha-$VaR定义为

$$\text{WC-VaR}_\alpha(\boldsymbol{x})\doteq\sup\{\zeta:\ \text{Prob}_{\text{WC}}\{-\rho(\boldsymbol{x})\ge\zeta\}\le\alpha\},$$

显然有

$$\text{WC-VaR}_\alpha(\boldsymbol{x})\le\gamma \quad\Leftrightarrow\quad \text{Prob}_{\text{WC}}\{-\rho(\boldsymbol{x})\ge\gamma\}\le\alpha. \tag{14.10}$$

而Chebyshev-Cantelli不等式表明，对于任意方差有限的随机变量z及$t>0$，有

$$\sup\ \mathrm{Prob}\{z-\mathbb{E}\{z\}\geq t\}=\frac{\mathrm{var}(z)}{\mathrm{var}(z)+t^2},$$

其中的上确界是对所有具有相同的均值和相同的协方差的分布来计算的。现在，考虑(14.9)中的随机变量$\rho(\boldsymbol{x})$，它的期望为$\hat{\boldsymbol{r}}^\top\boldsymbol{x}$，方差为$\left\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{x}\right\|_2^2$，将Chebyshev-Cantelli 不等式应用于$\rho(\boldsymbol{x})$，我们有

$$\begin{aligned}\mathrm{Prob}_{\mathrm{WC}}\{-\rho(\boldsymbol{x})\geq\zeta\}\quad&=\mathrm{Prob}_{\mathrm{WC}}\{-\rho(\boldsymbol{x})+\hat{\boldsymbol{r}}^\top\boldsymbol{x}\geq\zeta+\hat{\boldsymbol{r}}^\top\boldsymbol{x}\}\\&=\frac{\left\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{x}\right\|_2^2}{\left\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{x}\right\|_2^2+(\zeta+\hat{\boldsymbol{r}}^\top\boldsymbol{x})^2}.\end{aligned}$$

因此，由(14.10)可知，对于$\alpha\in(0,0.5)$,

$$\text{WC-VaR}_\alpha(\boldsymbol{x})\leq\gamma\ \Leftrightarrow\quad\kappa(\alpha)\left\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{x}\right\|_2\leq\hat{\boldsymbol{r}}^\top\boldsymbol{x}+\gamma,$$

其中$\kappa(\alpha)\doteq\sqrt{\frac{1-\alpha}{\alpha}}$. 如上，在VaR优化问题(14.5)中，将SOC 约束的系数$\Phi^{-1}(1-\alpha)$替换成$\kappa(\alpha)$，即得对应的分布不确定性下的鲁棒VaR优化问题。

§14.2.2.2 矩不确定性下的鲁棒VaR

除了分布不确定性外，如果参数$\hat{\boldsymbol{r}}$和$\boldsymbol{\Sigma}$的取值也具有不确定性，那么我们可以用如下方法[6] 来确定鲁棒的VaR 证券组合。假设描述$\hat{\boldsymbol{r}},\boldsymbol{\Sigma}$的不确定性的集合是区间型的，即

$$\begin{aligned}\mathcal{U}_{\boldsymbol{r}}=&\ \{\hat{\boldsymbol{r}}:\boldsymbol{r}_{\min}\leq\hat{\boldsymbol{r}}\leq\boldsymbol{r}_{\max}\},\\\mathcal{U}_{\boldsymbol{\Sigma}}=&\ \{\boldsymbol{\Sigma}:\boldsymbol{\Sigma}_{\min}\preceq\boldsymbol{\Sigma}\preceq\boldsymbol{\Sigma}_{\max}\},\end{aligned}$$

并且存在一个$\boldsymbol{\Sigma}\in\mathcal{U}_{\boldsymbol{\Sigma}}$使得$\boldsymbol{\Sigma}\succ\mathbf{0}$。然后，定义$\text{WC-VaR}_\alpha(\boldsymbol{x})$ 是$\mathrm{VaR}_\alpha(\boldsymbol{x})$关于所有$\hat{\boldsymbol{r}}\in\mathcal{U}_r,\boldsymbol{\Sigma}\in\mathcal{U}_{\boldsymbol{\Sigma}}$，以及所有均值为$\hat{\boldsymbol{r}}$，协方差阵为$\boldsymbol{\Sigma}$ 的分布取上确界，可以证明$\text{WC-VaR}_\alpha(\boldsymbol{x})\leq\gamma$ 当且仅当存在对称矩阵$\boldsymbol{\Lambda}_+,\boldsymbol{\Lambda}_-\in\mathbb{S}_+^n$，向量$\boldsymbol{\lambda}_+,\boldsymbol{\lambda}^-\in\mathbb{R}^n$和标量$v\in\mathbb{R}$,使得

$$\begin{gathered}\mathrm{trace}(\boldsymbol{\Lambda}_+\boldsymbol{\Sigma}_{\max})-\mathrm{trace}(\boldsymbol{\Lambda}_-\boldsymbol{\Sigma}_{\min})+k^2(\alpha)v+\boldsymbol{\lambda}_+^\top\boldsymbol{r}_{\max}-\boldsymbol{\lambda}_-^\top\boldsymbol{r}_{\min}\leq\gamma,\\\begin{bmatrix}\boldsymbol{\Lambda}_+-\boldsymbol{\Lambda}_- & \boldsymbol{x}/2\\ \boldsymbol{x}^\top/2 & v\end{bmatrix}\succeq\mathbf{0},\\\boldsymbol{\lambda}_+\geq\mathbf{0},\boldsymbol{\lambda}_-\geq\mathbf{0},\boldsymbol{\Lambda}_+\succeq\mathbf{0},\boldsymbol{\Lambda}_-\succeq\mathbf{0},\\\boldsymbol{x}=\boldsymbol{\lambda}_--\boldsymbol{\lambda}_+.\end{gathered}\tag{14.11}$$

[6]关于该方法的描述详见论文：EI Ghaoui, Oks and Oustry, Worst-case value-at-risk and robust portfolio optimization: a conic programming approach, Operations Research, 2003.

在分布和矩的双重不确定性下，与VaR优化问题(14.5)对应的鲁棒VaR优化可以显式地表述成如下的SDP，即

$$\begin{array}{ll}\underset{\boldsymbol{u},\boldsymbol{\Lambda}_+,\boldsymbol{\Lambda}_-,\boldsymbol{\lambda}_+,\boldsymbol{\lambda}_-,v}{\text{minimize}} & \hat{\boldsymbol{r}}^\top \boldsymbol{x} \\ \text{subject to} & (14.11), \\ & \boldsymbol{x} \in \mathcal{X}.\end{array}$$

§14.3 多期投资组合最优化

前两节中的投资分配模型仅仅考虑一个投资期，因此关注的是投资组合在投资期末的某种业绩。然而，在实践中，一名投资者在本投资期结束时，将再次面临如何在下一个周期重新分配他的资产的问题。当然，他可以再次求解单一投资期的分配问题，然后依次重复下去。然而，从长远目标来看，这种基于按期策略的决策有可能是“目光短浅的”(myopic)，即如果投资者提前知道他的投资目标是设定在自此向前的$T(\geq 1)$个投资期，投资者一开始就考虑自己投资问题的多周期性有利于得到更有远见的策略。

在接下来的几节，我们将讨论两种在多期投资组合上用以优化投资决策的**数据驱动**(data-driven)技术。第一种技术我们称之为**开环**(open-loop)方法，其目标是在时刻$k = 0$ 确定整个未来的投资组合调整量，这包括从时刻 0 到 最终时刻 T 上的投资组合调整量序列。第二种技术我们称之为**闭环**(colsed-loop)策略，它增加了未来决策的灵活性，从而可看成第一种策略的提升. 具体地，它将未来决策作为观测收益的函数(所谓的**方法**)，由此以减少未来不确定性的影响。两种方法都是数据驱动的，即它们是基于收益流的N种仿真情景的可得性。

我们从一些记号和准备性工作开始：资产集合记为$a_1, \ldots, a_n$，时刻$k\Delta$时资产a_i的市场价格记为$p_i(k)$，其中k是一个带符号的整数，Δ 是一个固定的时间段。对资产i的投资在第k期(从$(k-1)\Delta$到$k\Delta$)的简单收益是

$$r_i(k) \doteq \frac{p_i(k) - p_i(k-1)}{p_i(k-1)}, i = 1, \ldots, n;\ k = 1, 2, \ldots,$$

对应的**增益**(gain)定义为

$$g_i(k) \doteq 1 + r_i(k), i = 1, \ldots, n;\ k = 1, 2, \ldots$$

第k期资产的收益向量记为$\boldsymbol{r}(k) \doteq [r_1(k) \cdots r_n(k)]^\top$，$\boldsymbol{g}(k)$是对应的增益向量。符号$\boldsymbol{G}(k) = \operatorname{diag}(\boldsymbol{g}(k))$ 表示对角线元素是$\boldsymbol{g}(k)$的分量的对角矩阵。假

定收益和增益向量是随机向量。我们用$T \geq 1$表示需要做分配决策的正向周期数目，并假设有一个**场景生成先哲**(scenario-generating oracle)，它能产生正向收益流$\{\boldsymbol{r}(1), \ldots, \boldsymbol{r}(T)\}$的$N$个独立同分布的样本。

§14.3.1 投资组合动力学

我们考虑T个时段(或**阶段**(stage))的决策问题。在每个时段我们有机会重新平衡我们的投资组合，在保证满足每个时段的的投资组合约束的前提下，使得最后一个时段的某种恰当的成本函数(稍后讨论)取最小值。如图14.4所示，第k 个决策时段从时刻$k-1$开始，到时刻k结束。

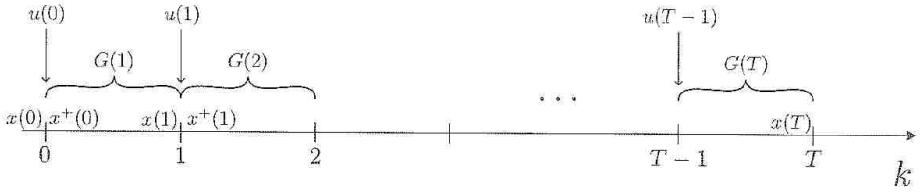

图 14.4: 投资决策水平和时段

我们用$x_i(k)$表示投资者时刻k在有价证券a_i上的投资占总资产的比例。时刻k的投资组合向量

$$\boldsymbol{x}(k) \doteq [x_1(k) \quad \cdots \quad x_n(k)]^\top.$$

时刻k投资者的总资产

$$w(k) \doteq \sum_{i=1}^{n} x_i(k) = \mathbf{1}^\top \boldsymbol{x}(k).$$

令$\boldsymbol{x}(0)$是时刻$k = 0$时给定的初始投资组合布局(例如，可以假设$\boldsymbol{x}(0)$仅一个分量非零，其代表初始的可用资金数目)。在时刻$k = 0$时，我们有机会在市场中进行交易，并且因此可以增加或减少对每项资产的投资数目，即调整投资组合。在交易过后，调整后的投资组合是$\boldsymbol{x}^+ = \boldsymbol{x}(0) + \boldsymbol{u}(0)$，如果我们增加了对第$i$个资产的投资，则$u_i > 0$；若减少了，则$u_i < 0$；若不变，则$u_i = 0$。现在假设投资组合在第一时段$\Delta$内固定不变。在这个时段结束时，投资组合布局是

$$\boldsymbol{x}(1) = \boldsymbol{G}(1)\boldsymbol{x}^+(0) = \boldsymbol{G}(1)\boldsymbol{x}(0) + \boldsymbol{G}(1)\boldsymbol{u}(0),$$

其中$\boldsymbol{G}(1) = \mathrm{diag}(g_1(1), \ldots, g_n(1))$是对角矩阵，对角线元素是时刻 0 到时刻 1 这个时段的资产增益。在时刻$k = 1$，我们再次对投资组合调整$\boldsymbol{u}(1)$，

即$\boldsymbol{x}^+(1)=\boldsymbol{x}(1)+\boldsymbol{u}(1)$，然后在紧接着的这个时段$\Delta$内持有这个更新后的投资组合。因此，时刻$k=2$ 的投资组合布局是

$$\boldsymbol{x}(2)=\boldsymbol{G}(2)\boldsymbol{x}^+(1)=\boldsymbol{G}(2)\boldsymbol{x}(1)+\boldsymbol{G}(2)\boldsymbol{u}(1).$$

对$k=0,1,2,\ldots$,依次按这种方式作用，我们得到了第$(k+1)$时段末的投资组合布局的迭代动力学方程

$$\boldsymbol{x}(k+1)=\boldsymbol{G}(k+1)\boldsymbol{x}(k)+\boldsymbol{G}(k+1)\boldsymbol{u}(k),k=0,\ldots,T-1 \tag{14.12}$$

和刚经过第$(k+1)$次交易后的投资组合布局方程(参见图14.4)

$$\boldsymbol{x}^+(k)=\boldsymbol{x}(k)+\boldsymbol{u}(k). \tag{14.13}$$

由(14.12)可得时刻$k=1,\ldots,T$的(随机的)投资组合布局是

$$\boldsymbol{x}(k)=\boldsymbol{\Phi}(1,k)\boldsymbol{x}(0)+\sum_{j=1}^{k}\boldsymbol{\Phi}(j,k)\boldsymbol{u}(j-1), \tag{14.14}$$

其中我们定义$\boldsymbol{\Phi}(v,k)(v\le k)$是从时段$v$开始到时段$k$结束的**复合增益**(compounded gain) 矩阵，即

$$\boldsymbol{\Phi}(v,k)\doteq\boldsymbol{G}(k)\boldsymbol{G}(k-1)\cdots\boldsymbol{G}(v),\quad\boldsymbol{\Phi}(k,k)\doteq\boldsymbol{G}(k).$$

投资组合的表达式可以重新紧促地重写为

$$\boldsymbol{x}(k)=\boldsymbol{\Phi}(1,k)\boldsymbol{x}(0)+\boldsymbol{\Omega}_k\boldsymbol{u},$$

其中

$$\begin{aligned}\boldsymbol{u}&\doteq[\boldsymbol{u}(0)^\top\ \cdots\ \boldsymbol{u}(T-2)^\top\ \boldsymbol{u}(T-1)^\top]^\top,\\ \boldsymbol{\Omega}_k&\doteq[\boldsymbol{\Phi}(1,k)\cdots\boldsymbol{\Phi}(k-1,k)\ \ \boldsymbol{\Phi}(k,k)|0\quad\cdots\quad 0].\end{aligned}$$

因此，我们的总资产

$$w(k)=\mathbf{1}^\top\boldsymbol{x}(k)=\boldsymbol{\phi}(1,k)^\top\boldsymbol{x}(0)+\boldsymbol{w}_k^\top\boldsymbol{u}$$

其中

$$\begin{aligned}\boldsymbol{\phi}(v,k)^\top&\doteq\mathbf{1}^\top\boldsymbol{\Phi}(v,k),\\ \boldsymbol{w}_k^\top&\doteq\mathbf{1}^\top\boldsymbol{\Omega}_k=[\boldsymbol{\phi}(1,k)^\top\ \cdots\ \boldsymbol{\phi}(k-1,k)^\top\ \boldsymbol{\phi}(k,k)^\top|\ 0\ \cdots\ 0].\end{aligned}$$

我们考虑自筹资金的投资组合，即

$$\sum_{i=1}^{n} u_i(k) = 0,\ k = 0, 1, \ldots, T-1,$$

并且通过强迫更新后的投资组合$\boldsymbol{x}^+(k)$属于多面体$\mathcal{X}(k)$允许我们在模型中包含一般的线性约束。投资在整体范围内的累积总收益是

$$\rho(\boldsymbol{u}) \doteq \frac{w(T)}{w(0)} = \frac{\mathbf{1}^\top \boldsymbol{x}(T)}{\mathbf{1}^\top \boldsymbol{x}(0)} = \frac{\boldsymbol{\phi}(1,T)^\top \boldsymbol{x}(0)}{\mathbf{1}^\top \boldsymbol{x}(0)} + \frac{1}{\mathbf{1}^\top \boldsymbol{x}(0)} \boldsymbol{w}_T^\top \boldsymbol{u}.$$

我们看到$\rho(\boldsymbol{u})$是决策变量$\boldsymbol{u}$的仿射函数，系数是随机向量$\boldsymbol{w}_T$，其依赖于T个时段上的随机增益。

§14.3.2 最优的开环策略

现在假设可以从场景生成先哲那儿得到各时段增益的N个独立同分布样本(场景)$\{\boldsymbol{G}^{(i)}(k), k = 1, \ldots, T\}$，$i = 1, \ldots, T$。这些样本因此对每个矩阵$\boldsymbol{\Omega}_k$，$k = 1, \ldots, T$，进而对$\boldsymbol{\omega}(k)$ 和$\boldsymbol{\phi}(1,k)$等向量依次产生了N 个场景。我们分别用$\boldsymbol{\Omega}_k^{(i)}, \boldsymbol{\omega}_k^{(i)}, \boldsymbol{\phi}^{(i)}(1,k)$表示这些场景，用$\boldsymbol{x}^{(i)}(k), \boldsymbol{w}^{(i)}(k), \rho^{(i)}(\boldsymbol{u})$ 分别表示第i个场景中，时刻k的投资组合布局，时刻k的总资产以及累积的期末收益。

利用采样的场景，我们可以构造几个可能的经验风险测度，将其作为我们最小化的目标。常用的目标如下：令γ是给定的期末收益水平(通常但不是必须地，将γ设置成累积期末收益的平均值，即$\gamma = \frac{1}{N}\sum_{i=1}^{N} \rho^{(i)}$)：

$$J_1 \doteq \frac{1}{N} \sum_{i=1}^{N} |\gamma - \rho^{(i)}|, \tag{14.15}$$

$$J_2 \doteq \frac{1}{N} \sum_{i=1}^{N} (\gamma - \rho^{(i)})^2, \tag{14.16}$$

$$J_{p1} \doteq \frac{1}{N} \sum_{i=1}^{N} \max(0, \gamma - \rho^{(i)}), \tag{14.17}$$

$$J_{p2} \doteq \frac{1}{N} \sum_{i=1}^{N} (\max(0, \gamma - \rho^{(i)}))^2. \tag{14.18}$$

这些目标表示收益与γ间偏差的经验平均值。特别地，(14.15)正比例于收益$\rho^{(i)}$与γ 的偏差的ℓ_1范数；(14.16)正比例于这些偏差的ℓ_2范数。(14.17)和(14.18)是非对称的测度(有时称之为**下偏距**(lower partial moments))：J_{p1}度量收益变量$\rho^{(i)}$ 低于水平γ的经验平均值；J_{p2}度量这种偏差的平方的平均

值。选择一阶还是二阶成本测度取决于投资者的**风险规避**(risk aversion)水平. 因为二阶成本是残量的平方，因此更看重成本，所以高阶测度反映了对风险规避的水平更高。

我们的开环多阶段分配策略即寻求投资组合调整参数$\boldsymbol{u} = (\boldsymbol{u}(0), \ldots, \boldsymbol{u}(T-1))$，它最小化上面描述的某个成本测度，同时受限制于整个范围内每个时段所给定的投资组合约束，即考虑

$$\begin{aligned} J^*(\gamma) = \quad & \underset{\boldsymbol{u}}{\text{minimize}} \quad J(\boldsymbol{u}) \\ & \text{subject to} \quad \boldsymbol{x}^{(i)+}(k) \in \mathcal{X}(k), k = 0, \ldots, T-1; i = 1, \ldots, N, \\ & \qquad\qquad \mathbf{1}^\top \boldsymbol{u}(k) = 0, k = 0, \ldots, T-1, \end{aligned}$$

其中$J(\boldsymbol{u})$是刚才提到的一个成本，$\boldsymbol{x}^{(i)+}(k)$是第i个采样场景中由(14.13)和(14.14)确定的。求解一个线性规划问题或者一个凸二次规划问题是确定这些最优分配数值解的有效方法。

§14.3.3 带仿射价格的闭环分配

因为所有的调整量$\boldsymbol{u}(0), \ldots, \boldsymbol{u}(T-1)$是在时刻$k = 0$计算出来的，所以上节研究的开环策略在实践中实施时可能是次优的。尽管第一个决策$\boldsymbol{u}(0)$必定立即实施(现时变量)，但是实际上可以等到获得向前推的一个时段的收益的实际结果后，再对未来的决策做出判断，这些观察结果可以减小不确定性。例如，在时刻$k \geq 1$，当我们需要实施$\boldsymbol{u}(k)$ 时，我们已经**观察到**(observed)了从时段1到时段k 的资产收益的一个实现。因此，我们可以利用这个信息来考虑该分配决策$\boldsymbol{u}(k)$的**条件**(conditional)分配决策，由此对之前时段的收益做出反应。这蕴含着，我们不应专注于固定的决策，而是确定一些适当的**方法**(policy)，其描述了基于时刻 1 至 k 的观测收益流以后如何作出实际决策。当确定投资方法的结构时，我们应该对所得到的最优化问题的普适性和可计算性进行折衷。一些研究工作[7]发现，线性或者仿射方法是一种有效的折衷方法，因为用凸优化技术可以有效地计算反应策略。在本节中，我们遵循该思路，考虑如下形式的仿射方法，即

$$\boldsymbol{u}(k) = \bar{\boldsymbol{u}}(k) + \boldsymbol{\Theta}(k)(\boldsymbol{g}(k) - \bar{\boldsymbol{g}}(k)), k = 1, \ldots, T-1, \tag{14.19}$$

且$\boldsymbol{u}(0) = \bar{\boldsymbol{u}}(0)$，这里$\bar{\boldsymbol{u}}(k) \in \mathbb{R}^n (k = 0, \ldots, T-1)$是“名义上的”分配决策变量，$\boldsymbol{g}(k)$是第$k$ 个时段的增益向量，$\bar{\boldsymbol{g}}(k)$是给定的对$\boldsymbol{g}(k)$ 的期望值的一

[7]比如参见：G. Calafiore, Multi-period portfolio optimization with linear control policies, *Automatica*, 2008.

个估计值，$\boldsymbol{\Theta}(k) \in \mathbb{R}^{n\times n}(k=0,\ldots,T-1)$是策略“反应矩阵”，它的作用是对名义分配的调整量和增益$\boldsymbol{g}(k)$与它的期望值的偏差成比例。因为预算平衡约束$\mathbf{1}^\top \boldsymbol{u}(k)=1$要对任意的增益的实现都满足，所以我们施加了限制

$$\mathbf{1}^\top \bar{\boldsymbol{u}}(k)=0,\ \mathbf{1}^\top \boldsymbol{\Theta}(k)=0, k=1,\ldots,T-1.$$

§14.3.3.1 仿射方法下的投资组合动力学

应用调整方法(14.19)，即将(14.19)代入投资组合动力学方程(14.12)和(14.13)，我们得到

$$\boldsymbol{x}^+(k)=\boldsymbol{x}(k)+\bar{\boldsymbol{u}}(k)+\boldsymbol{\Theta}(k)(\boldsymbol{g}(k)-\bar{\boldsymbol{g}}(k)), \tag{14.20}$$

$$\boldsymbol{x}(k+1)=\boldsymbol{G}(k+1)\boldsymbol{x}^+(k),\quad k=1,\ldots,T-1, \tag{14.21}$$

令$\boldsymbol{\Theta}(0)\doteq\mathbf{0}$。反复应用公式(14.20)和(14.21)，我们得到瞬时时刻$k=1,\ldots,T-1$处投资组合的表达式：

$$\boldsymbol{x}(k)=\boldsymbol{\Phi}(1,k)\boldsymbol{x}(0)+\boldsymbol{\Omega}_k\bar{\boldsymbol{u}}+\sum_{t=1}^{k}\boldsymbol{\Phi}(t,k)\boldsymbol{\Theta}(t-1)\widetilde{\boldsymbol{g}}(t-1), \tag{14.22}$$

其中$\boldsymbol{\Theta}(0)=\mathbf{0}$，

$$\bar{\boldsymbol{u}}\doteq[\bar{\boldsymbol{u}}(0)^\top\cdots\bar{\boldsymbol{u}}(T-2)^\top\ \ \bar{\boldsymbol{u}}(T-1)^\top]^\top,$$

$$\tilde{\boldsymbol{g}}(k)\doteq\boldsymbol{g}(k)-\bar{\boldsymbol{g}}(k),\quad k=1,\ldots,T.$$

一个关键的观察所得是$\boldsymbol{x}(k)$是决策变量$\bar{\boldsymbol{u}}(k)$和$\boldsymbol{\Theta}(k),k=1,\ldots,T-1$的仿射函数。在整个过程中，投资收益的累积总收益是

$$\begin{aligned}\rho(\bar{\boldsymbol{u}},\Theta)&=\frac{w(T)}{w(0)}=\frac{\mathbf{1}^\top\boldsymbol{x}(T)}{\mathbf{1}^\top\boldsymbol{x}(0)}\\&=\frac{1}{\mathbf{1}^\top\boldsymbol{x}(0)}(\boldsymbol{\phi}(1,T)^\top\boldsymbol{x}(0)+\boldsymbol{\omega}_T^\top\bar{\boldsymbol{u}}+\sum_{t=1}^{\top}\boldsymbol{\Phi}(t,T)\boldsymbol{\Theta}(t-1)\widetilde{\boldsymbol{g}}(t-1),\end{aligned}$$

这也是一个关于变量$\bar{\boldsymbol{u}}$和$\boldsymbol{\Theta}\doteq[\boldsymbol{\Theta}(1)\cdots\boldsymbol{\Theta}(T-1)]$的仿射函数.

§14.3.3.2 使用仿射方法的最佳策略

给定从场景生成先哲那儿得到的各时段增益$\{\boldsymbol{G}(k),k=1,\ldots,T\}$的$N$个独立同分布样本(场景)$\boldsymbol{G}^{(i)}(k),k=1,\ldots,T,\ i=1,\ldots,T$。则我们可以

将(14.15)-(14.18)最为最优化的目标函数，即求解如下凸优化问题以来确定最优政策：

$$\begin{aligned} J^*_{\mathrm{cl}}(\gamma) = \quad & \underset{\bar{\boldsymbol{u}},\boldsymbol{\Theta}}{\text{minimize}} \quad J(\boldsymbol{u},\boldsymbol{\Theta}) \\ & \text{subject to} \quad \boldsymbol{x}^{(i)+}(k) \in \mathcal{X}(k), k = 0,\dots,T-1;\ i = 1,\dots,N, \\ & \qquad\qquad \mathbf{1}^\top \bar{\boldsymbol{u}}(k) = 0, k = 0,\dots,T-1, \\ & \qquad\qquad \mathbf{1}^\top \boldsymbol{\Theta}(k) = 0, k = 1,\dots,T-1, \end{aligned}$$

其中$\boldsymbol{x}^{(i)+}(k)$由(14.20)确定，但是其中的$\boldsymbol{x}(k)$应该换成$\boldsymbol{x}^{(i)}(k)$，即将第i个采样场景代入(14.22)得到的与第i个采样场景相匹配的投资组合。

§14.4 稀疏指标跟踪

我们再次考虑在例9.13中介绍的replicating(跟踪)指标问题。这个问题可以视为有约束的最小二乘问题，即

$$\begin{aligned} \underset{\boldsymbol{x}\in\mathbb{R}^n}{\text{minimize}} \quad & \|\boldsymbol{R}\boldsymbol{x} - \boldsymbol{y}\|_2^2 \\ \text{subject to} \quad & \mathbf{1}^\top \boldsymbol{x} = 1, \\ & \boldsymbol{x} \geq \mathbf{0}, \end{aligned}$$

其中$\boldsymbol{R} \in \mathbb{R}^{T\times n}$，它的第$i(i = 1,\dots,n)$列数据表示组合中的资产$i$在$T(> n)$个时期内的历史收益，$\boldsymbol{y} \in \mathbb{R}^T$，它包含了$T$个时期投资者的参考收益。我们的目标是找到资产组合的比例向量$\boldsymbol{x}$，使其对应的收益尽可能地匹配收益流的参考指标。然而，在实践中，n 可能很大，并且用户更喜欢用组分资产中的一个**小**(small) 的子集来replicating指标，因为资产越少越便于管理，并且需要的交易成本更少。因此，客户愿意放弃一些跟踪精度来换取解$\boldsymbol{x}$ 的稀疏性。我们在前面的章节中提到，一种通常的提高解的稀疏性的作法是向目标中添加一个ℓ_1范数正则项，即考虑形如$\|\boldsymbol{R}\boldsymbol{x}-\boldsymbol{y}\|_2^2+\lambda\|\boldsymbol{x}\|_1$的目标。遗憾的是，这种方法在我们的背景下不能工作，因为这里的决策变量属于标准单纯形$\mathcal{X} = \{\boldsymbol{x} : 1^\top \boldsymbol{x} = 1, \boldsymbol{x} \geq \mathbf{0}\}$，因此所有可行的$\boldsymbol{x}$的$\ell_1$范数是常值，均等于1。我们接下来介绍一个当变量被限制在单形$\mathcal{X}$中取值时，可以获得稀疏解的另一个有效的方法。首先考虑理想的问题，即我们要求解

$$p^* = \min_{\boldsymbol{x}\in\mathcal{X}} f(\boldsymbol{x}) + \lambda\|\boldsymbol{x}\|_0, \tag{14.23}$$

其中$f(\boldsymbol{x}) \doteq \|\boldsymbol{R}\boldsymbol{x} - \boldsymbol{y}\|_2^2$，$\|\boldsymbol{x}\|_0$是$\boldsymbol{x}$的基数，$\lambda \geq 0$是一个折衷参数。因为我们知道这是一个非凸的问题，难以求解，所以我们要寻找一种可以有效

求解的松弛方法。为此，我们研究

$$\|\boldsymbol{x}\|_1 = \sum_{i=1}^{n} \mid x_i \mid \leq \| \boldsymbol{x}\|_0 \max_{i=1,\ldots,n} \mid x_i \mid \leq \|\boldsymbol{x}\|_0 \|\boldsymbol{x}\|_\infty.$$

对$\boldsymbol{x} \in \mathcal{X}$，上面表达式左边部分等于1。因此我们得到

$$\boldsymbol{x} \in \mathcal{X} \Rightarrow \|\boldsymbol{x}\|_0 \geq \frac{1}{\|\boldsymbol{x}\|_\infty} = \frac{1}{\max\limits_{i=1,\ldots,n} x_i}.$$

所以，对$\boldsymbol{x} \in \mathcal{X}$，

$$f(\boldsymbol{x}) + \lambda \|\boldsymbol{x}\|_0 \geq f(\boldsymbol{x}) + \lambda \frac{1}{\max\limits_{i=1,\ldots,n} x_i}.$$

因此，解决问题

$$p^*_\infty = \min_{\boldsymbol{x} \in \mathcal{X}} f(\boldsymbol{x}) + \lambda \frac{1}{\max\limits_{i=1,\ldots,n} x_i} \tag{14.24}$$

后，我们就得到了原始最优目标值的下界p^*_∞，即$p^*_\infty \leq p^*$。进一步，记(14.24)的最优解为$\boldsymbol{x}^*_\infty$，则我们立即得到

$$f(\boldsymbol{x}^*_\infty) + \lambda \parallel \boldsymbol{x}^*_\infty \parallel_0 \geq p^* \geq p^*_\infty,$$

其中第一个不等式源于(14.23)。这个关系允许我们进行如下操作：如果我们获得了一个解$\boldsymbol{x}^*_\infty$，那么去检查它的**后验性**(a posteriori)，即它的次优性水平，因为该关系表明“真正”的最优值p^*一定包含于区间$[p^*_\infty, f(\boldsymbol{x}^*_\infty) + \lambda\|\boldsymbol{x}^*_\infty\|_0]$。

尚未解决的问题是如何得到问题(14.24)的解，因为这个问题关于$\boldsymbol{x}$是非凸的(函数$\frac{1}{\max_i x_i}$在$\mathcal{X}$ 上是凹函数)，所以它的解并不是唾手可得的。然而我们可以进行如下推导[8]

$$\begin{aligned} p^*_\infty &= \min_{\boldsymbol{x} \in \mathcal{X}} f(\boldsymbol{x}) + \min_{i=1,\ldots,n} \frac{\lambda}{x_i} \\ &= \min_{i=1,\ldots,n} \min_{\boldsymbol{x} \in \mathcal{X}} f(\boldsymbol{x}) + \frac{\lambda}{x_i} \\ &= \min_{i=1,\ldots,n} \min\{f(\boldsymbol{x}) + t : \boldsymbol{x} \in \mathcal{X},\ t \geq 0, t x_i \geq \lambda\}. \end{aligned}$$

[8]参见Pilanci, EI Ghaoui, and Chandrasekaran, Recovery of sparse probability measures via convex programming, Proc. Conference on Neural Information Processing Systems, 2012.

