机器学习的算法方面

Ankur Moitra

Contents

Preface

该专著基于2013年秋季、2015年春季和2017年秋季在麻省理工学院讲授的“机器学习的算法方面”课程。感谢所有参与此课程并使教学成为一次美妙体验的学生和博士后。

To Diana and Olivia, the sunshine in my life

Chapter 1

Introduction

机器学习开始接管我们生活中许多方面的决策，包括：

(a) 在自动驾驶汽车中确保我们日常通勤的安全

(b) 根据我们的症状和病史进行准确诊断

(版权) 复杂证券的定价和交易

(d) 发现新的科学，例如各种疾病的遗传基础。

但是令人震惊的真相是，这些算法在没有任何可证明的行为保证的情况下工作。当它们面对一个优化问题时，它们实际上是否找到了最佳解决方案，甚至是一个相当不错的解决方案？当它们提出一个概率模型时，它们能否纳入新的证据并从真实的后验分布中进行采样？机器学习在实践中工作得非常出色，但这并不意味着我们理解why它为什么工作得如此出色。

如果您已经上过传统的算法课程，您通常接触算法思考的方式是通过最坏情况分析。当您有一个排序算法时，您根据它在最坏可能的输入上需要多少操作来衡量它的运行时间。这是一种方便的类型界限，因为它意味着您可以在不担心通常给出的输入类型的情况下，有意义地谈论您的算法需要多长时间。

但是，分析机器学习算法——尤其是现代算法——之所以具有挑战性，是因为它们试图解决的问题在最坏情况输入下实际上是 NP-难。当你将寻找最适合您数据的参数的问题视为一个优化问题时，存在一些情况，找到良好的拟合是 NP-难的。当你提出一个概率模型并希望使用它进行推理时，也存在一些情况，这同样是 NP-难的。

在这本书中，我们将通过尝试寻找更符合我们数据现实性的模型来解决为机器学习提供可证明保证的问题。在许多应用中，我们可以基于问题出现时的上下文做出合理的假设，这可以帮助我们绕过这些最坏情况下的障碍，并允许我们严格分析实践中使用的启发式方法，以及设计解决机器学习中一些核心、反复出现的问题的根本新方法。

退一步说，超越最坏情况分析的构想与理论计算机科学本身一样古老[1] [95]。事实上，关于理解算法在“典型”实例上的行为意味着什么，有许多不同的观点，包括：

[1]After all, heuristics performing well on real life inputs are old as well (long predating modern machine learning) and hence so is the need to explain them.

(a)概率模型针对您的输入——甚至混合模型，这些模型结合了最坏情况和平均情况分析的因素，如半随机模型[38, 71]或平滑分析[39, 130]

(b) 测量您问题复杂度的方法，并要求算法在简单输入上快速，如参数化复杂性[66]

(c) 尝试阐述哪些问题实例具有有意义的答案，并且是您实际上想要解决的问题的稳定性概念 [20, 32]

这不是一个详尽的主题或参考文献列表。无论如何，在这本书中，我们将带着这些关于如何克服不可行性的见解来处理机器学习问题。

最终，我们希望理论计算机科学和机器学习还有很多可以互相教授的东西。理解为什么像期望最大化或非凸函数上的梯度下降这样的启发式方法在实践中表现得如此之好，是理论计算机科学的一个重大挑战。但是，要在这类问题上取得进展，我们需要了解在机器学习背景下哪些类型的模型和假设是有意义的。另一方面，如果我们能在这些难题上取得进展，并对why启发式方法工作得如此之好有新的见解，我们就可以希望将它们工程化得更好。我们甚至可以希望发现解决机器学习中一些重要问题的全新方法，特别是通过利用我们算法工具箱中的现代工具。

在这本书中，我们将涵盖以下主题：

(a) 非负矩阵分解

(b) 主题建模

(c) 张量分解 (d) 稀疏恢复 (e) 稀疏编码 (f) 学习混合模型 (g) 矩阵补全

我希望在以后的版本中增加更多章节，因为随着该领域的发展和新发现的出现。

Chapter 2

Nonnegative Matrix Factorization

在这一章中，我们将探讨非负矩阵分解问题。首先将其与更熟悉的奇异值分解进行比较将是有帮助的。在最坏的情况下，非负矩阵分解问题是 NP-难（真的，你还能期待什么？）但我们将做出特定领域的假设（称为 *separability*），这将使我们能够为它的重要特殊情况提供可证明的算法。然后我们将我们的算法应用于学习主题模型参数的问题。这将是我们在面对计算不可行性时如何不退缩的第一个案例研究，并找到绕过它的方法。

2.1 Introduction

为了更好地理解非负矩阵分解问题的动机以及它在应用中的有用性，首先介绍奇异值分解并将它们进行比较将是有益的。最终，我们将

将这两个都应用于本节后面的文本分析。

The Singular Value Decomposition

奇异值分解（SVD）是线性代数中最有用的工具之一。给定一个 $m \times n$ 矩阵 M, 其奇异值分解可表示为

$$M = U\Sigma V^T$$

在 U 和 V 是正交归一的情况下，Σ 是一个只有对角线及其对角线元素非零的矩形矩阵，且其元素非负。或者我们可以写成

$$M = \sum_{i=1}^{r} \sigma_i u_i v_i^T$$

在 u_i 是 U 的 i^{th} 列，v_i 是 V 的 i^{th} 列，σ_i 是 Σ 的 i^{th} 对角元素的情况下。在本节中，我们将固定约定 $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_r > 0$。在这种情况下，$M$ 的秩恰好是 r。

在整个课程中，我们将有机会使用这种分解以及（可能更熟悉的）特征分解。如果 M 是一个 $n \times n$ 矩阵并且是可对角化的，其特征分解可表示为

$$M = PDP^{-1}$$

D 是对角线。目前需要记住的重要事实是:

(1) **Existence:** 每个矩阵都有一个奇异值分解，即使它是一个

矩形。相比之下，一个矩阵必须是方阵才能有特征分解。即使如此，并非所有方阵都可以对角化，但一个充分条件是M的所有特征值都不同。

(2) **Algorithms:** 这两种分解都可以高效地计算。计算奇异值分解的最佳通用算法在 $m \geq n$ 的情况下运行时间为 $O(mn^2)$。还有针对稀疏矩阵的更快算法。有算法可以在 $O(n^3)$ 时间内计算特征分解，并且基于快速矩阵乘法还有进一步的改进，尽管尚不清楚这些算法是否具有相同的稳定性和实用性。

(3) **Uniqueness:** 单值分解是唯一的，当且仅当其奇异值互不相同。同样，特征分解是唯一的，当且仅当其特征值互不相同。在某些情况下，我们只需要非零奇异值/特征值互不相同，因为我们可以忽略其他值。

Two Applications

两个最重要的奇异值分解特性是它可以用来找到最佳秩 k 近似，以及它可以用于降维。我们接下来探讨这些。首先，让我们正式化我们所说的最佳秩 k 近似问题。一种方法是使用弗罗贝尼乌斯范数:

Definition 2.1.1 (Frobenius norm) $\|M\|_F = \sqrt{\sum_{i,j} M^{2i,j}}$

可以看出，Frobenius范数在旋转下是不变的。例如，这可以通过单独考虑M的每一列作为一个向量来得出。矩阵的Frobenius范数的平方是其列范数平方的和。然后，左乘一个正交矩阵可以保持其每一列的范数。对于右乘一个正交矩阵（但处理行而不是列）也有相同的论据。这种不变性使我们能够给出Frobenius范数的另一种描述，这非常有用：

$$\|M\|_F = \|U^T M V\|_F = \|\Sigma\|_F = \sqrt{\sum \sigma_i^2}$$

第一个等式是所有动作发生的地方，并使用了我们在上面建立的旋转不变性属性。

然后，Eckart-Young 定理断言，在 Frobenius 范数) 的意义上，某些矩阵 M (的最佳秩 k 近似由其截断奇异值分解给出：

Theorem 2.1.2 (Eckart-Young) $\underset{\text{rank}(B)\leq k}{\text{argmin}} \|M - B\|_F = \sum_{i=1}^{k} \sigma_i u_i v_i^T$

设 M_k 为最佳秩 k 近似。然后，根据我们关于 Frobenius 范数的替代定义，可以立即得出 $\|M - M_k\|_F = \sqrt{\sum_{i=k+1}^{r} \sigma_i^2}$。

实际上，相同的陈述——即 k 对 M 的最佳秩 {v*} 近似是其截断奇异值分解——对于在旋转下不变的空间 any 范数也成立。作为另一个应用，考虑算子范数：

Definition 2.1.3 (operator norm) $\|M\| = \max_{\|x\|\leq 1} \|Mx\|$

它很容易看出算子范数在旋转下也是不变的，并且$\|M\| = \sigma_1$，再次使用σ_1是最大的奇异值的惯例。然后关于算子范数的Eckart-Young定理断言：

Theorem 2.1.4 (Eckart-Young) $\underset{\text{rank}(B)\leq k}{\text{argmin}} \|M - B\| = \sum_{i=1}^{k} \sigma_i u_i v_i^T$

再次令 M_k 为最佳秩 k 近似。然后 $\|M - M_k\| = \sigma_{k+1}$。作为一个快速检查，如果 $k \geq r$ 则 $\sigma_{k+1} = 0$，并且最佳秩 k 近似是精确的且没有误差（正如它应该的那样）。你应该将此视为你可以用任何计算 M 的奇异值分解的算法来做的事情——你可以找到任何旋转不变范数下的最佳秩 k 近似。事实上，令人惊讶的是，在许多不同的范数下，最佳秩 k 近似是一致的！此外，M 的最佳秩 k 近似可以直接从其最佳秩 $k + 1$ 近似中获得。这并不总是如此，正如我们将在下一章中与张量一起工作时将看到的。

接下来，我们在数据分析的背景下给出奇异值分解的另一种完全不同的应用，在我们继续探讨其在文本分析中的应用之前。回想一下，M是一个$m \times n$矩阵。我们可以将其视为定义在n维向量上的分布，这些向量是通过随机均匀选择其列之一获得的。进一步假设$\mathbb{E}[x] = 0$ - 即列的和为零向量。令$\mathcal{P}_k$为所有投影到k维子空间的空间。

Theorem 2.1.5 $\underset{P \in \mathcal{P}_k}{\text{argmax}} \mathbb{E}[\|Px\|^2] = \sum_{i=1}^{k} u_i u_i^T$

这是关于奇异值分解的另一个基本定理，从中我们可以轻松计算出最大化...的 k-维投影

预测方差。这个定理在可视化中经常被引用，其中可以通过将其投影到一个更易于管理的、低维子空间来可视化高维向量数据。

Latent Semantic Indexing

现在我们已经开发了一些关于奇异值分解背后的直觉，我们将看到它在文本分析中的应用。这个领域的一个核心问题（我们将在很多地方回到这个问题）是，给定大量文档，我们想要提取一些隐藏的 *thematic* 结构。Deerwester 等人 [60] 为了这个目的发明了潜在语义索引（LSI），他们的方法是将奇异值分解应用于通常称为词-文档矩阵的内容：

Definition 2.1.6 *The term-by-document matrix M is an $m \times n$ matrix where each row represents a word, each column represents a document where*

$$M_{i,j} = \frac{\text{count of word } i \text{ in document } j}{\text{total number of words in document } j}$$

有许多流行的归一化惯例，在这里我们选择将矩阵归一化，使其每一列的和为1。这样，我们可以将每个文档解释为词语上的概率分布。在构建词-文档矩阵时，我们还忽略了词语出现的顺序。这被称为*bag-of-words representation*，其合理性来源于一个思想实验。假设我给你一个文档中包含的词语，但顺序是混乱的。仍然应该能够确定文档

关于，因此忘记所有语法和语法的概念，将文档表示为向量会丢失一些结构，但仍然应该保留足够的信息，以便在文本分析中仍然可以完成许多基本任务。

一旦我们的数据以向量形式存在，我们就可以利用线性代数的工具。我们如何衡量两份文档之间的相似度？直观的方法是依据它们共有多少个单词来衡量相似度。让我们试试：

$$\langle M_i, M_j \rangle$$

这个数量计算从文档 i 中随机选择的单词 w 和从文档 j 中随机选择的单词 w' 是否相同的概率。但使这个度量变得糟糕的是，当文档稀疏时，它们可能只是因为每个作者选择使用特定的单词来描述相同类型的事物而偶然没有很多共同单词。更糟糕的是，一些文档可能被认为相似，因为它们都包含许多相同的常见单词，而这些单词与文档实际上关于的内容关系不大。

Deerwester等人[60]提出使用M的奇异值分解来计算一个更合理的相似度度量，这种度量在术语-文档矩阵稀疏时（通常如此）似乎效果更好。设$M = U\Sigma V^T$，设$U_{1...k}$和$V_{1...k}$分别为U和V的前k列。该方法计算如下：

$$\langle U^T_{1...k} M_i, U^T_{1...k} M_j \rangle$$

对于每对文档。直觉是，有一些$topics$在文档集合中反复出现。如果我们能够表示每个

文档 M_i 在主题的基础上，其在该基底的内在积将产生一个更有意义的相似度度量。有一些模型——即关于数据如何随机生成的假设——可以证明这种方法可以证明性地恢复真正的主题[118]。这是理论与实践的理想互动——我们有（某种程度上）有效的工作技术，我们可以分析/证明它们。

然而，潜在语义索引存在许多缺陷，这促使人们寻求替代方法。如果我们把最高阶的奇异向量与主题关联起来，那么：

(1) 主题是正交归一的

然而，像 $politics$ 和 $finance$ 这样的主题实际上包含许多共同词汇。

(2) 主题包含负值

因此，如果一个文档包含这样的词，它们对（该主题）的贡献可能会抵消其他词的贡献。此外，可以判断一对文档是相似的，因为它们都不涉及特定的主题。

Nonnegative Matrix Factorization

对于我们在上一节中描述的恰好是这些缺陷，非负矩阵分解在文本分析的许多应用中是奇异值分解的一个流行替代方案。然而，它也有自己的不足之处。与奇异值分解不同，它是 NP-难计算的。而在实践中普遍采用的方法是依赖于没有可证明保证的启发式方法。

Definition 2.1.7 *A nonnegative matrix factorization of inner-dimension r is a decomposition*

$$M = AW$$

where A is $n \times r$, W is $r \times n$ and both are entry-wise nonnegative. Moreover let the nonnegative rank of M – denoted by 排名$^+(M)$ – be the minimum r so that such a factorization exists.

我们将会看到，当这个分解应用于词-文档矩阵时，可以找到更多可解释的主题。除了文本分析之外，它在机器学习和统计学中还有许多其他应用，包括协同过滤和图像分割。现在，让我们在文本分析的具体背景下对非负矩阵分解进行解释。假设我们将其应用于词-文档矩阵。结果发现，我们总能将其置于一个方便的正则形式：设D为一个对角矩阵，其中

$$D_{j,j} = \sum_{i=1}^{m} A_{i,j}$$

并且进一步假设每个 $D_{j,j} > 0$。那么

Claim 2.1.8 *Set $\widetilde{A} = AD^{-1}$ and $\widetilde{W} = DW$. Then*

(1) $\widetilde{A}, \widetilde{W}$ are entry-wise nonnegative and $M = \widetilde{A}\widetilde{W}$

(2) The columns of $\widetilde{A}$ and the columns of $\widetilde{W}$ each sum to one

我们把这个命题的证明留给读者作为练习，但提示是，因为 M 的列之和也等于一，所以性质（2）成立。

因此，我们可以不失一般性地假设我们的非负矩阵分解 $M = AW$ 满足 A 的列和 W 的列各自之和为 1。然后我们可以这样解释这种分解：每份文档本身是一个关于单词的分布，而我们找到的是：

(1) 一组r主题的集合——A的列——这些主题本身是关于单词的分布

(2) 对于每个文档 i，它的表示 – 由 W_i 给出 – 是 r 主题的凸组合，以便我们恢复其在单词上的原始分布

稍后，我们将了解为什么非负矩阵分解是 NP-难的原因。但实际计算这种分解通常采用什么方法呢？通常的方法是 *alternating minimization*：

Alternating Minimization for NMF

Input: $M \in \mathbb{R}^{m \times n}$

Output: $M \approx A^{(N)}W^{(N)}$

Guess entry-wise nonnegative $A^{(0)}$ of dimension $m \times r$

For $i = 1$ to N

Set $W^{(i)} \leftarrow \text{argmin}_W \|M - A^{(i-1)}W\|_F^2$ s.t. $W \geq 0$

Set $A^{(i)} \leftarrow \text{argmin}_A \|M - AW^{(i)}\|_F^2$ s.t. $A \geq 0$

End

交替最小化相当通用，在整个课程中我们都会遇到

返回多次并发现我们感兴趣的问题在实践中使用上述基本方法的一些变体得到解决。然而，在传统意义上，它没有可证明的保证。它可能会失败，陷入一个比全局最优解差得多的局部最优解。事实上，这是不可避免的，因为它试图解决的问题实际上是 NP-难。

然而，在许多情况下，我们可以通过使用适当的随机模型来取得进展，其中我们可以证明它收敛到全局最优解。本课程的一个主要主题是不将实践中看似有效的启发式方法视为理所当然，因为能够分析它们本身将提供新的见解，了解它们何时以及为什么有效，以及可能出错的情况以及如何改进它们。

2.2 Algebraic Algorithms

在上一节中，我们介绍了非负矩阵分解问题，并描述了它在机器学习和统计学中的一些应用。事实上，（由于问题的代数性质）在最坏情况下，是否存在任何有限时间算法来计算它尚不清楚。在这里，我们将探讨解决多项式方程组的一些基本结果，并从中推导出非负矩阵分解的算法。

Rank vs. Nonnegative Rank

回忆起，$\text{rank}^+(M)$ 是满足 M 具有内维数 r 的非负矩阵分解 $M = AW$ 的最小值 r。容易看出以下

这是另一个等效的定义：

Claim 2.2.1 *$rank^+(M)$ is the smallest r such that there are r entry-wise nonnegative rank one matrices $\{M_i\}$ that satisfy $M = \sum_i M_i$.*

我们现在可以比较秩和非负秩。当然，矩阵秩有许多等价定义，但比较这两个定义最方便的是以下定义：

Claim 2.2.2 *$rank(M)$ is the smallest r such that there are r rank one matrices $\{M_i\}$ that satisfy $M = \sum_i M_i$.*

这两个定义之间的唯一区别在于，前者规定分解中的所有一阶矩阵的元素都是非负的，而后者则没有这样的规定。因此，可以立即得出以下结论：

Fact 2.2.3 *$rank^+(M) \geq rank(M)$*

矩阵的非负秩能否远大于其秩？我们鼓励读者在继续之前思考这个问题。这相当于问，对于每个元素都是非负的矩阵 M，是否可以不失一般性地要求其在秩分解中的因子也必须是每个元素都是非负的。对于秩为1的矩阵，这无疑是正确的，而且对于秩为2的矩阵也证明是正确的，但...

一般来说，非负秩不能由秩的任何函数来界定。事实上，秩与非负秩之间的关系（或缺乏关系）在理论领域的许多领域中具有根本重要性。

计算机科学。幸运的是，有一些简单的例子可以说明这两个参数可以相距甚远：

Example: 设 M 为一个 $n \times n$ 矩阵，其中 $M_{ij} = (i-j)^2$。

它很容易看出 M 的列空间由以下三个向量张成

$$\begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 2 \\ \vdots \\ n \end{bmatrix}, \begin{bmatrix} 1 \\ 4 \\ \vdots \\ n^2 \end{bmatrix}$$

因此，rank(M) ≤ 3。（事实上，rank(M) $= 3$）。然而，M 在对角线上有零，而在对角线之外有非零值。此外，对于任何秩为1，逐项非负矩阵 M_i，其零和非零的模式是一个 *combinatorial rectangle* —— 即某些行和列的交集 —— 并且可以证明至少需要 $\log n$ 个矩形来覆盖 M 的非零值，而不覆盖其任何零值。因此：

Fact 2.2.4 $rank^+(M) \geq$ 对数 n

A word of caution: 对于这个例子，许多作者错误地试图证明一个更强的下界（例如，rank$^+(M) = n$）。事实上（也许有些令人惊讶）结果是rank$^+(M) \leq 2 \log n$。通常的错误在于认为因为矩阵的秩是最大的 r，使得它有 r 线性无关的列，所以非负秩是最大的 r，使得没有列是其他 r 列的凸组合 $r-1$。这并不正确！

Systems of Polynomial Inequalities

我们可以将判断rank$^+(M) \leq r$的问题重新表述为寻找特定多项式不等式系统可行解的问题。更具体地说，rank$^+(M) \leq r$当且仅当：

$$\begin{cases} M & = AW \\ A & \geq 0 \\ W & \geq 0 \end{cases} \tag{2.1}$$

具有解。该系统由二次等式约束（每个M的条目一个）和线性不等式组成，这些不等式要求A和W逐项非负。在我们担心快速算法之前，我们应该问一个更基本的问题（答案并不明显）：

Question 1 *Is there any finite time algorithm for deciding if rank$^+(M) \leq r$?*

这是等价于决定上述线性系统是否有解，但困难在于即使有一个解，A和W的项也可能是无理数。这与3-SAT不同，在3-SAT中有一个简单的穷举算法。相比之下，对于非负矩阵分解，设计在有限时间内运行的算法相当具有挑战性。

但确实存在算法（在固定时间内运行）来决定一个多项式不等式系统在实RAM模型中是否有解。解决多项式不等式系统的第一个有限时间算法源于Tarski的开创性工作，并且一直有一系列

基于越来越强大的代数分解的改进。这一系列工作最终导致了Renegar的以下算法：

Theorem 2.2.5 *[126] Given a system of m polynomial inequalities in k variables, whose maximum degree is D and whose bit complexity is L, there is an algorithm whose running time is*

$$(nDL)^{O(k)}$$

and decides whether the system has a solution. Moreover, if it does have a solution then it outputs a polynomial and an interval (one for each variable) in which there is only one root, which is the value of the variable in the true solution.

注意，此算法找到一个隐式解的表示，因为你可以通过执行二分搜索根来找到你想要的任何数量的解位。此外，此算法本质上是最优的，改进它将产生3-SAT的亚指数时间算法。

我们可以使用这些算法来解决非负矩阵分解，这立即意味着存在一个算法可以决定$\text{rank}^+(M) \leq r$，该算法在指数时间内运行。然而，在朴素表示中，我们需要$nr + mr$个变量，每个变量对应于A或W中的每个条目。因此，即使$r = O(1)$，我们仍然需要一个线性数量的变量，运行时间仍然是指数级的。事实证明，尽管朴素表示使用了许多变量，但存在一种更巧妙的表示，它使用的变量要少得多。

Variable Reduction

这里我们探讨了找到一个多项式方程组来表示非负矩阵分解问题的想法，该方程组使用许多更少的变量。在[13, 112]中，Arora等人以及Moitra给出一个具有$f(r) = 2r^2$变量的多项式不等式系统，当且仅当rank$^+(M) \leq r$时，该系统有解。这立即给出一个多项式时间算法来计算任何$r = O(1)$的内维r (的非负矩阵分解（如果存在)）。这些算法在最坏情况下本质上是最优的，在此工作之前，即使是对于$r = 4$的情况，已知的最优算法也是指数时间复杂度。

我们将关注一个特殊情况，以说明变量减少背后的基本思想。假设rank(M) $= r$，我们的目标是决定是否rank$^+(M) = r$。这被称为单纯形分解问题。我们能否找到一个使用许多更少变量的多项式不等式系统来表示这个决策问题？以下这个简单但有用的观察将铺平道路:

Claim 2.2.6 *In any solution to the simplicial factorization problem, A and W must have full column and row rank respectively.*

Proof: 如果 $M = AW$，则 A 的列跨度必须包含 M 的列，同样地，W 的行跨度必须包含 M 的行。由于 rank(M) $= r$，我们得出结论，A 和 W 必须分别具有 r 线性无关的列和行。由于 A 有 r 列，W 有 r 行，这表明了该命题。■

因此我们知道 A 有一个左伪逆 A^+，而 W 有一个右伪逆 W^+，使得 $A^+A = WW^+I_r$，其中 I_r 是 $r \times r$ 单位矩阵。我们将

利用这些伪逆来减少我们多项式不等式系统的变量数量。特别是:

$$A^+AW = W$$

因此，我们可以从 M 的列的线性变换中恢复 W 的列。同样，我们可以从 M 的行的线性变换中恢复 A 的行。这导致以下替代的多项式不等式系统:

$$\begin{cases} MW^+A^+M & = M \\ MW^+ & \geq 0 \\ A^+M & \geq 0 \end{cases} \tag{2.2}$$

在先验条件下，我们并不清楚是否取得了进展，因为该系统也有与A^+和W^+的条目相对应的$nr + mr$个变量。然而，考虑矩阵MW^+。如果我们把S^+表示为一个$n \times r$矩阵，那么我们正在描述它在所有向量上的作用，但关键观察结果是，我们只需要知道S^+如何作用于由M生成的行，这些行跨越一个r维的空间。因此，我们可以应用基变换来写出

$$M_C = MU$$

在 U 是一个具有右伪逆的 $n \times r$ 矩阵的情况下。同样，我们可以写出

$$M_R = VM$$

在 V 是一个具有左伪逆的 $r \times m$ 矩阵的情况下。现在我们得到一个新的系统：

$$\begin{cases} M_C S T M_R & = M \\ M_C S & \geq 0 \\ T M_R & \geq 0 \end{cases} \tag{2.3}$$

注意，S 和 T 都是 $r \times r$ 矩阵，因此总共有 $2r^2$ 个变量。此外，这个公式在以下意义上等价于单纯形分解问题：

Claim 2.2.7 *If* $rank(M) = rank^+(M) = r$ *then (2.3) has a solution.*

Proof: 使用上述符号，我们可以设定 $S = U^+W^+$ 和 $T = A^+V^+$。然后 $M_C S = MUU^+W^+ = A$，同样地 $TM_R = A^+V^+VM = W$，这暗示了该命题。 ■

这通常被称为 *completeness*，因为如果原问题有解，我们希望我们的重新表述也有有效解。我们还需要证明 *soundness*，即任何对重新表述的解都对应原问题的有效解：

Claim 2.2.8 *If there is a solution to (2.3) then there is a solution to (2.1).*

Proof: 对于(2.3)的任何解，我们可以设定$A = M_C S$和$W = TM_R$，从而有 $A, W \geq$ = 0和$M = AW$。 ■

扩展上述思想到一般的非负矩阵分解相当复杂。[112]中的主要思想是首先建立一个新的

非负矩阵分解的正常形式，并利用观察结果，尽管A可能有指数级数量的线性无关列的最大集合，但它们的伪逆在代数上是相互依赖的，可以使用Cramer法则在公共的r^2变量集中表示。此外，Arora等人[13]表明，任何在$(nm)^{o(r)}$时间内解决甚至单纯形分解问题的算法，都会得到一个3-SAT的亚指数时间算法，因此，在标准复杂度假设下，上述算法几乎是最佳的。

Further Remarks

在本文节中，我们给出了一个简单的例子，说明了秩和非负秩之间的分离。事实上，在理论计算机科学中，存在更多有趣的分离例子。一个自然的问题是，将一个具有指数多个面的高维多面体 Q 投影到一个具有多项式多个面的特定多面体 P 中，该多面体在 n 维度中。这被称为 $extended\ formulation$，Yannakakis 的一个深刻结果是，任何这样的 Q（称为 P 的 $extension\ complexity$）的最小面数恰好等于与顶点和面之间的几何排列有关的某个矩阵的非负秩。然后，存在显式多面体 P 的扩展复杂度是指数的，这与找到显示其秩和非负秩之间大分离的显式矩阵密切相关。

此外，非负秩在通信复杂度中也有重要应用，其中最重要的未解问题之一——$log\text{-}rank\ conjecture$ [108]——可以重新表述为询问：给定一个布尔矩阵 M，是否

对数秩$^+(M) \leq$ (对数秩(M))$^{O(1)}$? 因此，在上面的例子中，非负秩不能被任何秩的函数所界定的这一事实，可能是因为M的项取了许多不同的值。

2.3 Stability and Separability

这里我们将给出非负矩阵分解的几何（与代数相对）解释，这将揭示为什么在最坏情况下它很难，以及哪些类型的特征使其变得容易。特别是，我们将超越最坏情况分析，并使用一个称为 *separability* 的新假设来工作，这将使我们能够给出一个多项式时间运行的算法（即使对于 r 的较大值）。这个假设最初被引入来理解非负矩阵分解问题具有唯一解的条件 [65]，这是算法设计中的一个常见主题。

Theme 1 *Looking for cases where the solution is unique and robust, will often point to cases where we can design algorithms with provable guarantees in spite of worst-case hardness results.*

Cones and Intermediate Simplicies

这里我们将发展一些关于非负矩阵分解的几何直觉——或者更确切地说，是我们在上一节中介绍的一个重要特殊情况，称为单纯形分解。首先，让我们引入锥的概念:

Definition 2.3.1 *Let A be an $m \times r$ matrix. Then the cone generated by the columns of A is*

$$\mathcal{C}_A = \{Ax | x \geq 0\}$$

我们可以立即将其与非负矩阵分解联系起来。

Claim 2.3.2 *Given matrix M, A of dimension $m \times n$ and $m \times r$ respectively, there is an entry-wise nonnegative matrix W of dimension $r \times n$ with $M = AW$ if and only if $\mathcal{C}_M \subseteq \mathcal{C}_A$.*

Proof: 在正向方向，假设 $M = AW$，其中 W 是逐项非负的。那么任何向量 $y \in \mathcal{C}_M$ 可以写成 $y = Mx$，其中 $x \geq \geqslant 0$，然后 $y = AWx$ 和向量 $Wx \geq \geqslant 0$，因此 $y \in \mathcal{C}_A$ 也同样。在反向方向，假设 $\mathcal{C}_M \subseteq \mathcal{C}_A$。那么任何列 $M_i \in \mathcal{C}_A$，我们可以写成 $M_i = AW_i$，其中 $W_i \geq \geqslant 0$。现在我们可以将 W 设置为列向量是 $\{W_i\}_i$ 的矩阵，这样就可以完成证明了。■

非负矩阵分解的难点在于 A 和 W 都未知（如果其中一个已知——比如说 A ——那么我们可以通过设置一个适当的线性规划来求解另一个，这相当于在 $\mathcal{C}_A$ 中表示 M 的每一列）。

Vavasis [139] 首次引入了单纯形分解问题，他的一个动机是因为它最终与一个关于在两个给定的多面体之间拟合单纯形纯粹几何问题相关联。这被称为中间单纯形问题：

Definition 2.3.3 *An instance of the intermediate simplex problem consists of P and Q with $P \subset Q \subset \mathbb{R}^{r-1}$ and P is specified by its vertices and Q is specified by its facets. The goal is to find a simplex K with $P \subseteq K \subseteq Q$.*

在下一段中，我们将证明单纯形分解问题和中间单纯形问题是等价的。

Reductions

我们将证明单纯复分解问题和中间单纯形问题是等价的，即在两个方向上都有多项式时间约简。我们将通过几个中间问题来完成这一证明。

假设我们给定了一个单纯形分解问题的实例。然后我们可以写出 $M = UV$，其中 U 和 V 的内部维度为 r，但它们不一定逐项非负。如果我们能找到一个可逆的 $r \times r$ 矩阵 T，其中 UT 和 $T^{-1}V$ 都是逐项非负的，那么我们就找到了一个有效的非负矩阵分解，其内部维度为 r。

Claim 2.3.4 *If $rank(M) = r$ and $M = UV$ and $M = AW$ are two factorizations that have inner-dimension r then*

(1) $colspan(U) = colspan(A) = colspan(M)$ and

(2) $rowspan(V) = rowspan(W) = rowspan(M)$.

这源于线性代数的基本事实，并且意味着任何两个这样的分解 $M = UV$ 和 $M = AW$ 都可以被线性变换为彼此

通过某个可逆 $r \times r$ 矩阵 T。因此，中间单纯形问题等价于：

Definition 2.3.5 *An instance of the problem* ***P1*** *consists of an* $m \times n$ *entry-wise nonnegative matrix* M *with rank*$(M) = r$ *and* $M = UV$ *with inner-dimension* r*. The goal is to find an invertible* $r \times r$ *matrix where both* UT *and* $T^{-1}V$ *are entry-wise nonnegative.*

Caution: 这个事实表明，你可以从一个任意的分解开始，并要求将其旋转成最小内维的非负矩阵分解，但你并没有把自己逼入死角，这仅限于单纯分解问题！在一般情况下，当rank(M) $<$ rank$^+(M)$时不成立。

现在我们可以给出 **P1** 的几何解释。

(1) 令 $u_1, u_2, ..., u_m$ 为 U 的行。

(2) 令 $t_1, t_2, ..., t_r$ 为 T 的列。

(3) 令 $v_1, v_2, ..., v_n$ 为 V 的列。

我们将首先处理一个中间锥问题，但它与中间单纯形问题的联系将是直接的。为此，设P是由$u_1, u_2, ..., u_m$生成的锥，设K是由$t_1, t_2, ..., t_r$生成的锥。最后，设Q是由以下生成的锥：

$$Q = \{x | \langle u_i, x \rangle \geq 0 \text{ for all } i\}$$

可以看出，Q 在某种意义上是一个锥体，因为它是由所有非负的生成。

有限向量集的组合（其极射线），但我们选择通过其支撑超平面（通过原点）来表示它。

Claim 2.3.6 *UT is entry-wise nonnegative if and only if $\{t_1, t_2, ..., t_r\} \subseteq Q$.*

这直接遵循于 Q 的定义，因为 U 的行是其支持超平面（通过原点）。因此，我们得到了约束 UT 在 **P1** 中逐项非负的几何重新表述。接下来，我们将解释另一个约束，即 $T^{-1}V$ 也逐项非负。

Claim 2.3.7 *$T^{-1}V$ is entry-wise nonnegative if and only if $\{v_1, v_2, ..., v_m\} \subseteq K$.*

Proof: 考虑 $x_i = T^{-1}v_i$。然后 $Tx_i = T(T^{-1})v_i = v_i$ 因此 x_i 是 v_i 作为 $\{t_1, t_2, ..., t_r\}$ 的线性组合的表示。此外，这是唯一的表示，因此这完成了证明。■

因此 **P1** 等价于以下问题：

Definition 2.3.8 *An instance of the intermediate cone problem consists of cones P and Q with $P \subset Q \subset \mathbb{R}^{r-1}$ and P is specified by its extreme rays and Q is specified by its supporting hyperplanes (through the origin). The goal is to find a cone K with r extreme rays and $P \subseteq K \subseteq Q$.*

此外，通过将其中圆锥与超平面相交，可以轻易地看出中间圆锥问题与中间单纯形问题等价，在这种情况下，具有极端射线的圆锥成为这些射线与超平面交点的凸包。

Geometric Gadgets

Vavasis利用上一节中的等价关系构建了某些几何装置来证明非负矩阵分解是NP-难。想法是构建一个二维装置，其中只有两种可能的中间三角形，然后可以用来表示变量x_i的真值分配。完整的归约描述及其正确性的证明是复杂的（参见[139]）。

Theorem 2.3.9 *[139] Nonnegative matrix factorization, simplicial factorization, intermediate simplex, intermediate cone and* ***P1*** *are all NP-hard.*

Arora等人[13]通过构建具有更多选择的低维小工具来改进这种简化。这使得他们能够将问题从d-SUM问题中减少，其中我们给定一个n数字的集合，目标是找到一个d的子集，使其和为零。已知的最优算法运行时间大约为$n^{\lceil d/2 \rceil}$。同样，完整的构建和正确性的证明都涉及在内。

Theorem 2.3.10 *Nonnegative matrix factorization, simplicial factorization, intermediate simplex, intermediate cone and* ***P1*** *all require time at least $(nm)^{\Omega(r)}$ unless there is a subexponential time algorithm for 3-SAT.*

在所有我们将要讨论的主题中，理解是什么使得问题困难，以便希望识别出是什么使得它容易，这是非常重要的。上述所有小工具的共同特征是，这些小工具本身非常不稳定，并且有多个解决方案，因此寻找答案本身既稳健又独特的情况，以识别出可以比最坏情况更有效地解决的问题实例，这是自然的。

Separability

实际上，Donoho 和 Stodden [64] 是最早探索何种条件意味着最小内维度的非负矩阵分解是唯一的一批人之一。他们的原始例子来自玩具问题图像分割，但似乎条件本身在文本分析的环境中可以最自然地被解释。

Definition 2.3.11 *We call A* 可分离 *if, for every column i of A, there is a row j where the only nonzero is in the i^{th} column. Furthermore, we call j an anchor word for column i.*

实际上，可分性在文本分析背景下相当自然。回想一下，我们将 A 的列解释为主题。我们可以将可分性视为这些主题附带 *anchor words* 的承诺；非正式地说，对于每个主题，都有一个未知的锚词，如果它在文档中出现，则该文档（部分）关于给定的主题。例如，*401k* 可能是主题 *personal finance* 的锚词。似乎自然语言包含许多这样的高度特定词汇。

我们现在将给出一个寻找锚词的算法，以及解决非负矩阵分解实例的算法，其中未知量 A 可以在多项式时间内分离。

Theorem 2.3.12 *[13] If $M = AW$ and A is separable and W has full row rank then the* **Anchor Words Algorithm** *outputs A and W (up to rescaling).*

为什么锚定词有帮助？很容易看出，如果 A 是可分离的，那么

行 W 在缩放) 后表现为行 M (。因此，我们只需确定哪些行 M 对应于锚定词。根据我们在第2.3节中的讨论，如果我们对 M、A 和 W 进行缩放，使它们的行和为1，那么 W 的行凸包包含 M 的行。但由于这些行也出现在 M 中，我们可以尝试通过迭代删除不改变其凸包的 M 行来找到 W。

让 M^i 表示 M 的 i 行，让 M^I 表示 $I \subseteq$ 中 I 行的 M 的限制。因此，现在我们可以使用以下简单程序找到锚文本:

Find Anchors [13]

输入：满足定理2.3.12中条件的矩阵$M \in \mathbb{R}^{m \times n}$ 输出：$W = M^I$

删除重复行 设置 $I = [n]$

对于 $i = 1, 2, ..., n$ 如果 $M^i \in \text{conv}(\{M^j | j \in I, j \neq i\})$
，设置 $I \leftarrow I - \{i\}$ 结束

这里在第一步，我们想要删除冗余行。如果两行是彼此的标量倍数，那么其中一行在由 W 的行生成的锥体中，意味着另一行也是，因此我们可以安全地删除其中一行。我们对所有行都这样做，以便在彼此是标量倍数的行等价类中，恰好保留一行。在讨论中，我们不会关注这个技术细节。

尽管如此。

可以看出，删除非锚词的 M 行不会改变剩余行的凸包，因此上述算法终止，得到只包含锚词的集合 I。此外，在终止时

$$\text{conv}(\{M^i | i \in I\}) = \text{conv}(\{M^j\}_j)$$

另一种情况是凸包与开始时相同。因此，被删除的锚定词是多余的，我们完全可以不用它们。

Anchor Words [13]

输入：满足定理2.3.12中条件的矩阵$M \in \mathbb{R}^{n \times m}$

输出：A, W

运行 **Find Anchors** 在 M 上，令 W 为输出

求解非负 A，使其最小化 $\|M - AW\|_F$ (凸规划) 结束

定理的证明直接从 **Find Anchors** 的正确性证明以及以下事实得出：当且仅当存在一个非负的 A (，其行和为 1) 时，conv($\{M^i\}_i$) $\subseteq$ conv($\{W^i\}_i$)。

上述算法如果天真地实现，将会非常慢。相反，上述算法已经有很多改进[33]，[100] [78]，我们将在[12]中描述其中一个。假设我们选择一个

行 M^i 随机。然后很容易看出，离 M^i 最远的行将是一个锚词。

类似地，如果我们找到了一个锚点词，那么离它最远的行将是另一个锚点词，依此类推。这样我们就可以贪婪地找到所有的锚点行，而且这种方法只依赖于成对距离和投影，因此我们可以在运行贪婪算法之前进行降维。这避免了在上面的算法的第一步中完全使用线性规划，而且第二步也可以快速实现，因为它涉及到将一个点投影到一个 $k-1$维单纯形中。

2.4 Topic Models

在这个部分，我们将使用用于生成文档集合的随机模型。这些模型被称为 *topic models*，我们的目标是学习它们的参数。存在多种类型的主题模型，但它们都符合以下抽象框架：

Abstract Topic Model

参数：主题矩阵 $A \in \mathbb{R}^{m \times r}$，在 $\mathbb{R}^r$ 简单形上的分布 μ

对于 $i = 1$ 到 n

样本 W_i 来自 μ

生成由从分布 AW_i 中独立同分布采样得到的 L 个词

结束

此过程生成长度为 L 的 n 个文档，我们的目标是根据观察此模型中的样本推断 A (和 μ)。设 $\widetilde{M}$ 为观察到的词-文档矩阵。我们将使用此符号来区分它与它的期望值

$$\mathbb{E}[\widetilde{M}|W] = M = AW$$

在非负矩阵分解的情况下，我们得到了 M 而不是 $\widetilde{M}$。然而，这些矩阵可能相距甚远！因此，尽管每份文档都被描述为词语上的分布，但我们只对这种分布的部分知识有所了解，即从其中抽取的 L 个样本。我们的目标是设计出在这些具有挑战性的模型中也能证明有效的算法。

现在是一个指出该模型包含许多已研究主题模型作为特殊情况的好时机。所有这些都与μ的不同选择相对应，这是用于生成W列的分布。其中一些最受欢迎的变体包括:

(a) **Pure Topic Model:** 每个文档只涉及一个主题，因此 μ 是在单纯形顶点上的分布，并且 W 的每一列恰好有

一个非零。

(b) **Latent Dirichlet Allocation [36] :** μ 是一个狄利克雷分布。特别是，可以通过从 r (not necessarily identical) 高斯分布中独立采样，然后重新归一化，使得它们的和为1，来生成一个狄利克雷分布的样本。这个主题模型允许文档涉及多个主题，但它的参数通常设置得使其倾向于相对稀疏的向量 W_i。

(c) **Correlated Topic Model [35] :** 某些主题对允许为正相关或负相关，且 μ 被限制为对数正态分布。

(d) **Pachinko Allocation Model [105] :** 这是对LDA的多级泛化，允许存在某些类型的结构化相关性。

在这个部分，我们将使用我们的可分离非负矩阵分解算法，以证明地学习任何（本质上）具有可分离主题矩阵的主题模型的参数。因此，即使存在主题之间的复杂关系，此算法也能正常工作。

The Gram Matrix

在这个子节中，我们将介绍两个矩阵 G 和 R —— 我们将分别称之为格拉姆矩阵和主题共现矩阵。这些矩阵的条目将根据各种事件发生的概率来定义。在整个这一节中，我们始终要牢记以下实验：我们从抽象主题模型中生成一个文档，并让 w_1 和 w_2 表示

随机变量分别对应其第一和第二个单词。考虑到这个实验，我们可以定义格拉姆矩阵：

Definition 2.4.1 *Let G denote the $m \times m$ matrix where*

$$G_{j,j'} = \mathbb{P}[w_1 = j, w_2 = j']$$

此外，对于每个词，我们不仅可以从 AW_i 中采样，还可以从 W_i 中采样，以选择从 A 的哪一列进行采样。此过程仍然生成来自同一分布 AW_i 的随机样本，但每个词 $w_1 = j$ 都会标注其来源的主题，即 $t_1 = i$ (，即我们从哪一列 A 中采样它)。现在我们可以定义主题共现矩阵：

Definition 2.4.2 *Let R denote the $r \times r$ matrix where*

$$R_{i,i'} = \mathbb{P}[t_1 = i, t_2 = i']$$

请注意，我们可以直接从我们的样本中估计 G 的项，但无法直接估计 R 的项。尽管如此，这些矩阵根据以下恒等式相关：

Lemma 2.4.3 $G = ARA^T$

Proof: 我们有

$$\begin{aligned} G_{j,j'} &= \mathbb{P}[w_1 = j, w_2 = j'] = \sum_{i,i'} \mathbb{P}[w_1 = j, w_2 = j' | t_1 = i, t_2 = i'] \mathbb{P}[t_1 = i, t_2 = i'] \\ &= \sum_{i,i'} \mathbb{P}[w_1 = j | t_1 = i] \mathbb{P}[w_2 = j' | t_2 = i'] \mathbb{P}[t_1 = i, t_2 = i'] \\ &= \sum_{i,i'} A_{j,i} A_{j',i'} R_{i,i'} \end{aligned}$$

在倒数第二行成立，因为根据它们的主题，w_1 和 w_2 从 A 的对应列独立采样。这完成了证明。■

关键观察是 $G = A(RA^T)$，其中 A 是可分离的，RA^T 是非负的。因此，如果我们将 G 的行重新归一化，使其总和为1，则锚词将是所有行构成的凸包的极点，我们可以通过我们的可分离非负矩阵分解算法来识别它们。我们能否推断出 A 的其余部分?

Recovery via Bayes Rule

考虑后验分布 $\mathbb{P}[t_1 | w_1 = j]$。这是在您对文档没有任何其他信息的情况下，由主题 $w_1 = j$ 生成的后验分布。后验分布只是 A 的重新归一化，使得行求和为1。然后假设 j 是主题 i 的锚词。我们将使用符号 $j = \pi(i)$。很容易看出

$$\mathbb{P}[t_1 = i' | w_1 = \pi(i)] = \begin{cases} 1, \text{ if } i' = i \\ 0 \text{ else} \end{cases}$$

现在我们可以展开:

$$\begin{aligned}\mathbb{P}[w_1 = j | w_2 = j'] &= \sum_{i'} \mathbb{P}[w_1 = j | w_2 = j', t_2 = i'] \mathbb{P}[t_2 = i' | w_2 = j'] \\ &= \sum_{i'} \mathbb{P}[w_1 = j | t_2 = i'] \mathbb{P}[t_2 = i' | w_2 = j']\end{aligned}$$

在最后一行，我们使用了以下恒等式:

Claim 2.4.4 $\mathbb{P}[w1 = j | w_2 = j', t_2 = i'] = \mathbb{P}[w_1 = j | t_2 = i']$

我们将此命题的证明留给读者作为练习。我们还将使用以下恒等式:

Claim 2.4.5 $\mathbb{P}[w1 = j | t_2 = i'] = \mathbb{P}[w_1 = j | w_2 = \pi(i')]$

Proof:

$$\begin{aligned}\mathbb{P}[w_1 = j | w_2 = \pi(i')] &= \sum_{i''} \mathbb{P}[w_1 = j | w_2 = \pi(i'), t_2 = i''] \mathbb{P}[t_2 = i'' | w_2 = \pi(i')] \\ &= \mathbb{P}[w_1 = j | w_2 = \pi(i'), t_2 = i']\end{aligned}$$

在最后一行成立，因为给定 w_2 是主题 i' 的锚词时，主题 $t_2 = i''$ 的后验分布等于一当且仅当 $i'' = i'$。最后，通过引用命题 2.4.4 进行证明。■

现在我们可以继续:

$$\mathbb{P}[w_1 = j | w_2 = j'] = \sum_{i'} \mathbb{P}[w_1 = j | w_2 = \pi(i')] \underbrace{\mathbb{P}[t_2 = i' | w_2 = j']}_{\text{unknowns}}$$

因此，这是一个关于变量 $\mathbb{P}[w_1 = j|w_2 = \pi(i')]$ 的线性系统，并且不难证明如果 R 具有满秩，则它有唯一解。

最后，通过贝叶斯法则，我们可以计算 A 的项

$$\begin{aligned}\mathbb{P}[w = j|t = i] &= \frac{\mathbb{P}[t = i|w = j]\mathbb{P}[w = j]}{\mathbb{P}[t = i]} \\ &= \frac{\mathbb{P}[t = i|w = j]\mathbb{P}[w = j]}{\sum_{j'} \mathbb{P}[t = i|w = j']\mathbb{P}[w = j']}\end{aligned}$$

将所有内容合并，我们得到以下算法:

Recover [14], [12]

输入：文档-术语矩阵 $M \in \mathbb{R}^{n\times m}$

输出：A, R

计算Gram矩阵G通过**Separable NMF**计算锚点
词求解$\mathbb{P}[t = i|w = j]$从贝叶斯规则计算$\mathbb{P}[$
$w = j|t = i]$

Theorem 2.4.6 *[14] There is a polynomial time algorithm to learn the topic matrix for any separable topic model, provided that R is full-rank.*

Remark 2.4.7 *The running time and sample complexity of this algorithm depend polynomially on $m, n, r, \sigma_{min}(R), p, 1/\epsilon,$* log 1
$/\delta$ where p is a lower bound on the probability of each anchor word, ϵ is the target accuracy and δ is the failure probability.

请注意，此算法适用于短文档，甚至适用于 $L = 2$。

Experimental Results

现在我们有了在可分性条件下的非负矩阵分解和主题建模的可证明算法。但是，$natural$主题模型是否可分或接近可分？考虑以下实验：

(1) **UCI Dataset**：300, 000 纽约时报文章的集合

(2) **MALLET**：一个流行的主题建模工具包

我们在UCI数据集上训练了MALLET，并发现当r =为200时，大约0.9的比例的主题具有近锚词——即$\mathbb{P}[t = i|w = j]$在某些主题上的值至少为0.9的词。事实上，我们给出的算法可以证明在存在一些适度错误的情况下——即与可分性假设的偏差——仍然可以工作。但是，它们能否在如此多的建模错误下工作呢?

然后我们进行了以下额外实验：

(1) 在UCI数据集上运行MALLET。学习一个主题矩阵($r = 200$) (2) 使用A从LDA模型中合成生成一组新文档 (3) 在一组新文档上运行MALLET和我们的算法，并将它们的输出与真实情况进行比较。特别是，计算估计列和真实列之间的最小成本匹配。

这是一个有偏见的实验——有偏见的*against*我们的算法！我们正在比较我们如何更好地找到隐藏的主题（在主题矩阵仅接近可分离的设置中）与MALLET如何更好地找到其自己的输出。而且，有了足够的文档，我们可以更准确地找到它，并且比以前快数百倍！这个新算法使我们能够探索比以前任何时候都大的文档集合。

2.5 Exercises

Problem 2-1: 以下哪些是非负秩的等价定义？对于每个定义，给出证明或反例。

(a) 最小的 r，使得 M 可以表示为 r 个秩为 1、非负矩阵的和 (b) 最小的 r，使得存在 r 个非负向量 $v_1, v_2, ..., v_r$，它们生成的锥包含 M 的所有列 (c) 最大的 r，使得存在 r 列的 M，$M_1, M_2, ..., M_r$ 列，使得集合中的任何列都不包含在由剩余的 $r-1$ 列生成的锥中

Problem 2-2: 设 $M \in \mathbb{R}^{n \times n}$ 其中 $M_{i,j} = (i-j)^2$。证明 rank(M) $\geqslant 3$ 且 rank$^+(M) \geq \leqslant \log_2 n$。*Hint:* 要证明 rank$^+(M)$ 的下界，只需考虑它在何处为零以及它在何处非零。

Problem 2-3: Papadimitriou 等人 [118] 考虑了以下文档模型：$M = AW$，并且 W 的每一列只有一个非零值，每个值的支撑集为

列 A 是不相交的。证明 M 的左奇异向量是经过缩放 A (后的列。可以假设 M 的所有非零奇异值都是不同的。*Hint:* MM^T 在对行和列应用排列 π 后是分块对角形的。

Problem 2-4: 考虑以下算法:

Greedy Anchorwords [13]

Input: matrix $M \in \mathbb{R}^{n \times m}$ satisfying the conditions in Theorem 2.3.12

Output: A, W

Set $S = \emptyset$

For $i = 2$ to r

    Project the rows of M orthogonal to the span of vectors in S

    Add the row with the largest ℓ_2 norm to S

End

让 $M = AW$，其中 A 是可分离的，*rows* 的 M、A 和 W 被归一化，总和为 1。还假设 W 具有满秩。证明 Greedy Anchorwords 找到所有锚词而没有其他。*Hint:* 的 ℓ_2 范数是严格凸的——即对于任何 $x \neq y$ 和 $t \in (0, 1)$，$\|tx + (1 - t)y\|_2 < t\|x\|_2 + (1 - t)\|y\|_2$。

Chapter 3

Tensor Decompositions: Algorithms

在这一章中，我们将研究张量以及我们可以对他们提出各种结构和计算问题。通常，在矩阵上容易解决的问题，在处理张量时可能会变得不适当或NP-困难。与流行观点相反，这并不是收拾行李回家的理由。实际上，我们可以从张量中获得一些从矩阵中无法获得的东西。我们只需小心我们试图解决的问题类型。更确切地说，在这一章中，我们将给出一个具有可证明保证的低秩张量分解算法——它在自然但受限的设置中工作——以及它的一些初步应用，例如因子分析。

3.1 The Rotation Problem

在研究围绕张量的算法问题之前，让我们首先了解它们为什么有用。为此，我们需要引入*factor analysis*的概念，其中使用张量而不是矩阵可以帮助我们绕过一个主要障碍。那么什么是因子分析呢？它是统计学中的一个基本工具，其目标是使用更少的隐藏变量（称为因子）来解释许多变量。但最好通过一个例子来理解它。为什么不从历史例子开始呢？它最初被用于查尔斯·斯皮尔曼的开拓性工作中，他有一个关于智力本质的理论——他相信存在两种基本类型的智力：*mathematical*和*verbal*。我不同意，但我们还是继续吧。

他设计了一个以下实验来检验他的理论：他测量了一千名学生在十个不同测试中的表现，并将他的数据整理成一个1000 $\times$ 10 矩阵 M。他相信学生在某个特定测试中的表现是由与学生和测试有关的某些隐藏变量决定的。想象一下，每个学生都可以用一个二维向量来描述，其中两个坐标分别给出了他的数学和语言智力的数值分数。同样，想象每个测试也可以用一个二维向量来描述，但其中的坐标代表它测试数学和语言推理的程度。Spearman试图找到这组二维向量，每个学生和一个测试对应一个，以便学生的测试表现可以通过他们两个相应向量的内积来给出。

让我们将问题翻译成更方便的语言。我们所要

寻找的是一个特定的分解

$$M = AB^T$$

在 A 的大小为1000 $\times$ 2 和 B 的大小为10 $\times$ 2 验证 Spearman 的理论。问题是，即使存在一个分解 $M = AB^T$，其中 A 的列和 B 的行可以给出一些 $meaningful$ 解释——这将证实 Spearman 的理论——我们如何找到它？可能有其他许多 M 的分解具有相同的内维数，但不是我们正在寻找的因子。为了使这个问题具体化，假设 O 是一个 2 $\times$ 2 正交矩阵。然后我们可以写出

$$M = AB^T = (AO)(O^T B^T)$$

并且我们同样可以找到因式分解 $M = \hat{A}\hat{B}^T$，其中$\hat{A} = AO$ 和$\hat{B} = BO$ 代替。所以即使存在一个有意义的因式分解可以解释我们的数据，也没有保证我们能找到它，而且在一般情况下，我们找到的可能是对其任意内旋转的任意内旋转，这本身也难以解释。这被称为 $rotation\ problem$。这就是我们之前提到的绊脚石，即如果我们使用矩阵技术进行因子分析时遇到的。

这里出了问题的是低秩矩阵分解不是唯一的。让我们具体阐述一下在这个上下文中我们所说的唯一是指什么。假设我们给定一个矩阵 M 并承诺它有一些有意义的低秩分解

$$M = \sum_{i=1}^{r} a^{(i)}(b^{(i)})^T$$

我们的目标是恢复因子 $a^{(i)}$ 和 $b^{(i)}$。问题是我们可以计算

单值分解 $M = U\Sigma V^T$ 并找到另一个低秩分解

$$M = \sum_{i=1}^{r} \sigma_i u^{(i)} (v^{(i)})^T$$

这些可能是两组非常不同的因素，恰好重构了相同的矩阵。实际上，向量 $u^{(i)}$ 必然是正交归一的，因为它们来自奇异值分解，尽管事先没有理由认为我们正在寻找的真正因素 $a^{(i)}$ 也是正交归一的。因此，现在我们可以定性回答我们一开始提出的问题。我们为什么对张量感兴趣？这是因为它们解决了旋转问题，并且它们的分解在比它们的矩阵分解对应物更弱的条件下是唯一的。

3.2 A Primer on Tensors

一个张量可能听起来很神秘，但它只是一组数字。让我们从我们将花费大部分时间的案例开始。一个三阶张量 T 有三个维度，有时分别称为 *rows*、*columns* 和 *tubes*。如果 T 的大小是 $n_1 \times n_2 \times n_3$，那么标准表示法是 $T_{i,j,k}$ 指的是 T 中第 i 行、第 j 列和第 k 管的数字。现在，一个矩阵只是一个二阶张量，因为它是一组由两个索引索引的数字集合。当然，你可以考虑任何阶数的张量。

我们可以从许多不同的角度来思考张量，并且所有这些观点在本章的不同点都将是有用的。也许将三阶张量 T 看作仅仅是 n_3 矩阵的集合，每个

大小为 $n_1 \times n_2$，彼此堆叠在一起。在进一步讨论之前，我们应该定义张量秩的概念。这将使我们能够探讨张量不仅仅是矩阵的集合，以及这些矩阵何时以及如何相互关联。

Definition 3.2.1 *A rank one, third-order tensor T is the tensor product of three vectors u, v and w, and its entries are*

$$T_{i,j,k} = u_i v_j w_k$$

Thus if the dimensions of u, v and w are n_1, n_2 and n_3 respectively, T is of size $n_1 \times n_2 \times n_3$. Moreover, we will often use the following shorthand

$$T = u \otimes v \otimes w$$

我们现在可以定义张量的秩:

Definition 3.2.2 *The rank of a third-order tensor T is the smallest integer r so that we can write*

$$T = \sum_{i=1}^{r} u^{(i)} \otimes v^{(i)} \otimes w^{(i)}$$

回忆，矩阵 M 的秩是满足 M 可以表示为 r 个秩一矩阵之和的最小整数 r。矩阵秩的美丽在于它有多少个等价定义。我们上面的是矩阵秩的许多定义之一对张量的自然推广。上述分解通常称为 CANDECOMP/PARFAC 分解。

现在我们手头有了秩的定义，让我们了解一个低秩张量不是*just*任意一组低秩矩阵。用$T_{\cdot,\cdot,k}$表示对应于张量k^{th}切片的$n_1 \times n_2$矩阵。

Claim 3.2.3 *Consider a rank r tensor*

$$T = \sum_{i=1}^{r} u^{(i)} \otimes v^{(i)} \otimes w^{(i)}$$

Then for all $1 \leq k \leq n3$,

$$colspan(T_{\cdot,\cdot,k}) \subseteq span(\{u^{(i)}\}_i)$$

and moreover

$$rowspan(T_{\cdot,\cdot,k}) \subseteq span(\{v^{(i)}\}_i)$$

我们将证明留给读者作为练习。实际上，这个陈述告诉我们为什么不是每个低秩矩阵的堆叠都产生一个低秩张量。确实，如果我们取一个低秩张量并观察其n_3不同的切片，我们得到的矩阵维度为$n_1 \times n_2$，秩最多为r。但我们知道的不止这些。它们的每一个列空间都包含在向量$u^{(i)}$的张成中。同样，它们的行空间包含在向量$v^{(i)}$的张成中。

直观上，旋转问题源于矩阵只是向量 *view* 和 $\{u^{(i)}\}_i$、$\{v^{(i)}\}_i$ 中的一个。但张量通过其每个切片给我们多个 *views*，这有助于我们解决不确定性。如果这还不完全清楚，那没关系。等你理解了 Jennrich 算法后再回来思考这个问题。

The Trouble with Tensors

在继续之前，消除你可能有的任何关于与张量工作将是与矩阵工作简单推广的神话是很重要的。那么，与张量工作有什么微妙之处呢？首先，线性代数之所以优雅和吸引人，在于诸如矩阵的秩M这样的概念可以接受多种等价定义。当我们定义张量的秩时，我们小心地说我们所做的是取矩阵秩的定义one，并将其自然推广到张量。但如果我们对矩阵的秩采用不同的定义，并以自然的方式推广它呢？我们会得到对张量相同的秩概念吗？通常不会！

让我们试试。与其将矩阵 M 的秩定义为得到 M 所需添加的最小一秩矩阵的数量，我们本可以通过其列/行空间的维度来定义秩。接下来的断言只是说我们会得到相同的秩概念。

Claim 3.2.4 *The rank of a matrix M is equal to the dimension of its column/row space. More precisely,*

$$rank(M) = dim(colspan(M)) = dim(rowspan(M))$$

这个关系对张量成立吗？差得远！作为一个简单的例子，让我们设$n_1 = k^2$，$n_2 = k$和$n_3 = k$。然后如果我们取T的n_1列作为$k^2 \times k^2$单位矩阵的列，我们知道T的$n_2 n_3$列都是线性无关的，并且维度为k^2。但是T的$n_1 n_3$行在最多k维的空间中，因此对于张量，维度

行跨度不一定等于列/管跨度维度。

事物只会越来越糟。张量秩有一些棘手的细微差别。首先，领域很重要。假设 T 是实值的。我们定义秩为 r 的最小值，这样我们可以将 T 写成 r 个秩一张量的和。但是，我们应该允许这些张量具有复数值，还是只有实数值？实际上，这种 can 改变了秩，如下面的例子所示。

考虑以下2 × 2 × 2 张量：

$$T = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}; \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$$

在第一个2 ×^2 矩阵是张量的第一个切片，第二个2 ×^2 矩阵是第二个切片。不难证明$\text{rank}_{\mathbb{R}}(T) \geq$ = 3。但很容易检查

$$T = \frac{1}{2}\left(\begin{bmatrix} 1 \\ -i \end{bmatrix} \otimes \begin{bmatrix} 1 \\ i \end{bmatrix} \otimes \begin{bmatrix} 1 \\ -i \end{bmatrix} + \begin{bmatrix} 1 \\ i \end{bmatrix} \otimes \begin{bmatrix} 1 \\ -i \end{bmatrix} \otimes \begin{bmatrix} 1 \\ i \end{bmatrix}\right)$$

所以即使 T 是实值的，如果我们允许使用复数，它可以写成 $fewer$ 个秩一张量的和。对于矩阵来说，这个问题从未出现过。如果 M 是实值的，并且存在一种方法将其写成 r 个秩一矩阵的和，这些矩阵的元素可能是复值的，那么总存在一种方法将其写成至多 r 个秩一矩阵的和，所有这些矩阵的元素都是实值的。现在我们面对的是秩与域相关的对象，这似乎是一个愉快的意外。

另一个令人担忧的问题是存在秩为3的张量，但可以用秩为2的张量任意良好地逼近。这导致我们定义边界秩：

Definition 3.2.5 *The* 边秩 *of a tensor T is the minimum r such that for any $\epsilon > 0$ there is a rank r tensor that is entry-wise ϵ-close to T.*

对于矩阵，秩和边界秩是相同的！如果我们固定一个秩为 r 的矩阵 M，那么我们通过一个秩为 $r' < r$ 的矩阵来逼近它的极限（取决于 M）是有限的。这一点可以从截断奇异值分解在低秩逼近中的最优性中得出。但对于张量，秩和边界秩确实可以不同，正如我们的最终示例所说明的那样。

考虑以下$2 \times 2 \times 2$ 张量：

$$T = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}; \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$$

很难证明$\text{rank}_{\mathbb{R}}(T) \geq = 3$。然而，它可以通过以下方案接受任意好的rank two近似。设

$$S_n = \begin{bmatrix} n & 1 \\ 1 & \frac{1}{n} \end{bmatrix}; \begin{bmatrix} 1 & \frac{1}{n} \\ \frac{1}{n} & \frac{1}{n^2} \end{bmatrix} \text{ and } R_n = \begin{bmatrix} n & 0 \\ 0 & 0 \end{bmatrix}; \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}.$$

两个 S_n 和 R_n 都是秩一，因此 $S_n - R_n$ 的秩最多为二。但请注意，$S_n - R_n$ 在元素上与 T 的距离为 $1/n$，随着 n 的增加，我们得到对 T 的任意好的近似。因此，尽管 T 的秩为 3，但其边界秩最多为 2。您可以看到这个例子利用了越来越大的抵消。它还

显示最佳低秩逼近的项的幅度不能作为 T 中项的幅度的函数进行界定。

矩阵的一个有用性质是，可以直接从其最佳秩 $k+1$ 近似得到 M 的最佳秩 k 近似。更精确地说，假设 $B^{(k)}$ 和 $B^{(k+1)}$ 分别是关于（例如）Frobenius 范数下 M 的最佳秩 k 和秩 $k+1$ 近似。然后我们可以得到 $B^{(k)}$ 作为 $B^{(k+1)}$ 的最佳秩 k 近似。然而，对于张量，T 的最佳秩 k 和秩 $k+1$ 近似不必共享任何 any 公共的秩 1 项。张量的最佳秩 k 近似难以处理。你必须担心它的域。你不能用输入来界定其元素的幅度。并且当改变 k 时，它会以复杂的方式变化。

对我来说，所有这些问题的根源最严重的问题是计算复杂性。当然，张量的秩不等于其列空间的维度。前者是NP-难（根据Hastad [85] 的结果）而后者容易计算。你必须小心处理张量。实际上，计算复杂性是一个如此普遍的问题，许多在矩阵上容易计算的问题在张量上却变成了NP-难，以至于Hillar和Lim [86] 的著名论文的标题总结了这一点：

"Most Tensor Problems are Hard"

为了支持这一点，Hillar和Lim [86]证明了其他一系列问题，如找到最佳低秩近似、计算谱范数以及判断张量是否为非负定形的难度也是NP-hard。如果这个部分让你有些悲观，请记住，我试图做的只是为你们搭建舞台，以便你们能像应该的那样兴奋，因为实际上确实有一些我们

可以用张量!

3.3 Jennrich's Algorithm

在这个部分，我们将介绍一种在自然但受限的设置中计算最小秩分解的算法。这个算法被称为Jennrich算法。有趣的是，它已经被多次重新发现（原因我们将在稍后进行推测），据我们所知，它首次出现在Harshman [84] 的工作论文中，作者将其归功于Robert Jennrich博士。

在以下内容中，我们将假设我们被给定一个张量 T，我们将假设它具有以下形式

$$T = \sum_{i=1}^{r} u^{(i)} \otimes v^{(i)} \otimes w^{(i)}$$

我们将把因子 $u^{(i)}$、$v^{(i)}$ 和 $w^{(i)}$ 称为 $hidden$ 因子，以强调我们不知道它们，但希望找到它们。这里我们需要小心。我们所说的“找到它们”是什么意思? 有一些歧义是我们永远无法解决的。我们只能希望恢复到任意重排（总和）和某些缩放，这些缩放不会改变秩为 1 的张量本身。这促使我们以下定义，它考虑了这些问题：

Definition 3.3.1 *We say that two sets of factors*

$$\left\{(u^{(i)}, v^{(i)}, w^{(i)})\right\}_{i=1}^{r} \text{ and } \left\{(\hat{u}^{(i)}, \hat{v}^{(i)}, \hat{w}^{(i)})\right\}_{i=1}^{r}$$

are equivalent if there is a permutation π $[r] \to [r]$ *such that for all* i

$$u^{(i)} \otimes v^{(i)} \otimes w^{(i)} = \hat{u}^{(\pi(i))} \otimes \hat{v}^{(\pi(i))} \otimes \hat{w}^{(\pi(i))}$$

重要的是，两组等价的因素产生两个分解

$$T = \sum_{i=1}^{r} u^{(i)} \otimes v^{(i)} \otimes w^{(i)} = \sum_{i=1}^{r} \hat{u}^{(i)} \otimes \hat{v}^{(i)} \otimes \hat{w}^{(i)}$$

具有相同和的一组秩一张量的集合。

本节的主要问题是：给定 T，我们能否高效地找到一个与隐藏因子等价的因素集？我们将陈述并证明一个比Jennrich算法更通用的版本，遵循Leurgans、Ross和Abel [103]的方法。

Theorem 3.3.2 *[84], [103] Suppose we are given a tensor of the form*

$$T = \sum_{i=1}^{r} u^{(i)} \otimes v^{(i)} \otimes w^{(i)}$$

where the following conditions are met:

(1) the vectors $\{u^{(i)}\}_i$ *are linearly independent,*

(2) the vectors $\{v^{(i)}\}_i$ *are linearly independent and*

(3) every pair of vectors in $\{w^{(i)}\}_i$ *are linearly independent*

Then there is an efficient algorithm to find a decomposition

$$T = \sum_{i=1}^{r} \hat{u}^{(i)} \otimes \hat{v}^{(i)} \otimes \hat{w}^{(i)}$$

and moreover the factors $(u^{(i)}, v^{(i)}, w^{(i)})$ *and* $(\hat{u}^{(i)}, \hat{v}^{(i)}, \hat{w}^{(i)})$ *are equivalent.*

原始结果由Jennrich [84] 提出，表述为一个 $uniqueness$ 定理，即在上述因素 $u^{(i)}$、$v^{(i)}$ 和 $w^{(i)}$ 的条件下，T 至多分解为 r 个秩一张量必须使用一个等价的因素集。碰巧的是，Jennrich证明这个唯一性定理的方法是给出一个寻找分解的算法，尽管在论文中从未这样表述。有趣的是，这似乎是结果被遗忘的主要原因。随后的许多文献引用了Kruskal的更强的唯一性定理，其证明是非构造性的，并且似乎忘记了Jennrich的较弱唯一性定理附带了一个算法。让我们以此为警示：如果你不仅证明了某个数学事实，而且你的论证可以轻易地得出一个算法，那么就说明这一点！

Jennrich's Algorithm [84]

输入：满足定理3.3.2中条件的张量 $T \in \mathbb{R}^{m \times n \times p}$

输出：因素 $\{u_i\}_i, \{v_i\}_i$ 和 $\{w_i\}_i$

选择 $a, b \in \mathbb{S}^{p-1}$ 均匀随机；设置

$$T^{(a)} = \sum_{i=1}^{p} a_i T_{\cdot,\cdot,i} \text{ and } T^{(b)} = \sum_{i=1}^{p} b_i T_{\cdot,\cdot,i}$$

计算 $T^{(a)}(T^{(b)})^+$ 和 $((T^{(a)})^+ T^{(b)})^T$ 的特征分解

U 和 V 是对应非零特征值的特征向量

配对 $u^{(i)}$ 和 $v^{(i)}$ 当且仅当它们的特征值互为倒数

求解 $w^{(i)}$ 在 $T = \sum_{i=1}^{r} u^{(i)} \otimes v^{(i)} \otimes w^{(i)}$ 中

结束

回忆一下，$T_{\cdot,\cdot,i}$ 表示通过 T 的 i^{th} 矩阵切片。因此，$T^{(a)}$ 只是 T 的矩阵切片的加权求和，每个切片的权重为 a_i。

分析的第一步是将 $T^{(a)}$ 和 $T^{(b)}$ 用隐藏因子表示。设 U 和 V 分别为大小为 $m \times r$ 和 $n \times r$ 的矩阵，其列分别为 $u^{(i)}$ 和 $v^{(i)}$。设 $D^{(a)}$ 和 $D^{(b)}$ 为 $r \times r$ 对角矩阵，其元素分别为 $\langle w^{(i)}, a \rangle$ 和 $\langle w^{(i)}, b \rangle$。然后

Lemma 3.3.3 $T^{(a)} = UD^{(a)}V^T$ *and* $T^{(b)} = UD^{(b)}V^T$

Proof: 由于从 T 计算出 $T^{(a)}$ 的操作是线性的，我们可以将其应用于 T 的低秩分解中的每个秩一张量。容易看出，如果我们给定秩一张量 $u \otimes v \otimes w$，那么对矩阵切片进行加权求和的效果，其中 i^{th} 切片由 a_i 加权，就是得到矩阵 $\langle w, a\rangle u \otimes v$ 。

因此，根据线性，我们有

$$T^{(a)} = \sum_{i=1}^{r} \langle w^{(i)}, a\rangle u^{(i)} \otimes v^{(i)}$$

这给出了词元的第一个部分。第二个部分类似地得出，其中 a 被替换为 b。■

结果显示，我们现在可以通过广义特征分解来恢复 U 的列和 V 的列。让我们做一个思想实验。如果我们被给定了形式为 $M = UDU^{-1}$ 的矩阵 M，其中对角矩阵 D 的条目是不同的且非零，那么 U 的列将是特征向量，但它们不一定是单位向量。由于 D 的条目是不同的，M 的特征分解是唯一的，这意味着我们可以恢复 U (的列，直到缩放)，作为 M 的特征向量。

现在如果我们被给出两个形式为 $A = UD^{(a)}V^T$ 和 $B = UD^{(b)}V^T$ 的矩阵，那么如果 $D^{(a)}(D^{(b)})^{-1}$ 的项是不同的且非零的，我们就可以通过 U 和 V (的特征分解再次恢复其列，直到缩放)。

$$AB^{-1} = UD^{(a)}(D^{(b)})^{-1}U^{-1} \text{ and } (A^{-1}B)^T = VD^{(b)}(D^{(a)})^{-1}V^{-1}$$

分别。实际上，我们不必形成上述矩阵，而是可以寻找所有满足 $Av = \lambda_v Bv$ 的向量 v，这被称为广义特征分解。无论如何，这是以下引理背后的主要思想，尽管我们需要小心，因为在我们的设置中，矩阵 U 和 V 不一定是方阵，更不用说可逆矩阵了。

Lemma 3.3.4 *Almost surely, the columns of U and V are the unique eigenvectors corresponding to non-zero eigenvalues of $T^{(a)}(T^{(b)})^+$ and $((T^{(a)})^+T^{(b)})^T$ respectively. Moreover the eigenvalue corresponding to $u^{(i)}$ is the reciprocal of the eigenvalue corresponding to $v^{(i)}$.*

Proof: 我们可以使用引理3.3.3中关于$T^{(a)}$和$T^{(b)}$的公式来计算

$$T^{(a)}(T^{(b)})^+ = UD^{(a)}(D^{(b)})^+U^+$$

$D^{(a)}(D^{(b)})^+$ 的元素为 $\langle w^{(i)}, a\rangle / \langle w^{(i)}, b\rangle$。然后因为 $\{w^{(i)}\}_i$ 中每对向量都是线性无关的，所以我们几乎可以肯定，在选择 a 和 b 时，$D^{(a)}(D^{(b)})^+$ 的对角线上的元素都将是非零且不同的。

现在回到上面的公式 $T^{(a)}(T^{(b)})^+$，我们看到它是一个特征分解，并且更进一步，非零特征值是不同的。因此，U 的列是具有非零特征值的 $T^{(a)}(T^{(b)})^+$ 的唯一特征向量，与 $u^{(i)}$ 对应的特征值是 $\langle w^{(i)}, a\rangle / \langle w^{(i)}, b\rangle$。一个相同的论据表明，$V$的列是以下矩阵的唯一特征向量

$$((T^{(a)})^+T^{(b)})^T = VD^{(b)}(D^{(a)})^+V^+$$

具有非零特征值。通过检查，我们发现与 $v^{(i)}$ 对应的特征值是 $\langle w^{(i)}, b\rangle / \langle w^{(i)}, a\rangle$，这完成了引理的证明。■

现在来完成定理的证明，注意到我们只恢复了 U 的列和 V 的列，直到缩放——即对于每一列，我们恢复了相应的单位向量。我们将用缺失的因子 $w^{(i)}$ 将这个缩放因子推进去。因此，算法的最后一步中的线性系统显然有解，剩下要证明的是这是唯一的解。

Lemma 3.3.5 *The matrices* $\left\{u^{(i)}(v^{(i)})^T\right\}_{i=1}^{r}$ *are linearly independent.*

Proof: 假设（为了矛盾的目的）存在一组系数，其中不全为零，其中

$$\sum_{i=1}^{r} \alpha_i u^{(i)} (v^{(i)})^T = 0$$

假设（不失一般性）$\alpha_1 \neq 0$。因为根据假设，向量 $\{v^{(i)}\}_i$ 线性无关，所以我们有一个向量 a 满足 $\langle v^{(1)}, a\rangle \neq 0$ 但与所有其他 $v^{(i)}$ 垂直。现在如果我们用 a 右乘上述恒等式，我们得到

$$\alpha_1 \langle v^{(1)}, a\rangle u^{(1)} = 0$$

这是矛盾的，因为左边是非零的。■

这立即意味着在 $w^{(i)}$ 上的线性系统具有唯一解。我们可以将线性系统写成 $mn \times r$ 矩阵，每个列代表一个矩阵 $u^{(i)}(v^{(i)})^T$，但以向量形式，乘以一个未知的 $r \times p$ 矩阵

其列表示向量 $w^{(i)}$。这两个矩阵的乘积被限制等于一个 $mn \times p$ 矩阵，其列表示通过张量 T 的每个 p 矩阵切片，但再次以向量形式。这完成了定理3.3.2的证明。

如果您想提出一个合适的问题，请注意，Jennrich算法中的条件只有在 $r \leq$ m in(n_1, n_2) 时才能成立，因为我们需要向量 $\{u^{(i)}\}_i$ 和 $\{v^{(i)}\}_i$ 是线性无关的。这被称为欠完备情况，因为秩被张量的最大维度所限制。当 r 大于 n_1, n_2 或 n_3 中的任何一个时，我们知道 T 的分解通常是唯一的。但是，对于分解通用的过完备三阶张量，是否有算法呢？即使 $r =$ 1.1 max(n_1, n_2, n_3)，这个问题仍然是开放的。

3.4 Perturbation Bounds

本节是良药。到目前为止，我们有一个算法（Jennrich算法），在因素的一些自然条件下分解三阶张量 T，但在假设我们知道 T $exactly$ 的情况下。在我们的应用中，这远远不够。我们需要处理噪声。本节的目标是回答以下问题：如果我们给定 $\widetilde{T} = T + E$ 而不是（可以将 E 视为表示采样噪声），我们如何近似隐藏因素?

我们的算法不会改变。我们仍然会使用Jennrich的算法。相反，在本节中，我们想要追踪错误是如何传播的。我们想要给出我们近似隐藏因子的定量界限，我们给出的界限将取决于E和T的性质。Jennrich的主要步骤是

算法是计算特征分解。自然地，这就是我们将花费大部分时间的地方——理解特征分解何时稳定。从这一点出发，我们很容易就能看出Jennrich算法在噪声存在时何时以及为什么有效。

Prerequisites for Perturbation Bounds

现在让我们更精确一些。我们感兴趣的主要问题是以下内容：

Question 2 *If $M = UDU^{-1}$ is diagonalizable and we are given $\widetilde{M} = M + E$, how well can we estimate U?*

自然的事情是计算一个对角化 $\widetilde{M}$ 的矩阵——即$\widetilde{U}$，其中 $\widetilde{M} = \widetilde{U}\widetilde{D}\widetilde{U}^{-1}$ ——并量化 $\widetilde{U}$ 作为 U 的估计有多好。但在我们深入之前，进行一个思想实验是好的。

在某些情况下，根本无法说 U 和$\widetilde{U}$ 是接近的。例如，如果 M 有两个非常接近的特征值，那么扰动 E 在原则上可能会将两个特征向量折叠成一个二维特征空间，我们就永远无法估计 U 的列。这意味着我们的扰动界限将不得不依赖于 M 中任意一对特征值之间的最小间隔。

就像这样，我们还可以进行一个思想实验，它告诉我们M的另一个属性，这个属性必须进入我们的扰动界限。但在我们到达那里之前，让我们在一个更简单的设置中理解这个问题。这带我们到一个数值线性代数的重要概念。

Definition 3.4.1 *The* 条件数 *of a matrix U is defined as*

$$\kappa(U) = \frac{\sigma_{\max}(U)}{\sigma_{\min}(U)}$$

where $\sigma_{\max}(U)$ and $\sigma_{\min}(U)$ are the maximum and minimum singular values of U, respectively.

条件数捕捉了解决线性方程组时误差放大的情况。让我们更精确一点：考虑求解 x 在 $Mx = b$ 中的问题。假设我们恰好给出了 M，但我们只知道 b 的估计值 $\widetilde{b} = b + e$。我们如何近似 x?

Question 3 *If we obtain a solution $\widetilde{x}$ that satisfies $M\widetilde{x} = \widetilde{b}$, how close is $\widetilde{x}$ to x?*

We have $\widetilde{x} = M^{-1}\widetilde{b} = x + M^{-1}e = x + M^{-1}(\widetilde{b} - b)$. So

$$\|x - \widetilde{x}\| \leq \frac{1}{\sigma_{\min}(M)}\|b - \widetilde{b}\|.$$

自 $Mx = b$ 以来，我们也有 $\|b\| \leq \sigma_{\max}(M)\|x\|$。因此，

$$\frac{\|x - \widetilde{x}\|}{\|x\|} \leq \frac{\sigma_{\max}(M)}{\sigma_{\min}(M)}\frac{\|b - \widetilde{b}\|}{\|b\|} = \kappa(M)\frac{\|b - \widetilde{b}\|}{\|b\|}.$$

$\|b - \widetilde{b}\|/\|b\|$这个术语通常被称为*relative error*，是数值线性代数中衡量接近度的流行距离。上述讨论告诉我们，条件数控制解线性系统时的相对误差。

现在让我们将这个内容与我们之前的讨论联系起来。结果发现，我们关于特征分解的扰动界限也将取决于条件

数字 U。直观上，这是因为给定 U 和 U^{-1}，求解 M 的特征值就像解决一个依赖于 U 和 U^{-1} 的线性系统。这可以更加精确，但我们在这里不会这样做。

Gershgorin's Disk Theorem and Distinct Eigenvalues

现在我们了解了哪些M的性质应该进入我们的扰动界限，我们可以继续证明它们。我们需要回答的第一个问题是：$\widetilde{M}$是否可对角化？我们的方法将是证明如果M有相异的特征值且E足够小，那么$\widetilde{M}$也有相异的特征值。我们证明中的主要工具是来自数值线性代数的一个有用事实，称为Gershgorin圆盘定理：

Theorem 3.4.2 *The eigenvalues of an $n \times n$ matrix M are all contained in the following union of disks in the complex plane:*

$$\bigcup_{i=1}^{n} D(M_{ii}, R_i)$$

where $D(a, b) := \{x \mid \|x - a\| \leq b\} \subseteq \mathbb{C}$ and $R_i = \sum_{j \neq i} |M_{ij}|$.

在特殊情况下考虑这个定理是有用的。如果 $M = I+E$，其中 I 是单位矩阵，E 是只有小元素的扰动，Gershgorin 圆盘定理告诉我们直观明显的事实，即 M 的特征值都接近于 1。定理中的半径给出了它们接近 1 的定量界限。现在来证明：

Proof: 设(x, λ)为一个特征向量-特征值对（注意，即使这样也是有效的）

当 M 不可对角化时。令 i 表示 x 的坐标中绝对值最大的。那么 $Mx = \lambda x$ 给出 $\sum_j M_{ij}x_j = \lambda x_i$。所以 $\sum_{j\neq i} M_{ij}x_j = \lambda x_i - M_{ii}x_i$。我们得出结论：

$$|\lambda - M_{ii}| = \left|\sum_{j\neq i} M_{ij}\frac{x_j}{x_i}\right| \leq \sum_{j\neq i} |M_{ij}| = R_i.$$

因此 $\lambda \in D(M_{ii}, R_i)$。■

现在我们可以回到证明 $\widetilde{M}$ 可对角化的任务。这个想法很简单，来源于对单个表达式的消化。考虑

$$U^{-1}\widetilde{M}U = U^{-1}(M+E)U = D + U^{-1}EU.$$

这个表达式告诉我们什么？右侧是一个对角矩阵的扰动，因此我们可以使用Gershgorin圆盘定理来说明其特征值接近于D的特征值。现在，因为左乘以U^{-1}和右乘以U是一种相似变换，这反过来又告诉我们关于$\widetilde{M}$的特征值。

让我们将此计划付诸实施，并应用Gershgorin圆盘定理来理解 $\widetilde{D} = D + U^{-1}EU$的特征值。首先，我们可以如下界定$\widetilde{E} = U^{-1}EU$的元素的大小。令$\|A\|_\infty$表示矩阵的最大范数，即$A$中任何元素的绝对值中的最大值。

Lemma 3.4.3 $\|\widetilde{E}\|_\infty \leq \kappa(U)\|E\|$

Proof: 对于任意的 i 和 j，我们可以将 $\widetilde{E}_{i,j}$ 视为 U^{-1} 的 i^{th} 行和 U 的 j^{th} 列在 E 上的二次型。现在 U 的 j^{th} 列的欧几里得范数至多为 $\sigma_{max}(U)$，同样地，U^{-1} 的 i^{th} 行的欧几里得范数至多为 $\sigma_{max}(U^{-1}) = 1/\sigma_{min}(U)$。共同作用，这给出了所需的界限。■

现在让我们证明在适当的条件下，$\widetilde{M}$的特征值是不同的。设$R =$为max $_i \sum_j |\widetilde{E}_{i,j}|$，设$\delta =$为min$_{i\neq j} |D_{i,i} - D_{j,j}|$是$D$特征值的最小间隔。

Lemma 3.4.4 *If $R < \delta/2$ then the eigenvalues of $\widetilde{M}$ are distinct.*

Proof: 首先，我们使用Gershgorin圆盘定理得出$\widetilde{D}$的特征值包含在互不重叠的圆盘中，每个圆盘对应一行。这里有一个小技术问题，Gershgorin圆盘定理使用的是行中所有元素的绝对值之和作为半径，除了对角线元素。但我们将其留作练习，以检查计算是否仍然有效。

实际上我们还没有完成[1]。即使Gershgorin圆盘定理意味着存在不相交的圆盘（每个圆盘对应一行），包含$\widetilde{D}$的特征值，我们如何知道没有圆盘包含超过一个特征值，以及没有圆盘包含特征值？事实证明，矩阵的特征值是项的连续函数，因此当我们绘制路径时

$$\gamma(t) = (1-t)D + t(\widetilde{D})$$

从 D 到$\widetilde{D}$，随着 t 从零到一的变化，Gershgorin圆盘定理中的圆盘总是不相交的，并且没有特征值可以从一个圆盘跳到另一个圆盘。因此，在$\widetilde{D}$ 我们知道每个圆盘中确实恰好有一个特征值，并且由于圆盘不相交，我们得到$\widetilde{D}$ 的特征值是互不相同的，正如所期望的那样。当然，$\widetilde{D}$ 和 $\widetilde{M}$ 的特征值是相同的，因为它们通过相似性相关联

转换。■

[1]Thanks to Santosh Vempala for pointing out this gap in an earlier version of this book. See also [79].