进一步，可以利用(10.2)将双曲型约束$tx_i \geq \lambda$表示为SOC约束，即

$$tx_i \geq \lambda, x_i \geq 0, t \geq 0 \Leftrightarrow \left\| \begin{bmatrix} \sqrt{\lambda} \\ x_i - t \end{bmatrix} \right\|_2 \leq x_i + t.$$

上述推导表明，我们可以通过求解n个凸优化问题，即对$i = 1, 2, \cdots, n$，求解

$$\begin{aligned} \underset{\boldsymbol{x},t}{\text{minimize}} \quad & f(\boldsymbol{x}) + t \\ \text{subject to} \quad & \boldsymbol{x} \in \mathcal{X}, \\ & t \geq 0, \\ & \left\| \begin{bmatrix} \sqrt{\lambda} \\ x_i - t \end{bmatrix} \right\|_2 \leq x_i + t. \end{aligned}$$

然后比较最优值，取最优值最小的一个问题对应的解，由此即获得非凸问题(14.24)的解. 下面用一个数值例子来说明，我们再次使用例9.13中的数据，它涉及到使用$n = 5$的组分指标来跟踪摩根士丹利世界(MSCI WORLD)指标。求解$\lambda = 0.1$时的问题(14.24)，我们获得只有两个非零元素的投资组合

$$\boldsymbol{x}_\infty^* = [95.19, 0, 4.81, 0, 0]^\top \times 0.01,$$

与此对应的跟踪误差$\|\boldsymbol{R}\boldsymbol{x}_\infty^* - \boldsymbol{y}\|_2^2 = 0.0206$。

注记14.1 我们这里讨论的稀疏性方法，除了应用到指标跟踪问题外，显然可以被应用于本文提到的其它决策变量被限制在标准单纯形中的投资组合分配问题。例如，一旦施加了非卖空条件，就可以用它来得到14.1.1节提到的均值方差最优化问题的稀疏解。相反的，当组合变量没有施加非负限制时(允许卖空)，可以使用标准的ℓ_1 范数法进行松弛，由此提升投资组合的稀疏性。

§14.5 习题

习题14.1. (**投资组合最优化问题**) 我们考虑单时段的涉及n种资产的最优化问题，其中决策向量$\boldsymbol{x} \in \mathbb{R}^n$包含了我们对每种资产的投资数量，确定以下哪些目标或约束可以使用凸优化进行建模。

(a) 风险水平(由投资组合的方差测度)等于一个给定的目标t(假设协方差矩阵是已知的)。

(b) 风险水平(由投资组合的方差测度)给给定的目标t之下。

(c) 夏普比率(定义为投资组合收益与投资组合标准差的比值)在目标$t \geq 0$之上。这里假设收益向量的期望和协方差矩阵都是已知的。

(d) 假设收益向量服从已知的高斯分布，确保投资组合收益低于目标t的概率小于3%。

(e) 假设收益向量$\boldsymbol{r} \in \mathbb{R}^n$有三种可能取值$\boldsymbol{r}^{(i)}, i = 1, 2, 3.$ 施加如下约束：这三个场景下投资收益的最小值要在目标水平t 之上。

(f) 在类似于(e)中的假设下：两个最小投资组合的收益的平均值均在目标水平t之上。**提示：**利用新变量$s_i = \boldsymbol{x}^\top r^{(i)} (i = 1, 2, 3)$ 并且考虑函数$\boldsymbol{s} \to s_{[2]} + s_{[3]}$，其中$s_{[k]}$表示$\boldsymbol{s}$的第$k$大元素，$k = 1, 2, 3$。

(g) 交易成本(在线性交易成本及初始投资组合$\boldsymbol{x}(0) = \mathbf{0}$的模型中)低于某个目标。

(h) 从初始投资组合$\boldsymbol{x}(0) = 0$到最优投资组合$\boldsymbol{x}$的交易数目低于某个目标。

(i) 投资组合收益的期望值与目标收益t之差的绝对值小于任意一个给定的正数ϵ(在这里，我们假设收益向量的期望$\bar{\boldsymbol{r}}$是已知的)。

(j) 投资组合收益的期望值或者在某一特定值t_{up}之上，或者在另一个特定值t_{low}之下。

习题14.2. (**中位风险**,Median risk) 我们考虑一个单时段关于n种资产的投资组合最优化问题。我们利用过去的样本，即单时段收益向量$\boldsymbol{r}_1, \ldots, \boldsymbol{r}_N$，其中$\boldsymbol{r}_t \in \mathbb{R}^n$ 代表着对资产从时段$t-1$ 到时段t 的投资组合。我们用

$$\hat{\boldsymbol{r}} \doteq (\frac{1}{N})(\boldsymbol{r}_1 + \cdots + \boldsymbol{r}_N)$$

来表示样本平均值向量，这是一个基于过去的样本对期望收益进行估计的估计量。

我们用以下量来度量风险。我们用$\rho_t(\boldsymbol{x})$来表示时刻t的收益(如果t时刻的投资组合是$\boldsymbol{x}$)。我们的风险测度是

$$\mathcal{R}_1(\boldsymbol{x}) \doteq \frac{1}{N} \sum_{t=1}^{n} |\rho_t(\boldsymbol{x}) - \hat{\rho}(\boldsymbol{x})|$$

其中$\hat{\rho}(\boldsymbol{x})$是投资组合的样本平均收益。

(a) 证明$\mathcal{R}_1(\boldsymbol{x}) = \|\boldsymbol{R}^\top \boldsymbol{x}\|_1$，你需要确定出这里的$\boldsymbol{R}$，它是一个$n \times N$ 矩阵。这个风险测度$\mathcal{R}_1$是凸的吗?

(b) 说明在投资组合收益率的样本平均值大于目标μ的条件下，极小化风险测度R_1可以用线性规划来表述. 将问题化成标准形式的线性规划，并且精确地定义变量和约束。

(c) 定量评价这种风险度量方式所得投资组合和使用更经典的，即基于方差

$$R_2(\boldsymbol{x}) \doteq \frac{1}{N}\sum_{t=1}^{n}(\rho_t(\boldsymbol{x}) - \hat{\rho}(\boldsymbol{x}))^2$$

度量风险所得投资组合的差异。

习题14.3. (**带因素模型的投资组合优化**-1)

(a) 考虑下面的投资组合优化问题：

$$\begin{aligned} p^* = \underset{\boldsymbol{x}}{\text{minimize}} \quad & \boldsymbol{x}^\top \boldsymbol{\Sigma} \boldsymbol{x} \\ \text{subject to} \quad & \hat{\boldsymbol{r}}^\top \boldsymbol{x} \geq \mu, \end{aligned}$$

其中$\hat{\boldsymbol{r}} \in \mathbb{R}^n$是收益向量的期望，$\boldsymbol{\Sigma} \in \mathbb{S}^n, \boldsymbol{\Sigma} \succeq \mathbf{0}$是收益的协方差矩阵，并且$\mu$ 是投资组合预期收益的一个目标值。假定随机收益向量$\boldsymbol{r}$ 服从简化因素模型，即

$$\boldsymbol{r} = \boldsymbol{F}(\boldsymbol{f} + \hat{\boldsymbol{f}}), \quad \hat{\boldsymbol{r}} \doteq \boldsymbol{F}\hat{\boldsymbol{f}},$$

其中$\boldsymbol{F} \in \mathbb{R}^{n\times k}(k \ll n)$是一个因素载荷矩阵，$\hat{\boldsymbol{f}} \in \mathbb{R}^k$是给定的，并且$\boldsymbol{f} \in \mathbb{R}^k$满足$\mathbb{E}\{\boldsymbol{f}\} = \mathbf{0}$ 和$\mathbb{E}\{\boldsymbol{f}\boldsymbol{f}^\top\} = \boldsymbol{I}$. 上述优化问题是一个包含$n$个决策变量的凸二次优化问题。解释如何将这个问题转化为等价的，且只包含k个决策变量的凸二次规划问题。从几何上解释既约后的问题. 给出问题的一个具有封闭形式的解。

(b) 考虑(a)中问题的变形，即

$$\begin{aligned} p^* = \underset{\boldsymbol{x}}{\text{minimize}} \quad & \boldsymbol{x}^\top \boldsymbol{\Sigma} \boldsymbol{x} - \gamma \hat{\boldsymbol{r}}^\top \boldsymbol{x} \\ \text{subject to} \quad & \boldsymbol{x} \geq \mathbf{0} \end{aligned}$$

其中$\gamma > 0$是一个折衷参数，用来权衡目标中的风险项和收益项的相关性。由于存在约束$\boldsymbol{x} \geq \mathbf{0}$，从而这个问题没有封闭形式的通解。

假定$\boldsymbol{r}$由如下形式的因素模型指定，即

$$\boldsymbol{r} = \boldsymbol{F}(\boldsymbol{f} + \hat{\boldsymbol{f}}) + \boldsymbol{e},$$

其中$\boldsymbol{F}, \boldsymbol{f}$以及$\hat{\boldsymbol{f}}$如(a)中所定义，$\boldsymbol{e}$是一个与$\boldsymbol{f}$不相关(即$\mathbb{E}\{\boldsymbol{f}\boldsymbol{e}^\top\} = \boldsymbol{0}$)的噪声项，并且满足$\mathbb{E}\{\boldsymbol{e}\} = \boldsymbol{0}$ 和$\mathbb{E}\{\boldsymbol{e}\boldsymbol{e}^\top\} = \boldsymbol{D}^2 \doteq \operatorname{diag}([d_1^2, \ldots, d_n^2]) \succ \boldsymbol{0}$. 假设我们期望去用12.3.1 节中讨论的对数障碍法来求解这个问题，解释如何利用收益的因素结构提高算法的数值性能。**提示：**添加恰当的松弛变量后，(加上障碍项的)目标的海森矩阵将是对角的。

习题14.4. (**带因素模型的投资组合优化**-2) 再次考虑练习14.3中的(b)。令$\boldsymbol{z} \doteq \boldsymbol{F}^\top \boldsymbol{x}$，并且验证(b)中的问题可以被重写成

$$\begin{aligned} p^* = \underset{\boldsymbol{x},\boldsymbol{z}}{\text{minimize}} \quad & \boldsymbol{x}^\top \boldsymbol{D}^2 \boldsymbol{x} + \boldsymbol{z}^\top \boldsymbol{z} - \gamma \hat{\boldsymbol{r}}^\top \boldsymbol{x} \\ \text{subject to} \quad & \boldsymbol{F}^\top \boldsymbol{x} = \boldsymbol{z}, \boldsymbol{x} \geq \boldsymbol{0}. \end{aligned}$$

考虑拉格朗日函数

$$L(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{\lambda}) = \boldsymbol{x}^\top D^2 \boldsymbol{x} + \boldsymbol{z}^\top \boldsymbol{z} - \gamma \hat{\boldsymbol{r}}^\top \boldsymbol{x} + \boldsymbol{\lambda}^\top (\boldsymbol{z} - \boldsymbol{F}^\top \boldsymbol{x})$$

以及对偶函数

$$g(\boldsymbol{\lambda}) \doteq \min_{\boldsymbol{x} \geq \boldsymbol{0}, \boldsymbol{z}} \quad L(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{\lambda}).$$

因为原始问题是凸的并且是严格可行的，所以强对偶成立，因此$p^* = d^* = \max_{\boldsymbol{\lambda}} g(\boldsymbol{\lambda})$.

(a) 找到对偶函数$g(\boldsymbol{\lambda})$的闭合形式的表达式。

(b) 用对偶最优变量$\boldsymbol{x}^*$来表示原始最优解$\boldsymbol{\lambda}^*$。

(c) 给出$-g(\boldsymbol{\lambda})$的一个次梯度。

习题14.5. (Kelly的投注策略) 一个赌徒的原始的资本是W_0，并且在一次赌局中将自己所有可得的资本下注，赌局中赢赌注的概率$p \in [0, 1]$ ，输掉赌注的概率是$1 - p$。赌局重复进行. 在k 次投注后，他的资本W_k是一个随机变量:

$$W_k = \begin{cases} 2^k W_0 & \text{依概率} p^k, \\ 0 & \text{以概率} 1 - p^k. \end{cases}$$

(a) 确定k次投注后该赌徒资产的期望。确定赌徒在某时刻k破产的概率。

(b) (a)中的结论描述的是毁灭性赌博策略。假设现在赌徒变得跟谨慎，即在每一步决将资本的一部分作为赌注。用w和ℓ 分别表示赌徒赢和输的赌局的次数，二者均是随机的. 则赌徒在k次投注后的资本

$$W_k = (1 + x)^w (1 - \boldsymbol{x})^\ell W_0,$$

其中$x \in [0,1]$是投注比例，并且$w + \ell = k$。定义赌徒资本的指数增长率为

$$G = \lim_{k\to\infty} \frac{1}{k} \log_2 \frac{W_k}{W_0}.$$

(i) 将指数增长率G表示成x的函数。这个函数是凹的吗？

(ii) 求$x \in [0,1]$使指数增长率G取到最大值。称这个x是最优的凯利投注策略。

(c) 考虑一个更加一般的情况，其中一个投资者可以在一个投资机会上面仅投资他的资金中的一部分，并且该投资以不同的概率具有不同的收益。具体地，如果投资了$W_0 x$ 美金，那么投资导致资产$W = (1 + rx)W_0$，其中r表示投资的收益，是一个离散随机变量，它以概率$p_1, \ldots, p_m$取$r_1, \ldots, r_m (p_i \geq 0, r_i \geq -1, i = 1, \ldots, m, \sum_{i=1}^m p_i = 1)$。

对于这种情况，(b)中引入的指数增长率就是投资的对数增益的期望值

$$G = \mathbb{E}\{\log(W/W_0)\} = \mathbb{E}\{\log(1 + rx)\}.$$

(b)中考虑的问题是这种推广问题的特殊情况，对应于$m = 2, r_1 = 1, r_2 = -1, p_1 = p, p_2 = 1 - p$。

(i) 给出指数增长函数G作为$x \in [0,1]$的函数的显式表达式。

(ii) 设计一个简单的计算方案以找到使G取最大值的最优的投资比例x。

习题14.6. (**多期投资**) 我们在n个时段上考虑一个多时段单一资产的投资决策问题。对任意给定的时段$i = 1, \ldots, n$，我们用y_i来表示预期收益，σ_i 表示对应的方差，u_i 表示投资的资金数量。假定我们的初始投资是$u_0 = w$，投资问题是

$$\phi(w) \doteq \max_{\boldsymbol{u}} \{\sum_{i=1}^{n+1} (y_i u_i - \lambda \sigma_i^2 u_i^2 - c|u_i - u_{i-1}|) : \ u_0 = w, u_{n+1} = 0\},$$

其中第一项代表利润，第二项代表风险，第三项是交易成本的近似。在这里，$c > 0$是单位交易成本，$\lambda > 0$是风险-收益折衷参数。(不失一般性我们假定$\lambda = 1$.)

(a) 确定这个问题的对偶问题.

(b) 证明函数ϕ是凹的，并且给出$-\phi$在w处的一个次梯度。如果函数ϕ在w处可微，它在w 处的梯度是什么？

(c) 函数ϕ关于初始投资w的灵敏度问题是什么?确切地，对于固定的$\boldsymbol{y}, \boldsymbol{\sigma}, c$和任意的$\epsilon > 0$，为$|\phi(w+\epsilon) - \phi(w)|$提供一个紧的上界. 你可以假设函数$\phi$对任意的$u \in [w, w+\epsilon]$是可微的。

习题14.7. (**个人理财问题**) 考虑如下的个人理财问题。在接下来的六个月内你将从事咨询工作，你因此而得到总共$C = 30000$美元的酬金。你打算用这笔报酬来清还信用卡在过去的债务，总共$D = 7000$美元。信用卡的APR(annual interest rate, 年利率)$r_1 = 15.95\%$。你要考虑下面的一些因素：

- 在每个月的一开始，你可以将信用卡上债务的一部分转移到另一张年利率较低的信用卡上，该张信用卡的年利率$r_2 = 2.9\%$。这种转移需要付交易费，交易费是总共转移数量的$r_3 = 0.2\%$。并且只能将债务从卡 1 转到卡 2。
- 雇主允许你自己选择付款时间表：你可以在至多六个月中分期付款。由于资金流动的原因，雇主限制每个月的付款量不超过$\frac{4}{3} \times (\frac{C}{6})$。
- 你每年的底薪$B = 70000$美金。你不能用你的底薪来还信用卡上的债务。然而它会影响你要缴纳的税金(如下所述)。
- 这六个月中的前三个月是本财年的最后三个月，后三个月是下一个财年的前三个月。所以，你要是想在本财年(咨询的前三个月)支取的酬金多一些，你就要交的税就很多。如果你在不同时期支取酬金的话，你要交的税将会少一些。具体的税费取决于你的年收入总额G，也就是你的底薪加上其它收入。边际税率表见表14.1。

表 14.1: 边际税收率表

所有总收入($)	边际税率	所有的税
$0 \le G \le 80,000$	10%	$10\% \times G$
$80,000 \le G$	28%	$28\% \times (G - 80,000) + 8000$

- 无风险利率(储蓄利率)是0。
- 事件的时间列表：所有事件发生在每个月的开始。也就是说，在每个月的开始，雇主支付给你想要的数量的酬金，同时你要决定还信用卡上债务的数量，并且决定从信用卡 1 转到信用卡 2 的债务的数量，并进行转移。任何未偿还债务的利率都将累积到该月底记账。

- 你的目标是在两个财政年结束时，还清信用卡上的所有债务，并且使得总资产最大。

(a) 将决策问题表述成一个优化问题。要保证精确地定义变量和约束。为了描述税收，我们使用下面的约束：

$$T_i = 0.1 \min(G_i, \alpha) + 0.28 \max(G_i - \alpha, 0) \tag{14.25}$$

其中T_i和G_i分别是第i个财年要缴的税金和总收入，$i = 1, 2$，$\alpha = 80000$是税收阈值参数。

(b) 这个问题是一个线性规划问题吗？请解释。

(c) 当α和G_i满足什么条件时，税收约束可以由下面的一组约束所代替？这是我们所要考虑的情况吗?在你的问题中可以用(14.26)来代替(14.25)吗？请解释。

$$\begin{aligned} T_i &= 0.1 d_{1,i} + 0.28 d_{2,i}, \\ d_{2,i} &\geq G_i - \alpha, \\ d_{2,i} &\geq 0, \\ d_{1,i} &\geq G_i - d_{2,i}, \\ d_{1,i} &\geq d_{2,i} - \alpha. \end{aligned} \tag{14.26}$$

(d) 这个问题的含(14.26)的新表述是凸的吗？解释你的答案。

(e) 用你喜爱的求解器求解这个问题。写出最优策略：收到酬金的方式(时间和数量)，支付/转移信用卡债务的时间表(时间和数量)，以及两年末最优总资产。最大资产W是多少？

(f) 对于$\alpha \in [70k, 90k]$，计算最优的W，并且画出W作为α的函数在这个范围的图形. 你能解释这个图形吗？

习题14.8. (**交易成本和市场影响**) 我们考虑下面的投资组合优化问题:

$$\max_{\boldsymbol{x}} \{\hat{\boldsymbol{r}}^\top \boldsymbol{x} - \lambda \boldsymbol{x}^\top \boldsymbol{C} \boldsymbol{x} - c \cdot T(\boldsymbol{x} - \boldsymbol{x}^0) : \ \boldsymbol{x} \geq \boldsymbol{0}, \ \boldsymbol{x} \in \mathcal{X}\} \tag{14.27}$$

其中C是经验协方差矩阵，$\lambda > 0$是风险参数，$\hat{\boldsymbol{r}}$表示在给定时段上每项资产收益的时间-平均。这里的约束集$\mathcal{X}$ 由下面的条件所决定。

- 不允许卖空。

- 存在预算约束$x_1 + \cdots + x_n = 1$。

在上面，T是函数，代表着交易费以及市场影响，$c \geq 0$是控制这些费用大小的参数，$\boldsymbol{x}^0 \in \mathbb{R}^n$是表示初始投资组合的向量。函数$T$ 形如

$$T(\boldsymbol{x}) = \sum_{i=1}^{n} B_M(\boldsymbol{x}),$$

其中函数B_M对于较小的$\boldsymbol{x}$来说是分段线性的，对于较大的$\boldsymbol{x}$来说是二次的；这样，我们寻求捕获事实：对于小宗交易(较小的$\boldsymbol{x}$)，交易成本是占主导地位的；同时对于那些大宗交易，市场影响占主导地位。确切地说，我们定义一个称之为“reverse Hüber”的截止参数为M的函数：对于一个标量z，函数值

$$B_M(z) \doteq \begin{cases} |z| & \text{当} |z| \leq M, \\ \frac{z^2+M^2}{2M} & \text{否则.} \end{cases}$$

标量$M > 0$描述直线形状的惩罚转变成二次形状的惩罚时的转变点。

(a) 说明B_M可以被表述为如下优化问题的解，即

$$B_M(z) = \min_{v,w \in \mathbb{R}} \{v + w + \frac{w^2}{2M} : \ |z| \leq v + w, v \leq M, w \geq 0\}.$$

上面的表述可以证明B_M是凸的. 请给出理由？

(b) 证明对于给定的$\boldsymbol{x} \in \mathbb{R}^n$:

$$T(\boldsymbol{x}) = \min_{\boldsymbol{w},\boldsymbol{v} \in \mathbb{R}^n} \{\mathbf{1}^\top(\boldsymbol{v}+\boldsymbol{w}) + \frac{1}{2M}\boldsymbol{w}^\top\boldsymbol{w} : \ \boldsymbol{v} \leq M\mathbf{1}, \boldsymbol{w} \geq \mathbf{0}, |\boldsymbol{x}-\boldsymbol{x}^0| \leq \boldsymbol{v}+\boldsymbol{w}\},$$

其中$\boldsymbol{w}, \boldsymbol{v}$现在表示$n$维向量，$\mathbf{1}$是分量全为 1 的向量，并且其中的$|\cdot|$和不等式都是逐分量意义的。

(c) 把最优化问题(14.27)表述成凸优化的形式。转化后的问题隶属于我们熟悉的那一类(LP,QP,SOCP 等)？

第十五章　控制问题

动力系统是时下热门的物理问题。典型动力系统的数学模型是常微分方程，其中指令与干扰信号是输入量，它们与已有的初始条件一起，决定了内在变量（状态）以及输出信号随时间如何演化。这种模型泛见于工程学，例如，用来描述飞机的飞行状态，内燃机和机械臂各自的运行情况，以及导弹的弹道等。

宽泛地讲，一个动力系统的**控制**问题就是通过确定适当的输入信号使系统向我们想要的方向演进，比如，遵循一个既定的轨道或在外界扰动下保持稳定等等。动力系统控制问题的入门介绍都需要一整本教科书。这里，我们仅考虑一类受限动力系统（即有限维线性定常系统）的几个具体方面。

我们首先介绍连续时间模型及其对应的离散模型。对离散时间模型来说，我们着重于在有限阶段下输入输出量的表现与由线性方程组描述的静态线性映射之间的联系。我们将看到，在这种情况下会自然而然地会涉及到最优化问题，并讨论这些问题在控制背景下的诠释。本章的第一部分处理所谓的基于最优化的控制问题(在这种问题中，通过求解一个有限阶段上的最优化问题来直接确定控制输入)及其在滑动阶段的迭代实现(此即所谓的模型预测控制(MPC)的范例)。本章的第二部分比较简略，我们反过来讨论一种更经典的控制方法，其基于诸如稳定性等这样的基于无限时间范畴的概念. 在这里，最优化方法(尤其是SDP)作为稳定性分析或设计稳定反馈控制器的工具，它是以间直接的方式出现在问题中的。

§15.1　连续及离散时间模型

§15.1.1　连续时间LTI 系统

先来考虑一个简单的例子。

例15.1. (**轨上小车**)如图15.1所示，考虑沿着一条水平轨道移动的质量为m的小车，这里小车受阻于阻尼系数为β的粘滞阻尼(即阻尼与速度成正比例). 记小车质心的位置为$p(t)$，记质心受到的力为$u(t)$，其中t是时间。

根据牛顿第二定律，有

$$u(t) - \beta\dot{p}(t) = m\ddot{p}(t),$$

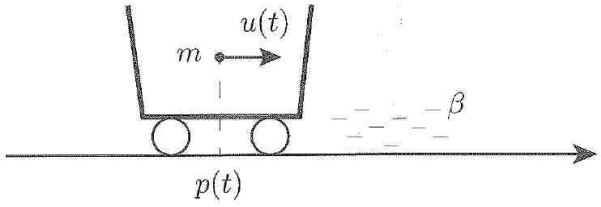

图 15.1: 轨上小车

这是一个二阶微分方程，它描述了这个系统的动力学。如果我们引入(状态)变量$x_1(t) = p(t)$，$x_2(t) = \dot{p}(t)$，则我们可以把牛顿方程重写为一阶常微分方程组：

$$\begin{aligned} \dot{x_1}(t) &= x_2(t), \\ \dot{x_2}(t) &= \alpha x_2(t) + bu(t), \end{aligned}$$

其中$\alpha \doteq -\frac{\beta}{m}, b \doteq \frac{1}{m}$. 这个系统可以用更紧凑的矩阵形式表示为

$$\dot{\boldsymbol{x}}(t) = \boldsymbol{A}_c\boldsymbol{x}(t) + \boldsymbol{B}_c u(t), \tag{15.1}$$

其中

$$\boldsymbol{A}_c = \begin{bmatrix} 0 & 1 \\ 0 & \alpha \end{bmatrix}, \quad \boldsymbol{B}_c = \begin{bmatrix} 0 \\ b \end{bmatrix}.$$

我们也可以给这个系统联立一个输出方程

$$y(t) = \boldsymbol{C}\boldsymbol{x}(t), \boldsymbol{C} = \begin{bmatrix} 1 & 0 \end{bmatrix} \tag{15.2}$$

以表示信号y是我们特别感兴趣的，这里表示小车自己的位置是我们特别关注的.

方程(15.1)和(15.2)实际上表述了一类相当有趣的动力系统，即所谓的(严格真(strictly proper)，有限维)连续时间的线性时不变(Linear Time-Invariant, LTI) 系统。如此的一个LTI 系统可以用$(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})$这样的三元矩阵对，以状态方程

$$\dot{\boldsymbol{x}} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}\boldsymbol{u} \tag{15.3}$$

$$\boldsymbol{y} = \boldsymbol{C}\boldsymbol{x} \tag{15.4}$$

来定义，这里$\boldsymbol{x} \in \mathbb{R}^n$是状态向量，$\boldsymbol{u} \in \mathbb{R}^m$是输入向量，$\boldsymbol{y} \in \mathbb{R}^p$是输出向量$(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{y}$都是时间$t \in \mathbb{R}$的函数，但为了简洁，我们省略对$t$ 的显式

依赖性)。给定t_0 时刻的状态的值，以及$t \geq t_0$时的输入向量$\boldsymbol{u}(t)$，则系统(15.3)的状态随时间的演化可以显式表示为(一般称之为拉格朗日公式)

$$\boldsymbol{x}(t) = \mathrm{e}^{\boldsymbol{A}(t-t_0)}\boldsymbol{x}(t_0) + \int_{t_0}^{t} \mathrm{e}^{\boldsymbol{A}(t-\tau)}\boldsymbol{B}\boldsymbol{u}(\tau)\boldsymbol{d}\tau, \quad t \geq t_0. \tag{15.5}$$

§15.1.2 离散时间LTI系统

对于分析特定的动力学现象，例如人口动力学或经济学，在**离散**的时间间隔上描述系统可能更加符合实际情况。例如，如果我们想要对一个国家的经济进行建模，当我们考虑要纳入模型的相关量(比如国内生产总值，或者失业率)时，我们会将离散的时间间隔Δ（Δ为一周，或一个月）处的数据，而非连续时间下的相关数据。通常用一个带符号的整型变量$k, k = \ldots, -1, 0, 1, 2, \ldots$,来表示“离散”的时间，这里$k$代表时刻$t = k\Delta$. (严格真，有限维的)离散时间线性时不变(LTI)系统用一阶差分方程组表示为

$$\boldsymbol{x}(k+1) = \boldsymbol{A}\boldsymbol{x}(k) + \boldsymbol{B}\boldsymbol{u}(k), \tag{15.6}$$

$$\boldsymbol{y}(k) = \boldsymbol{C}\boldsymbol{x}(k). \tag{15.7}$$

给定系统在时刻k_0的状态，以及$k \geq k_0$的输入$\boldsymbol{u}(k)$，通过递归应用方程(15.6)，易验证有

$$\boldsymbol{x}(k) = \boldsymbol{A}^{k-k_0}\boldsymbol{x}(k_0) + \sum_{i=k_0}^{k-1} \boldsymbol{A}^{k-i-1}\boldsymbol{B}\boldsymbol{u}(i), k \geqslant k_0. \tag{15.8}$$

§15.1.2.1 离散化

在实践中，常常会出现一个连续时间动力系统必须被数字器件(如DSP和计算机)分析或控制的情况。数字器件本身是离散时间对象，只能在离散时刻$k\Delta$与外界交互，这里的Δ是一个很小的时间间隔，代表I/O的时钟频率。因此很自然地，我们需要把一个形如(15.3)和(15.4)的连续时间系统“转化”为对应的离散时间的形式，通过在$t = k\Delta$这些时刻“给系统拍照”，这里的Δ称为**采样间隔**(sampling interval)，并假定输入信号$\boldsymbol{u}(t)$在两个相继的取样时刻之间是不变的，即

$$\boldsymbol{u}(t) = \boldsymbol{u}(k\Delta), \forall t \in [k\Delta, (k+1)\Delta).$$

这样的离散时间转换可以根据如下步骤完成：给定系统(15.3)在时刻$t = k\Delta$（这里我们把$\boldsymbol{x}(k\Delta)$记为$\boldsymbol{x}(k)$）的状态，我们可以用方程(15.5) 来计算

该状态在时刻$(k+1)\Delta$的值：

$$\begin{aligned}\boldsymbol{x}(k+1) &= e^{\boldsymbol{A}\Delta}\boldsymbol{x}(k) + \int_{k\Delta}^{(k+1)\Delta} e^{\boldsymbol{A}((k+1)\Delta-\tau)}\boldsymbol{B}\boldsymbol{u}(\tau)\boldsymbol{d}\tau \\ &= e^{\boldsymbol{A}\Delta}\boldsymbol{x}(k) + \int_{k\Delta}^{(k+1)\Delta} e^{\boldsymbol{A}((k+1)\Delta-\tau)}\boldsymbol{B}\boldsymbol{d}\tau\boldsymbol{u}(k) \\ &= e^{\boldsymbol{A}\Delta}\boldsymbol{x}(k) + \int_{0}^{\Delta} e^{\boldsymbol{A}\tau}\boldsymbol{B}\boldsymbol{d}\tau\boldsymbol{u}(k)\end{aligned}$$

对于连续时间系统(15.3)和(15.4)的采样版本，这意味着根据形如(15.6), (15.7)的离散时间递归时，有

$$\begin{aligned}\boldsymbol{x}(k+1) &= \boldsymbol{A}_\Delta\boldsymbol{x}(k) + \boldsymbol{B}_\Delta\boldsymbol{u}(k), \\ \boldsymbol{y}(k) &= \boldsymbol{C}\boldsymbol{x}(k),\end{aligned}$$

这里

$$\boldsymbol{A}_\Delta = e^{\boldsymbol{A}\Delta}, \boldsymbol{B}_\Delta = \int_0^\Delta e^{\boldsymbol{A}\tau}\boldsymbol{B}\boldsymbol{d}\tau. \tag{15.9}$$

例15.2. 考虑例15.1中轨上小车问题的连续时间模型，并假定$m = 1\text{kg}$, $\beta = 0.1\text{Ns/m}$. 采样间隔$\Delta = 0.1s$，对系统(15.1)离散化，我们得到离散时间系统

$$\boldsymbol{x}(k+1) = \boldsymbol{A}\boldsymbol{x}(k) + \boldsymbol{B}\boldsymbol{u}(k), \tag{15.10}$$

$$\boldsymbol{y}(k) = \boldsymbol{C}\boldsymbol{x}(k), \tag{15.11}$$

这里$\boldsymbol{A}, \boldsymbol{B}$由(15.9)按如下方式计算得到. 通过直接计算，我们首先观察到

$$\boldsymbol{A}_c^2 = \boldsymbol{A}_c\boldsymbol{A}_c = \begin{bmatrix}0 & \alpha \\ 0 & \alpha^2\end{bmatrix}, \boldsymbol{A}_c^3 = \boldsymbol{A}_c^2\boldsymbol{A}_c = \begin{bmatrix}0 & \alpha^2 \\ 0 & \alpha^3\end{bmatrix}, \ldots,$$

$$\boldsymbol{A}_c^k = \boldsymbol{A}_c^{k-1}\boldsymbol{A}_c = \begin{bmatrix}0 & \alpha^{k-1} \\ 0 & \alpha^k\end{bmatrix},$$

因此

$$\begin{aligned}\mathrm{e}^{\boldsymbol{A}_c t} &= \boldsymbol{I} + \sum_{k=1}^{\infty}\frac{1}{k!}\boldsymbol{A}_c^k t^k = \boldsymbol{I} + \sum_{k=1}^{\infty}\frac{1}{k!}\begin{bmatrix}0 & \alpha^{k-1} \\ 0 & \alpha^k\end{bmatrix}t^k \\ &= \begin{bmatrix}1 & \sum_{k=1}^{\infty}\frac{1}{k!}\alpha^{k-1}t^k \\ 0 & 1+\sum_{k=1}^{\infty}\frac{1}{k!}\alpha^k t^k\end{bmatrix} \\ &= \begin{bmatrix}1 & \frac{1}{\alpha}(\mathrm{e}^{\alpha t}-1) \\ 0 & \mathrm{e}^{\alpha t}\end{bmatrix}\end{aligned}$$

故并且

$$\boldsymbol{A} = \mathrm{e}^{\boldsymbol{A}_c\Delta} = \begin{bmatrix} 1 & \frac{1}{\alpha}(\mathrm{e}^{\alpha\Delta} - 1) \\ 0 & \mathrm{e}^{\alpha\Delta} \end{bmatrix} = \begin{bmatrix} 1 & 0.0995017 \\ 0 & 0.9900498 \end{bmatrix}.$$

类似地，对$\boldsymbol{B}$我们有

$$\begin{aligned} \boldsymbol{B} &= \int_0^\Delta \mathrm{e}^{\boldsymbol{A}_c\tau}\boldsymbol{B}_c d\tau = \int_0^\Delta \begin{bmatrix} 1 & \frac{1}{\alpha}(\mathrm{e}^{\alpha\tau} - 1) \\ 0 & \mathrm{e}^{\alpha\tau} \end{bmatrix} \begin{bmatrix} 0 \\ b \end{bmatrix} d\tau \\ &= b\int_0^\Delta \begin{bmatrix} \frac{1}{\alpha}(e^{\alpha\tau} - 1) \\ e^{\alpha\tau} \end{bmatrix} d\tau \\ &= \frac{b}{\alpha}\begin{bmatrix} \frac{1}{\alpha}(e^{\alpha\Delta} - 1) - \Delta \\ e^{\alpha\Delta} - 1 \end{bmatrix} = \begin{bmatrix} 0.0049834 \\ 0.0995016 \end{bmatrix} \end{aligned}$$

§15.2 基于最优化的控制综合问题

§15.2.1 状态跟踪的控制命令的综合

在这个章节中，我们集中讨论形如(15.6)和(15.7)的一个通用的离散时间系统，其有一个标量输入$u(k)$和一个标量输出$y(k)$(称这样的系统为SISO，其代表单输入单输出(Single-Input Single-Output))。这样的模型可能来自某系统的一个原始的离散时间表示，也可能是原始的连续时间系统的一个离散版本。我们讨论以下的控制问题：给定一个初始状态$\boldsymbol{x}(0) = \boldsymbol{x}_0$，一个目标状态$x_T$和一个目标整数时刻$T > 0$，确定对系统的一列控制动作$u(k)$，$k = 0, \ldots, T-1$使得在$T$ 时刻时系统状态为x_T。也就是说，我们寻找一个可以使系统状态在任意设定的时刻达到目标状态的控制命令列。

利用式(15.8)来表述这个问题，得到

$$\begin{aligned}\boldsymbol{x}(T) &= \boldsymbol{A}^\top \boldsymbol{x}_0 + \sum_{i=0}^{T-1} \boldsymbol{A}^{T-i-1}\boldsymbol{B}u(i) \\ &= \boldsymbol{A}^\top \boldsymbol{x}_0 + \begin{bmatrix}\boldsymbol{A}^{T-1}\boldsymbol{B} & \boldsymbol{A}^{T-2}\boldsymbol{B} & \dots & \boldsymbol{A}\boldsymbol{B} & \boldsymbol{B}\end{bmatrix} \begin{bmatrix} u(0) \\ u(1) \\ \cdot \\ \cdot \\ \cdot \\ u(T-2) \\ u(T-1)\end{bmatrix} \\ &= \boldsymbol{A}^\top \boldsymbol{x}_0 + \boldsymbol{R}_T \boldsymbol{\mu}_T,\end{aligned}$$

其中我们定义了T-**可达矩阵**(T-reachability matrix)，即$\boldsymbol{R}_T \doteq \begin{bmatrix}\boldsymbol{A}^{T-1}\boldsymbol{B} & \cdots & \boldsymbol{A}\boldsymbol{B} & \boldsymbol{B}\end{bmatrix}$，和包含了控制作用列的向量$\boldsymbol{\mu}_T$。因此原控制问题等价于寻找一个向量$\boldsymbol{\mu}_T \in \mathbb{R}^T$满足

$$\boldsymbol{R}_T \boldsymbol{\mu}_T = \boldsymbol{\xi}_T, \quad \boldsymbol{\xi}_T \doteq \boldsymbol{x}_T - \boldsymbol{A}^\top \boldsymbol{x}_0, \tag{15.12}$$

其中$\boldsymbol{R}_T \in \mathbb{R}^{n\times T}$是系统的$T-$可达矩阵。式(15.12)非常有趣，它告诉我们离散时间LTI系统在T时刻的状态是一个时间从 0 到$T-1$的输入命令列的线性方程，如图15.2所示.

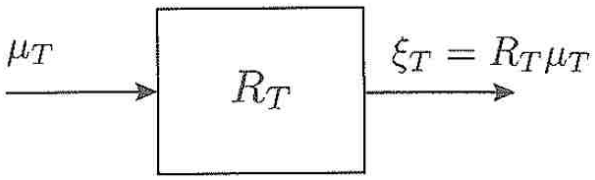

图 15.2: 从命令序列到状态的线性映射

决定一个使系统达到目标状态的输入序列相当于解一个关于$\boldsymbol{\mu}_T$的形如(15.12)的线性方程组。我们假设$T \geq n$，且$\boldsymbol{R}_T$是行满秩的(用控制工程语言描述，这个系统是完全$T-$可达的)。然后，式(15.12)总是有一个解；同时，如果$T > n$，它会有去穷多的可行解。这意味着这里有无限个可行的输入序列使得系统能达到指定的目标状态。在所有的可能性中，很自然地会选择一个用“最小的气力”达到目标状态的控制序列，其中“气力”(effort)可以定义为包含这些输入序列的向量$\boldsymbol{\mu}_T$ 的某种范数。

§15.2.1.1 最小能量控制

举例来说，有限时间间隔$\{0,\cdots,T-1\}$上的输入信号$\{u(0),\cdots,u(T-1)\}$的能量(energy)定义为

$$\|\boldsymbol{\mu}_T\|_2^2=\sum_{k=0}^{T-1}|u(k)|^2.$$

因此找能达到理想目标的能量最小的命令序列，就相当于寻找一个线性方程组的2-范数最小的解(见§6.3.2)，即

$$\begin{aligned}\underset{\boldsymbol{\mu}_T}{\text{minimize}}\quad&\|\boldsymbol{\mu}_T\|_2^2\\ \text{subject to}\quad&\boldsymbol{R}_T\boldsymbol{\mu}_T=\boldsymbol{\xi}_T.\end{aligned}$$

此系统有显式解(在完全可达性的假设下)

$$\boldsymbol{\mu}_T^*=\boldsymbol{R}_T^+\boldsymbol{\xi}_T=\boldsymbol{R}_T^\top(\boldsymbol{R}_T\boldsymbol{R}_T^\top)^{-1}\boldsymbol{\xi}_T,\tag{15.13}$$

称其中的矩阵$\boldsymbol{G}_T\doteq\boldsymbol{R}_T\boldsymbol{R}_T^\top$为系统的$T$可达Gram矩阵。

与最小能量控制问题相关的另一个问题是确定用单位能量的输入序列时，时刻T可以达到的所有可能状态的集合，即设法描述集合

$$\mathcal{E}_T=\{\boldsymbol{x}=\boldsymbol{R}_T\boldsymbol{\mu}_T+\boldsymbol{A}^\top\boldsymbol{x}_0:\|\boldsymbol{\mu}_T\|_2\le 1\}.$$

上述集合是单位球$\{\boldsymbol{\mu}_T:\|\boldsymbol{\mu}_T\|_2\le 1\}$ 在由矩阵$\boldsymbol{R}_T$描述的线性映射下的像集加上一个常数偏差项为$\boldsymbol{A}^\top\boldsymbol{x}_0$的平移。因此，根据引理6.4，我们推出单位能量的输入的可达状态集是以$\boldsymbol{c}=\boldsymbol{A}^\top\boldsymbol{x}_0$为中心，以$\boldsymbol{G}_T=\boldsymbol{R}_T\boldsymbol{R}_T^\top$ 为构型矩阵的椭球，即

$$\mathcal{E}_T=\{\boldsymbol{x}:(\boldsymbol{x}-\boldsymbol{c})^\top\boldsymbol{G}_T^{-1}(\boldsymbol{x}-\boldsymbol{c})\le 1\}.$$

§15.2.1.2 最小燃料控制

确定控制序列的另一种方法是最小化命令序列的1-范数而不是2-范数。$\boldsymbol{\mu}_T$的1-范数与产生输入命令的所需的“燃料”消耗成正比例。比如，在航天领域的应用中，控制的输入通常是指推进器通过喷射压缩气体而产生的推力，因此$\|\boldsymbol{\mu}_T\|_1$与控制驱动所需气体的总质量成正比例。于是问题现在变成

$$\begin{aligned}\underset{\boldsymbol{\mu}_T}{\text{minimize}}\quad&\|\boldsymbol{\mu}_T\|_1\\ \text{subject to}\quad&\boldsymbol{R}_T\boldsymbol{\mu}_T=\boldsymbol{\xi}_T,\end{aligned}\tag{15.14}$$

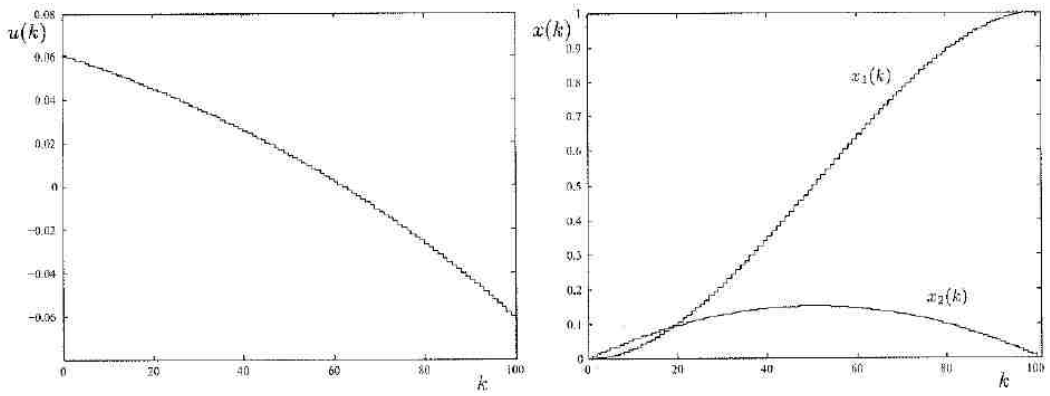

图 15.3: 最小能量控制信号(左)和相应的状态轨道(右)

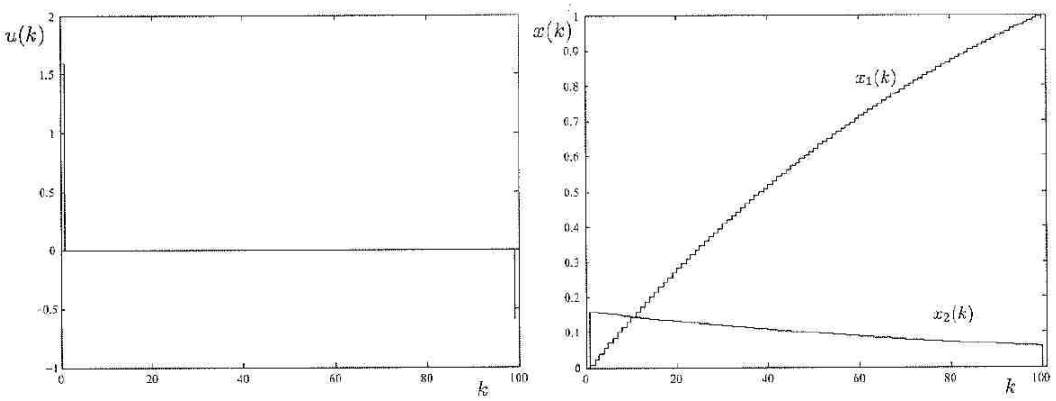

图 15.4: 最小燃料控制信号(左)和相应的状态轨道(右)

这可被重写成一个等价的线性规划

$$\begin{aligned} \underset{\boldsymbol{\mu}_T, \boldsymbol{s}}{\text{minimize}} \quad & \textstyle\sum_{k=0}^{T-1} s_k \\ \text{subject to} \quad & |u(k)| \leq s_k, k = 0, \ldots, T-1, \\ & \boldsymbol{R}_T \boldsymbol{\mu}_T = \boldsymbol{\xi}_T. \end{aligned}$$

例15.3. 考虑例15.2中轨上小车的离散模型。给定初始条件$\boldsymbol{x}(0) = [0 \quad 0]^\top$，我们寻找最小能量和最小燃料的输入序列使得小车在$t = 10s$ 时到达位置$p(t) = 1$，同时速度为0。因为采样时间是$\Delta = 0.1$s，从而知道最终的整

型目标时刻是$T = 100s$，因此我们的输入序列有100个未知量，即

$$\boldsymbol{\mu}_T = [u(0), \cdots, u(T-1)]^\top.$$

由式(15.13)可以找到最小能量控制序列，其结果如图15.3所示。最小燃料控制序列可以通过求解与(15.14)等价的LP问题得到，其结果如图15.4所示。从定性的角度来看，我们观察到从最小能量方法和最小燃料方法得到的输入信号的形状非常不同。尤其是最小燃料方法的解是**稀疏**(sparse)的，意味着在这种情况下，除了开始和结束的时刻，控制作用处处为0（在控制术语中，这叫做**开关**(bang-bang)控制序列).

§15.2.2 轨迹追踪控制命令综合

在§15.2.1中我们讨论了如何确定控制序列以达到一个目标状态(target state)的问题，但是我们并没有特别地关注初始状态和目标状态之间的状态的运动轨迹。在这里，我们研究如何确定控制序列，其使得一个离散时间LTI 系统的输出$y(k)$在给定的有限时间阶段$k \in \{1, \cdots, T\}$上尽可能地接近指定的参考轨道$y_{\text{ref}}(k)$. 我们接下来假设这个系统是SISO，且$\boldsymbol{x}(0) = \mathbf{0}$,$y_{\text{ref}}(0) = 0$.

根据方程(15.8)和$\boldsymbol{x}_0 = \mathbf{0}$，我们得到

$$\boldsymbol{x}(k) = \boldsymbol{A}^{k-1}\boldsymbol{B}u(0) + \cdots + \boldsymbol{A}\boldsymbol{B}u(k-2) + \boldsymbol{B}u(k-1).$$

然后，考虑输出方程$y(k) = \boldsymbol{C}\boldsymbol{x}(k)$我们得到

$$y(k) = \boldsymbol{C}\boldsymbol{A}^{k-1}\boldsymbol{B}u(0) + \cdots + \boldsymbol{C}\boldsymbol{A}\boldsymbol{B}u(k-2) + \boldsymbol{C}\boldsymbol{B}u(k-1), k = 1, \cdots, T.$$

将上式重写为矩阵形式，我们得到

$$\begin{bmatrix} y(1) \\ y(2) \\ \vdots \\ y(T) \end{bmatrix} = \begin{bmatrix} \boldsymbol{C}\boldsymbol{B} & 0 & \dots & 0 \\ \boldsymbol{C}\boldsymbol{A}\boldsymbol{B} & \boldsymbol{C}\boldsymbol{B} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{C}\boldsymbol{A}^{T-1}\boldsymbol{B} & \dots & \boldsymbol{C}\boldsymbol{A}\boldsymbol{B} & \boldsymbol{C}\boldsymbol{B} \end{bmatrix} \begin{bmatrix} u(0) \\ u(1) \\ \vdots \\ u(T-1) \end{bmatrix}$$

由此，我们显然可以定义输出序列$\boldsymbol{y}_T$和“变换矩阵”$\boldsymbol{\Phi}_T \in \mathbb{R}^{T\times T}$满足

$$\boldsymbol{y}_T = \boldsymbol{\Phi}_T\boldsymbol{\mu}_T. \tag{15.15}$$

矩阵$\boldsymbol{\Phi}_T$具有特殊结构(对角线元素是常数)，称为Toeplitz结构；此外，$\boldsymbol{\Phi}_T$是下三角矩阵，因此当$\boldsymbol{C}\boldsymbol{B} \neq 0$时，$\boldsymbol{\Phi}_T$是可逆的。更进一步，称$\boldsymbol{\Phi}_T$的第一

列的元素为系统的“**脉冲响应**”，因为它们代表了当输入信号是$u(0) = 1$和$u(k) = 0$(所有$k \geq 1$)这样的离散脉冲时，对应的系统的输出值序列。因此我们可以观察到，在一个固定的有限的时域中，一个离散时间LTI系统的输入输出行为可以用变换矩阵$\boldsymbol{\Phi}_T$决定的线性映射(15.15) 来描述。现在，如果想要得到的输出序列被指定为某一参考序列$y_{\text{ref}}(k)$，其中$k \in \{1, \cdots, T\}$，我们可以由式(15.15)解出输入序列$\boldsymbol{\mu}_T$，其能产生目标输出序列(当$\boldsymbol{BC} \neq 0$时，这样的解存在且唯一)。定义$\boldsymbol{y}_{\text{ref}} = [y_{\text{ref}}(1) \cdots y_{\text{ref}}(T)]^\top$，这也等价于找到$\boldsymbol{\mu}_T$满足

$$\min_{\boldsymbol{\mu}_T} \ \|\boldsymbol{\Phi}_T \boldsymbol{\mu}_T - \boldsymbol{y}_{\text{ref}}\|_2^2. \tag{15.16}$$

这个最小二乘表述的好处是当(15.15)是奇异的时，也可以为我们提供一个输入序列。在这种情况($\boldsymbol{\Phi}_T$奇异)下，我们找到这样一个输入序列$\boldsymbol{\mu}_T$ 使得对应的输出$\boldsymbol{y}_T$ 是“封闭”的，虽然在最小二乘意义上与$\boldsymbol{y}_{\text{ref}}$不完全相同。

把问题(15.16)从数学语言转换成一个控制设计“规范”(specification)，即要找到这样一个控制输入，它控制的输出与给定的参考输出$y_{\text{ref}}(k)$的**跟踪误差**(tracking error)最小。然而，对于控制系统想要找到仅有这样的规范的控制律是很难的。比如说注意到问题(15.16)所有的关注点都在跟踪误差上，然而我们完全忽视了**命令的活动力**(command activity)，也即输入序列$\boldsymbol{\mu}_T$ 的行为表现。事实上，有些问题必须考虑，比如说最大输入幅度的限制，转换速率约束或者至少让输入信号的能量或燃料消耗保持在可控范围中。举例来说，即使当Φ_T是可逆的，考虑下列正则化的轨迹追踪问题也是很好的实践，即

$$\min_{\boldsymbol{\mu}_T} \ \|\boldsymbol{\Phi}_T \boldsymbol{\mu}_T - \boldsymbol{y}_{\text{ref}}\|_2^2 + \gamma \|\boldsymbol{\mu}_T\|_2^2. \tag{15.17}$$

在这个问题里面我们引入了一个与输入信号能量成比正例的惩罚项. 这样的表述保证了追踪精度和输入能量间的一个权衡：典型的情况是为了达到一个小的跟踪精度需要一个高能量的输入信号。如果计算所得到的输入(当$\gamma = 0$时得到的)的能量过高，我们就需要尝试更大的$\gamma > 0$并再次求解这个问题，以此类推，直到找到一个令人满意的折中方案。

当然，以上问题里面可以有很多的变化。比如说，若想将1-范数纳入考量，那么对应的正则化问题是

$$\min_{\boldsymbol{\mu}_T} \ \|\boldsymbol{\Phi}_T \boldsymbol{\mu}_T - \boldsymbol{y}_{\text{ref}}\|_2^2 + \gamma \|\boldsymbol{\mu}_T\|_1,$$

或者想要给(15.17)添加关于控制信号的振幅的显式约束，即

$$\begin{aligned}\underset{\boldsymbol{\mu}_T}{\text{minimize}} \quad & \|\boldsymbol{\Phi}_T\boldsymbol{\mu}_T - \boldsymbol{y}_{\text{ref}}\|_2^2 + \gamma\|\boldsymbol{\mu}_T\|_2^2 \\ \text{subject to} \quad & \|\boldsymbol{\mu}_T\|_\infty \le u_{\max},\end{aligned}$$

得到的这个问题是凸二次规划。也易于处理关于信号的瞬时速率的改变(转换速率)的约束，通过观察可得

$$\begin{bmatrix} u(1)-u(0) \\ u(2)-u(1) \\ \vdots \\ u(T-1)-u(T-2) \end{bmatrix} = \boldsymbol{D}\boldsymbol{\mu}_T, \qquad \boldsymbol{D} \doteq \begin{bmatrix} -1 & 1 & 0 & \cdots & 0 \\ 0 & -1 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & -1 & 1 \end{bmatrix}$$

如果用$s_{\max}$表示对输入的转换速率的上限，则具有振幅限制和输入信号转换速率限制的正则化问题为

$$\begin{aligned}\underset{\boldsymbol{\mu}_T}{\text{minimize}} \quad & \|\boldsymbol{\Phi}_T\boldsymbol{\mu}_T - \boldsymbol{y}_{\text{ref}}\|_2^2 + \gamma\|\boldsymbol{\mu}_T\|_2^2 \\ \text{subject to} \quad & \|\boldsymbol{\mu}_T\|_\infty \le u_{\max}, \ \ \|\boldsymbol{D}\boldsymbol{\mu}_T\|_\infty \le s_{\max}.\end{aligned}$$

这是一个凸二次规划问题.