Comparing the Eigendecompositions

我们现在知道 $\widetilde{M}$ 有不同的特征值，所以我们最终可以写出 $\widetilde{M} = \widetilde{U}\widetilde{D}\widetilde{U}^{-1}$，因为 $\widetilde{M}$ 是对角化的。让我们转向我们的最后一步。M 的特征值与 $\widetilde{M}$ 的特征值之间存在自然对应关系，因为前一小节中的证明告诉我们，存在一组不相交的圆盘，每个圆盘恰好包含 M 的一个特征值和 $\widetilde{M}$ 的一个特征值。因此，让我们对 $\widetilde{M}$ 的特征向量进行排列，使我们的符号更简单。事实上，为什么不更进一步简化呢。让我们假设（不失一般性）所有特征向量都是单位向量。

现在假设我们给定 $(\widetilde{u}_i, \widetilde{\lambda}_i)$ 和 (u_i, λ_i)，它们是对应的特征向量 - 特征值对分别为 $\widetilde{M}$ 和 M。设 $\sum_j c_j u_j = \widetilde{u}_i.$，我们知道存在一组 c_j 的选择使得这个表达式成立，因为 u_j 是一组基。我们想要证明的是，在这个表达式中，对于所有 $j \neq i$，c_j 都很小。这将会意味着 u_i 和 $\widetilde{u}_i$ 非常接近。

Lemma 3.4.5 *For any $j \neq i$ we have*

$$|c_j| \leq \frac{\|E\|}{\sigma_{min}(U)(\delta - R)}.$$

Proof: 我们将通过操作表达式 $\sum_j c_j u_j = \widetilde{u}_i.$ 来得到这个结果。首先，将方程的两边乘以 $\widetilde{M}$ 并利用 $\{u_i\}_i$ 是 M 的特征向量以及 $\{\widetilde{u}_i\}_i$ 是 $\widetilde{M}$ 的特征向量的性质，我们得到

$$\sum_j c_j \lambda_j u_j + E\widetilde{u}_i = \widetilde{\lambda}_i \widetilde{u}_i$$

哪个重新排列项得到表达式 $\sum_j c_j(\lambda_j - \widetilde{\lambda}_i)u_j = -E\widetilde{u}_i$.

现在我们想要做的是只挑选出左边的一个系数，并使用右边来限制它。为了做到这一点，让 w_j^T 成为 U^{-1} 的 j 行，并将上述表达式的两边都乘以这个向量，我们得到

$$c_j(\lambda_j - \widetilde{\lambda}_i) = -w_j^T E\widetilde{u}_i.$$

现在让我们将这个表达式中的项进行界定。首先，对于任意的 $i \neq j$，我们使用 Gershgorin 圆盘定理得到 $|\lambda_j - \widetilde{\lambda}_i| \geq |\lambda_j - \lambda_i| - R \geq \delta - R$。其次，根据假设 $\widetilde{u}_i$ 是一个单位向量，并且 $\|w_j\| \leq 1/\sigma_{min}(U)$。利用这些界定并重新排列项，现在可以证明引理。■

三个我们已证明的词元可以组合起来，给出关于U与$\widetilde{U}$之间接近程度的定量界限，这是我们一开始的目标。

Theorem 3.4.6 *Let M be an $n \times n$ matrix with eigendecomposition $M = UDU^{-1}$. Let $\widetilde{M} = M + E$. Finally let*

$$\delta = \min_{i \neq j} |D_{i,i} - D_{j,j}|$$

i.e. the minimum separation of eigenvalues of M.

(1) If $\kappa(U)\|E\|n < \frac{\delta}{2}$ then $\widetilde{M}$ is diagonalizable.

(2) Moreover if $\widetilde{M} = \widetilde{U}\widetilde{D}\widetilde{U}^{-}1$ then there is a permutation $\pi : [n] \to [n]$ such that for all i

$$\|u_i - \widetilde{u}_{\pi(i)}\| \leq \frac{2\|E\|n}{\sigma_{min}(U)(\delta - \kappa(U)\|E\|n)}$$

where $\{u_i\}_i$ *are the columns of* U *and* $\{\widetilde{u}_i\}_i$ *are the columns of* $\widetilde{U}$.

Proof: 定理的第一部分通过结合引理3.4.3和引理3.4.4得出。对于定理的第二部分，我们固定 i 并令 P 为 u_i 的正交补的投影。然后利用初等几何和特征向量都是单位向量的事实，我们有

$$\|u_i - \widetilde{u}_{\pi(i)}\| \leq 2\|P\widetilde{u}_{\pi(i)}\|.$$

此外，我们可以将右侧边界限定为

$$\|P\widetilde{u}_{\pi(i)}\| = \|\sum_{j\neq i} c_j P u_j\| \leq \sum_{j\neq i} |c_j|.$$

引理3.4.5提供了系数c_j的界限，从而完成了定理的证明。■

您在早期就被警告过，这个界限会很混乱！它也绝对没有优化。但您应该吸取的定性推论是我们所追求的：如果 $\|E\| \leq \text{poly}(1/n, \sigma_{min}(U), 1/\sigma_{max}(U), \delta)$（即如果采样噪声与矩阵维度相比足够小，那么 U 和最小分离的条件数）那么 U 和 $\widetilde{U}$ 是接近的。

Back to Tensor Decompositions

现在让我们回到Jennrich的算法。我们已经给出了足够多的混乱边界，对我来说已经足够了。所以从现在开始，让我们作弊，并使用以下方法隐藏混乱边界

符号：设

$$A \xrightarrow{E\to 0} B$$

表示当 E 趋向于零时，A 以逆多项式速率收敛到 B。我们将使用此符号作为占位符。每次您看到它时，都应该想到您可以通过代数运算来找出 A 与 B 在 E 以及我们沿途收集的各种其他因素方面的接近程度。

使用这种记号，我们想要做的是 $qualitatively$ 跟踪 Jennrich 算法中错误的传播。如果我们让$\widetilde{T} = T + E$ 那么$\widetilde{T} \xrightarrow{E\to 0} T$和$\widetilde{T}^{(a)} \xrightarrow{E\to 0} T^{(a)}$ 其中$\widetilde{T}^{(a)} = \sum_i a_i \widetilde{T}_{\cdot,\cdot,i}$。我们将其留作读者的练习，检查是否存在自然条件，其中

$$(\widetilde{T}^{(b)})^+ \xrightarrow{E\to 0} (T^{(b)})^+.$$

作为一个提示，这种收敛需要依赖于$T^{(b)}$的最小奇异值。或者换句话说，如果E与$T^{(b)}$的最小奇异值相比不算小，那么一般来说，我们不能说$(T^{(b)})^+$和$(\widetilde{T}^{(b)})^+$是接近的。

在任何情况下，结合这些事实，我们有

$$\widetilde{T}^{(a)}(\widetilde{T}^{(b)})^+ \xrightarrow{E\to 0} T^{(a)}(T^{(b)})^+.$$

现在我们处于良好状态。右手边的特征向量是 U 的列。让 $\widetilde{U}$ 的列是左手边的特征向量。由于左手边以逆多项式速率收敛到右手边，我们可以调用我们的特征分解扰动界限（定理3.4.6）来得出结论，它们的特征向量也以逆多项式速率收敛。

特别地$\widetilde{U} \xrightarrow{E\to 0} U$，因为我们在这里滥用符号，因为上述收敛性只有在我们对$\widetilde{U}$的列应用适当的排列之后才会出现。同样，我们有$\widetilde{V} \xrightarrow{E\to 0} V$。

最后我们通过求解$\widetilde{U}$和$\widetilde{V}$中的线性系统来计算$\widetilde{W}$。可以证明$\widetilde{W} \xrightarrow{E\to 0} W$，利用$\widetilde{U}$和$\widetilde{V}$接近于条件良好的矩阵$U$和$V$的事实，这意味着从$\widetilde{U}$中的$i^{th}$列与$\widetilde{V}$中的$i^{th}$列的张量积得到的线性系统也是条件良好的。

这是证明Jennrich算法在噪声存在下表现良好的全部、血腥细节。如果我们使生活变得简单，在以下内容中我们分析我们的学习算法在无噪声情况($E = 0$)下，我们总是可以求助于各种扰动界限来对特征分解进行操作，并追踪所有错误如何传播，以限制我们找到的因子与真实隐藏因子之间的接近程度。这就是我所说的良药。您在使用张量分解时不必每次都考虑这些扰动界限，但您应该知道它们存在，因为它们确实是使用张量分解来解决存在采样噪声的学习问题的正当理由。

3.5 Exercises

Problem 3-1:

(a{v*} 假设我们想要求解线性方程组 $Ax = b$ (其中 $A \in \mathbb{R}^{n\times n}$ 是方阵且可逆) 但我们只能访问到一个满足的噪声向量 $\tilde{b}$

$$\frac{\|b - \tilde{b}\|}{\|b\|} \leq \varepsilon$$

并且一个满足 $\|A - \tilde{A}\| \leq \delta$（在算子范数）下的噪声矩阵$\tilde{A}$。设 $\tilde{x}$ 为$\tilde{A}\tilde{x} = \tilde{b}$ 的解。证明

$$\frac{\|x - \tilde{x}\|}{\|x\|} \leq \frac{\varepsilon\, \sigma_{\max}(A) + \delta}{\sigma_{\min}(A) - \delta}$$

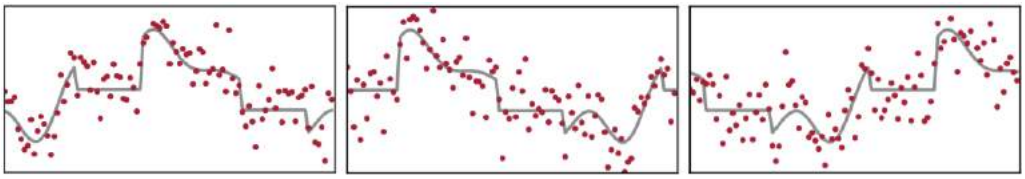
提供 $\delta < \sigma_{\min}(A)$。

(b) 现在，假设我们确切地知道 A，但 A 可能条件很差，甚至奇异。我们想证明仍然可能恢复 x 的一个特定坐标 x_j。设 $\tilde{x}$ 是 $A\tilde{x} = \tilde{b}$ 的任意解，并让 a_i 表示 A 的第 i 列。证明

$$|x_j - \tilde{x}_j| \leq \frac{\|b - \tilde{b}\|}{C_j}$$

C_j 是 a_j 投影到 span($\{a_i\}_{i \neq j}$) 正交补的范数。

Problem 3-2: 在 *multi-reference alignment* 问题中，我们观察到许多相同的未知信号 $x \in \mathbb{R}^d$ 的噪声副本，但每个副本都经过随机偏移的循环移位（如图所示）。



图：真信号 x 的三个平移副本以灰色显示。噪声样本 y_i 以红色显示。（图源：[23]）

形式上，对于 $i = 1, 2, \ldots, n$ 我们观察到

$$y_i = R_{\ell_i} x + \xi_i$$

在以下情况下：ℓ_i从$\{0, 1, \ldots, d-1\}$中均匀独立抽取；R_ℓ是循环平移向量ℓ个索引的算子；$\xi_i \sim \mathcal{N}(0, \sigma^2 I_{d\times d})$与$\{\xi_i\}_i$独立；$\sigma > 0$是一个已知的常数。将$d$、$x$和$\sigma$视为固定，而$n \to \infty$。目标是恢复$x$ (或x)的循环平移。

(a) 考虑张量 $T(x) = \frac{1}{d}\sum_{\ell=0}^{d-1}(R_\ell x)\otimes(R_\ell x)\otimes(R_\ell x)$。展示如何使用样本 y_i 来估计 T (，误差趋向于零的 $n \to \infty$)。对具有重复索引的条目（例如 T_{aab}, T_{aaa}）要格外小心。

(b) 给定 $T(x)$，证明Jennrich算法可以用来恢复 x (，直到循环移位)。在以下意义上假设 x 是 $generic$：令 $x' \in \mathbb{R}^d$ 为任意值，并令 x 通过向第一个元素添加一个小扰动 $\delta \sim \mathcal{N}(0, \epsilon)$ 从 x' 得到。*Hint:* 以 $\{R_\ell x\}_{0\le\ell<d}$ 为行形成一个矩阵，排列使得对角线项都是 x_1。

Chapter 4

Tensor Decompositions: Applications

许多令人兴奋的问题都符合以下范式：首先，我们选择一些参数化的分布族，它们足够丰富，可以模拟进化、写作和社会网络的形成。其次，我们设计算法来学习未知参数——你应该将其视为在我们数据中寻找隐藏结构的代理，就像解释物种如何相互演化的生命树，或解释文档集合背后的主题，或社交网络中强连接个体的社区。在本章中，我们所有的算法都将基于张量分解。我们将从分布的矩构造一个张量，并应用Jennrich算法来找到隐藏因子，进而揭示我们模型中的未知参数。

4.1 Phylogenetic Trees and HMMs

我们的第一次应用张量分解是用于学习系统发育树。在我们深入模型细节之前，了解动机是有帮助的。进化生物学中的一个核心问题是拼凑出描述物种如何相互演化的*tree of life*。更确切地说，它是一棵二叉树，其叶子代表*extant*物种（即目前存活的物种），而其内部节点代表*extinct*物种。当一个内部节点有两个子节点时，它代表了一次物种分化事件，其中两个种群分裂成不同的物种。

我们将与定义在这棵树上的随机模型合作，其中每条边都引入其自身的随机性，代表突变。更确切地说，我们的模型具有以下组件：

(a) 以 r (为根的有根二叉树，其叶子节点不一定具有相同的深度)

(b) 一组状态 Σ，例如 $\Sigma = \{A, C, G, T\}$。令 $k = |\Sigma|$。

(©) 树上的马尔可夫模型；即根状态上的分布 π_r 以及每条边上的转移矩阵 P^{uv} (u, v)。

我们可以按照以下方式从模型中生成一个样本：我们根据 π_r 选择根状态，对于每个具有父状态 u 的节点 v，我们根据 P^{uv} 的第 i 行定义的分布选择 v 的状态，其中 i 是 u 的状态。或者，我们可以将 $s(\cdot) : V \to \Sigma$ 视为一个随机函数，该函数分配

状态到顶点，其中边缘分布 $s(r)$ 为 π_r 且

$$P_{ij}^{uv} = \mathbb{P}(s(v) = j | s(u) = i),$$

注意，在树中从 v 到 t 的（唯一）最短路径经过 u 的条件下，$s(v)$ 与 $s(t)$ 是独立的。

在这个部分，我们的主要目标是学习在给定足够模型样本的情况下根树和转移矩阵。现在是将这个问题与生物学联系起来的时候了。从这个模型中抽取的样本代表什么？如果我们已经对现存物种进行了测序，并且这些序列已经得到了适当的对齐，那么我们可以将这些序列中的 i^{th} 符号视为上述模型样本中叶子的状态配置。当然，这只是一个对生物学问题的过度简化，但它仍然捕捉到了许多有趣的现象。

实际上有两个独立的任务：(a) 学习拓扑结构 (b) 估计转移矩阵。我们寻找拓扑结构的方法将遵循Steel [133] 和 Erdos, Steel, Szekely, 以及 Warnow [69] 的基础工作。一旦我们知道拓扑结构，我们可以应用张量分解来找到转移矩阵，遵循Chang [47] 和 Mossel 及 Roch [115] 的方法。

Learning the Topology

这里我们将关注学习树拓扑结构的问题。由于Steel [133]提出的惊人想法是，有一种方法可以定义一个*evolutionary distance*。这个距离的重要之处在于它（a）为每个

树边和（b）可以仅根据它们的联合分布对任何一对节点进行评估。那么，具有这些特性的神奇函数是什么？首先，对于任何一对节点 a 和 b，让 F^{ab} 是一个表示它们联合分布的 $k \times k$ 矩阵：

$$F^{ab}_{ij} = \mathbb{P}(s(a) = i, s(b) = j).$$

Definition 4.1.1 *Steel's evolutionary distance on an edge* (u, v) *is*

$$\nu_{uv} = -\ln|det(P^{uv})| + \frac{1}{2}\ln\left(\prod_{i\in[k]}\pi_u(i)\right) - \frac{1}{2}\ln\left(\prod_{i\in[k]}\pi_v(i)\right).$$

钢铁[133]证明了该距离函数的两个基本性质，以下引理中进行了描述：

Lemma 4.1.2 *Steel's evolutionary distance satisfies:*

(a) ν_{uv} *is nonnegative and*

(b) for any pair of nodes a *and* b*, we have*

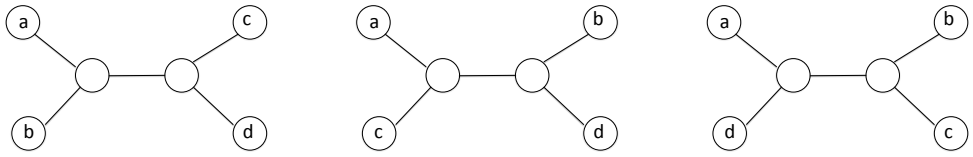
$$\psi_{ab} := -\ln|det(F^{ab})| = \sum_{(u,v)\in p_{ab}} \nu_{uv}$$

where p_{ab} *is the shortest path connecting* a *and* b *in the tree.*

这使得这个距离对我们目的非常有用，因为对于任何一对叶子 a 和 b，我们可以从我们的样本中估计 F^{ab}，因此我们可以在叶子上（近似地）计算 ψ_{ab}。所以从现在起，我们可以想象在树的边上有某种非负函数，并且我们有一个计算连接任何两个叶子的路径上距离总和的的或acles。

Reconstructing Quartets

现在我们将使用Steel的进化距离来通过每次拼接四个节点来计算拓扑。



(a) (b) (c)

图4.1：可能的四元拓扑

我们的目标是确定在给定的成对距离下，这些诱导拓扑中哪一个是真正的拓扑。

Lemma 4.1.3 *If all distances in the tree are strictly positive, then it is possible to determine the induced topology on any four nodes a, b, c and d given an oracle that can compute the distance between any pair of them.*

Proof: 证明是通过情况分析。考虑图4.1中给出的这些节点之间可能的三种诱导拓扑。在这里，我们所说的诱导拓扑是指删除任何一对四个叶子节点之间最短路径上的边，并在可能的情况下将路径合并为单一条边。

它很容易验证，在拓扑（a）下，我们有

$$\psi(a,b)+\psi(c,d) < \min\left\{\psi(a,c)+\psi(b,c), \psi(a,d)+\psi(b,d)\right\}.$$

但是，在拓扑(b)或(c)下，这个不等式不成立。有一种类似的方法可以识别其他每个拓扑与它们之间的区别。这意味着我们可以简单地计算三个值$\psi(a,b)+\psi(c,d)$、$\psi(a,c)+\psi(b,c)$和$\psi(a,d)+\psi(b,d)$。其中最小的一个决定了诱导拓扑分别是(a)、(b)或(c)。■

确实，仅从这些四重测试中，我们就可以恢复树的拓扑结构。

Lemma 4.1.4 *If for any quadruple of leaves a, b, c and d we can determine the induced topology, it is possible to determine the topology of the tree.*

Proof: 方法首先确定哪些叶子节点具有相同的父节点，然后确定哪些叶子节点具有相同的曾祖父母，依此类推。首先，固定一对叶子节点 a 和 b。很容易看出，它们具有相同的父节点当且仅当对于其他任何选择叶子节点 c 和 d，四重测试返回拓扑（a）。现在，如果我们想确定一对叶子节点 a 和 b 是否具有相同的曾祖父母，我们可以修改方法如下：它们具有相同的曾祖父母当且仅当对于其他任何选择叶子节点 c 和 d，这两个叶子节点都不是 a 或 b 的兄弟，四重测试返回拓扑（a）。本质上，我们是通过首先找到最近的节点来构建树。■

一个重要的技术点是，我们只能从我们的样本中近似 F^{ab}。当 a 和 b 接近时，这转化为对 ψ_{ab} 的良好近似，但当 a 和 b 相距较远时，则会产生噪声。最终，Erdos、Steel、Szekely 和 Warnow 在 [69] 中提出的方法是仅使用所有距离都较短的四分位数测试。

Estimating the Transition Matrices

现在我们将假设我们知道树的拓扑结构，并将我们的目标设定为估计转移矩阵。我们的方法是使用张量分解。为此，对于任何由叶子 a、b 和 c 组成的三元组，令 T^{abc} 为以下定义的 $k \times k \times k$ 张量：

$$T^{abc}_{ijk} = \mathbb{P}(s(a) = i, s(b) = j, s(c) = k).$$

这些是我们可以从样本中估计的分布的三阶矩。在本节中，我们将假设转移矩阵是满秩的。这意味着我们可以任意重新根树。现在考虑位于 a、b 和 c 之间所有最短路径上的唯一节点。让我们将其作为根。然后

$$\begin{aligned} T^{abc} &= \sum_{\ell} \mathbb{P}(s(r) = \ell)\mathbb{P}(s(a) = \cdot | s(r) = \ell) \otimes \mathbb{P}(s(b) = \cdot | s(r) = \ell) \otimes \mathbb{P}(s(c) = \cdot | s(r) = \ell) \\ &= \sum_{\ell} \mathbb{P}(s(r) = \ell) P^{ra}_{\ell} \otimes P^{rb}_{\ell} \otimes P^{rc}_{\ell} \end{aligned}$$

我们在其中使用 P^{rx}_{ℓ} 表示转换矩阵 P^{rx} 的 ℓ 行。

我们现在可以将第3.3节中的算法应用于计算T的张量分解，其因子在缩放范围内是唯一的。此外，这些因子是概率分布，因此我们可以计算它们的适当归一化。我们将此过程称为$star\ test$。（实际上，第3.3节中张量分解的算法已被多次重新发现，也被称为Chang引理[47]）。

在[115]中，Mossel和Roch使用这种方法来找到{v*}的转移矩阵。

一个系统发育树，给定树拓扑如下。假设 u 和 v 是内部节点，而 w 是一个叶节点。此外，假设 v 位于 u 和 w 之间的最短路径上。基本思想是编写

$$P^{uw} = P^{uv}P^{vw}$$

并且如果我们能通过上述星号测试找到 P^{uw} 和 P^{vw} (，那么我们可以计算 $P^{uv} = P^{uw}(P^{vw})^{-1}$，因为我们假设转移矩阵是可逆的。

然而，存在两个严重的并发症:

(a) 如同寻找拓扑的情况一样，长路径非常嘈杂。

Mossel和Roch表明，仅使用对短路径的查询也可以恢复转移矩阵。

(b) 我们只能恢复到重新标记的张量分解。

在上述星号测试中，我们可以对 r 的状态应用任何排列，并相应地排列转移矩阵 P^{ra}、P^{rb} 和 P^{rc} 的行，使得 a, b 和 c 上的联合分布保持不变。

然而，Mossel和Roch的方法是在Valiant [138]的$probably\ approximately\ correct$学习框架下工作，其目标是学习一个生成模型，该模型在叶子节点上产生几乎相同的联合分布。特别是，如果有多种方式对内部节点进行标记以在叶子节点上产生相同的联合分布，我们对它们是无所谓的。

Remark 4.1.5 *Hidden Markov models are a special case of phylogenetic trees where the underlying topology is a caterpillar. But note that for the above algorithm, we need that the transition matrices* 并且 *the observation matrices are full-rank.*

更精确地说，我们要求转移矩阵是可逆的，并且那些行空间对应隐藏节点、列空间对应输出符号的观测矩阵每个都具有满秩。

Beyond Full Rank?

该算法假设所有转移矩阵都是满秩的。实际上，如果我们去掉这个假设，那么嵌入一个经典的难学习问题实例 *noisy parity problem* [37] 就变得容易了。让我们首先定义这个问题 *without noise*:

让 $S \subset [n]$，并独立且均匀地从 ran- dom 中选择 $X^{(j)} \in \{0,1\}^n$，对于 $j = 1, \ldots, m$。给定 $X^{(j)}$ 和 $b^{(j)} = \chi_S(X^{(j)}) := \sum_{i \in S} X_i^{(j)} \bmod 2$ 对于每个 j，目标是恢复 S。

这是一个相当简单的问题：设A为矩阵，其j行是$X^{(j)}$，设b为一个列向量，其j个元素是$b^{(j)}$。很容易看出1_S是线性系统$Ax = b$的一个解，其中1_S是S的指示函数。此外，如果我们选择$\Omega(n \log n)$个样本，那么A以高概率具有满列秩，因此这个解是唯一的。然后我们可以通过在$GF(2)$上求解线性系统来找到S。

然而，上述问题中微小的变化并没有改变样本复杂度，但使得问题变得极其困难。噪声奇偶校验问题是

与上面相同，但对于每个 j，我们独立地给出值 $b^{(j)} = \chi_S(X^{(j)})$，可能是 2/3，否则 $b^{(j)} = 1 - \chi_S(X_j)$。挑战在于我们不知道哪些标签已经被翻转。

Claim 4.1.6 *There is an exponential time algorithm that solves the noisy parity problem using $m = O(n$ 对数 $n)$ samples*

Proof: 对于每个 T，计算与观测标签一致的样本比例 — 即

$$\frac{1}{m}\sum_{j=1}^{m} \mathbb{1}_{\chi T(X^{(j)})=b^{(j)}}$$

从标准浓度界限可以得出，这个值以高概率大于（比如说）3/5，当且仅当 $S = T$. ■

最先进的由于Blum，Kalai和Wasserman[37]，其运行时间和样本复杂度为2 $^{n/\log n}$。人们普遍认为，即使在给定任何多项式数量的样本的情况下，也没有多项式时间算法可以用于噪声奇偶校验。*This is an excellent example of a problem whose sample complexity and computational complexity are (conjectured) to be wildly different.*

接下来，我们展示如何将噪声奇偶校验问题中的样本嵌入到HMM中，然而要做到这一点，我们将使用非满秩的转移矩阵。考虑一个具有n个隐藏节点的HMM，其中第i个隐藏节点用于表示X的第i个坐标以及运行奇偶校验

$$\chi_{S_i}(X) := \sum_{i' \le i, i' \in S} X(i') \mod 2.$$

因此，每个节点有四种可能的状态。我们可以定义以下转移矩阵。设$s(i) = (x_i, s_i)$为i第v3个内部节点的状态。

我们可以定义以下转移矩阵：

$$\text{if } i+1 \in S \qquad P^{i,i+1} = \begin{cases} \frac{1}{2} & (0, s_i) \\ \frac{1}{2} & (1, s_i + 1 \mod 2) \\ 0 & \text{otherwise} \end{cases}$$

$$\text{if } i+1 \notin S \qquad P^{i,i+1} = \begin{cases} \frac{1}{2} & (0, s_i) \\ \frac{1}{2} & (1, s_i) \\ 0 & \text{otherwise} \end{cases} .$$

在每个内部节点我们观察到 x_i，在最后一个节点我们也观察到 $\chi_S(X)$，概率为 2/3，否则为 $1 - \chi_S(X)$。每个从有噪声奇偶校验问题中抽取的样本是这个 HMM 的观察集，如果我们能学习它的转移矩阵，我们必然会学习 S 并解决有噪声奇偶校验问题。

请注意，这里的观测矩阵肯定不是满秩的，因为我们只观察到两种可能的发射，尽管每个内部节点有四种可能的状态！因此，当转移（或观测）矩阵不是满秩时，这些问题变得更加困难！

4.2 Community Detection

这里我们介绍了张量方法在社区检测中的应用。有许多设置，我们希望发现 *communities*——即，分组

强连通个体。在这里，我们将关注图论方法，其中我们将社区视为一组节点，这些节点之间的连接比与集合外节点的连接更好。我们可以用许多方式来形式化这个概念，每种方式都会导致不同的优化问题，例如*sparsest cut*或*k-densest subgaph*。

然而，这些优化问题中的每一个都是 NP-难，甚至更糟糕的是难以近似。相反，我们将以平均情况模型来构建我们的问题，其中存在一个用于生成随机图的潜在社区结构，我们的目标是以高概率从图中恢复出真实的社区。

Stochastic Block Model

这里我们介绍了块随机模型，该模型用于在 V 上生成具有 $|V| = n$ 的随机图。此外，该模型由参数 p 和 q 以及由函数 π 指定的分区指定：

- π $V \to [k]$ 将顶点 V 划分为 k *disjoint* 组（我们稍后会放宽这个条件）；

- 每个可能的边（u, v）是以以下方式选择的（*independently*）：

$$\mathbb{P}[(u,v) \in E] = \begin{cases} q & \pi(u) = \pi(v) \\ p & \text{otherwise} \end{cases}.$$

在我们的设置中，我们将设置 $q > p$，这被称为 *assortative* 情况，但这个模型

也适用于称为 *disassortative* 情况的 $q < p$。例如，当 $q =$ 为 0 时，我们正在生成一个具有植入 k 着色的随机图。无论如何，我们观察从上述模型生成的随机图，我们的目标是恢复由 π 描述的划分。

何时是这种情况 *information theoretically*? 事实上，即使是对于 $k = 2$，其中 π 是二分，我们仍然需要

$$q - p > \Omega\Big(\sqrt{\frac{\log n}{n}}\Big)$$

为了确保真正的二分是唯一最小的分割，该分割以高概率将随机图 G 分割。如果 $q - p$ 更小，那么在信息论上甚至不可能找到 π。实际上，我们还应该要求分割的每一部分都很大，为了简单起见，我们将假设 $k = O(1)$ 和 $|\{u|\pi(u) = i\}| = \Omega(n)$。

在块随机模型中对随机图进行划分的研究已经有一系列，最终在McSherry [109] 的工作中达到高潮：

Theorem 4.2.1 *[109] There is an efficient algorithm that recovers π (up to relabeling) if*

$$\frac{q-p}{q} > c\sqrt{\frac{\log n/\delta}{qn}}$$

and succeeds with probability at least $1 - \delta$.

此算法基于谱聚类，我们将观察到的邻接矩阵视为一个秩 k 矩阵的和，该矩阵编码 π，以及一个误差项。如果误差很小，则可以通过找到邻接矩阵的最佳秩 k 近似来恢复接近真实秩 k 矩阵。有关详细信息，请参阅 [109]。

我们将遵循Anandkumar等人[9]提出的方法，该方法利用张量分解。事实上，他们的算法在$mixed\ membership$模型中也适用，其中我们允许每个节点是[k]上的分布。然后，如果π^u和π^v分别是u和v的概率分布，则边的概率(u, v)是$\sum_i \pi_i^u \pi_i^v q + \sum_{i \neq j} \pi_i^u \pi_j^v p$。我们可以将这个概率解释为：$u$和$v$分别根据$\pi^u$和$\pi^v$选择一个社区，如果他们选择相同的社区，则存在边的概率为q，否则存在边的概率为p。

Counting Three Stars

当我们使用张量分解时，实际上我们是在寻找条件独立的随机变量。这就是我们使用它们来学习系统树过渡矩阵时所做的事情。在那里，一旦我们根据唯一节点 r 的状态进行条件化，a、b 和 c 的状态就是独立的。我们在这里将做类似的事情。

如果我们有四个节点 a、b、c 和 x，并且我们根据节点属于哪个社区 x 进行条件假设，那么 (a, x)、(b, x) 和 (c, x) 是否是图中边都是独立的随机变量。当所有三条边都存在时，这被称为 $three\ star$。我们将设置一个计数三个星号的张量，如下所示。首先将 V 划分为四个集合 X、A、B 和 C。现在让 $\Pi \in \{0, 1\}^{V \times k}$ 表示节点到社区的（未知）分配，使得 Π 的每一行恰好包含一个 1。最后让 R 是一个 $k \times k$ 矩阵，其条目是连接概率。在

特定

$$(R)_{ij} = \begin{cases} q & i = j \\ & \\ p & i \neq j \end{cases}.$$

考虑乘积 ΠR。ΠR 的 i^{th} 列编码了从社区 i 中的一个节点到对应给定行的节点的边的概率。

$$(\Pi R)_{xi} = \Pr[(x, a) \in E | \pi(a) = i].$$

我们将使用 $(\Pi R)_i^A$ 来表示矩阵 ΠR 在 i^{th} 列和 A 行上的限制，对于 B 和 C 也同样适用。此外，设 p_i 为 X 中属于社区 i 的节点比例。然后，我们的算法围绕以下张量展开

$$T = \sum_i p_i (\Pi R)_i^A \otimes (\Pi R)_i^B \otimes (\Pi R)_i^C.$$

关键主张是:

Claim 4.2.2 *Let $a \in A$, $b \in B$ and $c \in C$, then*

$$T_{a,b,c} = \mathbb{P}[(x, a), (x, b), (x, c) \in E]$$

where the randomness is over x chosen uniformly at random from X and the edges included in G.

这是从上面的讨论中直接得出的。有了这个张量在手，我们需要证明的关键点是:

(a) 因子 $\{(\Pi R)_i^A\}_i$、$\{(\Pi R)_i^B\}_i$ 和 $\{(\Pi R)_i^B\}_i$ 线性无关

(b) 我们可以从 $\{(\Pi R)_i^A\}_i$ 中恢复分区 π，直到重新标记哪个社区是哪个。

我们将忽略准确估计 T 的问题，但大致来说，这相当于选择 X 远大于 A、B 或 C 并应用适当的浓度界限。无论如何，现在让我们弄清楚隐藏因素为什么是线性无关的。

Lemma 4.2.3 *If A, B and C have at least one node from each community then the factors $\{(\Pi R)_i^A\}_i$, $\{(\Pi R)_i^B\}_i$, and $\{(\Pi R)_i^B\}_i$ are each linearly independent.*

Proof: 首先容易看出 R 是满秩的。现在如果 A 至少包含每个社区的一个节点，那么 R 的每一行都出现在 $(\Pi R)^A$ 中，这意味着它具有满列秩。对于 B 和 C 也同样适用。■

实际上，我们需要因子不仅要有满秩，还要有良好的条件。与前面引理中相同类型的论证表明，只要每个社区在 A、B 和 C（中都有良好的代表性，这在 A、B 和 C 足够大且随机选择的情况下几乎总是发生），那么因子 $\{(\Pi R)_i^A\}_i$、$\{(\Pi R)_i^B\}_i$ 和 $\{(\Pi R)_i^B\}_i$ 将会有良好的条件。

现在让我们从隐藏因素中恢复社区结构：首先，如果我们有 $\{(\Pi R)_i^A\}_i$，那么我们可以通过将具有相同行的节点分组来将 A 划分为社区。反过来，如果 A 足够大，那么我们可以将这种划分扩展到整个图：我们仅在节点 $x \notin A$ 与 i 中连接的节点 $a \in A$ 的比例接近 q 时，将其添加到社区 i。如果 A 足够大并且我们已经恢复了其社区

结构正确，那么以高概率，此过程将恢复整个图中的真实社区。

对于算法的全面分析，包括其样本复杂度和准确性，请参阅[9]。Anandkumar等人还给出了一种混合成员模型算法，其中每个π_u从Dirichlet分布中选择。我们不会涵盖这个后续扩展，因为我们将转而在主题模型的设置中解释这些类型的技巧。

Discussion

我们注意到，存在对随机块模型的强大扩展，被称为 $semi\text{-}random\ \ models$。大致来说，这些模型允许在同一个簇中的节点之间添加边，并在 G 生成后删除簇之间的边。这听起来像是对手只是通过加强社区内的联系和打破它们之间的联系来使你的生活更容易。如果真实的社区结构是将 G 划分为 k 个部分，这些部分切割的边最少，那么在变化之后这更是如此。有趣的是，许多张量和谱算法在半随机模型中失效，但即使在更一般的设置中，也有优雅的技术来恢复 π（参见 [71]，[72]）。

This is some food for thought and begs the question: How much are we exploiting brittle properties of our stochastic model?

4.3 Extensions to Mixed Models

许多我们迄今为止研究的模型可以推广到所谓的$mixed\ membership\ models$。例如，我们可以将文档建模为多个主题的混合，而不是仅仅关于一个主题。同样，我们可以将个人建模为属于多个社区的混合，而不是仅仅属于一个社区。在这里，我们将利用混合成员资格设置中的张量分解。

Pure Topic Model

作为热身，我们首先看看如何使用张量分解来发现纯主题模型中的主题，在这种模型中，每篇文档只涉及一个主题。我们的方法将遵循Anandkumar等人[10]的方法。回想一下，在纯主题模型中，存在一个未知的$m \times r$主题矩阵A，并且每篇文档都是根据以下随机过程生成的：

(a) 以概率 p_i 选择文档 j 的主题 i

(b) 根据分布 A_i 选择 N_j 个词

在2.4节中，我们构建了表示成对单词联合分布的Gram矩阵。在这里，我们将使用单词三元组的联合分布。令w_1、w_2和w_3分别表示其第一个、第二个和第三个单词的随机变量。

Definition 4.3.1 *Let T denote the $m \times m \times m$ tensor where*

$$T_{a,b,c} = \mathbb{P}[w_1 = a, w_2 = b, w_3 = c].$$

我们可以将 T 用未知主题矩阵表示如下:

$$T = \sum_{i=1}^{r} p_i A_i \otimes A_i \otimes A_i$$

因此，我们如何从纯主题模型中给出的样本恢复主题矩阵？我们可以构建一个估计 $\widetilde{T}$，其中 $\widetilde{T}_{a,b,c}$ 计算我们样本中第一词、第二词和第三词分别是 a、b 和 c 的文档的分数。如果文档数量足够大，那么 $\widetilde{T}$ 将收敛到 T。

现在我们可以应用Jennrich的算法。假设A具有满列秩，我们将恢复分解中的真实因子，直到缩放。然而，由于A中的每一列都是一个分布，我们可以适当地归一化我们找到的任何隐藏因子，并计算p_i的值。要真正使这起作用，我们需要分析为了使$\widetilde{T}$接近T我们需要多少文档，然后应用第3.4节中的结果，在那里我们分析了Jennrich算法的噪声容忍度。重要的是，我们估计的列$\widetilde{A}$以与给定样本数量成反多项式速率收敛到A的列，其中收敛速率取决于诸如A的列条件如何等因素。

Latent Dirichlet Allocation

现在让我们继续讨论混合成员模型。到目前为止，驱动我们看到的张量分解的所有应用的都是条件独立随机变量。在纯主题模型的情况下，当我们根据用于生成文档的主题进行条件化时，前三个词的分布是独立的。然而，在混合模型中情况不会这么简单。我们从可用数据构建低秩三阶张量的方式将更复杂地结合低阶统计量。

我们将研究Blei等人开创性工作中引入的潜在狄利克雷分配模型[36]。令 $\Delta := \{x \in \mathbb{R}^r : x \geq 0, \sum_i x_i = 1\}$ 表示 r 维单纯形。然后每个文档根据以下随机过程生成:

(a) 根据Dirichlet分布Dir($\{\alpha_i\}_i$)，为文档j选择主题混合 $w_j \in \Delta$

(b) 重复 N_j 次：从 w_j 中选择一个主题 i，并根据分布 A_i 选择一个单词。

Dirichlet分布定义为

$$p(x) \propto \prod_i x_i^{\alpha_i - 1} \text{ for } x \in \Delta$$

此模型在以下方面已经更加真实。当文档较长（例如 $N_j > m \log m$）时，在纯主题模型中，文档对必然会有几乎相同的单词经验分布。但情况已不再如此

在混合模型中，如上所示。

基本问题是将我们的张量分解方法扩展到学习混合模型时，现在满足以下表达式的第三阶张量用于计算三元词对的联合分布

$$T = \sum_{ijk} D_{ijk} A_i \otimes A_j \otimes A_k$$

在随机文档中，$D_{i,j,k}$ 是前三个单词分别由主题 i、j 和 k 生成的概率。在纯主题模型中，$D_{i,j,k}$ 是对角线，但对于混合模型则不是！

Definition 4.3.2 *A Tucker decomposition of T is*

$$T = \sum_{i,j,k} D_{i,j,k} a_i \otimes b_j \otimes c_k$$

where D is $r_1 \times r_2 \times r_3$. We call D the core tensor.

结果显示，您可以计算一个Tucker分解，其中r_1、r_2和r_3尽可能小（它们实际上是列、行和管的维数）。然而，一个最小的Tucker分解通常不是唯一的，因此即使我们给出了T并计算了一个最小的Tucker分解，我们也无法保证其因子是主题模型中的隐藏主题。我们需要找到另一种方法，这相当于从T和我们能访问的较低阶矩中构造一个低秩的三阶张量。

因此，我们如何将张量分解方法扩展到适用于潜在狄利克雷分配模型？Anandkumar等人[8]的优雅方法为

基于以下想法:

Lemma 4.3.3

$$T = \sum_{ijk} D_{ijk} A_i \otimes A_j \otimes A_k$$

$$S = \sum_{ijk} \widetilde{D}_{ijk} A_i \otimes A_j \otimes A_k$$

$$\implies T - S = \sum_{ijk} (D_{ijk} - \widetilde{D}_{ijk}) A_i \otimes A_j \otimes A_k$$

Proof: 证明是多元线性代数中的一个简单练习。■

因此，如果我们能访问其他张量 S，这些张量可以使用与其 Tucker 分解中相同的因子 $\{A_i\}_i$ 来表示，我们可以减去 T 和 S，并希望使核心张量对角化。我们可以将 D 视为我们设置中狄利克雷分布的第三阶矩。我们还能访问哪些其他张量?