例15.4. 在此考虑例15.2中轨上小车的离散模型。给定给定初始状态$\boldsymbol{x}(0) = [0 \quad 0]^\top$和参考输出轨道

$$y_{\text{ref}}(k) = \sin(\omega k\Delta), \omega = \frac{2\pi}{10}, k = 1, \cdots, 100,$$

我们寻找使得系统(15.10)的输出尽可能地接近y_{ref}的输入信号$u(k), k \in \{1, \cdots, 100\}$。为此，我们考虑式(15.17)给出的正则化最小二乘表述。我们首先求解$\gamma = 0$ 时对应的问题。在这种情况下，因为$\boldsymbol{CB} = 0.005 \neq 0$，此时从原理上讲，可以可得输入控制其使得输出信号精确地等于参考信号，即可以做到精确跟踪，结果如图15.5所示：我们观察到跟踪误差的确是数值上的0(注意右图中$e(k)$ 的量级是10^{-15})。然而，实现这个结果的代价是输入信号$u(k)$ 剧烈震荡，且振幅很大，见图15.5中的左图。产生这种现象的原因之一是：当$\boldsymbol{CB} \neq 0$时，尽管$\boldsymbol{\Phi}_T$ 是可逆的，但是$\boldsymbol{\Phi}_T$ 的条件数随T 迅速上升(在这个问题中，$\boldsymbol{\Phi}_T$ 的条件数大约为10^6)。

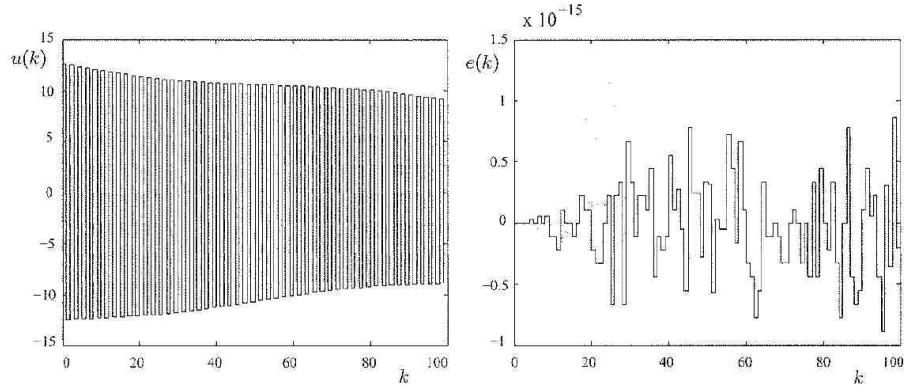

图 15.5: 输入信号$u(k)$(左)和$\gamma = 0$时的跟踪误差$e(k) = y(k) - y_{\text{ref}}(k)$(右)

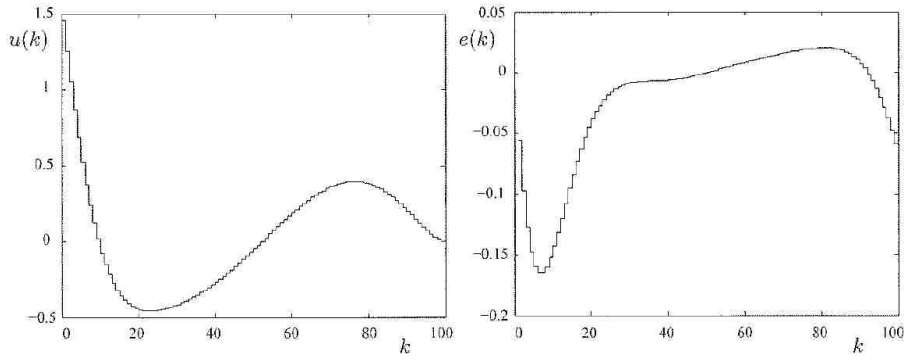

图 15.6: 输入信号$u(k)$(左)和$\gamma = 0.1$时的跟踪误差$e(k) = y(k) - y_{\text{ref}}(k)$(右)

如果考虑输入信号的能量的非零惩罚，则可以改变(减小)最小二乘问题的法方程的系数矩阵的条件数(见§6.7.3.1)，从而大幅提升输入信号的性能。举例来说，对于上例，取$\gamma = 0.1$并求解正则化的最小二乘问题(15.17)，所得结果如图15.6所示。我们观察到这里的跟踪误差比之前的高(但是可能仍然在可接受范围内，这个需要根据具体额问题具体分析)，然而与$\gamma = 0$时的情况相比，输入信号振幅变小了，也平缓得多。

§15.2.3 模型预测控制

模型预测控制(Model Predict Control, MPC)是一种非常有效的，且被广泛应用于动力系统的控制方法，它是基于最优化技术的。我们在这里简单地阐述LTI 离散时间系统中的MPC方法。前几小节中讨论的控制设计方法的共同特点是：需要事先(也就是在时刻 0 时)计算定义在时长

为T的给定时刻上的**整个**控制序列$u(0),\cdots,u(T-1)$，然后再将其应用到系统中。在实际操作的过程中，因为将计算所得的输入应用于真实的系统时，这个系统因干扰可能随时间发生变化，从而有可能出现某些问题。如果时间阶段T很大，即使我们计算得到了控制序列，然后靠后的瞬时时刻的系统状态因为干扰可能完全不同于我们在时刻 0 (即计算控制序列)时所事先预测的那样。像这种需要在多个时间段上进行决策(这里就是确定控制的输入)是一个很普遍的问题。

直觉表明以下作法更合理. 我们在时刻$k=0$ 计算序列

$$\boldsymbol{u}^{(0)}\doteq(u_{0|0},\cdots,u_{T-1|0}),$$

而后仅仅把这些控制中的第一个应用于$k=0$时的系统，即$u(0)=u_{0|0}$。然后，我们**静观其变**(wait and see)，即我们观察真实系统对第一个输入如何响应(即我们测量实际的状态$\boldsymbol{x}(1)$)), 然后重新计算向前平移时域一步后的新序列(记作$\boldsymbol{u}^{(1)}\doteq(u_{0|1},\cdots,u_{T-1|1})$。因此，在时刻$k=1$ 时，我们给系统施加这个序列的第一个值，即$u(1)=u_{0|1}$，然后不停地重复这个过程(即我们得到结果$\boldsymbol{x}(2)$，计算新的序列$\boldsymbol{u}^{(2)}=(u_{0|2},\cdots,u_{T-1|2})$，将$u(2)=u_{0|2}$施加给系统，以此类推)。称这种方法是MPC 控制率的**滑动时域**(sliding-horizon)实现.

下面考虑确定预测控制的一种基本模式. 对给定的时刻k，首先测量真实系统在时刻k的状态变量值，记为$\boldsymbol{x}(k)$. 然后以包含了需要计算的时刻k的预测序列$\boldsymbol{u}_{0|k}$的变量

$$\boldsymbol{u}^{(k)}\doteq(\boldsymbol{u}_{0|k},\cdots,\boldsymbol{u}_{T|k})$$

作为决策变量，并用$\boldsymbol{x}_{j|k},j=0,\cdots,T$,表示预测的状态序列，则其需要满足系统的动力学方程

$$\boldsymbol{x}_{j+1|k}=\boldsymbol{A}\boldsymbol{x}_{j|k}+\boldsymbol{B}\boldsymbol{u}_{j|k},j=0,\cdots,T-1. \tag{15.18a}$$

$$\boldsymbol{x}_{0|k}=\boldsymbol{x}(k). \tag{15.18b}$$

最后，MPC的滑动时域实现求解如下形式的最优化问题来得到$\boldsymbol{u}^{(k)}$，即

$$\begin{aligned}
\underset{\boldsymbol{u}^{(k)}}{\text{minimize}}\quad & J=\textstyle\sum_{j=0}^{T-1}(\boldsymbol{x}_{j|k}^{\top}\boldsymbol{Q}\boldsymbol{x}_{j|k}+\boldsymbol{u}_{j|k}^{\top}\boldsymbol{R}\boldsymbol{u}_{j|k})+\boldsymbol{x}_{T|k}^{\top}\boldsymbol{S}\boldsymbol{x}_{T|k}\\
\text{subject to}\quad & \boldsymbol{u}_{1b}^{(j)}\le\boldsymbol{u}_{j|k}\le\boldsymbol{u}_{ub}^{(j)},j=0,\cdots,T-1,\\
& \boldsymbol{F}^{(j)}\boldsymbol{x}_{j|k}\le\boldsymbol{g}^{(j)},j=1,\cdots,T,\\
& \boldsymbol{x}_{j+1|k}=\boldsymbol{A}\boldsymbol{x}_{j|k}+\boldsymbol{B}\boldsymbol{u}_{j|k},j=0,\cdots,T-1\\
& \boldsymbol{x}_{0|k}=\boldsymbol{x}(k),
\end{aligned} \tag{15.19}$$

其中$\boldsymbol{Q}\in\mathbb{S}^n_+,\boldsymbol{S}\in\mathbb{S}^n_+,\boldsymbol{R}\in\mathbb{S}^m_{++}$是给定的权重矩阵，$\boldsymbol{u}^{(j)}_{ub},\boldsymbol{u}^{(j)}_{lb}\in\mathbb{R}^m$ 是关于预测控制输入序列的给定的上界和下界向量，$\boldsymbol{F}^{(j)}\in\mathbb{R}^{q\times n},\boldsymbol{g}^{(j)}\in\mathbb{R}^q$描述了可能的关于系统状态的多面集约束。

如果我们从$i=0$到j递归使用方程(15.18a)，并注意到(15.18b)，得

$$\boldsymbol{x}_{j|k}=\boldsymbol{A}^j\boldsymbol{x}(k)+\boldsymbol{R}_j\boldsymbol{u}^{(k)},\qquad \boldsymbol{R}_j=[\boldsymbol{A}^{j-1}\boldsymbol{B}\cdots\boldsymbol{A}\boldsymbol{B}\quad\boldsymbol{B}\quad 0\cdots 0].$$

这表明$\boldsymbol{x}_{j|k}$是决策变量$\boldsymbol{u}^{(k)}$的线性函数。将这个表述代入问题(15.19)的目标和前两组约束，易于看到所得问题是关于变量$\boldsymbol{u}^{(k)}$ 的凸二次规划。

因此，在每一个时刻$k=0,1,\cdots$,我们观测$\boldsymbol{x}(k)$(这需要状态变量可测量)，然后通过求解(15.19)来找到最佳的预测序列$\boldsymbol{u}^{(k)}=(\boldsymbol{u}_{0|k},\cdots,\boldsymbol{u}_{T|k})$，最后将控制输入$\boldsymbol{u}(k)=\boldsymbol{u}_{0|k}$施加给真实的系统，以此类推，进行迭代。

问题(15.19)的目标函数量化了状态所蕴含能量和控制信号所蕴含能量的折衷($\boldsymbol{x}^\top_{j|k}\boldsymbol{Q}_{j|k}\boldsymbol{x}_{j|k}$和$\boldsymbol{u}^\top_{j|K}\boldsymbol{R}\boldsymbol{u}_{j|k}$分别代表着状态向量和控制向量的加权2-范数)，外加一个与此相似的关于最终状态的惩罚项$\boldsymbol{x}^\top_{T|k}\boldsymbol{S}\boldsymbol{x}_{T|k}$。这种控制策略的目的是驱使(调节)系统的状态渐近地逼近 0。MPC的一个特色是它允许我们**直接**对控制信号施加限制(比如上下界等)，也可以直接对容许状态的轨道施加限制。

§15.3 分析和控制器设计中的最优化

在本节中我们将简要概要讨论利用最优化分析系统的性质(比如稳定性)，或者为系统设计更合适的控制器。这些方法通常着重考虑系统的渐进性质. 与上一节我们所讨论的利用最优化直接确定控制序列不同，这些技术是间接的，即最优化不是用来直接确定控制序列，而是用来决定一些参数的值(例如控制器的增益)，这些参数与连接在系统上(例如一个反馈装置)以便对系统进行控制的理想器件(控制器)有关。以控制系统的分析和设计的最优化方法(尤其是基于SDP)为主题的文献现在已经非常多了. 这里将通过考虑一些关于LTI 系统的稳定性和反馈稳定性等非常基本的问题，让大家体会这一类结论的特点。

§15.3.1 连续时间的Lyapunov 稳定性分析

称连续时间LTI系统

$$\dot{\boldsymbol{x}}(t)=\boldsymbol{A}\boldsymbol{x}(t)+\boldsymbol{B}\boldsymbol{u}(t)\tag{15.20}$$

是(渐进)稳定的，如果对$\boldsymbol{u}(t)=0$，系统对所有的初态$\boldsymbol{x}(0)=\boldsymbol{x}_0$ 都有$\lim_{t\to\infty}\boldsymbol{x}(t)=0$成立. 用文字描述，系统的**自由响应**(free response) 渐进地趋于 0，并且

与初值无关。从Lyapunov稳定性理论，我们知道一个系统稳定的充分必要条件是存在一个矩阵$\boldsymbol{P} \succ \boldsymbol{0}$ 使得

$$\boldsymbol{A}^\top \boldsymbol{P} + \boldsymbol{P}\boldsymbol{A} \prec \boldsymbol{0}, \tag{15.21}$$

这是一个针对矩阵变量$\boldsymbol{P}$的(严格)LMI条件。存在满足条件的矩阵$\boldsymbol{P}$为形如二次Lyapunov 函数$V(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{P}\boldsymbol{x}$的系统稳定性提供了**判据**(certificate)。注意到$\boldsymbol{P} \succ \boldsymbol{0}$ 与(15.21)这两个条件是齐次的。这意味着如果存在某个$\bar{\boldsymbol{P}}$满足条件，即

$$\bar{\boldsymbol{P}} \succ \boldsymbol{0}, \quad \boldsymbol{A}^\top \bar{\boldsymbol{P}} + \bar{\boldsymbol{P}}\boldsymbol{A} = -\bar{\boldsymbol{Q}}, \bar{\boldsymbol{Q}} \succ \boldsymbol{0}$$

那么形如$\boldsymbol{P} = \alpha\bar{\boldsymbol{P}}, \alpha > 0$的这些矩阵同样满足这些条件. 特别的，如果我们选取

$$\alpha = \max\{\lambda_{min}^{-1}(\bar{\boldsymbol{P}}), \lambda_{min}^{-1}(\bar{\boldsymbol{Q}})\},$$

那么

$$\boldsymbol{P} = \alpha\bar{\boldsymbol{P}} \succeq \boldsymbol{I}, \quad \boldsymbol{A}^\top \boldsymbol{P} + \boldsymbol{P}\boldsymbol{A} \preceq -\boldsymbol{I}.$$

因此，稳定性等价于满足下面两个关于$\boldsymbol{P}$的非严格LMI条件：

$$\boldsymbol{P} \succeq \boldsymbol{I}, \quad \boldsymbol{A}^\top \boldsymbol{P} + \boldsymbol{P}\boldsymbol{A} \preceq -\boldsymbol{I}.$$

这样，寻找一个Lyapunov稳定性判据$\boldsymbol{P}$的问题可以表述成形如

$$\begin{aligned} \underset{\boldsymbol{P},\nu}{\text{minimize}} \quad & \nu \\ \text{subject to} \quad & \boldsymbol{A}^\top \boldsymbol{P} + \boldsymbol{P}\boldsymbol{A} \preceq -\boldsymbol{I}, \\ & \boldsymbol{I} \preceq \boldsymbol{P} \preceq \nu\boldsymbol{I}, \end{aligned}$$

的SDP问题. 该表述的最后一个LMI约束表示$\boldsymbol{P}$的条件数不超过ν，因此目标函数表明我们要寻找一个条件数最小的Lyapunov矩阵。

一种等价的方法是用逆Lyapunov矩阵$\boldsymbol{W} = \boldsymbol{P}^{-1}$进行另一种表述。事实上，给(15.21)的两边同时左乘和右乘$\boldsymbol{P}^{-1}$，我们得到另一个稳定性条件：$\exists \boldsymbol{W} \succ \boldsymbol{0}$,使得$\boldsymbol{W}\boldsymbol{A}^\top + \boldsymbol{A}\boldsymbol{W} \prec \boldsymbol{0}$，与上面关于$\boldsymbol{P}$的基于齐次性的推理类似，这些条件等价于

$$\exists \boldsymbol{W} \succeq \boldsymbol{I} : \boldsymbol{W}\boldsymbol{A}^\top + \boldsymbol{A}\boldsymbol{W} \preceq -\boldsymbol{I}. \tag{15.22}$$

从而，求解SDP问题

$$\begin{aligned} \underset{\boldsymbol{W},\nu}{\text{minimize}} \quad & \nu \\ \text{subject to} \quad & \boldsymbol{W}\boldsymbol{A}^\top + \boldsymbol{A}\boldsymbol{W} \preceq -\boldsymbol{I}, \\ & \boldsymbol{I} \preceq \boldsymbol{W} \preceq \nu\boldsymbol{I} \end{aligned}$$

可以找到一个最优的(最小条件数)逆Lyapunov判据.

§15.3.2 镇定状态反馈设计

能够将之前的方法从稳定性分析推广到反馈式镇定控制律设计。假设控制输入是**状态反馈**(state-feedback)式的，即

$$\boldsymbol{u}(t) = \boldsymbol{K}\boldsymbol{x}(t), \tag{15.23}$$

其中$\boldsymbol{K} \in \mathbb{R}^{m\times n}$是状态反馈**增益**(gain)矩阵(或者控制器)，需要设计$\boldsymbol{K}$使得被控系统是稳定的。将(15.23)代入状态方程(15.20)，我们得到被控制系统

$$\dot{\boldsymbol{x}} = (\boldsymbol{A} + \boldsymbol{B}\boldsymbol{K})\boldsymbol{x}.$$

这个系统是稳定的当且仅当闭环系统的矩阵$\boldsymbol{A}_{cc} = \boldsymbol{A}+\boldsymbol{B}\boldsymbol{K}$满足条件(15.22)，即当且仅当

$$\exists \boldsymbol{W} \succeq \boldsymbol{I}: \quad \boldsymbol{W}\boldsymbol{A}^\top + \boldsymbol{A}\boldsymbol{W} + (\boldsymbol{K}\boldsymbol{W})^\top\boldsymbol{B}^\top + \boldsymbol{B}(\boldsymbol{K}\boldsymbol{W}) \preceq -\boldsymbol{I}.$$

引入新变量$\boldsymbol{Y} \doteq \boldsymbol{K}\boldsymbol{W}$，则可将这些条件改写为

$$\exists \boldsymbol{W} \succeq \boldsymbol{I}: \quad \boldsymbol{W}\boldsymbol{A}^\top + \boldsymbol{A}\boldsymbol{W} + \boldsymbol{Y}^\top\boldsymbol{B}^\top + \boldsymbol{B}\boldsymbol{Y} \preceq -\boldsymbol{I}.$$

这是关于变量$\boldsymbol{M}$和$\boldsymbol{Y}$的LMI条件。我们先求解一个以上述LMI为约束的凸优化问题，然后由$\boldsymbol{K} = \boldsymbol{Y}\boldsymbol{W}^{-1}$得到镇定反馈增益$\boldsymbol{K}$。例如，我们可以将$\boldsymbol{W}$的条件数和$\boldsymbol{Y}$ 的范数的折衷作为最优化问题的目标，即考虑

$$\begin{aligned}
\underset{\boldsymbol{W},\boldsymbol{Y},\nu}{\text{minimize}} \quad & \nu + \eta\|\boldsymbol{Y}\|_2 \\
\text{subject to} \quad & \boldsymbol{W}\boldsymbol{A}^\top + \boldsymbol{A}\boldsymbol{W} + \boldsymbol{Y}^\top\boldsymbol{B}^\top + \boldsymbol{B}\boldsymbol{Y} \preceq -\boldsymbol{I}, \\
& \boldsymbol{I} \preceq \boldsymbol{W} \preceq \nu\boldsymbol{I},
\end{aligned}$$

其中$\eta \geq 0$是折衷参数。

§15.3.2.1 鲁棒的反馈设计

上面描述的基于LMI的反馈稳定化是相当灵活的，因为可以将其推广到**不确定**(uncertain)系统的稳定化。在这里，我们只讨论不确定性影响系统时最简单的情况，即所谓的**场景**(scenario)或者多面体不确定性。考虑系统

$$\dot{\boldsymbol{x}}(t) = \boldsymbol{A}\boldsymbol{x}(t) + \boldsymbol{B}\boldsymbol{u}(t), \tag{15.24}$$

其中系统矩阵$\boldsymbol{A}, \boldsymbol{B}$可以在由给定的矩阵$(\boldsymbol{A}_i, \boldsymbol{B}_i), i = 1, \ldots, N$,张成的多面体内变化，即

$$(\boldsymbol{A}, \boldsymbol{B}) \in \mathrm{co}\{(\boldsymbol{A}_1, \boldsymbol{B}_1), \ldots, (\boldsymbol{A}_N, \boldsymbol{B}_N)\}$$

矩阵对$(\boldsymbol{A}_i, \boldsymbol{B}_i)$是多面体的顶点，并且可以将它们解释为场景，或者一个不确定的设备的不同的可能实现，也可以将它们解释为在不同的操作条件或者时间点下的测量得到。可以证明不确定系统(15.24)稳定的**充分条件** 是对所有顶点系统存在一个公共的二次Lyapunov 函数，即

$$\exists \boldsymbol{W} \succeq \boldsymbol{I} : \boldsymbol{W}\boldsymbol{A}_i^\top + \boldsymbol{A}_i\boldsymbol{W} \preceq -\boldsymbol{I}, i = 1, \ldots, N.$$

基于这个充分条件，我们可以设计形如(15.23)的反馈控制律以便能鲁棒地镇定不确定系统。可将待控系统描述为

$$\dot{\boldsymbol{x}}(t) = (\boldsymbol{A} + \boldsymbol{B}\boldsymbol{K})\boldsymbol{x}(t).$$

如果

$$\exists \boldsymbol{W} \succeq \boldsymbol{I} : \boldsymbol{W}(\boldsymbol{A}_i + \boldsymbol{B}_i k)^\top + (\boldsymbol{A}_i + \boldsymbol{B}_i\boldsymbol{K})\boldsymbol{W} \preceq -\boldsymbol{I}, i = 1, \ldots, N,$$

则待控系统是稳定的. 引入新变量$\boldsymbol{Y} = \boldsymbol{K}\boldsymbol{W}$，则求解如下优化问题可得镇定控制器的反馈增益矩阵$\boldsymbol{K}$. 为此，先求解

$$\begin{aligned} \underset{\boldsymbol{W},\boldsymbol{Y},\nu}{\text{minimize}} \quad & \nu + \eta\|\boldsymbol{Y}\|_2 \\ \text{subject to} \quad & \boldsymbol{W}\boldsymbol{A}_i^\top + \boldsymbol{A}_i\boldsymbol{W} + \boldsymbol{Y}^\top\boldsymbol{B}_i^\top + \boldsymbol{B}_i\boldsymbol{Y} \preceq -\boldsymbol{I}, i = 1, \ldots, N, \\ & \boldsymbol{I} \preccurlyeq \boldsymbol{W} \preccurlyeq \nu\boldsymbol{I}, \end{aligned} \tag{15.25}$$

然后令$\boldsymbol{K} = \boldsymbol{Y}\boldsymbol{W}^{-1}$。

例15.5. (**倒立摆的鲁棒镇定**) 考虑一个简化的倒立摆模型，如图15.7所示，系统由长为ℓ的刚性杆和顶端的质量为m的点组成，其中$\theta(t)$表示杆关于竖直坐标轴的角位移，$u(t)$是施加在连接点的输入扭矩，g 是重力加速度。针对该系统牛顿平衡方程是

$$m\ell^2\ddot{\theta}(t) = mg\ell\sin\theta(t) + u(t).$$

这个二阶非线性微分方程描述了系统的动力学。然而，如果角位移$\theta(t)$持续很小时，即$|\theta(t)| \simeq 0$时，我们可以利用$\sin\theta(t) \simeq \theta(t)$这个事实，用一个线性微分方程

$$m\ell^2\ddot{\theta}(t) = mg\ell\theta(t) + u(t)$$

来近似这个非线性方程. 引入状态变量$x_1(t) = \theta(t), x_2 = \dot{\theta}(t)$，这个连续时间LTI系统为

$$\dot{x} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}\boldsymbol{u}, \boldsymbol{A} = \begin{bmatrix} 0 & 1 \\ \alpha & 0 \end{bmatrix}, \boldsymbol{B} = \begin{bmatrix} 0 \\ \beta \end{bmatrix},$$

其中$\alpha \doteq \frac{g}{\ell}, \beta \doteq \frac{1}{m\ell^2}$. 对于参数任何取值(正的)，这个系统都是不稳定的。假设给定了如下的数值：

$$g = 9.8\mathrm{m/s}^2,\ m = 0.5\mathrm{kg}, \ell \in [0.9, 1.1]\mathrm{m},$$

即我们不能确定杆的长度。我们希望设计一个状态控制反馈律$u(t) = Kx(t)$能鲁棒地镇定该系统。为此，我们观察到系统矩阵在由两个顶点系统定义的线段上变化，具体的顶点系统

$$\boldsymbol{A}_1 = \begin{bmatrix} 0 & 1 \\ 10.8889 & 0 \end{bmatrix}, \boldsymbol{B}_1 = \begin{bmatrix} 0 \\ 2.4691 \end{bmatrix};$$

$$\boldsymbol{A}_2 = \begin{bmatrix} 0 & 1 \\ 8.9091 & 0 \end{bmatrix}, \boldsymbol{B}_2 = \begin{bmatrix} 0 \\ 1.6529 \end{bmatrix}.$$

因此，由多面体系统的关于稳定性的充分条件，取$N = 2, \eta = 1$并求解凸规划问题(15.25)得到控制增益$\boldsymbol{K}$。利用CVX(一种求解凸优化的软件)求得数值解，得到鲁棒镇定控制器$\boldsymbol{K} = -[6.7289, 2.3151]$及Lyapunov判据

$$\boldsymbol{P} = \boldsymbol{W}^{-1} = \begin{bmatrix} 0.8796 & 0.2399 \\ 0.2399 & 0.5218 \end{bmatrix}.$$

注记15.1 **频域分析**. 前面几小节的分析和设计方法都是基于系统的**时域**(time-domain)表示。另一种可选的经典方法是在变换域上分析系统，通常称这个域为**频域**(frequency domain)。一个特殊的映射将时域信号变换成复变量的函数，从而完成信号从时域到频域的转变，称这个映射为(单边)**拉普拉斯变换**(Laplace transform)，其将时域信号$\boldsymbol{w}(t) \in \mathbb{C}^n$映射为函数

$$W(s) \doteq \int_0^\infty \boldsymbol{w}(t) e^{-st} \boldsymbol{d}t,$$

其中$s \in \mathbb{C}$. 如果对所有的s，上面的积分均收敛到一个有限值，则$W(s)$是有定义的。将$W(s)$在s取值为纯虚数时的限制记为

$$\hat{W}(\omega) \doteq W(s)|_{s=\jmath\omega} = \int_0^\infty \boldsymbol{w}(t) e^{-\jmath\omega t} \boldsymbol{d}t,$$

而且当$\boldsymbol{w}(t)$是一个**因果信号**[1]，那么该限制与$\boldsymbol{w}(t)$ 的傅立叶变换是一致的。如果$\boldsymbol{w}$ 是实数，那么$\hat{\boldsymbol{W}}(-\omega)^\top = \hat{\boldsymbol{W}}(\omega)^*$，这里*代表转置共轭。由

[1]即当$t < 0$时取值全为零的信号. 通常假定控制中的信号是因果的，因为习惯上假定瞬态$t = 0$为系统启动的时间.

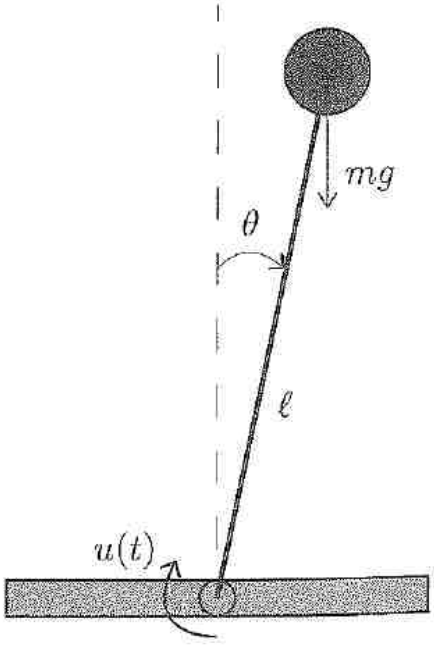

图 15.7: 一个倒立摆

此知道，对一个实信号，它的范数的平方$\|\hat{\boldsymbol{W}}(\omega)\|_2^2 = \hat{\boldsymbol{W}}(-\omega)^*\hat{\boldsymbol{W}}(\omega) = \hat{\boldsymbol{W}}(-\omega)^\top\hat{\boldsymbol{W}}(\omega)$ 关于ω 是中心对称的。

对于一个连续时间LTI系统

$$\dot{\boldsymbol{x}}(t) = \boldsymbol{A}\boldsymbol{x}(t) + \boldsymbol{B}\boldsymbol{u}(t), \boldsymbol{x}(0) = \boldsymbol{0}, \tag{15.26}$$

$$\boldsymbol{y}(t) = \boldsymbol{C}\boldsymbol{x}(t) + \boldsymbol{D}\boldsymbol{u}(t), \tag{15.27}$$

其中$\boldsymbol{x}(t) \in \mathbb{R}^n$是状态，$\boldsymbol{u}(t) \in \mathbb{R}^p$是输入，$\boldsymbol{y}(t) \in \mathbb{R}^m$是输出，并且系统矩阵$(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{D})$的维数是相容的。定义系统的**传递矩阵**(transfer matrix)为

$$\boldsymbol{H}(s) \doteq \boldsymbol{C}(sI - \boldsymbol{A})^{-1}\boldsymbol{B} + \boldsymbol{D}.$$

在Laplace域中，可以将输入和输出的关系间接地表述为

$$\boldsymbol{Y}(s) = \boldsymbol{H}(s)\boldsymbol{U}(s),$$

其中$\boldsymbol{U}(s), \boldsymbol{Y}(s)$分别是$\boldsymbol{u}(t), \boldsymbol{y}(t)$的Laplace变换。从而传递矩阵能够简单地描述系统在Laplace域的某些方面的行为. 矩阵$\boldsymbol{H}(s)$的元素$H_{ij}(s)$是实信号$h_{ij}(t)$的Laplace变换，它表示着当输入向量的第$_\mathrm{J}$ 个元素$u_j(t)$是一个脉冲(狄拉克δ函数)，其它输入$\boldsymbol{u}_k(t), k \neq {}_j$，均为 0 时，输出向量的第$_i$个

元$y_i(t)$。当然，如果输入信号的傅里叶变换是$\boldsymbol{U}(\omega)$，并且系统是稳定的，那么输出信号的傅立叶变换是

$$\hat{\boldsymbol{Y}}(\omega) = \boldsymbol{H}(\jmath\omega)\hat{\boldsymbol{U}}(\omega).$$

这里的$\hat{\boldsymbol{H}}(\omega) \doteq \boldsymbol{H}(\jmath\omega), \omega \geq 0$，就是所谓的系统的**频率响应**(frequency response)。

§15.3.3 离散时间的Lyapunov稳定性和镇定

和连续时间的情况类似，称一个离散时间 LTI 系统

$$\boldsymbol{x}(k+1) = \boldsymbol{A}\boldsymbol{x}(k) + \boldsymbol{B}\boldsymbol{u}(k) \tag{15.28}$$

是(渐进)稳定的，如果它的自由响应趋于零，并与初值无关。系统(15.28)是稳定的当且仅当存在 $\boldsymbol{P} \succ \boldsymbol{0}$ 满足

$$\boldsymbol{A}^\top \boldsymbol{P}\boldsymbol{A} - \boldsymbol{P} \prec \boldsymbol{0}, \tag{15.29}$$

或者，由齐次性，等价地有

$$\exists \boldsymbol{P} \succeq \boldsymbol{I} : \boldsymbol{A}^\top \boldsymbol{P}\boldsymbol{A} - \boldsymbol{P} \preceq -\boldsymbol{I}.$$

同样，定义 $\boldsymbol{W} = \boldsymbol{P}^{-1}$ 并且给方程(15.29)左乘 $\boldsymbol{W}$，右乘 $\boldsymbol{W}$ ，我们得到条件

$$\exists \boldsymbol{W} \succ \boldsymbol{0} : \boldsymbol{W} - \boldsymbol{W}\boldsymbol{A}^\top \boldsymbol{W}^{-1}\boldsymbol{A}\boldsymbol{W} \succ \boldsymbol{0}.$$

并将其去齐次化，得

$$\exists \boldsymbol{W} \succeq \boldsymbol{I} : \boldsymbol{W} - \boldsymbol{W}\boldsymbol{A}^\top \boldsymbol{W}^{-1}\boldsymbol{A}\boldsymbol{W} \succeq \boldsymbol{I}.$$

根据Schur补的性质，可将上面的条件写成仅仅是关于 $\boldsymbol{W}$ 的LMI条件

$$\begin{bmatrix} \boldsymbol{W} - \boldsymbol{I} & \boldsymbol{W}\boldsymbol{A}^\top \\ \boldsymbol{A}\boldsymbol{W} & \boldsymbol{W} \end{bmatrix} \succeq \boldsymbol{0}. \tag{15.30}$$

该表述在反馈控制设计中很有用：假设我们希望为系统(15.28)设计一个状态反馈律 $\boldsymbol{u}(k) = \boldsymbol{K}\boldsymbol{x}(k)$，那么受控系统是

$$\boldsymbol{x}(k+1) = (\boldsymbol{A} + \boldsymbol{B}\boldsymbol{K})\boldsymbol{x}(k),$$

该系统是稳定的当且仅当(应用(15.30))

$$\begin{bmatrix} \boldsymbol{W} - \boldsymbol{I} & \boldsymbol{W}(\boldsymbol{A} + \boldsymbol{B}\boldsymbol{K})^\top \\ (\boldsymbol{A} + \boldsymbol{B}\boldsymbol{K})\boldsymbol{W} & \boldsymbol{W} \end{bmatrix} \succeq \boldsymbol{0}.$$

定义新变量 $\boldsymbol{Y}=\boldsymbol{K}\boldsymbol{W}$，上式就变成一个关于 $\boldsymbol{W}$ 和 $\boldsymbol{Y}$ 的 LMI 条件。从而，求解关于 $\boldsymbol{W},\boldsymbol{Y}$ 的最优化问题，即

$$\begin{aligned}&\underset{\boldsymbol{W},\boldsymbol{Y},\nu}{\text{minimize}} \quad \nu+\eta\|\boldsymbol{Y}\|_2\\ &\text{subject to} \quad \begin{bmatrix}\boldsymbol{W}-\boldsymbol{I} & \boldsymbol{W}\boldsymbol{A}^\top+\boldsymbol{Y}^\top\boldsymbol{B}^\top\\ \boldsymbol{A}\boldsymbol{W}+\boldsymbol{B}\boldsymbol{Y} & \boldsymbol{W}\end{bmatrix}\succeq\boldsymbol{0},\\ &\qquad\qquad \boldsymbol{W}\preceq\nu\boldsymbol{I},\end{aligned} \tag{15.31}$$

并且令 $\boldsymbol{K}=\boldsymbol{Y}\boldsymbol{W}^{-1}$，即得镇定控制增益$\boldsymbol{K}$。用和之前连续时间系统基本相似的思路，也可以将这种方法推广以解决不确定的多面体系统的鲁棒镇定问题。

§15.4 练习

习题15.1. (稳定性和特征值) 证明连续时间系统(15.20)是渐进稳定的(或者简称为稳定的)当且仅当$\boldsymbol{A}$矩阵的所有特征值$\lambda_i(\boldsymbol{A})$的实部是(严格)负的, $i=1,\cdots,n$。

证明离散时间系统(15.28)是稳定的当且仅当矩阵$\boldsymbol{A}$的所有特征值$\lambda_i(A)$的模(严格)小于1, $i=1,\cdots,n$。

提示： 利用表达式$\boldsymbol{x}(t)=\mathrm{e}^{\boldsymbol{A}t}\boldsymbol{x}_0$和$\boldsymbol{x}(k)=\boldsymbol{A}^k\boldsymbol{x}_0$ 分别表示连续时间系统和离散时间系统的自由响应。你可以在$\boldsymbol{A}$是可对角化的假设下推导你的证明。

习题15.2. (信号范数) 一个连续时间信号$\boldsymbol{w}(t)$映射$t\in\mathbb{R}$到$\boldsymbol{w}(t)\in\mathbb{C}^m$ 或者 $\mathbb{R}^m$。信号的**能量**(energy)定义为

$$E(\boldsymbol{w})\doteq\|\boldsymbol{w}\|_2^2=\int_{-\infty}^{\infty}\quad\|\boldsymbol{w}(t)\|_2^2\boldsymbol{d}t,$$

其中$\|\boldsymbol{w}\|_2$是信号的2范数。能量有限的信号类指上面的2范数有限的信号全体。

周期信号是典型的能量无限的信号。对一个周期为T的信号，我们定义它的**功率**(power)为

$$P(\boldsymbol{w})\doteq\frac{1}{T}\int_{t_0}^{t_0+T}\quad\|\boldsymbol{w}(t)\|_2^2\boldsymbol{d}t.$$

(a) 计算谐波信号$\boldsymbol{w}(t)=\boldsymbol{v}\mathrm{e}^{j\omega t},\boldsymbol{v}\in\mathbb{R}^m$，和因果指数信号$\boldsymbol{w}(t)=\boldsymbol{v}\mathrm{e}^{at},a<0,t\geq0,\omega(t)=0,t<0$,的能量.

(b) 计算谐波信号$\boldsymbol{w}(t)=\boldsymbol{v}\mathrm{e}^{j\omega t}$和正弦信号$\boldsymbol{w}(t)=\boldsymbol{v}\sin(\omega t)$的功率.

习题15.3. (系统的状态演化的能量上界) 考虑一个连续时间LTI系统$\dot{\boldsymbol{x}}(t)=\boldsymbol{A}\boldsymbol{x}(t), t\geq 0$，没有输入(称这种系统是**自治的**(autonomous))，输出为$\boldsymbol{y}(t)=\boldsymbol{C}\boldsymbol{x}$。我们希望求出系统的输出信号所包含的能量，其由指标

$$J(\boldsymbol{x}_0)\doteq\int_0^\infty \quad \boldsymbol{y}(t)^\top\boldsymbol{y}(t)\boldsymbol{d}t=\int_0^\infty \quad \boldsymbol{x}(t)^\top\boldsymbol{Q}\boldsymbol{x}(t)\boldsymbol{d}t,$$

所度量，这里$\boldsymbol{Q}\doteq\boldsymbol{C}^\top\boldsymbol{C}\succeq\boldsymbol{0}$.

(a) 证明：如果系统稳定，那么对所有的$\boldsymbol{x}_0$，$J(\boldsymbol{x}_0)<\infty$.

(b) 证明：如果系统稳定，并且存在一个矩阵$\boldsymbol{P}\succeq\boldsymbol{0}$,使得

$$\boldsymbol{A}^\top\boldsymbol{P}+\boldsymbol{P}\boldsymbol{A}+\boldsymbol{Q}\preceq\boldsymbol{0},$$

那么$J(\boldsymbol{x}_0)\leq\boldsymbol{x}_0^\top\boldsymbol{P}\boldsymbol{x}_0$成立。提示：考虑二次型$V(\boldsymbol{x}(t))=\boldsymbol{x}(t)^\top\boldsymbol{P}\boldsymbol{x}(t)$，计算它对时间的导数。

(c) 对于给定的初始条件，解释如何计算一个系统的状态能量的最小上界。

习题15.4. (系统增益)一个系统的**增益**(gain)是指从输入信号到输出信号的最大的能量放大率。任意具有有限能量的输入信号$\boldsymbol{u}(t)$通过稳定系统后，会被映射成一个同样具有有限能量的输出信号$\boldsymbol{y}(t)$。**帕塞瓦恒等式**(Parseval identity)给出了信号$\boldsymbol{w}(t)$在时域的能量和信号在频域的能量(参考注记15.1)的关系，即

$$E(\boldsymbol{w})\doteq\|\boldsymbol{w}\|_2^2=\int_{-\infty}^{\infty}\|\boldsymbol{w}(t)\|_2^2\boldsymbol{d}t=\frac{1}{2\pi}\int_{-\infty}^{\infty}\|\hat{\boldsymbol{W}}(\omega)\|_2^2\boldsymbol{d}\omega\doteq\|\hat{\boldsymbol{W}}\|_2^2.$$

系统(15.26)的**能量增益**(energy gain)定义为

$$\mathrm{EG}\doteq\sup_{\boldsymbol{u}(t):\|\boldsymbol{u}\|_2<\infty,\boldsymbol{u}\neq 0}\frac{\|\boldsymbol{y}\|_2^2}{\|\boldsymbol{u}\|_2^2},$$

(a) 利用上面的信息证明：对于一个稳定的系统，

$$\mathrm{EG}\leq\sup_{\omega\geq 0}\|\boldsymbol{H}(j\omega)\|_2^2,$$

其中$\|\boldsymbol{H}(j\omega)\|_2$是系统(15.26)的传递矩阵在$s=j\omega$处的谱范数。称系统的能量增益的平方根为$\mathcal{H}_\infty$范数，记作$\|\boldsymbol{H}\|_\infty$

提示：运用帕塞瓦恒等式并证明某积分有上界。注意等式实际上蕴含在前面的公式中，但是并没有要求你去证明该事实。

(b) 假设系统(15.26)是稳定的，$\boldsymbol{x}(0)=\boldsymbol{0},\boldsymbol{D}=\boldsymbol{0}$。证明：如果存在$\boldsymbol{P}\succeq\boldsymbol{0}$使得

$$\begin{bmatrix}\boldsymbol{A}^\top\boldsymbol{P}+\boldsymbol{P}\boldsymbol{A}+\boldsymbol{C}^\top\boldsymbol{C} & \boldsymbol{P}\boldsymbol{B}\\ \boldsymbol{B}^\top\boldsymbol{P} & -\gamma^2\boldsymbol{I}\end{bmatrix}\preceq\boldsymbol{0}, \tag{15.32}$$

那么

$$\|\boldsymbol{H}\|_\infty\leq\gamma.$$

设计一种计算格式，其能得到系统能量增益的最小可能的上界γ^*。

提示：定义二次函数$V(\boldsymbol{x})=\boldsymbol{x}^\top\boldsymbol{P}\boldsymbol{x}$，并观察 V 对时间t的导数. 沿着系统(15.26)的轨道，这个导数是

$$\frac{\boldsymbol{d}V(\boldsymbol{x})}{\boldsymbol{d}t}=\boldsymbol{x}^\top\boldsymbol{P}\dot{\boldsymbol{x}}+\dot{\boldsymbol{x}}^\top\boldsymbol{P}\boldsymbol{x}.$$

然后说明LMI条件(15.32)等价于条件:

$$\frac{\boldsymbol{d}V(\boldsymbol{x})}{\boldsymbol{d}t}+\|\boldsymbol{y}\|^2-\gamma^2\|\boldsymbol{u}\|^2\leq 0,\forall\boldsymbol{x},\boldsymbol{u},\text{satisfying (15.26)},$$

并且这反过来蕴含着$\|\boldsymbol{H}\|_\infty\leq\gamma$。

习题15.5. (扩展超稳定矩阵)设矩阵$\boldsymbol{A}\in\mathbb{R}^{n\times n}$，如果存在$\boldsymbol{d}\in\mathbb{R}^n$使得

$$\sum_{j\neq i}|a_{ij}|d_j<-a_{ii}d_i,d_i>0,i=1,\cdots,n,$$

则称矩阵$\boldsymbol{A}$是连续时间**扩展超稳定的**(extended wuperstable)[2](记为$\boldsymbol{A}\in E_c$). 类似的，如果存在$\boldsymbol{d}\in\mathbb{R}^n$使得

$$\sum_{j=1}^{n}|a_{ij}|d_j<d_i,d_i>0,i=1,\cdots,n,$$

则称矩阵$\boldsymbol{A}$是离散时间扩展超稳定的(记为$\boldsymbol{A}\in E_d$).

如果$\boldsymbol{A}\in E_c$，那么它所有的特征值的实部都小于0，因此相应的连续时间LTI系统$\dot{\boldsymbol{x}}=\boldsymbol{A}\boldsymbol{x}$是稳定的。类似的，如果$\boldsymbol{A}\in E_d$，那么$\boldsymbol{A}$的所有特征值的模小于 1，因此相应的离散时间LTI 系统$\boldsymbol{x}(k+1)=\boldsymbol{A}\boldsymbol{x}(k)$是稳定的。扩展超稳定提供了稳定性的充分条件，它的优势在于仅需检验一列线性不等式的可行性。

[2]见B.T. Polyak, Extended superstability in control theory, *Automation and Remote Control*, 2004.

(a) 给定一个连续时间系统$\dot{\boldsymbol{x}} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}\boldsymbol{u}, \boldsymbol{x} \in \mathbb{R}^n, \boldsymbol{u} \in \mathbb{R}^m$，描述一种有效方法，其设计的形为$\boldsymbol{u} = -\boldsymbol{K}\boldsymbol{x}$的状态反馈控制律使得受控系统是扩展超稳定的。

(b) 给定离散时间系统$\boldsymbol{x}(k+1) = \boldsymbol{A}\boldsymbol{x}(k) + \boldsymbol{B}\boldsymbol{u}(k)$，假设矩阵$\boldsymbol{A}$具有区间不确定性，即

$$a_{ij} = \hat{a}_{ij} + \delta_{ij}, i, j = 1, \cdots, n$$

其中$\hat{a}_{ij}$是名义上的元素，δ_{ij} 是一个不确定项. 对给定的$r_{ij} \geq 0$，仅知道$|\delta_{ij}| \leq \rho r_{ij}$。定义扩展超稳定的半径是使得$\boldsymbol{A}$对$\rho \geq 0$时所有可容许的不确定量都是扩展超稳定的那个最大的，记为ρ^*。给出一种确定ρ^*的计算方法。

第十六章　工程设计

凸模型和凸优化在过去几十年对工程设计产生巨大的影响. 随着可靠的最优化技术(起先的线性规划，linear programming，后来的二次规划，quadratic programming，和锥规划)的出现，工程师开始回顾各种分析和设计问题，发现可以用凸模型来表述它们，因此可以被有效求解. 整个领域，诸如自动控制、电路设计，在二十世纪九十年代因为引入了凸规划的方法论(特别是SDP)而发生了革命性的变化. 今天，凸模型常被用来解决相关问题中的一部分，例如结构力学、识别、过程控制、滤波器设计、电子电路的宏建模、物流和管理、网络设计等等. 在这一章中，将详述从这些应用的一部分.

§16.1　数字滤波器设计

一个单输入和单输出的数字滤波器是一个动力系统，具有标量输入信号$u(t)$和标量输出信号$y(t)$，这里t代表了(离散的)时间变量. 有限冲击响应(finite-impulse response, FIR)滤波器是一种特殊的滤波器，形如

$$y(t)=\sum_{i=0}^{n-1}h_iu(t-i),\quad t\in\mathbb{Z},$$

这里$h_0,\dots,h_{n-1}$叫做滤波器的**脉冲响应**(impulse response)，该滤波器得名源于如下事实：滤波器对离散时间脉冲

$$u(t)=\begin{cases}1 & t=0,\\ 0 & \text{其它},\end{cases}$$

的时间响应刚好是有限支集信号

$$y(t)=\begin{cases}h_t & 0\le t\le n-1,\\ 0 & \text{其它}.\end{cases}$$

脉冲响应的离散傅里叶变换是一个复值函数$H:\mathbb{R}\to\mathbb{C}$，其值

$$H(\omega)=\sum_{t=0}^{n-1}h_t\mathrm{e}^{-\jmath\omega t},\quad \omega\in[-\pi,\pi].$$

该函数非常重要，因为它表明了滤波器如何对周期信号做出响应. 准确地说，如果输入是复指数$u(t)=\mathrm{e}^{\jmath\omega_0t}$，则输出信号将被放缩成复指数$y(t)=$

$H(\omega_0)\mathrm{e}^{\jmath\omega_0 t}$. 因为$H(\omega)$以$2\pi$ 为周期并且$H^*(\omega)=H(-\omega)$，我们可以将分析限定在规范区间$[0,\pi]$上. 一个简单的FIR滤波器是滑动平均滤波器：长度为2的滑动平均滤波形如

$$y(t)=\frac{1}{2}(u(t)+u(t-1)),\quad t\in\mathbb{Z}.$$

§16.1.1 线性相位FIR滤波器

I型线性相位滤波器(Type I linear-phase filter)是一类特殊的FIR滤波器，它的特征是项数为奇数，即$n=2N+1$，并且脉冲响应关于中点对称，即

$$h_t=h_{n-1-t},\quad t=0,\dots,n-1.$$

这里的修饰语“线性相位”源于如下事实：对这类滤波器，它的频率响应形如

$$H(\omega)=\mathrm{e}^{-\jmath\omega N}\tilde{H}(\omega),\quad \omega\in[0,\pi],$$

其中$\tilde{H}(\omega)$是**实值**函数

$$\tilde{H}(\omega)=h_N+2\sum_{t=0}^{N-1}h_t\cos((N-t)\omega).$$

称函数$\tilde{H}(\omega)$为滤波器的**振幅响应**(amplitude response). 需要注意的是$H(\omega)$的实际相位也许是不连续的(因为实际相位为$-\mathrm{sgn}(\tilde{H}(\omega))\omega N$，并且$H(\omega)$的模可以表示为

$$|H(\omega)|=|\tilde{H}(\omega)|.$$

有些情况下，为了方便工程中采用“连续线性相位”$\theta(\omega)=-\omega N$和振幅响应函数$\tilde{H}(\omega)$(它是实的，但是能够取正值或者负值)来表示$H(\omega)$，而不是直接用$H(\omega)$实际的相位和模来表示. 例如，我们例子中的$\tilde{H}(\omega)$是设计参数$\boldsymbol{h}$的一个简单线性函数，即，我们可以写为

$$\tilde{H}(\omega)=\boldsymbol{a}^\top(\omega)\boldsymbol{h},\ \boldsymbol{h}=\begin{bmatrix}h_0\\h_1\\\vdots\\h_{N-1}\\h_N\end{bmatrix},\boldsymbol{a}(\omega)=\begin{bmatrix}2\cos\omega N\\2\cos\omega(N-1)\\\vdots\\2\cos\omega\\1\end{bmatrix}.$$

我们进一步发现，不失一般性，可选取脉冲响应向量$\boldsymbol{h}$，使得$\tilde{H}(0)>0$.

§16.1.2 低通FIR设计规范

与FIR滤波器有关的设计问题通常涉及选取滤波器的脉冲响应$\boldsymbol{h}$，以便获得具有理想形状的振幅响应. 例如，要确保滤波器阻止高频信号，但是允许低频信号通过(低通滤波器). 可以将这些要求对应到下面规定的滤波器的振幅响应，也可以参考图16.1所示的振幅设计约束.

- 阻带约束：

$$-\delta_s \le \tilde{H}(\omega) \le \delta_s, \quad \omega \in [\Omega_s, \pi],$$

其中Ω_s是阻带边沿频率，$\delta_s > 0$对应着我们对高频需要获取的衰减水平.

- 通带约束：

$$1 - \delta_p \le \tilde{H}(\omega) \le 1 + \delta_p, \quad \omega \in [0, \Omega_p],$$

其中Ω_p是通带截止频率，$\delta_p > 0$界定了低频通带波纹.

注意到，对极小的δ_s和δ_p，上述约束近似等同于下面的两个关于$H(\omega)$的模取对数(以10为底)意义下的约束：

$$\log|H(\omega)| \le \log\delta_s, \quad \omega \in [\Omega_s, \pi],$$
$$-\log\delta_p \le \log|H(\omega)| \le \log\delta_p, \quad \omega \in [0, \Omega_p].$$

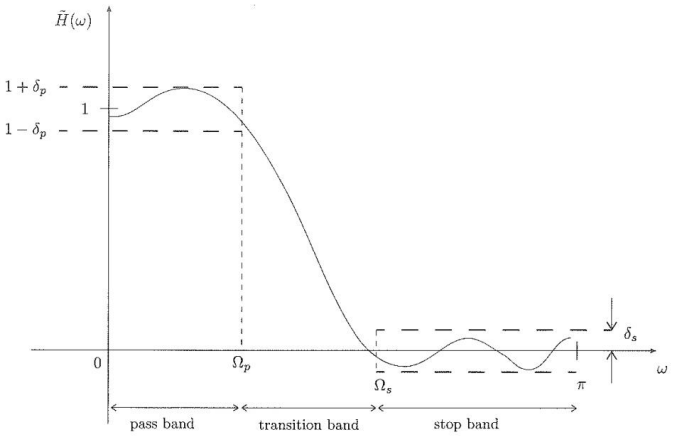

图 16.1: 低通FIR设计的模板

§16.1.3 通过线性规划设计FIR

注意到上述设计约束涉及到各自区间上关于区间中**每一个频率**ω处的线性不等式约束集. 因此，它们实际上涉及到**无限多个**线性约束. 为了克服该问题，这里简单地将频率区间进行离散处理. 做法就是在对应的区间上取有限个不同的频率点，对这些点处的频率要求约束满足，而不是强迫区间上的每个频率都满足约束. 将在高频区间$[\Omega_s, \pi]$取的点记为$\omega_i,\ i = 1, \ldots, N_s$，并由**有限个**关于$\boldsymbol{h}$ 的线性不等式

$$-\delta_s \le \boldsymbol{a}^\top(\omega_i)\boldsymbol{h} \le \delta_s, \quad i = 1, \ldots, N_s$$

来近似阻带约束. 同样的，将在低频区间$[0, \Omega_p]$上取的点记为$\omega_i,\ i = N_s + 1, \ldots, N_s + N_p$，得到通带约束的近似：

$$1 - \delta_p \le \boldsymbol{a}^\top(\omega_i)\boldsymbol{h} \le 1 + \delta_p, \quad i = N_s + 1, \ldots, N_s + N_p.$$

在前文提出的假设条件下，低通FIR滤波器设计问题可以通过各种方式表述成线性规划. 例如，我们限定要求的阻带衰减水平$\delta_s > 0$，寻找$\boldsymbol{h}$和δ_s 来最小化通带波纹，即

$$\begin{aligned}
&\underset{\boldsymbol{h}\in\mathbb{R}^{N+1},\delta_p\in\mathbb{R}}{\text{minimize}} \quad \delta_p \\
&\text{subject to} \ -\delta_s \le \boldsymbol{a}^\top(\omega_i)\boldsymbol{h} \le \delta_s, \qquad i = 1, \ldots, N_s, \\
&\qquad\qquad 1 - \delta_p \le \boldsymbol{a}^\top(\omega_i)\boldsymbol{h} \le 1 + \delta_p, \qquad i = N_s + 1, \ldots, N_s + N_p.
\end{aligned}$$

同样地，还可以限定通带波纹$\delta_p > 0$，来最小化阻带衰减水平δ_s，得到的LP问题形如

$$\begin{aligned}
&\underset{\boldsymbol{h}\in\mathbb{R}^{N+1},\delta_s\in\mathbb{R}}{\text{minimize}} \quad \delta_s \\
&\text{subject to} \ -\delta_s \le \boldsymbol{a}^\top(\omega_i)\boldsymbol{h} \le \delta_s, \qquad i = 1, \ldots, N_s, \\
&\qquad\qquad 1 - \delta_p \le \boldsymbol{a}^\top(\omega_i)\boldsymbol{h} \le 1 + \delta_p, \qquad i = N_s + 1, \ldots, N_s + N_p.
\end{aligned}$$

同样地，还可以对参数$\mu > 0$取不同值，来最小化δ_p与δ_s的加权值，即

$$\begin{aligned}
&\underset{\boldsymbol{h}\in\mathbb{R}^{N+1},\delta_s\in\mathbb{R},\delta_p\in\mathbb{R}}{\text{minimize}} \quad \delta_s + \mu\delta_p \\
&\text{subject to} \ -\delta_s \le a^\top(\omega_i)h \le \delta_s, \qquad i = 1, \ldots, N_s, \\
&\qquad\qquad 1 - \delta_p \le a^\top(\omega_i)h \le 1 + \delta_p, \quad i = N_s + 1, \ldots, N_s + N_p.
\end{aligned}$$

这里向读者介绍一个受约束于给定通带波纹δ_p，最小化阻带衰减水平δ_s的数值实例. 这里选取参数$N = 10$(因此滤波器的抽头数$n = 2N+1 =$

21)，$\Omega_p = 0.35\pi, \Omega_s = 0.5\pi, \delta_p = 0.02$(即$-33.98$ dB)，且以$N_s = N_p = 100$将频率区间离散化，线性等距取点. 求解相应的LP问题，得到的最优阻带衰减$\delta_s = 0.0285$(即-30.9 dB). 振幅响应曲线见图16.2，相应的$|H(\omega)|$的对数见图16.3，滤波器系数(脉冲响应)见图16.4.

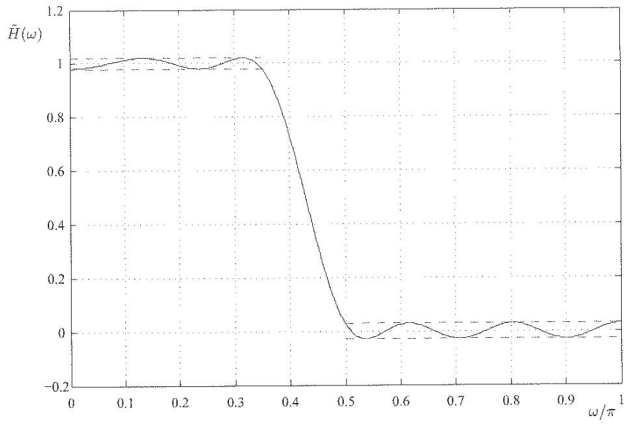

图 16.2: FIR滤波器振幅响应

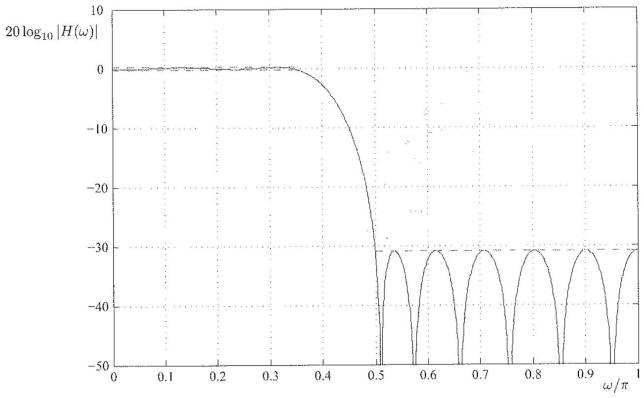

图 16.3: FIR滤波器的模作为(归一化)频率的函数

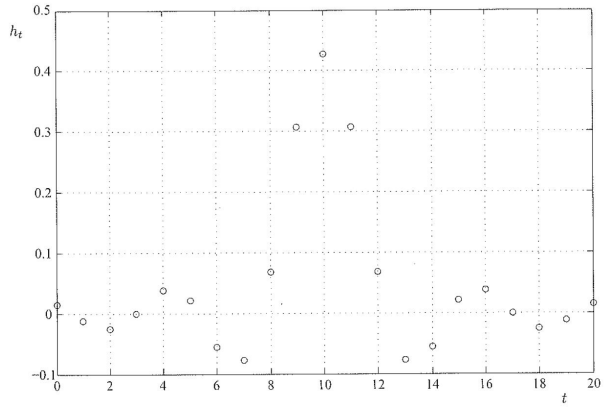

图 16.4: FIR滤波器的脉冲响应

§16.1.4 逼近参考匹配的滤波器设计

设计线性相位滤波器的一个可行方案，是提供一个“参考的”理想幅度响应，然后尝试寻找滤波器参数$\boldsymbol{h}$使得滤波器响应与参考响应在所有频率处尽可能地接近.

§16.1.4.1 最小二乘设计

设$\tilde{H}_{\mathrm{ref}}(\omega)$是给定的振幅响应，并且将频率区间$[0,\pi]$离散化为$\omega_i,\ i=1,\dots,M$. 一种可能的方法是寻找滤波器系数$\boldsymbol{h}$来极小化错配度量

$$\underset{\boldsymbol{h}}{\text{minimize}}\ \sum_{i=1}^{M}(\tilde{H}(\omega_i)-\tilde{H}_{\mathrm{ref}}(\omega_i))^2.$$

因为$\tilde{H}(\omega)=\boldsymbol{a}^{\top}(\omega)\boldsymbol{h}$，因此显然是一个最小二乘(least-square, LS)问题：

$$\underset{\boldsymbol{h}}{\text{minimize}}\ \|\boldsymbol{A}\boldsymbol{h}-\boldsymbol{b}\|_2^2,$$

这里

$$\boldsymbol{A}=\begin{bmatrix}\boldsymbol{a}^{\top}(\omega_1)\\ \vdots\\ \boldsymbol{a}^{\top}(\omega_M)\end{bmatrix},\quad \boldsymbol{b}=\begin{bmatrix}\tilde{H}_{\mathrm{ref}}(\omega_1)\\ \vdots\\ \tilde{H}_{\mathrm{ref}}(\omega_M)\end{bmatrix}.$$

这种方法的一个变形是对不同频率处的错配度量引入权重：选定一组权重$w_i\geq 0,\ i=1,\dots,M$，并且求解修正问题

$$\underset{\boldsymbol{h}}{\text{minimize}}\ \sum_{i=1}^{M}w_i^2\left(\tilde{H}(\omega_i)-\tilde{H}_{\mathrm{ref}}(\omega_i)\right)^2.$$

在这种设置下，ω_i的一个相对高的值意味着减小频率ω_i处的错配是很重要的. 相反，ω_i的一个相对小值则蕴含着ω_i处的错配是不太重要的. 由这种方法得到加权LS问题

$$\underset{\boldsymbol{h}}{\text{minimize}} \ \|\boldsymbol{W}(\boldsymbol{A}\boldsymbol{h}-\boldsymbol{b})\|_2^2, \quad \boldsymbol{W}=\text{diag}(w_1,\ldots,w_M).$$

用一个实例来说明这种方法. 考虑一个$N = 10$，$M = 200$的线性等距离散化频率的滤波器，当频率小于通带截止频率$\Omega_p = 0.35\pi$时，参考响应等于1；当$\omega > \Omega_s = 0.5\pi$时参考响应等于0. 在参考幅度在过渡带$[\omega_p, \omega_s]$从1线性递减到0. 利用常数(单位)频率权重，由LS 最优化问题得到的结果的振幅响应见图16.5，相应的模图见图16.6.