Other Tensors

我们根据以下实验描述了张量 T: 设 $T_{a,b,c}$ 为随机文档中前三个单词分别是 a、b 和 c 的概率。但我们可以考虑其他替代实验。为了给出 LDA 的张量谱算法，我们还需要进行以下两个实验:

(a) 随机选择三份文档，并查看每份文档的第一个单词

(b) 随机选择两份文档，并查看第一份文档的前两个单词以及第二份文档的第一个单词

这两个新实验与旧实验结果相结合，产生了三个张量，它们的Tucker分解使用相同的因子，但核心张量不同。

Definition 4.3.4 *Let μ, M and D be the first, second and third order moments of the Dirichlet distribution.*

更精确地说，设μ_i为随机文档中第一个单词生成自主题i的概率。设$M_{i,j}$为随机文档中第一个和第二个单词分别生成自主题i和j的概率。并且，如前所述，设$D_{i,j,k}$为随机文档中前三个单词分别生成自主题i、j和k的概率。然后设T^1、T^2和T^3分别为第一个（选择三份文档）、第二个（选择两份文档）和第三个（选择一份文档）实验的期望。

Lemma 4.3.5 *(a)* $T1 = \sum_{i,j,k}[\mu \otimes \mu \otimes \mu]_{i,j,k} A_i \otimes A_j \otimes A_k$

(b) $T^2 = \sum_{i,j,k}[M \otimes \mu]_{i,j,k} A_i \otimes A_j \otimes A_k$

(c) $T^3 = \sum_{i,j,k} D_{i,j,k} A_i \otimes A_j \otimes A_k$

Proof: 让 w_1 表示第一个词，让 t_1 表示 w_1 (的主题，对于其他词) 也同样处理。我们可以将 $\mathbb{P}[w_1 = a, w_2 = b, w_3 = c]$ 展开为：

$$\sum_{i,j,k} \mathbb{P}[w_1 = a, w_2 = b, w_3 = c | t_1 = i, t_2 = j, t_3 = k] \mathbb{P}[t_1 = i, t_2 = j, t_3 = k]$$

现在词元是直接的。■

注意 $T^2_{a,b,c} \neq T^2_{a,c,b}$，因为两个单词来自同一文档。尽管如此，我们可以以自然的方式对称化 T^2：设置 $S^2_{a,b,c} = T^2_{a,b,c} + T^2_{b,c,a} + T^2_{c,a,b}$。因此，对于任何排列 π: $\{a,b,c\} \to \{a,b,c\}$。

我们的主要目标是证明以下恒等式:

$$\alpha_0^2 D + 2(\alpha_0+1)(\alpha_0+2)\mu^{\otimes 3} - \alpha_0(\alpha_0+2)M \otimes \mu(\text{all three ways}) = \text{diag}(\{p_i\}_i)$$

在 $\alpha_0 = \sum_i \alpha_i$。因此，我们有

$$\alpha_0^2 T^3 + 2(\alpha_0+1)(\alpha_0+2)T^1 - \alpha_0(\alpha_0+2)S^2 = \sum_i p_i A_i \otimes A_i \otimes A_i$$

重要的一点是我们可以从我们的样本中估计左侧的项（如果我们假设我们知道 α_0 ），并且我们可以应用 Jennrich 的算法到右侧的张量以恢复主题模型，前提是 A 具有满列秩。实际上，我们可以从我们的样本中计算 α_0（参见 [8]），但我们将专注于证明上述恒等式。

Moments of the Dirichlet

我们要建立的主要恒等式只是关于狄利克雷分布矩的一个陈述。事实上，我们可以将狄利克雷分布视为由以下组合过程定义:

(a) 初始时，每种颜色的球有 α_i 个

(b) 重复 C 次：随机选择一个球，将其放回并再放回一个与其颜色相同的球

这个过程给出了Dirichlet分布的另一种特征描述，从中可以轻松计算出：

(a) $\mu = [\frac{\alpha_1}{\alpha_0}, \frac{\alpha_2}{\alpha_0}, ..., \frac{\alpha_r}{\alpha_0}]$

(b) $M_{i,j} = \begin{cases} \frac{\alpha_i(\alpha_i+1)}{\alpha_0(\alpha_0+1)} & i = j \\ \frac{\alpha_i\alpha_j}{\alpha_0(\alpha_0+1)} & \text{otherwise} \end{cases}$.

(c) $T_{i,j,k} = \begin{cases} \frac{\alpha_i(\alpha_i+1)(\alpha_i+2)}{\alpha_0(\alpha_0+1)(\alpha_0+2)} & i = j = k \\ \frac{\alpha_i(\alpha_i+1)\alpha_k}{\alpha_0(\alpha_0+1)(\alpha_0+2)} & i = j \neq k \\ \frac{\alpha_i\alpha_j\alpha_k}{\alpha_0(\alpha_0+1)(\alpha_0+2)} & i, j, k \text{ distinct} \end{cases}$.

例如，对于 $T_{i,i,k}$，这是前两个球颜色为 i，第三个球颜色为 k 的概率。第一个球颜色为 i 的概率是 $\frac{\alpha_i}{\alpha_0}$，由于我们将其放回并放入一个同色的球，第二个球颜色为 i 的概率也是 $\frac{\alpha_i+1}{\alpha_0+1}$。第三个球颜色为 k 的概率是$\frac{\alpha_k}{\alpha_0+2}$。在其它情况下也容易验证上述公式。

注意，在上面的公式中只考虑分子要容易得多。如果我们能证明以下关于仅分子的关系

$$D + 2\mu^{\otimes 3} - M \otimes \mu(\text{all three ways}) = \text{diag}(\{2\alpha_i\}_i)$$

我们可以很容易地验证，通过将 $\alpha_0^3(\alpha_0 + 1)(\alpha_0 + 2)$ 乘以所有项，我们将得到我们想要的公式。

Definition 4.3.6 *Let* $R = num(D) + num(2_{\mu^{\otimes 3}}) - num(M \otimes \mu)$*(all three ways)*

然后，主要引理是：

Lemma 4.3.7 $R = diag(\{2\alpha_i\}_i)$

我们将通过案例分析来建立这一点：

Claim 4.3.8 *If* i, j, k *are distinct then* $R_{i,j,k} = 0$

这是立即的，因为 i, j, k 分子 D、$\mu^{\otimes 3}$ 和 $M \otimes \mu$ 都是 $\alpha_i\alpha_j\alpha_k$。

Claim 4.3.9 $R_{i,i,i} = 2\alpha_i$

这是立即的，因为 i, i, i 分子 D 是 $\alpha_i(\alpha_i + 1)(\alpha_i + 2)$，同样地，$\mu^{\otimes 3}$ 的分子是 α_i^3。最后，i, i, i 分子 $M \otimes \mu$ 是 $\alpha_i^2(\alpha_i + 1)$。需要小心处理的情况是：

Claim 4.3.10 *If* $i \neq k$, $R_{i,i,k} = 0$

这个案例之所以棘手，是因为这三个方式)的项并不都同等计数。如果我们沿着张量的第三维度考虑μ，那么i^{th}主题在同一文档中出现了两次，但如果我们将μ视为沿着张量的第一或第二维度，即使i^{th}主题出现了两次，它也不会在同一文档中重复出现。因此，$M \otimes \mu$(所有三种方式)的分子是$\alpha_i(\alpha_i + 1)\alpha_k + 2\alpha_i^2\alpha_k$。同样，$D$的分子是$\alpha_i(\alpha_i + 1)\alpha_k$，而$\mu^{\otimes 3}$的分子再次是$\alpha_i^2\alpha_k$。

这三个断言共同建立了上述引理。即使我们可以在纯主题模型中立即分解的张量 T^3 在混合模型中不再具有对角核心张量，至少在LDA的情况下，我们仍然可以找到一个公式（我们可以从我们的样本中估计出每个项）来对核心张量进行对角化。这导致:

Theorem 4.3.11 *[8] There is a polynomial time algorithm to learn a topic matrix $\widetilde{A}$ whose columns are ϵ-close in Euclidean distance to the columns of A in a Latent Dirichlet Allocation model, provided we are given at least poly$(m, 1/\epsilon, 1/\sigma_r, 1/\alpha_{min})$ documents of length at least thee, where m is the size of the vocabulary and σ_r is the smallest singular value of A and α_{min} is the smallest α_i.*

Epilogue

Anandkumar等人[9]的混合成员随机块模型学习算法遵循相同的模式。再次，狄利克雷分布扮演了关键角色。与通常的随机块模型中每个节点只属于一个社区不同，每个节点由一个分布π_u描述，该分布覆盖社区，其中π_u从狄利克雷分布中选择。主要思想是计算三个星号，并通过从低阶子图计数中构建的张量来加减，以使自然Tucker分解的核心张量对角化。

这些技术似乎专门针对狄利克雷分布。正如我们所见，条件独立随机变量在张量分解中起着关键作用。在混合成员模型中，找到这样的随机变量具有挑战性。但狄利克雷分布与独立性有多远？尽管坐标不是独立的，但结果却表明它们

几乎都是。你可以通过独立地从每个坐标的beta分布中采样，然后重新归一化向量，使其位于 r-维简单形中，来从Dirichlet分布中采样。向前发展的一个有趣的概念问题是：*Are tensor decomposition methods fundamentally limited to settings where there is some sort of independence?*

4.4 Independent Component Analysis

我们可以将我们开发的张量方法视为一种使用高阶矩来学习分布参数（例如，用于系统发育树、HMM、LDA、社区检测）的方法，通过张量分解。在这里，我们将通过将方法应用于Comon [53] 介绍过的 *independent component analysis* 来给出另一种使用矩方法的方式。

这个问题定义简单：假设我们给出了以下形式的样本

$$y = Ax + b$$

在已知变量 x_i 独立且线性变换 (A, b) 未知的情况下。目标是高效地从多项式数量的样本中学习 A, b。这个问题有着悠久的历史，其典型动机是考虑一个被称为 *cocktail party problem* 的假设情况。

我们有 n 个麦克风和 n 次对话在进行中的房间内。每个麦克风都听到由 A 对应行给出的对话的叠加。如果我们把对话视为独立的

并且无记忆，我们能否将它们解开?

此类问题也常被称为 $blind\ source\ separation$。我们将遵循 Frieze、Jerrum 和 Kannan [74] 的方法。他们方法真正巧妙的地方在于它使用了非凸优化。

A Canonical Form and Rotational Invariance

首先将我们的问题转换成一个更方便的规范形式。结果是我们可以假设我们得到了以下样本的{v*}

$$y = Ax + b$$

但是，对于所有 i，$E[x_i] = 0, \mathbb{E}[x_i^2] = 1$。想法是，如果任何变量 x_i 不是均值为零，我们可以使其均值为零，并向 b 添加一个校正。同样，如果 x_i 不是方差为1，我们可以重新缩放它和相应的 A 列，使其方差为1。这些变化只是符号上的，它们不会影响我们观察到的样本分布。因此，从现在开始，让我们假设我们得到的样本是上述规范形式。

我们将给出一种基于非凸优化的算法来估计 A 和 b。但首先让我们讨论我们需要哪些假设。我们将做出两个假设：(a) A 是非奇异的，(b) 每个变量满足 $\mathbb{E}[x_i^4] \neq 3$。你现在应该已经习惯了非奇异假设（这是我们每次使用 Jennrich 算法时所需要的）。但第二个假设呢？它从哪里来？实际上，它实际上是非常自然的，并且需要排除一个有问题的情况。

Claim 4.4.1 *If each x_i is an independent standard Gaussian, then for any orthogonal transformation R, x and Rx and consequently*

$$y = Ax + b \text{ and } y = ARx + b$$

have identical distributions.

Proof: 标准 n-维高斯分布是旋转不变的。■

这意味着当我们的独立随机变量是标准高斯分布时，从信息论的角度来看，无法区分 A 和 AR。实际上，n 维高斯分布是问题所在。还有其他旋转不变分布，如 $\mathbb{S}^{n-1}$ 上的均匀分布，但其坐标不独立。标准 n 维高斯分布是唯一一个坐标独立的旋转不变分布。

鉴于这次讨论，我们可以理解我们关于第四矩的假设从何而来。对于标准高斯分布，其均值是零，其方差是一，其第四矩是三。因此，我们对每个 x_i 的第四矩的假设只是说它是明显非高斯的一种方式。

Whitening

通常，我们无法仅从二阶矩中学习 A。这实际上是我们在讨论旋转问题时出现的问题。在张量分解的情况下，我们直接使用三阶矩通过Jennrich算法学习 A 的列。在这里，我们将学习我们能从一阶和

二次矩，然后继续到四次矩。特别是，我们将使用前两个矩来学习 b 和学习 A 直到旋转：

Lemma 4.4.2 $\mathbb{E}[y] = b$ *and* $\mathbb{E}[yy^T] = AA^T$

Proof: 第一个恒等式显然。对于第二个，我们可以计算

$$\mathbb{E}[yy^T] = \mathbb{E}[Axx^TA^T] = A\,\mathbb{E}[xx^T]A^T = AA^T$$

在最后一个等式成立的情况下，条件是 $E[x_i] = 0$ 和 $E[x_i^2] = 1$，并且每个 x_i 是独立的。■

这意味着我们可以通过取足够的样本来估计 b 和 $M = AA^T$ 的任意精度。我所主张的是，这确定了 A 直到旋转。由于 $M \succ 0$，我们可以找到 B 使得 $M = BB^T$ 使用 Cholesky 分解。但是 B 和 A 之间有什么关系呢?

Lemma 4.4.3 *There is an orthogonal transformation* R *so that* $BR = A$

Proof: 回忆起我们假设 A 是非奇异的，因此 $M = AA^T$ 和 B 也是非奇异的。所以我们可以写出

$$BB^T = AA^T \Rightarrow B^{-1}AA^T(B^{-1})^T = I$$

这表明 $B^{-1}A = R$ 是正交的，因为每当一个方阵乘以其自身的转置是单位矩阵时，该矩阵就定义为正交。这完成了

证明。■

现在我们已经学习了A直到一个未知的旋转，我们可以开始使用更高阶矩来学习这个未知的旋转。首先，我们将对样本应用一个仿射变换：

$$z = B^{-1}(y - b) = B^{-1}Ax = Rx$$

这被称为 *whitening* (白噪声)，因为它使我们的分布的第一矩为零，第二矩在所有方向上均为一。我们分析的关键是以下泛函

$$F(u) = \mathbb{E}[(u^T z)^4] = \mathbb{E}[(u^T Rx)^4]$$

我们将要在单位球面上最小化它。当 u 在单位球面上变化时，$v^T = u^T R$ 也随之变化。因此，我们的优化问题等价于最小化

$$H(v) = \mathbb{E}[(v^T x)^4]$$

在单位球体上。这是一个非凸优化问题。一般来说，找到非凸函数的最小值或最大值是NP-难。但结果是，可以找到一个*local minimum*，并且这些足够好以学习R。

Lemma 4.4.4 *If for all i, $\mathbb{E}[x_i^4] < 3$ then the only local minima of $H(v)$ are at $v = \pm e_i$ where e_i are the standard basis vectors.*

Proof: 我们可以计算

$$\begin{aligned}\mathbb{E}\left[(v^T x)^4\right] &= \mathbb{E}\left[\sum_i (v_i x_i)^4 + 6\sum_{i<j}(v_i x_i)^2 (v_j x_j)^2\right] \\ &= \sum_i v_i^4\, \mathbb{E}(x_i^4) + 6\sum_{i<j} v_i^2 v_j^2 + 3\sum_i v_i^4 - 3\sum_i v_i^4 \\ &= \sum_i v_i^4 \left(\mathbb{E}\left[x_i^4\right] - 3\right) + 3\end{aligned}$$

从该表达式可以轻松检查，$H(v)$ 的局部极小值正好对应于将 $v = \pm e_i$ 设置为某些 i。■

回忆 $v^T = u^T R$ 以及因此这种特征化意味着 $F(u)$ 的局部最小值对应于将 u 设置为 $\pm R$ 的一个列。算法通过使用梯度下降（以及Hessian的下界）来证明可以快速找到 $F(u)$ 的局部最小值。直观上，如果你继续跟随陡峭的梯度，你正在减小目标值。最终，你必须卡在一个梯度很小的点上，这是一个近似局部最小值。任何这样的 u 必须接近 $\pm R$ 的某个列，然后我们可以递归到我们找到的向量的正交补中，以找到 R 的其他列。这个想法需要小心证明错误不会积累得太严重，参见 [74]，[140]，[17]。注意，当 $\mathbb{E}[x_i^4] \neq 3$ 而不是更强的假设 $\mathbb{E}[x_i^4] < 3$ 时，我们可以遵循相同的方法，但我们需要考虑 $F(u)$ 的局部最小值和局部最大值。此外，V empala 和 Xiao [140] 给出了一种在较弱的条件下工作的算法，即当存在一个与标准高斯不同的常数阶矩时。

我们上面遇到的奇怪表达式实际上被称为 *cumulants*，并且是分布矩的替代基。有时累积量

与它们满足吸引人的性质相比，更容易处理，即独立变量 X_i 和 X_j 的和的 k-阶累积量是 X_i 和 X_j 的 k-阶累积量的和。这个事实实际上在结合 Jennrich 的算法时，为解决独立成分分析提供了另一种更直观的方法，但它涉及到对高维累积量的一定程度的偏离。我们将此作为练习留给读者。

4.5 Exercises

Problem 4-1: Let $u \odot v$ 表示两个向量的 Khatri-Rao 积，其中如果 $u \in \mathbb{R}^m$ 和 $v \in \mathbb{R}^n$ 则 $u \odot v \in \mathbb{R}^{mn}$ 并且对应于按列将矩阵 uv^T 展平。此外，回忆一下，向量集合 $u_1, u_2, ..., u_m \in \mathbb{R}^n$ 的 Kruskal 稀疏秩 k-rank 是最大的 k，使得 $every$ 集合的 k 个向量线性无关。

在这个问题中，我们将探讨Khatr -Rao积的性质，并利用它来设计分解高阶张量的算法。

(a) 设k_u和k_v分别是$u_1, u_2, ..., u_m$和$v_1, v_2, ..., v_m$的k-秩。证明 $u_1 \odot v_1, u_2 \odot v_2, ..., u_m \odot v_m$的k-秩至少为min($k_u + k_v - 1, m$)。

(b) 构造一个例子族，其中 $u_1 \odot u_1, u_2 \odot u_2, ..., u_m \odot u_m$ 的 k-秩恰好为 $2k_u - 1$，而不是更大。为了使这个问题非平凡，你必须使用一个 $m > 2k_u - 1$ 的例子。

(c) 给定一个 $n \times n \times n \times n \times n$ 五阶张量 $T = \sum_{i=1}^{r} a_i^{\otimes 5}$，在适当条件下给出一个适用于 $r = 2n - 1$ 的寻找其因子的算法

在因素 $a_1, a_2, ..., a_r$ 上。*Hint:* 降至三阶情况。

实际上，对于随机或扰动的向量，Khatri-Rao积对它们的Kruskal秩有更强的$multiplying$效应。这些类型的性质可以用来获得在r是n的某个多项式的高度过完备情况下的高阶张量分解算法。

Problem 4-2: 在4.4节中，我们看到了如何使用非凸优化来解决独立成分分析。在本问题中，我们将看到如何使用张量分解来解决它。假设我们观察到许多形式为$y = Ax$的样本，其中A是一个未知的非奇异方阵，并且x的每个坐标都是独立的，并满足$\mathbb{E}[x_j] = 0$和$\mathbb{E}[x_j^4] \neq 3\,\mathbb{E}[x_j^2]^2$。$x_j$的分布是未知的，并且可能对于所有$j$并不相同。

(a) 用 A 和 x 的矩表示 $\mathbb{E}[y^{\otimes 4}]$ 和 $(\mathbb{E}[y^{\otimes 2}])^{\otimes 2}$ 的表达式。（期望中不应包含任何 A 。）

(b) 使用(a)部分，展示如何使用y的矩来生成形式为$\sum_j c_j a_j^{\otimes 4}$的张量，其中$a_j$表示$A$的第$j$列，$c_j$是非零标量。

(c) 展示如何使用 Jennrich 算法恢复 A (的列，直到排列和标量倍数)。

Chapter 5

Sparse Recovery

在这一章中，我们将首次见证 *sparsity* 的力量。让我们了解一下它有什么好处。考虑求解一个欠定线性系统 $Ax = b$ 的问题。如果我们给定 A 和 b，就没有机会唯一地恢复 x，对吧？好吧，如果我们知道 x 是稀疏的，情况就不同了。在这种情况下，A 上存在自然条件，我们实际上将能够恢复 x，即使 A 的行数与 x 的稀疏性相当，而不是其维度。在这里，我们将介绍稀疏恢复的理论。如果你好奇的话，这是一个不仅有一些理论瑰宝，而且对实践也有重大影响的领域。

5.1 Introduction

在信号处理（尤其是成像）中，我们经常面临给定信号的线性测量值来恢复某些未知信号的任务。让我们固定我们的符号。在本章中，我们将关注解决一个线性系统 $Ax = b$

在 A 是一个 $m \times n$ 矩阵，且 x 和 b 分别是 n 和 m 维向量的情况下。在我们的设置中，A 和 b 都是已知的。您可以将 A 视为我们所使用的某些测量设备的输入-输出功能表示。

现在如果 $m < n$，那么我们无法期望唯一地恢复 x。最多我们只能找到一些满足 $Ay = b$ 的解 y，并且我们有保证 $x = y + z$，其中 z 属于 A 的核。这告诉我们，如果我们想恢复一个 n 维的信号，我们至少需要 n 个线性测量。这是相当自然的。有时你会听到这被称为香农-奈奎斯特速率，尽管我发现这是一种相当晦涩的方式来描述正在发生的事情。将我们拯救的惊人想法是，如果 x 是稀疏的——即 b 是 A 的少数几列的线性组合——那么我们真的可以用更少的线性测量来完成任务，并且仍然能够精确地重建 x。

本节我想解释为什么你实际上不应该对此感到惊讶。如果你忽略算法（我们稍后不会这样做），这实际上相当简单。结果是，仅假设 x 是稀疏的本身是不够的。我们始终必须对 A 也做出一些结构假设。让我们考虑以下概念：

Definition 5.1.1 *The* Kruskal等级 *of a set of vectors* $\{A_i\}_i$ *is the maximum* r *such that all subsets of at most* r *vectors are linearly independent.*

如果你给定一个 n 维的 n 个向量集合，它们可以全部线性无关，在这种情况下，它们的 Kruskal 等价秩是 n。但是，如果你有 m 维的 n 个向量——就像我们取我们的感知矩阵 A 的列一样——并且 m 小于 n，向量不能全部线性无关，但它们仍然可以有 Kruskal 等价秩 m。事实上，这是常见的情况：

Claim 5.1.2 *If $A_1, A_2 \ldots A_n$ are chosen uniformly at random from $\mathbb{S}^{m-1}$ then almost surely their Kruskal rank is m.*

现在让我们证明关于稀疏恢复的第一个主要结果。设$\|x\|_0$为x中非零项的数量。我们将对以下高度非凸优化问题感兴趣:

$$(P_0) \min \|w\|_0 \text{ s.t. } Aw = b$$

让我们证明如果我们能解决(P_0)，我们就能从比n少得多的线性测量中找到x:

Lemma 5.1.3 *Let A be an $m \times n$ matrix whose columns have Kruskal rank at least r. Let x be an $r/2$-sparse vector and let $Ax = b$. Then the unique optimal solution to (P_0) is x.*

Proof: 我们知道 x 是 $Ax = b$ 的一个解，其目标值为 $\|x\|_0 = r/\text{^}2$。现在假设存在另一个满足 $Ay = b$ 的解 y。

考虑这些解之间的差异 $z = x - y$。我们知道 z 在 A 的核中。然而 $\|z\|_0 \geq r + 1$ 因为根据假设，A 中最多 r 列的每一组都是线性无关的。最后我们有

$$\|y\|_0 \geq \|z\|_0 - \|x\|_0 \geq r/2 + 1$$

这表明 y 的目标值大于 x。这就完成了证明。■

因此，如果我们选择我们的感知矩阵的列是随机的 m 维

向量，然后从仅 m 线性测量中，原则上可以唯一恢复任何 $m/2$- 稀疏向量。但有一个巨大的陷阱。求解 (P_0) —— 即找到线性方程组的稀疏解是 NP-难的。事实上，这是一个简单且重要的简化，值得一看。遵循Khachiyan [97]，让我们从子集和问题开始，这是一个标准的 NP-难问题：

Problem 1 *Given distinct values $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$, does there exist a set $I \subseteq [n]$ so that $|I| = m$ and $\sum_{i \in I} \alpha_i = 0$?*

我们将此问题的实例嵌入到在给定子空间中寻找最稀疏非零向量的问题中。我们将使用以下映射，称为 *weird moment curve*：

$$\Gamma'(\alpha_i) = [1, \alpha_i, \alpha_i^2, \ldots, \alpha_i^{m-2}, \alpha_i^m]$$

这个与标准矩曲线之间的区别在于最后一个项，我们使用 α_i^m 而不是 α_i^{m-1}。

Lemma 5.1.4 *A set I with $|I| = m$ has $\sum_{i \in I} \alpha_i = 0$ if and only if the vectors $\{\Gamma'(\alpha_i)\}_{i \in I}$ are linearly dependent.*

Proof: 考虑由列向量 $\{\Gamma'(\alpha_i)\}_{i \in I}$ 构成的矩阵的行列式。然后证明基于以下观察：

(a) 行列式是关于变量 α_i 的多项式，总次数为 $\binom{m}{2} + 1$，这可以通过将行列式写成其拉普拉斯展开式（例如参见[88]）来看到。

(b) 此外，行列式可被 $\prod_{i<j} \alpha_i - \alpha_j$ 整除，因为如果任何 $\alpha_i = \alpha_j$，行列式为零。

因此我们可以将行列式写为

$$\Big(\prod_{\substack{i<j \\ i,j\in I}} (\alpha_i - \alpha_j)\Big)\Big(\sum_{i\in I} \alpha_i\Big)$$

我们假设 α_i 是不同的，因此当且仅当 $\alpha_i =$ 之和为零时，行列式为零。■

我们现在可以证明一个双重打击。不仅解决(P_0) NP是困难的，计算Kruskal秩也是困难的：

Theorem 5.1.5 *Both computing the Kruskal rank and finding the sparsest solution to a system of linear equations are NP-hard.*

Proof: 首先让我们证明计算Kruskal秩是NP-难。考虑向量$\{\Gamma'(\alpha_i)\}_i$。根据引理5.1.4，如果存在一个集合I，它满足$|I| = m$，则$\sum_{i\in I} \alpha_i = 0$，那么$\{\Gamma'(\alpha_i)\}_i$的Kruskal秩至多为$m-1$，否则它们的Kruskal秩正好是$m$。由于子集和是$NP$-难，因此判断Kruskal秩是否为$m$或至多为$m-1$也是$NP$-难。

现在让我们继续展示找到线性系统的最稀疏解是 NP-难的问题。我们将使用一对一多映射。对于每个 j，考虑以下优化问题：

$$(P_j) \qquad \min \|w\|_0 \text{ s.t. } \Big[\Gamma'(\alpha_1), \ldots, \Gamma'(\alpha_{j-1}), \Gamma'(\alpha_{j+1}), \ldots, \Gamma'(\alpha_n)\Big] w = \Gamma'(\alpha_j)$$

可以看出，$\{\Gamma'(\alpha_i)\}_i$的Kruskal秩至多为$m-1$，当且仅当存在某个j，使得(P_j)有解，其目标值至多为$m-2$。因此(P_0)也是NP-难。■

在本章的其余部分，我们将专注于算法。我们将给出简单的贪婪方法以及基于凸规划松弛的方法。这些算法将在比仅其列具有大Kruskal秩更严格的假设下工作于感知矩阵A。尽管如此，我们所做的所有假设都将由随机选择的A以及许多其他矩阵满足。我们给出的算法甚至还会提供在噪声存在的情况下有意义的更强保证。

5.2 Incoherence and Uncertainty Principles

在1965年，Logan [107] 发现了一个显著现象。如果你在稀疏的一组位置上对一个带限信号进行破坏，就有可能唯一地恢复原始信号。这实际上是一个伪装的稀疏恢复问题。让我们将其形式化：

Example 1 *The spikes-and-sines matrix A is a $n \times 2n$ matrix*

$$A = [I, D]$$

where I is the identity matrix and D is the discrete Fourier transform matrix, i.e.

$$D_{a,b} = \frac{\omega^{(a-1)(b-1)}}{\sqrt{n}}$$

and $\omega = e^{2\pi i/n}$ *is the* n^{th} *root of unity.*

设 x 为一个稀疏的 $2n$-维向量。第一个 n 坐标中的非零值表示损坏的位置。最后一个 n 坐标中的非零值表示原始信号中存在的频率。因此，我们知道 A 和 b ，并且有保证存在一个解 x 来解决 $Ax = b$，其中 x 是稀疏的。直到 Donoho 和 Stark [64] 的工作数年后，他们意识到这种现象并不仅限于尖峰和正弦矩阵。实际上，这是一个相当普遍的现象。关键是不可协调性的概念:

Definition 5.2.1 *The columns of* $A \in \mathbb{R}^{n\times m}$ *are* μ- 不连贯 *if for all* $i \neq j$*:*

$$|\langle A_i, A_j\rangle| \leq \mu\|A_i\| \cdot \|A_j\|$$

在整个本节中，我们将仅关注当 A 的列是单位向量的情况。因此，一个矩阵是 μ 非相干的，如果对于所有 $i \neq j$，$|\langle A_i, A_j\rangle| \leq \mu$。然而，我们在这里推导的所有结果都可以扩展到一般 A，当列不一定是单位向量时。就像我们对于 Kruskal秩所做的那样，让我们证明随机向量是非相干的:

Claim 5.2.2 *If* $A_1, A_2 \ldots A_m$ *are chosen uniformly at random from* $\mathbb{S}^{n-1}$ *then with high probability they will be* μ*-incoherent for*

$$\mu = O\Big(\sqrt{\frac{\log m}{n}}\Big).$$

您也可以检查尖峰-正弦矩阵与$\mu = 1/\sqrt{n}$是μ非相干的。这样，我们在这里得出的结果将包含Logan现象

一个特殊情况。无论如何，现在让我们证明如果 A 是不连贯的，并且如果 x 足够稀疏，那么它将是 $Ax = b$ 的唯一最稀疏解。

Lemma 5.2.3 *Let A be an $n \times m$ matrix that is μ-incoherent and whose columns are unit norm. If $Ax = b$ and $\|x\|_0 < \frac{1}{2\mu}$, then x is the uniquely sparsest solution to the linear system.*

Proof: 假设为了矛盾的目的，我们还有另一个满足 $Ay = b$ 和 $\|y\|_0 < \frac{1}{2\mu}$ 的解 y。然后我们可以考虑这些解之间的差 $z = x - y$，它满足 $\|z\|_0 < \frac{1}{\mu}$，并考虑以下表达式

$$z^T A^T A z = 0.$$

如果我们让 S 表示 z 的支撑集——即它非零的位置——那么 $A^T A$ 限制在 S 中的行和列是奇异的。令这个矩阵为 B。那么 B 的对角线元素为 1，而对角线外的元素绝对值被 μ 所限制。但是根据 Gershgorin 圆盘定理，我们知道 B 的所有特征值都包含在复平面上以 1 为中心、半径为 $\mu|S| <$ 的圆盘内。因此 B 是非奇异的，我们得到了一个矛盾。■

实际上，当 A 是两个正交归一基的并集时，我们可以证明一个更强的唯一性结果，正如 spikes-and-sines 矩阵的情况。让我们首先证明以下结果，我们将神秘地称之为不确定性原理：

Lemma 5.2.4 *Let $A = [U, V]$ be a $n \times 2n$ matrix that is μ-incoherent where U and V are $n \times n$ orthogonal matrices. If $b = U\alpha = V\beta$, then $\|\alpha\|_0 + \|\beta\|_0 \geq \frac{2}{\mu}$.*

Proof: 由于 U 和 V 是正交归一的，因此我们有 $\|b\|_2 = \|\alpha\|_2 = \|\beta\|_2$。我们可以将 b 重写为 $U\alpha$ 或 $V\beta$，从而得到 $\|b\|_2^2 = |\beta^T(V^TU)\alpha|$。因为 A 是非相干的，所以我们可以得出结论，V^TU 的每个元素绝对值不超过 $\mu(A)$，因此 $|\beta^T(V^TU)\alpha| \leq \mu(A)\|\alpha\|_1\|\beta\|_1$。使用柯西-施瓦茨不等式，可以得出 $\|\alpha\|_1 \leq \sqrt{\|\alpha\|_0}\|\alpha\|_2$，从而

$$\|b\|_2^2 \leq \mu(A)\sqrt{\|\alpha\|_0\|\beta\|_0}\|\alpha\|_2\|\beta\|_2$$

重新排列，我们得到 $\frac{1}{\mu(A)} \leq \sqrt{\|\alpha\|_0\|\beta\|_0}$。最后，应用算术平均数-几何平均数不等式，我们得到 $\frac{2}{\mu} \leq \|\alpha\|_0 + \|\beta\|_0$，这完成了证明。■

这个证明简短且简单。可能唯一令人困惑的部分是为什么我们称之为不确定性原理。让我们通过应用引理5.2.4来阐明这一点。如果我们把A设为尖峰和正弦矩阵，我们得到任何非零信号在标准基或傅里叶基中至少有$\sqrt{n}$个非零值。这意味着没有任何信号可以在时间和频率域同时稀疏！值得退一步思考。如果我们只是证明了这个结果，你自然会将其与海森堡不确定性原理联系起来。但事实是，真正驱动它的是我们信号的时频基的不相干性，它同样适用于许多其他基的对。

让我们用我们的不确定性原理来证明一个更强的唯一性结果：

Claim 5.2.5 *Let $A = [U, V]$ be a $n \times 2n$ matrix that is μ-incoherent where U and V are $n \times n$ orthogonal matrices. If $Ax = b$ and $\|x\|_0 < \frac{1}{\mu}$, then x is the uniquely sparsest solution to the linear system.*

Proof: 考虑任何替代解 $A\widetilde{x} = b$。设定 $y = x - \widetilde{x}$，在这种情况下 $y \in \ker(A)$。将 y 写作 $y = [\alpha_y, \beta_y]^T$，由于 $Ay = 0$，因此我们有 $U\alpha_y = -V\beta_y$。现在我们可以应用不确定性原理并得出结论 $\|y\|_0 = \|\alpha_y\|_0 + \|\beta_y\|_0 \geq \frac{2}{\mu}$。容易看出 $\|\widetilde{x}\|_0 \geq \|y\|_0 - \|x\|_0 > \frac{1}{\mu}$，因此 $\widetilde{x}$ 的非零项比 x 多，这就完成了证明。■

我们可以将不一致性追溯到我们关于Kruskal秩的原始讨论。结果证明，具有列不一致的矩阵只是证明Kruskal秩下界的一种简单易检查的方法。以下命题的证明与引理5.2.3的证明基本相同。我们将它留给读者作为练习。

Claim 5.2.6 *If A is μ-incoherent then the Kruskal rank of the columns of A is at least $1/\mu$.*

在下一节中，我们将给出一个简单的贪婪算法，用于解决在非相干矩阵上的稀疏恢复问题。该算法将如何证明它在不断取得进展，并找到x的正确非零位置，将围绕我们刚刚证明的唯一性结果背后的相同思想展开。

5.3 Pursuit Algorithms

存在一类重要的稀疏恢复算法，称为追踪算法。这些算法是贪婪和迭代的。它们与不相关矩阵一起工作，寻找在 {v*} 中解释观察到的向量 {v*} 的列

他们减去一个多倍列并继续处理余数。第一个这样的算法是在Mallat和Zhang的一篇有影响力的论文[111]中引入的，被称为匹配追踪。在本节中，我们将分析其一个变种，称为*orthogonal matching pursuit*。后者特别方便的是，该算法将保持一个不变量，即余数与到目前为止我们已选择的A的所有列正交。这在每一步中更昂贵，但更容易分析和理解其背后的直觉。

在整个本节中，设 A 为一个 $n \times m$ 矩阵，它是 μ-非相干的。设 x 是 k-稀疏的，具有 $k < 1/(2\mu)$，并设 $Ax = b$。最后，我们将使用 T 来表示 x 的支撑集——即 x 中非零元素的位置。现在让我们正式定义正交匹配追踪：

Orthogonal Matching Pursuit

Input: matrix $A \in \mathbb{R}^{n \times m}$, vector $b \in \mathbb{R}^n$, desired number of nonzero entries $k \in \mathbb{N}$.

Output: solution x with at most k nonzero entries.

Initialize: $x^0 = 0$, $r^0 = Ax^0 - b$, $S = \emptyset$.

For $\ell = 1, 2, \ldots, k$

Choose column j that maximizes $\frac{|\langle A_j, r^{\ell-1} \rangle|}{\|A_j\|_2^2}$.

Add j to S.

Set $r^\ell = \text{proj}_{U^\perp}(b)$, where $U = \text{span}(A_S)$.

If $r^\ell = 0$, break.

End

Solve for x_S: $A_S x_S = b$. Set $x_{\bar{S}} = 0$.

我们的分析将专注于建立以下两个不变量：

(a) 算法选择的每个索引 j 都在 T 中。

(b) 每个索引 j 最多选择一次。

这两个属性立即意味着正交匹配追踪可以恢复真实解 x，因为残差误差 r^ℓ 将在 $S = T$ 之前不为零，而且线性系统 $A_T x_T = b$ 有唯一解（这是我们从前一节中知道的）。

属性 (b) 是直接的，因为一旦在算法的每一步都有 $j \in S$，我们就会得到 $r^\ell \perp U$，其中 $U = \text{span}(A_S)$，所以 $\langle r^\ell, A_j \rangle =$ 为 0 如果 $j \in S$。我们的主要目标是建立属性 (a)，我们将通过归纳法来证明。主要引理是：

Lemma 5.3.1 *If $S \subseteq T$ at the start of a stage, then orthogonal matching pursuit selects $j \in T$.*

我们将首先证明一个辅助引理：

Lemma 5.3.2 *If $r^{\ell-1}$ is supported in T at the start of a stage, then orthogonal matching pursuit selects $j \in T$.*

Proof: 设 $r^{\ell-1} = \sum_{i \in T} y_i A_i$。然后我们可以重新排列 A 的列，使得前 k' 列对应于 y 的 k' 个非零项，按大小顺序递减：

$$\underbrace{|y_1| \geq |y_2| \geq \cdots \geq |y_{k'}| > 0}_{\text{corresponds to first } k' \text{ columns of } A}, \quad |y_{k'+1}| = 0, |y_{k'+2}| = 0, \ldots, |y_m| = 0.$$

在 $k' \leq k$。因此，$\text{supp}(y) = \{1, 2, \ldots, k'\} \subseteq T$。然后，为了确保我们选择 $j \in T$，一个充分条件是

$$|\langle A_1, r^{\ell-1}\rangle| > |\langle A_i, r^{\ell-1}\rangle| \qquad \text{for all } i \geq k'+1. \tag{5.1}$$

我们可以降低(5.1)的左侧界限：

$$|\langle r^{\ell-1}, A_1\rangle| = \left|\left\langle \sum_{\ell=1}^{k'} y_\ell A_\ell, A_1 \right\rangle\right| \geq |y_1| - \sum_{\ell=2}^{k'} |y_\ell||\langle A_\ell, A_1\rangle| \geq |y_1| - |y_1|(k'-1)\mu \geq |y_1|(1-k'\mu+\mu),$$

该假设下，$k' \leq k < 1/(2\mu)$严格下界于 $|y_1|(1/2+\mu)$。

我们可以对(5.1)的右侧进行上界估计：

$$|\langle r^{\ell-1}, A_i\rangle| = \left|\left\langle \sum_{\ell=1}^{k'} y_\ell A_\ell, A_i \right\rangle\right| \leq |y_1| \sum_{\ell=1}^{k'} |\langle A_\ell, A_i\rangle| \leq |y_1| k'\mu,$$

该假设下，$k' \leq k < 1/(2\mu)$严格上界于$|y_1|/2$。由于$|y_1|(1/2+\mu) > |y_1|/2$，我们得出结论，条件（5.1）成立，保证算法选择$j \in T$，从而完成证明。■

现在我们可以证明引理5.3.1：

Proof: 假设阶段开始时的 $S \subseteq T$。然后残差 $r^{\ell-1}$ 在 T 中得到支持，因为我们可以将其写成

$$r^{\ell-1} = b - \sum_{i \in S} z_i A_i, \text{ where } z = \arg\min \|b - A_S z_S\|_2$$

应用上述引理，我们得出结论，该算法选择 $j \in T$. ■

这通过归纳建立了属性 (a)，并完成了正交匹配追踪正确性的证明，我们下面总结如下：

Theorem 5.3.3 *Let A be an $n \times m$ matrix that is μ-incoherent and whose columns are unit norm. If $Ax = b$ and $\|x\|_0 < \frac{1}{2\mu}$, then the output of orthogonal matching pursuit is exactly x.*

注意，此算法在恰好达到我们建立唯一性的阈值处工作。然而，在A =[U, V]和U以及V正交的情况下，我们证明了比常数因子更好的唯一性结果。虽然存在比上述算法更好的算法（例如，参见[67]），但尚无已知算法能匹配那里已知的最佳唯一性界限。

它还值得提及其他追求算法的不同之处。例如，在匹配追求中，我们不会在每个阶段结束时重新计算 x_i 对于 $i \in S$ 的系数。我们只保留它们当前的设置，并希望当我们向 S 中添加新索引 j 时，这些系数不需要调整太多。这就是匹配追求在实践中更快的原因，然而分析更为繁琐，因为我们需要跟踪由于没有将 b 投影到我们迄今为止选择的列的正交补中而产生的误差是如何累积的。

5.4 Prony's Method

广泛存在一种误解，认为稀疏恢复算法是现代发明。实际上，稀疏恢复可以追溯到1795年，一个称为Prony方法的算法。它将给我们几乎所有我们想要的东西。我们将有一个明确的$2k \times n$

感知矩阵 A，对于该矩阵我们能够精确地并且使用高效算法恢复任何 k-稀疏信号。它还有这样的好处，即我们可以使用快速傅里叶变换在 $O(n \log n)$ 时间内计算矩阵-向量乘积 Ax。

此方法的缺点是它非常不稳定，因为它涉及到对Vandermonde矩阵进行求逆，而Vandermonde矩阵可能非常病态。所以当你听到压缩感知可以突破香农-奈奎斯特极限时，你应该记住Prony的方法已经做到了这一点。我们将要研究的算法之所以与众不同，是因为它们在存在噪声的情况下也能工作。这就是使它们在实践上如此相关的关键方面。尽管如此，Prony的方法在理论方面非常有用，并且你可以从中得到的结果往往在其他名称下被重新发现。

Properties of the Discrete Fourier Transform

普诺伊方法将关键地利用离散傅里叶变换的各种性质。回想一下，作为一个矩阵，这种变换的项为

$$F_{a,b} = \left(\frac{1}{\sqrt{n}}\right) \exp\left(\frac{i2\pi(a-1)(b-1)}{n}\right)$$

与我们之前所做的一样，我们将简化符号，用 $\omega = e^{i2\pi/n}$ 表示单位根 n^{th}。使用这个符号，行 a 列 b 的项是 $\omega^{(a-1)(b-1)}$。

矩阵 F 具有许多重要性质，包括:

(a) F 是正交归一：$F^H F = FF^H$，其中 F^H 是 F 的厄米共轭转置

(b{v*}对卷积算子进行对角化

我们还没有定义卷积，所以现在让我们来定义它。实际上，让我们通过其相应的线性变换来做这件事：

Definition 5.4.1 (Circulant matrix) *For a vector $c = [c1, c_2, \ldots, c_n]$, let*

$$M^c = \begin{bmatrix} c_n & c_{n-1} & c_{n-2} & \ldots & c_1 \\ c_1 & c_n & c_{n-1} & \ldots & c_2 \\ \vdots & & & & \vdots \\ c_{n-1} & \ldots & \ldots & \ldots & c_n \end{bmatrix}.$$

然后矩阵向量乘积 $M^c x$ 是通过卷积 c 和 x 得到的向量，我们将它表示为 $c * x$。直观上，如果你将 c 和 x 视为表示离散随机变量的概率分布，那么 $c * x$ 表示通过将这两个概率分布相加并在 n 上使用模运算进行环绕得到的随机变量的分布。

我们如上所述，可以使用 F 对 M^c 进行对角化。更正式地说，我们有以下事实，我们将直接使用而不加证明：

Claim 5.4.2 *$M^c = F^H diag(\widehat{c}) F$, where $\widehat{c} = Fc$.*

这告诉我们，我们可以将卷积视为在傅里叶表示中按坐标乘法来考虑。更精确地说：

Corollary 5.4.3 *Let $z = c * x$; then $\widehat{z} = \widehat{c} \odot \widehat{x}$, where $\odot$ indicates coordinate-wise multiplication.*

Proof: 我们可以写出 $z = M^c x = F^H \text{diag}(\widehat{c}) F x = F^H \text{diag}(\widehat{c}) \widehat{x} = F^H(\widehat{c} \odot \widehat{x})$，这就完成了证明。■

The Helper Polynomial

Prony的方法围绕以下辅助多项式展开:

Definition 5.4.4 (Helper polynomial)

$$\begin{aligned} p(z) &= \prod_{b \in \text{supp}(x)} \omega^{-b}(\omega^b - z) \\ &= 1 + \lambda_1 z + \ldots + \lambda_k z^k. \end{aligned}$$

Claim 5.4.5 *If we know $p(z)$, we can find $supp(x)$.*

Proof: 实际上，一个索引 b 在 x 的支持中当且仅当 $p(\omega^b) = 0$。因此，我们可以在 ω 的幂上评估 p，而 p 评估为非零的指数正好是 x 的支持。■

Prony方法的基本思想是使用离散傅里叶变换的前$2k$个值来找到p，从而确定x的支持。然后我们可以解一个线性系统来实际找到x的值。我们的第一个目标是找到辅助多项式。让

$$v = [1, \lambda_1, \lambda_2, \ldots, \lambda_k, 0, \ldots, 0], \text{ and } \widehat{v} = Fv$$

可以看出，索引 $b+1$ 处的 $\widehat{v}$ 的值正好是 $p(\omega^b)$。

Claim 5.4.6 $supp(\widehat{v}) = \overline{supp(x)}$

即，$\widehat{v}$ 的零对应于 p 的根，因此是 x 的非零元素。反之，$\widehat{v}$ 的非零元素对应于 x 的零。我们得出结论 $x \odot \widehat{v} = = 0$，因此:

Corollary 5.4.7 $M^{\widehat{x}}v = 0$

Proof: 我们可以应用5.4.2主张来重写$x \odot \widehat{v}$ =0为$\widehat{x} * v = \widehat{\ }0$ =0，并且这蕴含该推论。■

让我们显式地写出这个线性系统:

$$M^{\widehat{x}} = \begin{bmatrix} \widehat{x}_n & \widehat{x}_{n-1} & \dots & \widehat{x}_{n-k} & \dots & \widehat{x}_1 \\ \widehat{x}_1 & \widehat{x}_n & \dots & \widehat{x}_{n-k+1} & \dots & \widehat{x}_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \widehat{x}_{k+1} & \widehat{x}_k & \dots & \widehat{x}_1 & \dots & \widehat{x}_{k+2} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \widehat{x}_{2k} & \widehat{x}_{2k-1} & \dots & \widehat{x}_k & \dots & \widehat{x}_{2k+1} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{bmatrix}$$

回忆，我们无法访问这个矩阵的所有条目，因为我们只知道 k 的前 $2k$ 个 DFT 的 x 值。然而，考虑一个 $k \times k+1$ 子矩阵，其左上角值为 $\widehat{x}_{k+1}$，右下角值为 $\widehat{x}_k$。这个矩阵只涉及我们知道的值!