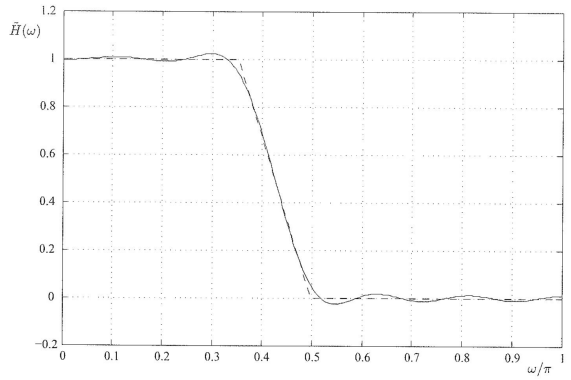

图 16.5: 由LS匹配得到的FIR滤波器的振幅响应，虚线表示的是参考响应$\tilde{H}_{\text{ref}}(\omega)$

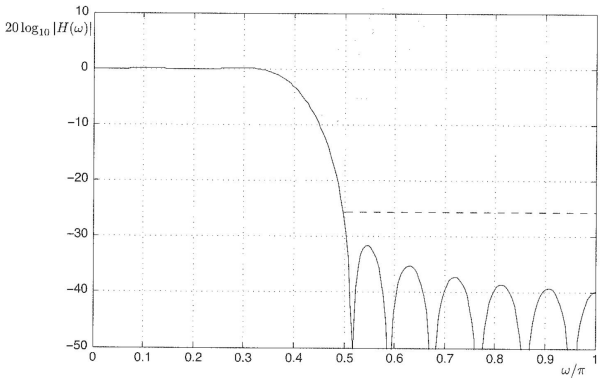

图 16.6: 由LS匹配得到的FIR滤波器的模，虚线显示了阻带边沿$\Omega_s = 0.5$处的模水平，其等于0.0518(即-25.71 dB)

§16.1.4.2 切比雪夫设计

易于计算LS设计的解. 然而，这种方法中我们没办法控制逐点的理想(参考)响应与滤波器实际的响应之间错配误差的最大值. 切比雪夫型设计便是将最小化最大加权错配作为目标，即

$$\underset{\boldsymbol{h}}{\text{minimize}} \max_{i=1,\ldots,M} w_i|\tilde{H}(\omega_i) - \tilde{H}_{\text{ref}}(\omega_i)|.$$

这一问题可以表述成如下LP：

$$\begin{aligned}
&\underset{\boldsymbol{h},\gamma}{\text{minimize}}\ \gamma \\
&\text{subject to } w_i(\boldsymbol{a}^\top(\omega_i)\boldsymbol{h} - \tilde{H}_{\text{ref}}(\omega_i)) \le \gamma, \qquad i = 1,\ldots,M, \\
&\qquad\qquad w_i(\boldsymbol{a}^\top(\omega_i)\boldsymbol{h} - \tilde{H}_{\text{ref}}(\omega_i)) \ge -\gamma, \qquad i = 1,\ldots,M.
\end{aligned}$$

将这一设计方法用于上文实例中的数据，得到的振幅响应见图16.7，对应的模图见图16.8. 由此得到的实际幅度响应与设定的振幅响应之间偏差的绝对值的最大者为$\gamma = 0.0313$(即-30.1 dB).

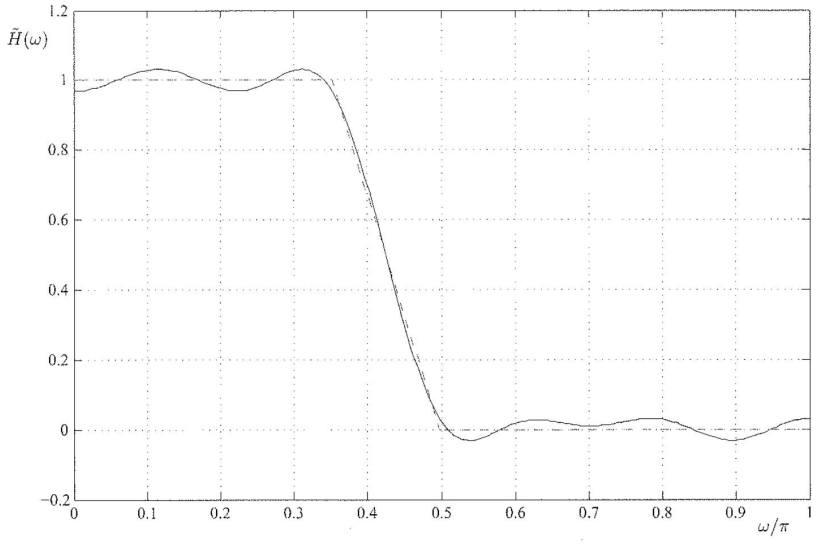

图 16.7: 由切比雪夫匹配得到的FIR滤波器的振幅响应，虚线是参考响应$\tilde{H}_{\text{ref}}(\omega)$

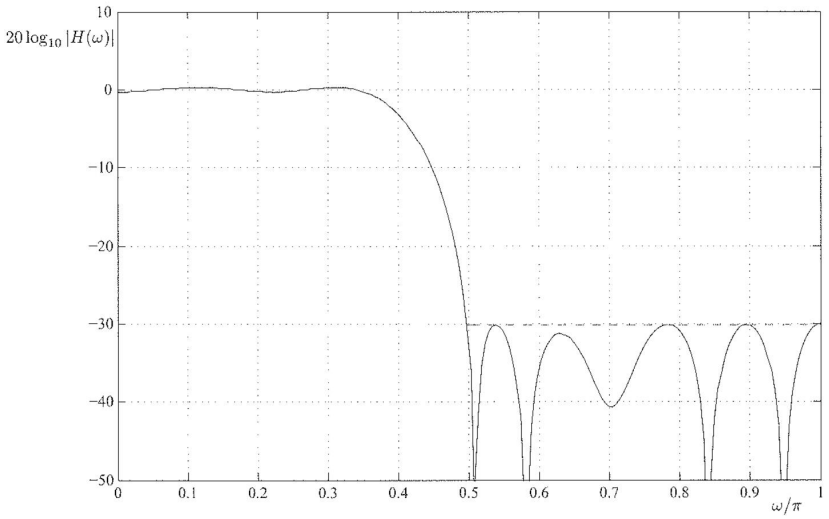

图 16.8: 由切比雪夫匹配得到的FIR滤波器的模，虚线表示的是阻带边沿$\Omega_s = 0.5$处的模水平，其等于0.0313(即−30.1 dB).

§16.2 天线阵列设计

在一个天线阵列中，对若干发射天线单元的输出进行线性组合，由此产生一个复合的阵列输出. 阵列输出具有方向图，其取决于在组合过程中使用的相对权重或比例因子. 权重设计的目标是去选取权重，以实现期望的方向图.

发射天线中的基本单元元件是各向同性的谐波振荡器，其发射波长为λ，频率为ω的球形单色波. 振荡器会产生电磁场，其在离天线距离为d的特定点$\boldsymbol{p}$处的电分量为

$$\frac{1}{d}\mathrm{Re}\left(\boldsymbol{z}\cdot\exp\left(\jmath\left(\omega t-\frac{2\pi d}{\lambda}\right)\right)\right),$$

其中$\boldsymbol{z}\in\mathbb{C}$是一个设计参数，用来缩放和改变电场的相位. 我们将这个复数称为天线单元的**权重**(weight). 现在，我们放置n个这样的振荡器，分别在点$\boldsymbol{p}_k\in\mathbb{R}^3,\ k=1,\dots,n$,处，复值权重分别为$\boldsymbol{z}_k\in\mathbb{C},\ k=1,\dots,n$. 这样，在点$\boldsymbol{p}\in\mathbb{R}^3$ 处的总电场强度由以下加权和给出：

$$E=\mathrm{Re}\left(\exp\left(\jmath\omega t\right)\cdot\sum_{k=1}^{n}\frac{1}{d_k}\boldsymbol{z}_k\cdot\exp\left(\frac{-2\pi\jmath d_k}{\lambda}\right)\right),$$

其中$d_k = \|\boldsymbol{p} - \boldsymbol{p}_k\|_2$是点$\boldsymbol{p}$到点$\boldsymbol{p}_k$的距离，$k = 1, \ldots, n$.

线性阵列的远场近似. 前面的公式在以下两个假设下可以近似简化：

(a) 振荡器形成线性阵列，即它们被放置在某一直线上的等距点上，如果在这条直线上建立x轴，那么这n个振荡器所处的网格点放置在x-轴上，即点可分别表示为$\boldsymbol{p}_k = \ell k \boldsymbol{e}_1,\ k = 1, \ldots, n$，其中$e_1 = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}$ (即$\mathbb{R}^3$的第一标准单位基向量)；

(b) 所考虑的点p离原点足够远：设$\boldsymbol{p} = r\boldsymbol{u}$，其中$\boldsymbol{u} \in \mathbb{R}^3$是给定方向的单位范数的向量，$r$ 是点$\boldsymbol{p}$离原点的距离，假定r足够大，以至于可以忽略天线阵列的几何尺寸，即$r \gg n\ell$，参见图16.9.

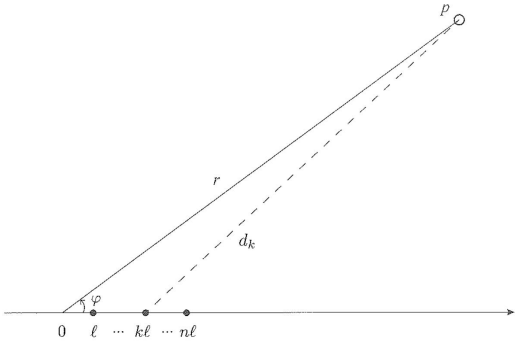

图 16.9: 线性天线阵列

对于线性阵列，电场强度E仅由阵列所在直线和上述假设(b)中的远点$\boldsymbol{p}$与原点连线之间的夹角φ近似决定. 因此，对足够小的$\frac{k\ell}{r}$，我们有

$$d_k = r\sqrt{1 + \left(\tfrac{k\ell}{r}\right)^2 + 2\left(\tfrac{k\ell}{r}\right)\cos\varphi} \simeq r + k\ell\cos\varphi.$$

这样，我们就得到了一个E的好的近似形式

$$E \simeq \frac{1}{r}\mathrm{Re}\left(\exp\left(\jmath\omega t - \frac{2\pi\jmath r}{\lambda}\right) \cdot D_{\boldsymbol{z}}(\varphi)\right),$$

其中函数$D_{\boldsymbol{z}} : [0, 2\pi] \to \mathbb{C}$被称为天线的**方向图**(diagram)：

$$D_{\boldsymbol{z}}(\varphi) \doteq \sum_{k=1}^{n} \boldsymbol{z}_k \cdot \exp\left(\frac{-2\pi\jmath k\ell\cos\varphi}{\lambda}\right). \tag{16.1}$$

我们使用下标“$\boldsymbol{z}$”是为了强调方向图由复值权重向量$\boldsymbol{z} = [\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n]$决定.

类似的结果适用于**接收器**(receiver)天线的线性阵列，如图16.10：具有频率ω和波长λ 的平面谐波以方向角φ入射并且通过阵列传播，信号的输出被转换成基带(复数)，由权重z_k进行加权，并求和，再次给出线性阵列波束方向图$D_{\boldsymbol{z}}(\varphi)$.

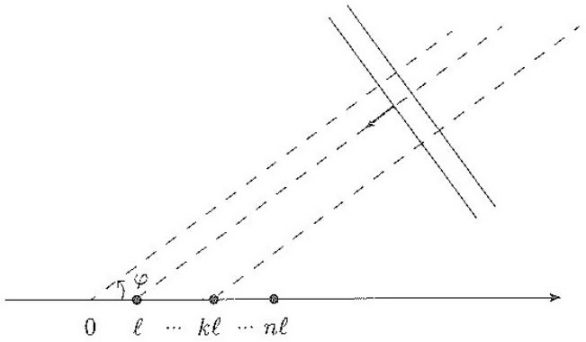

图 16.10: 一个接收器天线的线性阵列

§16.2.1 天线方向图赋形

可证明，天线方向图的模平方$|D_{\boldsymbol{z}}(\varphi)|^2$与天线发送的电磁场能量的方向密度成比例. 因此，我们感兴趣的是给天线方向图的模$|D_{\boldsymbol{z}}(\cdot)|$“赋形”（通过选择$\boldsymbol{z}$来实现），以便满足一些方向性要求. 注意到，当$\varphi$固定时，$D_{\boldsymbol{z}}(\varphi)$是权重向量$\boldsymbol{z}$的实部与虚部的线性函数. 特别地，由式(16.1)，我们有

$$D_{\boldsymbol{z}}(\varphi)=\boldsymbol{a}^{\top}(\varphi)\,\boldsymbol{z},\quad \boldsymbol{a}^{\top}(\varphi)=[a_1(\varphi)\cdots a_n(\varphi)],$$
$$a_k(\varphi)=\exp\left(\frac{-2\pi\jmath k\ell\cos\varphi}{\lambda}\right),\quad k=1,\ldots,n.$$

进一步注意到，$D_{\boldsymbol{z}}(\varphi)$的函数值是一个复数，因此，定义

$$\boldsymbol{a}_R(\varphi)=\mathrm{Re}(\boldsymbol{a}(\varphi)),\boldsymbol{a}_I(\varphi)=\mathrm{Im}(\boldsymbol{a}(\varphi)),\boldsymbol{\zeta}=\begin{bmatrix}\mathrm{Re}(\boldsymbol{z})\\ \mathrm{Im}(\boldsymbol{z})\end{bmatrix},$$
$$\boldsymbol{C}(\varphi)=\begin{bmatrix}\boldsymbol{a}_R^{\top}(\varphi) & -\boldsymbol{a}_I^{\top}(\varphi)\\ \boldsymbol{a}_I^{\top}(\varphi) & \boldsymbol{a}_R^{\top}(\varphi)\end{bmatrix},$$

我们得到

$$\begin{aligned}D_{\boldsymbol{z}}(\varphi) &= \begin{bmatrix}\boldsymbol{a}_R^{\top}(\varphi) & -\boldsymbol{a}_I^{\top}(\varphi)\end{bmatrix}\boldsymbol{\zeta}+\jmath\begin{bmatrix}\boldsymbol{a}_I^{\top}(\varphi) & \boldsymbol{a}_R^{\top}(\varphi)\end{bmatrix}\boldsymbol{\zeta},\\ |D_{\boldsymbol{z}}(\varphi)| &= \|\boldsymbol{C}(\varphi)\,\boldsymbol{\zeta}\|_2.\end{aligned}$$

一个典型的要求是天线沿着期望的方向（在给定角度上或其附近）很好地发射（或接收），而沿其它角度效果不佳. 以这种方式，发出的能量会集中在给定的“目标”方向上，比如说$\varphi_{\text{target}} = 0°$，并且在偏离该方向稍大的方向上能量较小. 另一类要求涉及由天线产生的热噪声功率.

归一化. 首先，我们归一化沿目标方向的能量. 当将所有权重乘以同一个非零常复数时，我们不改变能量的方向分布. 因此，通过归一化权重，即使得

$$D_{\boldsymbol{z}}(0) = 1,$$

我们没有任何损失. 该约束等价于决策变量$\boldsymbol{z} \in \mathbb{C}^n$的实部和虚部中的两个线性等式约束：

$$\begin{bmatrix} \boldsymbol{a}_R^\top(0) & -\boldsymbol{a}_I^\top(0) \end{bmatrix} \boldsymbol{\zeta} = 1, \quad \begin{bmatrix} \boldsymbol{a}_I^\top(0) & \boldsymbol{a}_R^\top(0) \end{bmatrix} \boldsymbol{\zeta} = 0.$$

旁瓣电平约束. 我们接下来定义“通带”$[-\phi, \phi]$，其中$\phi > 0$给定，在这个区间内，能量被集中；相应的“阻带”是该区间的外部. 为了满足能量被集中的需求，我们要求

$$|D_{\boldsymbol{z}}(\varphi)| \leq \delta, \quad \forall \varphi : |\varphi| \geq \phi,$$

其中δ是阻带上的期望衰减电平（有时也被称为旁瓣电平）. 事实上，旁瓣电平约束需要无穷多个约束. 处理这种连续无限多约束的一个实际可行方法是简单地离散化它们，即施加

$$|D_{\boldsymbol{z}}(\varphi_i)| \leq \delta, \quad i = 1, \ldots, N.$$

其中$\varphi_1, \ldots, \varphi_N$是在阻带中取值的$N$个间隔均匀的离散角度. 这是$N$个关于$\boldsymbol{z}$的实部和虚部的二次锥(second-order cone, SOC) 约束：

$$\|\boldsymbol{C}(\varphi_i)\boldsymbol{\zeta}\|_2 \leq \delta, \quad i = 1, \ldots, N.$$

举个例子，如图16.11所示，天线方向图的模必须通过右侧$\phi = 0°$上的点，否则将被包含在白色区域中. 在阻带（阴影区域）中，至少在离散点处，天线方向图的模必须保持低于δ.

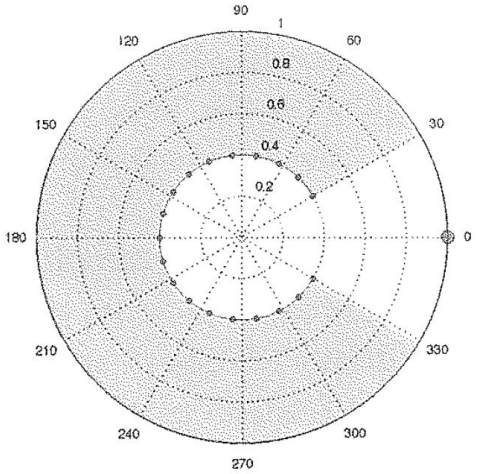

图 16.11: 关于天线方向图的约束

热噪声功率约束. 通常我们还希望控制由发射天线产生的热噪声功率. 结果证明，该功率与（复）向量z的欧几里得范数的平方成正比，即：

$$\text{热噪声功率} = \alpha \|\boldsymbol{z}\|_2^2 = \alpha \sum_{i=1}^{n} |\boldsymbol{z}_i|^2 .$$

§16.2.2 最小二乘设计法

第一种简化方法是最小二乘法，其通过考虑旁瓣电平衰减和热噪声功率之间的权衡来解决天线设计问题. 为了取代在每个离散角度φ_i处对阻带水平施加的约束，我们在目标函数中添加一个总阻带电平平方和的惩罚项，即考虑问题

$$\underset{\boldsymbol{z}}{\text{minimize}} \ \|\boldsymbol{z}\|_2^2 + \mu \sum_{i=1}^{N} |D_{\boldsymbol{z}}(\varphi_i)|^2, \quad \text{subject to } D_{\boldsymbol{z}}(0) = 1,$$

其中$\mu \geq 0$是权衡参数. 这是一个有等式约束的LS问题. 利用含$\boldsymbol{z}$的实部与虚部的变量$\boldsymbol{\zeta}$，可以把该问题更明晰地表示为：

$$\underset{\boldsymbol{\zeta}}{\text{minimize}} \ \|\boldsymbol{\zeta}\|_2^2 + \mu \boldsymbol{\zeta}^\top \boldsymbol{A} \boldsymbol{\zeta}, \quad \text{subject to } \boldsymbol{C}(0)\boldsymbol{\zeta} = \begin{bmatrix} 1 \\ 0 \end{bmatrix},$$

其中我们定义

$$\boldsymbol{A} = \sum_{i=1}^{N} \boldsymbol{C}^\top(\varphi_i)\boldsymbol{C}(\varphi_i).$$

由于$\boldsymbol{A}$是半正定的，所以该问题是一个凸二次规划问题. 通过分解$\boldsymbol{A} = \boldsymbol{F}^\top \boldsymbol{F}$，我们进一步可以把该问题写成形如

$$\min_{\boldsymbol{\zeta}} \left\| \begin{bmatrix} \boldsymbol{I} \\ \sqrt{\mu}\boldsymbol{F} \end{bmatrix} \boldsymbol{\zeta} \right\|_2^2, \quad \text{subject to } \boldsymbol{C}(0)\,\boldsymbol{\zeta} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

的最小二乘形式. 注意，这种惩罚的方法不需要知道阻带电平的一个**先验**(a priori)期望阈值δ. 我们仅希望，当μ足够大时，平方和中的所有项将小于期望的临界值δ. 因此，设计所达到的阻带衰减电平只能通过**后验**检查确定.

考虑一个数值例子，我们设定以下参数：天线数目$n = 16$，波长$\lambda = 8$，通带宽度$\phi = \frac{\pi}{6}$，天线的间距$\ell = 1$，离散的角度数目$N = 100$，权衡参数$\mu = 0.5$. 使用CVX求解所得到的解见图16.12.

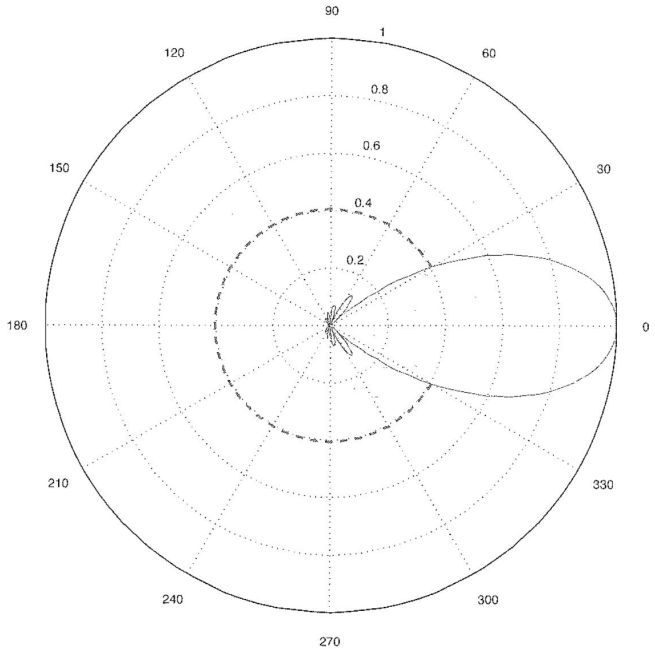

图 16.12: 由最小二乘的解得到的天线方向图

热噪声功率(平方根)是$\|\boldsymbol{z}\|_2 = 0.5671$，最大阻带电平是0.4050 (在图16.12中用虚线圆弧来突出显示). 此外，我们多次增加μ的值并重复求解该问题，绘制出了热噪声功率的平方根$\|\boldsymbol{z}\|_2$与阻带衰减电平之间的相应的权衡曲线，见图16.13.

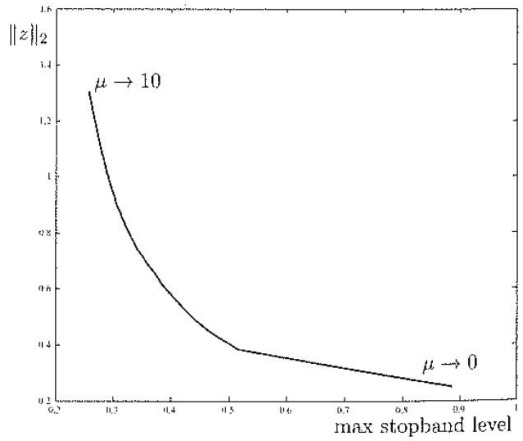

图 16.13: μ在区间$[0, 10]$上变动时对应的权衡曲线

§16.2.3 SOCP设计法

实际的天线设计问题，对离散角度上的阻带电平有明确的约束，可以直接表述为二次锥规划(second-order cone programming, SOCP). 一种可能的方法是在旁瓣电平约束下最小化热噪声功率，由此得到以含有$\boldsymbol{z}$的实部和虚部的向量$\boldsymbol{\zeta}$为变量的显式SOCP：

$$\begin{aligned} \underset{\boldsymbol{\zeta}\in\mathbb{R}^{2n},\gamma}{\text{minimize}} \quad & \gamma \\ \text{subject to} \quad & \boldsymbol{C}(0)\boldsymbol{\zeta} = \begin{bmatrix} 1 & 0 \end{bmatrix}^{\top}, \\ & \|\boldsymbol{C}(\varphi_i)\boldsymbol{\zeta}\|_2 \le \delta,\ i = 1, \ldots, N, \\ & \|\boldsymbol{\zeta}\|_2 \le \gamma. \end{aligned} \tag{16.2}$$

或者，可以通过给定热噪声功率的阈值γ来最小化旁瓣电平衰减δ. 更进一步，我们可以通过考虑利用如下SOCP来求出阻带衰减和噪声之间的最佳权衡：

$$\begin{aligned} \underset{\boldsymbol{\zeta}\in\mathbb{R}^{2n},\gamma,\delta}{\text{minimize}} \quad & \delta + w\gamma \\ \text{subject to} \quad & C(0)\boldsymbol{\zeta} = \begin{bmatrix} 1 & 0 \end{bmatrix}^{\top}, \\ & \|C(\varphi_i)\boldsymbol{\zeta}\|_2 \le \delta,\ i = 1, \ldots, N, \\ & \|\boldsymbol{\zeta}\|_2 \le \gamma. \end{aligned} \tag{16.3}$$

其中$w \ge 0$是给定的权衡参数.

作为数值示例，求解$\delta = 0.35$时的问题(16.2)，得到$\|z\|_2 = 0.435$，相应的方向图见图16.14.

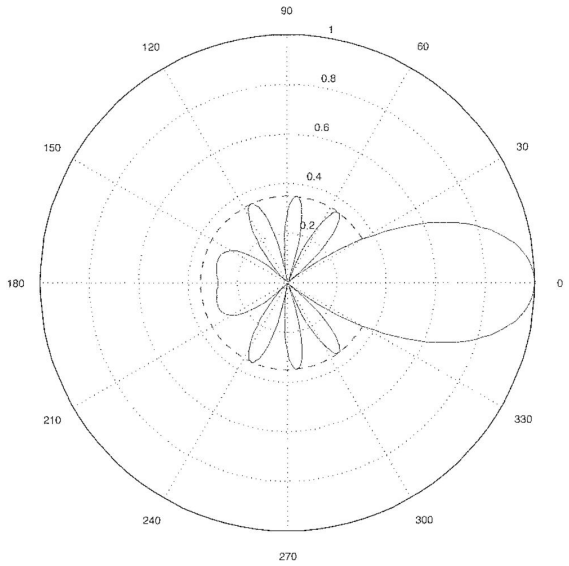

图 16.14: $\delta = 0.35$时求解问题(16.2)所的结果的天线方向图

同样，我们多次增加w的值并重复求解问题(16.3)，得到权衡曲线. 图16.15显示了$w \in [0.2, 10]$的权衡曲线.

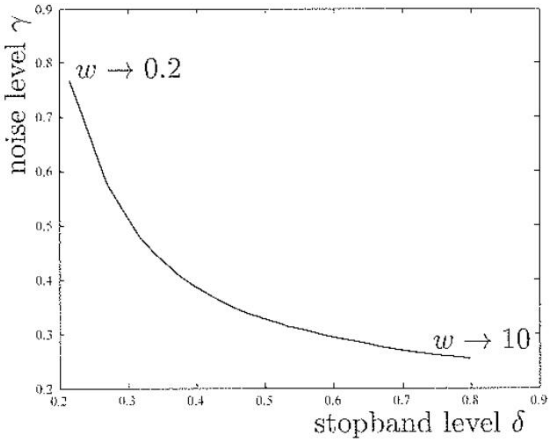

图 16.15: w在区间$[0.2, 10]$上变动时的权衡曲线

§16.3 数字电路设计

我们把数字电路设计问题考虑为一个**组合逻辑模块**. 在这样一个电

路中，基本构建模块是**门**：一个执行着一些简单布尔函数的电路，例如**逆变器**(inverter)，它表现为逻辑非，或者一些更为复杂的函数，如“与非(not-AND或NAND)”，它把两个布尔输入值A, B 转换为“A且B”整体的非. 数字电路设计的基本思想是设计门(确定门的大小)，使得电路运行快而且占据的面积小. 设计问题中还涉及其他因素，比如功率，但在这里我们不考虑这些. 设计变量是比例因子，它决定了每一个门的大小，同时也是门的基本电学参数. 这些参数连同电路的拓扑(这里固定)一同影响电路的速度. 这一节，我们将用几何规划(geometric programming, GP)模型来处理电路设计问题. 电路设计中GP在有一个很长的历史. [1]

§16.3.1 电路拓扑

组合电路由有主输入和输出的连通门组成. 我们假设在相应的电路拓扑图中没有环. 对于每一个门，我们可以定义**扇入**(fan-in)，即电路图中该门的前驱的集合，以及**扇出**(fan-out)，它是该门的后继的集合. 电路由n个“门”(由一些输入和一个输出的逻辑模块)组成，这些门连接着主输入(在图16.16中标记为$\{8, 9, 10\}$)和主输出(在图中标记为$\{11, 12\}$). 每一个门都用一个指定其类型的符号代表，例如，标记为$\{1, 3, 6\}$ 的门是逆变器. 对于这个电路，门4的扇入和扇出分别是

$$\mathrm{FI}(4) = \{1, 2\}, \quad \mathrm{FO}(4) = \{6, 7\}.$$

根据定义，主输入扇入为空，主输出扇出为空.

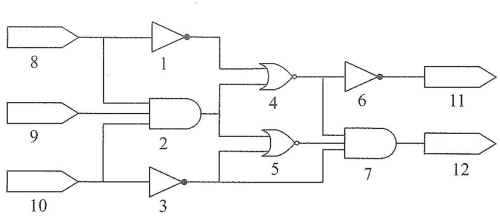

图 16.16: 一个数字电路的例子

§16.3.2 设计变量

我们模型中的设计变量是**比例因子**$x_i,\ i = 1, \ldots, n$，它们粗略的决定了每一个门的尺寸. 这些比例因子满足$x_i \geq 1,\ i = 1, \ldots, n$，其中$x_i = 1$表

[1]这里给出的介绍以及示例来自论文Digital circuit optimization via geometric programming(S. Boyd, S.-J. Kim, D. Patil, and M. Horowitz, *Operations Research*, 2005)，其中也包含了许多参考文献.

示对应门的尺寸最小，当一个比例因子$x_i = 16$时表示门中所有装置的总宽度是最小门的16倍. 比例因子决定了门的大小和各种电学特征，例如门的电阻和电导. 它们之间的关系可以被很好的近似如下.

- 门i的**面积**$A_i(\boldsymbol{x})$与比例因子x_i成正比，即$A_i(\boldsymbol{x}) = a_i x_i$对某个$a_i > 0$成立.
- 门i的**固有电容**形如
$$C_i^{\text{intr}}(\boldsymbol{x}) = C_i^{\text{intr}} x_i,$$
其中C_i^{intr}是正系数.
- 门i的**负载电容**是关于门i的扇出中门的比例因子的线性函数：
$$C_i(\boldsymbol{x}) = \sum_{j \in \text{FO(i)}} C_j x_j,$$
其中C_j是正系数.
- 每一个门都有**电阻**，它与比例因子成反比(门越大，通过它的电流越大)：
$$R_i(\boldsymbol{x}) = \frac{r_i}{x_i},$$
其中r_i是正系数.
- **门延迟**是衡量门执行它应该执行的逻辑操作的速度的量，这个延迟可以近似为
$$D_i(\boldsymbol{x}) = 0.7 R_i(\boldsymbol{x}) \left(C_i^{\text{intr}}(\boldsymbol{x}) + C_i(\boldsymbol{x}) \right).$$

我们可以看到，上述所有量都是关于(正的)设计向量$\boldsymbol{x}$的正多项式函数.

§16.3.3 设计目标

一个可能的设计目标是最小化电路的总延迟D. 总延迟可表示为

$$D = \max_{1 \le i \le n} T_i, \tag{16.4}$$

其中T_i表示门i直到成功输出一次所消耗的时间，并假设主输入信号在$t = 0$时传递. 即，T_i是在所有始于主输入并终于门i的所有路上延迟的最大值. 我们可以递归地表示T_i为

$$T_i = \max_{j \in \text{FI(i)}} \left(T_j + D_i \right). \tag{16.5}$$

D的计算过程仅包括加法和逐点取最大. 由于每一个D_i是$\boldsymbol{x}$的正多项式，我们可以将总延迟表示为一个$\boldsymbol{x}$的广义正多项式. 对于图16.16中的电路，总延迟D可表示为

$$\begin{aligned} T_i &= D_i, \quad i = 1, 2, 3; \\ T_4 &= \max(T_1, T_2) + D_4; \\ T_5 &= \max(T_2, T_3) + D_5; \\ T_6 &= T_4 + D_6; \\ T_7 &= \max(T_3, T_4, T_5) + D_7; \\ D &= \max(T_6, T_7). \end{aligned}$$

§16.3.4 一个电路设计问题

我们现在考虑问题：在面积约束条件下，选择比例因子x_i使得总延迟最小，即

$$\underset{\boldsymbol{x}}{\text{minimize}}\, D(\boldsymbol{x}) \quad \text{subject to } A_i(\boldsymbol{x}) \le A_{\max},\ x_i \ge 1, \quad i = 1, \ldots, n,$$

其中$A_{\max}$是每个门的面积上限. 因为D是$\boldsymbol{x}$的广义正多项式，所以上述问题可以被看做是一个GP问题. 为了找到一个紧凑且明确的GP表示，我们可以使用在总延迟的定义中出现的中间变量T_i. 通过观察可知，由不等式关系

$$T_i \ge \max_{j \in \mathrm{FI}(i)} (T_j + D_i) \tag{16.6}$$

代替等式关系(16.5)不改变优化问题的最优解. 其原因为：由于式(16.4)中的目标D相对于T_i是非递减的，并且T_i随着$T_j,\ j \in \mathrm{FI}(i)$的增加而增加，所以最优解将选择所有尽可能小的$T_i$，因此(16.6)是以等式成立的. 更进一步，上述不等式等价于

$$T_i \ge T_j + D_i, \quad \forall j \in \mathrm{FI}(i).$$

之后，我们得到设计问题的如下显式GP表述：

$$\begin{array}{lll} \underset{\boldsymbol{x}, T_i > 0, D}{\text{minimize}} & D & \\ \text{subject to} & A_i(x) \le A_{\max},\ x_i \ge 1, & i = 1, \ldots, n; \\ & D \ge T_i, & i = 1, \ldots, n; \\ & T_i \ge T_j + D_i(x),\ \forall j \in \mathrm{FI}(i), & i = 1, \ldots, n. \end{array}$$

我们考虑图16.16中的电路给出的数值实例，其中$A_{\max}=16$，其他数据见表16.1.

门	C_i	C_i^{intr}	r_i	a_i
1,3,6	3	3	0.48	3
2,7	4	6	0.48	8
4,5	5	6	0.48	10

表 16.1: 图16.16中电路的数据

利用CVX求得最优比例因子

$$\boldsymbol{x}^*=\begin{bmatrix}4.6375\\2.0000\\4.8084\\1.6000\\1.0000\\1.0000\\1.0000\end{bmatrix},$$

对应的最小延迟$D^*=7.686$.

§16.4 航天器设计

近年来，我们已经注意到，一些关于航空器结构与运行方面的问题能够表述为几何规划的形式，从而，通过凸优化被高效的求解. 我们在这里展示一个简单的机翼设计问题的例子. [2]

我们的设计变量是总面积S，机翼展弦比$A=b^2/S$，其中b为翼展，和巡航速度V(m/s)，目标是最小化阻力

$$D=\frac{1}{2}\rho V^2 C_D S, \tag{16.7}$$

其中ρ是空气密度，C_D是阻力系数，见图16.17.

[2]我们的陈述来自论文Geometric programming for aircraft design optimization, by W. Hoburg and P. Abbeel, in proc. *Structures, Structural Dynamics and Materials Conference*, 2012，读者可以阅读该文以获得更多细节.

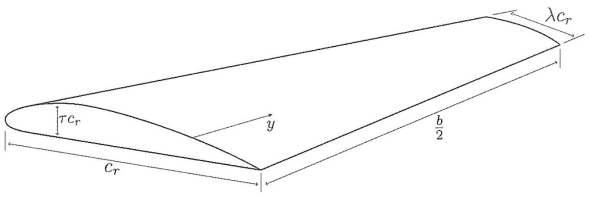

图 16.17: 单锥形机翼几何：c_r是根弦，$b/2$是半翼展，λ是弦锥因子，τ是翼型厚弦比.

当飞机稳定飞行时，它必须满足两个基本的平衡条件，正像图16.18所展示的：升力L必须抵消飞机的重力W，并且推力T 与阻力D必需是平衡的，即

$$L = W, \quad T = D.$$

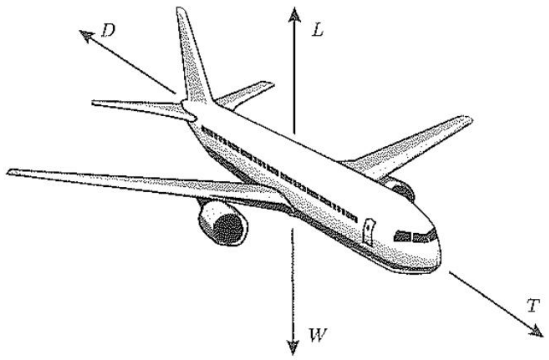

图 16.18: 在稳态飞行下，升力L等于重力W，推力T等于阻力D

用C_L表示升力系数，升力可以表示为

$$L = \frac{1}{2}\rho V^2 C_L S.$$

式(16.7)中的阻力系数可建模为三项的和：机身寄生阻力，机翼寄生阻力，诱导阻力，即

$$C_D = \frac{\mathrm{CDA}_0}{S} + \kappa C_f \frac{S_{\mathrm{wet}}}{S} + \frac{C_L^2}{\pi A e}, \tag{16.8}$$

其中CDA_0是机身阻力面积，κ是导致压差阻力的形式因子，S_{wet}是受潮面积(即机翼曲面的实际面积，而S是二维投影面积或是机翼阴影面积)，C_L是

升力系数，e是Oswald效率因子. 对于一个完全湍流边界层，式(16.8)中的表面摩擦系数C_f 可以粗略地近似为

$$C_f = \frac{0.074}{\mathrm{Re}^{0.2}},$$

其中

$$\mathrm{Re} = \frac{\rho V}{\mu}\sqrt{S/A}$$

是平均翼弦[3]$c = \sqrt{S/A}$处的雷诺数，μ是空气粘度. 航空器的总重量W被建模为一个固定重量W_0与机翼重量W_w之和，后者被建模为

$$W_w = \kappa_s S + \kappa_l \frac{N_{\mathrm{ult}} b^3 \sqrt{W_0 W}}{S\tau},$$

这里κ_s,κ_l是合适的常数，N_{ult}是结构尺寸决定的极限载荷因子，τ是翼型厚弦比(注意到W_w是W 自己的函数). 在稳态飞行条件下，约束将重力方程耦合到阻力方程中，升力必须要等于重力，即$L = W$. 因此，必须满足

$$\begin{aligned} \frac{1}{2}\rho V^2 C_L S &= W, \\ W_0 + W_w &= W. \end{aligned}$$

最后，飞机在落地时，必须没有拖延地能以最小速度$V_{\min}$飞行. 这个要求由下面的限制得以满足：

$$\frac{1}{2}\rho V_{\min}^2 C_L^{\max} S \geq W,$$

这里，$C_L^{\max}$是着陆(放下副翼)时的最大升力系数.

我们必须选择S, A和V的值，使得阻力最小化且满足上述**所有**关系. 表16.2给定了常量参数. 这个问题看起来是一个很难的优化问题，涉及非线性耦合变量. 然而，这个问题中的一个关键点是：我们能够将正项多项式**等式**约束松弛成正项多项式**不等式**约束，且保持问题的最优目标值不变. 比如，如果C_D没有出现在其他任何单项等式限制中，且目标与不等式限制

$$C_D \geq \frac{\mathrm{CDA}_0}{S} + kC_f\frac{S_{\mathrm{wet}}}{S} + \frac{C_L^2}{\pi A e} \tag{16.9}$$

关于C_D都是单调递增的(或者是常数)，则我们可用不等式关系(16.9)来等价替换等式(16.8). 在这些条件下，如果等式关系(16.8)在最优解处不成立，显然，我们能减小C_D直到等式满足，而不会增大目标值或是将解移

[3]平均翼弦是区间$[\lambda c_r, c_r]$中且满足$S = bc$的值c.

表 16.2: 航天器设计实例中的常数

量	值	单位	描述
CDA_0	0.0306	m^2	机身阻力面积
ρ	1.23	$\mathrm{Kg/m}^3$	空气密度
μ	1.78×10^{-5}	Kg/ms	空气粘度
S_{wet}/S	2.05		受潮面积比
κ	1.2		形式因子
e	0.96		Oswald效率因子
W_0	4940	N	除机翼外的机身重量
N_{ult}	2.5		极限载荷因子
τ	0.12		翼型厚弦比
$V_{\min}$	22	m/s	着陆速度
$C_L^{\max}$	2.0		(放下副翼)时C_L的最大值
κ_s	45.42	$\mathrm{N/m}^2$	
κ_l	8.71×10^{-5}	m^{-1}	

出可行集. 因为在目前的背景下这些条件都是满足的，所以我们能够将设计问题写成如下显示的GP形式：

$$\begin{array}{ll} \underset{S,A,V,C_D,C_L,C_f,\mathrm{Re},W,W_w}{\text{minimize}} & \frac{1}{2}\rho V^2 C_D S \\ \text{subject to} & \frac{0.074}{C_f \mathrm{Re}^{0.2}} = 1,\ \frac{2W}{\rho V^2 C_L S} = 1, \\ & \frac{2W}{\rho V_{\min}^2 C_L^{\max} S} \le 1,\ \frac{\rho V}{\mu \mathrm{Re}}\sqrt{S/A} = 1, \\ & \frac{\mathrm{CDA}_0}{C_D S} + \kappa \frac{C_f}{C_D}\frac{S_{\mathrm{wet}}}{S} + \frac{C_L^2}{C_D \pi A e} \le 1, \\ & \kappa_s \frac{S}{W_w} + \kappa_l \frac{N_{\mathrm{ult}} A^{3/2}\sqrt{W_0 W S}}{W_w \tau} \le 1, \\ & \frac{W_0}{W} + \frac{W_w}{W} \le 1. \end{array} \tag{16.10}$$

常量参数见表16.2. 求解GP(16.10)得到的最优设计结果见表16.3. 值得指出的是，这里的S, A, V是独立设计变量，其余的变量都是由这三个独立变量和常量参数决定的.

GP模型使我们很容易地得到设计参数的全局最优值. 但是，在一个现实的设计背景下，设计者需要考虑在相互冲突的目标间考虑一系列可能的折衷解(比如增加着陆速度和(或)巡航速度). 对一定范围内的一系列$V_{\min}$和V的值，求解GPGP(16.10)很易于得到这些数值折衷解. 并且，最

表 16.3: 航天器实例中的最有设计

参数	结果	单位	描述
A	12.7		机翼展弦比
S	12.08	m^2	机翼表面积
V	38.55	m/s	巡航速度
C_D	0.0231		阻力系数
C_L	0.6513		升力系数
C_f	0.0039		表面摩擦系数
Re	2.5978×10^6		雷诺数
W	7189.1	N	总重量
W_w	2249.1	N	机翼重量

优化模型中也包含了几个其他方面. 例如由模型可以计算出燃料的质量，即通过

$$W = (W_0 + W_w)(1 + \theta_{\text{fuel}}),$$

其中θ_{fuel}是燃料的质量比. 进一步，Brequet的航程方程

$$R = \frac{h_{\text{fuel}}}{g}\eta_0 \frac{L}{D}\log(1 + \theta_{\text{fuel}}) \tag{16.11}$$

描述了燃料质量比与航空器的最大航程之间的关系，这里假定升阻比L/D是常数，η_0是全体燃料能量对推力能量的效率因子. 等式$P_{\text{fuel}} = \dot{m}_{\text{fuel}} h_{\text{fuel}}$描述了$h_{\text{fuel}}$与燃料质量流速率与燃料能量的关系. 在稳态飞行条件下，必须满足

$$TV \leq \eta_0 P_{\text{fuel}}.$$

虽然Brequet的航程方程(16.11)并不是直接表述成正的多项式，但可以用泰勒级数展开来很好地逼近它. 由(16.11)解得

$$1 + \theta_{\text{fuel}} = \exp\left(\frac{gRD}{h_{\text{fuel}}\eta_0 L}\right),$$

且指数函数的级数展式具有正项式结构. 因此，利用

$$\begin{aligned} z &= \frac{gRD}{h_{\text{fuel}}\eta_0 L}, \\ \theta_{\text{fuel}} &\geq z + \frac{z^2}{2!} + \frac{z^3}{3!} + \cdots \end{aligned}$$

可将Brequet的航程方程近似地引入到GP模型中.

§16.5 供应链管理

本节讨论一个生产工程中出现的问题，即在不确定的需求下如何控制库存量. [4]这个问题由T个时段的订购、库存和存储决策组成，目标是在满足需求的前提下最小化成本. 总成本由实际采购成本、持有与缺货带来的成本组成. 考虑一个单一商品，此商品在时段k的库存量记作$x(k)$. 那么，商品库存量随着(离散)时刻的基本演化可以写作：

$$x(k+1) = x(k) + u(k) - w(k), \quad k = 0, 1, \ldots, T-1,$$

这里$x(0) = x_0$是一个给定的初始库存量，$u(k)$是时段k的采购量，$w(k)$是时段k到$k+1$的需求量. 我们假设一个单位的仓储价格是h，一个单位的短缺成本是p，购买单位商品的价格是c. 那么时段k的成本由

$$cu(k) + \max(hx(k+1), -px(k+1))$$

确定. 进一步假设任一次订货规模的上界是M，我们可以将T个时段的库存控制问题写作：

$$\begin{array}{ll} \underset{u(0),\ldots,u(T-1)}{\text{minimize}} & \sum_{k=0}^{T-1} cu(k) + \max(hx(k+1), -px(k+1)) \\ \text{subject to} & 0 \le u(k) \le M,\ k = 0, \ldots, T-1, \end{array}$$

这里

$$x(k) = x_0 + \sum_{i=0}^{k-1} (u(i) - w(i)), \quad k = 1, ..., T.$$

引入松弛变量$y(0), ..., y(T-1)$，此问题可表述成如下的线性规划：

$$\begin{array}{ll} \underset{u(0),\ldots,u(T-1),y(0),\ldots,y(T-1)}{\text{minimize}} & \sum_{k=0}^{T-1} y(k) \\ \text{subject to} & cu(k) + hx(k+1) \le y(k),\ k = 0, \ldots, T-1, \\ & cu(k) - px(k+1) \le y(k),\ k = 0, \ldots, T-1, \\ & x(k) = x_0 + \sum_{i=0}^{k-1} (u(i) - w(i)), \quad k = 1, ..., T, \\ & 0 \le u(k) \le M,\ k = 0, \ldots, T-1. \end{array}$$

接着，我们定义向量$\boldsymbol{u} = (u(0), \ldots, u(T-1)), \boldsymbol{y} = (y(0), \ldots, y(T-1)), \boldsymbol{w} =$

[4]这里展示的处理方法所采用的设置与记号源于Bertsimas 和Thiele的论文：A robust optimization approach to supply chain management, *Operation Research*, 2006.

$(w(0),\dots,w(T-1)), \boldsymbol{x}=(x(0),\dots,x(T))$. 可用紧凑的记号将问题重写为：

$$\begin{array}{ll}\underset{\boldsymbol{u},\boldsymbol{y}}{\text{minimize}} & \mathbf{1}^\top \boldsymbol{y} \\ \text{subject to} & c\boldsymbol{u}+h\boldsymbol{x}\le \boldsymbol{y},\ c\boldsymbol{u}-p\boldsymbol{x}\le \boldsymbol{y}, \\ & \boldsymbol{x}=x_0\mathbf{1}+\boldsymbol{U}\boldsymbol{u}-\boldsymbol{U}\boldsymbol{w}, \\ & \mathbf{0}\le \boldsymbol{u}\le M\mathbf{1},\end{array} \tag{16.12}$$

其中

$$\boldsymbol{U}\doteq\begin{bmatrix}1&0&0&\cdots&0\\1&1&0&\cdots&0\\\vdots&\vdots&\ddots&\ddots&\vdots\\1&1&1&\cdots&1\end{bmatrix}.$$

在这种设置下，我们假设在所有时段$k=0,\dots,T-1$，需求$w(k)$是确定且已知的，并且时段$k=0$的所有决策都是已知的. 在下一节中，我们将讨论如何放松这两个假设.

§16.5.1 针对区间不确定需求的鲁棒法

这里我们假设可能能精准地知道需求$\boldsymbol{w}$. 特别地，考虑需求在名义上的预期需求$\hat{\boldsymbol{w}}\ge\mathbf{0}$ 周围有$\rho\%$的波动的情况，即

$$\boldsymbol{w}\in W\doteq\{\boldsymbol{w}:\boldsymbol{w}_{\text{lb}}\le\boldsymbol{w}\le\boldsymbol{w}_{\text{ub}}\}, \tag{16.13}$$

其中

$$\boldsymbol{w}_{\text{lb}}\doteq(1-\rho/100)\hat{\boldsymbol{w}},\quad \boldsymbol{w}_{\text{ub}}\doteq(1+\rho/100)\hat{\boldsymbol{w}}.$$

然后，我们寻求一个订货序列$\boldsymbol{u}$，使得在所有可能的需求下的最坏情况的成本最小，即

$$\begin{array}{ll}\underset{u,y}{\text{minimize}} & \mathbf{1}^\top \boldsymbol{y} \\ \text{subject to} & c\boldsymbol{u}+h\boldsymbol{x}\le \boldsymbol{y},\ \forall \boldsymbol{w}\in W, \\ & c\boldsymbol{u}-p\boldsymbol{x}\le \boldsymbol{y},\ \forall \boldsymbol{w}\in W, \\ & \boldsymbol{x}=x_0\mathbf{1}+\boldsymbol{U}\boldsymbol{u}-\boldsymbol{U}\boldsymbol{w},\ \forall \boldsymbol{w}\in W, \\ & \mathbf{0}\le \boldsymbol{u}\le M\mathbf{1}.\end{array} \tag{16.14}$$

因为$\boldsymbol{x}=x_0\mathbf{1}+\boldsymbol{U}\boldsymbol{u}-\boldsymbol{U}\boldsymbol{w}$，其中$\boldsymbol{U}$的元素都是非负的，并且$\boldsymbol{w}$中的每个元素是属于一个已知区间的，约束$c\boldsymbol{u}+h\boldsymbol{x}\le\boldsymbol{y},\ \forall\boldsymbol{w}\in W$ 成立当且仅当

$$c\boldsymbol{u}+h(x_0\mathbf{1}+\boldsymbol{U}\boldsymbol{u}-\boldsymbol{U}\boldsymbol{w}_{\text{lb}})\le\boldsymbol{y},$$

约束$c\boldsymbol{u}-p\boldsymbol{x}\leq\boldsymbol{y},\forall\boldsymbol{w}\in W$成立当且仅当

$$c\boldsymbol{u}-p(x_0\mathbf{1}+\boldsymbol{U}\boldsymbol{u}-\boldsymbol{U}\boldsymbol{w}_{\mathrm{ub}})\leq\boldsymbol{y}.$$

因此可以通过求解LP

$$\begin{array}{ll}\underset{\boldsymbol{u},\boldsymbol{y}}{\text{minimize}} & \mathbf{1}^{\top}\boldsymbol{y}\\ \text{subject to} & c\boldsymbol{u}+h(x_0\mathbf{1}+\boldsymbol{U}\boldsymbol{u}-\boldsymbol{U}\boldsymbol{w}_{\mathrm{lb}})\leq\boldsymbol{y},\\ & c\boldsymbol{u}-p(x_0\mathbf{1}+\boldsymbol{U}\boldsymbol{u}-\boldsymbol{U}\boldsymbol{w}_{\mathrm{ub}})\leq\boldsymbol{y},\\ & \mathbf{0}\leq\boldsymbol{u}\leq M\mathbf{1}\end{array}\tag{16.15}$$

得到最坏情况下的最优决策.

§16.5.2 区间不确定性下的仿射订购策略

在以上的表述中，是在时段$k=0$且以一种“开环”的方式来计算所有向前的订货决策$u(0),\ldots,u(T-1)$. 但在实际中，人们只会执行第一次决策$u(0)$(（所谓的“此时此地”决策)，然后等待并观察接下来发生的事，直到下一个决定时段. 在时段$k=1$，可以观察到不确定需求的实际实现. 因此，当确定决策$u(1)$时可以获取这个信息. 显然，这些信息会使新的决策$u(1)$受益，因此原先的忽略这些信息的方法是次优的.

一般地，在时段k将执行的决策$u(k)$ 将受益于从时段0 到$k-1$之间**观察**到的需求信息(这些需求在时间k时已经是确定的了)，也就是说$u(k)$是过去需求的泛函，即$u(k)=\varphi_k(w(0),\ldots,w(k-1))$，在这样的反映式或者“闭环”的方法下，这个优化问题相当于是寻找最优函数φ_k(通常称为**政策**)，使得最坏情况成本最小. 但是寻找所有泛函中搜索使得问题变得极为困难(需要在一个由所有可能函数φ_k组成的“无限维”集合中进行搜寻). 因此，一种常用方法是固定φ_k中的参数，然后优化有限维的参数集合(参考14.3.3节描述了一个金融领域中的多期优化问题的类似方法). 举例来说，有一种有效的方法是，考虑**仿射式**参数化，也就是考虑$k=1,\ldots,T-1$时决策φ_k形如

$$u(k)=\varphi_k(w(0),\ldots,w(k-1))=\bar{u}(k)+\sum_{i=0}^{k-1}\alpha_{k,i}(w(i)-\hat{w}(i)),$$

这里$\bar{u}(k)$是“名义”决策，$u(0)=\bar{u}(0)$，对$k\geq1,\alpha_{k,i}$是我们修正名义决策使其和需求$w(i)$与名义需求$\hat{w}(i)$之间偏差成正比例的系数. 使用矩阵记号，

我们写成

$$\boldsymbol{u} = \bar{\boldsymbol{u}} + \boldsymbol{A}(\boldsymbol{w} - \hat{\boldsymbol{w}}), \quad \boldsymbol{A} \doteq \begin{bmatrix} 0 & 0 & \cdots & 0 \\ \alpha_{1,0} & 0 & \cdots & 0 \\ \alpha_{2,0} & \alpha_{2,1} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{T-1,0} & \cdots & \alpha_{T-1,T-2} & 0 \end{bmatrix},$$

其中$\bar{\boldsymbol{u}} = (\bar{u}(0), \ldots, \bar{u}(T-1))$. 特别的，在区间不确定模型(16.13)中，需求波动$\tilde{\boldsymbol{w}} \doteq \boldsymbol{w} - \hat{\boldsymbol{w}}$属于对称向量区间

$$-\bar{\boldsymbol{w}} \le \tilde{\boldsymbol{w}} \le \bar{\boldsymbol{w}}; \quad \bar{\boldsymbol{w}} \doteq \frac{\rho}{100}\hat{\boldsymbol{w}} \ge \boldsymbol{0}.$$

将这里的$\boldsymbol{u}$的表达式代入问题(16.14)，我们得到

$$\begin{array}{ll} \underset{\bar{\boldsymbol{u}},\boldsymbol{y},\boldsymbol{A}}{\text{minimize}} & \boldsymbol{1}^\top \boldsymbol{y} \\ \text{subject to} & (c\bar{\boldsymbol{u}} + c\boldsymbol{A}\tilde{\boldsymbol{w}}) + h(x_0\boldsymbol{1} + \boldsymbol{U}(\bar{\boldsymbol{u}} + \boldsymbol{A}\tilde{\boldsymbol{w}}) - \boldsymbol{U}\boldsymbol{w}) \le \boldsymbol{y}, \ \forall \boldsymbol{w} \in W \\ & (c\bar{\boldsymbol{u}} + c\boldsymbol{A}\tilde{\boldsymbol{w}}) - p(x_0\boldsymbol{1} + \boldsymbol{U}(\bar{\boldsymbol{u}} + \boldsymbol{A}\tilde{\boldsymbol{w}}) - \boldsymbol{U}\boldsymbol{w}) \le \boldsymbol{y}, \ \forall \boldsymbol{w} \in W \\ & \boldsymbol{0} \le \boldsymbol{u} + \boldsymbol{A}\tilde{\boldsymbol{w}} \le M\boldsymbol{1}, \forall \boldsymbol{w} \in W. \end{array} \tag{16.16}$$

如果$\boldsymbol{v}$是一个向量并且$\bar{\boldsymbol{w}} \ge \boldsymbol{0}$，易于验证

$$\begin{aligned} \underset{-\bar{\boldsymbol{w}} \le \tilde{\boldsymbol{w}} \le \bar{\boldsymbol{w}}}{\text{maximize}} \boldsymbol{v}^\top \tilde{\boldsymbol{w}} &= |\boldsymbol{v}|^\top \bar{\boldsymbol{w}}, \\ \underset{-\bar{\boldsymbol{w}} \le \tilde{\boldsymbol{w}} \le \bar{\boldsymbol{w}}}{\text{minimize}} \boldsymbol{v}^\top \tilde{\boldsymbol{w}} &= -|\boldsymbol{v}|^\top \bar{\boldsymbol{w}}, \end{aligned}$$

这里$|\boldsymbol{v}|$是由$\boldsymbol{v}$分量的绝对值组成的向量. 针对问题(16.16)的约束逐行应用这个规则，即可得到这个鲁棒问题等价于

$$\begin{array}{ll} \underset{\bar{\boldsymbol{u}},\boldsymbol{y},\boldsymbol{A}}{\text{minimize}} & \boldsymbol{1}^\top \boldsymbol{y} \\ \text{subject to} & c\bar{\boldsymbol{u}} + h\boldsymbol{U}\bar{\boldsymbol{u}} + hx_0\boldsymbol{1} - h\boldsymbol{U}\hat{\boldsymbol{w}} + |c\boldsymbol{A} + h\boldsymbol{U}\boldsymbol{A} - h\boldsymbol{U}|\bar{\boldsymbol{w}} \le \boldsymbol{y}, \\ & c\bar{\boldsymbol{u}} - pU\bar{\boldsymbol{u}} - px_0\boldsymbol{1} + p\boldsymbol{U}\hat{\boldsymbol{w}} + |c\boldsymbol{A} - p\boldsymbol{U}\boldsymbol{A} + p\boldsymbol{U}|\bar{\boldsymbol{w}} \le \boldsymbol{y}, \\ & \bar{\boldsymbol{u}} + |\boldsymbol{A}|\bar{\boldsymbol{w}} \le M\boldsymbol{1}, \\ & \bar{\boldsymbol{u}} - |\boldsymbol{A}|\bar{\boldsymbol{w}} \ge \boldsymbol{0}. \end{array}$$

引入三个由松弛变量组成的下三角矩阵$\boldsymbol{Z}_1, \boldsymbol{Z}_2, \boldsymbol{Z}_3$，我们将此问题重新表

述成LP

$$\begin{array}{ll}\underset{\bar{\boldsymbol{u}},\boldsymbol{y},\boldsymbol{A},\boldsymbol{Z}_1,\boldsymbol{Z}_2,\boldsymbol{Z}_3}{\text{minimize}} & \mathbf{1}^\top \boldsymbol{y} \\ \text{subject to} & c\bar{\boldsymbol{u}} + h\boldsymbol{U}\bar{u} + hx_0\mathbf{1} - h\boldsymbol{U}\hat{\boldsymbol{w}} + \boldsymbol{Z}_1\bar{\boldsymbol{w}} \le \boldsymbol{y}, \\ & c\bar{\boldsymbol{u}} - p\boldsymbol{U}\bar{u} - px_0\mathbf{1} + p\boldsymbol{U}\hat{\boldsymbol{w}} + \boldsymbol{Z}_2\bar{\boldsymbol{w}} \le \boldsymbol{y}, \\ & \bar{\boldsymbol{u}} + \boldsymbol{Z}_3\bar{\boldsymbol{w}} \le M\mathbf{1}, \\ & \bar{\boldsymbol{u}} - \boldsymbol{Z}_3\bar{\boldsymbol{w}} \ge 0, \\ & |c\boldsymbol{A} + h\boldsymbol{U}\boldsymbol{A} - h\boldsymbol{U}| \le \boldsymbol{Z}_1, \\ & |c\boldsymbol{A} - p\boldsymbol{U}\boldsymbol{A} + p\boldsymbol{U}| \le \boldsymbol{Z}_2, \\ & |\boldsymbol{A}| \le \boldsymbol{Z}_3. \end{array} \tag{16.17}$$

一旦求解了这个问题，即可获得订货策略和不确定的库存量

$$\begin{aligned} \boldsymbol{u} &= \bar{\boldsymbol{u}} + \boldsymbol{A}\tilde{\boldsymbol{w}}, \\ \boldsymbol{x} &= x_0\mathbf{1} + \boldsymbol{U}(\bar{\boldsymbol{u}} - \hat{\boldsymbol{w}}) + (\boldsymbol{U}\boldsymbol{A} - \boldsymbol{U})\tilde{\boldsymbol{w}}. \end{aligned} \tag{16.18}$$

然后，我们能获得订购和库存量的上下极限，如下

$$\begin{aligned} \boldsymbol{u}_{\text{lb}} &= \bar{\boldsymbol{u}} - |\boldsymbol{A}|\bar{\boldsymbol{w}}, \\ \boldsymbol{u}_{\text{ub}} &= \bar{\boldsymbol{u}} + |\boldsymbol{A}|\bar{\boldsymbol{w}}, \\ \boldsymbol{x}_{\text{lb}} &= x_0\mathbf{1} + \boldsymbol{U}(\bar{\boldsymbol{u}} - \hat{\boldsymbol{w}}) - |\boldsymbol{U}\boldsymbol{A} - \boldsymbol{U}|\bar{\boldsymbol{w}}, \\ \boldsymbol{x}_{\text{ub}} &= x_0\mathbf{1} + \boldsymbol{U}(\bar{\boldsymbol{u}} - \hat{\boldsymbol{w}}) + |\boldsymbol{U}\boldsymbol{A} - \boldsymbol{U}|\bar{\boldsymbol{w}}. \end{aligned}$$

实践中，将根据策略(16.18)在线计算时段k时的实际订购量$u(k)$. 随着时间的演进，将观察到$\tilde{w}(i),\ i = 0, \dots, k-1,$ 的实际值. 我们已知的先验信息是：当$k = 0, \dots, T-1$时，值$u(k)$将包含于区间$[u_{\text{lb}}(k), u_{\text{ub}}(k)]$.