考虑

$$\begin{bmatrix} \widehat{x}_k & \widehat{x}_{k-1} & \dots & \widehat{x}_1 \\ \vdots & & & \\ \widehat{x}_{2k-1} & \widehat{x}_{2k-1} & \dots & \widehat{x}_k \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_k \end{bmatrix} = - \begin{bmatrix} \widehat{x}_{k+1} \\ \vdots \\ \vdots \\ \widehat{x}_{2k} \end{bmatrix}$$

结果证明这个线性系统是满秩的，因此 λ 是该方程组的唯一解

线性系统（证明留给读者[1]）。λ中的项是p的系数，因此一旦我们解出λ，我们就可以在ω^b上评估辅助多项式以找到x的支持。剩下的只是找到x的值。实际上，让M是F对S列及其前$2k$行的限制。M是一个Vandermonde矩阵，因此$Mx_S = \widehat{x}_{1,2,\ldots 2k}$有唯一解，我们可以解这个线性系统以找到$x$的非零值。

Prony方法的保证总结在以下定理中：

Theorem 5.4.8 *Let A be the $2k \times n$ matrix obtained from taking the first $2k$ rows of the discrete Fourier transform matrix F. Then for any k-sparse signal x, Prony's method recovers x exactly from Ax.*

在您好奇的情况下，这又是稀疏恢复中的一个主题，我们可以将其与Kruskal秩联系起来。很容易证明A的列具有等于$2k$的Kruskal秩。事实上，无论我们选择F中的哪$2k$行，这都是正确的。此外，结果发现，在某些设置下，Prony方法及相关方法可以在噪声存在的情况下被证明是有效的，但仅在某些关于x中非零位置的分离条件下。有关更多详细信息，请参阅Moitra [113]。

5.5 Compressed Sensing

在这个部分，我们将介绍关于我们的感知矩阵 A 的一项强大新假设，称为限制等距性质。你可以将其视为Kruskal秩的鲁棒模拟，我们不仅希望（比如说）A 的每一组 $2k$ 列都是线性无关的，还希望它们具有良好的条件数。我们将展示

这是一个简单的凸规划松弛非常有效。对于A的良好选择，我们将能够从$O(k \log(n/k))$线性测量中恢复一个k稀疏信号。该算法在多项式时间内运行，并且它对噪声具有鲁棒性，即即使x不是k稀疏，我们仍然能够近似恢复其k最大的坐标。这是一种更强的保证。毕竟，自然信号不是k稀疏。但能够恢复它们的k最大坐标通常已经足够了。

现在让我们定义限制等距性质：

Definition 5.5.1 *A matrix A satisfies the (k, δ)-restricted isometry property if for all k-sparse vectors x we have:*

$$(1 - \delta)\|x\|_2^2 \leq \|Ax\|_2^2 \leq (1 + \delta)\|x\|_2^2$$

与其他我们考虑的假设一样，限制等距性质在随机选择的感知矩阵中具有很高的概率成立：

Lemma 5.5.2 *Let A be an $m \times n$ matrix where each entry is an independent standard Gaussian random variable. Provided that $m \geq 10k \log n/k$ then with high probability A satisfies the $(k, 1/3)$-restricted isometry property.*

接下来，让我们将我们所说的“近似恢复 k 个最大坐标”的形式化。我们的目标将以以下函数的形式来表述：

Definition 5.5.3 $\gamma_k(x) = \min_{w \text{ s.t. } \|w\|_0 \leq k} \|x - w\|_1$

将此用更简单的话来说，$\gamma_k(x)$ 是除了 k 个最大幅度条目之外的所有条目的绝对值之和。如果 x 确实是 k-稀疏的，那么 $\gamma_k(x) =$ 为 0。

我们的目标是找到一个与任何 k-稀疏向量几乎一样好的 w 来逼近 x。更正式地说，我们希望找到一个满足 $\|x-w\|_1 \leq C\gamma_k(x)$ 的 w，并且我们希望尽可能少地使用线性测量来实现这一点。这个学习目标已经包含了我们之前章节中的其他精确恢复结果，因为当 x 是 k-稀疏时，正如我们讨论的，$\gamma_k(x)$ 是零，所以我们别无选择，只能恢复 $w = x$。

在这个部分，我们的方法将基于凸规划松弛。而不是试图解决 NP-难优化问题 (P_0)，我们将考虑现在著名的 ℓ_1-松弛：

$$(P_1) \qquad \min \|w\|_1 \text{ s.t. } Aw = b$$

首先陈述一些关于使用(P_1)进行稀疏恢复的已知结果：

Theorem 5.5.4 *[43] If* $\delta_{2k} + \delta_{3k} < 1$ *then if* $\|x\|_0 \leq k$ *we have* $w = x$.

Theorem 5.5.5 *[42] If* $\delta_{3k} + 3\delta_{4k} < 2$ *then*

$$\|x - w\|_2 \leq \frac{C}{\sqrt{k}}\gamma_k(x)$$

以上保证与其他的略有不同（并且通常更强），因为界限是以误差 ℓ_2 的 $x - w$ 范数来表示的。

Theorem 5.5.6 *[51] If* $\delta_{2_k} < 1/3$ *then*

$$\|x - w\|_1 \leq \frac{2 + 2\delta_{2k}}{1 - 3\delta_{2k}} \gamma_k(x)$$

我们不会证明这些结果的确切性。但我们将按照Kashin和Temlyakov[96]的方法证明类似的结果，这种方法（在我看来）极大地简化了这些分析。但在我们分析(P_1)之前，我们需要引入一个来自泛函分析的概念，称为几乎欧几里得子集。

Almost Euclidean Subsections

非正式地说，一个几乎欧几里得子空间是一个子空间，其中在缩放后，ℓ_1和ℓ_2范数几乎等价。我们将只断言随机子空间几乎以高概率是几乎欧几里得子空间。相反，我们将大部分时间用于建立关于它们的各种几何性质，这些性质在我们回到压缩感知时将使用。关键的定义如下：

Definition 5.5.7 *A subspace* $\Gamma \subseteq \mathbb{R}^n$ *is a* C*-almost Euclidean subsection if for all* $v \in \Gamma$*,*

$$\frac{1}{\sqrt{n}} \|v\|_1 \leq \|v\|_2 \leq \frac{C}{\sqrt{n}} \|v\|_1$$

实际上，第一个不等式是平凡的。对于任何向量，总有$\frac{1}{\sqrt{n}}\|v\|_1 \leq \|v\|_2$ 成立。所有动作都发生在第二个不等式中。第一次看到它们时，并不明显存在这样的子空间。事实上，Garnaev 和 Gluskin [75] 证明了存在大量的几乎欧几里得子区段：

Theorem 5.5.8 *If Γ is a subspace chosen uniformly at random with $dim(\Gamma) = n-m$ then for*

$$C = O\Big(\sqrt{\frac{n}{m}\log\frac{n}{m}}\Big)$$

we have that Γ will be a C-almost Euclidean subsection with high probability.

让我们以一幅美好的图片结束，以便记住。考虑 ℓ_1 范数的单位球。有时它被称为交叉多面体，为了可视化它，你可以将其视为向量 $\{\pm e_i\}_i$ 的凸包，其中 e_i 是标准基向量。然后，一个子空间 Γ 几乎欧几里得，如果当我们将其与交叉多面体相交时，我们得到一个几乎球形的凸体。

Geometric Properties of Γ

这里我们将建立一些重要的C-近欧几里得子区段的几何性质。在本节中，设 $S = n/C^2$。首先我们证明Γ不能包含任何稀疏、非零向量:

Claim 5.5.9 *Let $v \in \Gamma$, then either $v = 0$ or $|supp(v)| \geq S$.*

Proof: 从柯西-施瓦茨不等式和 C-近欧几里得性质，我们有:

$$\|v\|_1 = \sum_{j\in \mathrm{supp}(v)} |v_j| \leq \sqrt{|\mathrm{supp}(v)|}\cdot\|v\|_2 \leq \sqrt{|\mathrm{supp}(v)|}\frac{C}{\sqrt{n}}\|v\|_1$$

证明现在通过重新排列项得出。■

它值得注意，这与线性纠错码有很好的类比，线性纠错码也是高维子空间（但是在 GF_2 上），我们希望每个

非零向量至少有常数分数的坐标是非零的。在任何情况下，让我们继续讨论一些

让我们继续探讨几乎欧几里得子区的一些更强的性质，这些性质与ℓ_1范数的分布情况有关。首先，我们给出一个有用的记号：

Definition 5.5.10 *For $\Lambda \subseteq [n]$, let v_Λ denote the restriction of v to coordinates in Λ. Similarly let v^Λ denote the restriction of v to $\bar{\Lambda}$.*

使用这个符号，让我们证明以下内容：

Claim 5.5.11 *Suppose $v \in \Gamma$ and $\Lambda \subseteq [n]$ and $|\Lambda| < S/16$. Then*

$$\|v_\Lambda\|_1 < \frac{\|v\|_1}{4}$$

Proof: 证明几乎与5.5.9声明的证明相同。再次使用柯西-施瓦茨不等式和C-近欧几里得性质，我们有：

$$\|v_\Lambda\|_1 \leq \sqrt{|\Lambda|}\|v_\Lambda\|_2 \leq \sqrt{|\Lambda|}\|v\|_2 \leq \sqrt{|\Lambda|}\frac{C}{\sqrt{n}}\|v\|_1$$

在插值术语中，完成了证明。■

现在我们拥有了所有需要的工具来给出关于(P_1)的第一个结果：

Lemma 5.5.12 *Let $w = x + v$ and $v \in \Gamma$ where $\|x\|_0 \leq S/16$. Then $\|w\|_1 > \|x\|_1$.*

Proof: 设置 $\Lambda = \text{supp}(x)$。然后

$$\|w\|_1 = \|(x+v)_\Lambda\|_1 + \|(x+v)^\Lambda\|_1 = \|(x+v)_\Lambda\|_1 + \|v^\Lambda\|_1$$

现在我们可以调用三角不等式:

$$\|w\|_1 \geq \|x_\Lambda\|1 - \|v_\Lambda\|1 + \|v^\Lambda\|1 = \|x\|1 - \|v_\Lambda\|1 + \|v^\Lambda\|1 = x_\Lambda\|1 - 2\|v_\Lambda\|_1 + \|v\|_1$$

然而 $\|v\|_1 - 2\|v_\Lambda\|_1 \geq \|v\|_1/2 > 0$ 使用权利要求 5.5.11。这完成了证明。

■

将定理5.5.8中的界限代入，我们已证明可以恢复一个维度为n的k-稀疏向量x

$$k \leq S/16 = \Omega(n/C^2) = \Omega\Big(\frac{m}{\log n/m}\Big)$$

从 m 线性测量中。

接下来我们将考虑稳定恢复。我们的主要定理是:

Theorem 5.5.13 *Let $\Gamma = ker(A)$ be a C-almost Euclidean subsection. Let $S = \frac{n}{C^2}$. If $Ax = Aw = b$ and $\|w\|_1 \leq \|x\|_1$ we have*

$$\|x - w\|_1 \leq 4\,\sigma_{\frac{S}{16}}(x)\,.$$

Proof: 让 $\Lambda \subseteq [n]$ 是最大化 $S/16$ 坐标的集合。我们希望上界 $\|x - w\|_1$。通过重复应用三角不等式,

$\|w\|_1 = \|w^\Lambda\|1 + \|w_\Lambda\|1 \le \|x\|1$ 以及 $\sigma_t(\cdot)$ 的定义，因此可以得出

$$\begin{aligned}\|x-w\|_1 &= \|(x-w)_\Lambda\|_1 + \|(x-w)^\Lambda\|_1 \\ &\le \|(x-w)_\Lambda\|_1 + \|x^\Lambda\|_1 + \|w^\Lambda\|_1 \\ &\le \|(x-w)_\Lambda\|_1 + \|x^\Lambda\|_1 + \|x\|_1 - \|w_\Lambda\|_1 \\ &\le 2\|(x-w)_\Lambda\|_1 + 2\|x^\Lambda\|_1 \\ &\le 2\|(x-w)_\Lambda\|_1 + 2\,\sigma_{\frac{S}{16}}(x)\,.\end{aligned}$$

自$(x-w) \in \Gamma$以来，我们可以应用5.5.11命题来得出结论$\|(x-w)_\Lambda\|_1 \le \frac{1}{4}\|x-w\|_1$。因此

$$\|x-w\|_1 \le \frac{1}{2}\|x-w\|_1 + 2\sigma_{\frac{S}{16}}(x)\,.$$

这完成了证明。■

Epilogue

最后，我们将以压缩感知中的一个主要开放问题结束，即给出满足限制等距性质的矩阵的 *deterministic* 构造：

Question 4 (Open) *Is there deterministic algorithm to construct a matrix with the restricted isometry property? Alternatively is there a deterministic algorithm to construct an almost Euclidean subsection* Γ*?*

Avi Wigderson喜欢将这些类型的问题称为“在干草堆里找草”。我们知道随机选择的A满足以下限制等距性质

高概率。它的核也是一个几乎欧几里得子空间，具有高概率。但我们能去除随机性吗？已知的最佳确定性构造归功于Guruswami、Lee和Razborov[82]：

Theorem 5.5.14 *[82] There is a polynomial time deterministic algorithm for constructing an almost Euclidean subspace* Γ *with parameter* $C \sim$ (对数 $n)^{\log\log\log n}$ 这一定太奇怪了，不可能是我们能做的最好的，对吧?

5.6 Exercises

Problem 5-1: 在这个问题中，我们将探讨稀疏恢复的唯一性条件，以及ℓ_1-最小化可证明有效的工作条件。

(a) 设 $A\widehat{x} = b$，并假设 A 有 n 列。进一步假设 $2k \leq m$。证明对于 *every* $\widehat{x}$，若 $\|\widehat{x}\|_0 \leq k$，则 $\widehat{x}$ 是线性系统的唯一最稀疏解，当且仅当 A 的列的 k-秩至少为 $2k$ 。

(b) 设 $U =$ 为 kernel(A)，且 $U \subset \mathbb{R}^n$。假设对于每个非零 $x \in U$，以及任何具有 $|S| \leq k$ 的集合 $S \subset [n]$,

$$\|x_S\|_1 < \frac{1}{2}\|x\|_1$$

x_S 表示将 x 限制在 S 的坐标上。证明如下：

$$(P1) \qquad \min \|x\|_1 \text{ s.t. } Ax = b$$

恢复 $x = \widehat{x}$，前提是 $A\widehat{x} = b$ 和 $\|\widehat{x}\|_0 \leq k$。

(c) **Challenge:** 你能构造一个维度为$\Omega(n)$的子空间 $U \subset \mathbb{R}^n$，使得每个非零 $x \in U$ 至少有 $\Omega(n)$ 个非零坐标吗？*Hint:* 使用一个扩张器。

Problem 5-2: 设 $\widehat{x}$ 是 k-稀疏向量，在 n-维度上。设 ω 是单位根的第 n 次根。假设我们给定 $v_\ell = \sum_{j=1}^{n} \widehat{x}_j \omega^{\ell j}$ 对于 $\ell = 0, 1, ..., 2k-1$。设 $A, B \in \mathbb{R}^{k \times k}$ 定义为 $A_{i,j} = v_{i+j-2}$ 和 $B_{i,j} = v_{i+j-1}$。

(a) 将 A 和 B 分别表示为 $A = VD_AV^T$ 和 $B = VD_BV^T$ 的形式，其中 V 是一个范德蒙德矩阵，D_A, D_B 是对角矩阵。

(b) 证明广义特征值问题 $Ax = \lambda Bx$ 的解可以用来恢复 $\widehat{x}$ 中非零元素的位置。

(c) 给定 $\widehat{x}$ 和 $v_0, v_1, ..., v_{k-1}$ 中非零元素的位置，给定一个算法来恢复 $\widehat{x}$ 中非零系数的值。

这被称为矩阵笔法。如果你眯着眼睛看，它看起来像Prony的方法（第5.4节）并且有类似的保证。两者在Vandermonde矩阵条件良好时（且仅在此条件下）对噪声都（某种程度上）稳健，而这种情况何时发生则是一个更长的故事。参见Moitra [113]。

Chapter 6

Sparse Coding

许多类型的信号在它们的自然基或手工设计的基（例如小波族）中都是稀疏的。但如果我们给定一组信号，而我们不知道它们稀疏的基，我们能否自动学习它？这个问题有各种名称，包括稀疏编码和字典学习。它在神经科学背景下被引入，用于解释神经元如何获得它们所具有的激活模式类型。它还应用于压缩和深度学习。在本章中，我们将给出利用凸规划松弛的稀疏编码算法以及迭代算法，我们将证明贪婪方法在适当的随机模型中成功最小化非凸函数。

6.1 Introduction

稀疏编码是由Olshausen和Field [117] 引入的，他们是神经科学家，对理解哺乳动物视觉皮层的特性感兴趣。他们能够测量神经元的感受野——即神经元对各种类型刺激的反应。但他们发现的结果让他们感到惊讶。反应模式总是:

(a) $spatially\ localized$ 表示每个神经元只对图像中特定区域的光敏感

(b) $bandpass$ 在添加高频成分对响应影响微乎其微的意义上

(c) $oriented$在该情况下，只有当边缘位于某些角度范围内时，旋转具有锐边的图像才会产生响应

令人惊讶的是，如果你收集了一组自然图像，并通过主成分分析找到一个k-维子空间将它们投影到该空间，你找到的方向将不具有这些属性。那么神经元是如何学习它们用来表示图像的基础的?

Olshausen和Field[117]提出的观点具有革命性。首先，神经元所使用的基的优点在于它们产生稀疏的激活模式。或者用我们的话说，它们在一个稀疏的基中表示自然图像的集合。其次，他们提出了存在自然和生物上合理的规则来学习这样的基。他们引入了一个简单的更新规则，即一起放电的神经元会加强它们彼此之间的连接。

其他。这被称为赫布学习规则。并且他们通过实验表明，当他们的迭代算法在自然图像上运行时，恢复了一个满足上述三个性质的基本集。
Thus algorithms can explain the emergence of certain biological properties of the visual cortex.

自那时起，稀疏编码和字典学习已成为信号处理和机器学习中的重要问题。我们假设我们被给定了由稀疏于公共基的示例 $b^{(1)}, b^{(2)}, \ldots, b^{(p)}$ 组成的集合。特别是有一个矩阵 A 和一组表示 $x^{(1)}, x^{(2)}, \ldots, x^{(p)}$，其中 $Ax^{(i)} = b^{(i)}$ 和每个 $x^{(i)}$ 都是稀疏的。让我们讨论两种流行的称为最优方向法和 k-SVD 的方法。

Method of Optimal Directions [68]

Input: Matrix B whose columns can be jointly sparsely represented

Output: A basis $\widehat{A}$ and representation $\widehat{X}$

Guess $\widehat{A}$

Repeat until convergence:

Given $\widehat{A}$, compute a column sparse $\widehat{X}$ so that $\widehat{A}\widehat{X} \approx B$ (using e.g. matching pursuit [111] or basis pursuit [50])

Given $\widehat{X}$, compute the $\widehat{A}$ that minimizes $\|\widehat{A}\widehat{X} - B\|_F$

End

为了简化我们的符号，我们将观测值 $b^{(i)}$ 组织成矩阵 B 的列，并将使用矩阵 $\widehat{X}$ 来表示我们的估计稀疏表示。另一种流行的方法如下：

K-SVD [5]

Input: Matrix B whose columns can be jointly sparsely represented

Output: A basis $\widehat{A}$ and representation $\widehat{X}$

Guess $\widehat{A}$

Repeat until convergence:

Given $\widehat{A}$, compute a column sparse $\widehat{X}$ so that $\widehat{A}\widehat{X} \approx B$ (using e.g. matching pursuit [111] or basis pursuit [50])

For each column $\widehat{A}_j$:

Group all samples $b^{(i)}$ where $\widehat{x}^{(i)}$ has a non-zero at index j. Subtract off components in the other directions

$$b^{(i)} - \sum_{j' \neq j} \widehat{A}_{j'} \widehat{x}_{j'}^{(i)}$$

Organize these vectors into a residual matrix and compute the top singular vector v and update the column $\widehat{A}_j$ to v

End

您应该将这些算法视为我们为非负矩阵分解给出的交替最小化算法的变体。它们遵循相同的风格

启发式。其区别在于k-SVD在更新时更巧妙地纠正了基$\widehat{A}$中其他列的贡献，这使得它在实践中成为首选的启发式方法。经验表明，这两种算法对其初始化都很敏感，但除了这个问题外，它们都工作得很好。

我们希望算法具有可证明的保证。因此，自然地关注于A是基且我们知道如何解决稀疏恢复问题的情形。因此，我们可以考虑A具有满列秩的欠完备情形，也可以考虑列数多于行数的过完备情形，其中A要么是不相干的，要么具有受限等距性质。这正是本章将要做的。我们还将假设一个关于$x^{(i)}$生成的随机模型，这有助于防止出现许多病理情况（例如，A中的一列永远不会被表示）。

6.2 The Undercomplete Case

在这个部分，我们将给出当A具有满列秩时的稀疏编码算法。我们的方法将基于凸规划松弛，以及我们在上一章中开发出的许多见解。我们将利用以下见解来找到我们的稀疏表示矩阵X：其行是样本矩阵B的行空间中稀疏度最高的向量。更正式地说，Spielman等人[131]的算法在以下自然生成模型下工作：

(a) 存在一个未知的字典 A，它是一个 $n \times m$ 矩阵并且具有满列秩(b) 每个样本 x 都有独立的坐标，这些坐标以非零概率独立

能力 θ。如果一个坐标非零，其值是从标准高斯分布中抽取的样本

(c) 我们观察到 b 其中 $Ax = b$。

因此，我们所有的样本在未知基下都是稀疏的。因此，我们希望找到 A，或者等价地，找到它的左伪逆 A^+，这是一个使所有样本变得稀疏的线性变换。参数 θ 控制每个表示 x 的平均稀疏度，并且它既不能太大也不能太小。更正式地，我们假设：

$$\frac{1}{n} \leq \theta \frac{1}{n^{1/2} \log n}$$

Spielman等人[131]给出了一种多项式时间算法，可以精确恢复A。这比我们后面将要看到的算法提供了更强的保证，后者仅仅可以近似恢复A或达到任意好的精度，但需要越来越多的样本才能做到这一点。然而，后面的算法将在过完备情况和存在噪声的情况下工作。还重要的是要注意，严格来说，如果x_i的坐标是独立的，我们可以使用独立成分分析算法[74]来恢复A。然而，这些算法对独立性假设非常敏感，而我们在这里所做的一切即使在更弱的条件（这些条件很难正确拼写出来）下也能工作。

我们将做出简化的假设，即A是可逆的。这实际上并没有给我们带来任何成本，但让我们将其留作读者的练习。无论如何，算法背后的主要洞察力包含在以下声明中：

Claim 6.2.1 *The row span of B and the row span of $A^{-1}B = X$ are the same*

Proof: 证明通过观察对任何向量 u,

$$u^T B = (u^T A)A^{-1}B = v^T X$$

因此，我们可以用 B 的行的一个对应线性组合来表示任何线性组合的行。显然，我们也可以反向进行。■

我们现在将非正式地陈述第二个主张：

Claim 6.2.2 *Given enough samples, with high probability the sparsest vectors in the row span of X are the rows of X*

希望这个说法直观明显。X的每一行都是独立随机向量，其平均稀疏度为θ。对于我们的选择θ，我们将有很少的冲突，这意味着X任意两行的稀疏度应该大约是单行稀疏度的两倍。

现在我们来到了需要凸规划松弛的需求。我们无法期望直接在任意子空间中找到最稀疏的向量。我们在定理5.1.5中证明了该问题是NP-难。但让我们利用我们从稀疏恢复中获得的认识，并使用凸规划松弛。考虑以下优化问题：

$$(P_1) \qquad \min \|w^T B\|_1 \text{ s.t. } r^T w = 1$$

这是用向量的ℓ_1范数替换其稀疏性的常用技巧。约束$r^T w = 1$只需要用来固定归一化，以防止我们返回

所有零向量作为解。我们将选择 r 作为 B 中的一个列向量，原因将在后面变得清晰。我们的目标是证明 (P_1) 的最优解是 X 的一个缩放行。实际上，我们可以将上述线性规划转化为一个更简单的形式，这将更容易分析：

$$(Q_1) \qquad \min \|z^T X\|_1 \text{ s.t. } c^T z = 1$$

Lemma 6.2.3 *Let $c = A^{-1}r$. Then there is a bijection between the solutions of (P_1) and the solutions of (Q_1) that preserves the objective value.*

Proof: 给定一个 (P_1) 的解 w，我们可以设置 $z = A^T w$。目标值相同，因为

$$w^T B = w^T AX = z^T X$$

并且线性约束再次得到满足

$$1 = r^T w = r^T (A^T)^{-1} z = r^T (A^{-1})^T z = c^T z$$

并且很容易检查，你可以从 (Q_1) 的一个解到 (P_1) 的一个解

以类似的方式。■

The Minimizers are Somewhat Sparse

这里我们将建立分析的关键步骤。我们将证明任何最优解 z_* 的支持包含在 c 的支持中。记住我们选择了 r 来

成为 B 的列。我们承诺稍后解释，所以现在我们可以问：我们为什么要这样做? 关键是如果 r 是 B 的列，那么我们的双射在 (P_1) 和 (Q_1) 的解之间工作的方式是，我们设定 $c = A^{-1}r$，因此 c 是 X 的列。在我们的模型中，c 是稀疏的，如果我们证明 z_* 的支撑包含在 c 的支撑中，那么我们就证明了 z_* 也是稀疏的。

现在让我们在本小节中陈述并证明主要引理。在接下来的内容中，我们将断言某些事情发生的概率很高，但不会过多关注使这些事情成为事实所需的样本数量。相反，我们将给出一个启发式论证，为什么集中界限应该以那种方式起作用，并且将重点放在与稀疏恢复的类比上。有关详细信息，请参阅Spielman等人[131]。

Lemma 6.2.4 *With high probability, any optimal solution z_* to (Q_1) satisfies $supp(z_*) \subseteq supp(c)$*

Proof: 让我们将 z_* 分解为两部分。设 $J =$ 为 supp(c) 并写出 $z_* = z_0 + z_1$，其中 z_0 在 J 中有支持，z_1 在 J 中有支持。然后我们有 $c^T z_0 = c^T z_*$。这意味着，由于 z_* 是 (Q_1) 的一个可行解，那么 z_0 也是。我们的目标是证明 z_0 是 (Q_1) 的一个比 z_* 更优的严格解。更正式地说，我们想要证明:

$$\|z_0^T X\|_1 < \|z_*^T X\|_1$$

设 S 为 X 中在 J 中有非零项的列集合。即

$$S = \{j | X_j^J \neq \vec{0}\}$$

我们现在计算：

$$\begin{aligned}\|z_*^T X\|_1 &= \|z_*^T X_S\|_1 + \|z_*^T X_{\overline{S}}\|_1 \\ &\geq \|z_0^T X_S\|_1 - \|z_1^T X_S\|_1 + \|z_1^T X_{\overline{S}}\|_1 \\ &\geq \|z_0^T X\|_1 - 2\|z_1^T X_S\|_1 + \|z_1^T X\|_1\end{aligned}$$

现在，让我们假设以下主张：

Claim 6.2.5 *With high probability, for any non-zero* z_1 *we have* $\|z^T 1\, X\|1 > 2\|z_1^T X_S\|_1$.

使用这个主张，我们有

$$\|z_*^T X\|_1 > \|z_0^T X\|_1$$

这完成了证明。■

现在我们来证明命题6.2.5：

Proof: 现在让我们作弊并假设 z_1 是固定的，并且是一个单位向量。那么 S 是一个随机集，如果我们从模型中抽取 p 个样本，我们就有：

$$\mathbb{E}[\|z_1^T X_S\|_1] = \frac{|S|}{p}\,\mathbb{E}[\|z_1^T X\|_1]$$

预期的 S 的大小为 $p \times \mathbb{E}[|\text{supp}(x_i)|] \times \theta = \theta^2 np = o(p)$。共同意味着

$$\mathbb{E}[\|z_1^T X\|_1 - 2\|z_1^T X_S\|_1] = \left(1 - \frac{2\,\mathbb{E}[|S|]}{p}\right)\mathbb{E}[\|z_1^T X\|_1]$$

是有界远离零的，从而证明了我们想要的界限

$$\|z_1^T X\|_1 - 2\|z_1^T X_S\|_1 > 0$$

对于任何固定的 z_1，以高概率成立。我们可以在所有可能的单位向量 z_1 的 ϵ-网中取并集，并通过缩放得出结论，该界限对所有非零 z_1 成立。■

The Minimizers are Rows of X

现在我们知道 (Q_1) 的解相对稀疏，因为它们的支撑集包含在 c 的支撑集中。但是，X 的行稀疏线性组合也将有很少的冲突，因此我们应该期望 ℓ_1 范数大致保持不变。更精确地说：

Lemma 6.2.6 *With high probability, for any vector z supported in a set J of size at most $10\theta n \log n$ we have*

$$\|z_J^T X^J\|_1 = (1 \pm o(1))C\frac{p}{|J|}\|z_J\|_1$$

where C is the expected absolute value of a non-zero in X.

我们在这里不会证明这个引理。有关详细信息，请参阅Spielman等人[131]。然而，直觉很容易理解。我们应该预期X_J的大多数列最多只有一个非零元素。分析这些列的期望贡献是直接的，而其余列只有低阶贡献。

这意味着对我们来说，我们不是考虑(Q_1)，而是可以考虑：

$$(R_1) \qquad \min \|z\|_1 \text{ s.t. } c^T z = 1$$

因为(Q_1)和(R_1)的可行域相同，并且在缩放后它们的目标值几乎相同。最终步骤如下：

Lemma 6.2.7 *If c has a unique coordinate of maximum value c_i, then the unique optimal solution to (R_1) satisfies $z_i = 1/c_i$ and $z_j = 0$ for all other coordinates j.*

现在我们可以陈述主要定理：

Theorem 6.2.8 *[131] Suppose A is an $n \times m$ matrix with full column rank and we are given a polynomial number of samples from the generative model. There is a polynomial time algorithm to recover A exactly (up to a permutation and rescaling of its columns) that succeeds with high probability.*

Proof: 定理可通过结合引理6.2.4、引理6.2.6和引理6.2.7得出。使用这些引理以及引理6.2.3中的双射，我们得出结论，对于(P_1)的任何最优解，出现在目标函数中的向量是

$$w^T B = z^T X$$

仅 i^{th} 坐标非零的 z。因此，它是 i^{th} 行 X 的缩放副本。现在，由于生成模型从标准高斯中选择了 x 的非零项，我们几乎可以肯定存在一个坐标，其绝对值是严格最大的。

实际上，甚至更多是正确的。对于任何固定的坐标 i，以高概率它将是 X 某一列的绝对值严格最大的坐标。这意味着如果我们通过将 r 设置为 B 的不同列来反复求解 (P_1)，那么以高概率 X 的每一行都会出现。现在一旦我们知道 X 的行，我们就可以如下求解 A。以高概率，如果我们取足够的样本，那么 X 将具有左伪逆，我们可以计算 $A = BX^+$，这将恢复 A，直到其列的排列和缩放。这完成了证明。■

6.3 Gradient Descent

梯度下降及其相关算法是机器学习中最为普遍的算法之一。传统上，我们面临的是最小化凸函数 $f : \mathbb{R}^n \to \mathbb{R}$ 的任务，要么在整个空间（无约束情况）上，要么在某些凸体 K 上。你能想到的最简单的算法——沿着最速下降方向前进——是有效的。实际上，根据你对函数的了解，存在各种收敛保证。它至少是二阶可微的吗？它的梯度是否平滑变化？你能在它下面拟合一个二次函数吗？甚至还有利用物理中的动量等联系来加速的方法，以获得更快的收敛速度。你可以写一本关于迭代方法的书。确实，有许多极好的资源，如Nesterov [116] 或Rockefellar [127]。

在这个部分，我们将证明关于梯度下降的一些基本结果，在f是二阶可微、β-光滑和α-强凸的最简单设置中。

我们将展示我们的目标函数当前值与最优值之间的差异呈指数衰减。最终，我们对梯度下降的兴趣将在于将其应用于非凸问题。一些最有趣的问题——如在一个深度网络中拟合参数——是非凸的。当面对一个非凸函数 f 时，你仍然可以运行梯度下降。

非常具有挑战性的是证明非凸优化的保证（除了能够达到局部最小值之类的事情）。尽管如此，我们针对过完备稀疏编码的方法将基于梯度下降分析的一种抽象。实际上，梯度始终指向全局最小解的方向。在非凸设置中，我们仍然可以利用这种直觉，通过表明在适当的随机假设下，即使简单的更新规则也能以类似的方式取得进展。无论如何，现在让我们定义梯度下降：

Gradient Descent

Given: a convex, differentiable function $f : \mathbb{R}^n \to \mathbb{R}$

Output: a point x_T that is an approximate minimizer of f

For $t = 1$ to T

$$x_{t+1} = x_t - \eta \nabla f(x_t)$$

End

参数 η 被称为学习率。你希望让它很大，但又不能太大以至于超出范围。我们对梯度下降的分析将取决于多变量

有能微积分。对我们有用的一个因素将是以下多元泰勒定理:

Theorem 6.3.1 *Let f $\mathbb{R}^n \to \mathbb{R}$ be a convex, differentiable function. Then*

$$f(y) = f(x) + (\nabla f(x))^T(y-x) + \frac{1}{2}(y-x)^T\nabla^2 f(x)(y-x) + o(\|y-x\|^2)$$

现在让我们精确地定义我们将对 f 施加的条件。首先我们需要梯度不要变化得太快:

Definition 6.3.2 *We will say that f is β-smooth if for all x and y we have*

$$\|\nabla f(y) - \nabla f(x)\| \leq \beta\|y-x\|$$

Alternatively if f is twice differentiable, the condition above is equivalent to $\|\nabla^2 f(x)\| \leq \beta$. for all x.

接下来我们需要能够在 f 下方拟合一个二次函数。我们需要这种条件的原因是排除 f 在长时间内基本平坦的情况，但我们需要移动很远才能达到全局最小值。如果你能在 f 下方拟合一个二次函数，那么你就知道全局最小值不可能离你当前的位置太远。

Definition 6.3.3 *We will say that a convex function f is α-strongly convex if for all x and y we have*

$$(y-x)^T\nabla^2 f(x)(y-x) \geq \alpha\|y-x\|^2$$

Or equivalently for all x and y, f satisfies

$$f(y) \geq f(x) + (\nabla f(x))^T(y-x) + \frac{\alpha}{2}\|y-x\|^2$$

现在让我们陈述本节将要证明的主要结果:

Theorem 6.3.4 *Let f be twice differentiable, β-smooth and α-strongly convex. Let x^* be the minimizer of f and $\eta \leq \frac{1}{\beta}$. Then gradient descent starting from x_1 satisfies*

$$f(x_t) - f(x^*) \leq \beta\Big(1 - \frac{\eta\alpha}{2}\Big)^{t-1}\|x_1 - x^*\|^2$$

我们将使用以下辅助引理:

Lemma 6.3.5 *If f is twice differentiable, β-smooth and α-strongly convex, then*

$$\nabla f(x_t)^T(x_t - x^*) \geq \frac{\alpha}{4}\|x_t - x^*\|^2 + \frac{1}{2\beta}\|\nabla f(x_t)\|^2$$

让我们回到它的证明。现在让我们看看它是如何被用来建立定理6.3.4的: {v *}

Proof: 让 $\alpha' = \frac{\alpha}{4}$ 和 $\beta' = \frac{1}{2\beta}$。然后我们有

$$\begin{aligned}
\|x_{t+1} - x^*\|^2 &= \|x_t - x^* - \eta\nabla f(x_t)\|^2 \\
&= \|x_t - x^*\|^2 - 2\eta\nabla f(x_t)^T(x_t - x^*) + \eta^2\|\nabla f(x_t)\|^2 \\
&\leq \|x_t - x^*\|^2 - 2\eta(\alpha'\|x_t - x^*\|^2 + \beta'\|\nabla f(x_t)\|) \\
&= (1 - 2\eta\alpha')\|x_t - x^*\|^2 + (\eta^2 - 2\eta\beta')\|\nabla f(x_t)\|^2 \\
&\leq (1 - 2\eta\alpha')\|x_t - x^*\|^2
\end{aligned}$$

第一个等式由梯度下降的定义得出。第一个不等式由引理6.3.5得出，最后一个不等式由学习率η的界限得出。为了完成证明，请注意

$$f(x_t) + \nabla f(x_t)^T (x^* - x_t) \leq f(x^*)$$

重新排列这个不等式并调用 β-平滑性，我们得到

$$f(x_t) - f(x^*) \leq \nabla f(x_t)^T (x_t - x^*) \leq \beta \|x_t - x^*\|^2$$

将所有内容合并，我们得到

$$f(x_t) - f(x^*) \leq \beta \Big(1 - 2\eta\alpha'\Big) \|x_t - x^*\|^2$$

这完成了证明。■

现在让我们整理我们的松散结尾并证明引理6.3.5：

Proof: 首先，通过强凸性我们有

$$f(x^*) \geq f(x) + \nabla f(x)^T (x^* - x) + \frac{\alpha}{2} \|x - x^*\|^2$$

使用事实 $f(x) \geq f(x^*)$ 并重新排列，我们得到

$$\nabla f(x)^T (x - x^*) \geq \frac{\alpha}{2} \|x - x^*\|^2$$

这是词元的二分之一。现在让我们将等式左边与梯度的范数联系起来。实际上，我们需要一个更方便的6.3.1定理的形式，它具有

拉格朗日余项:

Theorem 6.3.6 *Let f $\mathbb{R}^n \to \mathbb{R}$ be a twice differentiable function. Then for some some $t \in [0,1]$ and $x' = ty + (1-t)x$ we have*

$$\nabla f(x) = \nabla f(y) + \nabla^2 f(x')(x-y)$$

这可以通过使用多元介值定理来证明。在任何情况下，通过设置 $y = x^*$ 并观察 $\nabla f(x^*) = 0$，我们得到

$$\nabla f(x) = \nabla^2 f(x')(x-x^*)$$

从其中我们得到

$$\nabla f(x)^T (\nabla^2 f(x'))^{-1} \nabla f(x) = \nabla f(x)^T (x-x^*)$$

对于某些 $x' = tx + (1-t)x^*$。现在 β-平滑性意味着 $(\nabla^2 f(x'))^{-1} \geq \frac{1}{\beta}\|\nabla f(x')\|^2$，这给我们带来$\frac{1}{\beta}\|\nabla f(x')\|^2$

$$\|\nabla f(x)^T (x-x^*)\| \geq \frac{1}{\beta}$$

取两个主要不等式的平均值完成证明。■

实际上，我们的证明即使在移动的方向只是梯度的近似时也成立。这是一个重要的捷径（例如），当 f 是一个依赖于大量训练样本的损失函数时。你不必计算 f 的梯度，而是可以采样一些训练样本，计算

您的损失函数仅针对这些，并跟随其梯度。这被称为 *stochastic gradient descent*。它移动的方向是一个随机变量，其期望是 f 的梯度。它的美妙之处在于，梯度下降法收敛的常规证明可以简单地迁移（前提是您的样本足够大）。

存在一个更进一步的抽象我们可以做出。如果你移动的方向不是梯度的随机近似，而是仅仅满足第6.3.5引理所示条件的某个方向，那会怎样？为了给它起个名字，我们称之为抽象梯度下降：

Abstract Gradient Descent

Given: a function $f : \mathbb{R}^n \to \mathbb{R}$

Output: a point x_T that is close to x^*

For $t = 1$ to T

$x_{t+1} = x_t - \eta g_t$

End

让我们介绍以下关键定义：

Definition 6.3.7 *We say that g_t is $(\alpha', \beta', \epsilon_t)$-correlated with a point x^* if for all t we have*

$$g_t^T(x_t - x^*) \geq \alpha' \|x_t - x^*\|^2 + \beta' \|g_t\|^2 - \epsilon_t$$

我们已经证明，如果 f 是两次可微的，β-光滑的，并且 α-强凸的，那么梯度与最优解 x^* 是 $(\frac{\alpha}{4}, \frac{1}{2\beta}, 0)$-相关的。它

结果证明，我们给出的定理6.3.4的证明可以立即推广到这个更抽象的设置：

Theorem 6.3.8 *Suppose that g_t is $(\alpha', \beta', \epsilon_t)$-correlated with a point x^* and moreover $\eta \leq 2\beta'$. Then abstract gradient descent starting from $x1$ satisfies*

$$\|x_t - x^*\|^2 \leq \left(1 - \frac{\eta\alpha'}{2}\right)^{t-1} \|x_1 - x^*\|^2 + \frac{\max_t \epsilon_t}{\alpha'}$$

现在我们拥有了进行过完备稀疏编码所需的工具。尽管他们试图最小化的基础函数是非凸的，我们仍将证明迭代方法的收敛界限。关键是使用上述框架并利用我们模型中的随机性质。

6.4 The Overcomplete Case

在这个部分，我们将给出一个适用于过完备字典的稀疏编码算法。像往常一样，我们将在一个随机模型中工作。更正式地说，x 是一个随机 k-稀疏向量，根据以下程序生成：

(a) 从 $[m]$ 中所有大小为 k 的子集中，随机均匀选择 x 的支持 (b) 如果第 j 个坐标非零，则其值独立选择为 $+1$ 或 -1（概率相等）

我们观察到 $Ax = b$ 的右侧。我们的目标是学习 A 的列，前提是从模型中获取足够的样本。实际上，我们已经做了一些简化

在上述模型中的假设，我们实际上并不需要。我们并不真的需要选择 x 的支持是均匀随机分布的，或者坐标是独立的。事实上，我们的算法甚至能够容忍加性噪声。尽管如此，我们的模型更容易思考，所以我们还是坚持使用它。

现在我们来到了主要的概念洞察。通常我们认为迭代算法是在非凸目标函数上执行交替最小化。例如，在稀疏编码的背景下，一个流行的能量函数如下：

$$\mathcal{E}(\widehat{A}, \widehat{X}) = \sum_{i=1}^{p} \|b^{(i)} - \widehat{A}\widehat{x}^{(i)}\|^2 + \sum_{i=1}^{p} L(\widehat{x}^{(i)})$$

在 $Ax^{(i)} = b^{(i)}$ 是我们的观测样本。此外，L 是一个惩罚非 k-稀疏向量 $\widehat{x}^{(i)}$ 的损失函数。你可以将其视为一个硬惩罚函数，当 x 有超过 k 个非零坐标时，其值为无穷大，否则为零。它也可能是你最喜欢的稀疏性诱导软惩罚函数。

许多迭代算法试图最小化一个能量函数，如上所述，该函数平衡了你的基如何解释每个样本以及每个表示的稀疏程度。问题是该函数是非凸的，所以如果你想给出可证明的保证，你就必须弄清楚所有各种各样的事情，比如为什么它不会陷入局部最小值，或者为什么它不会花太多时间在鞍点周围缓慢移动。

Question 5 *Instead of viewing iterative methods as attempting to minimize a known, non-convex function, can we view them as minimizing an unknown, convex function?*

我们所说的意思是：如果我们不是插入 $\widehat{x}$，而是插入真正的稀疏

表示 x? 我们的能量函数变为:

$$\mathcal{E}(\widehat{A}, X) = \sum_{i=1}^{p} \|b^{(i)} - \widehat{A}x^{(i)}\|^2$$

这是凸的，因为只有基 A 是未知的。此外，在我们的随机模型（以及可能许多其他模型）中，$\mathcal{E}(\widehat{A}, X)$ 的最小化值收敛到真正的基 A 是自然的预期。因此，我们现在有一个凸函数，其中存在一条从我们的初始解到最优解的路径，通过梯度下降。问题是，由于 X 是未知的，我们无法评估或计算函数 $\mathcal{E}(\widehat{A}, X)$ 的梯度。

本节我们将展示简单、迭代的稀疏编码算法在接近 $\mathcal{E}(\widehat{A}, X)$ 梯度的方向上移动。更确切地说，我们将证明在我们的随机模型下，我们的更新规则移动的方向满足定义6.3.7中的条件。这就是我们的行动计划。我们将研究以下迭代算法:

Hebbian Rule for Sparse Coding

输入：样本 $b = Ax$ 和一个估计 $\widehat{A}$ 输出：
一个改进的估计 $\widehat{A}$

对于 $t = 0$ 到 T

使用当前词典解码：

$$\widehat{x}^{(i)} = \text{threshold}_{1/2}(\widehat{A}^T b^{(i)})$$

更新字典：

$$\widehat{A} \leftarrow \widehat{A} + \eta \sum_{i=qt+1}^{q(t+1)} (b^{(i)} - \widehat{A}\widehat{x}^{(i)}) \, \text{sign}(\widehat{x}^{(i)})^T$$

结束

我们使用了以下符号：

Definition 6.4.1 *Let* 标记
denote the entry-wise operation that sets positive coordi- nates to $+1$
, negative coordinates to -1 *and zero to zero. Also let* 阈值$_C$ *denote the entry-wise operation that zeros out coordinates whose absolute value is less than* $C/2$ *and keeps the rest of the coordinates the same.*

更新规则在另一个意义上也是自然的。在神经科学背景下，字典 A 通常代表两个相邻神经元层之间的连接权重。然后，更新规则具有这样的性质：当你设置一个计算稀疏表示的神经网络时，它会加强同时放电的神经元对之间的连接。回想一下，这些被称为 *Hebbian* 规则。

现在让我们定义我们将用于衡量我们的估计 $\widehat{A}$ 与真实字典 A 接近程度的度量。像往常一样，我们无法希望恢复哪一列是哪一列或正确的符号，因此我们需要考虑这一点：

Definition 6.4.2 *We will say that two $n \times m$ matrices $\widehat{A}$ and A whose columns are unit vectors are (δ, κ)-close if there is a permutation and sign flip of the columns of $\widehat{A}$ that results in a matrix B that satisifes*

$$\|B_i - A_i\| \leq \delta$$

for all i, and furthermore $\|B - A\| \leq \kappa \|A\|$.

首先让我们分析算法的解码步骤：

Lemma 6.4.3 *Suppose that A an $n \times m$ matrix that is μ-incoherent and that $Ax = b$ is generated from the stochastic model. Further suppose that*

$$k \leq \frac{1}{10\mu \log n}$$

and $\widehat{A}$ is $(1/\log n, 2)$-close to A. Then decoding succeeds — i.e.

$$\text{sign}(\text{threshold}_{1/2}(\widehat{A}^T b)) = \text{sign}(x)$$

with high probability.

我们在这里不会证明这个引理。思路是这样的，对于任意的 j，我们可以写出：

$$(\widehat{A}^T b)_j = A_j^T A_j x_j + (\widehat{A}_j - A_j)^T A_j x_j + \widehat{A}_j^T \sum_{i \in S \setminus \{j\}} A_i x_i$$

在 $S = \text{supp}(x)$ 的地方。第一项是 x_j。第二项的绝对值至多为 $1/\log n$。第三项是一个随机变量，其方差可以适当界定。有关详细信息，请参阅 Arora 等人 [16]。请注意，对于非相干字典，我们认为 $\mu = 1/\sqrt{n}$。

设 γ 表示任何范数可忽略的向量（例如 $n^{-\omega(1)}$）。我们将使用 γ 收集各种小的误差项，而无需担心最终表达式的外观。考虑当我们的赫布更新被限制在某些列 j 时，预期的方向。我们有

$$g_j = \mathbb{E}[(b - \widehat{A}\widehat{x})\ \text{sign}(\widehat{x}_j)]$$

在期望是来自我们模型的样本 $Ax = b$ 的情况下。这是一个先验上复杂的表达式，因为 b 是我们模型中的随机变量，$\widehat{x}$ 是从我们的解码规则中产生的随机变量。我们主要引理如下：

Lemma 6.4.4 *Suppose that $\widehat{A}$ and A are $(1/\log n, 2)$-close. Then*

$$g_j = p_j q_j (I - \widehat{A}_j \widehat{A}_j^T) A_j + p_j \widehat{A}_{-j} Q \widehat{A}_{-j}^T A_j \pm \gamma$$

where $q_j = \mathbb{P}[j \in S]$, $q_{i,j} = \mathbb{P}[i, j \in S]$ and $p_j = \mathbb{E}[x_j \text{sign}(x_j) | j \in S]$. Moreover $Q = \text{diag}(\{q_{i,j}\}_i)$.

Proof: 使用解码步骤以高概率恢复 x 的正确符号这一事实，我们可以对解码成功与否的指示变量进行各种操作，以用 x 代替 $\widehat{x}$。现在让我们陈述以下内容

声明，我们将在后面证明：

Claim 6.4.5 $g_j = \mathbb{E}[(I - \widehat{A}_S\widehat{A}_S^T)Ax\ sign(x_j))] \pm \gamma$

现在让 $S = \text{supp}(x)$。我们首先想象采样 x 的支持集，然后选择其非零项的值。因此，我们可以使用子条件重写期望值：

$$\begin{aligned} g_j &= \mathop{\mathbb{E}}_{S}[\mathop{\mathbb{E}}_{x_S}[(I - \widehat{A}_S\widehat{A}_S^T)Ax\ \text{sign}(x_j))]|S] \pm \gamma \\ &= \mathop{\mathbb{E}}_{S}[\mathop{\mathbb{E}}_{x_S}[(I - \widehat{A}_S\widehat{A}_S^T)A_j x_j\ \text{sign}(x_j))]|S] \pm \gamma \\ &= p_j \mathop{\mathbb{E}}_{S}[(I - \widehat{A}_S\widehat{A}_S^T)A_j] \pm \gamma \\ &= p_j q_j (I - \widehat{A}_j\widehat{A}_j^T)A_j + p_j\widehat{A}_{-j}Q\widehat{A}_{-j}^T A_j \pm \gamma \end{aligned}$$

第二个等式使用了在支持 S 条件下坐标不相关的性质。第三个等式使用了 p_j 的定义。第四个等式是通过将 j 对所有其他坐标的贡献分离出来得到的，其中 A_{-j} 表示删除 j 列后得到的矩阵。这现在完成了主要引理的证明。■

因此，为什么这个引理告诉我们我们的更新规则满足定义6.3.7中的条件？当$\widehat{A}$和A接近时，你应该将表达式视为以下：

$$g_j = \underbrace{p_j q_j (I - \widehat{A}_j\widehat{A}_j^T)A_j}_{\approx p_j q_j (A_j - \widehat{A}_j)} + \underbrace{p_j\widehat{A}_{-j}Q\widehat{A}_{-j}^T A_j}_{\text{systemic error}} \pm \gamma$$

因此，更新规则移动的预期方向几乎就是

理想的向量 $A_j - \widehat{A}_j$ 指向真实解的方向。这告诉我们，有时绕过非凸性的方法是拥有一个合理的随机模型。即使在最坏的情况下，你仍然可能陷入局部最小值，但在平均情况下，你通常会在每一步都取得进展。我们尚未讨论如何初始化它的问题。但结果是，存在简单的谱算法来找到良好的初始化。参见Arora等人[16]的详细内容，以及整体算法的保证。

让我们通过证明命题6.4.5来结束：

Proof: Let F 表示解码恢复 x 的正确符号的事件。从引理6.4.3我们知道 F 以高概率成立。首先，让我们使用事件 F 的指示变量，以添加一个可忽略的错误项为代价，将符号函数内的 $\widehat{x}$ 替换为 x：

$$\begin{aligned} g_j &= \mathbb{E}[(b - \widehat{A}\widehat{x})\ \mathrm{sign}(\widehat{x}_j)\mathbb{1}_F] + \mathbb{E}[(b - \widehat{A}\widehat{x})\ \mathrm{sign}(\widehat{x}_j)\mathbb{1}_{\overline{F}}] \\ &= \mathbb{E}[(b - \widehat{A}\widehat{x})\ \mathrm{sign}(x_j)\mathbb{1}_F] \pm \gamma \end{aligned}$$

等式使用了当事件 F 发生时，$\mathrm{sign}(\widehat{x}_j) = \mathrm{sign}(x_j)$ 的性质。现在让我们将 $\widehat{x}$ 代入：

$$\begin{aligned} g_j &= \mathbb{E}[(b - \widehat{A}\ \mathrm{threshold}_{1/2}(\widehat{A}^T b))\ \mathrm{sign}(x_j)\mathbb{1}_F] \pm \gamma \\ &= \mathbb{E}[(b - \widehat{A}_S \widehat{A}_S^T b)\ \mathrm{sign}(x_j)\mathbb{1}_F] \pm \gamma \\ &= \mathbb{E}[(I - \widehat{A}_S \widehat{A}_S^T) b\ \mathrm{sign}(x_j)\mathbb{1}_F] \pm \gamma \end{aligned}$$

这里我们使用了这样一个事实：当事件 F 发生时，阈值 $_{1/2}(\widehat{A}^T b)$ 保持所有坐标在 S 中不变，并将其余部分置为零。现在我们可以再玩一些

技巧以消除指示变量：

$$\begin{aligned} g_j &= \mathbb{E}[(I - \widehat{A}_S \widehat{A}_S^T) b \operatorname{sign}(x_j)] - \mathbb{E}[(I - \widehat{A}_S \widehat{A}_S^T) b \operatorname{sign}(x_j) \mathbb{1}_{\overline{F}}] \pm \gamma \\ &= \mathbb{E}[(I - \widehat{A}_S \widehat{A}_S^T) b \operatorname{sign}(x_j)] \pm \gamma \end{aligned}$$

这完成了对断言的证明。逐行来看，操作是平凡的，但得到了一个对更新规则的有用表达式。■

存在其他更早的过完备稀疏编码算法。Arora等人[15]提出了一种基于重叠聚类的算法，该算法适用于几乎达到稀疏恢复问题具有唯一解的阈值（类似于引理5.2.3）的不相干字典。Agarwal等人[2]、[3]给出了适用于过完备、不相干字典的算法，这些算法的工作阈值比多项式因子更差。Barak等人[25]给出了基于平方和层次结构的算法，这些算法具有近线性的稀疏性，但多项式的次数取决于所需的精度。

6.5 Exercises

Problem 6-1: 考虑稀疏编码模型 {v*}，其中 A 是一个固定的 $n \times n$ 矩阵，具有正交归一列 a_i，而 x 具有从分布中抽取的独立同分布坐标

$$x_i = \begin{cases} +1 & \text{with probability } \alpha/2, \\ -1 & \text{with probability } \alpha/2, \\ 0 & \text{with probability } 1 - \alpha. \end{cases}$$

目标是从符号和排列) 恢复 A (的列，给定许多独立样本 y。构建矩阵

$$M = \mathbb{E}_y\left[\langle y^{(1)}, y\rangle\langle y^{(2)}, y\rangle y y^T\right]$$

在 $y^{(1)} = Ax^{(1)}$ 和 $y^{(2)} = Ax^{(2)}$ 是稀疏编码模型的两个固定样本的情况下，期望是在稀疏编码模型的第三个样本 y 上。设 $\hat{z}$ 为与最大的（绝对值）特征值对应的 M 的（单位范数）特征向量。

(a) 用 $\alpha, x^{(1)}, x^{(2)}, \{a_i\}$ 表示 M 的表达式。 (b) 为了简化，假设 $x^{(1)}$ 和 $x^{(2)}$ 的支持大小恰好为 αn，并且它们的支持在单个坐标 i^* 处相交。证明在极限 $\alpha \to 0$ 时，$\langle \hat{z}, a_{i^*}\rangle^2 \geq 1 - O(\alpha^2 n)$。

这种方法可用于找到交替最小化一个好的起点。

Chapter 7

Gaussian Mixture Models

许多自然统计——例如人们身高的分布——可以建模为高斯混合。混合的成分代表来自不同亚群体的分布部分。但如果我们事先不知道亚群体，我们能否找出它们是什么以及学习它们的参数？然后我们能否根据样本可能来自哪个亚群体进行分类？在本章中，我们将介绍以逆多项式速率学习高斯混合参数的第一个算法。一维情况由Karl Pearson引入，他是统计学创始人之一。我们将展示他方法的第一项可证明的保证。在此基础上，我们将解决高维学习问题。在这个过程中，我们将发展关于多项式方程系统和它们如何用于参数学习的见解。

7.1 Introduction

卡尔·皮尔逊是统计学界的杰出人物之一，并帮助为其奠定基础。他引入了革命性的新思想和方法，例如：

(a) p-值，现在是衡量统计显著性的实际方式

(b) 卡方检验，用于衡量与高斯分布的拟合优度

(c) 皮尔逊相关系数

(d) 分布参数的矩估计法

(e) 用于建模亚群体存在的混合模型

相信与否，最后两个是在同一项有影响力的1894年研究中引入的，该研究代表了皮尔逊首次涉足生物统计学[120]。让我们了解是什么引导皮尔逊走上了这条路。在度假期间，他的同事沃尔特·韦尔顿和他的妻子仔细收集了一千只那不勒斯蟹，并测量了每只蟹的23个不同的物理特征。但数据中隐藏着一个惊喜。除了一个之外，所有这些统计数据都近似于高斯分布。那么，为什么它们不是全部都是高斯分布的呢?

每个人都很困惑，直到皮尔逊提供了一个解释：*Maybe the Naples crab is not one species but rather two species?* 然后自然地，将观察到的分布建模为两个高斯分布的混合，而不是只有一个。让我们更正式一点。回忆一下，一维高斯分布的密度函数

具有均值 μ 和方差 σ^2 的:

$$\mathcal{N}(\mu, \sigma^2, x) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{\frac{-(x-\mu)^2}{2\sigma^2}\right\}$$

对于两个高斯混合，它是:

$$F(x) = w_1 \underbrace{\mathcal{N}(\mu_1, \sigma_1^2, x)}_{F_1(x)} + (1 - w_1) \underbrace{\mathcal{N}(\mu_2, \sigma_2^2, x)}_{F_2(x)}$$

我们将使用 F_1 和 F_2 来表示混合中的两个高斯分布。你也可以这样想：如何从这个分布中生成一个样本：取一个概率为 w_1 出头的偏硬币和剩余概率为 $1 - w_1$ 的尾巴。然后对于每个样本，你抛掷这枚硬币——即决定你的样本来自哪个子群体——如果是正面，你输出来自第一个高斯分布的样本，否则你输出来自第二个高斯分布的样本。

这是一个强大且灵活的统计模型。但皮尔逊并没有止步于此。他想找到两个高斯混合的最佳参数，以拟合观察到的数据来检验他的假设。当只有一个高斯分布时，这很简单，因为你可以将 μ 和 σ^2 分别设置为经验均值和经验方差。但当有五个未知参数，并且每个样本都有一个表示其来自哪个子群体的隐藏变量时，你应该怎么做呢？皮尔逊使用了矩估计法，我们将在下一小节中解释。他找到的参数似乎是一个很好的拟合，但仍有许多未解决的问题，比如，矩估计法是否总是能在存在解的情况下找到好的解?

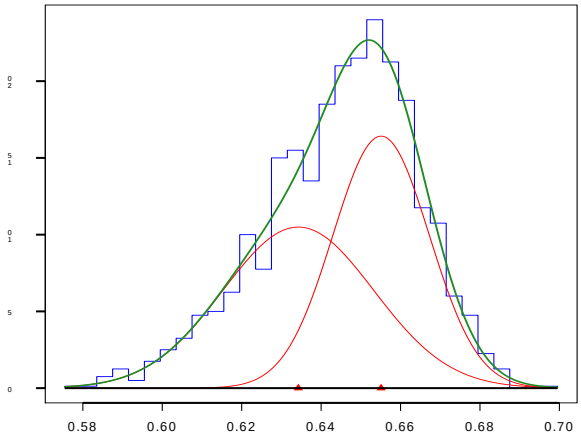


图7.1：使用R由Peter Macdonald创建的两个一元高斯混合对那不勒斯蟹的皮尔逊数据进行拟合

Method of Moments

这里我们将解释皮尔逊如何使用矩估计法来找到未知参数。关键观察是高斯混合的矩本身是未知参数的多项式。让我们用 M_r 表示高斯的原矩 r^{th}：

$$\mathop{\mathbb{E}}_{x \leftarrow F_1(x)}[x^r] = M_r(\mu, \sigma)$$

它很容易计算 $M_1(\mu, \sigma) = \mu$ 和 $M_2(\mu, \sigma) = \mu^2 + \sigma^2$ 等等，并检查 M_r 是 μ 和 σ 上的一个次数为 r 的多项式。现在我们有

$$\mathop{\mathbb{E}}_{x \leftarrow F(x)}[x^r] = w_1 M_r(\mu_1, \sigma_1) + (1 - w_1) M_r(\mu_2, \sigma_2) = P_r(w_1, \mu_1, \sigma_1, \mu_2, \sigma_2)$$

因此，两个高斯混合的 {v*} 原始矩本身是一个 $r+$ 次多项式——我们用 P_r 表示——在我们要学习的参数中。

Pearson's Sixth Moment Test: 我们可以从随机样本中估计 $\mathbb{E}_{x\leftarrow F}[x^r]$：设 S 为我们的样本集。然后我们可以计算：

$$\widetilde{M}_r = \frac{1}{|S|}\sum_{x\in S} x^r$$

给定一个多项式数量的样本（对于任何 $r = O(1)$），$\widetilde{M}_r$ 将加性接近于 $\mathbb{E}_{x\leftarrow F(x)}[x^r]$. 皮尔逊的方法是：

- 设置一个多项式方程组

$$\left\{P_r(w_1,\mu_1,\sigma_1,\mu_2,\sigma_2) = \widetilde{M}_r\right\},\ r = 1,2,\ldots,5$$

- 解这个系统。每个解都是所有五个参数的设置，这些参数解释了前五个经验矩。

皮尔逊解决了上述多项式方程组 *by hand*，并找到了多个候选解。每个解都对应一种设置所有五个未知参数的方法，使得混合物的矩与经验矩相匹配。但我们如何从这些候选解中选择呢？其中一些解显然是不正确的；一些解的方差为负值，或者混合权重不在零和一之间。但即使消除了这些解，皮尔逊仍然有多个候选解。他的方法是选择预测值与经验值最接近的候选解。

第六矩 $\widetilde{M}_6$。这被称为 *sixth moment test*。

Expectation Maximization

现代统计学中的工作马是 *maximum likelihood estimator*，它设置参数以最大化混合生成观察样本的概率。这个估计量具有许多美好的特性。在一定的技术条件下，它是 *asymptotically efficient*，这意味着没有其他估计量可以达到作为样本数量函数的渐近更小的方差。甚至其分布定律也可以被描述，并且已知它是正态分布的，其方差与所谓的费舍尔信息相关。不幸的是，对于我们将要感兴趣的大多数问题，它是 NP-hard to compute [19]。

流行的替代方案被称为 *expectation-maximization*，并在 Dempster、Laird、Rubin [61] 的一篇有影响力的论文中提出。重要的是要认识到，这只是一个用于计算最大似然估计的启发式方法，并不继承其任何统计保证。期望最大化是一种处理潜在变量的通用方法，我们在给定当前参数估计潜在变量和更新参数之间交替。在两个高斯混合的情况下，它重复以下步骤直到收敛

- 对于每个 $x \in S$，计算后验概率：

$$\widehat{w}_1(x) = \frac{\widehat{w}_1 \widehat{F}_1(x)}{\widehat{w}_1 \widehat{F}_1(x) + (1 - \widehat{w}_1)\widehat{F}_2(x)}$$

- 更新混合权重:

$$\widehat{w}_1 \leftarrow \frac{\sum_{x\in S} \widehat{w}_1(x)}{|S|}$$

- 重新估计参数:

$$\widehat{\mu}_i \leftarrow \frac{\sum_{x\in S} \widehat{w}_i(x)x}{\sum_{x\in S} \widehat{w}_i(x)},\ \widehat{\Sigma}_i \leftarrow \frac{\sum_{x\in S} \widehat{w}_i(x)(x-\widehat{\mu}_i)(x-\widehat{\mu}_i)^T}{\sum_{x\in S} \widehat{w}_i(x)}$$

实际上，它似乎工作得很好。但它可能会陷入似然函数的局部最大值。更糟糕的是，它对初始化方式非常敏感（例如，参见[125]）。

7.2 Clustering-Based Algorithms

我们的基本目标将是提供能够证明计算高斯混合真参数的算法，前提是给定多项式数量的随机样本。这个问题在Dasgupta的开创性论文[56]中被提出，第一代算法主要关注高维情况，其中成分之间足够远，以至于它们基本上没有$overlap$。下一代算法基于代数洞察，并完全避免聚类。

The High-Dimensional Geometry of Gaussians

在继续之前，我们将讨论一些高维高斯分布的逆直觉性质。首先，$\mathbb{R}^n$中的多维高斯密度是

给定如下:

$$\mathcal{N}(\mu, \Sigma) = \frac{1}{(2\pi)^{n/2} det(\Sigma)^{1/2}} exp\left\{\frac{-(x-\mu)^\top \Sigma^{-1}(x-\mu)}{2}\right\}$$

这里 Σ 是协方差矩阵。如果 $\Sigma = \sigma^2 I_n$ 和 $\mu = \vec{0}$，那么分布就是：$\mathcal{N}(0, \sigma^2) \times \mathcal{N}(0, \sigma^2) \times ... \times \mathcal{N}(0, \sigma^2)$，我们称之为球形高斯，因为密度函数是旋转不变的。

Fact 7.2.1 *The maximum value of the density function is at* $x = \mu$.

Fact 7.2.2 *For a spherical Gaussian, almost all of the weight of the density function has* $\|x - \mu\|22 = \sigma\text{^}2n \pm \sigma\text{^}2\sqrt{n \log n}$

起初，这些事实可能看起来不一致。第一个告诉我们样本最可能的价值在零。第二个告诉我们几乎所有样本都远离零。最容易想到的是在球坐标系中发生的事情。密度函数的最大值发生在半径 $R = 0$。但是球体表面积增加的速度比密度函数减少的速度快得多，直到我们达到半径为 $R = \sigma\sqrt{n}$。实际上，我们应该将高维球面高斯视为本质上是一个薄球壳。

The Cluster-then-Learn Paradigm

聚类算法都是基于以下策略:

- 将所有样本 S 根据它们是否由第一个或第二个组件生成，聚类成两组 S_1 和 S_2。

- 输出每个 S_i 的经验均值和协方差，以及经验混合权重 $\frac{|S_1|}{|S|}$。

我们如何实施第一步的细节以及我们需要施加的条件类型将因算法而异。但首先让我们看看，如果我们能够设计一个以高概率成功的聚类算法，我们找到的参数将是对真实参数的证明良好的估计。这由以下引理所捕捉。设$|S| = m$为样本数量。

Lemma 7.2.3 *If* $m \geq C\frac{\text{对数}\, 1/\delta}{\epsilon^2}$ *and clustering succeeds then*

$$|\widehat{w}_1 - w_1| \leq \epsilon$$

with probability at least $1 - \delta$.

现在让 $w_{min} = \min(w_1, 1 - w_1)$。然后

Lemma 7.2.4 *If* $m \geq C^n\frac{\text{对数}\, 1/\delta}{w_{min}\epsilon^2}$ *and clustering succeeds then*

$$\|\widehat{\mu}_i - \mu_i\|_2 \leq \epsilon$$

for each i, *with probability at least* $1 - \delta$.

最后让我们证明经验协方差也非常接近：

Lemma 7.2.5 *If* $m \geq C^n\frac{\text{对数}\, 1/\delta}{w_{min}\epsilon^2}$ *and clustering succeeds then*

$$\|\widehat{\Sigma}_i - \Sigma_i\| \leq \epsilon$$

for each i, with probability at least $1 - \delta$.

所有这些引理都可以通过标准的集中界限来证明。前两个来自于标量随机变量的集中界限，而第三个需要更高级的矩阵集中界限。然而，通过证明$\widehat{\Sigma}_i$和Σ_i的每个元素都接近，并使用并集界限，可以很容易地证明一个更差但仍然与n的多项式相关的版本。这些引理共同告诉我们，如果我们真的能够解决聚类问题，那么我们确实能够证明估计未知参数。

Dasgupta [56] – $\widetilde{\Omega}(\sqrt{n})$ Separation

Dasgupta 提出了学习高斯混合的第一个可证明的算法，并要求 $\|\mu_i - \mu_j\|_2 \geq \widetilde{\Omega}(\sqrt{n}\sigma_{max})$，其中 σ_{max} 是任何方向上任何高斯分布的最大方差（例如，如果分量不是球形的）。请注意，分离中的常数取决于 w_{min}，我们假设我们知道这个参数（或它的下界）。

算法背后的基本思想是将混合物随机均匀地投影到 {v*} 对数维度上。这种投影将以高概率保留每对中心 μ_i 和 μ_j 之间的距离，但会缩短来自同一组件的样本之间的距离，使每个组件更接近球形，从而更容易进行聚类。非正式地说，我们可以将这种分离条件视为：如果我们把每个高斯视为一个球形球体，那么如果组件之间足够远，那么这些球体将 *disjoint*。

Arora and Kannan [19], Dasgupta and Schulman [64] – $\widetilde{\Omega}(\{v^*\})$ Separation

我们将详细描述[19]中的方法。基本问题是，如果$\sqrt{n}$分离是我们可以将组件视为不相交的阈值，那么我们如何学习当组件非常接近时的情况？事实上，即使组件仅分离为$\widetilde{\Omega}(n^{1/4})$，仍然成立，即来自同一组件的*every*对样本比来自不同组件的*every*对样本更接近。这怎么可能？解释是，尽管代表每个组件的球体不再不相交，但我们仍然非常不可能从它们的重叠区域中采样。

考虑 $x, x' \leftarrow F_1$ 和 $y \leftarrow F_2$。

Claim 7.2.6 *All of the vectors $x-\mu_1$, $x'-\mu_1$, $\mu_1-\mu_2$, $y-\mu_2$ are nearly orthogonal (whp)*

这个断言是直接的，因为向量 $x-\mu_1$、$x'-\mu_1$、$y-\mu_2$ 从球面上均匀分布，而 $\mu_1-\mu_2$ 是唯一的固定向量。实际上，任何除了一个之外的所有向量都从球面上均匀随机选取的向量集几乎都是正交的。

现在我们可以计算：

$$\begin{aligned}\|x-x'\|^2 &\approx \|x-\mu_1\|^2 + \|\mu_1-x'\|^2 \\ &\approx 2n\sigma^2 \pm 2\sigma^2\sqrt{n\log n}\end{aligned}$$

同样地：

$$\begin{aligned}\|x-y\|^2 &\approx \|x-\mu_1\|^2 + \|\mu_1-\mu_2\|^2 + \|\mu_2-y\|^2 \\ &\approx 2n\sigma^2 + \|\mu_1-\mu_2\|^2 \pm 2\sigma^2\sqrt{n\log n}\end{aligned}$$

因此，如果 $\|\mu_1-\mu_2\| = \tilde{\Omega}(n^{1/4}, \sigma)$，则 $\|\mu_1-\mu_2\|^2$ 大于误差项，并且来自同一组件的每个样本对将比来自不同组件的每个样本对更接近。实际上，我们可以找到正确的阈值 τ 并正确地聚类所有样本。再次，我们可以输出每个聚类的经验均值、经验协方差和相对大小，这些将是真实参数的良好估计。

Vempala and Wang [141] – $\widetilde{\Omega}(k^{1/4})$ Separation

Vempala和Wang [141] 移除了对 n 的依赖，并用一个依赖于 k（即组件数量）的分离条件来替换它。其想法是，如果我们能将混合投影到由 $\{\mu_1, \ldots, \mu_k\}$ 张成的子空间 T 中，我们就会保留每个组件对之间的分离，但会降低环境维度。

因此，我们如何找到由均值张成的子空间 T 呢？我们将限制我们的讨论范围到一个具有共同方差 $\sigma^2 I$ 的球形高斯混合。设 $x \sim F$ 是混合的一个随机样本，那么我们可以写成 $x = c + z$，其中 $z \sim N(0, \sigma^2 I_n)$，而 c 是一个随机向量，对于每个 $i \in [k]$，它以概率 w_i 取值 μ_i。所以：

$$\mathbb{E}[xx^T] = \mathbb{E}[cc^T] + \mathbb{E}[zz^T] = \sum_{i=1}^{k} w_i \mu_i \mu_i^\top + \sigma^2 I_n$$

因此，$\mathbb{E}[xx^T]$ 的左上角奇异向量，其奇异值严格大于 σ^2，正好张成 T。然后，我们可以从足够多的随机样本中估计 $\mathbb{E}[xx^T]$，计算其奇异值分解，并将混合物投影到 T 上，并调用 [19] 中的算法。

Brubaker and Vempala [40] – Separating Hyperplane

如果任何组件的最大方差远大于组件之间的分离，会怎样？Brubaker和Vempala [40]观察到，对于看起来像一对*parallel pancakes*的混合物，现有的所有算法都失败了。在这个例子中，存在一个超平面可以分离混合物，使得几乎一个组件的所有部分都在一边，而另一个组件的几乎所有部分都在另一边。[40]给出了一种算法，只要存在这样的分离超平面，该算法就能成功，然而，对于三个或更多高斯混合物的条件更为复杂。对于三个组件，我们很容易构建我们希望学习的混合物，但是当没有超平面可以分离一个组件与其他组件时。

7.3 Discussion of Density Estimation

我们迄今为止讨论的所有算法都依赖于聚类。但有些情况下，这种策略根本无法奏效，因为聚类在信息论上是不可能的。更准确地说，我们将在下面展示，如果$d_{TV}(F_1, F_2) = 1/2$，那么我们将很快遇到一个我们无法确定其由哪个组件生成的样本，即使我们知道了真实参数。

让我们通过耦合的概念来形式化这一点：

Definition 7.3.1 *A coupling between F and G is a distribution on pairs (x, y) so that the marginal distribution on x is F and the marginal distribution on y is G. The error is the probability that $x \neq y$.*

所以最佳耦合的错误是什么？很容易看出，它正好是总变分距离：

Claim 7.3.2 *There is a coupling with error ε between F and G if and only if $d_{TV}(F, G) \leq \varepsilon$.*

实际上，这是一种思考总变差距离的好方法。从操作上讲，上界总变差距离告诉我们存在良好的耦合。以类似的方式，你可以将KL散度解释为你使用最佳代码从另一个分布最优地编码样本时所付出的代价（以期望编码长度来衡量）。

返回到从两个高斯混合物中聚类的样本的问题，假设我们有 $d_{TV}(F_1, F_2) = 1/2$ 以及

$$F(x) + 1/2F_1(x) + 1/2F_2(x)$$

使用上述断言，我们知道F_1和F_2之间存在耦合，概率为$1/2$。因此，我们不必像通常那样考虑从两个高斯混合物中采样（选择哪个成分，然后从中选择一个随机样本），我们可以选择以下方式进行采样：

1. 从F_1和F_2的最佳耦合中选择(x, y) 2. 如果$x = y$，以$1/2$的概率输出x，否则输出y

3. 否则以概率$1/2$输出x，否则输出y

此过程从 F 中生成一个随机样本，就像之前一样。重要的是，如果你达到第二步，你输出的值不依赖于哪个

组件样本来源。因此，你无法比随机猜测更好地预测它。这是思考基于聚类的算法所做假设的有用方式。有些比其他更强，但至少它们需要至少 n 个样本并将它们全部正确聚类。为了使这成为可能，我们必须有

$$d_{TV}(F_1, F_2) \geq 1 - 1/n$$

但是谁说学习算法必须首先聚类？我们能否希望在组件几乎完全重叠的情况下学习参数，例如当 $d_{TV}(F_1, F_2) = 1/n$？

现在是一个讨论我们可能追求的目标类型及其相互关系的良好时机：

(a) Improper Density Estimation

这是最弱的学习目标。如果我们从某个类别 $\mathcal{C}$（中的某个分布 F 获得样本，例如 $\mathcal{C}$ 可能是两个高斯分布的所有混合，那么我们希望找到任何其他满足 $d_{TV}(F, \widehat{F}) \leq \varepsilon$ 的分布 $\widehat{F}$。我们不需要 $\widehat{F}$ 也属于类别 $\mathcal{C}$。关于不适当密度估计的重要信息是，在一维中它很容易。只要 F 是平滑的，您就可以使用核密度估计来解决它。

这里是如何进行核密度估计的。首先，你取很多样本并构建一个经验点质量分布 G。现在 G 并不接近 F。它甚至不光滑，怎么可能呢？但是你可以通过与具有小方差的高斯函数卷积来修复这个问题。特别是，如果你设置 $\widehat{F} = G * \mathcal{N}(0, \sigma^2)$ 并适当地选择参数和样本数量，你得到的结果将满足

$d_{TV}(F, \widehat{F}) \leq \varepsilon$ 高概率。此方案不使用太多关于分布 F 的信息，但它以高维为代价。问题是您将无法获得足够接近的样本。一般来说，核密度估计需要样本数量以维度的指数形式增加才能工作。

(b) Proper Density Estimation

适当的密度估计要求$\widehat{F} \in \mathcal{C}$，因此它更强。有时，你可以通过将$\widehat{F}$约束在包含$\mathcal{C}$的某个更大的类别中，在不适定密度估计和适当密度估计之间进行插值。还值得注意的是，有时你只需取核密度估计或解决不适定密度估计问题的任何其他方法，并寻找与你的不适定估计最接近的$\widehat{F} \in \mathcal{C}$。这肯定会起作用，但问题是算法上通常不清楚如何找到某个类别中与某个难以处理的目标分布最接近的分布。最后，我们达到了最强大的目标类型：

(c) Parameter Learning

这里我们不仅需要 $d_{TV}(F, \widehat{F}) \leq \varepsilon$ 和$\widehat{F} \in \mathcal{C}$，还希望$\widehat{F}$ 是
F on a component-by-component basis 的良好估计。例如，我们的目标专门针对两个高斯混合的情况是：

Definition 7.3.3 *We will say that a mixture $\widehat{F} = \widehat{w}_1\widehat{F}_1 + \widehat{w}_2\widehat{F}_2$ is ε-close (on a component-by-component basis) to F if there is a permutation $\pi : \{1, 2\} \to \{1, 2\}$ so that for all $i \in \{1, 2\}$:*

$$\left|w_i - \widehat{w}_{\pi(i)}\right|, d_{TV}(F_i, \widehat{F}_{\pi(i)}) \leq \varepsilon$$

注意，F 和 $\widehat{F}$ 作为混合物也必须足够接近：$d_{TV}(F, \widehat{F}) \leq 4\varepsilon$。然而，我们可以有混合物 F 和 $\widehat{F}$，它们都是 k 高斯分布的混合，作为分布接近，但在逐个成分的基础上并不接近。那么，我们为什么要追求这样一个具有挑战性的目标呢？结果证明，如果 $\widehat{F}$ 与 ε 相对于 F 接近，那么给定一个典型样本，我们可以准确估计后验概率[94]。这意味着即使你不能将所有样本聚类到它们所属的成分中，你仍然可以确定哪些样本可以自信地确定。这是参数学习相对于一些较弱的学习目标的主要优势之一。

它很好，能够实现你所能希望的 strongest 学习目标，但你应该记住，这些 strong 学习目标的下限（例如参数学习）并不意味着较弱问题的下限（例如适当的密度估计）。我们将给出学习 k 高斯混合参数的算法，这些算法对于任何 $k = O(1)$ 都在多项式时间内运行，但与 k 的指数相关。但这在以下情况下是必要的，即存在 k 高斯混合的 F 和 $\widehat{F}$ 对，它们在逐个组件的基础上并不接近，但具有 $d_{TV}(F, \widehat{F}) \leq 2^{-k}$ [114]。因此，任何参数学习算法都能够区分它们，但这至少需要通过耦合论据再次使用 2^k 个样本。但也许对于适当的密度估计，可以得到一个在所有参数上都是多项式的算法：

Open Question 1 *Is there a* $\mathrm{poly}(n, k, 1/\varepsilon)$ *time algorithm for proper density estimation for mixtures of* k *Gaussians in* n *dimensions? What about in one dimension?*