下面用一个例子来说明这些想法. 考虑有$T = 12$(比如月)个时段的决策，相应的数据为：购买成本$c = 5$，存储成本$h = 4$，短缺成本$p = 6$，订购上界$M = 110$. 进一步，假设名义需求服从正弦曲线

$$\hat{w}(k) \doteq 100 + 20\sin\left(2\pi\frac{k}{T-1}\right), \quad k = 0, \dots, T-1, \tag{16.19}$$

并且初始库存量$x_0 = 100$. 在没有不确定性的情况下(需求量$\boldsymbol{w}$等于名义需求量$\hat{\boldsymbol{w}}$)，求解问题(16.12)得到的最优成本是5,721.54，并且订购与存储量的图形见图16.19.

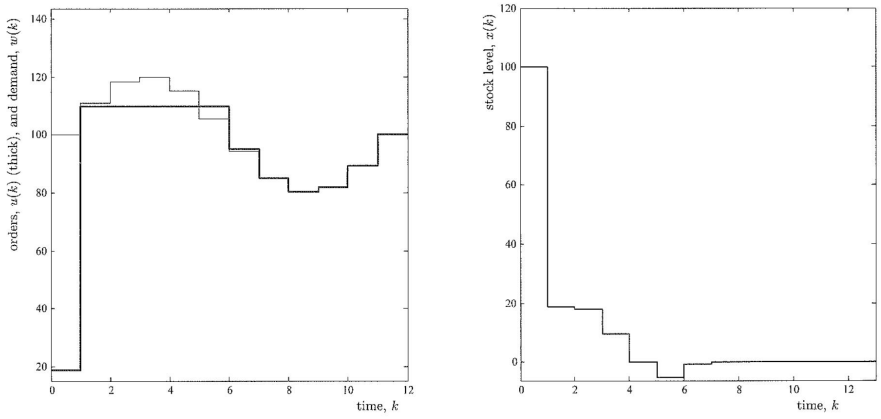

图 16.19: 名义需求下的最优订购策略. 左图：需求和订购曲线. 右图：库存量曲线. 最优的名义成本为$5,721.54$.

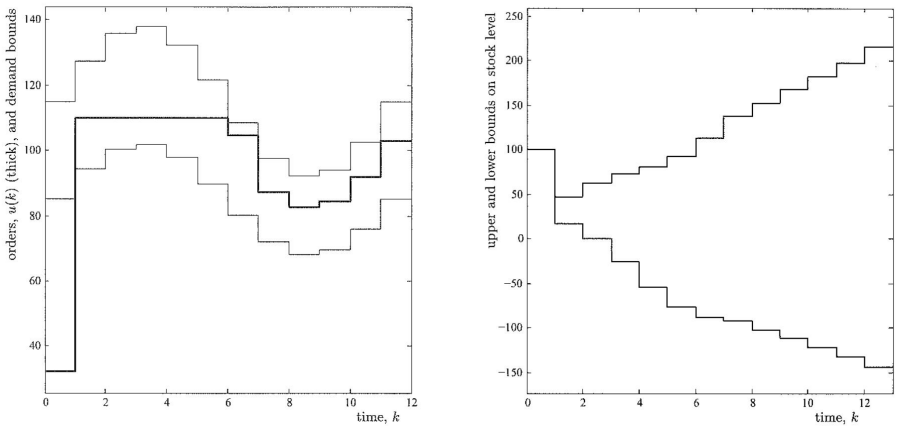

图 16.20: 在$\rho = 15\%$不确定性下的最优“开环”订购. 左图：需求和订购曲线，右图：库存量曲线. 最坏情况成本为$11,932.50$.

接下来考虑需求量存在$\rho = 15\%$的不确定性的情况. 我们求解鲁棒的“开环”问题(16.15)，得到的最坏情况的成本是11,932.50. 订购与库存量曲线如图16.20所示.

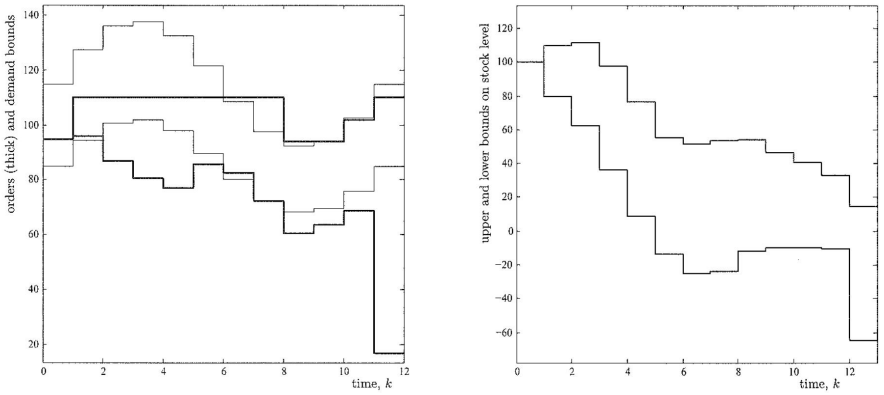

图 16.21: 不确定性为$\rho = 15\%$时的最优“闭环”订购. 左图，需求和订购曲线：右图：库存量曲线. 最坏情况的成本为$8,328.64$

在需求量同样有15%的不确定性的情况下，我们求解鲁棒的“闭环”问题(16.17)，得到最坏情况下的成本是8,328.64. 订购与库存量的曲线如图16.21所示. 进一步，由(16.18)得到的订购策略的最优参数为

$$\bar{\boldsymbol{u}}^T = \begin{bmatrix} 94.69 & 103.0 & 98.36 & 95.26 & 93.49 & 97.83 & 96.26 & 91.25 & 77.33 & 78.68 & 85.2 & 63.4 \end{bmatrix},$$

$$\boldsymbol{A} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.4697 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.2506 & 0.4743 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.1332 & 0.2518 & 0.4828 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.07069 & 0.1324 & 0.254 & 0.4868 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.007817 & 0.01524 & 0.04958 & 0.1856 & 0.4395 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.009959 & 0.0148 & 0.03499 & 0.0915 & 0.2258 & 0.4529 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.02248 & 0.05136 & 0.08535 & 0.114 & 0.1627 & 0.2676 & 0.4913 & 0 & 0 & 0 & 0 & 0 \\ 0.01838 & 0.03096 & 0.04785 & 0.06251 & 0.0879 & 0.1427 & 0.2594 & 0.5104 & 0 & 0 & 0 & 0 \\ 0.009341 & 0.01578 & 0.02445 & 0.03202 & 0.04514 & 0.07356 & 0.1343 & 0.2649 & 0.5463 & 0 & 0 & 0 \\ 0.004745 & 0.008067 & 0.01258 & 0.01654 & 0.02343 & 0.03848 & 0.07088 & 0.1412 & 0.2945 & 0.6839 & 0 & 0 \\ 0.01105 & 0.02035 & 0.03276 & 0.04322 & 0.06086 & 0.09775 & 0.1717 & 0.3123 & 0.547 & 0.8799 & 1.372 & 0 \end{bmatrix}.$$

通过使用“反应式的”(或“闭环”)决策取代静态的“开环”法，就“最坏情况下的成本”而言，获得的改进超过30%.

§16.5.3 用于一般的随机不确定性的场景法

当需求量$w(k)$的不确定性是随机的，且不能建模为前面的区间模型时，我们可以使用一种虽然是近似但却简单而有效的方法，即基于不确定性的场景抽样. 该法中，我们假设需求量$\boldsymbol{w}$(包含需求$w(0), \ldots, w(T-1)$)是一个随机向量，它的期望$\hat{\boldsymbol{w}} = \mathbb{E}\{\boldsymbol{w}\}$是已知的，它的概率分布也是已知的，或者至少得到了它的N个独立同分布的样本$\boldsymbol{w}^{(1)}, \ldots, \boldsymbol{w}^{(N)}$. 在这种设

定下，$w(k)$不必有界，且基于所设定的分布，有可能是时间相关的.

这种基于场景的方法[5]的简单思想假定：如果N 足够大，则N个生成的场境的全体将会合理地呈现出不确定性. 这样，这里的场境－鲁棒问题相当于求解形如(16.16) 的优化问题，但是要用$\forall \boldsymbol{w} \in \{\boldsymbol{w}^{(1)}, \ldots, \boldsymbol{w}^{(N)}\}$代替$\forall \boldsymbol{w} \in W$. 也就是说，我们的目的不是对所有可能的不确定性的实现均要求约束成立，而是对样本场境值要求这些约束成立. 显然，这样所获得的解在绝对的最坏情况下将不再是鲁棒的，但在非严格的概率情况下是鲁棒的. 我们可以通过选择足够大的N以获得一个好的鲁棒性. 对给定的$\alpha \in (0,1)$，由等式(10.30)，如果已经确定了场境数量，那么

$$N \geq \frac{2}{1-\alpha}(n+10), \tag{16.20}$$

其中n是优化问题中决策变量的总数目，那么这个场境解将是水平为α的“概率鲁棒的”(然而，通常对于即使较小的情境数目N，也能获得好的结果）.

按以上描述，我们需要求解一个如下的线性优化问题：

$$\begin{array}{ll} \underset{\bar{\boldsymbol{u}},\boldsymbol{y},\boldsymbol{A}}{\text{minimize}} & \mathbf{1}^{\top}\boldsymbol{y} \\ \text{subject to} & c\boldsymbol{u}^{(i)} + h\boldsymbol{x}^{(i)} \leq \boldsymbol{y},\ i = 1, \ldots, N, \\ & c\boldsymbol{u}^{(i)} - p\boldsymbol{x}^{(i)} \leq \boldsymbol{y},\ i = 1, \ldots, N, \\ & \boldsymbol{u}^{(i)} = \bar{\boldsymbol{u}} + \boldsymbol{A}(\boldsymbol{w}^{(i)} - \hat{\boldsymbol{w}}), \quad i = 1, \ldots, N, \\ & \boldsymbol{x}^{(i)} = \ x_0\mathbf{1} + \boldsymbol{U}\boldsymbol{u}^{(i)} - \boldsymbol{U}\boldsymbol{w}^{(i)}, \quad i = 1, \ldots, N, \\ & 0 \leq \boldsymbol{u}^{(i)} \leq M\mathbf{1},\ i = 1, \ldots, N. \end{array} \tag{16.21}$$

作为一个数值例子，我们再次考虑16.5.2节中的数据. 此时，问题(16.21)的决策变量数目

$$n = T + T + \frac{T(T-1)}{2} = 90.$$

考虑鲁棒性水平$\alpha = 0.9$，由公式(16.20)计算得知，在问题中需要使用$N = 2,000$个情境. 我们进一步假设期望需求由(16.19)给定，向量$\boldsymbol{w}$满足正态分布，其协方差矩阵

$$\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_0^2, \ldots, \sigma_{T-1}^2),$$

其中方差随时间演进的方式(刻画不确定性随时间增加的情况)为

$$\sigma_k^2 = (1+k)\bar{\sigma}^2; \quad k = 0, \ldots, T-1, \quad \text{其中} \bar{\sigma}^2 = 1.$$

[5] 见G.Calafiore的Random convex programs, *SIAM J. Opt.*, 2010.

将这些参数代入场境(16.21)，求解所得的场景问题，得到最坏情况下的最优成本是6,296.80，最优决策如下

$$\bar{\boldsymbol{u}}^T = \begin{bmatrix} 24.4939 & 108.77 & 108.603 & 108.808 & 110.0 & 110.0 & 98.7917 & 83.7308 & 80.6833 & 81.5909 & 89.8084 & 94.6564 \end{bmatrix},$$

$$\boldsymbol{A} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.3774 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.02567 & 0.2845 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -0.1184 & 0.2025 & 0.1344 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.4868 & 0.03163 & 0.496 & 0.4467 & 0.4006 & 0.5332 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.3849 & -0.1964 & 0.254 & 0.2024 & 0.4895 & 0.2111 & 0.6726 & 0 & 0 & 0 & 0 & 0 \\ 0.135 & 0.072 & -0.1547 & 0.1603 & -0.03367 & 0.02405 & 0.04367 & 0.5894 & 0 & 0 & 0 & 0 \\ -0.2695 & 0.6738 & 0.1186 & -0.2117 & 0.08057 & 0.001069 & 0.1795 & 0.2256 & 0.6008 & 0 & 0 & 0 \\ 0.2583 & 0.1248 & 0.04462 & 0.07893 & 0.05989 & 0.09069 & 0.007373 & 0.3262 & 0.2858 & 0.5088 & 0 & 0 \\ 0.4861 & 0.6754 & 0.009505 & 0.3034 & 0.3853 & 0.08153 & 0.1048 & 0.4221 & 0.6108 & 0.3069 & 1.086 & 0 \end{bmatrix}.$$

然后，我们可以通过Monte Carlo法，也就是生成新的场境(与优化中的不同)，来仿真这个决策在随随机需求下的**后验**行为. 使用1,500 个新的需求场境，我们得到了图16.22中的结果. 由各个情境产生的成本的直方图见图16.23. 这个仿真得到的最大成本是$6,190.22$，而场境法中的最坏情况成本为6,296.80. 显然前者要低于后者.

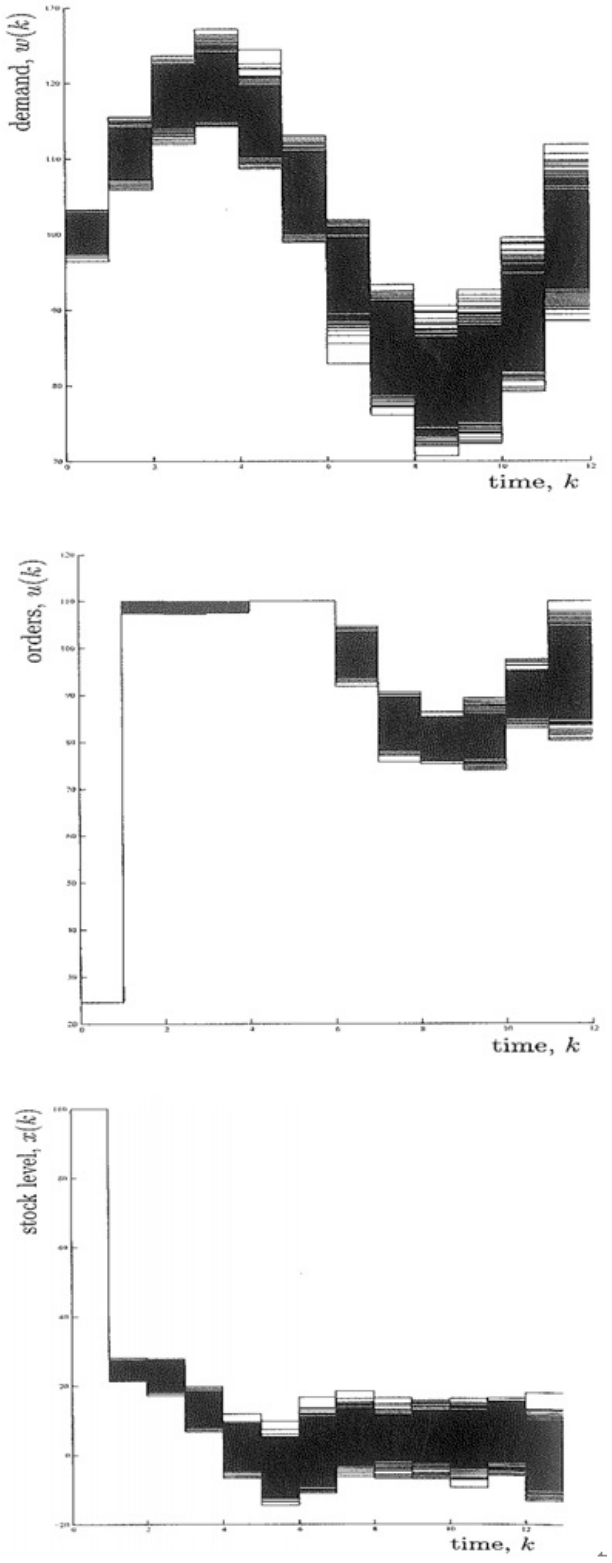

图 16.22: 1,500个随机生成的需求下的最优场境策略的Monte Carlo仿真. 上图：随机需求. 中图：订购曲线. 下图：库存量曲线.

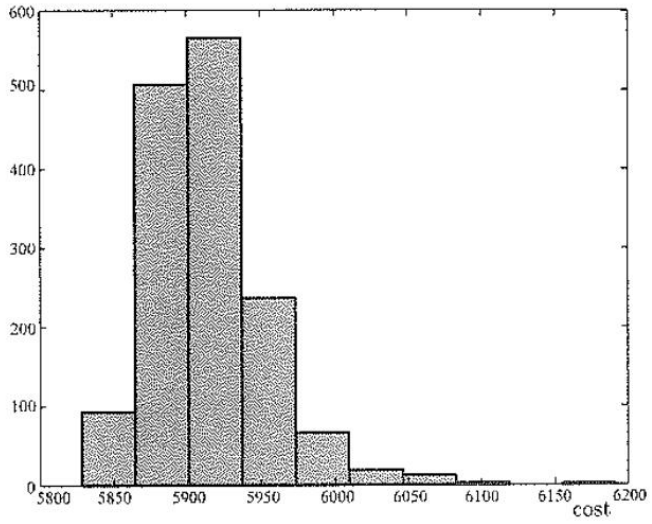

图 16.23: 由1,500个随机生成需求下最优场景策略的Monte Carlo仿真所产生的成本的直方图.

§16.6 练习

习题16.1.（网络拥塞控制）

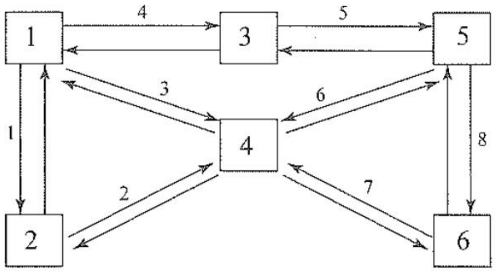

图 16.24: 一个小型网络

图16.24显示了由$n = 6$个计算机组成的P2P网络. 每一台计算机可以在图中所示链接以特定的速率上传和下载数据. 令$\boldsymbol{b}^+ \in \mathbb{R}^8$为一个向量，表示图中标号对应的连接上的数据传输速率，令$\boldsymbol{b}^- \in \mathbb{R}^8$为一向量，表示反向连接的数据传输速率，这里必须有$\boldsymbol{b}^+ \geq \mathbf{0}, \boldsymbol{b}^- \geq \mathbf{0}$.

为此网络定义节点弧关联矩阵

$$\boldsymbol{A} \doteq \begin{bmatrix} 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 \\ 0 & -1 & -1 & 0 & 0 & -1 & -1 & 0 \\ 0 & 0 & 0 & 0 & -1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \end{bmatrix},$$

令$\boldsymbol{A}_+ \doteq \max(\boldsymbol{A}, 0)$($\boldsymbol{A}$的正部)，$A_- \doteq \min(\boldsymbol{A}, 0)$($\boldsymbol{A}$的负部). 由此得到节点的总输出(上传)速率$\boldsymbol{v}_{\text{upl}} = \boldsymbol{A}_+\boldsymbol{b}^+ - \boldsymbol{A}_-\boldsymbol{b}^-$，总输入(下载)速率$\boldsymbol{v}_{\text{dwl}} = \boldsymbol{A}_+\boldsymbol{b}^- - \boldsymbol{A}_-\boldsymbol{b}^+$. 因此节点上网络的出流可以表示为

$$\boldsymbol{v}_{\text{net}} = \boldsymbol{v}_{\text{upl}} - \boldsymbol{v}_{\text{dwl}} = \boldsymbol{A}\boldsymbol{b}^+ - \boldsymbol{A}\boldsymbol{b}^-,$$

此外还需要满足流平衡方程$[v_{\text{net}}]_i = f_i$，如果计算机i既没有产生数据也没有接收数据(它仅仅是传递接收到的数据，即它的表现像一个中继站)，则$f_i = 0$；如果计算机i产生包，则$f_i > 0$；若它接收包，则$f_i < 0$，其中f_i是已分配的速率.

每一台计算机下载数据的最大速率$\bar{v}_{\text{dwl}} = 20$ Mbit/s，上传数据的最大速率$\bar{v}_{\text{upl}} = 10$ Mbit/s (是对每台计算机所有上传或下载连接的速率之和的限制). 每一条连接的拥堵水平定义为

$$c_j = \max(0, (b_j^+ + b_j^- - 4)), \quad j = 1, \ldots, 8.$$

假设点1必须以$f_1 = 9$ Mbit/s的速率传输数据到点5，点2必须以$f_2 = 8$ Mbit/s的速率传输数据到点6. 求出网络平均拥挤水平最小时所有连接的数据传输速率.

习题16.2. （**水库设计问题**）我们需要设计一个用于蓄水和蓄能的水库. 如图16.25所示.

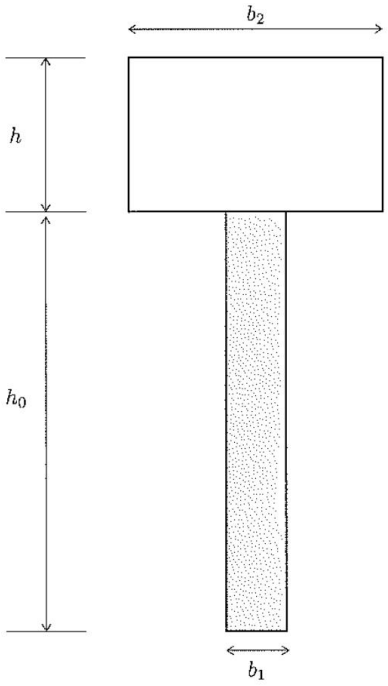

图 16.25: 混凝土基座上的蓄水池

混凝土基座截面为长方形，长为b_1，高为h_0，水库本身截面也为长方形，长为b_2，高为h. 表格16.4中给出了一些需要用到的数据.

表 16.4: 水库问题的数据

参数	结果	单位	描述
g	9.8	$\mathrm{m/s^2}$	重力加速度
E	30×10^9	$\mathrm{N/m^2}$	基座弹性模量
ρ_w	10×10^3	$\mathrm{N/m^3}$	水的密度
ρ_b	25×10^3	$\mathrm{N/m^3}$	基座密度
J	$b_1^4/12$	$\mathrm{m^4}$	基座惯性矩
N_{cr}	$\pi^2 JE/(2h_0)^2$	N	基座的临界载荷限制

基座的临界负载限制为N_{cr}，它至少为水的重量的两倍，结构要满足规范$h_0/b_1^2 \le 35$. 蓄水池的形状应满足$1 \le b_2/h \le 2$. 结构的总高度不能够超过30 m. 结构的总重量(基座和装满水的蓄水池)不应超过9.8×10^5 N. 问

题是找到维度b_1, b_2, h_0, h使得蓄水池中水的势能P_w最大(假设$P_w = (\rho_w h b_2^2) h_0$). 请问这个问题能否被建模为凸优化问题？如果可以的话，请给出建立的模型和解释，并求解所得模型给出最优设计.

习题16.3. （**电路中的导线尺寸问题**） 现代电子芯片间的互连可以被建模为置于基质之上的导电表面积. 因此可以将一根“导线”看成图16.26所示的矩形段序列.

图 16.26: 将导线表示成基质上的矩形表面. 每个片段的长度ℓ_i是固定的，宽度x_i是决策变量. 该例中有三个导线段.

我们假设每个矩形段的长度是固定的，而宽度需要依据下文说明的准则来确定. 一个通用的方法是把导线模型化为数个RC段的级联，其中对每个段而言，$S_i = 1/R_i$和C_i分别是第i个矩形段的电导和电容，参见图16.27.

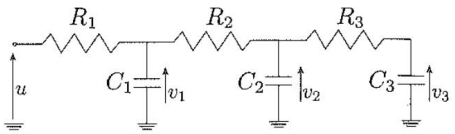

图 16.27: 三分段的导线RC模型

S_i和C_i的值与导线段的表面积成比例，由于已知长度ℓ_i是定值，因此电容和电导是宽度的仿射函数，即

$$S_i = S_i(x_i) = \sigma_i^{(0)} + \sigma_i x_i, \quad C_i = C_i(x_i) = c_i^{(0)} + c_i x_i,$$

这里$\sigma_i^{(0)}, \sigma_i, c_i^{(0)}, c_i$是给定的正常数. 对于图16.27中的三分段导线模型，可以写出如下的动力方程组来描述节点间电压$v_i(t),\ i = 1, \ldots, 3,$ 随时间的变化，即

$$\begin{bmatrix} C_1 & C_2 & C_3 \\ 0 & C_2 & C_3 \\ 0 & 0 & C_3 \end{bmatrix} \dot{\boldsymbol{v}}(t) = - \begin{bmatrix} S_1 & 0 & 0 \\ -S_2 & S_2 & 0 \\ 0 & -S_3 & S_3 \end{bmatrix} \boldsymbol{v}(t) + \begin{bmatrix} S_1 \\ 0 \\ 0 \end{bmatrix} u(t),$$

其中$u(t)$是输入.

如果我们进行换元

$$\boldsymbol{v}(t)=\boldsymbol{Q}\boldsymbol{z}(t),\quad \boldsymbol{Q}=\begin{bmatrix}1&0&0\\1&1&0\\1&1&1\end{bmatrix},$$

则上述方程组就可行表述成更有用的形式. 具体地，利用这里引入的新变量$\boldsymbol{z}(t)$，上述方程变成

$$\boldsymbol{C}(\boldsymbol{x})\dot{\boldsymbol{z}}(t)=-\boldsymbol{S}(x)\boldsymbol{z}(t)+\begin{bmatrix}S_1\\0\\0\end{bmatrix}u(t),$$

其中

$$\boldsymbol{C}(x)\doteq\begin{bmatrix}C_1+C_2+C_3&C_2+C_3&C_3\\C_2+C_3&C_2+C_3&C_3\\C_3&C_3&C_3\end{bmatrix},\quad \boldsymbol{S}(x)\doteq\operatorname{diag}(\mathrm{S}_1,\mathrm{S}_2,\mathrm{S}_3).$$

$\boldsymbol{C}(x),\boldsymbol{S}(x)$显然是对称矩阵，他们的每一个分量是决策变量$\boldsymbol{x}=(x_1,x_2,x_3)$的仿射函数. 此外我们可以看出：只要$\boldsymbol{x}\geq\boldsymbol{0},\boldsymbol{C}(\boldsymbol{x})$就是非奇异的(我们问题中的物理情形就是这样的)，因此$\boldsymbol{z}(t)$的演化可以被表示为(我们接下来假设$u(t)=0$，也就是说我们仅考虑系统自由响应的时间演化)

$$\dot{\boldsymbol{z}}(t)=-\boldsymbol{C}(\boldsymbol{x})^{-1}\boldsymbol{S}(\boldsymbol{x})\boldsymbol{z}(t).$$

定义电路的**主时间常数**(dominant time constant)为

$$\tau=\frac{1}{\lambda_{\min}(\boldsymbol{C}(\boldsymbol{x})^{-1}\boldsymbol{S}(\boldsymbol{x}))'},$$

这是电路“速度”的一种量测(τ越小，电路的响应越快).

在保证主时间常数不超过一个指定水平$\eta>0$的前提下，描述一个可以有效确定导线宽度$\boldsymbol{x}$的方法，其使得导线所占总面积最小.