7.4 Clustering-Free Algorithms

我们的目标是学习一个与F的ε-close的$\widehat{F}$。首先，我们将定义推广到k高斯混合模型：

Definition 7.4.1 *We will say that a mixture* $\widehat{F} = \sum_{i=1}^{k} \widehat{w}_i \widehat{F}_i$ *is* ε*-close (on a component-by-component basis) to* F *if there is a permutation* $\pi : \{1, 2, \ldots, k\} \to \{1, 2, \ldots, k\}$ *so that for all* $i \in \{1, 2, \ldots, k\}$*:*

$$\left| w_i - \widehat{w}_{\pi(i)} \right|, d_{TV}(F_i, \widehat{F}_{\pi(i)}) \leq \varepsilon$$

我们何时可以期望在多项式$(n, 1/\varepsilon)$样本中学习到ε的近似值？有两种情况下这是不可能的。最终我们的算法将表明，这些是唯一出错的情况：

(a) 如果 $w_i = 0$，我们永远无法学习接近 F_i 的$\widehat{F}_i$，因为我们从未从 F_i 获取任何样本。

实际上，我们需要对每个 w_i 有一个量化的下界，比如说 $w_i \geq \varepsilon$，这样如果我们取一个合理的样本数量，我们至少会得到每个成分的一个样本。

(b) 如果 $d_{TV}(F_i, F_j) = 0$，我们永远无法学习 w_i 或 w_j，因为 F_i 和 F_j 完全重叠。

再次，我们需要对 $d_{TV}(F_i, F_j)$ 有一个量化的下界，比如说对每个 $i \neq j$ 有 $d_{TV}(F_i, F_j) \geq \varepsilon$，这样如果我们取一个合理的样本数量，我们至少会得到一个来自各个组件对之间非重叠区域的样本。

Theorem 7.4.2 *[94], [114] If $w_i \geq \varepsilon$ for each i and $d_{TV}(F_i, F_j) \geq \varepsilon$ for each $i \neq j$, then there is an efficient algorithm that learns an ε-close estimate $\widehat{F}$ to F whose running time and sample complexity are 多$(n, 1/\varepsilon, \log 1/\delta)$ and succeeds with probability $1 - \delta$.*

注意，多项式的次数多项式地依赖于 k。Kalai、Moitra 和 Valiant [94] 提出了第一个学习两个高斯混合的算法，没有分离条件。随后，Moitra 和 Valiant [114] 提出了一个 k 个高斯混合的算法，同样没有分离条件。

在独立和并行工作中，Belkin和Sinha[28]也为k高斯混合物提供了一个多项式时间算法，然而，由于他们的工作依赖于希尔伯特基定理，该定理在本质上无效)，因此没有给出关于运行时间作为k (函数的显式界限。此外，[94]和[114]的目标是学习$\widehat{F}$，使其分量在总变差距离上接近F的分量，这通常是一个比要求参数在加性接近上更强的目标，而[28]的目标是参数在加性接近。好处是[28]中的算法适用于一维设置中更一般的学习问题，我们将在本章末尾解释他们算法的思想。

在整个本节中，我们将关注 $k = 2$ 情况，因为该算法在概念上要简单得多。事实上，我们的目标是实现一个更弱的学习目标：我们将说$\widehat{F}$ 与 F 的距离是 *additively* ε，如果对于所有 i 都有 $|w_i - \widehat{w}_{\pi(i)}|, \|\mu_i - \widehat{\mu}_{\pi(i)}\|, \|\Sigma_i - \widehat{\Sigma}_{\pi(i)}\|_F \leq \varepsilon$。我们希望找到这样的$\widehat{F}$。结果证明，我们可以假设 F 在以下意义上是归一化的：

Definition 7.4.3 *A distribution F is in isotropic position if*

(a) $\mathbb{E}_{x \leftarrow F}[x] = 0$

(b) $\mathbb{E}_{x \leftarrow F}[xx^T] = I$

第二个条件意味着在 *every* 方向上的方差为1。实际上，只要没有方差为零的方向，将分布置于各向同性位置很容易。更精确地说：

Claim 7.4.4 *If* $\mathbb{E}_{x \leftarrow F}[xx^T]$ *is full-rank, then there is an affine transformation that places* F *in isotropic position*

Proof: 让 $\mu = E_{x \leftarrow F}[x]$。然后

$$E_{x \leftarrow F}[(x - \mu)(x - \mu)^T] = M = BB^T$$

这遵循因为 M 因此具有 Cholesky 分解。根据假设 M 具有满秩，因此 B 也具有满秩。现在如果我们设置

$$y = B^{-1}(x - \mu)$$

很容易看出 $\mathbb{E}[y] = 0$ 和 $\mathbb{E}[yy^T] = B^{-1}M(B^{-1})^T = I$ 如所期望。 ■

我们的目标是学习一个对 ε 的加性 F 近似，并且我们假设 F 已经预处理过，使其处于各向同性位置。

Outline

我们现在可以描述算法的基本轮廓，尽管还有很多细节需要填充：

(a) 考虑一系列投影到一维（b）运行单变量学习算法（c）在高维参数上建立线性方程组，并回代求解

Isotropic Projection Lemma

我们需要克服许多障碍才能实现这个计划，但让我们先了解当我们将高斯函数沿某个方向 r 投影时，其参数会发生什么变化：

Claim 7.4.5 $proj_r[\mathcal{N}(\mu, \Sigma)] = \mathcal{N}(r^T\mu, r^T\Sigma r)$

这个简单的声明已经告诉我们一些重要的事情：假设我们想要学习高维高斯分布的参数 μ 和 Σ。如果我们将其投影到方向 r 并学习结果一维高斯分布的参数，那么我们真正学习到的是对 μ 和 Σ 的线性约束。如果我们对许多不同的方向 r 进行多次操作，我们希望得到足够的线性约束 μ 和 Σ，从而可以简单地求解它们。更自然地，我们希望需要大约 n^2 个方向，因为 Σ 有那么多参数。但现在我们遇到了第一个问题，我们需要找到一种方法来绕过它。让我们引入一些符号：

Definition 7.4.6 $d_p(\mathcal{N}(\mu_1, \sigma^{21}), \mathcal{N}(\mu_2, \sigma^{22})) = |\mu_1 - \mu_2| + |\sigma^{21} - \sigma^{22}|$

我们将将其称为参数距离。最终，我们将给出一个学习高斯混合的单变量算法，并希望在$\text{proj}_r[F]$上运行它。

Problem 2 *But what if $d_p(proj_r[F1], proj_r[F_2])$ is exponentially small?*

这会成为一个问题，因为我们需要以指数级的精度运行我们的单变量算法，只是为了看到有两个组成部分而不是一个！我们如何解决这个问题？我们将证明当F处于各向同性位置时，这个问题实际上根本不会出现。为了直观理解，考虑两种情况：

(a) 假设 $\|\mu_1 - \mu_2\| \geq$ 多项式$(1/n, \varepsilon)$。

您可以将这个条件理解为只是说 $\|\mu_1 - \mu_2\|$ 不是指数级小的。无论如何，我们知道将一个向量投影到随机方向通常会将其范数减少一个因子 $\sqrt{n}$，并且其投影长度会集中在这个值附近。这告诉我们，以高概率 $\|r^T\mu_1 - r^T\mu_2\|$ 至少是 poly(1/n, ε)。因此，proj$_r$[F_1] 和 proj$_r$[F_2] 的参数将因它们的均值差异而明显不同。

(b) 否则 $\|\mu_1 - \mu_2\| \leq$ 多项式$(1/n, \varepsilon)$。

关键思想是，如果 $d_{TV}(F_1, F_2) \geq \varepsilon$ 及其均值指数接近，那么它们在随机方向 r 上的协方差 Σ_1 和 Σ_2 必然有显著差异。在这种情况下，proj$_r$[F_1] 和 proj$_r$[F_2] 将因方差差异而有显著不同的参数。这正是以下引理背后的直觉：

Lemma 7.4.7 *If F is in isotropic position and $w_i \geq \varepsilon$ and $d_{TV}(F_1, F_2) \geq \varepsilon$, then with high probability for a direction r chosen uniformly at random*

$$d_p(proj_r[F_1], proj_r[F_2]) \geq \varepsilon_3 = \text{poly}(1/n, \varepsilon)$$

这个引理在 F 不处于各向同性位置时是错误的（例如，考虑平行煎饼的例子！）它也在推广到 $k > 2$ 高斯混合时失败，即使混合处于各向同性位置。问题在于，存在这样的例子，将投影到几乎所有的方向 r 实质上导致具有严格更少成分的混合！[114] 中的方法是学习较少高斯混合作为真实混合的代理，然后找到可以用来分离已合并的成对成分的方向。

Pairing Lemma

接下来我们将遇到第二个问题：假设我们将投影到方向 r 和 s 并分别学习 $\widehat{F}^r = \frac{1}{2}\widehat{F}_1^r + \frac{1}{2}\widehat{F}_2^r$ 和 $\widehat{F}^s = \frac{1}{2}\widehat{F}_1^s + \frac{1}{2}\widehat{F}_2^s$。然后 $\widehat{F}_1^r$ 的均值和方差对两个高维高斯分布中的一个施加线性约束，对 $\widehat{F}_1^s$ 也是类似。

Problem 3 *How do we know that they yield constraints on the* 相同 *high-dimensional component?*

最终，我们希望建立一个线性约束系统来解决 F_1 的参数，但当我们将 F 投影到不同的方向（例如，r 和 s）时，我们需要将这两个方向上的分量配对。关键观察结果是，当我们改变 r 到 s 时，混合物的参数会连续变化。见图 7.2。因此，当我们投影到 r 时，根据各向同性投影引理，我们知道这两个分量将要么具有明显不同的均值或方差。假设它们的均值相差 ε_3；那么如果 r 和 s 相对 ε_1 较接近，混合物中每个分量的参数变化不会很大，

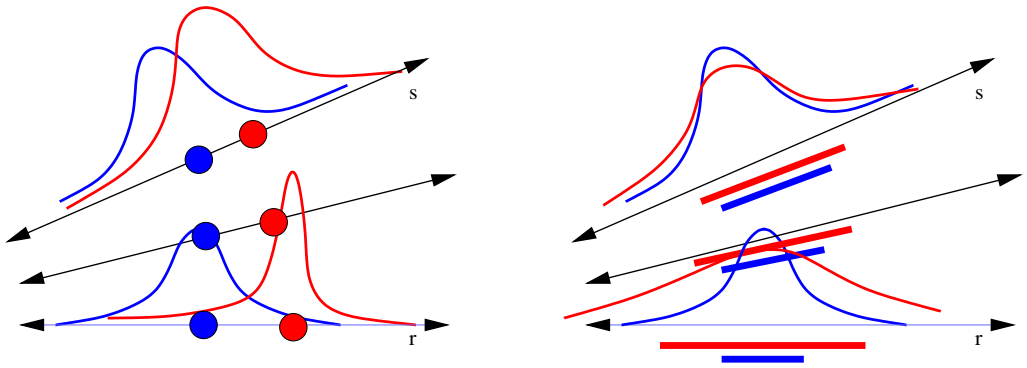


图7.2：当我们从r扫到s时，预测均值和预测方差连续变化。

组件在$\text{proj}_r[F]$中具有较大均值的将对应于$\text{proj}_s[F]$中具有较大均值的相同组件。当方差至少相差ε_3时，也适用类似的陈述。

Lemma 7.4.8 *If* $\|r - s\| \leq \varepsilon_2 =$ 多项式$(1/n, \varepsilon_3)$ *then*

(a) *If* $|r^T\mu_1 - r^T\mu 2| \geq \varepsilon 3$ *then the components in* $proj_r[F]$ *and* $proj_s[F]$ *with the larger mean correspond to the same high-dimensional component*

(b) *Else if* $|r^T\Sigma_1 r - r^T\Sigma 2r| \geq \varepsilon 3$ *then the components in* $proj_r[F]$ *and* $proj_s[F]$ *with the larger variance correspond to the same high-dimensional component*

因此，如果我们随机选择 r 并仅搜索具有 $\|r - s\| \leq \varepsilon_2$ 的方向 s，我们就能正确地将不同一维混合物中的组件配对。

Condition Number Lemma

现在我们在高维情况下遇到了最终问题：假设我们随机选择 r，对于 $s_1, s_2, \ldots, s_p$ 我们学习 F 在这些方向上的投影参数，并正确配对组件。我们只能希望在这些投影上学习参数，达到某个加性精度 ε_1 (，并且我们的单变量学习算法将具有运行时间和样本复杂度 poly$(1/\varepsilon_1)$)。

Problem 4 *How do these errors in our univariate estimates translate to errors in our high dimensional estimates for* $\mu_1, \Sigma_1, \mu_2, \Sigma_2$*?*

回忆一下，*condition number*控制这个。在高维情况下，我们需要最终的引理是：

Lemma 7.4.9 *The condition number of the linear system to solve for* $\mu_{1,}\Sigma 1$ *is* poly$(1/\varepsilon_2, n)$ *where all pairs of directions are* ε_2 *apart.*

直观上，当 r 和 $s_1, s_2, \ldots, s_p$ 更接近时，系统的条件数会更差（因为线性约束更接近冗余），但关键事实是条件数被一个关于 $1/\varepsilon_2$ 和 n 的固定多项式所界定，因此如果我们选择 $\varepsilon_1 =$ poly$(\varepsilon_2, n)\varepsilon$，那么我们对高维参数的估计将在一个加性 ε 范围内。注意，每个参数 $\varepsilon, \varepsilon_3, \varepsilon_2, \varepsilon_1$ 是早期参数（以及 $1/n$）的一个固定多项式，因此我们只需要在我们的单变量学习算法上运行多项式数量的混合，以逆多项式精度学习一个 ε-close 估计 $\widehat{F}$!

但我们仍然需要设计一个一元算法，接下来我们回到皮尔逊的原始问题!

7.5 A Univariate Algorithm

这里我们将给出一个一元算法，用于学习两个高斯混合的参数，直到加性精度 ε ，其运行时间和样本复杂度为 poly$(1/\varepsilon)$。我们的第一个观察结果是所有参数都是有界的：

Claim 7.5.1 *Let $F = w_1F_1+w_2F_2$ be a mixture of two Gaussians that is in isotropic position. Suppose that $w_1, w_2 \geq \varepsilon$. Then*

(a) $\mu_1, \mu_2 \in [-1/\sqrt{\varepsilon}, 1/\sqrt{\varepsilon}]$

(b) $\sigma_1^2, \sigma_2^2 \in [0, 1/\varepsilon]$

想法是，如果任一条件被违反，则意味着混合物的方差严格大于一。一旦我们知道参数是有界的，自然的方法是尝试网格搜索：

Grid Search

输入：来自 $F(\Theta)$ 的样本 输出：参数
$\widehat{\Theta} = (\widehat{w}_1, \widehat{\mu}_1, \widehat{\sigma}_1^2, \widehat{\mu}_2, \widehat{\sigma}_2^2)$

对于所有有效的 $\widehat{\Theta}$，其中参数是 ε^C 的倍数，使用样本
测试 $\widehat{\Theta}$，如果通过则输出 $\widehat{\Theta}$ 结束

有许多方法我们可以考虑来测试我们的估计与模型真实参数的接近程度。例如，我们可以通过经验估计

源样本中$F(\Theta)$的前六个矩，如果$\widehat{\Theta}$的前六个矩每个都在经验矩的某些加性公差τ内，则通过$\widehat{\Theta}$。（这实际上是皮尔逊第六矩检验的一个变体）。很容易看出，如果我们取足够的样本并适当地设置τ，那么如果我们将真实参数Θ四舍五入到任何有效网格点，其参数是ε^C的倍数，那么结果$\widehat{\Theta}$将很可能通过我们的测试。这被称为*completeness*。更具挑战性的部分是建立*soundness*；毕竟，为什么除了接近Θ的参数集$\widehat{\Theta}$之外，没有其他参数集可以通过我们的测试?

另一种情况，我们想要证明任何两个参数 *do not* 在加性 ε 范围内匹配的混合物 F 和 $\widehat{F}$ 必须有它们的前六个矩中有一个明显不同。主要引理是:

Lemma 7.5.2 (Six Moments Suffice) *For any F and $\widehat{F}$ that are not ε-close in parameters, there is an $r \in \{1, 2, \ldots, 6\}$ where*

$$\left|M_r(\Theta) - M_r(\widehat{\Theta})\right| \geq \varepsilon^{O(1)}$$

where Θ and $\widehat{\Theta}$ are the parameters of F and $\widehat{F}$ respectively, and M_r is the r^{th} raw moment.

让 $\widetilde{M_r}$ 为经验矩。然后

$$\left|M_r(\widehat{\Theta}) - M_r(\Theta)\right| \leq \underbrace{\left|\widetilde{M_r}(\widehat{\Theta}) - \widetilde{M_r}\right|}_{\leq\tau} + \underbrace{\left|\widetilde{M_r} - M_r(\Theta)\right|}_{\leq\tau} \leq 2\tau$$

在第一个项最多为 τ，因为测试通过，第二个项很小，因为我们可以取足够的样本（但仍然是 poly$(1/\tau)$），所以经验

瞬间和真实瞬间很接近。因此，我们可以将上述引理应用于逆命题，并得出结论：如果网格搜索输出 $\widehat{\Theta}$，那么 Θ 和 $\widehat{\Theta}$ 必须在参数上与 ε 接近，这为我们提供了一个高效的单变量算法！

所以我们的主要目标是证明，如果F和$\widehat{F}$不是ε-close，那么它们的前六个矩中至少有一个明显不同。事实上，即使$\varepsilon =$为0的情况也具有挑战性：如果F和$\widehat{F}$是两种高斯分布的不同混合，为什么它们的前六个矩中必然有一个不同？我们的主要目标是使用$heat\ equation$来证明这个陈述。

实际上，让我们考虑以下思想实验。设 $f(x) = F(x) - \widehat{F}(x)$ 为密度函数 F 和 $\widehat{F}$ 的逐点差。那么，问题的关键是：我们能否证明 $f(x)$ 至多在 x-轴上交叉六次？见图7.3。

Lemma 7.5.3 *If $f(x)$ crosses the x-axis at most six times, then one of the first six moments of F and $\widehat{F}$ are different*

Proof: 实际上，我们可以构造一个（非零）次数最多为六的多项式 $p(x)$，它与 $f(x)$ 的符号一致——即对于所有 x，有 $p(x)f(x) \geq 0$。然后

$$\begin{aligned} 0 < \Big| \int_x p(x)f(x)dx \Big| &= \Big| \int_x \sum_{r=1}^{6} p_r x^r f(x)dx \Big| \\ &\leq \sum_{r=1}^{6} |p_r| \Big| M_r(\Theta) - M_r(\widehat{\Theta}) \Big| \end{aligned}$$

如果 F 和 $\widehat{F}$ 的前六个矩量完全匹配，则右侧为零，这是矛盾的。■

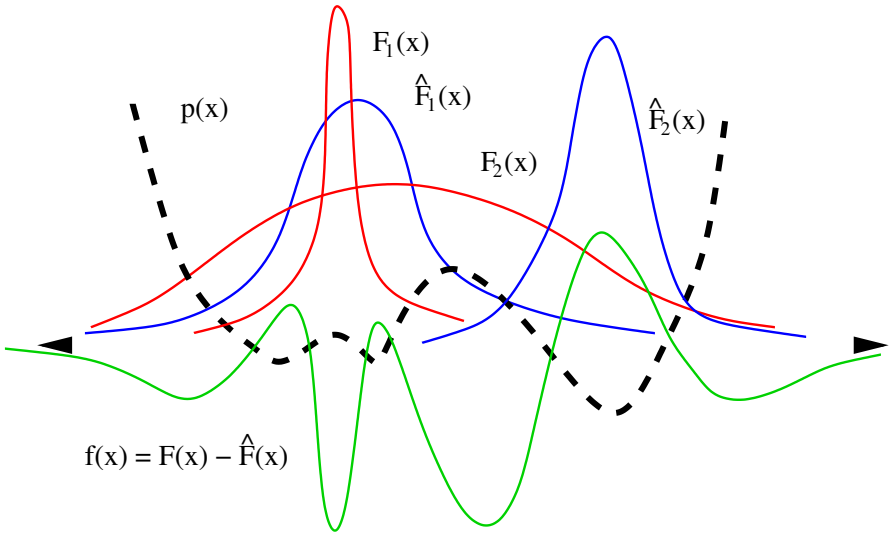


图7.3：如果$f(x)$至多有六个零点穿越，我们可以找到一个至多为六次的与它符号一致的多项式

所以我们需要证明的是 $F(x) - \widehat{F}(x)$ 至多有六个零交叉。让我们通过归纳法证明一个更强的引理:

Lemma 7.5.4 *Let* $f(x) = \sum_{i=1}^{k} \alpha_i \mathcal{N}(\mu_i, \sigma_i^2, x)$ *be a linear combination of* k *Gaussians (*α_i *can be negative). Then if* $f(x)$ *is not identically zero,* $f(x)$ *has at most* $2k - 2$ *zero crossings.*

我们将依赖以下工具:

Theorem 7.5.5 *Given* $f(x) : \mathbb{R} \to \mathbb{R}$, *that is analytic and has* n *zero crossings, then for any* $\sigma^2 > 0$, *the function* $g(x) = f(x) * \mathcal{N}(0, \sigma^2)$ *has at most* n *zero crossings.*

这个定理有物理意义。如果我们把 $f(x)$ 视为无限一维棒的温度分布，那么在某个时间之后的温度分布是什么样的呢？实际上，它恰好是适当选择的 σ^2 的 $g(x) = f(x) * \mathcal{N}(0, \sigma^2)$。或者，高斯是热方程的 *Green's function*。因此，我们对扩散的许多物理直觉对卷积也有影响——通过高斯卷积一个函数可以使其平滑，并且它不能创建新的局部极大值（并且相关地，它不能创建新的零交叉）。

最后我们回顾一个基本事实:

Fact 7.5.6 $\mathcal{N}(0, \sigma_1^2) * \mathcal{N}(0, \sigma_2^2) = \mathcal{N}(0, \sigma_1^2 + \sigma_2^2)$

现在我们准备证明上述引理并得出结论，如果我们知道两个高斯混合 *exactly* 的前六个矩，那么我们也会确切地知道其参数。让我们通过归纳法证明上述引理，并假设对于 $k = 3$ 个高斯函数的任意线性组合，零交叉的数量是

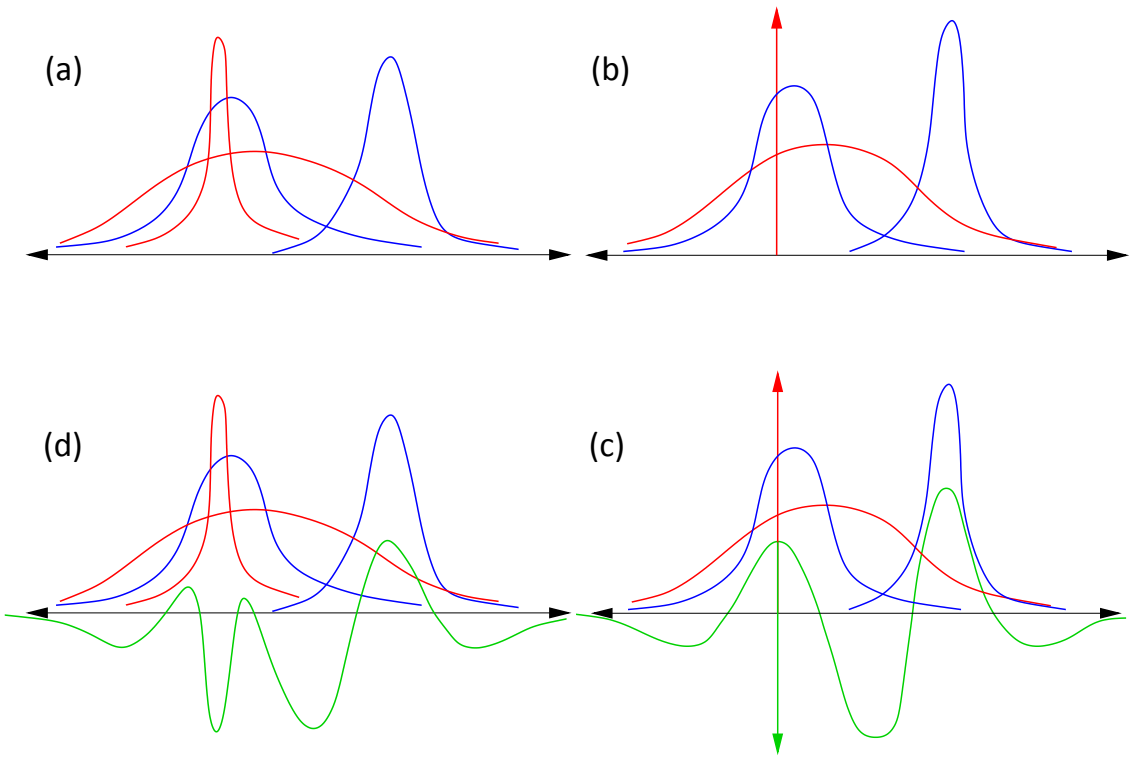


图7.4：(a) 四个高斯函数的线性组合 (b) 从每个方差中减去 σ^2 (c) 添加回 delta 函数 (d) 通过 $\mathcal{N}(0, \sigma^2)$ 卷积以恢复原始线性组合

至多四个。现在考虑四个高斯函数的任意线性组合，并令 σ^2 为任何成分的最小方差。见图7.4(a)。我们可以考虑一个相关的混合，我们从每个成分的方差中减去 σ^2 。见图7.4(b)。

现在如果我们忽略狄拉克δ函数，我们有一个三个高斯函数的线性组合，通过归纳我们知道它最多有四个零交叉。但是当我们重新加入狄拉克δ函数时，我们可以添加多少个零交叉呢？我们

最多可以添加两个，一个在上升过程中，一个在下降过程中（这里为了便于展示，我们忽略了一些与delta函数相关的实际分析复杂性）。参见图7.4(c)。现在我们可以通过$\mathcal{N}(0, \sigma^2)$卷积该函数来恢复原始的四个高斯函数的线性组合，但这一步并不会增加零交叉的数量！参见图7.4(d)。

这证明了

$$\left\{ M_r(\widehat{\Theta}) = M_r(\Theta) \right\},\ r = 1, 2, \ldots, 6$$

只有两个解（真实参数，我们也可以互换哪个是哪个分量）。事实上，这个多项式方程组也是 $stable$，并且存在多项式方程组的条件数类似物，它意味着我们刚刚证明的定量版本：如果 F 和 $\widehat{F}$ 不在 ε-close 范围内，那么它们的前六个矩中有一个明显不同。这为我们提供了我们的单变量算法。

7.6 A View from Algebraic Geometry

这里我们将介绍Belkin和Sinha[28]提出的另一种单变量学习算法，该算法也使用了矩估计方法，但使用代数几何工具进行了更一般化的分析。

Polynomial Families

我们将分析以下分布类的方法矩：

Definition 7.6.1 *A class of distributions $F(\Theta)$ is called a polynomial family if*

$$\forall r,\ \mathbb{E}_{X\in F(\Theta)}\left[X^r\right] = M_r(\Theta)$$

where $M_r(\Theta)$ is a polynomial in $\Theta = (\theta_1, \theta_2, \ldots., \theta_k)$.

这个定义捕捉了一类广泛的分布，例如成分是均匀、指数、泊松、高斯或伽马函数的混合模型。我们需要对分布添加另一个（温和的）条件，以确保它由其所有矩来表征。

Definition 7.6.2 *The moment generating function (mgf) of a random variable X is defined as*

$$f(t) = \sum_{n=0}^{\infty} \mathbb{E}\left[X^n\right] \frac{t^n}{n!}$$

Fact 7.6.3 *If the moment generating function of X converges in a neighborhood of zero, it uniquely determines the probability distribution, i.e.*

$$\forall r,\ M_r(\Theta) = M_r(\widehat{\Theta}) \implies F(\Theta) = F(\widehat{\Theta}).$$

我们的目标是证明对于任何多项式族，其*finite*个矩足够。首先，我们引入相关的定义：

Definition 7.6.4 *Given a ring R, an ideal I generated by $g_1, g_2, \cdots, g_n \in R$ denoted by $I = \langle g_1, g_2, \cdots, g_n\rangle$ is defined as*

$$I = \left\{ \sum_i r_i g_i \ \textit{where}\ r_i \in R \right\}.$$

Definition 7.6.5 *A Noetherian ring is a ring such that for any sequence of ideals*

$$I_1 \subseteq I_2 \subseteq I_3 \subseteq \cdots,$$

there is N such that $I_N = I_{N+1} = I_{N+2} = \cdots$.

Theorem 7.6.6 *[Hilbert's Basis Theorem] If R is a Noetherian ring, then $R[X]$ is also a Noetherian ring.*

它很容易看出 $\mathbb{R}$ 是一个诺特环，因此我们知道 $\mathbb{R}[x]$ 也是诺特环。现在我们可以证明，对于任何多项式族，有限个矩足以唯一识别该族中的任何分布：

Theorem 7.6.7 *Let $F(\Theta)$ be a polynomial family. If the moment generating function converges in a neighborhood of zero, there exists N such that*

$$F(\Theta) = F(\widehat{\Theta}) \text{ if and only if } M_r(\Theta) = M_r(\widehat{\Theta}) \ \forall r \in 1, 2, \cdots, N$$

Proof: 设 $Q_r(\Theta, \widehat{\Theta}) = M_r(\Theta) - M_r(\widehat{\Theta})$。设 $I_1 = \langle Q_1 \rangle$，$I_2 = \langle Q_1, Q_2 \rangle, \cdots$。这是我们 $\mathbb{R}[\Theta, \widehat{\Theta}]$ 中的理想升链。我们可以调用希尔伯特基定理并得出结论，$\mathbb{R}[X]$ 是一个诺特环，因此存在 N 使得 $I_N = I_{N+1} = \cdots$。所以对于所有 $N + j$，我们有

$$Q_{N+j}(\Theta, \widehat{\Theta}) = \sum_{i=1}^{N} p_{ij}(\Theta, \widehat{\Theta}) Q_i(\Theta, \widehat{\Theta})$$

对于某些多项式 $p_{ij} \in \mathbb{R}[\Theta, \widehat{\Theta}]$。因此，如果对于所有 $r \in 1, 2, \cdots, N$，则 $M_r(\Theta) = M_r(\widehat{\Theta})$ 对于所有 r，并且从事实7.6.3我们可以得出结论 $F(\Theta) = F(\widehat{\Theta})$。

定理的另一面显然。■

上述定理没有给出 N 的任何有限上界，因为基定理也没有。这是因为基定理是通过反证法证明的，但更基本的是，不可能只依赖于环的选择来给出 N 的上界。考虑以下例子

Example 2 *Consider the Noetherian ring* $\mathbb{R}[x]$*. Let* $I_i = \left\langle x^{N-i} \right\rangle$ *for* $i = 0, \cdots, N$*. It is a strictly ascending chain of ideals for* $i = 0, \cdots, N$*. Therefore, even if the ring* $\mathbb{R}[x]$ *is fixed, there is no universal bound on* N*.*

界限，如定理7.6.7中的那些，通常被称为*ineffective*。考虑将上述结果应用于高斯混合：从上述定理中，我们得知，如果且仅当这些混合在它们的第一N矩上相同时，任何两个F和$\widehat{F}$的k高斯混合是相同的。在这里N是k的函数，N是有限的，但我们无法使用上述工具将N作为k的函数写出任何显式的界限。尽管如此，这些工具的应用范围比我们之前用于证明上一节中$4k-2$矩足以用于k高斯混合的热方程专门工具要广泛得多。

Systems of Polynomial Inequalities

通常，我们无法精确地访问分布的矩，而只有噪声近似。我们的主要目标是证明先前结果的定量版本，该结果表明，任何两个在它们的第一个 N 矩上接近的分布 F 和 $\widehat{F}$，它们的参数也接近。关键事实是我们可以界定

多项不等式系统的条件数；有几种方法可以做到这一点，但我们将使用 *quantifier elimination*。回忆：

Definition 7.6.8 *A set S is semi-algebraic if there exist multivariate polynomials $p_1, \ldots, p_n$ such that*

$$S = \{x_1, \ldots, x_r | p_i(x_1, \ldots, x_r) \geq 0\}$$

or if S is a finite union or intersection of such sets.

当一组可以通过多项式等式定义时，我们称其为 *algebraic*

Theorem 7.6.9 *[Tarski] The projection of a semi-algebraic set is semi-algebraic.*

有趣的是，代数集的投影不一定是代数的。你能想出一个例子吗？一个投影不仅对应于通过多项式不等式定义一个集合，还对应于定义一个∃算子。结果是，你甚至可以取一个∃和∀算子的序列，得到的集合仍然是半代数的。

使用此工具，我们定义以下辅助集：

$$H(\varepsilon, \delta) = \left\{ \forall (\Theta, \widehat{\Theta}) \ : \ |M_r(\Theta) - M_r(\widehat{\Theta})| \leq \delta \text{ for } r = 1, 2, \ldots N \Longrightarrow d_p(\Theta, \widehat{\Theta}) \leq \varepsilon \right\}.$$

这里 $d_p(\Theta, \widehat{\Theta})$ 是 Θ 和 $\widehat{\Theta}$ 之间的某些参数距离。我们选择什么并不重要，只要它能通过参数的多项式表达，并且对待产生相同分布的参数相同——例如，通过取 $F(\Theta)$ 中组件与 $F(\widehat{\Theta})$ 中组件的所有匹配中的最小值并求和，来计算组件间的参数距离。

现在令 $\varepsilon(\delta)$ 为 ε 关于 δ 的最小值。使用 Tarski 定理，我们可以证明以下关于矩估计方法的稳定性界限：

Theorem 7.6.10 *There are fixed constants* C_1, C_2, s *such that if* $\delta \leq 1/C_1$ *then* $\varepsilon(\delta) \leq C_2\delta^{1/s}$.

Proof: 可以容易地看出，我们可以将 $H(\varepsilon, \delta)$ 定义为半代数集的投影，因此根据Tarski定理，我们得出结论，$H(\varepsilon, \delta)$ 也是半代数的。关键观察是，因为 $H(\varepsilon, \delta)$ 是半代数的，所以我们可以选择 ε 作为 δ 的函数的最小值本身就是 δ 的多项式函数。这里有一些注意事项，因为我们需要证明对于固定的 δ，我们可以选择 ε 大于零，并且多项式关系在 ε 和 δ 之间仅在 δ 足够小的情况下成立。然而，这些问题可以通过更多的工作得到解决，参见[28]。■

现在我们得到了主要结果：

Corollary 7.6.11 *If* $|M_r(\Theta) - M_r(\widehat{\Theta})| \leq \left(\frac{\varepsilon}{C_2}\right)^s$ *then* $d_p(\Theta, \widehat{\Theta}) \leq \varepsilon$.

因此，存在一个多项式时间算法来学习任何一元多项式族（其矩生成函数在零的邻域内收敛）的参数，其加性精度为 ε，其运行时间和样本复杂度为 poly($1/\varepsilon$)；我们可以取足够的样本来估计 N 个第一矩，并在参数网格上搜索，任何与每个矩匹配的参数集在参数距离上必然接近真实参数。

7.7 Exercises

Problem 7-1: 假设我们给定了一个由两个高斯分布组成的混合物，其中每个成分的方差相等——即

$$F(x) = w_1\mathcal{N}(\mu_1, \sigma^2, x) + (1 - w_1)\mathcal{N}(\mu_2, \sigma^2, x)$$

证明四个矩足以唯一确定混合分布的参数。

Problem 7-2: 假设我们有权访问一个预言机，对于任何方向 r，它返回投影的均值和方差，即一个分量对应的 $r^T\mu_1$ 和 $r^T\Sigma_1 r$，以及 $r^T\mu_2$ 和 $r^T\Sigma_2 r$。问题是您不知道哪些参数对应哪个分量。

(a) 设计一个算法以恢复 μ_1 和 μ_2 (，直到可以交换哪个分量是哪个)，使得最多进行 $O(d^2)$ 次查询到预言机，其中 d 是维度。*Hint:* 恢复 $(\mu_1 - \mu_2)(\mu_1 - \mu_2)^T$ 的条目。

(b) **Challenge:** 设计一个算法以恢复 Σ_1 和 Σ_2 (，直到可以交换这些组件的顺序) ，使得 $O(1)$ 查询到预言机时 $d = 2$。

请注意，我们在此处并未对投影均值或方差在某些方向r上的距离做出任何假设。

Chapter 8

Matrix Completion

在早期章节中，我们看到了稀疏性的力量。从比其维度少得多的测量中恢复稀疏向量是可能的。如果我们不知道我们的向量稀疏的基，只要有足够的例子，我们就可以学习它。但稀疏性只是开始。有许多其他方法可以使我们处理的对象具有低复杂性。在本章中，我们将研究矩阵补全问题，其目标是即使只观察到矩阵的一小部分条目，也能重建矩阵。在没有对矩阵的任何假设的情况下，这是不可能的，因为自由度太多。但是当矩阵是低秩和非相干时，结果是有简单的凸程序可以工作。你可以将这些想法进一步发展，并通过凸程序研究各种结构恢复问题，例如将矩阵分解为稀疏矩阵和低秩矩阵之和。我们在这里不会涉及这些，但会提供文献指南。

8.1 Introduction

2006年，Netflix向机器学习社区发起了一项重大挑战：将我们的推荐算法在推荐电影给用户方面的预测准确率提高超过十个百分点，我们将给你一百万美元。经过几年时间，挑战最终被攻克，Netflix也支付了奖金。在这段时间里，我们都学到了很多关于如何构建良好的推荐系统。在本章中，我们将介绍一个主要成分，即称为矩阵补全问题。

起点是将我们的预测电影评分问题建模为从观察到的矩阵中预测未观察到的条目的问题。更确切地说，如果用户 i 给电影 j (评分为一至五星级)，我们将 $M_{i,j}$ 设置为数值分数。我们的目标是使用我们观察到的条目 $M_{i,j}$ 来预测我们不知道的条目。如果我们能准确预测这些，这将为我们提供一种方式，以我们认为他们可能喜欢的方式向用户推荐电影。事先没有理由相信你可以做到这一点。如果我们考虑通过强制每个用户对每部电影进行评分而得到的整个矩阵 M（在Netflix数据集中有480, 189个用户和17, 770部电影），那么原则上我们观察到的条目 $M_{i,j}$ 可能对我们了解未观察到的条目毫无帮助。

我们处于与讨论压缩感知时相同的困境。从先验知识来看，没有理由相信你可以对向量 x 进行比其维度更少的线性测量并重建 x。我们需要的是关于结构的某种假设。在压缩感知中，我们假设 x 是稀疏的或近似稀疏的。在矩阵补全中，我们将假设 M 是低秩的或近似低秩的。重要的是要考虑这个假设的适用范围。

来自。如果 M 是低秩的，我们可以将其写为

$$M = u^{(1)}(v^{(1)})^T + u^{(2)}(v^{(2)})^T \dots u^{(r)}(v^{(r)})^T$$

希望这些排名第一的术语代表一些电影类别。例如，第一个术语可能代表类别 $drama$，而$u^{(1)}$中的条目可能代表每个用户喜欢剧情片的程度？然后$v^{(1)}$中的每个条目将代表每部电影对喜欢剧情片的人有多大的吸引力？这就是低秩假设的来源。我们希望数据中存在一些类别，使得可以填充缺失的条目。当我有一个用户对每个类别的电影的评分时，我就可以利用我从其他用户那里获得的数据，为他推荐他喜欢的类别中的其他电影。

The Model and Main Results

现在让我们正式一点。假设有 n 个用户和 m 部电影，使得 M 是一个 $n \times m$ 矩阵。令 $\Omega \subseteq [n] \times [m]$ 为我们观察到值 $M_{i,j}$ 的索引。我们的目标是，在 M 是低秩或近似低秩的假设下，填补缺失的项。问题是，在这个一般性水平上，找到与我们的观察一致的最低秩矩阵 M 是 NP-难的问题。然而，有一些现在已经标准化的假设，在这些假设下，我们将能够给出有效的算法来精确恢复 M:

(a) 我们观察到的条目是从 $[n] \times [m]$ 中均匀随机选择的

(b) M 的秩为 r

(c) M 的单变量向量与标准基无关（这样的矩阵称为 $incoherent$，我们稍后定义这个）

在这一章中，我们的主要结果是，如果 $m \geq n$ 和 rank(M) $\leq r$，则存在有效的算法可以精确恢复 M，当且仅当 $m \approx mr \log m$。这与压缩感知类似，我们能够从 $O(k \log n/k)$ 个线性测量中恢复一个 k-稀疏信号 x，这比 x 的维度小得多。在这里，我们也可以从比 M 的维度小得多的观测值中恢复一个低秩矩阵 M。

让我们检验上述假设。应该让我们感到不安的假设是Ω是均匀随机的。这在某种程度上是不自然的，因为如果我们观察到的概率$M_{i,j}$取决于其本身值，那么这会更可信。或者，用户更有可能对电影$if\ he\ actually\ liked\ it$进行评分。

我们已经讨论了第二个假设。为了理解第三个假设，假设我们的观察确实是均匀随机的。考虑

$$M = \Pi \left[\begin{array}{c|c} I_r & 0 \\ \hline 0 & 0 \end{array}\right] \Pi^T$$

Π 是一个均匀随机排列矩阵。M 是低秩的，但除非我们观察到对角线上的所有元素，否则我们无法唯一地恢复 M。实际上，M 的前几个奇异向量是标准基向量；但如果我们假设 M 的奇异向量与标准基不相关，我们就可以避免这个陷阱，因为我们在 M 的低秩分解中的向量分布在许多行和列上。

Definition 8.1.1 *The coherence μ of a subspace $U \subseteq \mathbb{R}^n$ of dimension $dim(u) = r$ is*

$$\frac{n}{r} \max_i \|P_U e_i\|^2,$$

where P_U denotes the orthogonal projection onto U, and e_i is the standard basis element.

它很容易看出，如果我们随机均匀地选择 U，那么 $\mu(U) = \widetilde{O}(1)$。此外，我们还有 $1 \leq \mu(U) \leq n/r$，并且当 U 包含任何 e_i 时达到上界。现在我们可以看到，如果我们把 U 设置为上述示例中的顶部奇异向量，那么 U 具有高一致性。我们将在 M 上需要以下条件：

(a) 设 $M = U\Sigma V^T$，则 $\mu(U), \mu(V) \leq \mu_0$。

(b) $\|UV^T\|_\infty \leq \frac{\mu_1\sqrt{r}}{\sqrt{nm}}$，其中 $||\cdot||_\infty$ 表示任何条目最大绝对值。

本章的主要结果是：

Theorem 8.1.2 *Suppose* Ω {v*}

$$|\Omega| \geq C \max(\mu_1^2, \mu_0) r(n+m) \log^2(n+m)$$

该定理中的算法基于对矩阵秩的凸松弛，称为*nuclear norm*。我们将在下一节中介绍它，并建立其一些性质，但可以将其视为我们在压缩感知中使用的ℓ_1最小化方法的类似物。这种方法是

首次在Fazel的论文[70]中提出，Recht、Fazel和Parrilo [124]证明了这种方法在 *matrix sensing*的设置中精确恢复M，这与我们在此处考虑的问题相关。

在一篇里程碑式的论文中，Candes和Recht [41] 证明了基于核范数的松弛方法也适用于矩阵补全，并引入了上述假设以证明他们的算法是有效的。此后，有一系列工作改进了对 m 的要求，上述定理和我们的阐述将遵循Recht [123] 的最近论文，该论文通过利用Bernstein界矩阵类似物并使用这些类似物在现在称为 *quantum golfing* 的程序中进行，大大简化了分析，该程序最初由Gross [80] 提出。

Remark 8.1.3 *We will restrict to* $M \in \mathbb{R}^{n\times n}$ *and assume* $\mu_0, \mu 1 = \widetilde{O}(1)$ *in our analysis, which will reduce the number of parameters we need to keep track of.*

8.2 Nuclear Norm

这里我们介绍核范数，它将成为我们矩阵补全算法的基础。我们将遵循与压缩感知平行的提纲。特别是，一个自然的起点是以下优化问题:

$$(P_0) \qquad \min \operatorname{rank}(X) \text{ s.t. } X_{i,j} = M_{i,j} \text{ for all } (i,j) \in \Omega$$

这个问题是 NP-难。如果 $\sigma(X)$ 是 X 的奇异值向量，那么我们可以将 X 的秩等价地视为 $\sigma(X)$ 的稀疏性。回想一下，在压缩感知中，我们面临了类似的障碍: 寻找最稀疏的解

一个线性方程组也是 NP-难，但我们考虑了 ℓ_1 松弛，并证明了在各种条件下，这个优化问题可以恢复最稀疏的解。同样，考虑 $\sigma(X)$ 的 ℓ_1-范数是自然的，这被称为核范数：

Definition 8.2.1 *The nuclear norm of X denoted by $\|X\|_*$ is $\|\sigma(X)\|_1$.*

我们将解决凸规划问题：

$$(P_1) \qquad \min \|X\|_* \text{ s.t. } X_{i,j} = M_{i,j} \text{ for all } (i,j) \in \Omega$$

我们的目标是证明满足以下条件时，(P_1) 的解恰好是 M。请注意，这是一个凸规划问题，因为 $\|X\|_*$ 是一个范数，并且存在多种高效的算法来解决上述规划问题。

实际上，就我们的目的而言，一个关键的概念是 *dual norm* 的概念。我们不需要这个概念在完全一般的情况下，所以我们只对核范数的特定情况进行陈述。这个概念为我们提供了一个方法来降低矩阵的核范数的下界：

Definition 8.2.2 *Let $\langle X, B\rangle = \sum_{i,j} X_{i,j}B_{i,j} = trace(X^T B)$ denote the matrix inner-product.*

Lemma 8.2.3 $\|X\|_* = \max_{\|B\|\leq 1}\langle X, B\rangle$.

为了感受这一点，考虑将 X 和 B 限制为对角线的情况。此外，设 X = 为 diag(x)，B = 为 diag(b)。然后 $\|X\|_* = \|x\|_1$ 和约束 $\|B\| \leq 1$（B 的谱范数至多为 1）等价于

$\|b\|_\infty \leq 1$. 因此，在对角矩阵的特殊情况下，我们可以恢复向量范数的更熟悉的特征描述：

$$\|x\|_1 = \max_{\|b\|_\infty \leq 1} b^T x$$

Proof: 我们只证明上述引理的一个方向。我们应该使用什么 B 来证明 X 的核范数。设 $X = U_X \Sigma_X V_X^T$，然后我们将选择 $B = U_X V_X^T$。然后

$$\langle X, B \rangle = \text{trace}(B^T X) = \text{trace}(V_X U_X^T U_X \Sigma_X V_X^T) = \text{trace}(V_X \Sigma_X V_X^T) = \text{trace}(\Sigma_X) = \|X\|_*$$

在本文中，我们使用了基本事实：tr(ABC) = tr(BCA)。因此，这证明了 $\|X\|_* \leq$ max$_{\|B\| \leq 1} \langle X, B \rangle$，而另一个方向并不困难得多（例如，参见[88]）。■

如何证明 (P_1) 的解是 M? 我们的基本方法将是反证法。假设不是这样，那么对于某个在 Ω 中有支撑的 Z，解是 $M + Z$。我们的目标将是构造一个谱范数至多为一的矩阵 B，使得

$$\|M + Z\|_* \geq \langle M + Z, B \rangle > \|M\|_*$$

因此 $M + Z$ 不会是 (P_1) 的最优解。这种策略与压缩感知中的策略类似，我们假设了一些其他解 w，该解与 x 在感知矩阵 A 的核 y 中不同。当时我们的策略是利用 ker(A) 的几何性质来证明 w 的 ℓ_1 范数比 x 严格更大。这里的证明将具有相同的风格，但将大大更

技术和复杂的。

让我们介绍一些基本投影算子，这些算子在我们证明中将至关重要。回忆一下，$M = U\Sigma V^T$，设$u_1, \ldots, u_r$是U的列，设$v_1, \ldots, v_r$是V的列。选择$u_{r+1}, \ldots, u_n$，使得$u_1, \ldots, u_n$构成$\mathbb{R}^n$的正规正交基，即$u_{r+1}, \ldots, u_n$是$U^\perp$的任意正规正交基。类似地，选择$v_{r+1}, \ldots, v_n$，使得$v_1, \ldots, v_n$构成$\mathbb{R}^n$的正规正交基。我们将对以下矩阵上的线性空间感兴趣：

Definition 8.2.4 $T = span\{u_i v_j^T \mid 1 \le i \le r \ or \ 1 \le j \le r \ or \ both\}$.

然后 $T^\perp = \text{span}\{u_i v_j^T \text{ s.t. } r+1 \le i, j \le n\}$.。我们有 dim($T$) $= r^2 + 2(n-r)r$ 和 dim ($T^\perp$) $= (n-r)^2$。此外，我们可以定义将分别投影到 T 和 $T^\perp$ 的线性算子：

$$P_{T^\perp}[Z] = \sum_{i=r+1}^{n} \sum_{j=r+1}^{n} \langle Z, u_i v_j^T \rangle \cdot u_i v_j^T = P_{U^\perp} Z P_{V^\perp}.$$

同样地

$$P_T[Z] = \sum_{(i,j) \in [n]\times[n] - [r+1,n]\times[r+1,n]} \langle Z, u_i v_j^T \rangle \cdot u_i v_j^T = P_U Z + Z P_V - P_U Z P_V.$$

我们现在准备描述定理8.1.2证明的大纲。证明将基于：

(a) 我们将假设存在一个特定的辅助矩阵 Y，并展示这足以对任何在 Ω 中支持的 Z 推出 $\|M + Z\|_* > \|M\|_*$

(b) 我们将使用量子高尔夫[80]构建这样的 Y。

Conditions for Exact Recovery

这里我们将陈述对辅助矩阵 Y 需要满足的条件，并证明如果存在这样的 Y，则 M 是 (P_1) 的解。我们要求 Y 在 Ω 上有支撑，并且

(a) $\|P_T(Y) - UV^T\|_F \leq \sqrt{r/8n}$
(b) $\|P_{T^\perp}(Y)\| \leq 1/2$.

我们想要证明对于任何在Ω中支持的Z，$\|M + Z\|_* > \|M\|_*$。回想，我们想要找到一个谱范数不超过一的矩阵B，使得$\langle M+Z, B\rangle > \|M\|_*$。设$U_\perp$和$V_\perp$是$P_{T^\perp}$[$Z$]的奇异向量。然后考虑

$$B = \begin{bmatrix} U & U_\perp \end{bmatrix} \cdot \begin{bmatrix} V^T \\ V_\perp^T \end{bmatrix} = UV^T + U_\perp V_\perp^T.$$

Claim 8.2.5 $\|B\| \leq 1$

Proof: 通过构造 $U^T U_\perp = 0$ 和 $V^T V_\perp = 0$，因此上述 B 的表达式是其奇异值分解，并且现在可以得出结论。■

因此，我们可以将我们的选择 B 代入并简化：

$$\begin{aligned} \|M + Z\|_* &\geq \langle M + Z, B\rangle \\ &= \langle M + Z, UV^T + U_\perp V_\perp^T\rangle \\ &= \underbrace{\langle M, UV^T\rangle}_{\|M\|_*} + \langle Z, UV^T + U_\perp V_\perp^T\rangle \end{aligned}$$

在最后一行中，我们使用了M与$U_\perp V_\perp^T$正交的事实。现在利用Y和Z具有不相交的支持的事实，我们可以得出：

$$\|M+Z\|_* \geq \|M\|_* + \langle Z, UV^T + U_\perp V_\perp^T - Y\rangle$$

因此，为了证明本节的主要结果，只需证明 $\langle Z, UV^T + U_\perp V_\perp^T - Y\rangle > 0$。我们可以将这个量在它投影到 T 和 $T^\perp$ 上的基础上进行展开，并简化如下：

$$\begin{aligned}\|M+Z\|_* - \|M\|_* &\geq \langle P_T(Z), P_T(UV^T + U_\perp V_\perp^T - Y)\rangle + \langle P_{T^\perp}(Z), P_{T^\perp}(UV^T + U_\perp V_\perp^T - Y)\rangle \\ &\geq \langle P_T(Z), UV^T - P_T(Y)\rangle + \langle P_{T^\perp}(Z), U_\perp V_\perp^T - P_{T^\perp}(Y)\rangle \\ &\geq \langle P_T(Z), UV^T - P_T(Y)\rangle + \|P_{T^\perp}(Z)\|_* - \langle P_{T^\perp}(Z), P_{T^\perp}(Y)\rangle\end{aligned}$$

在最后一行中，我们使用了$U_\perp$和$V_\perp$是$P_{T^\perp}[Z]$的奇异向量的事实，因此$\langle U_\perp V_\perp^T, P_{T^\perp}[Z]\rangle = \|P_{T^\perp}[Z]\|_*$。

现在我们可以调用本节中假设的 Y 的属性，以证明右侧的下界。根据 Y 的属性 (a)，我们有 $\|P_T(Y) - UV^T\|_F \leq \sqrt{\frac{r}{2n}}$。因此，我们知道第一项 $\langle P_T(Z), UV^T - P_T(Y)\rangle \geq -\sqrt{\frac{r}{8n}}\|P_T(Z)\|_F$。根据 Y 的属性 (b)，我们知道 $P_T^\perp(Y)$ 的算子范数不超过 1/2。因此，第三项 $\langle P_{T^\perp}(Z), P_{T^\perp}(Y)\rangle$ 至多为 $\frac{1}{2}\|P_{T^\perp}(Z)\|_*$。因此

$$\|M+Z\|_* - \|M\|_* \geq -\sqrt{\frac{r}{8n}}\|P_T(Z)\|_F + \frac{1}{2}\|P_{T^\perp}(Z)\|_* \overset{?}{>} 0$$

我们将证明，在Ω的选择上，以高概率满足以下不等式

确实成立。我们推迟这个最后事实的证明，因为它和辅助矩阵 Y 的构造都将使用我们在下一节中提出的矩阵Bernstein不等式。

8.3 Quantum Golfing

剩余的任务是构建一个辅助矩阵 Y 并证明在 Ω 上的高概率下，对于任何在 Ω 中支持的矩阵 Z，以完成我们在上一节开始的证明。我们将使用 Gross [80] 介绍的方法，并遵循 Recht 在 [123] 中的证明，其中策略是迭代构建 Y。在每一阶段，我们将调用矩阵值随机变量的集中结果来证明 Y 的误差部分以几何级数减少，并在构建良好的辅助矩阵方面取得快速进展。

首先，我们将介绍将在几个设置中应用的关键浓度结果。以下矩阵值伯恩斯坦不等式首次出现在Ahlswede和Winter关于量子信息理论的工作中[6]。

Theorem 8.3.1 *[Non-commutative Bernstein Inequality] Let* $X_1 \dots X_l$ *be independent mean 0 matrices of size* $d \times d$. *Let* $\rho^2_k = \max\{\|\mathbb{E}[X_k X_k^T]\|, \|\mathbb{E}[X_k^T X_k]\|\}$ *and suppose* $\|X_k\| \leq M$ *almost surely. Then for* $\tau > 0$,

$$\Pr\left[\left\|\sum_{k=1}^{l} X_k\right\| > \tau\right] \leq 2d \exp\left\{\frac{-\tau^2/2}{\sum_k \rho_k^2 + M\tau/3}\right\}$$

如果 $d = 1$ 这就是标准的Bernstein不等式。如果 $d > 1$ 并且矩阵 X_k 是

对角线，然后可以从并集界和标准的伯恩斯坦不等式再次得到这个不等式。然而，为了建立直观，考虑以下玩具问题。设u_k是$\mathbb{R}^d$中的随机单位向量，并设$X_k = u_k u_k^T$。那么很容易看出$\rho_k^2 = 1/d$。我们需要多少次试验才能使$\sum_k X_k$接近单位矩阵（在缩放后）？我们应该预期需要$\Theta(d \log d)$次试验；这甚至对于u_k从标准基向量$\{e_1 \dots e_d\}$中均匀抽取也是正确的，因为存在优惠券收集问题。事实上，上述界证实了我们的直觉，即$\Theta(d \log d)$是必要且充分的。

现在我们将应用上述不等式来构建完成证明所需的工具。

Definition 8.3.2 *Let $R_\Omega$be the operator that zeros out all the entries of a matrix except those in Ω.*

Lemma 8.3.3 *If* Ω {v*} log {v*} {v*}

$$\frac{n^2}{m} \left\| P_T R_\Omega P_T - \frac{m}{n^2} P_T \right\| < \frac{1}{2}$$

Remark 8.3.4 *Here we are interested in bounding the operator norm of a linear operator on matrices. Let T be such an operator, then $\|T\|$ is defined as*

$$\max_{\|Z\|_F \leq 1} \|T(Z)\|_F$$

我们将解释这个界限如何融入矩阵Bernstein不等式的框架中，但完整的证明请参见[123]。请注意，$\mathbb{E}[P_T R_\Omega P_T] = P_T \mathbb{E}[R_\Omega] P_T = \frac{m}{n^2} P_T$，所以我们只需证明$P_T R_\Omega P_T$没有偏离得太远。

期望。设 $e_1, e_2, \ldots, e_d$ 为标准基向量。然后我们可以展开:

$$\begin{aligned} P_T(Z) &= \sum_{a,b} \langle P_T(Z), e_a e_b^T \rangle e_a e_b^T \\ &= \sum_{a,b} \langle Z, P_T(e_a e_b^T) \rangle e_a e_b^T \end{aligned}$$

因此 $R_\Omega P_T(Z) = \sum_{(a,b)\in\Omega} \left\langle Z, P_T(e_a e_b^T) \right\rangle e_a e_b^T$，最后我们得出结论

$$P_T R_\Omega P_T(Z) = \sum_{(a,b)\in\Omega} \left\langle Z, P_T(e_a e_b^T) \right\rangle P_T(e_a e_b^T)$$

我们可以将 $P_T R_\Omega P_T$ 视为形式为 $\tau_{a,b} : Z \to \left\langle Z, P_T(e_a e_b^T) \right\rangle P_T(e_a e_b^T)$ 的随机算子的和，通过将矩阵Bernstein不等式应用于随机算子 $\sum_{(a,b)\in\Omega} \tau_{a,b}$，引理随之得出。

我们现在可以完成部分(a)的延迟证明:

Lemma 8.3.5 *If Ω is chosen uniformly at random and $m \geq nr \log n$ then with high probability for any Z supported in $\overline{\Omega}$ we have*

$$\|P_{T^\perp}(Z)\|_* > \sqrt{\frac{r}{2n}} \|P_T(Z)\|_F$$

Proof: 使用引理8.3.3和算子范数的定义（参见注释）

我们有

$$\left\langle Z, P_T R_\Omega P_T Z - \frac{m}{n^2} P_T Z \right\rangle \geq -\frac{m}{2n^2} \|Z\|_F^2$$

此外，我们可以将左侧上界为:

$$\begin{aligned}\langle Z, P_T R_\Omega P_T Z\rangle = \langle Z, P_T R_\Omega^2 P_T Z\rangle &= \|R_\Omega(Z - P_{T^\perp}(Z))\|_F^2 \\ &= \|R_\Omega(P_{T^\perp}(Z))\|_F^2 \leq \|P_{T^\perp}(Z)\|_F^2\end{aligned}$$

在最后一行中我们使用了 Z 在 Ω 中是支持的，因此 $\overline{R_\Omega}(Z) = 0$。因此我们有

$$\|P_{T^\perp}(Z)\|_F^2 \geq \frac{m}{n^2}\|P_T(Z)\|_F^2 - \frac{m}{2n^2}\|Z\|_F^2$$

我们可以使用事实 $\|Z\|_F^2 = \|P_{T^\perp}(Z)\|_F^2 + \|P_T(Z)\|_F^2$ 并得出结论 $\|P_{T^\perp}(Z)\|_F^2 \geq \frac{m}{4n^2}\|P_T(Z)\|_F^2$。现在

$$\begin{aligned}\|P_{T^\perp}(Z)\|_*^2 \geq \|P_{T^\perp}(Z)\|_F^2 &\geq \frac{m}{4n^2}\|P_T(Z)\|_F^2 \\ &> \frac{r}{2n}\|P_T(Z)\|_F^2\end{aligned}$$

该命题得证。■

所有剩下的就是证明我们使用的辅助矩阵 Y 确实存在（以高概率）。回想一下，我们要求 Y 在 Ω 中有支撑，并且 $\|P_T(Y) - UV^T\|_F \leq \sqrt{r/8n}$ 和 $\|P_{T^\perp}(Y)\| \leq 1/2$。基本思路是将 Ω 分解为不相交的集合 $\Omega_1, \Omega_2, \ldots \Omega_p$，其中 $p = \log n$，并使用每个观测集来推进剩余的 $P_T(Y) - UV^T$。更确切地说，初始化 $Y_0 = 0$，在这种情况下，剩余的是 $W_0 = UV^T$。然后设置

$$Y_{i+1} = Y_i + \frac{n^2}{m}R_{\Omega_{i+1}}(W_i)$$

并且更新 $W_{i+1} = UV^T - P_T(Y_{i+1})$。很容易看出 $\mathbb{E}[\frac{n^2}{m}R_{\Omega_{i+1}}] = I$。直观上这意味着在每一步 $Y_{i+1} - Y_i$ 是 W_i 的无偏估计量，因此我们应该期望剩余部分迅速减少（这里我们将依赖于我们从非交换伯恩斯坦不等式推导出的集中界限）。现在我们可以解释命名法 *quantum golfing*；在每一步，我们朝着洞的方向击高尔夫球，但在这里我们的目标是逼近矩阵 UV^T，由于各种原因，这类问题是量子力学中出现的类型。

它很容易看出 $Y = \sum_i Y_i$ 在 Ω 中得到支持，并且对于所有 i，有 $P_T(W_i) = W_i$。因此，我们可以计算

$$\begin{aligned}\|P_T(Y_i) - UV^T\|_F &= \left\| P_T\frac{n^2}{m}R_{\Omega_i}W_{i-1} - W_{i-1}\right\|_F = \left\| P_T\frac{n^2}{m}R_{\Omega_i}P_TW_{i-1} - P_TW_{i-1}\right\|_F \\ &= \frac{n^2}{m}\left\| P_TR_\Omega P_T - \frac{m}{n^2}P_T\right\| \le \frac{1}{2}\|W_{i-1}\|_F\end{aligned}$$

最后的不等式由引理8.3.3得出。因此，余项的Frobenius范数呈几何级数下降，并且很容易保证Y满足条件(a)。

更技术性的部分是证明 Y 也满足条件 (b)。然而直觉是 $\|P_{T^\perp}(Y_1)\|$ 本身并不太大，由于余项 W_i 的范数以几何级数减少，我们预计 $\|P_{T^\perp}(Y_i)\|$ 也会如此，因此对贡献的大部分应该来自

$$\|P_{T^\perp}(Y)\| \le \sum_i \|P_{T^\perp}(Y_i)\|$$

来自第一项。有关详细信息，请参阅[123]。这完成了证明，计算凸程序的解确实找到M *exactly*，前提是

那 M 不连贯且 $|\Omega| \geq \max(\mu_1^2, \mu_0) r(n+m) \log^2(n+m)$。

Further Remarks

有许多其他矩阵补全的方法。上述论点在技术上的复杂性在于我们想要解决精确的矩阵补全问题。当我们的目标是恢复对 M 的近似时，就更容易展示 (P_1) 的性能界限。Srebro 和 Shraibman [132] 使用 Rademacher 复杂度和矩阵集中界限来证明 (P_1) 恢复的解接近 M。此外，他们的论点可以简单地扩展到当 M 仅在逐项上接近低秩时，在实践上更有相关性的情况。Jain 等人 [93] 和 Hardt [83] 为交替最小化提供了可证明的保证。这些保证在依赖 M 的连贯性、秩和条件数方面更差，但交替最小化具有更好的运行时间和空间复杂度，并且在实践中是最受欢迎的方法。Barak 和 Moitra [26] 研究了噪声张量补全，并表明可以比简单地将其展平为矩阵更好地补全张量，并基于拒绝随机约束满足问题的难度展示了下界。

在矩阵补全工作之后，凸规划已被证明在许多其他相关问题上很有用，例如将矩阵分解为低秩和稀疏部分的和[44]。Chandrasekaran等人[46]为分析线性逆问题的凸规划提供了一个通用框架，并将其应用于许多设置。一个有趣的方向是使用约简和凸规划层次结构作为探索计算与统计权衡的框架[29, 45, 24]。

Bibliography

[1] D. Achlioptas 和 F. McSherry. 关于分布混合的谱学习方法。在 *COLT*，第 458–469 页，2005 年。 [2] A. Agarwal，A. Anandkumar，P. Jain，P. Netrapalli，R. Tandon 通过交替最小化学习稀疏使用的完备字典 *arXiv:1310.7991*，2013 年 [3] A. Agarwal，A. Anandkumar，P. Netrapalli 稀疏使用完备字典的精确恢复 *arXiv:1309.1952*，2013 年 [4] M. Aharon。 *Overcomplete Dictionaries for Sparse Representation of Signals*。博士论文，2006 年。 [5] M. Aharon，M. Elad 和 A. Bruckstein。K-SVD：一种用于设计稀疏表示完备字典的算法。*IEEE Trans. on Signal Processing*，54(11)：4311–4322，2006 年。 [6] R. Ahlswede 和 A. Winter。通过量子信道进行识别的强逆。 *IEEE Trans. Inf. Theory* 48(3)：569–579，2002 年。 [7] Noga Alon。 *Tools from Higher Algebra.* 在 *Handbook of Combinatorics*，第 1749–1783 页，1996 年。

[8] A. Anandkumar, D. Foster, D. Hsu, S. Kakade, Y. Liu. 基于谱的潜在狄利克雷分配算法。在 *NIPS*，第 926–934 页，2012 年。[9] A. Anandkumar, R. Ge, D. Hsu 和 S. Kakade。学习混合成员社区模型的张量谱方法。在 *COLT*，第 867–881 页，2013 年。[10] A. Anandkumar, D. Hsu 和 S. Kakade。隐马尔可夫模型和多视角混合模型的矩方法。在 *COLT*，第 33.1–33.34 页，2012 年。[11] J. Anderson, M. Belkin, N. Goyal, L Rademacher 和 J. Voss。越多越好：高维学习大高斯混合的祝福。*arXiv:1311.2891*，2013 年。[12] S. Arora, R. Ge, Y. Halpern, D. Mimno, A. Moitra, D. Sontag, Y. Wu 和 M. Zhu。具有可证明保证的主题建模实用算法。在 *ICML*，第 280–288 页，2013 年。[13] S. Arora, R. Ge, R. Kannan 和 A. Moitra。计算非负矩阵分解 – 可证明的 In *STOC*，第 145–162 页，2012 年。[14] S. Arora, R. Ge 和 A. Moitra。学习主题模型 - 超越 SVD。在 *FOCS*，第 1–10 页，2012 年。[15] S. Arora, R. Ge 和 A. Moitra。学习非一致和过度完备字典的新算法。*arXiv:1308.6273*，2013 年。[16] S. Arora, R. Ge, T. Ma 和 A. Moitra。用于稀疏编码的简单、高效和神经算法。在 *COLT*，第 113–149 页，2015 年。

[17] S. Arora, R. Ge, A. Moitra 和 S. Sachdeva. 具有未知高斯噪声的可证明独立成分分析及其在高斯混合模型和自编码器中的意义。在 *NIPS*，第 2384–2392 页，2012。

[18] S. Arora, R. Ge, S. Sachdeva 和 G. Schoenebeck. 在社交网络中寻找重叠社区：迈向严格的方法。在 *EC*，2012。

[19] S. Arora 和 R. Kannan. 学习分离的非球形高斯混合模型。*Annals of Applied Probability*，第 69-92 页，2005 年。

[20] M. Balcan, A. Blum 和 A. Gupta. 近似稳定性下的聚类。*Journal of the ACM*，2013。

[21] M. Balcan, A. Blum 和 N. Srebro. 关于具有相似度函数的学习理论。*Machine Learning*，第 89–112 页，2008 年。

[22] M. Balcan, C. Borgs, M. Braverman, J. Chayes 和 S-H Teng. 寻找内生形成的社区。在 *SODA*，2013。

[23] A. Bandeira, P. Rigollet 和 J. Weed. 多参考对齐的估计最优速率。*arXiv:1702.08546*，2017。

[24] B. Barak, S. Hopkins, J. Kelner, P. Kothari, A. Moitra 和 A. Potechin. 对于植入 clique 问题的一个几乎紧的平方和下界。在 *FOCS*，第 428–437 页，2016 年。

[25] B. Barak, J. Kelner 和 D. Steurer. 通过平方和法进行字典学习和张量分解。在 *STOC*，第 143–151 页，2015 年。

[26] B. Barak 和 A. Moitra. 通过平方和层次结构进行噪声张量补全。在 *COLT*，第 417–445 页，2016 年。

[27] M. Belkin 和 K. Sinha. 向学习任意分离的高斯混合模型迈进。在 *COLT*，第 407–419 页，2010 年。

[28] M. Belkin 和 K. Sinha. 分布族的多项式学习。在 *FOCS*，第 103–112 页，2010 年。

[29] Q. Berthet 和 P. Rigollet. 稀疏主成分检测的复杂度理论下界。在 *COLT*，第 1046–1066 页，2013。

[30] A. Bhaskara，M. Charikar和A. Vijayaraghavan. 张量分解的唯一性与多项式可识别性应用。在 *COLT*，第742–778页，2014年。[31] A. Bhaskara，M. Charikar，A. Moitra和A. Vijayaraghavan。张量分解的平滑分析。在 *STOC*，第594–603页，2014年。[32] Y. Bilu和N. Linial。稳定的实例容易吗？在 *Combinatorics, Probability and Computing*，第21(5):643–660，2012年。[33] V. Bittorf，B. Recht，C. Re和J. Tropp。使用线性规划分解非负矩阵。在 *NIPS*，2012年。[34] D. Blei。概率主题模型导论。*Communications of the ACM*，第77–84页，2012年。

[35] D. Blei 和 J. Lafferty. 科学相关主题模型。*Annals of Applied Statistics*，第 17–35 页，2007 年。

[36] D. Blei, A. Ng 和 M. Jordan. 潜在狄利克雷分配。*Journal of Machine Learning Research*，第 993–1022 页，2003 年。

[37] A. Blum, A. Kalai 和 H. Wasserman. 噪声容忍学习，奇偶性问题以及统计查询模型。*Journal of the ACM* 50: 506-519, 2003. [38] A. Blum 和 J. Spencer. 随机和半随机 k-着色图。在 *Journal of Algorithms*，19(2):204–234，1995。 [39] K. Borgwardt. *The Simplex Method: a probabilistic analysis* Springer，2012。 [40] S. C. Brubaker 和 S. Vempala. 各向同性 PCA 和仿射不变聚类。在 *FOCS*，第 551–560 页，2008。 [41] E. Candes 和 B. Recht. 通过凸优化实现精确矩阵补全。
Foundations of Computational Math.，第 717–772 页，2008。 [42] E. Candes，J. Romberg 和 T. Tao. 从不完整和不精确的测量中恢复稳定信号。
Communications of Pure and Applied Math.，第 1207–1223 页，2006。 [43] E. Candes 和 T. Tao. 通过线性规划解码。*IEEE Trans. on Information Theory*，51(12): 4203–4215，2005。 [44] E. Candes，X. Li，Y. Ma 和 J. Wright. 鲁棒主成分分析? *Journal of the ACM*，58(3):1–37，2011。 [45] V. Chandrasekaran 和 M. Jordan. 通过凸松弛实现的计算和统计权衡。
Proceedings of the National Academy of Sciences，110(13)E1181–E1190，2013。

[46] V. Chandrasekaran, B. Recht, P. Parrilo 和 A. Willsky. 线性逆问题的凸几何学。*Foundations of Computational Math.*，12(6)805–849，2012。[47] J. Chang. 在进化树上的马尔可夫模型的全重建：可识别性和一致性。*Mathematical Biosciences*，137(1):51–73，1996。[48] K. Chaudhuri 和 S. Rao。使用相关性和独立性学习产品分布的混合。在 *COLT*，第 9–20 页，2008。[49] K. Chaudhuri 和 S. Rao。超越高斯：学习重尾分布混合的谱方法。在 *COLT*，第 21–32 页，2008。[50] S. Chen，D. Donoho 和 M. Saunders。基于基追踪的原子分解。*SIAM J. on Scientific Computing*，20(1):33–61，1998。[51] A. Cohen，W. Dahmen 和 R. DeVore。压缩感知和最佳 k-项逼近。*Journal of the AMS*，第 211–231 页，2009。[52] J. Cohen 和 U. Rothblum。非负秩，非负矩阵的分解和因子分解。*Linear Algebra and its Applications*，第 149–168 页，1993。[53] P. Comon。独立成分分析：一个新概念？*Signal Process- ing*，第 287–314 页，1994。[54] A. Dasgupta。*Asymptotic Theory of Statistics and Probability.* Springer，2008。[55] A. Dasgupta，J. Hopcroft，J. Kleinberg 和 M. Sandler。关于学习重尾分布混合。在 *FOCS*，第 491–500 页，2005。

[56] S. Dasgupta. 高斯混合学习。在 *FOCS*，第 634–644 页，1999。

[57] S. Dasgupta 和 L. J. Schulman. 高斯混合模型的EM两轮变体。在 *UAI*，第 152-159 页，2000 年。

[58] G. Davis, S. Mallat 和 M. Avellaneda. Greedy adaptive approximations. *J. of Constructive Approximation*, 13:57–98, 1997.

[59] L. De Lathauwer, J Castaing 和 J. Cardoso. 基于四阶累积量的欠定混合盲识别。*IEEE Trans. on Signal Pro- cessing*，55(6):2965–2973，2007。[60] S. Deerwester, S. Dumais，T. Landauer，G. Furnas 和 R. Harshman。基于潜在语义分析的索引。*JASIS*，第 391–407 页，1990。[61] A.P. Dempster，N.M. Laird，和 D.B. Rubin。通过 EM 算法从不完全数据中估计最大似然。
Journal of the Royal Statistical Society Series B，第 1–38 页，1977。[62] D. Donoho 和 M. Elad。通过 ℓ_1-最小化在一般（非正交）字典中的最优稀疏表示。*PNAS*，100(5):2197–2202，2003。[63] D. Donoho 和 X. Huo。不确定性原理和理想原子分解。*IEEE Trans. on IT*，47(7):2845–2862，1999。

[64] D. Donoho 和 P. Stark. 不确定性原理与信号恢复。*SIAM J. on Appl. Math.*，49(3):906–931，1989。

[65] D. Donoho 和 V. Stodden. 非负矩阵分解何时能给出正确的部分分解？在 *NIPS*，2003。

[66] R. Downey 和 M. Fellows. *Parameterized complexity* Springer，2012。

[67] M. Elad. *Sparse and Redundant Representations*. Springer，2010。

[68] K. Engan, S. Aase 和 J. Hakon-Husoy. 框架设计的最优方向法。*ICASSP*，5:2443–2446，1999。

[69] P. Erdos, M. Steel, L. Szekely 和 T. Warnow. 几个日志足以构建（几乎）所有树。I. *Random Structures and Algorithms* 14:153-184, 1997. [70] M. Fazel. *Matrix Rank Minimization with Applications*。博士论文，斯坦福大学，2002。 [71] U. Feige 和 J. Kilian. 半随机图问题的启发式算法。*JCSS*，第 639–671 页，2001。 [72] U. Feige 和 R. Krauthgamer。在半随机图中找到和证明一个大隐藏团。*Random Structures and Algorithms*，第 195–208 页，2009。 [73] J. Feldman，R. A. Servedio 和 R. O’Donnell。无分离假设下 PAC 学习轴对齐高斯混合。在 *COLT* 中，第 20–34 页，2006。 [74] A. Frieze，M. Jerrum，R. Kannan。学习线性变换。在 *FOCS* 中，第 359–368 页，1996。

[75] A. Garnaev 和 E. Gluskin. 欧几里得球的宽度。*Sovieth Math. Dokl.*，第 200–204 页，1984 年。

[76] R. Ge 和 T. Ma. 使用平方和算法分解过完备的三阶张量。在 *RANDOM*，第 829–849 页，2015 年。

[77] A. Gilbert, S. Muthukrishnan 和 M. Strauss. 使用一致性近似冗余字典上的函数。在 *SODA*，2003。[78] N. Gillis. 线性规划模型 hotttopixx 的鲁棒性分析，用于分解非负矩阵。*arXiv:1211.6687*，2012。[79] N. Goyal, S. Vempala 和 Y. Xiao. 傅里叶 PCA。在 *STOC*，2014。[80] D. Gross. 从任何基中从少数系数中恢复低秩矩阵。*arXiv:0910.1879*，2009。[81] D. Gross, Y-K Liu, S. Flammia, S. Becker 和 J. Eisert. 通过压缩感知进行量子态层析成像。*Physical Review Letters*，105(15)，2010。[82] V. Guruswami, J. Lee 和 A. Razborov. 通过扩张码通过 ℓ_1^n 的几乎欧几里得子空间。*Combinatorica*，30(1):47–68，2010。[83] M. Hardt. 理解交替最小化在矩阵补全中的应用。*FOCS*，第 651–660 页，2014。[84] R. Harshman. PARFAC 程序的基础：解释性多模式因子分析的模式和条件。*UCLA Working Papers in Phonetics*，第 1–84 页，1970。[85] J. Håstad. 张量秩是 NP-完备的。*Journal of Algorithms*，11(4):644-654，1990。[86] C. Hillar 和 L-H. Lim. 大多数张量问题都是 NP-难。*arXiv:0911.1393v4*，2013

[87] T. Hofmann. 概率潜在语义分析。在 UAI，第 289–296 页，1999。

[88] R. Horn 和 C. Johnson. *Matrix Analysis*. 剑桥大学出版社，1990年。

[89] Hsu D. 和 Kakade S. 学习球面高斯混合：矩方法和谱分解。在 $ITCS$，第 11–20 页，2013 年。

[90] P. J. Huber. 投影寻踪。*Annals of Statistics* 13:435–475，1985。

[91] R. A. Hummel 和 B. C. Gidas. 零交叉和热方程。*Courant Institute of Mathematical Sciences* TR-111，1984。[92] R. Impagliazzo 和 R. Paturi。关于 k-SAT 的复杂性。*J. Computer and System Sciences* 62(2):pp. 367–375，2001。[93] P. Jain，P. Netrapalli 和 S. Sanghavi。使用交替最小化进行低秩矩阵补全。$STOC$，第 665-674 页，2013。[94] A. T. Kalai，A. Moitra 和 G. Valiant。高效学习两个高斯混合。在 $STOC$ 中，第 553-562 页，2010。[95] R. Karp。一些组合搜索问题的概率分析。在 *Algorithms and Complexity: New Directions and Recent Results* 中，第 1-19 页，1976。[96] B. Kashin 和 V. Temlyakov。关于压缩感知的一个注释。手稿，2007。[97] L. Khachiyan。在矩阵中近似极值行列式的复杂性。*Journal of Complexity*，第 138-153 页，1995。

[98] D. Koller 和 N. Friedman. *Probabilistic Graphical Models*. MIT Press, 2009. [99] J. Kruskal. 三维数组：三线性分解的秩和唯一性及其在算术复杂性和统计学中的应用。*Linear Algebra and its Applications*，第 95-138 页，1997。 [100] A. Kumar、V. Sindhwani 和 P. Kambadur。快速锥形 hull 算法用于近似可分离非负矩阵分解。在 *ICML* 中，第 231-239 页，2013。 [101] D. Lee 和 H. Seung。通过非负矩阵分解学习物体的部分。*Nature*，第 788-791 页，1999。 [102] D. Lee 和 H. Seung。非负矩阵分解的算法。在 *NIPS* 中，第 556-562 页，2000。 [103] S. Leurgans、R. Ross 和 R. Abel。三维数组的分解。*SIAM Journal on Matrix Analysis and Applications*，第 14(4):1064–1083，1993。 [104] M. Lewicki 和 T. Sejnowski。学习超完备表示。*Neural Computation*，第 12:337–365，2000。 [105] W. Li 和 A. McCallum。Pachinko 分配：主题相关性的 DAG 结构混合模型。*ICML*，第 633-640 页，2007。 [106] B. Lindsay。*Mixture Models: Theory, Geometry and Applications*。数学统计研究所，1995。 [107] B. F. Logan。*Properties of High-Pass Signals*。博士论文，哥伦比亚大学，1965。

[108] L. Lovász和M. Saks. 通信复杂性与组合格理论。
Journal of Computer and System Sciences，第322–349页，1993年。

[109] F. McSherry. 随机图的谱划分。在 *FOCS*，第 529–537 页，2001 年。

[110] S. Mallat. *A Wavelet Tour of Signal Processing*. 学术出版社，1998年。

[111] S. Mallat 和 Z. Zhang. 基于时频字典的匹配追踪
IEEE Trans. on Signal Processing，41(12)：3397–3415，1993。[112] A. Moitra. 计算非负秩的几乎最优算法。在 *SODA*，第 1454–1464 页，2013。[113] A. Moitra. 超分辨率，极值函数和 Vandermonde 矩阵的条件数。在 *STOC*，第 821–830 页，2015。[114] A. Moitra 和 G. Valiant. 设置高斯混合的多项式可学习性。在 *FOCS*，第 93–102 页，2010。[115] E. Mossel 和 S. Roch. 学习非奇异系统发育树和隐马尔可夫模型。在 *STOC*，第 366–375 页，2005。

[116] Y. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Springer，2004。

[117] B. Olshausen 和 B. Field. 使用过完备基集的稀疏编码：V1 使用的策略?
Vision Research，37(23)：331–3325，1997。

[118] C. Papadimitriou, P. Raghavan, H. Tamaki 和 S. Vempala. 隐语义索引：概率分析。*JCSS*，第 217–235 页，2000 年。

[119] Y. Pati, R. Rezaiifar, P. Krishnaprasad. 正交匹配追踪：递归函数逼近及其在小波分解中的应用。*Asilo- mar Conference on Signals, Systems and Computers*，第40-44页，1993年。[120] K. Pearson. 对进化数学理论的贡献。*Philosophical Transactions of the Royal Society A*，1894年。[121] Y. Rabani，L. Schulman和C. Swamy。在大离散域上学习任意分布的混合。在*ITCS* 2014年。[122] R. Raz. 张量秩和算术公式的下界。在*STOC*，第659-666页，2010年。[123] B. Recht。矩阵补全的简单方法。*Journal of Machine Learning Research*，第3413-3430页，2011年。[124] B. Recht，M. Fazel和P. Parrilo。通过核范数最小化保证矩阵方程的最小秩解。*SIAM Review*，第471-501页，2010年。[125] R. A. Redner和H. F. Walker。混合密度、最大似然估计和EM算法。*SIAM Review*，26(2)：195-239，1984年。[126] J. Renegar。一阶实数理论的计算复杂性和几何。符号计算杂志，第255-352页，1991年。[127] T. Rockefellar。*Convex Analysis*。普林斯顿大学出版社，1996年。[128] A. Seidenberg。初等代数的新决策方法。*Annals of Math*，第365-374页，1954年。

[129] V. de Silva 和 L-H Lim. 张量秩与最佳低秩逼近问题的病态性。*SIAM Journal on Matrix Analysis and Applications*，30(3)：1084–1127，2008。[130] D. Spielman 和 S-H Teng。算法的平滑分析：为什么单纯形算法通常需要多项式时间。在*Journal of the ACM*，51(3)：385–463，2004。[131] D. Spielman，H. Wang 和 J. Wright。稀疏字典的精确恢复。*Journal of Machine Learning Research*，2012。[132] N. Srebro 和 A. Shraibman。秩、迹范和最大范。在*COLT*，第545–560页，2005。[133] M. Steel。从马尔可夫模型下生成的叶颜色中恢复树。*Appl. Math. Lett.* 7：19-24，1994。[134] A. Tarski。初等代数和几何的决策方法。*University of California Press*，1951。[135] H. Teicher。混合物的可识别性。*Annals of Mathematical Statistics*，第244–248页，1961。[136] J. Tropp。贪婪是好：稀疏逼近的算法结果。*IEEE Trans. on IT*，50(10)：2231–2242，2004。[137] J. Tropp，A. Gilbert，S. Muthukrishnan 和 M. Strauss。在准非相关字典上的改进稀疏逼近。*IEEE International Conf. on Image Processing*，2003。[138] L. Valiant。可学习理论。*Comm. ACM*，27(11)：1134–1142，1984。

[139] S. Vavasis. 非负矩阵分解的复杂性。*SIAM Journal on Optimization*，第1364-1377页，2009年。[140] S. Vempala，Y. Xiao。从局部最优解中提取结构：通过高阶PCA学习子空间联合。*arXiv:*abs/1108.3329，2011年。[141] S. Vempala和G. Wang。学习混合模型的谱算法。*Journal of Computer and System Sciences*，第841-860页，2004年。[142] M. Wainwright和M. Jordan。
Graphical Models, Exponential Families, and Variational Inference. Foundations and Trends in Machine Learning，第1-305页，2008年。[143] P. Wedin。与奇异值分解相关的扰动界限。*BIT*，第12卷，第99-111页，1972年。[144] M. Yannakakis。用线性规划表达组合优化问题。*Journal of Computer and System Sciences*，第441-466页，1991年。

Index